Introduction to Statistics
Class 10, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the three overlapping “phases” of statistical practice.

2. Know what is meant by the term *statistic*.

2 Introduction to statistics

Statistics deals with data. Generally speaking, the goal of statistics is to make inferences based on data. We can divide this the process into three phases: collecting data, describing data and analyzing data. This fits into the paradigm of the scientific method. We make hypotheses about what’s true, collect data in experiments, describe the results, and then infer from the results the strength of the evidence concerning our hypotheses.

2.1 Experimental design

The design of an experiment is crucial to making sure the collected data is useful. The adage ‘garbage in, garbage out’ applies here. A poorly designed experiment will produce poor quality data, from which it may be impossible to draw useful, valid inferences. To quote R.A. Fisher one of the founders of modern statistics,

> To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.

2.2 Descriptive statistics

Raw data often takes the form of a massive list, array, or database of labels and numbers. To make sense of the data, we can calculate summary statistics like the mean, median, and interquartile range. We can also visualize the data using graphical devices like histograms, scatterplots, and the empirical cdf. These methods are useful for both communicating and exploring the data to gain insight into its structure, such as whether it might follow a familiar probability distribution.

2.3 Inferential statistics

Ultimately we want to draw inferences about the world. Often this takes the form of specifying a statistical model for the random process by which the data arises. For example, suppose the data takes the form of a series of measurements whose error we believe follows a normal distribution. (Note this is always an approximation since we know the error must have some bound while a normal distribution has range $(-\infty, \infty)$.) We might then use the data to provide evidence for or against this hypothesis. Our focus in 18.05 will be on how to use data to draw inferences about model parameters. For example, assuming gestational length follows a $N(\mu, \sigma)$ distribution, we'll use the data of the gestational lengths of, say, 500 pregnancies to draw inferences about the values of the parameters μ and σ. Similarly, we may model the result of a two-candidate election by a Bernoulli(p) distribution, and use poll data to draw inferences about the value of p.

We can rarely make definitive statements about such parameters because the data itself comes from a random process (such as choosing who to poll). Rather, our statistical evidence will always involve probability statements. Unfortunately, the media and public at large are wont to misunderstand the probabilistic meaning of statistical statements. In fact, researchers themselves often commit the same errors. In this course, we will emphasize the meaning of statistical statements alongside the methods which produce them.

Example 1. To study the effectiveness of new treatment for cancer, patients are recruited and then divided into an experimental group and a control group. The experimental group is given the new treatment and the control group receives the current standard of care. Data collected from the patients might include demographic information, medical history, initial state of cancer, progression of the cancer over time, treatment cost, and the effect of the treatment on tumor size, remission rates, longevity, and quality of life. The data will be used to make inferences about the effectiveness of the new treatment compared to the current standard of care.

Notice that this study will go through all three phases described above. The experimental design must specify the size of the study, who will be eligible to join, how the experimental and control groups will be chosen, how the treatments will be administered, whether or not the subjects or doctors know who is getting which treatment, and precisely what data will be collected, among other things. Once the data is collected it must be described and analyzed to determine whether it supports the hypothesis that the new treatment is more (or less) effective than the current one(s), and by how much. These statistical conclusions will be framed as precise statements involving probabilities.

As noted above, misinterpreting the exact meaning of statistical statements is a common source of error which has led to tragedy on more than one occasion.

Example 2. In 1999 in Great Britain, Sally Clark was convicted of murdering her two children after each child died weeks after birth (the first in 1996, the second in 1998). Her conviction was largely based on a faulty use of statistics to rule out sudden infant death syndrome. Though her conviction was overturned in 2003, she developed serious psychiatric problems during and after her imprisonment and died of alcohol poisoning in 2007. See https://en.wikipedia.org/wiki/Sally_Clark

This TED talk discusses the Sally Clark case and other instances of poor statistical intuition: https://www.youtube.com/watch?v=kLmzxmRcUTo

2.4 What is *a* statistic?

We give a simple definition whose meaning is best elucidated by examples.

Definition. A statistic is anything that can be computed from the collected data.

Example 3. Consider the data of 1000 rolls of a die. All of the following are statistics: the average of the 1000 rolls; the number of times a 6 was rolled; the sum of the squares of the rolls minus the number of even rolls. It's hard to imagine how we would use the last example, but it is a statistic. On the other hand, the probability of rolling a 6 is *not* a statistic, whether or not the die is truly fair. Rather this probability is a property of the die (and the way we roll it) which we can estimate using the data. Such an estimate is given by the statistic 'proportion of the rolls that were 6'.

Example 4. Suppose we treat a group of cancer patients with a new procedure and collect data on how long they survive post-treatment. From the data we can compute the average survival time of patients in the group. We might employ this statistic as an estimate of the average survival time for future cancer patients following the new procedure. The "expected survival time" for the new procedure (if that even has a meaning) is *not* a statistic.

Example 5. Suppose we ask 1000 residents whether or not they support the proposal to legalize marijuana in Massachusetts. The proportion of the 1000 who support the proposal is a statistic. The proportion of all Massachusetts residents who support the proposal is *not* a statistic since we have not queried every single one (note the word "collected" in the definition). Rather, we hope to draw a statistical conclusion about the state-wide proportion based on the data of our random sample.

The following are two general types of statistics we will use in 18.05.

1. Point statistics: a single value computed from data, such as the sample average $\overline{x}_n$ or the sample standard deviation s_n.
2. Interval statistics: an interval $[a, b]$ computed from the data. This is really just a pair of point statistics, and will often be presented in the form $\overline{x} \pm s$.

3 Review of Bayes' theorem

We cannot stress strongly enough how important Bayes' theorem is to our view of inferential statistics. Recall that Bayes' theorem allows us to 'invert' conditional probabilities. That is, if H and D are events, then Bayes' theorem says

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}.$$

In scientific experiments we start with a hypothesis and collect data to test the hypothesis. We will often let H represent the event 'our hypothesis is true' and let D be the collected data. In these words Bayes' theorem says

$$P(\text{hypothesis is true} \mid \text{data}) = \frac{P(\text{data } | \text{hypothesis is true}) \cdot P(\text{hypothesis is true})}{P(\text{data})}$$

The left-hand term is the probability our hypothesis is true given the data we collected. This is precisely what we'd like to know. When all the probabilities on the right are known exactly, we can compute the probability on the left exactly. This will be our focus next week. Unfortunately, in practice we rarely know the exact values of all the terms on the right. Statisticians have developed a number of ways to cope with this lack of knowledge and still make useful inferences. We will be exploring these methods for the rest of the course.

Example 6. Screening for a disease redux
Suppose a screening test for a disease has a 1% false positive rate and a 1% false negative rate. Suppose also that the rate of the disease in the population is 0.002. Finally suppose a randomly selected person tests positive. In the language of hypothesis and data we have:
Hypothesis: H = 'the person has the disease'
Data: D = 'the test was positive.'
What we want to know: $P(H|D) = P(\text{the person has the disease} \mid \text{a positive test})$

In this example all the probabilities on the right are known so we can use Bayes' theorem to compute what we want to know.

$$\begin{aligned}
P(\text{hypothesis} \mid \text{data}) &= P(\text{the person has the disease} \mid \text{a positive test}) \\
&= P(H|D) \\
&= \frac{P(D|H)P(H)}{P(D)} \\
&= \frac{0.99 \cdot 0.002}{0.99 \cdot 0.002 + 0.01 \cdot 0.998} \\
&= 0.166
\end{aligned}$$

Before the test we would have said the probability the person had the disease was 0.002. After the test we see the probability is 0.166. That is, the positive test provides some evidence that the person has the disease.

Maximum Likelihood Estimates
Class 10, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to define the likelihood function for a parametric model given data.
2. Be able to compute the maximum likelihood estimate of unknown parameter(s).

2 Introduction

Suppose we know we have data consisting of values $x_1, \dots, x_n$ drawn from an exponential distribution. The question remains: which exponential distribution?!

We have casually referred to *the* exponential distribution or *the* binomial distribution or *the* normal distribution. In fact the exponential distribution $\exp(\lambda)$ is not a single distribution but rather a one-parameter family of distributions. Each value of λ defines a different distribution in the family, with pdf $f_\lambda(x) = \lambda e^{-\lambda x}$ on $[0, \infty)$. Similarly, a binomial distribution $\text{bin}(n, p)$ is determined by the two parameters n and p, and a normal distribution $N(\mu, \sigma^2)$ is determined by the two parameters μ and σ^2 (or equivalently, μ and σ). Parameterized families of distributions are often called parametric distributions or parametric models.

We are often faced with the situation of having random data which we know (or believe) is drawn from a parametric model, whose parameters we do not know. For example, in an election between two candidates, polling data constitutes draws from a Bernoulli(p) distribution with unknown parameter p. In this case we would like to use the data to estimate the value of the parameter p, as the latter predicts the result of the election. Similarly, assuming gestational length follows a normal distribution, we would like to use the data of the gestational lengths from a random sample of pregnancies to draw inferences about the values of the parameters μ and σ^2.

Our focus so far has been on computing the probability of data arising from a parametric model with known parameters. Statistical inference flips this on its head: we will estimate the probability of parameters given a parametric model and observed data drawn from it. In the coming weeks we will see how parameter values are naturally viewed as hypotheses, so we are in fact estimating the probability of various hypotheses given the data.

3 Maximum Likelihood Estimates

There are many methods for estimating unknown parameters from data. We will first consider the maximum likelihood estimate (MLE), which answers the question:

> For which parameter value does the observed data have the biggest probability?

The MLE is an example of a point estimate because it gives a single value for the unknown parameter (later our estimates will involve intervals and probabilities). Two advantages of the MLE are that it is often easy to compute and that it agrees with our intuition in simple examples. We will explain the MLE through a series of examples.

Example 1. A coin is flipped 100 times. Given that there were 55 heads, find the maximum likelihood estimate for the probability p of heads on a single toss.

Before actually solving the problem, let's establish some notation and terms.

We can think of counting the number of heads in 100 tosses as an experiment. For a given value of p, the probability of getting 55 heads in this experiment is the binomial probability

$$P(55 \text{ heads}) = \binom{100}{55} p^{55}(1-p)^{45}.$$

The probability of getting 55 heads depends on the value of p, so let's include p in by using the notation of conditional probability:

$$P(55 \text{ heads} \,|\, p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

You should read $P(55 \text{ heads} \,|\, p)$ as:

'the probability of 55 heads given p,'

or more precisely as

'the probability of 55 heads given that the probability of heads on a single toss is p.'

Here are some standard terms we will use as we do statistics.

- Experiment: Flip the coin 100 times and count the number of heads.
- Data: The data is the result of the experiment. In this case it is '55 heads'.
- Parameter(s) of interest: We are interested in the value of the unknown parameter p.
- Likelihood, or likelihood function: this is $P(\text{data} \,|\, p)$. Note it is a function of both the data and the parameter p. In this case the likelihood is

  $$P(55 \text{ heads} \,|\, p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

Notes: **1.** The likelihood $P(\text{data } |\, p)$ changes as the parameter of interest p changes.

2. Look carefully at the definition. One typical source of confusion is to mistake the likelihood $P(\text{data } |\, p)$ for $P(p \,|\, \text{data})$. We know from our earlier work with Bayes' theorem that $P(\text{data} \,|\, p)$ and $P(p \,|\, \text{data})$ are usually very different.

Definition: Given data the maximum likelihood estimate (MLE) for the parameter p is the value of p that maximizes the likelihood $P(\text{data } |\, p)$. That is, the MLE is the value of p for which the data is most likely.

Solution: For the problem at hand, we saw above that the likelihood

$$P(55 \text{ heads} \,|\, p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

We'll use the notation $\hat{p}$ for the MLE. We use calculus to find it by taking the derivative of the likelihood function and setting it to 0.

$$\frac{d}{dp}P(\text{data }|p) = \binom{100}{55}(55p^{54}(1-p)^{45} - 45p^{55}(1-p)^{44}) = 0.$$

Solving this for p we get

$$\begin{aligned}
&55p^{54}(1-p)^{45} = 45p^{55}(1-p)^{44} \\
&55(1-p) = 45p \\
&55 = 100p \\
&\text{the MLE is } \hat{p} = 0.55
\end{aligned}$$

Note: **1.** The MLE for p turned out to be exactly the fraction of heads we saw in our data.

2. The MLE is computed from the data. That is, it is a statistic.

3. Officially we need to check that this critical point is actually the maximum. We could use the second derivative test. Another way is to notice that we are interested only in $0 \le p \le 1$; that the probability is bigger than zero for $0 < p < 1$; and that the probability is equal to zero for $p = 0$ and for $p = 1$. From these facts it follows that the critical point must be the unique maximum.

3.1 Log likelihood

If is often easier to work with the natural log of the likelihood function. For short this is simply called the log likelihood. Since $\ln(x)$ is an increasing function, the maxima of the likelihood and log likelihood coincide.

Example 2. Redo the previous example using log likelihood.

Solution: We had the likelihood $P(55 \text{ heads} \,|\, p) = \binom{100}{55}p^{55}(1-p)^{45}$. Therefore the log likelihood is

$$\ln(P(55 \text{ heads} \,|\, p) = \ln\left(\binom{100}{55}\right) + 55\ln(p) + 45\ln(1-p).$$

Maximizing likelihood is the same as maximizing log likelihood. We check that calculus gives us the same answer as before:

$$\begin{aligned}
\frac{d}{dp}(\text{log likelihood}) &= \frac{d}{dp}\left[\ln\left(\binom{100}{55}\right) + 55\ln(p) + 45\ln(1-p)\right] \\
&= \frac{55}{p} - \frac{45}{1-p} = 0 \\
&\Rightarrow 55(1-p) = 45p \\
&\Rightarrow \hat{p} = 0.55
\end{aligned}$$

3.2 Maximum likelihood for continuous distributions

For continuous distributions, we use the probability density function to define the likelihood. We show this in a few examples. In the next section we explain how this is analogous to what we did in the discrete case.

Example 3. Light bulbs
Suppose that the lifetime of *Badger* brand light bulbs is modeled by an exponential distribution with (unknown) parameter λ. We test 5 bulbs and find they have lifetimes of 2, 3, 1, 3, and 4 years, respectively. What is the MLE for λ?

Solution: We need to be careful with our notation. With five different values it is best to use subscripts. Let X_i be the lifetime of the i^{th} bulb and let x_i be the value X_i takes. Then each X_i has pdf $f_{X_i}(x_i) = \lambda \mathrm{e}^{-\lambda x_i}$. We assume the lifetimes of the bulbs are independent, so the joint pdf is the product of the individual densities:

$$f(x_1, x_2, x_3, x_4, x_5 \mid \lambda) = (\lambda \mathrm{e}^{-\lambda x_1})(\lambda \mathrm{e}^{-\lambda x_2})(\lambda \mathrm{e}^{-\lambda x_3})(\lambda \mathrm{e}^{-\lambda x_4})(\lambda \mathrm{e}^{-\lambda x_5}) = \lambda^5 \mathrm{e}^{-\lambda(x_1+x_2+x_3+x_4+x_5)}.$$

Note that we write this as a conditional density, since it depends on λ. Viewing the data as fixed and λ as variable, this density is the likelihood function. Our data had values

$$x_1 = 2,\ x_2 = 3,\ x_3 = 1,\ x_4 = 3,\ x_5 = 4.$$

So the likelihood and log likelihood functions with this data are

$$f(2, 3, 1, 3, 4 \mid \lambda) = \lambda^5 \mathrm{e}^{-13\lambda}, \quad \ln(f(2, 3, 1, 3, 4 \mid \lambda) = 5\ln(\lambda) - 13\lambda$$

Finally we use calculus to find the MLE:

$$\frac{d}{d\lambda}(\text{log likelihood}) = \frac{5}{\lambda} - 13 = 0 \ \Rightarrow \ \boxed{\hat{\lambda} = \frac{5}{13}}.$$

Note: **1.** In this example we used an uppercase letter for a random variable and the corresponding lowercase letter for the value it takes. This will be our usual practice.

2. The MLE for λ turned out to be the reciprocal of the sample mean $\bar{x}$, so $X \sim \exp(\hat{\lambda})$ satisfies $E[X] = \bar{x}$.

The following example illustrates how we can use the method of maximum likelihood to estimate multiple parameters at once.

Example 4. Normal distributions
Suppose the data $x_1, x_2, \ldots, x_n$ is drawn from a $\mathrm{N}(\mu, \sigma^2)$ distribution, where μ and σ are unknown. Find the maximum likelihood estimate for the pair (μ, σ^2).

Solution: Let's be precise and phrase this in terms of random variables and densities. Let uppercase $X_1, \ldots, X_n$ be i.i.d. $\mathrm{N}(\mu, \sigma^2)$ random variables, and let lowercase x_i be the value X_i takes. The density for each X_i is

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Since the X_i are independent their joint pdf is the product of the individual pdf's:

$$f(x_1, \ldots, x_n \mid \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \mathrm{e}^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}.$$

For the fixed data $x_1, \ldots, x_n$, the likelihood and log likelihood are

$$f(x_1, \ldots, x_n|\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \mathrm{e}^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}, \quad \ln(f(x_1, \ldots, x_n|\mu,\sigma)) = -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}.$$

Since $\ln(f(x_1, \ldots, x_n|\mu,\sigma))$ is a function of the two variables μ, σ we use partial derivatives to find the MLE. The easy value to find is $\hat{\mu}$:

$$\frac{\partial f(x_1, \ldots, x_n|\mu,\sigma)}{\partial \mu} = \sum_{i=1}^n \frac{(x_i-\mu)}{\sigma^2} = 0 \;\Rightarrow\; \sum_{i=1}^n x_i = n\mu \;\Rightarrow\; \hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \overline{x}.$$

To find $\hat{\sigma}$ we differentiate and solve for σ:

$$\frac{\partial f(x_1, \ldots, x_n|\mu,\sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^n \frac{(x_i-\mu)^2}{\sigma^3} = 0 \;\Rightarrow\; \hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i-\mu)^2}{n}.$$

We already know $\hat{\mu} = \overline{x}$, so we use that as the value for μ in the formula for $\hat{\sigma}$. We get the maximum likelihood estimates

$$\begin{aligned} \hat{\mu} &= \overline{x} && = \text{the mean of the data} \\ \hat{\sigma}^2 &= \sum_{i=1}^n \frac{1}{n}(x_i-\hat{\mu})^2 = \sum_{i=1}^n \frac{1}{n}(x_i-\overline{x})^2 && = \text{the unadjusted variance of the data.} \end{aligned}$$

(Later we will learn that the sample variance is $\dfrac{\sum_{i=1}^n (x_i-\hat{\mu})^2}{n-1}$.)

Example 5. Uniform distributions
Suppose our data $x_1, \ldots x_n$ are independently drawn from a uniform distribution $U(a,b)$. Find the MLE for a and b.

Solution: This example is different from the previous ones in that we won't use calculus to find the MLE. The density for $U(a,b)$ is $\frac{1}{b-a}$ on $[a,b]$. Therefore our likelihood function is

$$f(x_1, \ldots, x_n \,|\, a, b) = \begin{cases} \left(\frac{1}{b-a}\right)^n & \text{if all } x_i \text{ are in the interval } [a,b] \\ 0 & \text{otherwise.} \end{cases}$$

This is maximized by making $b-a$ as small as possible. The only restriction is that the interval $[a,b]$ must include all the data. Thus the MLE for the pair (a,b) is

$$\hat{a} = \min(x_1, \ldots, x_n) \qquad \hat{b} = \max(x_1, \ldots, x_n).$$

Example 6. Capture/recapture method

The capture/recapture method is a way to estimate the size of a population in the wild. The method assumes that each animal in the population is equally likely to be captured by a trap.

Suppose 10 animals are captured, tagged and released. A few months later, 20 animals are captured, examined, and released. 4 of these 20 are found to be tagged. Estimate the size of the wild population using the MLE for the probability that a wild animal is tagged.

Solution: Our unknown parameter n is the number of animals in the wild. Our data is that 4 out of 20 recaptured animals were tagged (and that there are 10 tagged animals). The likelihood function is

$$P(\text{data} \,|\, n \text{ animals}) = \frac{\binom{n-10}{16}\binom{10}{4}}{\binom{n}{20}}$$

(The numerator is the number of ways to choose 16 animals from among the $n-10$ untagged ones times the number of was to choose 4 out of the 10 tagged animals. The denominator is the number of ways to choose 20 animals from the entire population of n.) We can use R to compute that the likelihood function is maximized when $n = 50$. This should make some sense. It says our best estimate is that the fraction of all animals that are tagged is 10/50 which equals the fraction of recaptured animals which are tagged.

Example 7. Hardy-Weinberg. Suppose that a particular gene occurs as one of two alleles (A and a), where allele A has frequency θ in the population. That is, a random copy of the gene is A with probability θ and a with probability $1-\theta$. Since a diploid genotype consists of two genes, the probability of each genotype is given by:

genotype	AA	Aa	aa
probability	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$

Suppose we test a random sample of people and find that k_1 are AA, k_2 are Aa, and k_3 are aa. Find the MLE of θ.

Solution: The likelihood function is given by

$$P(k_1, k_2, k_3 \,|\, \theta) = \binom{k_1+k_2+k_3}{k_1}\binom{k_2+k_3}{k_2}\binom{k_3}{k_3}\theta^{2k_1}(2\theta(1-\theta))^{k_2}(1-\theta)^{2k_3}.$$

So the log likelihood is given by

$$\text{constant} + 2k_1\ln(\theta) + k_2\ln(\theta) + k_2\ln(1-\theta) + 2k_3\ln(1-\theta)$$

We set the derivative equal to zero:

$$\frac{2k_1+k_2}{\theta} - \frac{k_2+2k_3}{1-\theta} = 0$$

Solving for θ, we find the MLE is

$$\hat{\theta} = \frac{2k_1+k_2}{2k_1+2k_2+2k_3},$$

which is simply the fraction of A alleles among all the genes in the sampled population.

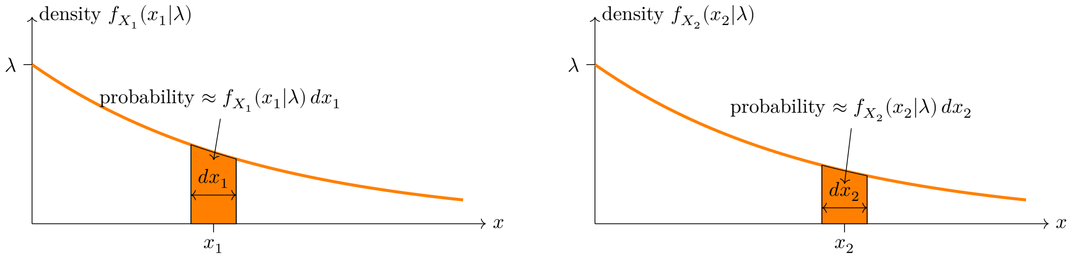
4 Why we use the density to find the MLE for continuous distributions

The idea for the maximum likelihood estimate is to find the value of the parameter(s) for which the data has the highest probability. In this section we 'll see that we're doing this is really what we are doing with the densities. We will do this by considering a smaller version of the light bulb example.

Example 8. Suppose we have two light bulbs whose lifetimes follow an exponential(λ) distribution. Suppose also that we independently measure their lifetimes and get data $x_1 = 2$ years and $x_2 = 3$ years. Find the value of λ that maximizes the probability of this data.

Solution: The main paradox to deal with is that for a continuous distribution the probability of a single value, say $x_1 = 2$, is zero. We resolve this paradox by remembering that a single measurement really means a range of values, e.g. in this example we might check the light bulb once a day. So the data $x_1 = 2$ years really means x_1 is somewhere in a range of 1 day around 2 years.

If the range is small we call it dx_1. The probability that X_1 is in the range is approximated by $f_{X_1}(x_1|\lambda)\,dx_1$. This is illustrated in the figure below. The data value x_2 is treated in exactly the same way.



The usual relationship between density and probability for small ranges.

Since the data is collected independently the joint probability is the product of the individual probabilities. Stated carefully

$$P(X_1 \text{ in range, } X_2 \text{ in range}|\lambda) \approx f_{X_1}(x_1|\lambda)\,dx_1 \cdot f_{X_2}(x_2|\lambda)\,dx_2$$

Finally, using the values $x_1 = 2$ and $x_2 = 3$ and the formula for an exponential pdf we have

$$P(X_1 \text{ in range, } X_2 \text{ in range}|\lambda) \approx \lambda e^{-2\lambda}\,dx_1 \cdot \lambda e^{-3\lambda}\,dx_2 = \lambda^2 e^{-5\lambda}\,dx_1\,dx_2.$$

Now that we have a genuine probability we can look for the value of λ that maximizes it. Looking at the formula above we see that the factor $dx_1\,dx_2$ will play no role in finding the maximum. So for the MLE we drop it and simply call the density the likelihood:

$$\text{likelihood } = f(x_1, x_2|\lambda) = \lambda^2 e^{-5\lambda}.$$

The value of λ that maximizes this is found just like in the example above. It is $\hat{\lambda} = 2/5$.

5 Appendix: Properties of the MLE

For the interested reader, we note several nice features of the MLE. These are quite technical and will not be on any exams.

The MLE behaves well under transformations. That is, if $\hat{p}$ is the MLE for p and g is a one-to-one function, then $g(\hat{p})$ is the MLE for $g(p)$. For example, if $\hat{\sigma}$ is the MLE for the standard deviation σ then $(\hat{\sigma})^2$ is the MLE for the variance σ^2.

Furthermore, under some technical smoothness assumptions, the MLE is asymptotically unbiased and has asymptotically minimal variance. To explain these notions, note that the MLE is itself a random variable since the data is random and the MLE is computed from the data. Let $x_1, x_2, \ldots$ be an infinite sequence of samples from a distribution with parameter p. Let $\hat{p}_n$ be the MLE for p based on the data $x_1, \ldots, x_n$.

Asymptotically unbiased means that as the amount of data grows, the mean of the MLE converges to p. In symbols: $E[\hat{p}_n] \to p$ as $n \to \infty$. Of course, we would like the MLE to be close to p with high probability, not just on average, so the smaller the variance of the MLE the better. Asymptotically minimal variance means that as the amount of data grows, the MLE has the minimal variance among all unbiased estimators of p. In symbols: for any unbiased estimator $\tilde{p}_n$ and $\epsilon > 0$ we have that $\mathrm{Var}(\tilde{p}_n) + \epsilon > \mathrm{Var}(\hat{p}_n)$ as $n \to \infty$.

Bayesian Updating with Discrete Priors
Class 11, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to apply Bayes' theorem to compute probabilities.
2. Be able to define and to identify the roles of prior probability, likelihood (Bayes term), posterior probability, data and hypothesis in the application of Bayes' Theorem.
3. Be able to use a Bayesian update table to compute posterior probabilities.

2 Review of Bayes' theorem

Recall that Bayes' theorem allows us to 'invert' conditional probabilities. If $\mathcal{H}$ and $\mathcal{D}$ are events, then:

$$P(\mathcal{H}\,|\,\mathcal{D}) = \frac{P(\mathcal{D}\,|\,\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}$$

Our view is that Bayes' theorem forms the foundation for inferential statistics. We will begin to justify this view today.

2.1 The base rate fallacy

When we first learned Bayes' theorem we worked an example about screening tests showing that $P(\mathcal{D}|\mathcal{H})$ can be very different from $P(\mathcal{H}|\mathcal{D})$. In the appendix we work a similar example. If you are not comfortable with Bayes' theorem you should read the example in the appendix now.

3 Terminology and Bayes' theorem in tabular form

We now use a coin tossing problem to introduce terminology and a tabular format for Bayes' theorem. This will provide a simple, uncluttered example that shows our main points.

Example 1. There are three types of coins which have different probabilities of landing heads when tossed.

- Type A coins are fair, with probability 0.5 of heads
- Type B coins are bent and have probability 0.6 of heads
- Type C coins are bent and have probability 0.9 of heads

Suppose I have a drawer containing 5 coins: 2 of type A, 2 of type B, and 1 of type C. I reach into the drawer and pick a coin at random. Without showing you the coin I flip it once and get heads. What is the probability it is type A? Type B? Type C?

Solution: Let A, B, and C be the event that the chosen coin was type A, type B, and type C. Let $\mathcal{D}$ be the event that the toss is heads. The problem asks us to find

$$P(A|\mathcal{D}), \quad P(B|\mathcal{D}), \quad P(C|\mathcal{D}).$$

Before applying Bayes' theorem, let's introduce some terminology.

- Experiment: pick a coin from the drawer at random, flip it, and record the result.
- Data: the result of our experiment. In this case the event $\mathcal{D}$ = 'heads'. We think of $\mathcal{D}$ as data that provides evidence for or against each hypothesis.
- Hypotheses: we are testing three hypotheses: the coin is type A, B or C.
- Prior probability: the probability of each hypothesis prior to tossing the coin (collecting data). Since the drawer has 2 coins of type A, 2 of type B and 1 of type C we have

  $$P(A) = 0.4, \qquad P(B) = 0.4, \qquad P(C) = 0.2.$$

- Likelihood: (This is the same likelihood we used for the MLE.) The likelihood function is $P(\mathcal{D}|\mathcal{H})$, i.e., the probability of the data assuming that the hypothesis is true. Most often we will consider the data as fixed and let the hypothesis vary. For example, $P(\mathcal{D}|A)$ = probability of heads if the coin is type A. In our case the likelihoods are

  $$P(\mathcal{D}|A) = 0.5, \qquad P(\mathcal{D}|B) = 0.6, \qquad P(\mathcal{D}|C) = 0.9.$$

  The name likelihood is so well established in the literature that we have to teach it to you. However in colloquial language likelihood and probability are synonyms. This leads to the likelihood function often being confused with the probability of a hypothesis. Because of this we'd prefer to use the name Bayes' term. However since we are stuck with 'likelihood' we will try to use it very carefully and in a way that minimizes any confusion.

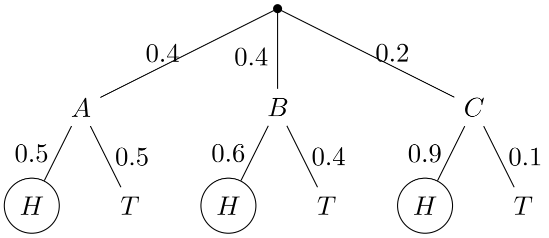
- Posterior probability: the probability (posterior to) of each hypothesis given the data from tossing the coin.

  $$P(A|\mathcal{D}), \qquad P(B|\mathcal{D}), \qquad P(C|\mathcal{D}).$$

  These posterior probabilities are what the problem asks us to find.

We now use Bayes' theorem to compute each of the posterior probabilities. We are going to write this out in complete detail so we can pick out each of the parts (Remember that the data $\mathcal{D}$ is that the toss was heads.)

First we organize the probabilities into a tree:



Probability tree for choosing and tossing a coin.

Bayes' theorem says, e.g. $P(A|\mathcal{D}) = \dfrac{P(\mathcal{D}|A)P(A)}{P(\mathcal{D})}$. The denominator $P(\mathcal{D})$ is computed using the law of total probability:

$$P(\mathcal{D}) = P(\mathcal{D}|A)P(A) + P(\mathcal{D}|B)P(B) + P(\mathcal{D}|C)P(C) = 0.5 \cdot 0.4 + 0.6 \cdot 0.4 + 0.9 \cdot 0.2 = 0.62.$$

Now each of the three posterior probabilities can be computed:

$$\begin{aligned} P(A|\mathcal{D}) &= \frac{P(\mathcal{D}|A)P(A)}{P(\mathcal{D})} = \frac{0.5 \cdot 0.4}{0.62} = \frac{0.2}{0.62} \\ P(B|\mathcal{D}) &= \frac{P(\mathcal{D}|B)P(B)}{P(\mathcal{D})} = \frac{0.6 \cdot 0.4}{0.62} = \frac{0.24}{0.62} \\ P(C|\mathcal{D}) &= \frac{P(\mathcal{D}|C)P(C)}{P(\mathcal{D})} = \frac{0.9 \cdot 0.2}{0.62} = \frac{0.18}{0.62} \end{aligned}$$

Notice that the total probability $P(\mathcal{D})$ is the same in each of the denominators and that it is the sum of the three numerators. We can organize all of this very neatly in a Bayesian update table:

hypothesis	prior	likelihood	Bayes numerator	posterior (numerator/$P(\mathcal{D})$)
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{D})$
A	0.4	0.5	0.2	0.3226
B	0.4	0.6	0.24	0.3871
C	0.2	0.9	0.18	0.2903
total	1	NO SUM	$P(\mathcal{D}) = 0.62$	1

The Bayes numerator is the product of the prior and the likelihood. We see in each of the Bayes' formula computations above that the posterior probability is obtained by dividing the Bayes numerator by $P(\mathcal{D}) = 0.62$. We also see that the law of law of total probability says that $P(\mathcal{D})$ is the sum of the entries in the Bayes numerator column.

Bayesian updating: The process of going from the prior probability $P(\mathcal{H})$ to the posterior $P(\mathcal{H}|\mathcal{D})$ is called Bayesian updating. Bayesian updating uses the data to alter our understanding of the probability of each of the possible hypotheses.

3.1 Important things to notice

1. There are two types of probabilities: Type one is the standard probability of data, e.g. the probability of heads is $p = 0.9$. Type two is the probability of the hypotheses, e.g. the probability the chosen coin is type A, B or C. This second type has prior (before the data) and posterior (after the data) values.

2. The posterior (after the data) probabilities for each hypothesis are in the last column. We see that coin B is now the most probable, though its probability has decreased from a prior probability of 0.4 to a posterior probability of 0.39. Meanwhile, the probability of type C has increased from 0.2 to 0.29.

3. The Bayes numerator column determines the posterior probability column. To compute the latter, we simply divided each numerator by $P(\mathcal{D})$, i.e. rescaled the Bayes numerators so that they sum to 1.

4. If all we care about is finding the most likely hypothesis, the Bayes numerator works as well as the normalized posterior.

5. The likelihood column does not sum to 1. The likelihood function is *not* a probability function.

6. The posterior probability represents the outcome of a 'tug-of-war' between the likelihood and the prior. When calculating the posterior, a large prior may be deflated by a small likelihood, and a small prior may be inflated by a large likelihood.

7. The maximum likelihood estimate (MLE) for Example 1 is hypothesis C, with a likelihood $P(\mathcal{D}|C) = 0.9$. The MLE is useful, but you can see in this example that it is not the entire story, since type B has the greatest posterior probability.

Terminology in hand, we can express Bayes' theorem in various ways:

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}$$

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})}$$

With the data fixed, the denominator $P(\mathcal{D})$ just serves to normalize the total posterior probability to 1. So we can also express Bayes' theorem as a statement about the proportionality of two functions of $\mathcal{H}$ (i.e, of the last two columns of the table).

$$P(\text{hypothesis}|\text{data}) \;\propto\; P(\text{data}|\text{hypothesis})P(\text{hypothesis})$$

This leads to the most elegant form of Bayes' theorem in the context of Bayesian updating:

$$\boxed{\text{posterior} \;\propto\; \text{likelihood} \times \text{prior}}$$

3.2 Prior and posterior probability mass functions

Earlier in the course we saw that it is convenient to use random variables and probability mass functions. To do this we had to assign values to events (head is 1 and tails is 0). We will do the same thing in the context of Bayesian updating.

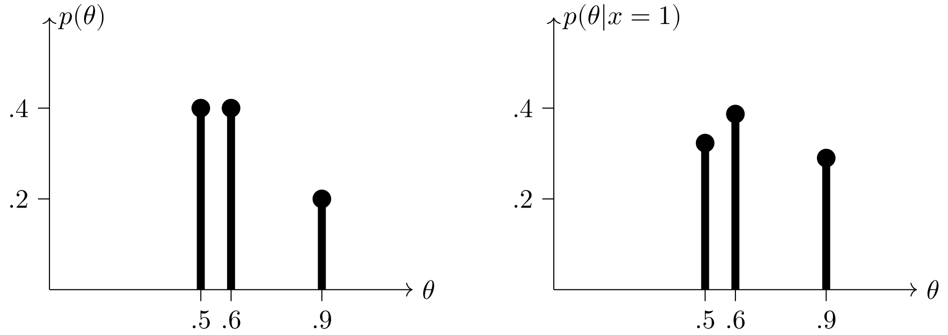
Our standard notations will be:

- θ is the value of the hypothesis.
- $p(\theta)$ is the prior probability mass function of the hypothesis.
- $p(\theta|\mathcal{D})$ is the posterior probability mass function of the hypothesis given the data.
- $p(\mathcal{D}|\theta)$ is the likelihood function. (This is not a pmf!)

In Example 1 we can represent the three hypotheses A, B, and C by $\theta = 0.5, 0.6, 0.9$. For the data we'll let $x = 1$ mean heads and $x = 0$ mean tails. Then the prior and posterior probabilities in the table define the prior and posterior probability mass functions.

Hypothesis	θ	prior pmf $p(\theta)$	posterior pmf $p(\theta\|x=1)$
A	0.5	$P(A) = p(0.5) = 0.4$	$P(A\|\mathcal{D}) = p(0.5\|x=1) = 0.3226$
B	0.6	$P(B) = p(0.6) = 0.4$	$P(B\|\mathcal{D}) = p(0.6\|x=1) = 0.3871$
C	0.9	$P(C) = p(0.9) = 0.2$	$P(C\|\mathcal{D}) = p(0.9\|x=1) = 0.2903$

Here are plots of the prior and posterior pmf's from the example.



Prior pmf $p(\theta)$ and posterior pmf $p(\theta|x=1)$ for Example 1

If the data was different then the likelihood column in the Bayesian update table would be different. We can plan for different data by building the entire likelihood table ahead of time. In the coin example there are two possibilities for the data: the toss is heads or the toss is tails. So the full likelihood table has two likelihood columns:

hypothesis	likelihood $p(x\|\theta)$	
θ	$p(x=0\|\theta)$	$p(x=1\|\theta)$
0.5	0.5	0.5
0.6	0.4	0.6
0.9	0.1	0.9

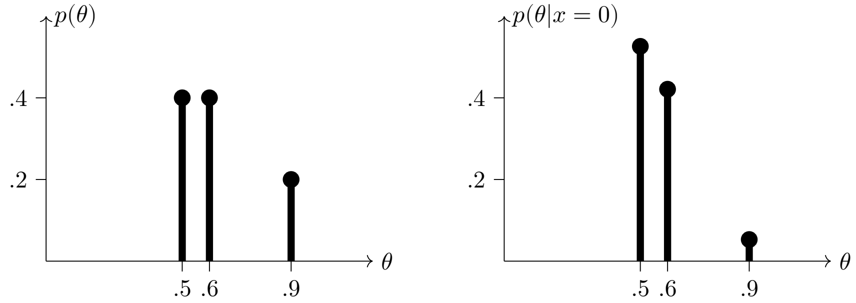
Important convention. Notice that in the above table we used the value of θ as the hypothesis. Of course, hypothesizing '$\theta = 0.5$' is exactly the same as hypothesizing 'the coin is type A'. It is also useful in settings where we haven't named all the possible hypotheses.

Example 2. Using the notation $p(\theta)$, etc., redo Example 1 assuming the flip was tails.

Solution: Since the data has changed, the likelihood column in the Bayesian update table is now for $x = 0$. That is, we must take the $p(x=0|\theta)$ column from the likelihood table.

hypothesis	prior	likelihood	Bayes numerator	posterior
θ	$p(\theta)$	$p(x=0\,\|\,\theta)$	$p(x=0\,\|\,\theta)p(\theta)$	$p(\theta\,\|\,x=0)$
0.5	0.4	0.5	0.2	0.5263
0.6	0.4	0.4	0.16	0.4211
0.9	0.2	0.1	0.02	0.0526
total	1	NO SUM	0.38	1

Now the probability that $\theta = 0.5$, i.e. the coin is type A, has increased from 0.4 to 0.5263, while the probability that $\theta = 0.9$, i.e the coin is type C, has decreased from 0.2 to only 0.0526. Here are the corresponding plots:



Prior pmf $p(\theta)$ and posterior pmf $p(\theta|x=0)$

3.3 Food for thought.

Suppose that in Example 1 you didn't know how many coins of each type were in the drawer. You picked one at random and got heads. How would you go about deciding which hypothesis (coin type) if any was most supported by the data?

4 Updating again and again

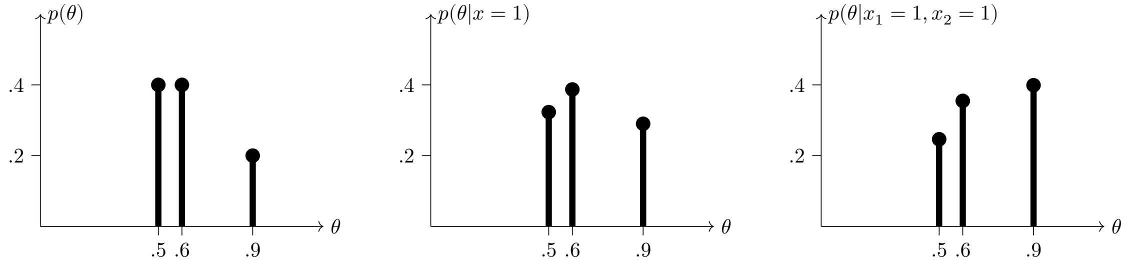
In life we are continually updating our beliefs with each new experience of the world. In Bayesian inference, after updating the prior to the posterior, we can take more data and update again! For the second update, the posterior from the first data becomes the prior for the second data.

Example 3. Suppose you have picked a coin as in Example 1. You flip it once and get heads. Then you flip the same coin and get heads again. What is the probability that the coin was type A? Type B? Type C?

Solution: As we update several times the table gets big, so we use a smaller font to fit it in:

hypothesis	prior	likelihood 1	Bayes numerator 1	likelihood 2	Bayes numerator 2	posterior 2
θ	$p(\theta)$	$p(x_1=1\|\theta)$	$p(x_1=1\|\theta)p(\theta)$	$p(x_2=1\|\theta)$	$p(x_2=1\|\theta)p(x_1=1\|\theta)p(\theta)$	$p(\theta\|x_1=1,x_2=1)$
0.5	0.4	0.5	0.2	0.5	0.1	0.2463
0.6	0.4	0.6	0.24	0.6	0.144	0.3547
0.9	0.2	0.9	0.18	0.9	0.162	0.3990
total	1	NO SUM		NO SUM	0.406	1

Note that the second Bayes numerator is computed by multiplying the first Bayes numerator and the second likelihood; since we are only interested in the final posterior, there is no need to normalize until the last step. As shown in the last column and plot, after two heads the type C hypothesis has finally taken the lead!



prior $p(\theta)$, first posterior $p(\theta|x_1=1)$, second posterior $p(\theta|x_1=1, x_2=1)$

5 Appendix: the base rate fallacy

Example 4. A screening test for a disease is both sensitive and specific. By that we mean it is usually positive when testing a person with the disease and usually negative when testing someone without the disease. Let's assume the true positive rate is 99% and the false positive rate is 2%. Suppose the prevalence of the disease in the general population is 0.5%. If a random person tests positive, what is the probability that they have the disease?

Solution: As a review we first do the computation using trees. Next we will redo the computation using tables.

Let's use notation established above for hypotheses and data: let $\mathcal{H}_+$ be the hypothesis (event) that the person has the disease and let $\mathcal{H}_-$ be the hypothesis they do not. Likewise, let $\mathcal{T}_+$ and $\mathcal{T}_-$ represent the data of a positive and negative screening test respectively. We are asked to compute $P(\mathcal{H}_+|\mathcal{T}_+)$.

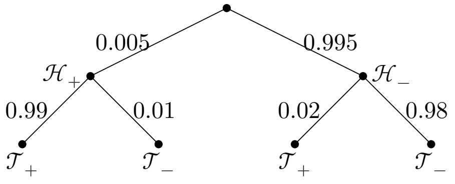
We are given

$$P(\mathcal{T}_+|\mathcal{H}_+) = 0.99, \quad P(\mathcal{T}_+|\mathcal{H}_-) = 0.02, \quad P(\mathcal{H}_+) = 0.005.$$

From these we can compute the false negative and true negative rates:

$$P(\mathcal{T}_-|\mathcal{H}_+) = 0.01, \quad P(\mathcal{T}_-|\mathcal{H}_-) = 0.98$$

All of these probabilities can be displayed quite nicely in a tree.



Bayes' theorem yields

$$P(\mathcal{H}_+|\mathcal{T}_+) = \frac{P(\mathcal{T}_+|\mathcal{H}_+)P(\mathcal{H}_+)}{P(\mathcal{T}_+)} = \frac{0.99 \cdot 0.005}{0.99 \cdot 0.005 + 0.02 \cdot 0.995} = 0.19920 \approx 20\%$$

Now we redo this calculation using a Bayesian update table:

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{T}_+\|\mathcal{H})$	$P(\mathcal{T}_+\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{T}_+)$
$\mathcal{H}_+$	0.005	0.99	0.00495	0.19920
$\mathcal{H}_-$	0.995	0.02	0.01990	0.80080
total	1	NO SUM	$P(\mathcal{T}_+) = 0.02485$	1

The table shows that the posterior probability $P(\mathcal{H}_+|\mathcal{T}_+)$ that a person with a positive test has the disease is about 20%. This is far less than the sensitivity of the test (99%) but much higher than the prevalence of the disease in the general population (0.5%).

Bayesian Updating: Probabilistic Prediction
Class 12, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to use the law of total probability to compute prior and posterior predictive probabilities.

2 Introduction

In the previous class we looked at updating the probability of hypotheses based on data. We can also use the data to update the probability of each possible outcome of a future experiment. In this class we will look at how this is done.

2.1 Probabilistic prediciton; words of estimative probability (WEP)

There are many ways to word predictions:

- Prediction: "It will rain tomorrow."
- Prediction using words of estimative probability (WEP): "It is likely to rain tomorrow."
- Probabilistic prediction: "Tomorrow it will rain with probability 60% (and not rain with probability 40%)."

Each type of wording is appropriate at different times.

In this class we are going to focus on probabilistic prediction and precise quantitative statements. You can see `https://en.wikipedia.org/wiki/Words_of_Estimative_Probability` for an interesting discussion about the appropriate use of words of estimative probability. The article also contains a list of *weasel words* such as 'might', 'cannot rule out', 'it's conceivable' that should be avoided as almost certain to cause confusion.

There are many places where we want to make a probabilistic prediction. Examples are

- Medical treatment outcomes
- Weather forecasting
- Climate change
- Sports betting
- Elections
- ...

These are all situations where there is uncertainty about the outcome and we would like as precise a description of what could happen as possible.

3 Predictive Probabilities

Probabilistic prediction simply means assigning a probability to each possible outcomes of an experiment.

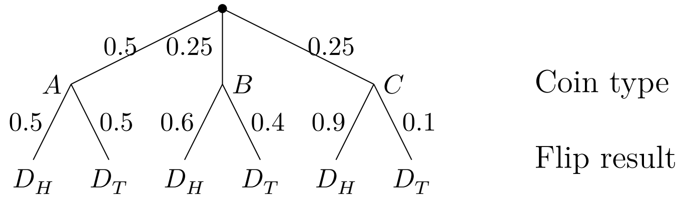
Recall the coin example from the previous class notes: there are three types of coins which are indistinguishable apart from their probability of landing heads when tossed.

- Type A coins are fair, with probability 0.5 of heads
- Type B coins have probability 0.6 of heads
- Type C coins have probability 0.9 of heads

You have a drawer containing 4 coins: 2 of type A, 1 of type B, and 1 of type C. You reach into the drawer and pick a coin at random. We let A stand for the event 'the chosen coin is of type A'. Likewise for B and C.

3.1 Prior predictive probabilities

Before taking data we can compute the probability that our chosen coin will land heads (or tails) if flipped. Let D_H be the event it lands heads and let D_T the event it lands tails. We can use the law of total probability to determine the probabilities of these events. Either by drawing a tree or directly proceeding to the algebra, we get:



$$\begin{aligned}
P(D_H) &= P(D_H|A)P(A) + P(D_H|B)P(B) + P(D_H|C)P(C) \\
&= 0.5 \cdot 0.5 + 0.6 \cdot 0.25 + 0.9 \cdot 0.25 = 0.625 \\
P(D_T) &= P(D_T|A)P(A) + P(D_T|B)P(B) + P(D_T|C)P(C) \\
&= 0.5 \cdot 0.5 + 0.4 \cdot 0.25 + 0.1 \cdot 0.25 = 0.375
\end{aligned}$$

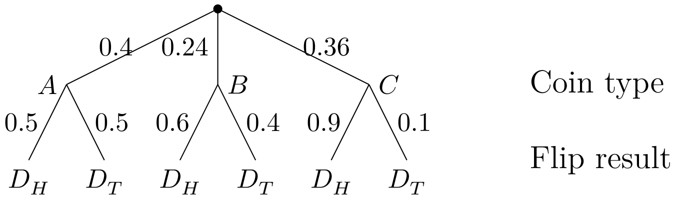
Definition: These probabilities give a (probabilistic) prediction of what will happen if the coin is tossed. Because they are computed before we collect any data they are called prior predictive probabilities.

3.2 Posterior predictive probabilities

Suppose we flip the coin once and it lands heads. We now have data D, which we can use to update the prior probabilities of our hypotheses to posterior probabilities. Last class we learned to use a Bayes table to facilitate this computation:

hypothesis	prior	likelihood	Bayes numerator	posterior
H	$P(H)$	$P(D\|H)$	$P(D\|H)P(H)$	$P(H\|D)$
A	0.5	0.5	0.25	0.4
B	0.25	0.6	0.15	0.24
C	0.25	0.9	0.225	0.36
total	1	NO SUM	0.625	1

Having flipped the coin once and gotten heads, we can compute the probability that our chosen coin will land heads (or tails) if flipped a second time. We proceed just as before, but using the posterior probabilities $P(A|D)$, $P(B|D)$, $P(C|D)$ in place of the prior probabilities $P(A)$, $P(B)$, $P(C)$.



$$\begin{aligned}P(D_H|D) &= P(D_H|A)P(A|D) + P(D_H|B)P(B|D) + P(D_H|C)P(C|D)\\ &= 0.5\cdot 0.4 + 0.6\cdot 0.24 + 0.9\cdot 0.36 = 0.668\\ P(D_T|D) &= P(D_T|A)P(A|D) + P(D_T|B)P(B|D) + P(D_T|C)P(C|D)\\ &= 0.5\cdot 0.4 + 0.4\cdot 0.24 + 0.1\cdot 0.36 = 0.332\end{aligned}$$

Definition: These probabilities give a (probabilistic) prediction of what will happen if the coin is tossed again. Because they are computed after collecting data and updating the prior to the posterior, they are called posterior predictive probabilities.

Note that heads on the first toss increases the probability of heads on the second toss.

3.3 Review

Here's a succinct description of the preceding sections that may be helpful:

Each hypothesis gives a different probability of heads, so the total probability of heads is a weighted average. For the prior predictive probability of heads, the weights are given by the prior probabilities of the hypotheses. For the posterior predictive probability of heads, the weights are given by the posterior probabilities of the hypotheses.

Remember: Prior and posterior probabilities are for hypotheses. Prior predictive and posterior predictive probabilities are for outcomes. To keep this straight, remember that the predictive probabilities are used to predict future outcomes, i.e. data.

Bayesian Updating: Odds
Class 12, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to convert between odds and probability.
2. Be able to update prior odds to posterior odds using Bayes factors.
3. Understand how Bayes factors measure the extent to which data provides evidence for or against a hypothesis.

2 Odds

When comparing two events, it common to phrase probability statements in terms of odds.

Definition The odds of event E versus event E' are the ratio of their probabilities $P(E)/P(E')$. If unspecified, the second event is assumed to be the complement E^c. So the odds of E are:

$$O(E) = \frac{P(E)}{P(E^c)}.$$

For example, $O(\text{rain}) = 2$ means that the probability of rain is twice the probability of no rain (2/3 versus 1/3). We might say 'the odds of rain are 2 to 1.'

Example. For a fair coin, $O(\text{heads}) = \frac{1/2}{1/2} = 1$. We might say the odds of heads are 1 to 1 or fifty-fifty.

Example. For a standard die, the odds of rolling a 4 are $\frac{1/6}{5/6} = \frac{1}{5}$. We might say the odds are '1 to 5 for' or '5 to 1 against' rolling a 4.

Example. The probability of a pair in a five card poker hand is 0.42257. So the odds of a pair are 0.42257/(1-0.42257) = 0.73181.

We can go back and forth between probability and odds as follows.

Conversion formulas: if $P(E) = p$ then $O(E) = \frac{p}{1-p}$. If $O(E) = q$ then $P(E) = \frac{q}{1+q}$.

Notes:
1. The second formula simply solves $q = p/(1-p)$ for p.
2. Probabilities are between 0 and 1, while odds are between 0 to ∞.
3. The property $P(E^c) = 1 - P(E)$ becomes $O(E^c) = 1/O(E)$.

Example. Let F be the event that a five card poker hand is a full house. Then $P(F) = 0.00145214$ so $O(F) = 0.0014521/(1 - 0.0014521) = 0.0014542$.

The odds not having a full house are $O(F^c) = (1 - 0.0014521)/0.0014521 = 687 = 1/O(F)$.

4. If $P(E)$ or $O(E)$ is small then $O(E) \approx P(E)$. This follows from the conversion formulas.

Example. In the poker example where F = 'full house' we saw that $P(F)$ and $O(F)$ differ only in the fourth significant digit.

3 Updating odds

3.1 Introduction

In Bayesian updating, we used the likelihood of data to update prior probabilities of hypotheses to posterior probabilities. In the language of odds, we will update prior odds to posterior odds. One of our key points will be that the data can provide evidence supporting or negating a hypothesis depending on whether its posterior odds are greater or less than its prior odds.

We'll begin by returning to our familiar example of a screening test for a disease.

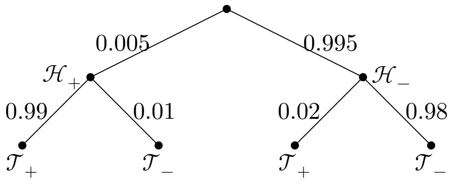
Example 1. Briefly, a screening test for a disease is both sensitive and specific. Assume the true positive rate is 99% and the false positive rate is 2%. Suppose the prevalence of the disease in the general population is 0.5%. For a randomly chosen person, what are the prior odds that they have the disease? Suppose they test positive, now what are the posterior odds that they have the disease? By what factor have the odds changed as a result of the test?

Solution: We'll use our, by now, standard notation:
$\mathcal{H}_+$ = have disease, $\mathcal{H}_-$ = do not have disease, $\mathcal{T}_+$ = test positive, $\mathcal{T}_-$ = test negative.

To start with the prior odds that they have the disease are

$$O(\mathcal{H}_+) = \frac{P(\mathcal{H}_+)}{P(\mathcal{H}_-)} = \frac{0.005}{0.995} \approx 0.005$$

For the posterior odds, we'll do the computation with trees and then repeat it with tables. Here is the tree describing the scenario.



Bayes' theorem yields

$$O(\mathcal{H}_+|\mathcal{T}_+) = \frac{P(\mathcal{H}_+|\mathcal{T}_+)}{P(\mathcal{H}_-|\mathcal{T}_+)} = \frac{P(\mathcal{T}_+|\mathcal{H}_+)P(\mathcal{H}_+)/P(\mathcal{T}_+)}{P(\mathcal{T}_+|\mathcal{H}_-)P(\mathcal{H}_-)/P(\mathcal{T}_+)} = \frac{P(\mathcal{T}_+|\mathcal{H}_+)P(\mathcal{H}_+)}{P(\mathcal{T}_+|\mathcal{H}_-)P(\mathcal{H}_-)}$$

This is great! For the odds, the total probability $P(\mathcal{T}_+)$ cancels and does not need to be computed. Putting in numbers we see the posterior odds are

$$O(\mathcal{H}_+|\mathcal{T}_+) = \frac{0.99 \cdot 0.005}{0.02 \cdot 0.995} \approx 50 \cdot 0.005 = 1/4.$$

We've structured the presentation so you can easily see that the posterior odds are only one in four. However, the posterior odds are about a factor of 50 greater than the prior odds.

Redoing this calculation using a Bayesian update table:

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{T}_+\vert\mathcal{H})$	$P(\mathcal{T}_+\vert\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\vert\mathcal{T}_+)$
$\mathcal{H}_+$	0.005	0.99	0.00495	0.19920
$\mathcal{H}_-$	0.995	0.02	0.01990	0.80080
total	1	NO SUM	$P(\mathcal{T}_+) = 0.02485$	1

The prior odds are computed using the prior column of the table. As above, they are $\dfrac{P(\mathcal{H}_+)}{P(\mathcal{H}_-)} = \dfrac{0.005}{0.995}$.

The posterior odds are computed using either the posterior or Bayes numerator columns of the table. We can use either column, because, they only differ by the normalazing factor of $P(\mathcal{T}_+)$ in the denominator of the posteriors. We get the same answer as above:

$$O(\mathcal{H}_+|\mathcal{T}_+) = \frac{0.00495}{0.01990} \approx \frac{5}{20}.$$

You should see that these odds come by multiplying the prior odds by the ratio of the likelihoods.

3.2 Example: Marfan syndrome

Marfan syndrome is a genetic disease of connective tissue that occurs in 1 of every 15000 people. The main ocular features of Marfan syndrome include bilateral ectopia lentis (lens dislocation), myopia and retinal detachment. About 70% of people with Marfan syndrome have a least one of these ocular features; only 7% of people without Marfan syndrome do. (We don't guarantee the accuracy of these numbers, but they will work perfectly well for our example.)

If a person has at least one of these ocular features, what are the odds that they have Marfan syndrome?

Solution: This is a standard Bayesian updating problem. Our hypotheses are:
M = 'the person has Marfan syndrome' M^c = 'the person does not have Marfan syndrome'

The data is:

F = 'the person has at least one ocular feature'.

We are given the prior probability of M and the likelihoods of F given M or M^c:

$$P(M) = 1/15000, \qquad P(F|M) = 0.7, \qquad P(F|M^c) = 0.07.$$

As before, we can compute the posterior probabilities using a table:

hypothesis	prior	likelihood	Bayes numerator	posterior
H	$P(H)$	$P(F\|H)$	$P(F\|H)P(H)$	$P(H\|F)$
M	0.000067	0.7	0.0000467	0.00066
M^c	0.999933	0.07	0.069995	0.99933
total	1	no sum	$P(F) = 0.07004$	1

First we find the prior odds:

$$O(M) = \frac{P(M)}{P(M^c)} = \frac{1/15000}{14999/15000} = \frac{1}{14999} \approx 0.000067.$$

The posterior odds are given by the ratio of the posterior probabilities or the Bayes numerators, since the normalizing factor will be the same in both numerator and denominator.

$$O(M|F) = \frac{P(M|F)}{P(M^c|F)} = \frac{P(F|M)P(M)}{P(F|M^c)P(M^c)} = 0.000667.$$

The posterior odds are a factor of 10 larger than the prior odds. In that sense, having an ocular feature is strong evidence in favor of the hypothesis M. However, because the prior odds are so small, it is still highly unlikely the person has Marfan syndrome.

4 Bayes factors and strength of evidence

The factor of 10 in the previous example is called a Bayes factor or a likelihood ratio. The exact definition is the following.

Definition: For a hypothesis H and data D, the Bayes factor is the ratio of the likelihoods:

$$\text{Bayes factor} = \frac{P(D|H)}{P(D|H^c)}.$$

This is also called the likelihood ratio.

Let's see exactly where the Bayes factor arises in updating odds. We have

$$\begin{aligned} O(H|D) &= \frac{P(H|D)}{P(H^c|D)} \\ &= \frac{P(D|H)\,P(H)}{P(D|H^c)P(H^c)} \\ &= \frac{P(D|H)}{P(D|H^c)} \cdot \frac{P(H)}{P(H^c)} \\ &= \frac{P(D|H)}{P(D|H^c)} \cdot O(H) \end{aligned}$$

posterior odds = **Bayes factor** × prior odds

From this formula, we see that the Bayes' factor (BF) tells us whether the data provides evidence for or against the hypothesis.

- If $BF > 1$ then the posterior odds are greater than the prior odds. So the data provides evidence for the hypothesis.
- If $BF < 1$ then the posterior odds are less than the prior odds. So the data provides evidence against the hypothesis.
- If $BF = 1$ then the prior and posterior odds are equal. So the data provides no evidence either way.

The following example is taken from the textbook *Information Theory, Inference, and Learning Algorithms* by David J. C. Mackay, who has this to say regarding trial evidence.

> In my view, a jury's task should generally be to multiply together carefully evaluated likelihood ratios from each independent piece of admissible evidence with an equally carefully reasoned prior probability. This view is shared by many statisticians but learned British appeal judges recently disagreed and actually overturned the verdict of a trial because the jurors *had* been taught to use Bayes' theorem to handle complicated DNA evidence.

Example 2. Two people have left traces of their own blood at the scene of a crime. A suspect , Oliver, is tested and found to have type 'O' blood. The blood groups of the two traces are found to be of type 'O' (a common type in the local population, having frequency 60%) and type 'AB' (a rare type, with frequency 1%). Does this data (type 'O' and 'AB' blood were found at the scene) give evidence in favor of the proposition that Oliver was one of the two people present at the scene of the crime?"

Solution: There are two hypotheses:
S = 'Oliver and another unknown person were at the scene of the crime
S^c = 'two unknown people were at the scene of the crime'

The data is:
D = 'type 'O' and 'AB' blood were found'

The Bayes factor for Oliver's presence is $BF_{\text{Oliver}} = \dfrac{P(D|S)}{P(D|S^c)}$. We compute the numerator and denominator of this separately.

The data says that both type O and type AB blood were found. If Oliver was at the scene then 'type O' blood would be there. So $P(D|S)$ is the probability that the other person had type AB blood. We are told this is 0.01, so $P(D|S) = 0.01$.

If Oliver was not at the scene then there were two random people one with type O and one with type AB blood. The probability of this is $2 \cdot 0.6 \cdot 0.01$.* Thus the Bayes factor for Oliver's presence is

$$BF_{\text{Oliver}} = \frac{P(D|S)}{P(D|S^c)} = \frac{0.01}{2 \cdot 0.6 \cdot 0.01} = 0.83.$$

Since $BF_{\text{Oliver}} < 1$, the data provides (weak) evidence against Oliver being at the scene.

*The factor of 2 is, perhaps surprising. The following careful counting will explain it. Suppose there are N people in the population, N_O have type O blood and N_{AB} have type AB. So $N_O/N = 0.6$

and $N_{AB}/N = 0.01$. We want the probability that a random choice of 2 people will pick one of type O and one of type AB. This is clearly

$$\frac{\binom{N_O}{1}\binom{N_{AB}}{1}}{\binom{N}{2}} = \frac{N_O N_{AB}}{N(N-1)/2} = 2 \cdot \frac{N_O}{N} \cdot \frac{N_{AB}}{N-1} \approx 2 \cdot 0.6 \cdot 0.01.$$

In the last approximation, we assumed that N is large enough the $N_{AB}/(N-1) \approx N_{AB}/N$.

Example 3. Another suspect Alberto is found to have type 'AB' blood. Do the same data give evidence in favor of the proposition that Alberto was one of the two people present at the crime?

Solution: Reusing the above notation with Alberto in place of Oliver we have:

$$BF_{\text{Alberto}} = \frac{P(D|S)}{P(D|S^c)} = \frac{0.6}{2 \cdot 0.6 \cdot 0.01} = 50.$$

Since $BF_{\text{Alberto}} \gg 1$, the data provides strong evidence in favor of Alberto being at the scene.

Notes:
1. In both examples, we have only computed the Bayes factor, not the posterior odds. To compute the latter, we would need to know the prior odds that Oliver (or Alberto) was at the scene based on other evidence.

2. Note that if 50% of the population had type O blood instead of 60%, then the Oliver's Bayes factor would be 1 (neither for nor against). More generally, the break-even point for blood type evidence is when the proportion of the suspect's blood type in the general population equals the proportion of the suspect's blood type among those who left blood at the scene.

4.1 Updating again and again

Suppose we collect data in two stages, first D_1, then D_2. We have seen in our dice and coin examples that the final posterior can be computed all at once or in two stages where we first update the prior using the likelihoods for D_1 and then update the resulting posterior using the likelihoods for D_2. The latter approach works whenever likelihoods multiply:

$$P(D_1, D_2|H) = P(D_1|H)P(D_2|H).$$

Since likelihoods are conditioned on hypotheses, we say that D_1 and D_2 are conditionally independent if the above equation holds for every hypothesis H.

Example. There are five dice in a drawer, with 4, 6, 8, 12, and 20 sides (these are the hypotheses). I pick a die at random and roll it twice. The first roll gives 7. The second roll gives 11. Are these results conditionally independent? Are they independent?

Solution: These results are conditionally independent. For example, for the hypothesis of the 8-sided die we have:

$$\begin{aligned} P(7 \text{ on roll } 1 \,|\, 8\text{-sided die}) &= 1/8 \\ P(11 \text{ on roll } 2 \,|\, 8\text{-sided die}) &= 0 \\ P(7 \text{ on roll } 1, 11 \text{ on roll } 2 \,|\, 8\text{-sided die}) &= 0 \end{aligned}$$

For the hypothesis of the 20-sided die we have:

$$\begin{aligned} P(7\text{ on roll }1\,|\,20\text{-sided die}) &= 1/20 \\ P(11\text{ on roll }2\,|\,20\text{-sided die}) &= 1/20 \\ P(7\text{ on roll }1,\ 11\text{ on roll }2\,|\,20\text{-sided die}) &= (1/20)^2 \end{aligned}$$

However, the results of the rolls are *not* independent. That is:

$$P(7\text{ on roll }1,\ 11\text{ on roll }2) \neq P(7\text{ on roll }1)P(11\text{ on roll }2).$$

Intuitively, this is because a 7 on the roll 1 allows us to rule out the 4- and 6-sided dice, making an 11 on roll 2 more likely. Let's check this intuition by computing both sides precisely. On the righthand side we have:

$$\begin{aligned} P(7\text{ on roll }1) &= \frac{1}{5}\cdot\frac{1}{8} + \frac{1}{5}\cdot\frac{1}{12} + \frac{1}{5}\cdot\frac{1}{20} = \frac{31}{600} \\ P(11\text{ on roll }2) &= \frac{1}{5}\cdot\frac{1}{12} + \frac{1}{5}\cdot\frac{1}{20} = \frac{2}{75} \end{aligned}$$

On the lefthand side we have:

$$\begin{aligned} P(7\text{ on roll }1,\ 11\text{ on roll }2) &= P(11\text{ on roll }2\,|\,7\text{ on roll }1)P(7\text{ on roll }1) \\ &= \left(\frac{30}{93}\cdot\frac{1}{12} + \frac{6}{31}\cdot\frac{1}{20}\right)\cdot\frac{31}{600} \\ &= \frac{17}{465}\cdot\frac{31}{600} = \frac{17}{9000} \end{aligned}$$

Here $\frac{30}{93}$ and $\frac{6}{31}$ are the posterior probabilities of the 12- and 20-sided dice given a 7 on roll 1. We conclude that, without conditioning on hypotheses, the rolls are not independent.

Returning the to general setup, if D_1 and D_2 are conditionally independent for H and H^c then it makes sense to consider each Bayes factor independently:

$$BF_i = \frac{P(D_i|H)}{P(D_i|H^c)}.$$

The prior odds of H are $O(H)$. The posterior odds after D_1 are

$$O(H|D_1) = BF_1 \cdot O(H).$$

And the posterior odds after D_1 and D_2 are

$$\begin{aligned} O(H|D_1, D_2) &= BF_2 \cdot O(H|D_1) \\ &= BF_2 \cdot BF_1 \cdot O(H) \end{aligned}$$

We have the beautifully simple notion that updating with new data just amounts to multiplying the current posterior odds by the Bayes factor of the new data.

Example 4. Other symptoms of Marfan Syndrome

Recall from the earlier example that the Bayes factor for a least one ocular feature (F) is

$$BF_F = \frac{P(F|M)}{P(F|M^c)} = \frac{0.7}{0.07} = 10.$$

The wrist sign (W) is the ability to wrap one hand around your other wrist to cover your pinky nail with your thumb. Assume 10% of the population have the wrist sign, while 90% of people with Marfan's have it. Therefore the Bayes factor for the wrist sign is

$$BF_W = \frac{P(W|M)}{P(W|M^c)} = \frac{0.9}{0.1} = 9.$$

We will assume that F and W are conditionally independent symptoms. That is, among people with Marfan syndrome, ocular features and the wrist sign are independent, and among people without Marfan syndrome, ocular features and the wrist sign are independent. Given this assumption, the posterior odds of Marfan syndrome for someone with both an ocular feature and the wrist sign are

$$O(M|F,W) = BF_W \cdot BF_F \cdot O(M) = 9 \cdot 10 \cdot \frac{1}{14999} \approx \frac{6}{1000}.$$

We can convert the posterior odds back to probability, but since the odds are so small the result is nearly the same:

$$P(M|F,W) \approx \frac{6}{1000+6} \approx 0.596\%.$$

So ocular features and the wrist sign are both strong evidence in favor of the hypothesis M, and taken together they are very strong evidence. Again, because the prior odds are so small, it is still unlikely that the person has Marfan syndrome, but at this point it might be worth undergoing further testing given potentially fatal consequences of the disease (such as aortic aneurysm or dissection).

Note also that if a person has exactly one of the two symptoms, then the product of the Bayes factors is near 1 (either 9/10 or 10/9). So the two pieces of data essentially cancel each other out with regard to the evidence they provide for Marfan's syndrome.

5 Log odds

In practice, people often find it convenient to work with the natural log of the odds in place of odds. Naturally enough these are called the log odds. The Bayesian update formula

$$O(H|D_1, D_2) = BF_2 \cdot BF_1 \cdot O(H)$$

becomes

$$\ln(O(H|D_1, D_2)) = \ln(BF_2) + \ln(BF_1) + \ln(O(H)).$$

We can interpret the above formula for the posterior log odds as the sum of the prior log odds and all the evidence $\ln(BF_i)$ provided by the data. Note that by taking logs, evidence in favor ($BF_i > 1$) is positive and evidence against ($BF_i < 1$) is negative.

To avoid lengthier computations, we will work with odds rather than log odds in this course. Log odds are nice because sums are often more intuitive then products. Log odds also play a central role in logistic regression, an important statistical model related to linear regression.

Bayesian Updating with Continuous Priors
Class 13, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Understand a parameterized family of distributions as representing a continuous range of hypotheses for the observed data.
2. Be able to state Bayes' theorem and the law of total probability for continous densities.
3. Be able to apply Bayes' theorem to update a prior probability density function to a posterior pdf given data and a likelihood function.
4. Be able to interpret and compute posterior predictive probabilities.

2 Introduction

Up to now we have only done Bayesian updating when we had a finite number of hypothesis, e.g. our dice example had five hypotheses (4, 6, 8, 12 or 20 sides). Now we will study Bayesian updating when there is a continuous range of hypotheses. The Bayesian update process will be essentially the same as in the discrete case. As usual when moving from discrete to continuous we will need to replace the probability mass function by a probability density function, and sums by integrals.

The first few sections of this note are devoted to working with pdfs. In particular we will cover the law of total probability and Bayes' theorem. We encourage you to focus on how these are essentially identical to the discrete versions. After that, we will apply Bayes' theorem and the law of total probability to Bayesian updating.

3 Examples with continuous ranges of hypotheses

Here are three standard examples with continuous ranges of hypotheses.

Example 1. Suppose you have a system that can succeed or fail with probability p. Then we can hypothesize that p is anywhere in the range $[0, 1]$. That is, we have a continuous range of hypotheses. We will often model this example with a 'bent' coin with unknown probability p of heads.

Example 2. The lifetime of a certain isotope is modeled by an exponential distribution $\exp(\lambda)$. In principle, the mean lifetime $1/\lambda$ can be any real number in $(0, \infty)$.

Example 3. We are not restricted to a single parameter. In principle, the parameters μ and σ of a normal distribution can be any real numbers in $(-\infty, \infty)$ and $(0, \infty)$, respectively. If we model gestational length for single births by a normal distribution, then from millions of data points we know that μ is about 40 weeks and σ is about one week.

In all of these examples we modeled the random process giving rise to the data by a distribution with parameters –called a parametrized distribution. Every possible choice of the parameter(s) is a hypothesis, e.g. we can hypothesize that the probability of succcess in Example 1 is $p = 0.7313$. We have a continuous set of hypotheses because we could take any value between 0 and 1.

4 Notational conventions

4.1 Parametrized models

As in the examples above our hypotheses often take the form a certain parameter has value θ. We will often use the letter θ to stand for an arbitrary hypothesis. This will leave symbols like p, f, and x to take there usual meanings as pmf, pdf, and data. Also, rather than saying 'the hypothesis that the parameter of interest has value θ' we will simply say the hypothesis θ.

4.2 Big and little letters

We have two parallel notations for outcomes and probability:

1. (Big letters) Event A, probability function $P(A)$.
2. (Little letters) Value x, pmf $p(x)$ or pdf $f(x)$.

These notations are related by $P(X = x) = p(x)$, where x is a value the discrete random variable X and '$X = x$' is the corresponding event.

We carry these notations over to the probabilities used in Bayesian updating.

1. (Big letters) From hypotheses $\mathcal{H}$ and data $\mathcal{D}$ we compute several associated probabilities

$$P(\mathcal{H}),\ P(\mathcal{D}),\ P(\mathcal{H}|\mathcal{D}),\ P(\mathcal{D}|\mathcal{H}).$$

In the coin example we might have $\mathcal{H}_{0.6}$ = 'the chosen coin has probability 0.6 of heads', $\mathcal{D}$ = '3 flips landed HHT', so $P(\mathcal{D}|\mathcal{H}_{0.6}) = (0.6)^2(0.4)$

2. (Small letters) Hypothesis values θ and data values x both have probabilities or probability densities:

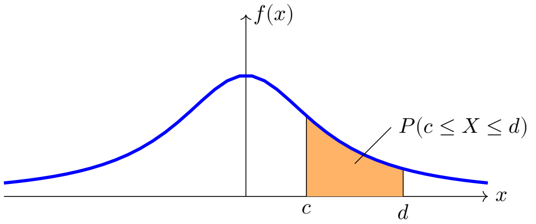
$$\begin{array}{cccc} p(\theta) & p(x) & p(\theta|x) & p(x|\theta) \\ f(\theta) & f(x) & f(\theta|x) & f(x|\theta) \end{array}$$

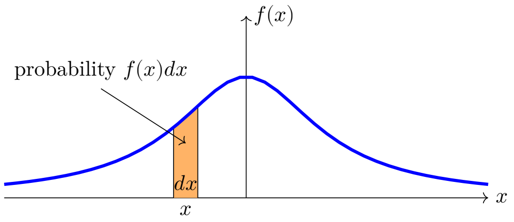
In the coin example we might have $\theta = 0.6$ and x is the sequence $1, 1, 0$. So, $p(x|\theta) = (0.6)^2(0.4)$. We might also write $p(x = 1, 1, 0|\theta = 0.6)$ to emphasize the values of x and θ.

Although we will still use both types of notation, from now on we will mostly use the small letter notation involving pmfs and pdfs. Hypotheses will usually be parameters represented by Greek letters $(\theta, \lambda, \mu, \sigma, \dots)$ while data values will usually be represented by English letters $(x, x_i, y, \dots)$.

5 Quick review of pdf and probability

Suppose X is a random variable with pdf $f(x)$. Recall $f(x)$ is a density; its units are probability/(units of x).





The probability that the value of X is in $[c, d]$ is given by

$$\int_c^d f(x)\,dx.$$

The probability that X is in an infinitesimal range dx around x is $f(x)\,dx$. In fact, the integral formula is just the 'sum' of these infinitesimal probabilities. We can visualize these probabilities by viewing the integral as area under the graph of $f(x)$.

In order to manipulate probabilities instead of densities in what follows, we will make frequent use of the notion that $f(x)\,dx$ is the probability that X is in an infinitesimal range around x of width dx. Please make sure that you fully understand this notion.

6 Continuous priors, discrete likelihoods

In the Bayesian framework we have probabilities of hypotheses –called prior and posterior probabilities– and probabilities of data given a hypothesis –called likelihoods. In earlier classes both the hypotheses and the data had discrete ranges of values. We saw in the introduction that we might have a continuous range of hypotheses. The same is true for the data, but for today we will assume that our data can only take a discrete set of values. In this case, the likelihood of data x given hypothesis θ is written using a pmf: $p(x|\theta)$.

We will use the following coin example to explain these notions. We will carry this example through in each of the succeeding sections.

Example 4. Suppose we have a bent coin with unknown probability θ of heads. In this case, we'll say the coin is of 'type θ' and we'll label the hypothesis that a random coin is of type θ by $\mathcal{H}_\theta$. The value of θ is random and could be anywhere between 0 and 1. For this and the examples that follow we'll suppose that the value of θ follows a distribution with continuous prior probability density $f(\theta) = 2\theta$. We have a discrete likelihood because tossing a coin has only two outcomes, $x = 1$ for heads and $x = 0$ for tails.

$$p(x = 1|\mathcal{H}_\theta) = \theta, \qquad p(x = 0|\mathcal{H}_\theta) = 1 - \theta.$$

As we stated earlier, we will often write θ for the hypothesis $\mathcal{H}_\theta$. So the above probabilities become

$$p(x = 1|\theta) = \theta, \qquad p(x = 0|\theta) = 1 - \theta.$$

Think: This can be tricky to wrap your mind around. We have a continuous range of types of coins –we identify the type by the value of the parameter θ. We are able to choose a coin at random and the type chosen has a probability density $f(\theta)$.

It may help to see that the discrete examples we did in previous classes are similar. In one example, we had three types of coin with probability of heads 0.5, 0.6, or 0.9. So, we called our hypotheses $H_{0.5}$, $H_{0.6}$, $H_{0.9}$ and these had prior probabilities $P(H_{0.5})$ etc. In other words, we had a type of coin with an unknown probability of heads, we had hypotheses about that probability and each of these hypotheses had a prior probability.

7 The law of total probability

The law of total probability for continuous probability distributions is essentially the same as for discrete distributions. We replace the prior pmf by a prior pdf and the sum by an integral. We start by reviewing the law for the discrete case.

Recall that for a discrete set of hypotheses $\mathcal{H}_1, \mathcal{H}_2, \ldots \mathcal{H}_n$ the law of total probability says

$$P(\mathcal{D}) = \sum_{i=1}^{n} P(\mathcal{D}|\mathcal{H}_i)P(\mathcal{H}_i). \tag{1}$$

This is the total prior probability of $\mathcal{D}$ because we used the prior probabilities $P(\mathcal{H}_i)$

In the little letter notation with $\theta_1, \theta_2, \ldots, \theta_n$ for hypotheses and x for data the law of total probability is written

$$p(x) = \sum_{i=1}^{n} p(x|\theta_i)p(\theta_i). \tag{2}$$

We also called this the prior predictive probability of the outcome x to distinguish it from the prior probability of the hypothesis θ.

Likewise, there is a law of total probability for continuous pdfs. We state it as a theorem using little letter notation.

Theorem. Law of total probability. Suppose we have a continuous parameter θ in the range $[a, b]$, and discrete random data x. Assume θ is itself random with density $f(\theta)$ and that x and θ have likelihood $p(x|\theta)$. In this case, the total probability of x is given by the formula.

$$p(x) = \int_a^b p(x|\theta)f(\theta)\,d\theta \tag{3}$$

Proof. Our proof will be by analogy to the discrete version: The probability term $p(x|\theta)f(\theta)\,d\theta$ is perfectly analogous to the term $p(x|\theta_i)p(\theta_i)$ in Equation 2 (or the term $P(\mathcal{D}|\mathcal{H}_i)P(\mathcal{H}_i)$ in Equation 1). Continuing the analogy: the sum in Equation 2 becomes the integral in Equation 3

As in the discrete case, when we think of θ as a hypothesis explaining the probability of the data we call $p(x)$ the prior predictive probability for x.

Example 5. (Law of total probability.) Continuing with Example 4. We have a bent coin with probability θ of heads. The value of θ is random with prior pdf $f(\theta) = 2\theta$ on $[0, 1]$.

Suppose I am about to flip the coin. What is the total probability of heads, i.e what is the prior predictive probability of heads?

Solution: In Example 4 we noted that the likelihoods are $p(x=1|\theta)=\theta$ and $p(x=0|\theta)=1-\theta$. So the total probability of $x=1$ is

$$p(x=1)=\int_0^1 p(x=1|\theta)\,f(\theta)\,d\theta=\int_0^1 \theta\cdot 2\theta\,d\theta=\int_0^1 2\theta^2 d\theta=\frac{2}{3}.$$

Since the prior is weighted towards higher probabilities of heads, so is the total probability of heads.

8 Bayes' theorem for continuous probability densities

The statement of Bayes' theorem for continuous pdfs is essentially identical to the statement for pmfs. We state it including $d\theta$ so we have genuine probabilities:

Theorem. Bayes' Theorem. Use the same assumptions as in the law of total probability, i.e. θ is a continuous parameter with pdf $f(\theta)$ and range $[a,b]$; x is random discrete data; together they have likelihood $p(x|\theta)$. With these assumptions:

$$f(\theta|x)\,d\theta=\frac{p(x|\theta)f(\theta)\,d\theta}{p(x)}=\frac{p(x|\theta)f(\theta)\,d\theta}{\int_a^b p(x|\theta)f(\theta)\,d\theta}. \qquad (4)$$

Proof. Since this is a statement about probabilities it is just the usual statement of Bayes' theorem. We hope this is clear.

It is important enough to spell out somewhat formally: Let Θ be the random variable that produces the value θ. Consider the events

$$\mathcal{H}=\text{‘}\Theta\text{ is in an interval of width } d\theta \text{ around the value } \theta\text{’}$$

and

$$\mathcal{D}=\text{‘the value of the data is } x\text{’}.$$

Then $P(\mathcal{H})=f(\theta)\,d\theta$, $P(\mathcal{D})=p(x)$, and $P(\mathcal{D}|\mathcal{H})=p(x|\theta)$. Now our usual form of Bayes' theorem becomes

$$f(\theta|x)\,d\theta=P(\mathcal{H}|\mathcal{D})=\frac{P(\mathcal{D}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}=\frac{p(x|\theta)f(\theta)\,d\theta}{p(x)}$$

Looking at the first and last terms in this equation we see the new form of Bayes' theorem.

Finally, we firmly believe that it is more conducive to careful thinking about probability to keep the factor of $d\theta$ in the statement of Bayes' theorem. But because it appears in the numerator on both sides of Equation 4 many people drop the $d\theta$ and write Bayes' theorem in terms of densities as

$$f(\theta|x)=\frac{p(x|\theta)f(\theta)}{p(x)}=\frac{p(x|\theta)f(\theta)}{\int_a^b p(x|\theta)f(\theta)\,d\theta}.$$

9 Bayesian updating with continuous priors

Now that we have Bayes' theorem and the law of total probability we can finally get to Bayesian updating. Before continuing with Example 4, we point out two features of the Bayesian updating table that appears in the next example:

1. The table for continuous priors is very simple: since we cannot have a row for each of an infinite number of hypotheses we'll have just one row which uses a variable to stand for all hypotheses $\mathscr{H}_\theta$.

2. By including $d\theta$, all the entries in the table are probabilities and all our usual probability rules apply.

Example 6. (Bayesian updating.) Continuing Examples 4 and 5. We have a bent coin with unknown probability θ of heads. The value of θ is random with prior pdf $f(\theta) = 2\theta$. Suppose we flip the coin three times and get the sequence HTT. Compute the posterior pdf for θ.

Solution: We make the usual update table, with an added column giving the range of values that θ can take. We make the first row an abstract version of Bayesian updating and the second row is Bayesian updating for this particular example. In later examples we will skip that abstract version.

hypothesis	range	prior	likelihood	Bayes numerator	posterior
$\mathscr{H}_\theta$	θ range	$f(\theta)\,d\theta$	$p(x=1,1,0\|\theta)$	$p(x=1,1,0\|\theta)f(\theta)\,d\theta$	$f(\theta\|x=1,1,0)\,d\theta$
$\mathscr{H}_\theta$	$[0,1]$	$2\theta\,d\theta$	$\theta^2(1-\theta)$	$2\theta^3(1-\theta)\,d\theta$	$20\theta^3(1-\theta)\,d\theta$
total	$[0,1]$	$\int_a^b f(\theta)\,d\theta = 1$	no sum	$p(x=1,1,0)$ $= \int_0^1 2\theta^3(1-\theta)\,d\theta = 1/10$	1

Therefore the posterior pdf (after seeing HHT) is $\boxed{f(\theta|x) = 20\theta^3(1-\theta)}$.

We have a number of comments:

1. Since we used the prior probability $f(\theta)\,d\theta$, the hypothesis should have been:
'the unknown paremeter is in an interval of width $d\theta$ around θ'.
Even for us that is too much to write, so you will have to think it everytime we write that the hypothesis is θ or $\mathscr{H}_\theta$.

2. The posterior pdf for θ is found by removing the $d\theta$ from the posterior probability in the table.

$$f(\theta|x) = 20\theta^3(1-\theta).$$

3. (i) As always $p(x)$ is the total probability. Since we have a continuous distribution instead of a sum we compute an integral.

(ii) Notice that by including $d\theta$ in the table, it is clear what integral we need to compute to find the total probability $p(x)$.

4. The table organizes the continuous version of Bayes' theorem. Namely, the posterior pdf is related to the prior pdf and likelihood function via:

$$f(\theta|x)\,d\theta = \frac{p(x|\theta)\,f(\theta)d\theta}{\int_a^b p(x|\theta)f(\theta)\,d\theta} = \frac{p(x|\theta)\,f(\theta)}{p(x)}\,d\theta$$

Removing the $d\theta$ in the numerator of both sides we have the statement in terms of densities.

5. Regarding both sides as functions of θ, we can again express Bayes' theorem in the form:

$$f(\theta|x) \propto p(x|\theta)\cdot f(\theta)$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

9.1 Flat priors

One important prior is called a flat or uniform prior. A flat prior assumes that every hypothesis is equally probable. For example, if θ has range $[0,1]$ then $f(\theta) = 1$ is a flat prior.

Example 7. (Flat priors.) We have a bent coin with unknown probability θ of heads. Suppose we toss it once and get heads. Assume a flat prior and find the posterior probability for θ.

Solution: This is similar Example 6 with a different prior and data.

hypothesis	range	prior	likelihood	Bayes numerator	posterior
θ	θ	$f(\theta)\,d\theta$	$p(x=1\|\theta)$		$f(\theta\|x=1)\,d\theta$
θ	$[0,1]$	$1\cdot d\theta$	θ	$\theta\,d\theta$	$2\theta\,d\theta$
total	$[0,1]$	$\int_a^b f(\theta)\,d\theta = 1$	no sum	$p(x=1) = \int_0^1 \theta\,d\theta = 1/2$	1

9.2 Using the posterior pdf

Example 8. In the previous example the prior probability was flat. First show that this means that a priori the coin is equally like to be biased towards heads or tails. Then, after observing one heads, what is the (posterior) probability that the coin is biased towards heads?

Solution: Since the parameter θ is the probability the coin lands heads, the first part of the problem asks us to show $P(\theta > 0.5) = 0.5$ and the second part asks for $P(\theta > 0.5\,|\,x = 1)$. These are easily computed from the prior and posterior pdfs respectively.

The prior probability that the coin is biased towards heads is

$$P(\theta > 0.5) = \int_{0.5}^1 f(\theta)\,d\theta = \int_{0.5}^1 1\cdot d\theta = \theta\Big|_{0.5}^1 = \frac{1}{2}.$$

The probability of 1/2 means the coin is equally likely to be biased toward heads or tails. The posterior probabilitiy that it's biased towards heads is

$$P(\theta > 0.5|x = 1) = \int_{0.5}^1 f(\theta|x=1)\,d\theta = \int_{0.5}^1 2\theta\,d\theta = \theta^2\Big|_{0.5}^1 = \frac{3}{4}.$$

We see that observing one heads has increased the probability that the coin is biased towards heads from 1/2 to 3/4.

10 Predictive probabilities

Just as in the discrete case we are also interested in using the posterior probabilities of the hypotheses to make predictions for what will happen next.

Example 9. (Prior and posterior prediction.) Continuing Examples 4, 5, 6: we have a coin with unknown probability θ of heads and the value of θ has prior pdf $f(\theta) = 2\theta$. Find the prior predictive probability of heads. Then suppose the first flip was heads and find the posterior predictive probabilities of both heads and tails on the second flip.

Solution: For notation let x_1 be the result of the first flip and let x_2 be the result of the second flip. The prior predictive probability is exactly the total probability computed in Examples 5 and 6.

$$p(x_1 = 1) = \int_0^1 p(x_1 = 1|\theta) f(\theta)\, d\theta = \int_0^1 2\theta^2\, d\theta = \frac{2}{3}.$$

The posterior predictive probabilities are the total probabilities computed using the posterior pdf. From Example 6 we know the posterior pdf is $f(\theta|x_1 = 1) = 3\theta^2$. So the posterior predictive probabilities are

$$\begin{aligned} p(x_2 = 1|x_1 = 1) &= \int_0^1 p(x_2 = 1|\theta, x_1 = 1) f(\theta|x_1 = 1)\, d\theta = \int_0^1 \theta \cdot 3\theta^2\, d\theta = 3/4 \\ p(x_2 = 0|x_1 = 1) &= \int_0^1 p(x_2 = 0|\theta, x_1 = 1) f(\theta|x_1 = 1)\, d\theta = \int_0^1 (1-\theta) \cdot 3\theta^2\, d\theta = 1/4 \end{aligned}$$

(More simply, we could have computed $p(x_2 = 0|x_1 = 1) = 1 - p(x_2 = 1|x_1 = 1) = 1/4$.)

11 (Optional) From discrete to continuous Bayesian updating

This section is optional. In it we will try to develop intuition for the transition from discrete to continuous Bayesian updating. We'll walk a familiar road from calculus. Namely we will:

(i) divide the continuous range of hypotheses into a finite number of short intervals.

(ii) create the discrete updating table for the finite number of hypotheses.

(iii) consider how the table changes as the number of hypotheses goes to infinity.

In this way, we'll see the prior and posterior pmfs converge to the prior and posterior pdfs.

Example 10. To keep things concrete, we will work with the same prior and data as in Example 7. We have a 'bent' coin with a flat prior $f(\theta) = 1$. Our data is we tossed the coin once and got heads.

Our goal is to go from discrete to continuous by increasing the number of hypotheses.

4 hypotheses. Suppose we have four types of coins that have probability of heads 1/8, 3/8, 5/8 and 7/8 respectively. If one coin is chosen at random, our hypotheses for its type are

$$\mathcal{H}_1 : \theta = 1/8, \quad \mathcal{H}_2 : \theta = 3/8, \quad \mathcal{H}_3 : \theta = 5/8, \quad \mathcal{H}_4 : \theta = 7/8.$$

To get this, we divided $[0,1]$ into 4 equal intervals: $[0,1/4]$, $[1/4,1/2]$, $[1/2,3/4]$, $[3/4,1]$. Each interval has width $\Delta\theta = 1/4$. We put our the value of θ for our coin types at the centers of the four intervals.

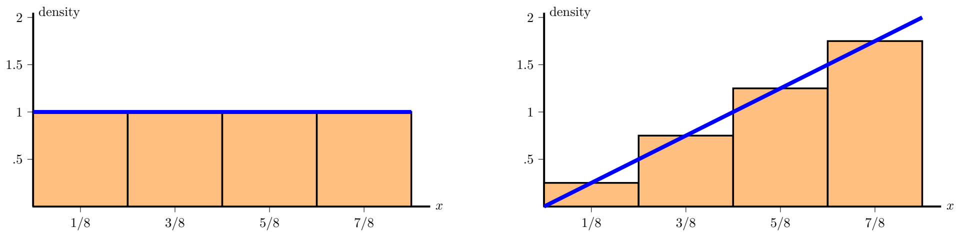
(Just as with forming Riemann sums in calculus, it's not important where in each interval we choose θ. The center is one easy choice.)

Let's name each of these values $\theta_j = j/8$, where $j = 1, 3, 5, 7$.

The flat prior gives each hypothesis a probability of $1/4 = 1 \cdot \Delta\theta$. We have the table:

hypothesis	prior	likelihood	Bayes num.	posterior
$\theta = \theta_1 = \frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4} \cdot \frac{1}{8}$	0.0625
$\theta = \theta_2 = \frac{3}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4} \cdot \frac{3}{8}$	0.1875
$\theta = \theta_3 = \frac{5}{8}$	$\frac{1}{4}$	$\frac{5}{8}$	$\frac{1}{4} \cdot \frac{5}{8}$	0.3125
$\theta = \theta_4 = \frac{7}{8}$	$\frac{1}{4}$	$\frac{7}{8}$	$\frac{1}{4} \cdot \frac{7}{8}$	0.4375
Total	1	–	$\sum_{i=1}^{n} \theta_i \, \Delta\theta$	1

Here are the density histograms of the prior and posterior pmf. The prior and posterior pdfs from Example 7 are superimposed on the histograms in red.

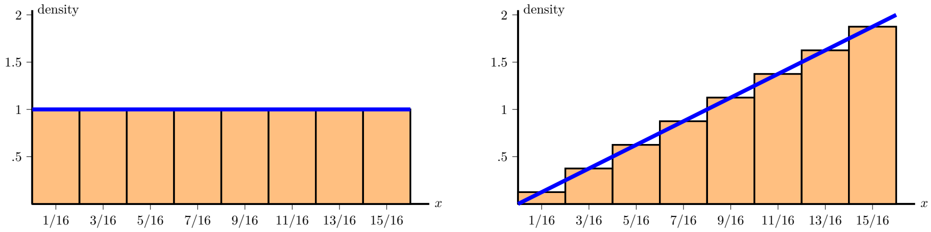


8 hypotheses. Next we slice [0,1] into 8 intervals each of width $\Delta\theta = 1/8$ and use the center of each slice for our 8 hypotheses θ_i.

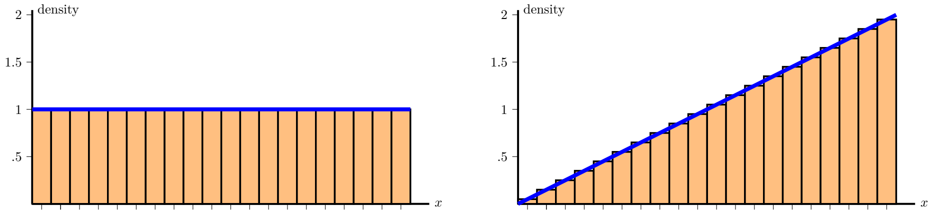
θ_1: '$\theta = 1/16$', θ_2: '$\theta = 3/16$', θ_3: '$\theta = 5/16$', θ_4: '$\theta = 7/16$'
θ_5: '$\theta = 9/16$', θ_6: '$\theta = 11/16$', θ_7: '$\theta = 13/16$', θ_8: '$\theta = 15/16$'

The flat prior gives each hypothesis the probablility $1/8 = 1 \cdot \Delta\theta$. Here are the table and density histograms.

hypothesis	prior	likelihood	Bayes num.	posterior
$\theta = \theta_1 = \frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{8} \cdot \frac{1}{16}$	0.0156
$\theta = \theta_2 = \frac{3}{16}$	$\frac{1}{8}$	$\frac{3}{16}$	$\frac{1}{8} \cdot \frac{3}{16}$	0.0469
$\theta = \theta_3 = \frac{5}{16}$	$\frac{1}{8}$	$\frac{5}{16}$	$\frac{1}{8} \cdot \frac{5}{16}$	0.0781
$\theta = \theta_4 = \frac{7}{16}$	$\frac{1}{8}$	$\frac{7}{16}$	$\frac{1}{8} \cdot \frac{7}{16}$	0.1094
$\theta = \theta_5 = \frac{9}{16}$	$\frac{1}{8}$	$\frac{9}{16}$	$\frac{1}{8} \cdot \frac{9}{16}$	0.1406
$\theta = \theta_6 = \frac{11}{16}$	$\frac{1}{8}$	$\frac{11}{16}$	$\frac{1}{8} \cdot \frac{11}{16}$	0.1719
$\theta = \theta_7 = \frac{13}{16}$	$\frac{1}{8}$	$\frac{13}{16}$	$\frac{1}{8} \cdot \frac{13}{16}$	0.2031
$\theta = \theta_8 = \frac{15}{16}$	$\frac{1}{8}$	$\frac{15}{16}$	$\frac{1}{8} \cdot \frac{15}{16}$	0.2344
Total	1	–	$\sum_{i=1}^{n} \theta_i \, \Delta\theta$	1



20 hypotheses. Finally we slice [0,1] into 20 pieces. This is essentially identical to the previous two cases. Let's skip right to the density histograms.



Looking at the sequence of plots we see how the prior and posterior density histograms converge to the prior and posterior probability density functions.

Notational conventions
Class 13, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to work with the various notations and terms we use to describe probabilities and likelihood.

2 Introduction

We've introduced a number of different notations for probability, hypotheses and data. We collect them here, to have them in one place.

3 Notation and terminology for data and hypotheses

The problem of labeling data and hypotheses is a tricky one. When we started the course we talked about outcomes, e.g. heads or tails. Then when we introduced random variables we gave outcomes numerical values, e.g. 1 for heads and 0 for tails. This allowed us to do things like compute means and variances. We need to do something similar now. Recall our notational conventions:

- Events are labeled with capital letters, e.g. A, B, C.
- A random variable is capital X and takes values small x.
- The connection between values and events: '$X = x$' is the event that X takes the value x.
- The probability of an event is capital $P(A)$.
- A discrete random variable has a probability mass function small $p(x)$ The connection between P and p is that $P(X = x) = p(x)$.
- A continuous random variable has a probability density function $f(x)$ The connection between P and f is that $P(a \le X \le b) = \int_a^b f(x)\,dx$.
- For a continuous random variable X the probability that X is in an infinitesimal interval of width dx around x is $f(x)\,dx$.

In the context of Bayesian updating we have similar conventions.

- We use capital letters, especially $\mathscr{H}$, to indicate a hypothesis, e.g. $\mathscr{H} =$ 'the coin is fair'.

- We use lower case letters, especially θ, to indicate the hypothesized value of a model parameter, e.g. the probability the coin lands heads is $\theta = 0.5$.
- We use upper case letters, especially $\mathcal{D}$, when talking about data as events. For example, $\mathcal{D} =$ 'the sequence of tosses was HTH.
- We use lower case letters, especially x, when talking about data as values. For example, the sequence of data was $x_1, x_2, x_3 = 1, 0, 1$.
- When the set of hypotheses is discrete we can use the probability of individual hypotheses, e.g. $p(\theta)$. When the set is continuous we need to use the probability for an infinitesimal range of hypotheses, e.g. $f(\theta)\,d\theta$.

The following table summarizes this for discrete θ and continuous θ. In both cases we are assuming a discrete set of possible outcomes (data) x. Tomorrow we will deal with a continuous set of outcomes.

	hypothesis	prior	likelihood	Bayes numerator	posterior
	$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{D})$
Discrete θ:	θ	$p(\theta)$	$p(x\|\theta)$	$p(x\|\theta)p(\theta)$	$p(\theta\|x)$
Continuous θ:	θ	$f(\theta)\,d\theta$	$p(x\|\theta)$	$p(x\|\theta)f(\theta)\,d\theta$	$f(\theta\|x)\,d\theta$

Remember the continuous hypothesis θ is really a shorthand for 'the parameter θ is in an interval of width $d\theta$ around θ'.

Continuous Data with Continuous Priors
Class 14, 18.05
Jeremy Orloff and Jonathan Bloom

This reading is not assigned. It goes into a little more detail on Bayesian updating where both hypotheses and data are continuous.

1 Learning Goals

1. Be able to construct a Bayesian update table for continuous hypotheses and continuous data.
2. Be able to recognize the pdf of a normal distribution and determine its mean and variance.

2 Introduction

We are now ready to do Bayesian updating when both the hypotheses and the data take continuous values. The pattern is the same as what we've done before, so let's first review the previous two cases.

3 Previous cases

1. Discrete hypotheses, discrete data

Notation

- Hypotheses $\mathcal{H}$
- Data x
- Prior $P(\mathcal{H})$
- Likelihood $p(x \,|\, \mathcal{H})$
- Posterior $P(\mathcal{H} \,|\, x)$.

Example 1. Suppose we have data x and three possible explanations (hypotheses) for the data that we'll call A, B, C. Suppose also that the data can take two possible values, -1 and 1.

In order to use the data to help estimate the probabilities of the different hypotheses we need a prior pmf and a likelihood table. Assume the prior and likelihoods are given in the following table. (For this example we are only concerned with the formal process of Bayesian updating. So we just made up the prior and likelihoods.)

hypothesis	prior
$\mathcal{H}$	$P(\mathcal{H})$
A	0.1
B	0.3
C	0.6

Prior probabilities

hypothesis	likelihood $p(x \mid \mathcal{H})$	
$\mathcal{H}$	$x = -1$	$x = 1$
A	0.2	0.8
B	0.5	0.5
C	0.7	0.3

Likelihoods

Naturally, each entry in the likelihood table is a likelihood $p(x \mid \mathcal{H})$. For instance the 0.2 row A and column $x = -1$ is the likelihood $p(x = -1 \mid A)$.

Question: Suppose we run one trial and obtain the data $x_1 = 1$. Use this to find the posterior probabilities for the hypotheses.

Solution: The data picks out one column from the likelihood table which we then use in our Bayesian update table.

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$p(x = 1 \mid \mathcal{H})$	$p(x \mid \mathcal{H})P(\mathcal{H})$	$P(\mathcal{H} \mid x) = \dfrac{p(x \mid \mathcal{H})P(\mathcal{H})}{p(x)}$
A	0.1	0.8	0.08	0.195
B	0.3	0.5	0.15	0.366
C	0.6	0.3	0.18	0.439
total	1	no sum	$p(x) = 0.41$	1

To summarize: the prior probabilities of hypotheses and the likelihoods of data given hypothesis were given; the Bayes numerator is the product of the prior and likelihood; the total probability $p(x)$ is the sum of the probabilities in the Bayes numerator column; and we divide by $p(x)$ to normalize the Bayes numerator.

Note: As usual, the term 'no sum' in the likelihood column is not literally true. What it means is that the sum is not meaningful to us. In particular, we don't expect the likelihood column to sum to 1.

2. Continuous hypotheses, discrete data

Now suppose that we have data x that can take a discrete set of values and a continuous parameter θ that determines the distribution the data is drawn from.

Notation

- Hypotheses θ
- Data x
- Prior $f(\theta)\,d\theta$
- Likelihood $p(x \mid \theta)$
- Posterior $f(\theta \mid x)\,d\theta$.

Note: Here we multiplied by $d\theta$ to express the prior and posterior as probabilities. As densities, we have the prior pdf $f(\theta)$ and the posterior pdf $f(\theta\,|\,x)$.

Example 2. Assume that $x \sim \text{Binomial}(5, \theta)$. So θ is in the range $[0, 1]$ and the data x can take six possible values, 0, 1, ..., 5.

Since there is a continuous range of values we use a pdf to describe the prior on θ. Let's suppose the prior is $f(\theta) = 2\theta$. We can still make a likelihood table, though it only has one row representing an arbitrary hypothesis θ.

hypothesis	likelihood $p(x\,\|\,\theta)$					
	$x = 0$	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$
θ	$\binom{5}{0}(1-\theta)^5$	$\binom{5}{1}\theta(1-\theta)^4$	$\binom{5}{2}\theta^2(1-\theta)^3$	$\binom{5}{3}\theta^3(1-\theta)^2$	$\binom{5}{4}\theta^4(1-\theta)$	$\binom{5}{5}\theta^5$

Likelihoods

Question: Suppose we run one trial and obtain the data $x = 2$. Use this to find the posterior pdf for the parameter (hypotheses) θ.

Solution: As before, the data picks out one column from the likelihood table which we can use in our Bayesian update table. Since we want to work with probabilities we write $f(\theta)d\,\theta$ and $f(\theta\,|\,x)\,d\theta$ for the pdfs.

hypothesis	prior	likelihood (for $x = 2$)	Bayes numerator	posterior
θ	$f(\theta)\,d\theta$	$p(x\,\|\,\theta)$	$p(x\,\|\,\theta)f(\theta)\,d\theta$	$f(\theta\,\|\,x)\,d\theta = \dfrac{p(x\,\|\,\theta)f(\theta)\,d\theta}{p(x)}$
θ	$2\theta\,d\theta$	$\binom{5}{2}\theta^2(1-\theta)^3$	$2\binom{5}{2}\theta^3(1-\theta)^3\,d\theta$	$f(\theta\,\|\,x)\,d\theta = \dfrac{7!}{3!\,3!}\theta^3(1-\theta)^3\,d\theta$
total	1	no sum	$p(x) \;= \int_0^1 2\binom{5}{2}\theta^3(1-\theta)^3\,d\theta$ $= 2\binom{5}{2}\frac{3!\,3!}{7!}$	1

To summarize: the prior probabilities of hypotheses and the likelihoods of data given hypothesis were given; the Bayes numerator is the product of the prior and likelihood; the total probability $p(x)$ is the integral of the probabilities in the Bayes numerator column; and we divide by $p(x)$ to normalize the Bayes numerator.

4 Continuous hypotheses and continuous data

When both data and hypotheses are continuous, the only change to the previous example is that the likelihood function uses a pdf $\phi(x\,|\,\theta)$ instead of a pmf $p(x\,|\,\theta)$. The general shape of the Bayesian update table is the same.

Notation

- Hypotheses θ. For continuous hypotheses, this really means that we hypothesize that the parameter is in a small interval of size $d\theta$ around θ.
- Data x. For continuous data, this really means that the data is in a small interval of size dx around x.
- Prior $f(\theta)d\theta$. This is our initial belief about the probability that the parameter is in a small interval of size $d\theta$ around θ.
- Likelihood $\phi(x \mid \theta)\,dx$. This is the (calculated) probability that the data is in a small interval of size dx around x, ASSUMING the hypothesis θ.
- Posterior $f(\theta \mid x)\,d\theta$. This is the (calculated) probability that the parameter is in a small interval of size $d\theta$ around θ, GIVEN the data x.

Simplifying the notation. In the previous cases we included $d\theta$ so that we were working with probabilities instead of densities. When both data and hypotheses are continuous we will need both $d\theta$ and dx. This makes things conceptually simpler, but notationally cumbersome. To simplify the notation we will sometimes allow ourselves to drop dx in our tables. This is fine because the data x is fixed in each calculation. We keep the $d\theta$ because the hypothesis θ is allowed to vary.

For comparison, we first show the general table in simplified notation followed immediately afterward by the table showing both infinitesimals.

			Bayes	
hypoth.	prior	likelihood	numerator	posterior
θ	$f(\theta)\,d\theta$	$\phi(x \mid \theta)$	$\phi(x \mid \theta)f(\theta)\,d\theta$	$f(\theta \mid x)\,d\theta = \dfrac{\phi(x \mid \theta)f(\theta)\,d\theta}{\phi(x)}$
total	1	no sum	$\phi(x) = \int \phi(x \mid \theta)f(\theta)\,d\theta$	1
(integrate over θ)			= prior prob. density for data x	

Bayesian update table without dx

			Bayes	
hypoth.	prior	likelihood	numerator	posterior
θ	$f(\theta)\,d\theta$	$\phi(x \mid \theta)\,dx$	$\phi(x \mid \theta)f(\theta)\,d\theta\,dx$	$f(\theta \mid x)\,d\theta = \dfrac{\phi(x \mid \theta)f(\theta)\,d\theta\,dx}{\phi(x)\,dx}$ $= \frac{\phi(x \mid \theta)f(\theta)\,d\theta}{\phi(x)}$
total	1	no sum	$\phi(x)\,dx = \left(\int \phi(x \mid \theta)f(\theta)\,d\theta\right)\,dx$	1

Bayesian update table with $d\theta$ and dx

To summarize: the prior probabilities of hypotheses and the likelihoods of data given hypothesis were given; the Bayes numerator is the product of the prior and likelihood; the total probability $\phi(x)\,dx$ is the integral of the probabilities in the Bayes numerator column; we divide by $\phi(x)\,dx$ to normalize the Bayes numerator.

5 A digression on notational messes

We have chosen to use the notation $\phi(x)$, $\phi(x\,|\,\theta)$ for the pdfs of data and $f(\theta)$, $f(\theta\,|\,x)$ for the pdfs of hypotheses. This is nice because ϕ is a Greek f, but the different symbols help us distinguish the two types of pdfs. Many, perhaps most, writers use the same letter f for both. This forces the reader to look at the arguments to the function to understand what is meant. That is, $f(x|\theta)$ is the probability of data given an hypothesis, i.e. the likelihood and $f(\theta|x)$ is the probability of an hypothesis given the data, i.e. the posterior pdf.

As mathematicians this makes us pull our hair out. But, to be fair, there is a philosophical underpinning to this notation. We can think of f as a universal probability density which gives the probability of absolutely any combination of things. Thus $f(x, y)$ is the joint probability density for the quantities denoted by x and y. If we just write $f(x)$ the implication is that this means the marginal density for x, i.e. the density for x when y is allowed to take any value. Similarly we can write $f(x, y|z)$ for the conditional density of x and y given z.

6 Normal hypothesis, normal data

A standard example of continuous hypotheses and continuous data assumes that both the data and prior follow normal distributions. The following example assumes that the variance of the data is known.

Example 3. Suppose we have data $x = 5$ which was drawn from a normal distribution with unknown mean θ and standard deviation 1.

$$x \sim \mathrm{N}(\theta, 1)$$

Suppose further, that our prior distribution for the unknown parameter θ is $\theta \sim \mathrm{N}(2, 1)$.

Let x represent an arbitrary data value.

(a) Make a Bayesian table with prior, likelihood, and Bayes numerator.

(b) Show that the posterior distribution for θ is normal as well.

(c) Find the mean and variance of the posterior distribution.

Solution: As we did with the tables above, a good compromise on the notation is to include $d\theta$ but not dx. The reason for this is that the total probability is computed by integrating over θ and the $d\theta$ reminds of us that.

Our prior pdf is

$$f(\theta) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(\theta-2)^2/2}.$$

The likelihood function is

$$\phi(x = 5\,|\,\theta) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(5-\theta)^2/2}.$$

We know we are going to multiply the prior and the likelihood, so we carry out that algebra first. In the very last step we give the complicated constant factor the name c_1.

$$\begin{aligned}
\text{prior} \cdot \text{likelihood} &= \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(\theta-2)^2/2} \cdot \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(5-\theta)^2/2} \\
&= \frac{1}{2\pi}\mathrm{e}^{-(2\theta^2-14\theta+29)/2} \\
&= \frac{1}{2\pi}\mathrm{e}^{-(\theta^2-7\theta+29/2)} \quad \text{(complete the square)} \\
&= \frac{1}{2\pi}\mathrm{e}^{-((\theta-7/2)^2+9/4)} \\
&= \frac{\mathrm{e}^{-9/4}}{2\pi}\mathrm{e}^{-(\theta-7/2)^2)} \\
&= c_1\mathrm{e}^{-(\theta-7/2)^2}
\end{aligned}$$

In the last step we named the complicated constant factor c_1.

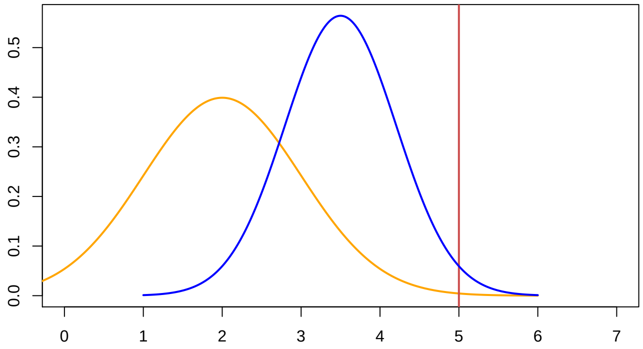
hypothesis	prior	likelihood	Bayes numerator	posterior $f(\theta\,\vert\,x=5)\,d\theta$
θ	$f(\theta)\,d\theta$	$\phi(x=5\,\vert\,\theta)$	$\phi(x=5\,\vert\,\theta)f(\theta)\,d\theta$	$\dfrac{\phi(x=5\,\vert\,\theta)f(\theta)\,d\theta}{\phi(x=5)}$
θ	$\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(\theta-2)^2/2}\,d\theta$	$\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(5-\theta)^2/2}$	$c_1\mathrm{e}^{-(\theta-7/2)^2}$	$c_2\mathrm{e}^{-(\theta-7/2)^2}$
total	1	no sum	$\phi(x=5)=\int\phi(x=5\,\vert\,\theta)f(\theta)\,d\theta$	1

We can see by the form of the posterior pdf that it is a normal distribution. Because the exponential for a normal distribution is $\mathrm{e}^{-(\theta-\mu)^2/2\sigma^2}$ we have mean $\mu = 7/2$ and $2\sigma^2 = 1$, so variance $\sigma^2 = 1/2$.

We don't need to bother computing the total probability; it is just used for normalization and we already know the normalization constant $\dfrac{1}{\sigma\sqrt{2\pi}}$ for a normal distribution. To summarize,

The posterior pdf follows a $\mathrm{N}(7/2, 1/2)$ distribution.

Here is the graph of the prior and the posterior pdfs for this example. Note how the data 'pulls' the prior (the wider bell on the left) towards the data. The posterior is the narrower bell on the right. After collecting data, we have a new opinion about the mean, and we are more sure of this new opinion.



prior = orange; posterior = blue; data = red line

Now we'll repeat the previous example for general x. When reading this if you mentally substitute 5 for x you will understand the algebra.

Example 4. Suppose our data x is drawn from a normal distribution with unknown mean θ and standard deviation 1.

$$x \sim \mathrm{N}(\theta, 1)$$

Suppose further, that our prior distribution for the unknown parameter θ is $\theta \sim \mathrm{N}(2, 1)$.

Solution: As before, we show the algebra used to simplify the Bayes numerator: The prior pdf and likelihood function are

$$f(\theta) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(\theta-2)^2/2} \qquad f(x\,|\,\theta) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(x-\theta)^2/2}.$$

The Bayes numerator is the product of the prior and the likelihood:

$$\begin{aligned}
\text{prior} \cdot \text{likelihood} &= \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(\theta-2)^2/2} \cdot \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(x-\theta)^2/2} \\
&= \frac{1}{2\pi}\mathrm{e}^{-(2\theta^2-(4+2x)\theta+4+x^2)/2} \\
&= \frac{1}{2\pi}\mathrm{e}^{-(\theta^2-(2+x)\theta+(4+x^2)/2)} \qquad \text{(complete the square)} \\
&= \frac{1}{2\pi}\mathrm{e}^{-((\theta-(1+x/2))^2-(1+x/2)^2+(4+x^2)/2)} \\
&= c_1\mathrm{e}^{-(\theta-(1+x/2))^2}
\end{aligned}$$

Just as in the previous example, in the last step we replaced all the constants, including the exponentials that just involve x, by by the simple constant c_1.

Now the Bayesian update table becomes

hypothesis	prior	likelihood	Bayes numerator	posterior $f(\theta \mid x)\,d\theta$
θ	$f(\theta)\,d\theta$	$\phi(x \mid \theta)$	$\phi(x \mid \theta)f(\theta)\,d\theta$	$\dfrac{\phi(x \mid \theta)f(\theta)\,d\theta}{\phi(x)}$
θ	$\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(\theta-2)^2/2}\,d\theta$	$\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(x-\theta)^2/2}$	$c_1\mathrm{e}^{-(\theta-(1+x/2))^2}$	$c_2\mathrm{e}^{-(\theta-(1+x/2))^2}$
θ	$\theta \sim \mathrm{N}(2,1)$	$x \sim \mathrm{N}(\theta,1)$		$\theta \sim \mathrm{N}(1+x/2, 1/2)$
total	1	no sum	$\phi(x) = \int \phi(x \mid \theta)f(\theta)\,d\theta$	1

As in the previous example we can see by the form of the posterior that it must be a normal distribution with mean $1+x/2$ and variance $1/2$. That is,

The posterior pdf follows a $\mathrm{N}(1+x/2, 1/2)$ distribution.

You should compare this with the case $x=5$ in the previous example.

7 Predictive probabilities

Since the data x is continuous it has prior and posterior predictive pdfs. The prior predictive pdf is the total probability density computed at the bottom of the Bayes numerator column:

$$\phi(x) = \int f(x|\theta)f(\theta)\,d\theta,$$

where the integral is computed over the entire range of θ.

The posterior predictive pdf has the same form as the prior predictive pdf, except it uses the posterior probabilities for θ:

$$\phi(x_2|x_1) = \int \phi(x_2|\theta, x_1)f(\theta|x_1)\,d\theta,$$

We usually assume that x_1 and x_2 are conditionally independent. That is,

$$\phi(x_2|\theta, x_1) = \phi(x_2|\theta).$$

In this case the formula for the posterior predictive pdf is a little simpler:

$$\phi(x_2|x_1) = \int \phi(x_2|\theta)f(\theta|x_1)\,d\theta.$$

Conjugate priors: Beta and normal
Class 15, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be familiar with the 2-parameter family of beta distributions and its normalization.
2. Understand the benefits of conjugate priors.
3. Be able to update a beta prior given a Bernoulli, binomial, or geometric likelihood.
4. Understand and be able to use the formula for updating a normal prior given a normal likelihood with known variance.

2 Introduction

Our main goal here is to introduce the idea of conjugate priors and look at some specific conjugate pairs. These simplify the job of Bayesian updating to simple arithmetic. We'll start by introducing the beta distribution and using it as a conjugate prior with a binomial likelihood. After that we'll look at other conjugate pairs.

3 Beta distribution

The beta distribution Beta(a, b) is a two-parameter distribution with range $[0, 1]$ and pdf

$$f(\theta) = \frac{(a+b-1)!}{(a-1)!(b-1)!}\theta^{a-1}(1-\theta)^{b-1}$$

We have made an applet so you can explore the shape of the beta distribution as you vary the parameters:

`https://mathlets.org/mathlets/beta-distribution/`.

As you can see in the applet, the beta distribution may be defined for any real numbers $a > 0$ and $b > 0$. In 18.05 we will stick to integers a and b, but you can get the full story here: `https://en.wikipedia.org/wiki/Beta_distribution`

In the context of Bayesian updating, a and b are often called hyperparameters to distinguish them from the unknown parameter θ representing our hypotheses. In a sense, a and b are 'one level up' from θ since they parameterize its pdf.

3.1 A simple but important observation!

If a pdf $f(\theta)$ with range $[0, 1]$ has the form $c\theta^{a-1}(1-\theta)^{b-1}$ then $f(\theta)$ is a Beta(a, b) distribution and the normalizing constant must be

$$c = \frac{(a+b-1)!}{(a-1)!\,(b-1)!}.$$

This follows because the constant c must normalize the pdf to have total probability 1. There is only one such constant and it is given in the formula for the beta distribution.

A similar observation holds for normal distributions, exponential distributions, and so on.

3.2 Beta priors and posteriors for binomial random variables

Example 1. Suppose we have a bent coin with unknown probability θ of heads. We toss it 12 times and get 8 heads and 4 tails. Starting with a flat prior, show that the posterior pdf is a Beta$(9, 5)$ distribution.

Solution: This is nearly identical to examples from the previous class. We'll call the data from all 12 tosses x_1. In the following table we call the leading constant factor in the posterior column c_2. Our simple observation will tell us that it has to be the constant factor from the beta pdf.

The data is 8 heads and 4 tails. Since this comes from a binomial(12, θ) distribution, the likelihood $p(x_1|\theta) = \binom{12}{8}\theta^8(1-\theta)^4$. Thus the Bayesian update table is

hypothesis	prior	likelihood	Bayes numerator	posterior
θ	$1 \cdot d\theta$	$\binom{12}{8}\,\theta^8(1-\theta)^4$	$\binom{12}{8}\,\theta^8(1-\theta)^4\,d\theta$	$c_2\,\theta^8(1-\theta)^4\,d\theta$
total	1		$T = \binom{12}{8}\int_0^1 \theta^8(1-\theta)^4\,d\theta$	1

For the posterior pdf, our simple observation holds with $a = 9$ and $b = 5$. Therefore the posterior pdf follows a Beta$(9, 5)$ distribution and we have

$$f(\theta|x_1) = c_2\theta^8(1-\theta)^4, \text{ where } c_2 = \frac{13!}{8!\,4!}.$$

Note: We explicitly included the binomial coefficient $\binom{12}{8}$ in the likelihood. We could just as easily have given it a name, say c_1 and not bothered making its value explicit.

Example 2. Now suppose we toss the same coin again, getting n heads and m tails. Using the posterior pdf of the previous example as our new prior pdf, show that the new posterior pdf is that of a Beta$(9 + n, 5 + m)$ distribution.

Solution: It's all in the table. We'll call the data of these $n+m$ additional tosses x_2. This time we won't make the binomial coefficient explicit. Instead we'll just call it c_3. Whenever we need a new label we will simply use c with a new subscript.

hyp.	prior	likelihood	Bayes numerator	posterior
θ	$c_2\theta^8(1-\theta)^4\,d\theta$	$c_3\,\theta^n(1-\theta)^m$	$c_2c_3\,\theta^{n+8}(1-\theta)^{m+4}\,d\theta$	$c_4\,\theta^{n+8}(1-\theta)^{m+4}\,d\theta$
total	1		$T = \int_0^1 c_2c_3\,\theta^{n+8}(1-\theta)^{m+4}\,d\theta$	1

Again our simple observation holds and therefore the posterior pdf

$$f(\theta|x_1, x_2) = c_4\theta^{n+8}(1-\theta)^{m+4}$$

follows a Beta$(n+9, m+5)$ distribution.

Note: Flat beta. The Beta$(1,1)$ distribution is the same as the uniform distribution on $[0,1]$, which we have also called the flat prior on θ. This follows by plugging $a = 1$ and $b = 1$ into the definition of the beta distribution, giving $f(\theta) = 1$.

Summary: If the probability of heads is θ, the number of heads in $n+m$ tosses follows a binomial$(n+m, \theta)$ distribution. We have seen that if the prior on θ is a beta distribution then so is the posterior; only the parameters a, b of the beta distribution change! We summarize precisely how they change in a table. We assume the data is n heads and m tails in $n+m$ tosses.

hypothesis	data	prior	likelihood	posterior
θ	$x = n,\ m$	Beta(a, b)	binomial$(n+m, \theta)$	Beta$(a+n, b+m)$
θ	$x = n,\ m$	$c_1\theta^{a-1}(1-\theta)^{b-1}\,d\theta$	$c_2\theta^n(1-\theta)^m$	$c_3\theta^{a+n-1}(1-\theta)^{b+m-1}\,d\theta$

4 Conjugate priors

The beta distribution is called a conjugate prior for the binomial distribution. This means that if the likelihood function is binomial, then a beta prior gives a beta posterior –this is what we saw in the previous examples. In fact, the beta distribution is a conjugate prior for the Bernoulli and geometric distributions as well.

We will soon see another important example: the normal distribution is its own conjugate prior. In particular, if the likelihood function is normal with known variance, then a normal prior gives a normal posterior.

Conjugate priors are useful because they reduce Bayesian updating to modifying the parameters of the prior distribution (so-called hyperparameters) rather than computing integrals. We saw this for the beta distribution in the last table. For many more examples see: `https://en.wikipedia.org/wiki/Conjugate_prior_distribution`

We now give a definition of conjugate prior. It is best understood through the examples in the subsequent sections.

Definition. Suppose we have data with likelihood function $\phi(x|\theta)$ depending on a hypothesized parameter θ. Also suppose the prior distribution for θ is one of a family of parametrized distributions. If the posterior distribution for θ is in this family then we say the the family of priors are conjugate priors for the likelihood.

This definition will be illustrated with specific examples in the sections below.

5 Beta priors

In this section, we will show that the beta distribution is a conjugate prior for binomial, Bernoulli, and geometric likelihoods.

5.1 Binomial likelihood

We saw above that the beta distribution is a conjugate prior for the binomial distribution. This means that if the likelihood function is binomial and the prior distribution is beta then the posterior is also beta.

More specifically, suppose that the likelihood follows a binomial(N, θ) distribution where N is known and θ is the (unknown) parameter of interest. We also have that the data x from one trial is an integer between 0 and N. Then for a beta prior we have the following table:

hypoth.	data	prior	likelihood	posterior
θ	x	Beta(a, b) $f(\theta) = c_1\theta^{a-1}(1-\theta)^{b-1}$	binomial(N, θ) $p(x\|\theta) = c_2\theta^x(1-\theta)^{N-x}$	Beta$(a + x, b + N - x)$ $f(\theta\|x) = c_3\theta^{a+x-1}(1-\theta)^{b+N-x-1}$

The table is simplified by writing the normalizing coefficients as c_1, c_2 and c_3 respectively. If needed, we can recover the values of the c_1 and c_2 by recalling (or looking up) the normalizations of the beta and binomial distributions.

$$c_1 = \frac{(a+b-1)!}{(a-1)!\,(b-1)!} \qquad c_2 = \binom{N}{x} = \frac{N!}{x!\,(N-x)!} \qquad c_3 = \frac{(a+b+N-1)!}{(a+x-1)!\,(b+N-x-1)!}$$

5.2 Bernoulli likelihood

The beta distribution is a conjugate prior for the Bernoulli distribution. This is actually a special case of the binomial distribution, since Bernoulli(θ) is the same as binomial(1, θ). We do it separately because it is slightly simpler and of special importance. In the table below, we show the updates corresponding to success $(x = 1)$ and failure $(x = 0)$ on separate rows.

hypothesis	data	prior	likelihood	posterior
θ	x	Beta(a, b)	Bernoulli(θ)	Beta$(a + 1, b)$ or Beta$(a, b + 1)$
θ	$x = 1$	$f(\theta) = c_1\theta^{a-1}(1-\theta)^{b-1}$	$p(x\|\theta) = \theta$	Beta$(a + 1, b)$: $f(\theta\|x) = c_3\theta^a(1-\theta)^{b-1}$
θ	$x = 0$	$f(\theta) = c_1\theta^{a-1}(1-\theta)^{b-1}$	$p(x\|\theta) = 1-\theta$	Beta$(a, b + 1)$: $f(\theta\|x) = c_4\theta^{a-1}(1-\theta)^b$

The constants c_1, c_3 and c_4 have the same formulas as in the previous (binomial likelihood case) with $N = 1$.

5.3 Geometric likelihood

Recall that the geometric(θ) distribution describes the probability of x successes before the first failure, where the probability of success on any single independent trial is θ. The corresponding pmf is given by $p(x) = \theta^x(1-\theta)$.

Now suppose that we have a data point x, and our hypothesis θ is that x is drawn from a geometric(θ) distribution. From the table we see that the beta distribution is a conjugate prior for a geometric likelihood as well:

hypothesis	data	prior	likelihood	posterior
θ	x	Beta(a, b) $= f(\theta) = c_1\theta^{a-1}(1-\theta)^{b-1}$	geometric(θ) $= p(x\|\theta) = \theta^x(1-\theta)$	Beta$(a + x, b + 1)$ $f(\theta\|x) = c_3\theta^{a+x-1}(1-\theta)^b$

At first it may seem strange that the beta distribution is a conjugate prior for both the binomial and geometric distributions. The key reason is that the geometric likelihood is proportional to a binomial likelihood as a function of θ. Let's illustrate this in a concrete example.

Example 3. While traveling through the Mushroom Kingdom, Mario and Luigi find some rather unusual coins. They agree on a prior of $f(\theta) \sim \text{Beta}(5{,}5)$ for the probability of heads, though they disagree on what experiment to run to investigate θ further.

(a) Mario decides to flip a coin 5 times. He gets four heads in five flips.

(b) Luigi decides to flip a coin until the first tails. He gets four heads before the first tail.

Show that Mario and Luigi will arrive at the same posterior on θ, and calculate this posterior.

Solution: We will show that both Mario and Luigi find the posterior pdf for θ is a $\text{Beta}(9, 6)$ distribution.

Mario's table

hypothesis	data	prior	likelihood	posterior
θ	$x = 4$	$\text{Beta}(5,5)$ $= c_1\theta^4(1-\theta)^4$	$\text{binomial}(5,\theta)$ $= \binom{5}{4}\theta^4(1-\theta)$	??? $= c_3\theta^8(1-\theta)^5$

Luigi's table

hypothesis	data	prior	likelihood	posterior
θ	$x = 4$	$\text{Beta}(5,5)$ $= c_1\theta^4(1-\theta)^4$	$\text{geometric}(\theta)$ $= \theta^4(1-\theta)$	??? $= c_3\theta^8(1-\theta)^5$

Since both Mario and Luigi's posteriors have the form of a $\text{Beta}(9, 6)$ distribution that's what they both must be. The normalizing factor must be the same in both cases because it's determined by requiring the total probability to be 1.

6 Bayesian updating with continuous hypotheses and continuous data

The idea here is essentially identical to the Bayesian updating we've already done. The only change is, with a continuous likelihood, we have to compute the total probability of the data (i.e. sum of the Bayes numerator column, i.e. normalizing factor) as an integral instead of a sum. We will cover this briefly. For those who are interested, a bit more detail is given in an optional note.

Notation

- Hypotheses θ. For continuous hypotheses, this really means that we hypothesize that the parameter is in a small interval of size $d\theta$ around θ.
- Data x. For continuous data, this really means that the data is in a small interval of size dx around x.
- Prior $f(\theta)d\theta$. This is our initial belief about the probability that the parameter is in a small interval of size $d\theta$ around θ.

- Likelihood $\phi(x \,|\, \theta)$. So the probability that the data is in a small interval of size dx around x, ASSUMING the hypothesis θ is $\phi(x \,|\, \theta)\, dx$
- Posterior $f(\theta \,|\, x)\, d\theta$. This is the (calculated) probability that the parameter is in a small interval of size $d\theta$ around θ, GIVEN the data x.

hypoth.	prior	likelihood	Bayes numerator	posterior
θ	$f(\theta)\, d\theta$	$\phi(x \,\|\, \theta)$	$\phi(x \,\|\, \theta) f(\theta)\, d\theta$	$f(\theta \,\|\, x)\, d\theta = \dfrac{\phi(x \,\|\, \theta) f(\theta)\, d\theta}{\phi(x)}$
total (integrate over θ)	1	no sum	$\phi(x) = \int \phi(x \,\|\, \theta) f(\theta)\, d\theta$ $=$ prior prob. density for data x	1

Continuous-continuous Bayesian update table

To summarize: the prior probabilities of hypotheses and the likelihoods of data given hypothesis were given; the Bayes numerator is the product of the prior and likelihood; the total likelihood $\phi(x)$ is the integral of the probabilities in the Bayes numerator column; we divide by $\phi(x)$ to normalize the Bayes numerator.

7 Normal begets normal

We now turn to an important example of coninuous-continuous updating: the normal distribution is its own conjugate prior. In particular, if the likelihood function is normal with known variance, then a normal prior gives a normal posterior. Now both the hypotheses and the data are continuous.

Suppose we have a measurement $x \sim N(\theta, \sigma^2)$ where the variance σ^2 is known. That is, the mean θ is our unknown parameter of interest and we are given that the likelihood comes from a normal distribution with variance σ^2. If we choose a normal prior pdf

$$f(\theta) \sim \mathrm{N}(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$$

then the posterior pdf is also normal: $f(\theta|x) \sim \mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$ where

$$\frac{\mu_{\text{post}}}{\sigma^2_{\text{post}}} = \frac{\mu_{\text{prior}}}{\sigma^2_{\text{prior}}} + \frac{x}{\sigma^2}, \qquad \frac{1}{\sigma^2_{\text{post}}} = \frac{1}{\sigma^2_{\text{prior}}} + \frac{1}{\sigma^2} \tag{1}$$

The following form of these formulas is easier to read and shows that μ_{post} is a weighted average between μ_{prior} and the data x.

$$a = \frac{1}{\sigma^2_{\text{prior}}} \qquad b = \frac{1}{\sigma^2}, \qquad \mu_{\text{post}} = \frac{a\mu_{\text{prior}} + bx}{a+b}, \qquad \sigma^2_{\text{post}} = \frac{1}{a+b}. \tag{2}$$

With these formulas in mind, we can express the update via the table:

hypothesis	data	prior	likelihood	posterior
θ	x	$f(\theta) \sim \mathrm{N}(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$ $= c_1 \exp\left(\frac{-(\theta-\mu_{\text{prior}})^2}{2\sigma^2_{\text{prior}}}\right)$	$\phi(x\|\theta) \sim \mathrm{N}(\theta, \sigma^2)$ $= c_2 \exp\left(\frac{-(x-\theta)^2}{2\sigma^2}\right)$	$f(\theta\|x) \sim \mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$ $= c_3 \exp\left(\frac{-(\theta-\mu_{\text{post}})^2}{2\sigma^2_{\text{post}}}\right)$

We leave the proof of the general formulas to the problem set. It is an involved algebraic manipulation which is essentially the same as the following numerical example.

Example 4. Suppose we have prior $\theta \sim \mathrm{N}(4, 8)$, and likelihood function likelihood $x \sim \mathrm{N}(\theta, 5)$. Suppose also that we have one measurement $x_1 = 3$. Show the posterior distribution is normal.

Solution: We will show this by grinding through the algebra which involves completing the square.

prior: $f(\theta) = c_1\,\mathrm{e}^{-(\theta-4)^2/16}$; likelihood: $\phi(x_1|\theta) = c_2\,\mathrm{e}^{-(x_1-\theta)^2/10} = c_2\,\mathrm{e}^{-(3-\theta)^2/10}$

We multiply the prior and likelihood to get the posterior:

$$\begin{aligned} f(\theta|x_1) &= c_3\,\mathrm{e}^{-(\theta-4)^2/16}\,\mathrm{e}^{-(3-\theta)^2/10} \\ &= c_3 \exp\left(-\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10}\right) \end{aligned}$$

We complete the square in the exponent

$$\begin{aligned} -\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10} &= -\frac{5(\theta-4)^2 + 8(3-\theta)^2}{80} \\ &= -\frac{13\theta^2 - 88\theta + 152}{80} \\ &= -\frac{\theta^2 - \frac{88}{13}\theta + \frac{152}{13}}{80/13} \\ &= -\frac{(\theta - 44/13)^2 + 152/13 - (44/13)^2}{80/13}. \end{aligned}$$

Therefore the posterior is

$$f(\theta|x_1) = c_3\,\mathrm{e}^{-\frac{(\theta-44/13)^2 + 152/13 - (44/13)^2}{80/13}} = c_4\,\mathrm{e}^{-\frac{(\theta-44/13)^2}{80/13}}.$$

This has the form of the pdf for $\mathrm{N}(44/13, 40/13)$. QED

For practice we check this against the formulas (2).

$$\mu_{\text{prior}} = 4, \quad \sigma^2_{\text{prior}} = 8, \quad \sigma^2 = 5 \;\Rightarrow\; a = \frac{1}{8}, \quad b = \frac{1}{5}.$$

Therefore

$$\begin{aligned} \mu_{\text{post}} &= \frac{a\mu_{\text{prior}} + bx}{a+b} = \frac{44}{13} = 3.38 \\ \sigma^2_{\text{post}} &= \frac{1}{a+b} = \frac{40}{13} = 3.08. \end{aligned}$$

7.1 A word on weighted averages

The updating formula 2 gives μ_{post} as a weighted average of the μ_{prior} and the data. The weight on μ_{prior} is $a/(a+b)$, and the weight on the data is $b/(a+b)$. These weights are always positive numbers summing to 1. If b is very large (that is, if the data has a tiny variance) then most of the weight is on the data. If a is very large (that is, σ^2_{prior} is small, i.e. if you are very confident in your prior) then most of the weight is on the prior.

In the above example the variance on the prior was bigger than the variance on the data, so a was smaller than b; so the weight was mostly on the data. The posterior 3.38 for the mean was closer to the data 3 than to the prior 4 for the mean.

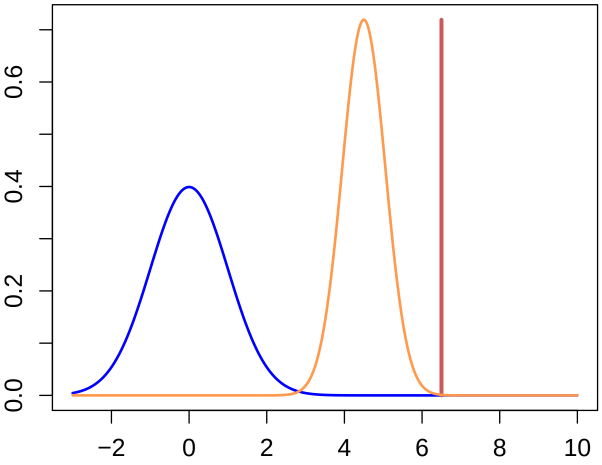
7.2 Examples of normal-normal updating

Example 5. Suppose that we know the data $x \sim \text{N}(\theta, 4/9)$ and we have prior $\text{N}(0,1)$. We get one data value $x = 6.5$. Describe the changes to the pdf for θ in updating from the prior to the posterior.

Solution: $\mu_{\text{prior}} = 0$, $\sigma^2_{\text{prior}} = 1$, $\sigma^2 = 4/9$. So, using the updating formulas 2 we have

$$a = 1, \quad b = \frac{1}{4/9} = \frac{9}{4}, \quad \mu_{\text{post}} = \frac{a\mu_{\text{prior}} + bx}{a+b} = 4.5, \quad \sigma^2_{\text{post}} = \frac{1}{a+b} = \frac{4}{13}.$$

Here is a graph of the prior and posterior pdfs with the data point marked by a red line.



Prior in blue, posterior in orange, data = red line

We see that the posterior mean is closer to the data point than the prior mean We also see that the posterior distribution is taller and narrower than the prior, i.e. it has a smaller variance. The smaller variance says that we are now more certain about where the value of θ lies.

Example 6. Use the formulas 2 to show that for normal-normal Bayesian updating we have:

1. The posterior mean is always between the data point and the prior mean.

2. The posterior variance is smaller than both the prior variance and σ. That is, our

posterior uncetainty is smaller than both our prior uncertainty and the uncertainty in the data.

Solution: Using the update formulas 2, we have The posterior mean is the weighted average of the prior mean and the data, so it must lie between the prior mean and the data.

Also, the posterior variance is

$$\sigma^2_{\text{post}} = \frac{1}{a+b} < \frac{1}{a} = \sigma^2_{\text{prior}}$$

That is the posterior has smaller variance than the prior, i.e. data makes us more certain about where in its range θ lies.

Likewise $\sigma^2_{\text{post}} = \dfrac{1}{a+b} < \dfrac{1}{b} = \sigma^2$. So, the posterior variance is smaller than σ^2.

7.3 More than one data point

Example 7. Suppose we have data x_1, x_2, x_3. Use the formulas (1) to update sequentially.

Solution: Let's label the prior mean and variance as μ_0 and σ_0^2. The updated means and variances will be μ_i and σ_i^2. In sequence we have

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}; \qquad \frac{\mu_1}{\sigma_1^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1}{\sigma^2}$$

$$\frac{1}{\sigma_2^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{2}{\sigma^2}; \qquad \frac{\mu_2}{\sigma_2^2} = \frac{\mu_1}{\sigma_1^2} + \frac{x_2}{\sigma^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1+x_2}{\sigma^2}$$

$$\frac{1}{\sigma_3^2} = \frac{1}{\sigma_2^2} + \frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{3}{\sigma^2}; \qquad \frac{\mu_3}{\sigma_3^2} = \frac{\mu_2}{\sigma_2^2} + \frac{x_3}{\sigma^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1+x_2+x_3}{\sigma^2}$$

The example generalizes to n data values $x_1, \ldots, x_n$:

Normal-normal update formulas for n data points

$$\frac{\mu_{\text{post}}}{\sigma^2_{\text{post}}} = \frac{\mu_{\text{prior}}}{\sigma^2_{\text{prior}}} + \frac{n\bar{x}}{\sigma^2}, \qquad \frac{1}{\sigma^2_{\text{post}}} = \frac{1}{\sigma^2_{\text{prior}}} + \frac{n}{\sigma^2}, \qquad \bar{x} = \frac{x_1 + \ldots + x_n}{n}. \tag{3}$$

Again we give the easier to read form, showing μ_{post} is a weighted average of μ_{prior} and the sample average $\bar{x}$:

$$a = \frac{1}{\sigma^2_{\text{prior}}} \qquad b = \frac{n}{\sigma^2}, \qquad \mu_{\text{post}} = \frac{a\mu_{\text{prior}} + b\bar{x}}{a+b}, \qquad \sigma^2_{\text{post}} = \frac{1}{a+b}. \tag{4}$$

Interpretation: μ_{post} is a weighted average of μ_{prior} and $\bar{x}$. If the number of data points is large then the weight b is large and $\bar{x}$ will have a strong influence on the posterior. If σ^2_{prior} is small then the weight a is large and μ_{prior} will have a strong influence on the posterior. To summarize:

1. Lots of data has a big influence on the posterior.
2. High certainty (low variance) in the prior has a big influence on the posterior.

The actual posterior is a balance of these two influences.

Choosing priors
Class 16, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Learn that the choice of prior affects the posterior.
2. See that too rigid a prior can make it difficult to learn from the data.
3. See that more data lessens the dependence of the posterior on the prior.
4. Be able to make a reasonable choice of prior, based on prior understanding of the system under consideration.

2 Introduction

Up to now we have always been handed a prior pdf. In this case, statistical inference from data is essentially an application of Bayes' theorem. When the prior is known there is no controversy on how to proceed. The art of statistics starts when the prior is not known with certainty. There are two main schools on how to proceed in this case: Bayesian and frequentist. For now we are following the Bayesian approach. Starting next week we will learn the frequentist approach.

Recall that given data D and a hypothesis H we used Bayes' theorem to write

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

$$\text{posterior} \propto \text{likelihood} \cdot \text{prior}.$$

Bayesian: Bayesians make inferences using the posterior $P(H|D)$, and therefore always need a prior $P(H)$. If a prior is not known with certainty the Bayesian must try to make a reasonable choice. There are many ways to do this and reasonable people might make different choices. In general it is good practice to justify your choices and to explore a range of priors to see if they all point to the same conclusion.

Frequentist: Very briefly, frequentists do not try to create a prior. Instead, they make inferences using the likelihood $P(D|H)$.

We will compare the two approaches in detail once we have more experience with each. For now we simply list two benefits of the Bayesian approach.

1. The posterior probability $P(H|D)$ for the hypothesis given the evidence is usually exactly what we'd like to know. The Bayesian can say something like 'the parameter of interest has probability 0.95 of being between 0.49 and 0.51.'
2. The assumptions that go into choosing the prior can be clearly spelled out.

More good data: It is always the case that more and better data allows for stronger conclusions and lessens the influence of the prior. The emphasis should be as much on better data (quality) as on more data (quantity).

3 Example: Dice

Suppose we have a drawer full of dice, each of which has either 4, 6, 8, 12, or 20 sides. This time, we do not know how many of each type are in the drawer. A die is picked at random from the drawer and rolled 5 times. The results in order are 4, 2, 4, 7, and 5.

3.1 Uniform prior

Suppose we have no idea what the distribution of dice in the drawer might be. In this case it's reasonable to use a flat prior. Here is the update table for the posterior probabilities that result from updating after each roll. In order to fit all the columns, we leave out the Bayes numerators.

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	1/5	1/4	0.370	1/4	0.542	1/4	0.682	0	0.000	0	0.000
H_6	1/5	1/6	0.247	1/6	0.241	1/6	0.202	0	0.000	1/6	0.000
H_8	1/5	1/8	0.185	1/8	0.135	1/8	0.085	1/8	0.818	1/8	0.876
H_{12}	1/5	1/12	0.123	1/12	0.060	1/12	0.025	1/12	0.161	1/12	0.115
H_{20}	1/5	1/20	0.074	1/20	0.022	1/20	0.005	1/20	0.021	1/20	0.009

This should look familiar. Given the data the final posterior is heavily weighted towards hypthesis H_8 that the 8-sided die was picked.

3.2 Other priors

To see how much the above posterior depended on our choice of prior, let's try some other priors. Suppose we have reason to believe that there are ten times as many 20-sided dice in the drawer as there are each of the other types. The table becomes:

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	0.071	1/4	0.222	1/4	0.453	1/4	0.650	0	0.000	0	0.000
H_6	0.071	1/6	0.148	1/6	0.202	1/6	0.193	0	0.000	1/6	0.000
H_8	0.071	1/8	0.111	1/8	0.113	1/8	0.081	1/8	0.688	1/8	0.810
H_{12}	0.071	1/12	0.074	1/12	0.050	1/12	0.024	1/12	0.136	1/12	0.107
H_{20}	0.714	1/20	0.444	1/20	0.181	1/20	0.052	1/20	0.176	1/20	0.083

Even here the final posterior is heavily weighted to the hypothesis H_8.

What if the 20-sided die is 100 times more likely than each of the others?

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	0.0096	1/4	0.044	1/4	0.172	1/4	0.443	0	0.000	0	0.000
H_6	0.0096	1/6	0.030	1/6	0.077	1/6	0.131	0	0.000	1/6	0.000
H_8	0.0096	1/8	0.022	1/8	0.043	1/8	0.055	1/8	0.266	1/8	0.464
H_{12}	0.0096	1/12	0.015	1/12	0.019	1/12	0.016	1/12	0.053	1/12	0.061
H_{20}	0.9615	1/20	0.889	1/20	0.689	1/20	0.354	1/20	0.681	1/20	0.475

With such a strong prior belief in the 20-sided die, the final posterior gives a lot of weight to the theory that the data arose from a 20-sided die, even though it extremely unlikely the

20-sided die would produce a maximum of 7 in 5 roles. The posterior now gives roughly even odds that an 8-sided die versus a 20-sided die was picked.

3.3 Rigid priors

Mild cognitive dissonance. Too rigid a prior belief can overwhelm any amount of data. Suppose I've got it in my head that the die has to be 20-sided. So I set my prior to $P(H_{20}) = 1$ with the other 4 hypotheses having probability 0. Look what happens in the update table.

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	0	1/4	0	1/4	0	1/4	0	0	0	0	0
H_6	0	1/6	0	1/6	0	1/6	0	0	0	1/6	0
H_8	0	1/8	0	1/8	0	1/8	0	1/8	0	1/8	0
H_{12}	0	1/12	0	1/12	0	1/12	0	1/12	0	1/12	0
H_{20}	1	1/20	1	1/20	1	1/20	1	1/20	1	1/20	1

No matter what the data, a hypothesis with prior probability 0 will have posterior probability 0. In this case I'll never get away from the hypothesis H_{20}, although I might experience some mild cognitive dissonance.

Severe cognitive dissonance. Rigid priors can also lead to absurdities. Suppose I now have it in my head that the die must be 4-sided. So I set $P(H_4) = 1$ and the other prior probabilities to 0. With the given data on the fourth roll I reach an impasse. A roll of 7 can't possibly come from a 4-sided die. Yet this is the only hypothesis I'll allow. My Bayes numerator is a column of all zeros which cannot be normalized.

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	Bayes numer_4	post_4
H_4	1	1/4	1	1/4	1	1/4	1	0	0	???
H_6	0	1/6	0	1/6	0	1/6	0	0	0	???
H_8	0	1/8	0	1/8	0	1/8	0	1/8	0	???
H_{12}	0	1/12	0	1/12	0	1/12	0	1/12	0	???
H_{20}	0	1/20	0	1/20	0	1/20	0	1/20	0	???

I must adjust my belief about what is possible or, more likely, I'll suspect you of accidently or deliberately messing up the data.

4 Example: Malaria

Here is a real example adapted from *Statistics, A Bayesian Perspective* by Donald Berry:

By the 1950s scientists had begun to formulate the hypothesis that carriers of the sickle-cell gene were more resistant to malaria than noncarriers. There was a fair amount of circumstantial evidence for this hypothesis. It also helped explain the persistance of an otherwise deleterious gene in the population. In one experiment scientists injected 30 African volunteers with malaria. Fifteen of the volunteers carried one copy of the sickle-cell gene and the other 15 were noncarriers. Fourteen out of 15 noncarriers developed malaria while only 2 out of 15 carriers did. Does this small sample support the hypothesis that the sickle-cell gene protects against malaria?

Let S represent a carrier of the sickle-cell gene and N represent a non-carrier. Let $D+$ indicate developing malaria and $D-$ indicate not developing malaria. The data can be put in a table.

	$D+$	$D-$	
S	2	13	15
N	14	1	15
	16	14	30

Before analysing the data we should say a few words about the experiment and experimental design. First, it is clearly unethical: to gain some information they infected 16 people with malaria. We also need to worry about bias. How did they choose the test subjects? Is it possible the noncarriers were weaker and thus more susceptible to malaria than the carriers? Berry points out that it is reasonable to assume that an injection is similar to a mosquito bite, but it is not guaranteed. This last point means that if the experiment shows a relation between sickle-cell and protection against injected malaria, we need to consider the hypothesis that the protection from mosquito transmitted malaria is weaker or non-existent. Finally, we will frame our hypothesis as 'sickle-cell protects against malaria', but really all we can hope to say from a study like this is that 'sickle-cell is correlated with protection against malaria'.

Model. For our model let θ_S be the probability that an injected carrier S develops malaria and likewise let θ_N be the probability that an injected noncarrier N develops malaria. We assume independence between all the experimental subjects. With this model, the likelihood is a function of both θ_S and θ_N:

$$P(\text{data}|\theta_S, \theta_N) = c\,\theta_S^2(1-\theta_S)^{13}\theta_N^{14}(1-\theta_N).$$

As usual we leave the constant factor c as a letter. (It is a product of two binomial coefficients: $c = \binom{15}{2}\binom{15}{14}$.)

Hypotheses. Each hypothesis consists of a pair (θ_N, θ_S). To keep things simple we will only consider a finite number of values for these probabilities. We could easily consider many more values or even a continuous range of hypotheses. Assume θ_S and θ_N are each one of 0, 0.2, 0.4, 0.6, 0.8, 1. This leads to two-dimensional tables.

First is a table of hypotheses. The color coding indicates the following:
1. Light blue squares along the diagonal are where $\theta_S = \theta_N$, i.e. sickle-cell makes no difference one way or the other.
2. Orange and darker blue squares above the diagonal are where $\theta_N > \theta_S$, i.e. sickle-cell provides some protection against malaria.
3. In the orange squares $\theta_N - \theta_S \geq 0.6$, i.e. sickle-cell provides a lot of protection.
4. White squares below diagonal are where $\theta_S > \theta_N$, i.e. sickle-cell actually increases the probability of developing malaria.

$\theta_N\backslash\theta_S$	0	0.2	0.4	0.6	0.8	1
1	(0,1)	(.2,1)	(.4,1)	(.6,1)	(.8,1)	(1,1)
0.8	(0,.8)	(.2,.8)	(.4,.8)	(.6,.8)	(.8,.8)	(1,.8)
0.6	(0,.6)	(.2,.6)	(.4,.6)	(.6,.6)	(.8,.6)	(1,.6)
0.4	(0,.4)	(.2,.4)	(.4,.4)	(.6,.4)	(.8,.4)	(1,.4)
0.2	(0,.2)	(.2,.2)	(.4,.2)	(.6,.2)	(.8,.2)	(1,.2)
0	(0,0)	(.2,0)	(.4,0)	(.6,0)	(.8,0)	(1,0)

Hypotheses on level of protection due to S:
orange = strong; darker blue = some; light blue = none; white = negative.

Next is the table of likelihoods. (Actually we've taken advantage of our indifference to scale and scaled all the likelihoods by $100000/c$ to make the table more presentable.) Notice that, to the precision of the table, many of the likelihoods are 0. The color coding is the same as in the hypothesis table. We've highlighted the biggest likelihoods with a thick black border.

$\theta_N\backslash\theta_S$	0	0.2	0.4	0.6	0.8	1
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	1.93428	0.18381	0.00213	0.00000	0.00000
0.6	0.00000	0.06893	0.00655	0.00008	0.00000	0.00000
0.4	0.00000	0.00035	0.00003	0.00000	0.00000	0.00000
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Likelihoods $p(\text{data}|\theta_S, \theta_N)$ scaled by $100000/c$

4.1 Flat prior

Suppose we have no opinion whatsoever on whether and to what degree sickle-cell protects against malaria. In this case it is reasonable to use a flat prior. Since there are 36 hypotheses each one gets a prior probability of 1/36. This is given in the table below. Remember each square in the table represents one hypothesis. Because it is a probability table we include the marginal pmfs.

$\theta_N\backslash\theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.8	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p(\theta_S)$	1/6	1/6	1/6	1/6	1/6	1/6	1

Flat prior $p(\theta_S, \theta_N)$: every hypothesis (square) has equal probability

To compute the posterior we simply multiply the likelihood table by the prior table and normalize. Normalization means making sure the entire table sums to 1.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N \vert \text{data})$
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	0.88075	0.08370	0.00097	0.00000	0.00000	0.96542
0.6	0.00000	0.03139	0.00298	0.00003	0.00000	0.00000	0.03440
0.4	0.00000	0.00016	0.00002	0.00000	0.00000	0.00000	0.00018
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$p(\theta_S \vert \text{data})$	0.00000	0.91230	0.08670	0.00100	0.00000	0.00000	1.00000

Posterior to flat prior: $p(\theta_S, \theta_N|\text{data})$

To decide whether S confers protection against malaria, we compute the posterior probabilities of 'some protection' and of 'strong protection'. These are computed by summing the corresponding squares in the posterior table.

$$\text{Some protection: } P(\theta_N > \theta_S) = \text{sum of orange and darker blue} = 0.99995$$

$$\text{Strong protection: } P(\theta_N - \theta_S > 0.6) = \text{sum of orange} = 0.88075$$

Working from the flat prior, it is effectively certain that sickle-cell provides some protection and very probable that it provides strong protection.

4.2 Informed prior

The experiment was not run without prior information. There was a lot of circumstantial evidence that the sickle-cell gene offered some protection against malaria. For example it was reported that a greater percentage of carriers survived to adulthood.

Here's one way to build an informed prior. We'll reserve a reasonable amount of probability for the hypotheses that S gives no protection. Let's say 24% split evenly among the 6 (light blue) cells where $\theta_N = \theta_S$. We know we shouldn't set any prior probabilities to 0, so let's spread 6% of the probability evenly among the 15 white cells below the diagonal. That leaves 70% of the probability for the 15 orange and darker blue squares above the diagonal.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N)$
1	0.04667	0.04667	0.04667	0.04667	0.04667	0.04000	0.27333
0.8	0.04667	0.04667	0.04667	0.04667	0.04000	0.00400	0.23067
0.6	0.04667	0.04667	0.04667	0.04000	0.00400	0.00400	0.18800
0.4	0.04667	0.04667	0.04000	0.00400	0.00400	0.00400	0.14533
0.2	0.04667	0.04000	0.00400	0.00400	0.00400	0.00400	0.10267
0	0.04000	0.00400	0.00400	0.00400	0.00400	0.00400	0.06000
$p(\theta_S)$	0.27333	0.23067	0.18800	0.14533	0.10267	0.06000	1.0

Informed prior $p(\theta_S, \theta_N)$: makes use of prior information that sickle-cell is protective.

We then compute the posterior pmf.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N \mid \text{data})$
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	0.88076	0.08370	0.00097	0.00000	0.00000	0.96543
0.6	0.00000	0.03139	0.00298	0.00003	0.00000	0.00000	0.03440
0.4	0.00000	0.00016	0.00001	0.00000	0.00000	0.00000	0.00017
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$p(\theta_S \mid \text{data})$	0.00000	0.91231	0.08669	0.00100	0.00000	0.00000	1.00000

Posterior to informed prior: $p(\theta_S, \theta_N \mid \text{data})$

We again compute the posterior probabilities of 'some protection' and 'strong protection'.

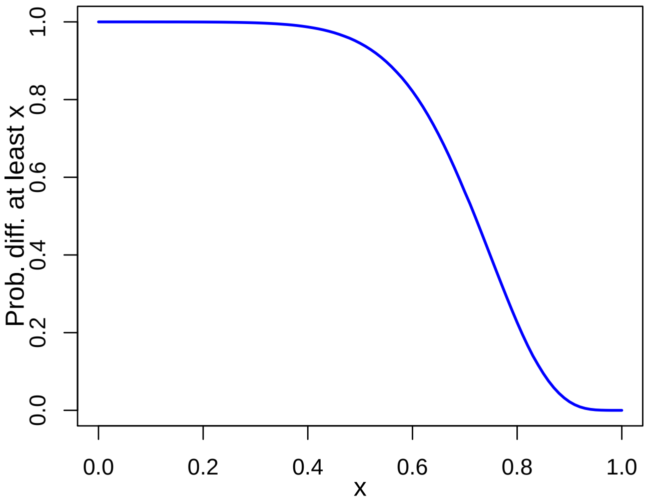
$$\text{Some protection: } P(\theta_N > \theta_S) = \text{sum of orange and darker blue} = 0.99996$$

$$\text{Strong protection: } P(\theta_N - \theta_S > 0.6) = \text{sum of orange} = 0.88076$$

Note that the informed posterior is nearly identical to the flat posterior.

4.3 PDALX

The following plot is based on the flat prior. For each x, it gives the probability that $\theta_N - \theta_S \geq x$. To make it smooth we used many more hypotheses.



Probability the difference $\theta_N - \theta_S$ is at least x (PDALX).

Notice that it is almost certain that the difference is at least 0.4.

Probability intervals
Class 16, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to find probability intervals given a pmf or pdf.

2. Understand how probability intervals summarize belief in Bayesian updating.

3. Be able to use subjective probability intervals to construct reasonable priors.

4. Be able to construct subjective probability intervals by systematically estimating quantiles.

2 Probability intervals

Suppose we have a pmf $p(\theta)$ or pdf $f(\theta)$ describing our belief about the value of an unknown parameter of interest θ.

Definition: A p-probability interval for θ is an interval $[a, b]$ with $P(a \le \theta \le b) = p$.

Notes.
1. In the discrete case with pmf $p(\theta)$, this means $\sum_{a \le \theta_i \le b} p(\theta_i) = p$.

2. In the continuous case with pdf $f(\theta)$, this means $\int_a^b f(\theta)\, d\theta = p$.

3. We may say 90%-probability interval to mean 0.9-probability interval. Probability intervals are also called credible intervals to contrast them with confidence intervals, which we will introduce in the frequentist unit.

Example 1. Between the 0.05 and 0.55 quantiles is a 0.5 probability interval. There are many 50% probability intervals, e.g. the interval from the 0.25 to the 0.75 quantiles.

In particular, notice that the p-probability interval for θ is not unique.

Q-notation. We can phrase probability intervals in terms of **quantiles**. Recall that the s-quantile for θ is the value q_s with $P(\theta \le q_s) = s$. So for $s \le t$, the amount of probability between the s-quantile and the t-quantile is just $t - s$. In these terms, a p-probability interval is any interval $[q_s, q_t]$ with $t - s = p$.

Example 2. We have 0.5 probability intervals $[q_{0.25}, q_{0.75}]$ and $[q_{0.05}, q_{0.55}]$.

Symmetric probability intervals.
The interval $[q_{0.25}, q_{0.75}]$ is symmetric because the amount of probability remaining on either side of the interval is the same, namely 0.25. If the pdf is not too skewed, the symmetric interval is usually a good default choice.

More notes.

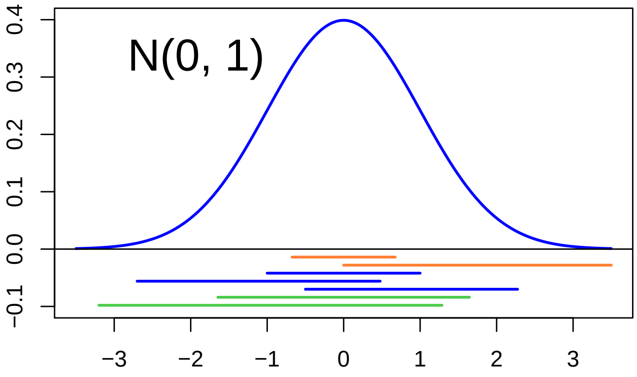
1. Different p-probability intervals for θ may have different widths. We can make the width smaller by centering the interval under the highest part of the pdf. Such an interval is usually a good choice since it contains the most likely values. See the examples below for normal and beta distributions.

2. Since the width can vary for fixed p, a larger p does not always mean a larger width. Here's what is true: if a p_1-probability interval is fully contained in a p_2-probability interval, then p_1 is smaller than p_2.

Probability intervals for a normal distribution. The figure shows a number of probability intervals for the standard normal.

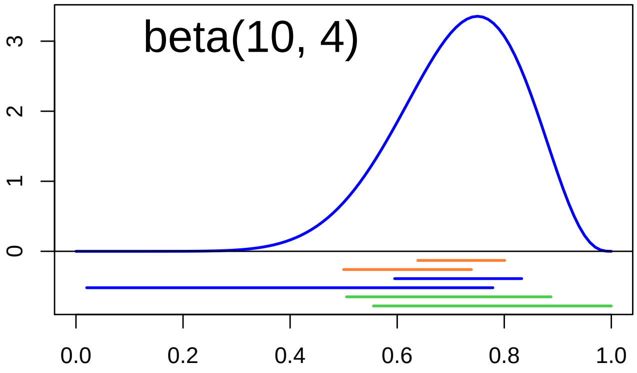
1. All of the blue bars span a 0.68-probability interval. Notice that the smallest blue bar runs between -1 and 1. This runs from the 16th percentile to the 84th percentile so it is a symmetric interval.

2. All the green bars span a 0.9-probability interval. They are longer than the blue bars because they include more probability. Note again that the shortest green bar is symmetric.



orange = 0.5, blue = 0.68, green = 0.9

Probabilitiy intervals for a beta distribution. The following figure shows probability intervals for a beta distribution. Notice how the two blue bars have very different lengths yet cover the same probability $p = 0.68$.



orange = 0.5, blue = 0.68, green = 0.9

3 Uses of probability intervals

3.1 Summarizing and communicating your beliefs

Probability intervals are an intuitive and effective way to summarize and communicate your beliefs. It's hard to describe an entire function $f(\theta)$ to a friend in words. If the function isn't from a parameterized family then it's especially hard. Even with a beta distribution, it's easier to interpret "I think θ is between 0.45 and 0.65 with 50% probability" than "I think θ follows a beta(8,6) distribution". An exception to this rule of communication might be the normal distribution, but only if the recipient is also comfortable with standard deviation. Of course, what we gain in clarity we lose in precision, since the function contains more information than the probability interval.

Probability intervals also play well with Bayesian updating. If we update from the prior $f(\theta)$ to the posterior $f(\theta|x)$, then the p-probability interval for the posterior will tend to be shorter than than the p-probability interval for the prior. In this sense, the data has made us more certain. See for example the election example below.

4 Constructing a prior using subjective probability intervals

Probability intervals are also useful when we do not have a pmf or pdf at hand. In this case, subjective probability intervals give us a method for constructing a reasonable prior for θ "from scratch". The thought process is to ask yourself a series of questions, e.g., 'what is my expected value for θ?'; 'my 0.5-probability interval?'; 'my 0.9-probability interval?' Then build a prior that is consistent with these intervals.

4.1 Estimating the intervals directly

Example 3. Building priors
In 2013 there was a special election for a congressional seat in a district in South Carolina. The election pitted Republican Mark Sanford against Democrat Elizabeth Colbert Busch. Let θ be the fraction of the population who favored Sanford. Our goal in this example is to build a subjective prior for θ. We'll use the following prior evidence.

- Sanford is a former S. Carolina Congressman and Governor
- In 2009, while Governor, he had to resign after he was discovered to be having an affair in Argentina while he claimed to be hiking the Appalachian trail.
- In 2013 Sanford won the Republican primary over 15 primary opponents.
- In the district in the 2012 presidential election the Republican Romney beat the Democrat Obama 58% to 40%.
- The Colbert bump: Elizabeth Colbert Busch is the sister of well-known comedian Stephen Colbert.

Our strategy will be to use our intuition to construct some probability intervals and then find a beta distribution that approximately matches these intervals. This is subjective so someone else might give a different answer.

Step 1. Use the evidence to construct 0.5 and 0.9 probability intervals for θ.

We'll start by thinking about the 90% interval. The single strongest prior evidence is the 58% to 40% of Romney over Obama. Given the negatives for Sanford we don't expect he'll win much more than 58% of the vote. So we'll put the top of the 0.9 interval at 0.65. With all of Sanford's negatives he could lose big. So we'll put the bottom at 0.3.

$$\text{0.9 interval:} \quad [0.3,\ 0.65]$$

For the 0.5 interval we'll pull these endpoints in. It really seems unlikely Sanford will get more votes than Romney, so we can leave 0.25 probability that he'll get above 57%. The lower limit seems harder to predict. So we'll leave 0.25 probability that he'll get under 42%.

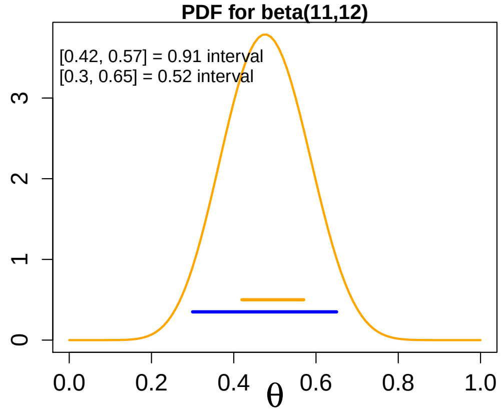
$$\text{0.5 interval:} \quad [0.42,\ 0.57]$$

Step 2. Use our 0.5 and 0.9 probability intervals to pick a beta distribution that approximats these intervals. We used the R function `pbeta` and a little trial and error to choose beta(11,12). Here is our R code.

```
a = 11
b = 12
pbeta(0.65, a, b) - pbeta(0.3, a, b)
pbeta(0.57, a, b) - pbeta(0.42, a, b)
```

This computed $P([0.3, 0.65]) = 0.91$ and $P([0.42, 0.57]) = 0.52$. So our intervals are actually 0.91 and 0.52-probability intervals. This is pretty close to what we wanted!

The plot below shows the density of beta(11,12). The horizontal orange line shows our interval [0.42, 0.57] and the blue line shows our interval [0.3, 0.65].



beta(11,12) fitting 0.5 and 0.9 probability intervals

4.2 Constructing a prior by estimating quantiles

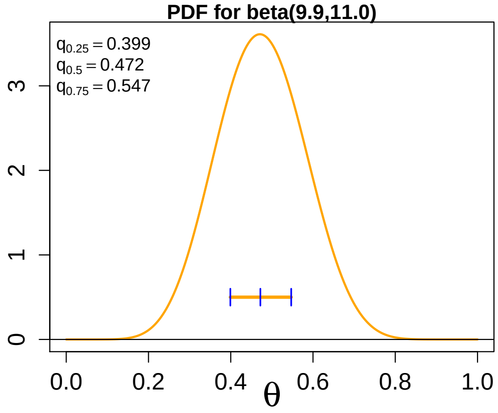
The method in Example 3 gives a good feel for building priors from probability intervals. Here we illustrate a slightly different way of building a prior by estimating quantiles. The basic strategy is to first estimate the median, then divide and conquer to estimate the first and third quantiles. Finally you choose a prior distribution that fits these estimates.

Example 4. Redo the Sanford vs. Colbert-Busch election example using quantiles.

Solution: We start by estimating the median. Just as before the single strongest evidence is the 58% to 40% victory of Romney over Obama. However, given Sanford's negatives and Busch's Colbert bump we'll estimate the median at 0.47.

In a district that went 58 to 40 for the Republican Romney it's hard to imagine Sanford's vote going a lot below 40%. So we'll estimate Sanford 25th percentile as 0.40. Likewise, given his negatives it's hard to imagine him going above 58%, so we'll estimate his 75th percentile as 0.55.

We used R to search through values of a and b for the beta distribution that matches these quartiles the best. Since the beta distribution does not require a and b to be integers we looked for the best fit to 1 decimal place. We found beta(9.9, 11.0). Above is a plot of beta(9.9,11.0) with its actual quartiles shown. These match the desired quartiles pretty well.



beta(9.9, 11.0) matching desired quartiles

Historic note. In the election Sanford won 54% of the vote and Busch won 45.2%. (Source: https://elections.huffingtonpost.com/2013/mark-sanford-vs-elizabeth-colbert-busch-sc1

The Frequentist School of Statistics
Class 17, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to explain the difference between the frequentist and Bayesian approaches to statistics.
2. Know our working definition of a statistic and be able to distinguish a statistic from a non-statistic.

2 Introduction

After much foreshadowing, the time has finally come to switch from Bayesian statistics to frequentist statistics. For much of the twentieth century, frequentist statistics has been the dominant school. If you've ever encountered confidence intervals, p-values, t-tests, or χ^2-tests, you've seen frequentist statistics. With the rise of high-speed computing and big data, Bayesian methods are becoming more common. After we've studied frequentist methods we will compare the strengths and weaknesses of the two approaches.

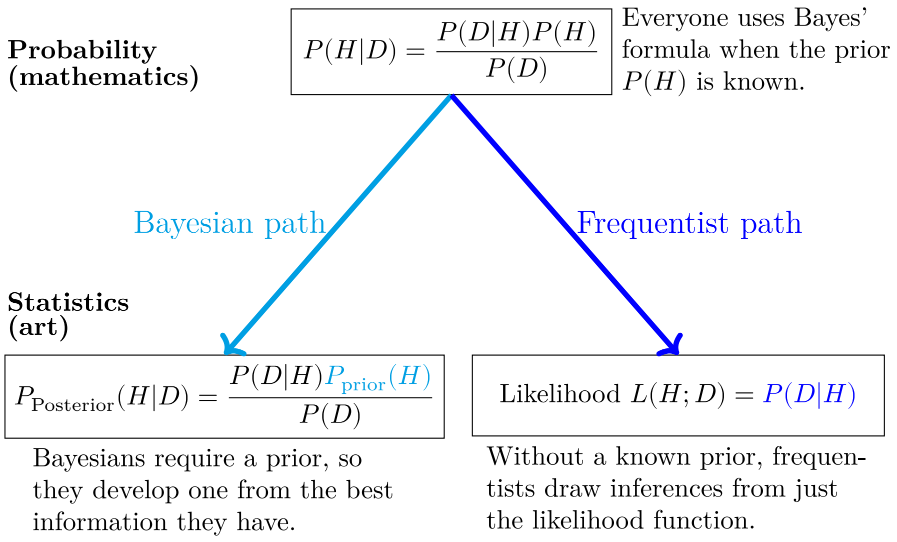
2.1 The fork in the road

Both schools of statistics start with probability. In particular both know and love Bayes' theorem:

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}.$$

When the prior is known exactly, all statisticians will use Bayes' formula. For Bayesian inference we take H to be a hypothesis and D some data. Over the last few weeks we have seen that, given a prior and a likelihood model, Bayes' theorem is a complete recipe for updating our beliefs in the face of new data. This works perfectly when the prior was known perfectly. We saw this in our dice examples. We also saw examples of a disease with a known frequency in the general population and a screening test of known accuracy.

In practice we saw that there is usually no universally accepted prior – different people will have different a priori beliefs – but we would still like to make useful inferences from data. Bayesians and frequentists take fundamentally different approaches to this challenge, as summarized in the figure below.


Probability (mathematics)
$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$
Everyone uses Bayes' formula when the prior $P(H)$ is known.
Bayesian path
Frequentist path
Statistics (art)
$P_{\text{Posterior}}(H|D) = \frac{P(D|H)P_{\text{prior}}(H)}{P(D)}$
Likelihood $L(H; D) = P(D|H)$
Bayesians require a prior, so they develop one from the best information they have.
Without a known prior, frequentists draw inferences from just the likelihood function.

The reasons for this split are both practical (ease of implementation and computation) and philosophical (subjectivity versus objectivity and the nature of probability).

2.2 What is probability?

The main philosophical difference concerns the meaning of probability. The term frequentist refers to the idea that probabilities represent long term frequencies of repeatable random experiments. For example, 'a coin has probability 1/2 of heads' means that the relative frequency of heads (number of heads out of number of flips) goes to 1/2 as the number of flips goes to infinity. This means the frequentist finds it nonsensical to specify a probability distribution for a parameter with a fixed value. While Bayesians are happy to use probability to describe their incomplete knowledge of a fixed parameter, frequentists reject the use of probability to quantify degree of belief in hypotheses.

Example 1. Suppose I have a bent coin with unknown probability θ of heads. The value of θ may be unknown, but it is a fixed value. Thus, to the frequentist there can be no prior pdf $f(\theta)$. By comparison the Bayesian may agree that θ has a fixed value, but interprets $f(\theta)$ as representing uncertainty about that value. Both the Bayesian and the frequentist are perfectly happy with $p(\text{heads} \,|\, \theta) = \theta$, since the longterm frequency of heads given θ is θ.

In short, Bayesians put probability distributions on everything (hypotheses and data), while frequentists put probability distributions on (random, repeatable, experimental) data given a hypothesis. For the frequentist when dealing with data from an unknown distribution only the likelihood has meaning. The prior and posterior do not.

3 Working definition of a statistic

Our view of statistics is that it is the art of drawing conclusions (making inferences) from data. With that in mind we can make a simple working definition of a statistic. There is a more formal definition, but we don't need to introduce it at this point.

Statistic. A statistic is anything that can be computed from data and known values. Sometimes to be more precise we'll say a statistic is a rule for computing something from data and the value of the statistic is what is computed. This can include computing likelihoods where we hypothesize values of the model parameters. But it does not include anything that requires we know the true value of a model parameter with unknown value.

Examples. 1. The mean of data is a statistic. It is a rule that says given data $x_1, \dots, x_n$ compute $\frac{x_1+\ldots+x_n}{n}$.

2. The maximum of data is a statistic. It is a rule that says to pick the maximum value of the data $x_1, \dots, x_n$.

3. Suppose $x \sim \mathrm{N}(\mu, 3^2)$ where μ is unknown. Then the likelihood

$$\phi(x|\mu = 7) \quad = \quad \frac{1}{3\sqrt{2\pi}}\,\mathrm{e}^{-\frac{(x-7)^2}{18}}$$

is a statistic. However, the distance of x from the true mean μ is not a statistic since we cannot compute it without knowing μ

4. If our data $x_1, \dots, x_n$ is drawn from $\mathrm{N}(\mu, 3^2)$, where μ is unknown, then $z = \dfrac{\overline{x} - 5}{3/\sqrt{n}}$ is a statistic, since it is computed from the data and known values. More generally, if μ_0 is a known value then $z = \dfrac{\overline{x} - \mu_0}{3/\sqrt{n}}$ is a statistic. However, since μ is not known, $z = \dfrac{\overline{x} - \mu}{3/\sqrt{n}}$ is not a statistic.

Note. We will usually stick with our Bayesian practice of using the symbol ϕ for continuous likelihoods. This will help remind us that Frequentists don't have prior and posterior probabilities for hypotheses.

Point statistic. A point statistic is a single value computed from data. For example, the mean and the maximum are both point statistics. The maximum likelihood estimate is also a point statistic since it is computed directly from the data based on a likelihood model.

Interval statistic. An interval statistic is an interval computed from data. For example, the range from the minimum to maximum of $x_1, \dots, x_n$ is an interval statistic, e.g. the data 0.5, 1.0, 0.2, 3.0, 5.0 has range [0.2, 5.0].

Set statistic. A set statistic is a set computed from data.

Example. Suppose we have five dice: 4, 6, 8, 12 and 20-sided. We pick one at random and roll it once. The value of the roll is the data. The set of dice for which this roll is possible is a set statistic. For example, if the roll is a 10 then the value of this set statistic is {12, 20}. If the roll is a 7 then this set statistic has value {8, 12, 20}.

It's important to remember that a statistic is itself a random variable since it is computed from random data. For example, if data is drawn from $\mathrm{N}(\mu, \sigma^2)$ then the mean of n data points follows the distribution $\mathrm{N}(\mu, \sigma^2/n))$.

Sampling distribution. The probability distribution of a statistic is called its sampling distribution.

Point estimate. We can use statistics to make a point estimate of a parameter θ. For example, if our data is drawn from a normal distribution with unknown mean θ. Then the data mean $\bar{x}$ is a point estimate of θ.

Null Hypothesis Significance Testing I
Class 17, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the definitions of the significance testing terms: null hypothesis, alternative hypothesis, NHST, simple hypothesis, composite hypothesis, significance level, power.

2. Be able to design and run a significance test for Bernoulli or binomial data.

3. Be able to compute a p-value for a normal hypothesis and use it in a significance test.

2 Introduction

Frequentist statistics is often applied in the framework of null hypothesis significance testing (NHST). We will look at the Neyman-Pearson paradigm which focuses on one hypothesis called the null hypothesis. There are other paradigms for hypothesis testing, but Neyman-Pearson is the most common. Stated simply, this method asks if the data is well outside the region where we would expect to see it under the null hypothesis. If so, then we reject the null hypothesis in favor of a second hypothesis called the alternative hypothesis. The reasoning is that such extreme data is very unlikely in a world where the null hypothesis is true.

We have said before that statistics is an art. We will see that this statement is certainly valid when we discuss choosing null and alternative hypotheses.

The computations done here all involve the likelihood function. There are two main differences between what we'll do here and what we did in Bayesian updating.

1. The evidence of the data will be considered purely through the likelihood function: it will not be weighted by our prior beliefs.

2. We will need a notion of extreme data, e.g. 95 out of 100 heads in a coin toss or a mayfly that lives for a month.

2.1 A suggestion of how to learn this material

There are seemingly a lot of terms with similar definitions. If you pay careful attention to the figures and how they are shaded and labeled we think you will find that is not so complicated.

2.2 Motivating examples

Example 1. Suppose you want to decide whether a coin is fair. If you toss it 100 times and get 85 heads, would you think the coin is likely to be unfair? What about 60 heads? Or 52 heads? Most people would guess that 85 heads is strong evidence that the coin is unfair, whereas 52 heads is no evidence at all. Sixty heads is less clear. Null hypothesis significance testing (NHST) is a frequentist approach to thinking quantitatively about these questions.

Example 2. Suppose you want to compare a new medical treatment to a placebo or to the current standard of care. What sort of evidence would convince you that the new treatment is better than the placebo or the current standard? Again, NHST is a quantitative framework for answering these questions.

3 Significance testing

We'll start by listing the ingredients for NHST. Formally they are pretty simple. There is an art to choosing good ingredients. We will explore the art in examples. If you have never seen NHST before just scan this list now and come back to it after reading through the examples and explanations given below.

3.1 Ingredients

- H_0: the null hypothesis. This is the default assumption for the model generating the data.
- H_A: the alternative hypothesis. If we reject the null hypothesis we accept this alternative as the best explanation for the data.
- X: the test statistic. We compute this from the data. It is a random variable, because it is computed from random data.
- Null distribution: the probability distribution of X assuming H_0.
- Rejection region: if X is in the rejection region we reject H_0 in favor of H_A.
- Non-rejection region: the complement to the rejection region. If X is in this region we do not reject H_0. Note that we say 'do not reject' rather than 'accept' because usually the best we can say is that the data does not support rejecting H_0.

The null hypothesis H_0 and the alternative hypothesis H_A play different roles. Typically we choose H_0 to be either a simple hypothesis or the default which we'll only reject if we have enough evidence against it. The examples below will clarify this.

4 NHST Terminology

In this section we will use one extended example to introduce and explore the terminology used in null hypothesis significance testing (NHST).

Example 3. To test whether a coin is fair we flip it 10 times. If we get an unexpectedly large or small number of heads we'll suspect the coin is unfair. To make this precise in the language of NHST we set up the ingredients as follows. Let θ be the probability that the coin lands heads when flipped.

1. Null hypothesis: H_0 = 'the coin is fair', i.e. $\theta = 0.5$.
2. Alternative hypothesis: H_A = 'the coin is not fair', i.e. $\theta \neq 0.5$
3. Test statistic: X = number of heads in 10 flips
4. Null distribution: This is the probability function based on the null hypothesis

$$p(x \,|\, \theta = 0.5) \sim \text{binomial}(10, 0.5).$$

Here is the probability table for the null distribution.

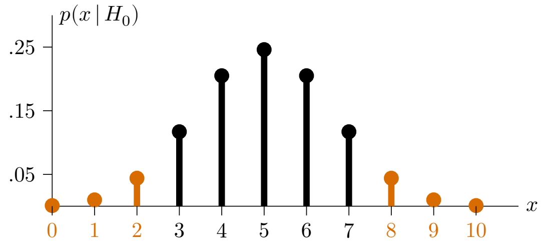
x	0	1	2	3	4	5	6	7	8	9	10
$p(x \,|\, H_0)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001

5. Rejection region: under the null hypothesis we expect to get about 5 heads in 10 tosses. Whereas, under the alternate hypothesis we expect the number of heads to be biased either above or below 5. So, we'll reject H_0 in favor of H_A if the number of heads is much fewer or greater than 5. Let's set the rejection region as $\{0, 1, 2, 8, 9, 10\}$. That is, if the number of heads in 10 tosses is in this region we will reject the hypothesis that the coin is fair in favor of the hypothesis that it is not.

We can summarize all this in the graph and probability table below. The rejection region consists of those values of x in orange, i.e. 0, 1, 2, 8, 9, 10. The probabilities corresponding to it are shaded in orange. We also show the null distribution as a stem plot with the rejection values of x in orange.

x	0	1	2	3	4	5	6	7	8	9	10
$p(x|H_0)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

Rejection region and null probabilities as a table for Example 3.



Rejection region and null probabilities as a stem plot for Example 3.

Notes for Example 3:
1. The null hypothesis is the cautious default: we won't claim the coin is unfair unless we have compelling evidence.
2. The rejection region consists of data that is extreme under the null hypothesis and more likely under the alternative hypothesis. That is, it consists of the outcomes that are in the tail of the null distribution away from the high probability center. As we'll discuss soon, how far away depends on the significance level α of the test.
3. If we get 3 heads in 10 tosses, then the test statistic is in the non-rejection region. The usual scientific language would be to say that the data 'does not support rejecting the null hypothesis'. Even if we got 5 heads, we would not claim that the data proves the null hypothesis is true.

Question: If we have a fair coin what is the probability that we will decide incorrectly it is unfair?

Solution: The null hypothesis is that the coin is fair. The question asks for the probability the data from a fair coin will be in the rejection region. That is, the probability that we will get 0, 1, 2, 8, 9 or 10 heads in 10 tosses. This is the sum of the probabilities in the table in shaded orange boxes. That is,

$$P(\text{rejecting } H_0 \,|\, H_0 \text{ is true}) = 0.11$$

Below we will continue with Example 3, define more terms used in NHST and see how to quantify properties of the significance test.

4.1 Simple and composite hypotheses

Definition: simple hypothesis: A simple hypothesis is one for which we can specify its distribution completely. A typical simple hypothesis is that a parameter of interest takes a specific value.

Definition: composite hypotheses: If its distribution cannot be fully specified, we say that the hypothesis is composite. A typical composite hypothesis is that a parameter of interest lies in a range of values.

In Example 3 the null hypothesis is that $\theta = 0.5$, so the null distribution is binomial$(10, 0.5)$. Since the null distribution is fully specified, H_0 is simple. The alternative hypothesis is that $\theta \neq 0.5$. This is really many hypotheses in one: θ could be 0.51, 0.7, 0.99, etc. Since the alternative distribution binomial$(10, \theta)$ is not fully specified, H_A is composite.

Example 4. Suppose we have data $x_1, \ldots, x_n$. Suppose also that our hypotheses are
H_0: the data is drawn from $N(0, 1)$
H_A: the data is drawn from $N(1, 1)$.
These are both simple hypotheses – each hypothesis completely specifies a distribution.

Example 5. (**Composite hypotheses.**) Now suppose that our hypotheses are
H_0: the data is drawn from a Poisson distribution of unknown parameter.
H_A: the data is not drawn from a Poisson distribution.
These are both composite hypotheses, as they don't fully specify the distribution.

Example 6. In an ESP experiment a subject is asked to identify the suits of 100 cards drawn (with replacement) from a deck of cards. Let T be the number of successes. The (simple) null hypothesis that the subject does not have ESP is given by

$$H_0\colon\ T \sim \text{binomial}(100, 0.25)$$

The (composite) alternative hypothesis that the subject has ESP is given by

$$H_A\colon\ T \sim \text{binomial}(100, p) \text{ with } p > 0.25$$

Another (composite) alternative hypothesis that something besides pure chance is going on, i.e. the subject has ESP or anti-ESP. This is given by

$$H_A\colon\ T \sim \text{binomial}(100, p), \text{ with } p \neq 0.25$$

Values of $p < 0.25$ represent hypotheses that the subject has a kind of anti-esp.

4.2 Types of error

There are two types of errors we can make. We can incorrectly reject the null hypothesis when it is true or we can incorrectly fail to reject it when it is false. These are unimaginatively labeled type I and type II errors. We summarize this in the following table.

		True state of nature	
		H_0	H_A
Our	Reject H_0	Type I error	correct decision
decision	'Don't reject' H_0	correct decision	Type II error

Type I: false rejection of H_0
Type II: false non-rejection ('acceptance') of H_0

4.3 Significance level and power

Significance level and power are used to quantify the quality of the significance test. Ideally a significance test would not make errors. That is, it would not reject H_0 when H_0 was true and would reject H_0 in favor of H_A when H_A was true. Altogether there are 4 important probabilities corresponding to the 2×2 table just above.

$$\begin{array}{ll} P(\text{reject } H_0|H_0) & P(\text{reject } H_0|H_A) \\ P(\text{do not reject } H_0|H_0) & P(\text{do not reject } H_0|H_A) \end{array}$$

The two probabilities we focus on are:

$$\begin{aligned} \text{Significance level} &= P(\text{reject } H_0|H_0) \\ &= \text{probability we incorrectly reject } H_0 \\ &= P(\text{type I error}). \end{aligned}$$

$$\begin{aligned} \text{Power} &= \text{probability we correctly reject } H_0 \\ &= P(\text{reject } H_0|H_A) \\ &= 1 - P(\text{type II error}). \end{aligned}$$

Ideally, a hypothesis test should have a small significance level (near 0) and a large power (near 1). Here are two analogies to help you remember the meanings of significance and power.

Some analogies

1. Think of H_0 as the hypothesis 'nothing noteworthy is going on', i.e. 'the coin is fair', 'the treatment is no better than placebo' etc. And think of H_A as the opposite: 'something interesting is happening'. Then power is the probability of detecting something interesting when it's present and significance level is the probability of mistakenly claiming something interesting has occured.

2. In the U.S. criminal defendents are presumed innocent until proven guilty beyond a reasonable doubt. We can phrase this in NHST terms as

H_0: the defendent is innocent (the default)
H_A: the defendent is guilty.

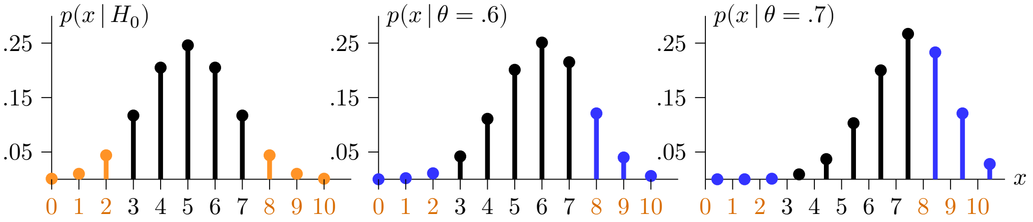
Significance level is the probability of finding an innocent person guilty. Power is the probability of correctly finding a guilty party guilty. 'Beyond a reasonable doubt' means we should demand the significance level be very small.

Composite hypotheses

H_A is composite in Example 3, so the power is different for different values of θ. We expand the previous probability table to include some alternate values of θ. We do the same with the stem plots. As always in the NHST game, we look at likelihoods: the probability of the data given a hypothesis.

The rejection range consists of the extreme values of x. The non-rejection range (3-7) is a set of center values of x.

x	0	1	2	3	4	5	6	7	8	9	10
$H_0 : p(x\|\theta = 0.5)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	.001
$H_A : p(x\|\theta = 0.6)$	0.000	0.002	0.011	0.042	0.111	0.201	0.251	0.215	0.121	0.040	0.006
$H_A : p(x\|\theta = 0.7)$	0.000	0.000	0.001	0.009	0.037	0.103	0.200	0.267	0.233	0.121	0.028



Rejection region and null and alternative probabilities for example 3

We use the probability table to compute the significance level and power of this test.

$$\begin{aligned}\text{Significance level} &= \text{probability we reject } H_0 \text{ when it is true}\\ &= \text{probability the test statistic is in the rejection region when } H_0 \text{ is true}\\ &= \text{probability the test stat. is in the rejection region of the } H_0 \text{ row of the table}\\ &= \text{sum of shaded orange boxes in the } \theta = 0.5 \text{ row}\\ &= 0.11\end{aligned}$$

$$\begin{aligned}\text{Power when } \theta = 0.6 &= \text{probability we reject } H_0 \text{ when } \theta = 0.6\\ &= \text{probability the test statistic is in the rejection region when } \theta = 0.6\\ &= \text{probability the test stat. is in the rejection region of the } \theta = 0.6 \text{ row of the table}\\ &= \text{sum of blue boxes in the } \theta = 0.6 \text{ row}\\ &= 0.180\end{aligned}$$

$$\begin{aligned}\text{Power when } \theta = 0.7 &= \text{probability we reject } H_0 \text{ when } \theta = 0.7\\ &= \text{probability the test statistic is in the rejection region when } \theta = 0.7\\ &= \text{probability the test stat. is in the rejection region of the } \theta = 0.7 \text{ row of the table}\\ &= \text{sum of blue boxes in the } \theta = 0.7 \text{ row}\\ &= 0.384\end{aligned}$$

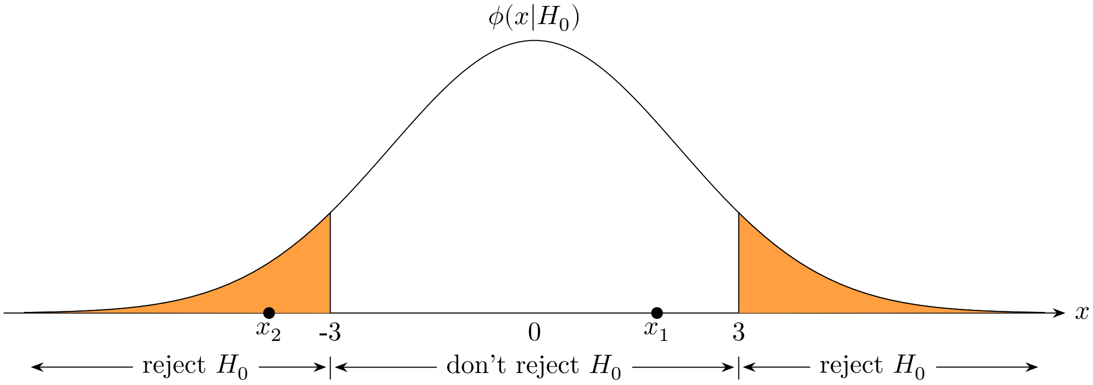
We see that the power is greater for $\theta = 0.7$ than for $\theta = 0.6$. This isn't surprising since we expect it to be easier to recognize that a 0.7 coin is unfair than is is to recognize 0.6 coin is unfair. Typically, we get higher power when the alternate hypothesis is farther from the null hypothesis. In Example 3, it would be quite hard to distinguish a fair coin from one with $\theta = 0.51$.

4.4 Conceptual sketches

We illustrate the notions of null hypothesis, rejection region and power with some sketches of the pdfs for the null and alternative hypotheses.

4.4.1 Null distribution: rejection and non-rejection regions

The first diagram below illustrates a null distribution with rejection and non-rejection regions. Also shown are two possible test statistics: x_1 and x_2.



The test statistic x_1 is in the non-rejection region. So, if our data produced the test statistic x_1 then we would not reject the null hypothesis H_0. On the other hand the test statistic x_2 is in the rejection region, so if our data produced x_2 then we would reject the null hypothesis in favor of the alternative hypothesis.

There are several things to note in this picture.

1. The rejection region consists of values far from the center of the null distribution.

2. The rejection region is two-sided. We will also see examples of one-sided rejection regions as well.

3. The probability of rejection (significance) is the shaded area under the curve, i.e. the probability of data being in the rejection region assuming H_0 is true.

4. The alternative hypothesis is not mentioned. We reject or don't reject H_0 based only on the likelihood $\phi(x|H_0)$, i.e. the probability of the test statistic conditioned on H_0. As we will see, the alternative hypothesis H_A should be considered when choosing a rejection region, but formally it does not play a role in rejecting or not rejecting H_0.

5. Sometimes we rather lazily call the non-rejection region the acceptance region. This is technically incorrect because we never truly accept the null hypothesis. We either reject or say the data does not support rejecting H_0. This is often summarized by the statement: you can never prove the null hypothesis.

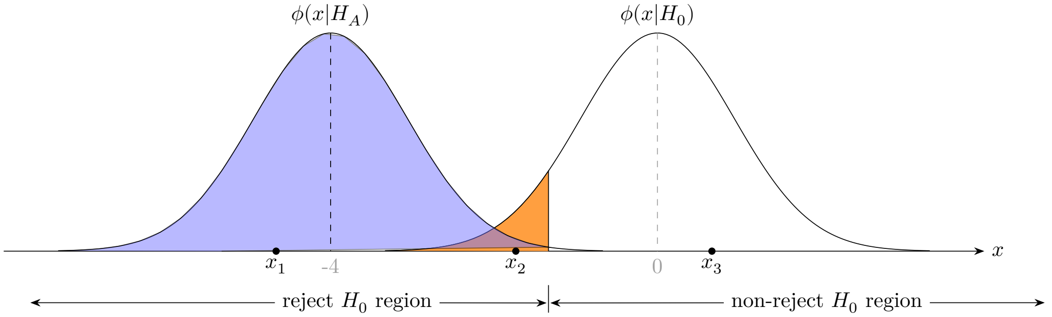
4.4.2 High and low power tests

The next two figures show high and low power tests. In both tests the null distributions are standard normal. Likewise, the alternative distributions are both normal with variance 1, but different means: -4.0 in the top figure and -0.4 in the bottom one.

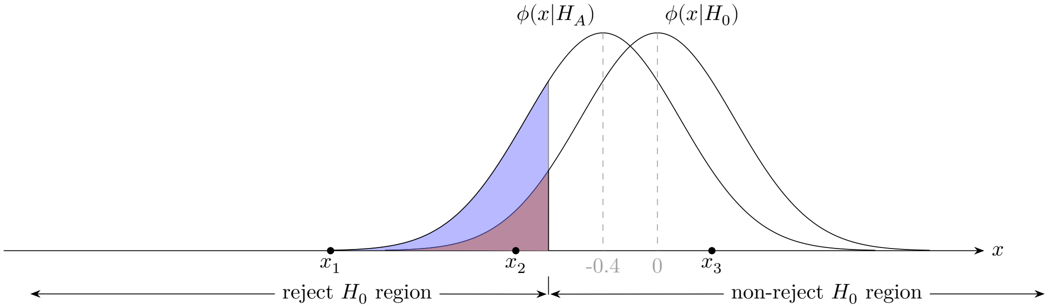
Since the alternative distribution is to the left of the null distribution we use a one-sided rejection region. It is the same for both tests. The shaded area under $\phi(x|H_0)$ represents the significance level – it is the same for both tests. Remember the significance level can be described two ways:

- The probability of falsely rejecting the null hypothesis when it is true.
- The probabilitiy the test statistic falls in the rejection region even though H_0 is true.

Likewise, the shaded area under $\phi(x|H_A)$ represents the power, i.e. the probability that the test statistic is in the rejection (of H_0) region when H_A is true. Both tests have the same significance level, but different powers. When $\phi(x|H_A)$ has considerable overlap with $\phi(x|H_0)$ the power will tend to be low (bottom figure) and when they are well separated it tends to be high (top figure). It is well worth your while to thoroughly understand these graphical representations of significance testing.



High power test



Low power test

In the top figure we see that the means of the null and alternative distributions are 4 standard deviations apart. Since the areas under the densities have very little overlap the test has high power. That is, if H_A was true, then seeing the data x_3 would be rare and surprising and similarly for any point in the non-rejection region. That is, if H_A is the true distribution we are extremely likely to correctly reject the null hypothesis, i.e. we are unlikely to make a type II error.

In the bottom figure we see that the means of the null and alternative distributions are just 0.4 standard deviations apart. Since the areas under the densities have a lot of overlap the test has low power. That is, if the data x is drawn from H_A it is highly likely to be in the non-rejection region. For example x_3 would be not be a very surprising outcome for the H_A

distribution. That is, if H_A is the true distribution, we are highly likely to make a type II error.

Typically we can increase the power of a test by increasing the amount of data and thereby decreasing the variance of the null and alternative distributions. In experimental design it is important to determine ahead of time the number of trials or subjects needed to achieve a desired power.

Example 7. Suppose a drug for a disease is being compared to a placebo. We choose our null and alternative hypotheses as

H_0 = the drug does not work better than the placebo

H_A = the drug works better than the placebo

The power of the hypothesis test is the probability that the test will conclude that the drug is better, if it is indeed truly better. The significance level is the probability that the test will conclude that the drug works better, when in fact it does not. We will look at this in more detail below.

5 Designing a hypothesis test

Formally all a hypothesis test requires is H_0, H_A, a test statistic and a rejection region. In practice the design is often done using the following steps.

1. Pick the null hypothesis H_0.
The choice of H_0 and H_A is not mathematics. It's art and custom. We often choose H_0 to be simple. Or we often choose H_0 to be the simplest or most cautious explanation, i.e. no effect of drug, no ESP, no bias in the coin.

2. Decide if H_A is one-sided or two-sided.
In Example 3 we wanted to know if the coin was unfair. An unfair coin could be biased for or against heads, so $H_A : \theta \neq 0.5$ is a two-sided hypothesis. If we only care whether or not the coin is biased for heads we could use the one-sided hypothesis $H_A : \theta > 0.5$.

3. Pick a test statistic.
For example, the sample mean, sample total, or sample variance. Often the choice is obvious. Some standard statistics that we will encounter are z, t, and χ^2. We will learn to use these statistics as we work examples over the next few classes. One thing we will say repeatedly is that the distributions that go with these statistics are always conditioned on the null hypothesis. That is, we will compute likelihoods such as $\phi(z \,|\, H_0)$.

4. Pick a significance level and determine the rejection region.
We will usually use α to denote the significance level. The Neyman-Pearson paradigm is to pick α in advance. Typical values are 0.1, 0.05, 0.01. Recall that the significance level is the probability of a type I error, i.e. of incorrectly rejecting the null hypothesis when it is true. The value we choose will depend on the consequences of a type I error.

Once the significance level is chosen we can determine the rejection region in the tail(s) of the null distribution. In Example 3, H_A is two sided so the rejection region is split between the two tails of the null distribution. This distribution is given in the following table:

x	0	1	2	3	4	5	6	7	8	9	10
$p(x\|H_0)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

If we set $\alpha = 0.05$ then the rejection region must contain at most 0.05 probability. For a two-sided rejection region this is split between the two tails, so we get

$$\{0, 1, 9, 10\}.$$

If we set $\alpha = 0.01$ the rejection region is

$$\{0, 10\}.$$

We show these in tables with the rejection region (values of x) inside the orange rectangle and the corresponding null likelihoods in the shaded orange boxes.

x	0	1	2	3	4	5	6	7	8	9	10
$p(x\|H_0)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

x	0	1	2	3	4	5	6	7	8	9	10
$p(x\|H_0)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

Tables with shaded rejection regions for $\alpha = 0.05$ (top) and $\alpha = 0.01$ (bottom)

Suppose we change H_A to 'the coin is biased in favor of heads'. We now have a one-sided hypothesis $\theta > 0.5$. Our rejection region will now be in the right-hand tail since we don't want to reject H_0 in favor of H_A if we get a small number of heads. Now if $\alpha = 0.05$ the rejection region is the one-sided range

$$\{9, 10\}.$$

If we set $\alpha = 0.01$ then the rejection region is

$$\{10\}.$$

As above we show the one sided rejection regions in tables.

x	0	1	2	3	4	5	6	7	8	9	10
$p(x\|H_0)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

x	0	1	2	3	4	5	6	7	8	9	10
$p(x\|H_0)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

Tables with shaded one-sided rejection regions for $\alpha = 0.05$ (top) and $\alpha = 0.01$ (bottom)

5. Determine the power(s).
As we saw in Example 3, once the rejection region is set we can determine the power of the test at various values of the alternate hypothesis.

Example 8. (**Consequences of significance**) If $\alpha = 0.1$ then we'd expect a 10% type I error rate. That is, we expect to reject the null hypothesis in 10% of those experiments where the null hypothesis is true. Whether 0.1 is a reasonable signficance level depends on the decisions that will be made using it.

For example, if you were running an experiment to determine if your chocolate is more than 72% cocoa then a 10% error type I error rate is probably okay. That is, falsely believing some 72% chocalate is greater that 72%, is probably acceptable. On the other hand, if your forensic lab is identifying fingerprints for a murder trial then a 10% type I error rate, i.e. mistakenly claiming that fingerprints found at the crime scene belonged to someone who was truly innocent, is definitely not acceptable.

Significance for a composite null hypothesis. If H_0 is composite then P(type I error) depends on which member of H_0 is true. In this case the significance level is defined as the maximum of these probabilities.

6 Critical values

Critical values are like quantiles except they refer to the probability to the right of the value instead of the left.

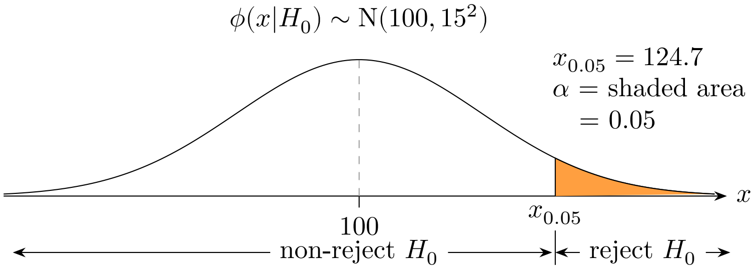
Example 9. Use R to find the 0.05 critical value for the standard normal distribution.

Solution: We label this critical value $z_{0.05}$. The critical value $z_{0.05}$ is just the 0.95 quantile, i.e. it has 5% probability to its right and therefore 95% probability to its left. We computed it with the R function qnorm: `qnorm(0.95, 0, 1)`, which returns 1.64.

In a typical significance test the rejection region consists of one or both tails of the null distribution. The value of the test statistic that marks the start of the rejection region is a critical value. We show this and the notation used in some examples.

Example 10. Critical values and rejection regions. Suppose our test statistic x has null distribution $\mathrm{N}(100, 15^2)$, i.e. $\phi(x|H_0) \sim \mathrm{N}(100, 15^2)$. Suppose also that our rejection region is right-sided and we have a significance level of 0.05. Find the critical value and sketch the null distribution and rejection region.

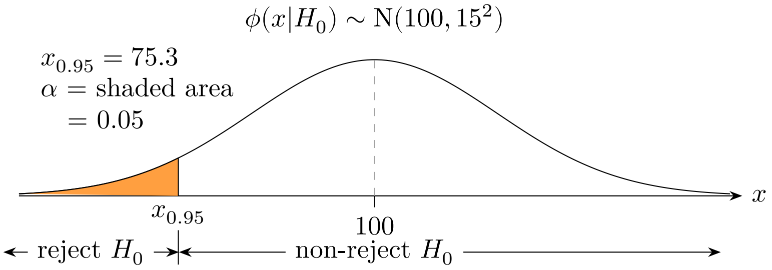
Solution: The notation used for the critical value with right tail containing probability 0.05 is $x_{0.05}$. The critical value $x_{0.05}$ is just the 0.95 quantile, i.e. it has 5% probability to its right and therefore 95% probability to its left. We computed it with the R function qnorm: `qnorm(0.95, 100, 15)`, which returned 124.7. This is shown in the figure below.



Example 11. Critical values and rejection regions. Repeat the previous example for a left-sided rejection region with significance level 0.05.

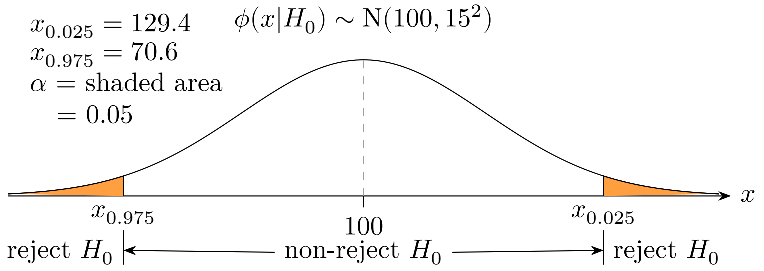
Solution: In this case the critical value has 0.05 probability to its left and therefore 0.95

probability to its right. So we label it $x_{0.95}$. Since it is the 0.05 quantile compute it with the R function: `qnorm(0.05, 100, 15)`, which returned 75.3.



Example 12. Critical values. Repeat the previous example for a two-sided rejection region. Put half the significance in each tail.

Solution: To have a total significance of 0.05 we put 0.025 in each tail. That is, the left tail starts at $x_{0.975} = q_{0.025}$ and the right tail starts at $x_{0.025} = q_{0.975}$. We compute these values with `qnorm(0.025, 100, 15)` and `qnorm(0.975, 100, 15)`. The values are shown in the figure below.



7 *p*-values

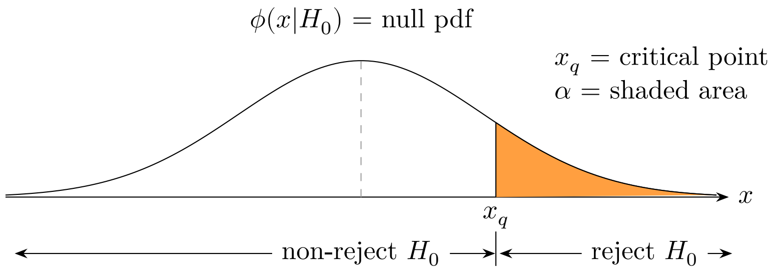
In practice people often specify the significance level and do the significance test using what are called p-values. We will first define p-value and then state the p-test. After that, we will illustrate it with figures and examples.

Definition. The p-value is the probability, assuming the null hypothesis, of seeing data at least as extreme as the experimental data. What 'at least as extreme' means depends on the experimental design.

P-test. If the p-value is less than the significance level α then we reject H_0. Otherwise we do not reject H_0.

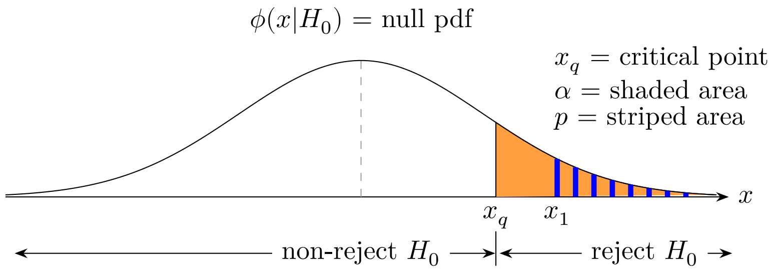
We first illustrate p-values graphically and then we will work a simple one-sided example. We will look at two-sided examples in later classes.

Suppose we have a right-sided alternate hypothesis, so our rejection region is in the right tail of the range of possible outcomes. This is illustrated in the following figure.



Right-sided rejection region

Suppose we get data x_1 which is in the rejection region. This is shown in the figure below.



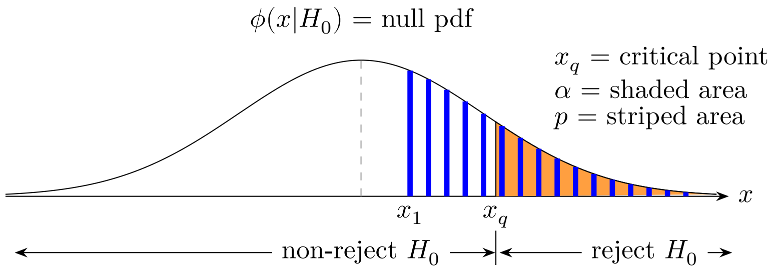
p-value = probability of data ‘more extreme’ than x_1: $p < \alpha$

Since the rejection region is right-sided, the phrase ‘at least as extreme’ means all values to the right of x_1. So, the p-value is the area of the striped region.

Here is the key to connecting the p-test, rejection and signifcance:

Since x_1 is in the rejection region, the striped area p is less than the shaded area α. That is, $p < \alpha$. In other words, we reject the null hypothesis when $p < \alpha$.

For completeness we show a figure where x_1 is not in the rejection region, so $p > \alpha$. That is, we do not reject H_0 when $p > \alpha$.



p-value = probability of data ‘more extreme’ than x_1: $p > \alpha$

7.1 Example: z-tests

When our test statistic is standard normal we will call it z, and the corresponding test for significance will be called a z-test.

Example 13. The z-test for normal hypotheses
IQ is normally distributed in the population according to a $\mathrm{N}(100, 15^2)$ distribution. We suspect that most MIT students have above average IQ so we frame the following hypothe-

ses.

H_0 = MIT student IQs are distributed identically to the general population
= MIT IQ's follow a $\text{N}(100, 15^2)$ distribution.
H_A = MIT student IQs tend to be higher than those of the general population
= the average MIT student IQ is greater than 100.

Notice that H_A is one-sided.

Suppose we test 9 students and find they have an average IQ of $\overline{x} = 112$. Can we reject H_0 at a significance level $\alpha = 0.05$?

Solution: To compute p we first standardize the data: Under the null hypothesis $\bar{x} \sim \text{N}(100, 15^2/9)$ and therefore

$$z = \frac{\bar{x} - 100}{15/\sqrt{9}} = \frac{36}{15} = 2.4 \sim \text{N}(0,1).$$

That is, the null distribution for z is standard normal. We call z a z-statistic, we will use it as our test statistic.

For a right-sided alternative hypothesis the phrase 'data at least as extreme' is a one-sided tail to the right of z. The p-value is then

$$p = P(Z \geq 2.4) = \texttt{1- pnorm(2.4,0,1) = 0.0081975}.$$

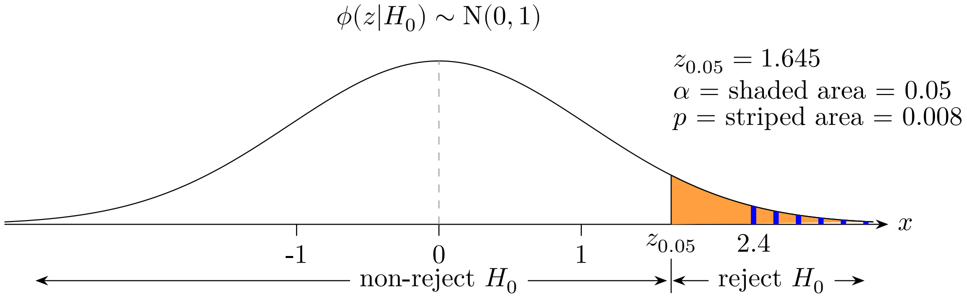
Since $p \leq \alpha$ we reject the null hypothesis. The reason this works is explained below. We phrase our conclusion as

> We reject the null hypothesis in favor of the alternative hypothesis that MIT students have higher IQs on average. We have done this at significance level 0.05 with a p-value of 0.008.

Notes: **1.** The average $\overline{x} = 112$ is random: if we ran the experiment again we could get a different value for $\overline{x}$.

2. We could use the statistic $\overline{x}$ directly. Standardizing is fairly standard because, with practice, we will have a good feel for the meaning of different z-values.

The justification for rejecting H_0 when $p \leq \alpha$ is given in the following figure.



In this example $\alpha = 0.05$, $z_{0.05} = 1.64$ and the rejection region is the range to the right of $z_{0.05}$. Also, $z = 2.4$ and the p-value is the probability to the right of z. The picture illustrates that

- $z = 2.64$ is in the rejection region

- is the same as z is to the right of $z_{0.05}$
- is the same as the probability to the right of z is less than 0.05
- which means $p < 0.05$.

8 More examples

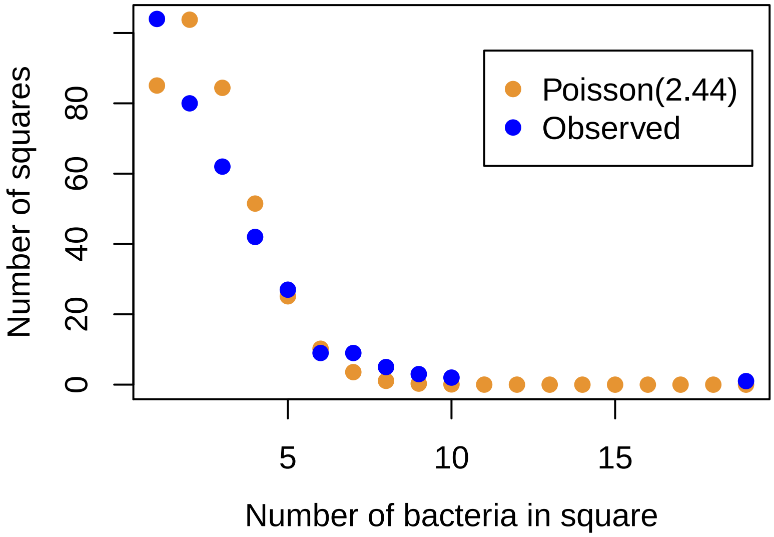
Hypothesis testing is widely used in inferential statistics. We don't expect that the following examples will make perfect sense at this time. Read them quickly just to get a sense of how hypothesis testing is used. We will explore the details of these examples in class.

Example 14. The chi-square statistic and goodness of fit. (Rice, example B, p.313)

To test the level of bacterial contamination, milk was spread over a grid with 400 squares. The amount of bacteria in each square was measured. We summarize in the table below. The bottom row of the table is the number of different squares that had a given amount of bacteria.

Amount of bacteria	0	1	2	3	4	5	6	7	8	9	10	19
Number of squares	56	104	80	62	42	27	9	9	5	3	2	1

We compute that the average amount of bacteria per square is 2.44. Since the Poisson(λ) distribution is used to model counts of relatively rare events and the parameter λ is the expected value of the distribution. we decide to see if these counts could come from a Poisson distribution. To do this we first graphically compare the observed frequencies with those expected from Poisson(2.44).



The picture is suggestive, so we do a hypothesis test with

H_0 : the samples come from a Poisson(2.44) distribution.

H_A : the samples come from a different distribution.

We use a chi-square statistic, so called because it (approximately) follows a chi-square distribution. To compute X^2 we first combine the last few cells in the table so that the minimum expected count is around 5 (a general rule-of-thumb in this game.)

The expected number of squares with a certain amount of bacteria comes from considering 400 trials from a Poisson(2.44) distribution, e.g., with $= 2.44$ the expected number of squares with 3 bacteria is $400 \times \mathrm{e}^{-}\frac{3}{3!} = 84.4$.

The chi-square statistic is $\sum \frac{(O_i - E_i)^2}{E_i}$, where O_i is the observed number and E_i is the expected number.

Number per square	0	1	2	3	4	5	6	> 6
Observed	56	104	80	62	42	27	9	20
Expected	34.9	85.1	103.8	84.4	51.5	25.1	10.2	5.0
Component of X^2	12.8	4.2	5.5	6.0	1.7	0.14	0.15	44.5

Summing up we get $X^2 = 74.9$.

Since the mean (2.44) and the total number of trials (400) are fixed, the 8 cells only have 6 degrees of freedom. So, assuming H_0, our chi-square statistic follows (approximately) a χ^2_6 distribution. Using this distribution, $P(X^2 > 74.59) = 0$ (to at least 6 decimal places). Thus we decisively reject the null hpothesis in favor of the alternate hypothesis that the distribution is not Poisson(2.44).

To analyze further, look at the individual components of X^2. There are large contributions in the tail of the distribution, so that is where the fit goes awry.

Example 15. Student's t test.

Suppose we want to compare a medical treatment for increasing life expectancy with a placebo. We give n people the treatment and m people the placebo. Let $X_1, \ldots, X_n$ be the number of years people live after receiving the treatment. Likewise, let $Y_1, \ldots, Y_m$ be the number of years people live after receiving the placebo. Let $\bar{X}$ and $\bar{Y}$ be the sample means. We want to know if the difference between $\bar{X}$ and $\bar{Y}$ is statistically significant. We frame this as a hypothesis test. Let μ_X and μ_Y be the (unknown) means.

$$H_0 : \mu_X = \mu_Y, \quad H_A : \mu_X \neq \mu_Y.$$

With certain assumptions and a proper formula for the pooled standard error s_p the test statistic $t = \frac{\bar{X} - \bar{Y}}{s_p}$ follow a t distribution with $n + m - 2$ degrees of freedom. So our rejection region is determined by a threshold t_0 with $P(t > t_0) = \alpha$.

Null Hypothesis Significance Testing II
Class 18, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to list the steps common to all null hypothesis significance tests.

2. Be able to define and compute the probability of Type I and Type II errors.

3. Be able to look up and apply one- and two-sample t-tests.

2 Introduction

We continue our study of significance tests. In these notes we will introduce two new tests: one-sample t-tests and two-sample t-tests. You should pay careful attention to the fact that every test makes some assumptions about the data – often that is drawn from a normal distribution. You should also notice that all the tests follow the same pattern. It is just the computation of the test statistic and the type of the null distribution that changes.

3 Review

3.1 Setting up and running a significance test

There is a fairly standard set of steps one takes to set up and run a null hypothesis significance test.

1. Design an experiment to collect data and choose a test statistic x to be computed from the data. The key requirement here is to know the null distribution $\phi(x|H_0)$. To compute power, one must also know the alternative distribution $\phi(x|H_A)$.

2. Decide if the test is one or two-sided based on H_A and the form of the null distribution.

3. Choose a significance level α for rejecting the null hypothesis.

4. Decide how much data you need to collect to achieve the desired power for the test.

5. Run the experiment to collect data $x_1, x_2, \ldots, x_n$.

6. Compute the test statistic x.

7. Compute the p-value corresponding to x using the null distribution.

8. If $p < \alpha$, reject the null hypothesis in favor of the alternative hypothesis.

Notes.
1. Rather than choosing a significance level, you could instead choose a rejection region and reject H_0 if x falls in this region. The corresponding significance level is then the probability, assuming H_0, that x falls in the rejection region.

2. The null hypothesis is often the 'cautious hypothesis'. The lower we set the significance level, the more "evidence" we will require before rejecting our cautious hypothesis in favor of a more sensational alternative. It is standard practice to publish the p value itself so that others may draw their own conclusions.

3. **A key point of confusion:** A significance level of 0.05 does not mean the test only makes mistakes 5% of the time. It means that if the null hypothesis is true, then the probability the test will mistakenly reject it is 5%. The power of the test measures the accuracy of the test when the alternative hypothesis is true. Namely, the power of the test is the probability of rejecting the null hypothesis if the alternative hypothesis is true. Therefore the probability of falsely failing to reject the null hypothesis is 1 minus the power.

4. **Another key point of confusion:** We use p-values, but conceptually the p-value is just a computational trick. After choosing a test statistic, the conceptual order is: first pick a significance level, then use this to define the rejection region. We reject the null hypothesis if the test statistic is in the rejection region. All the p-value does is tell us in one computation whether or not the test stastic is in the rejection region.

Errors. We can summarize these two types of errors and their probabilities as follows:

$$\begin{array}{lcl} \text{Type I error} & = & \text{rejecting } H_0 \text{ when } H_0 \text{ is true.} \\ \text{Type II error} & = & \text{failing to reject } H_0 \text{ when } H_A \text{ is true.} \end{array}$$

$$\begin{array}{lcl} \text{P(type I error)} & = & \text{probability of falsely rejecting } H_0 \\ & = & \text{P(test statistic is in the rejection region} \mid H_0) \\ & = & \text{significance level of the test} \end{array}$$

$$\begin{array}{lcl} \text{P(type II error)} & = & \text{probability of falsely not rejecting } H_0 \\ & = & \text{P(test statistic is in the acceptance region} \mid H_A) \\ & = & \text{1 - power.} \end{array}$$

Helpful analogies.

In terms of medical testing for a disease: a Type I error is a false positive and a Type II error is a false negative.

In a jury trial, a Type I error is convicting an innocent defendant and a Type II error is acquitting a guilty defendant.

3.2 Power

We discussed power in the Class 17 notes. Power is the probabilitiy of correctly rejecting the null hypothesis. It depends on the alternative hypothesis H_A being considered.

The ideal test has power equal to 1.0 and significance equal to 0.0. Of course, in general, this is impossible. And we want to find some compromise where power is high and signficance is low.

In symbols: power $= P(\text{data is in the rejection region} \mid H_A)$.

Compare this with: signficance $= P(\text{data is in the rejection region} \mid H_0)$.

4 Understanding a significance test

Questions to ask:

1. How did they collect data? What is the experimental setup?
2. What are the null and alternative hypotheses?
3. What type of significance test was used?
   Does the data match the criteria needed to use this type of test?
   How robust is the test to deviations from these criteria?
4. For example, some tests comparing two groups of data assume that the groups are drawn from distributions that have the same variance. This needs to be verified before applying the test. Often the check is done using another significance test designed to compare the variances of two groups of data.
5. How is the p-value computed?
   A significance test comes with a test statistic and a null distribution. In most tests the p-value is

   $$p = P(\text{data at least as extreme as what we got} \mid H_0)$$

   What does 'data at least as extreme as the data we saw' mean? For example, is the test one or two-sided?
6. What is the significance level α for this test? If $p < \alpha$ then the experimenter will reject H_0 in favor of H_A.
7. What is the power of the test?

5 t tests

Many significance tests assume that the data are drawn from a normal distribution, so before using such a test you should examine the data to see if the normality assumption is reasonable. We will describe how to do this in more detail later, but plotting a histogram is a good start. Like the z-test, the one-sample and two-sample t-tests that we consider below start from this normality assumption.

We don't expect you to memorize all the computational details of these tests and those to follow. In real life, you have access to textbooks, google, and wikipedia; on the exam, you'll have your notecard. Instead, you should be able to identify when a t-test is appropriate and apply this test after looking up the details and using a table or software like R.

5.1 z-test

Let's first review the z-test.

- Data: we assume $x_1, x_2, \ldots, x_n \sim N(\mu, \sigma^2)$, where μ is unknown and σ is known.
- Null hypothesis: $\mu = \mu_0$ for some specific value μ_0
- Test statistic: $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$ = standardized mean
- Null distribution: $\phi(z \mid H_0)$ is the pdf of $Z \sim N(0,1)$
- One-sided p-value (right side): $p = P(Z \geq z \mid H_0)$
  One-sided p-value (left side): $p = P(Z \leq z \mid H_0)$
  Two-sided p-value: $p = \begin{cases} 2P(Z \geq z) & \text{if } z > 0 \\ 2P(Z \leq z) & \text{if } z < 0. \end{cases}$.

  Because of the symmetry of the distribution around 0, we can also write this as $p = P(|Z| \geq |z|)$.

  See Example 1b for the rationale for this.

Example 1. Suppose that we have data that follows a normal distribution of unknown mean μ and known variance 4. Let the null hypothesis H_0 be that $\mu = 2$. Let the alternative hypothesis H_A be that $\mu > 2$. Suppose we collect the following data:

$$3,\ 2,\ 5,\ 7,\ 1$$

At a significance level of $\alpha = 0.05$, should we reject the null hypothesis?

Solution: There are 5 data points with average $\overline{x} = 3.6$. Because we have normal data with a known variance we should use a z test. Our z statistic is

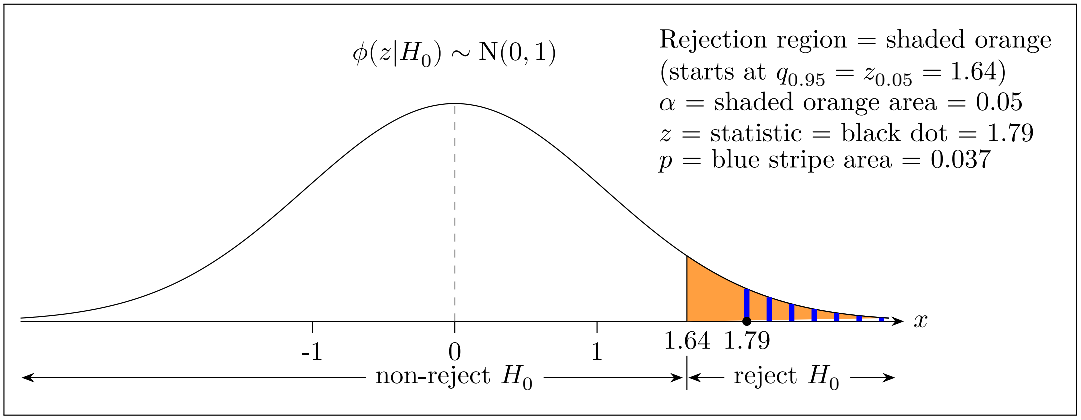
$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{3.6 - 2}{2/\sqrt{5}} = 1.79$$

Our test is one-sided because the alternative hypothesis is one-sided. So (using R) our p-value is

$$p = P(Z > z) = P(Z > 1.79) = 0.037$$

Since $p < \alpha = 0.05$, we reject the null hypothesis in favor of the alternative hypothesis that $\mu > 2$.

We can visualize the test as follows:



Example 1b. Repeat Example 1 as a two-sided test, i.e. with H_A being $\mu \neq 2$.

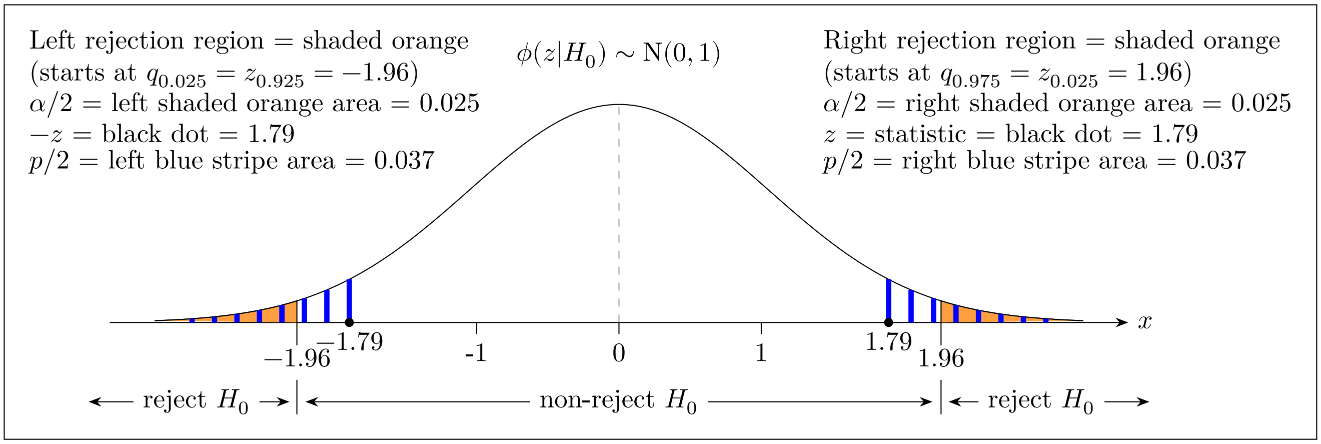
Solution: Let's do the test and then we'll explain the rationale behind the computation of the p-value.

Since $z > 0$, $p = 2P(Z > z) = 0.074$. Since, $p > \alpha = 0.05$, the data does not support rejecting the null hypothesis in favor of H_A.

Reason for the factor of 2 in the computation of p

The reason is essentially arithmetic. Remember, the purpose of the p-value is that $p \leq \alpha$ indicates that the test statistic is in the rejection region.

The picture below illustrates the following. For a two-sided test, each side of the rejection region has probability $\alpha/2$. So, if the test statistic is on the right, then it is in the rejection region if $P(Z > z) \leq \alpha/2$, i.e. if $p = 2P(Z > z) \leq \alpha$

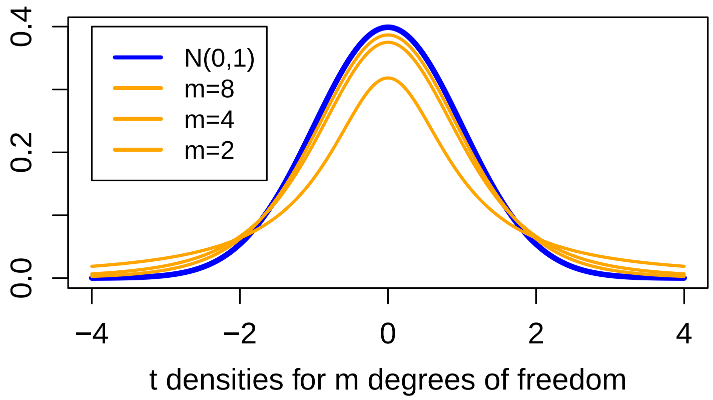


5.2 The Student t distribution

'Student' is the pseudonym used by the William Gosset who first described this test and distribution. See https://en.wikipedia.org/wiki/Student's_t-test

The t-distribution is symmetric and bell-shaped like the normal distribution. It has a parameter df which stands for degrees of freedom. For df small the t-distribution has more probability in its tails than the standard normal distribution. As df increases $t(df)$ becomes more and more like the standard normal distribution.

Here is a simple applet that shows $t(df)$ and compares it to the standard normal distribution: https://mathlets.org/mathlets/t-distribution/



As degrees of freedom increases the t-distribution becomes normal

5.3 R

As usual in R, the functions `pt, dt, qt, rt` correspond to cdf, pdf, quantiles, and random sampling for a t distribution. Remember that you can type `?dt` in RStudio to view the help file specifying the parameters of `dt`. For example, `pt(1.65,3)` computes the probability that x is less than or equal 1.65 given that x is sampled from the t distribution with 3 degrees of freedom, i.e. $P(x \leq 1.65)$ given that $x \sim t(3)$).

5.4 One sample t-test

For the z-test, we assumed that the variance of the underlying distribution of the data was known. However, it is often the case that we don't know σ and therefore we must estimate it from the data. In these cases, we use a one sample t-test instead of a z-test and the studentized mean in place of the standardized mean

- Data: we assume $x_1, x_2, \ldots, x_n \sim N(\mu, \sigma^2)$, where both μ and σ are unknown.
- Null hypothesis: $\mu = \mu_0$ for some specific value μ_0
- Test statistic:
$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$
where
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$
Here t is called the Studentized mean and s^2 is called the sample variance. The latter is an estimate of the true variance σ^2.
- Null distribution: $\phi(t \mid H_0)$ is the pdf of $T \sim t(n-1)$, the t distribution with $n-1$ degrees of freedom.*

- One-sided p-value (right side): $p = P(T \geq t \,|\, H_0)$
  One-sided p-value (left side): $p = P(T \leq t \,|\, H_0)$
  Two-sided p-value: $p = \begin{cases} 2P(T \geq t) & \text{if } t > 0 \\ 2P(T \leq t) & \text{if } t < 0. \end{cases}$

  Because of the symmetry of the distribution around 0, we can also write this as $p = P(|T| \geq |t|)$.

***Important note.** This is a good example of how we will work with significance tests. Once we know the distribution of the test statistic, all the tests have the same basic form. In this case, we make use of a theorem that says, for normal data the Studentized mean follows a t-distribution. We will not prove this in 18.05, but you can look up the proof if you want: `https://en.wikipedia.org/wiki/Student's_t-distribution#Derivation`

Example 2. Now suppose that in Example 1 the variance is unknown. That is, we have data that follows a normal distribution of unknown mean μ and and unknown variance σ. Suppose we collect the same data as before:

$$1,\ 2,\ 3,\ 6,\ -1$$

As above, let the null hypothesis H_0 be that $\mu = 0$ and the alternative hypothesis H_A be that $\mu > 0$. At a significance level of $\alpha = 0.05$, should we reject the null hypothesis?

Solution: There are 5 data points with average $\overline{x} = 2.2$. Because we have normal data with unknown mean and unknown variance we should use a one-sample t test. Computing the sample variance we get

$$s^2 = \frac{1}{4}\left((1-2.2)^2 + (2-2.2)^2 + (3-2.2)^2 + (6-2.2)^2 + (-1-2.2)^2\right) = 6.7$$

Our t-statistic is the Studentized mean:

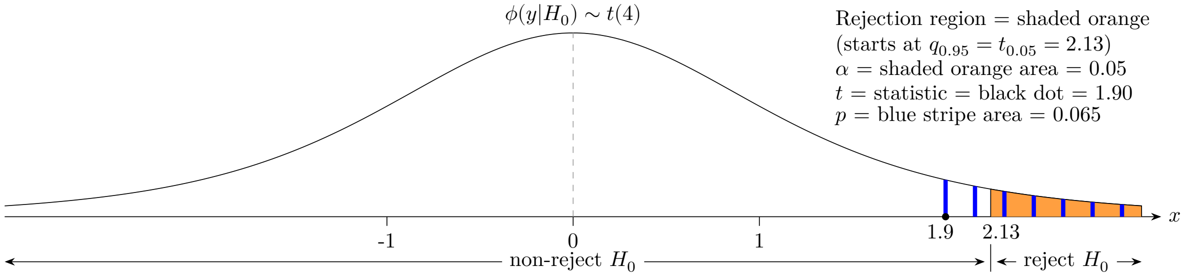
$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{2.2 - 0}{\sqrt{6.7}/\sqrt{5}} = 1.901$$

Our test is one-sided because the alternative hypothesis is one-sided. So (using R) the p-value is

$$p = P(T > t) = P(T > 1.901) = \texttt{1-pt(1.901,4)} = 0.065$$

Since $p > 0.05$, we do not reject the null hypothesis.

We can visualize the test as follows:



5.5 Two-sample t-test with equal variances

We next consider the case of comparing the means of two samples. For example, we might be interested in comparing the mean efficacies of two medical treatments.

- Data: We assume we have two sets of data drawn from normal distributions

  $$x_1, x_2, \ldots, x_n \sim N(\mu_1, \sigma^2)$$
  $$y_1, y_2, \ldots, y_m \sim N(\mu_2, \sigma^2)$$

  where the means μ_1 and μ_2 and the variance σ^2 are all unknown. Notice the assumption that the two distributions have the same variance. Also notice that there are n samples in the first group and m samples in the second.

- Null hypothesis: $\mu_1 = \mu_2$ (the values of μ_1 and μ_2 are not specified)

- Test statistic:

  $$t = \frac{\overline{x} - \overline{y}}{s_p},$$

  where s_p^2 is the pooled variance

  $$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right)$$

  Here s_x^2 and s_y^2 are the sample variances of the x_i and y_j respectively. The expression for t is somewhat complicated, but the basic idea remains the same and it still results in a known null distribution.

- Null distribution: $\phi(t \,|\, H_0)$ is the pdf of $T \sim t(n+m-2)$.

- One-sided p-value (right side): $p = P(T > t \,|\, H_0)$
  One-sided p-value (left side): $p = P(T < t \,|\, H_0)$
  Two-sided p-value: $p = P(|T| > |t|)$.

Note 1: Some authors use a different notation. They define the pooled variance as

$$s^2_{p\text{-other-authors}} = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

and what we called the pooled variance they point out is the estimated variance of $\overline{x} - \overline{y}$. That is,

$$s_p^2 \;=\; s_{p\text{-other-authors}} \times (1/n + 1/m) \;\approx\; s^2_{\overline{x}-\overline{y}}$$

Note 2: There is a version of the two-sample t-test that allows the two groups to have different variances. In this case the test statistic is a little more complicated but R will handle it with equal ease.

Note 3: We reiterate our 'important note' from above: It can be proved that under the assumptions on the data (independent samples, normal data, equal variances), the null distribution is a t-distribution. We won't prove this in 18.05. But knowing it, we can work with and understand the gist of the two-sample t-test in exactly the same way we can understand other significance tests.

Example 3. The following data comes from a real study in which 1408 women were admitted to a maternity hospital for (i) medical reasons or through (ii) unbooked emergency admission. The duration of pregnancy is measured in complete weeks from the beginning of the last menstrual period. We can summarize the data as follows:

Medical: 775 observations with $\bar{x}_M = 39.08$ and $s_M^2 = 7.77$.

Emergency: 633 observations with $\bar{x}_E = 39.60$ and $s_E^2 = 4.95$

Set up and run a two-sample t-test to investigate whether the mean duration differs for the two groups.

What assumptions did you make?

Solution: The pooled variance for this data is

$$s_p^2 = \frac{774(7.77) + 632(4.95)}{1406}\left(\frac{1}{775} + \frac{1}{633}\right) = 0.0187$$

The t statistic for the null distribution is

$$\frac{\bar{x_M} - \bar{y_E}}{s_p} = -3.8064$$

We have 1406 degrees of freedom. Using R to compute the two-sided p-value we get

$$p = P(|T| > |t|) = \texttt{2*pt(-3.8064, 1406) = 0.00015}$$

p is very small, much smaller than $\alpha = 0.05$ or $\alpha = 0.01$. Therefore we reject the null hypothesis in favor of the alternative that there is a difference in the mean durations.

Rather than compute the two-sided p-value exactly using a t-distribution we could have noted that with 1406 degrees of freedom the t distribution is essentially standard normal and 3.8064 is almost 4 standard deviations. So

$$P(|t| \geq 3.8064) \approx P(|z| \geq 3.8064) < 0.001$$

We assumed the data was normal and that the two groups had equal variances. Given the large difference between the sample variances this assumption may not be warranted.

In fact, there are other significance tests that test whether the data is approximately normal and whether the two groups have the same variance. In practice one might apply these first to determine whether a t test is appropriate in the first place. We don't have time to go into normality tests here, but we will see the F distribution used for equality of variances next week.

https://en.wikipedia.org/wiki/Normality_test
https://en.wikipedia.org/wiki/F-test_of_equality_of_variances

Null Hypothesis Significance Testing III
Class 19, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Given hypotheses and data, be able to identify an appropriate significance test from a list of common ones.
2. Given hypotheses, data, and a suggested significance test, know how to look up details and apply the significance test.

2 Introduction

In these notes we will collect together some of the most common significance tests, though by necessity we will leave out many other useful ones. Still, all significance tests follow the same basic pattern in their design and implementation, so by learning the ones we include you should be able to apply other ones as needed.

Designing a null hypothesis significance test (NHST):

- Specify null and alternative hypotheses.
- Choose a test statistic whose null distribution and alternative distribution(s) are known.
- Specify a rejection region. Very often this is done implicitly by specifying a significance level α and a method for computing p-values based on the tails of the null distribution.
- Compute power using the alternative distribution(s).

Running a NHST:

- Collect data and compute the test statistic.
- Check if the test statistic is in the rejection region. Most often this is done implicitly by checking if $p < \alpha$. If so, we 'reject the null hypothesis in favor of the alternative hypothesis'. Otherwise we conclude 'the data does not support rejecting the null hypothesis'.

Note the careful phrasing: when we fail to reject H_0, we do not conclude that H_0 is true. The failure to reject may have other causes. For example, we might not have enough data to clearly distinguish H_0 and H_A, whereas more data might indicate that we should reject H_0.

3 Population parameters and sample statistics

Example 1. If we randomly select 10 men from a population and measure their heights we say that we have sampled the heights from the population. In this case the sample mean, say $\overline{x}$, is the mean of the sampled heights. It is a statistic and we know its value explicitly. On the other hand, the true average height of the population, say μ, is unknown and we can only estimate its value. We call μ a population parameter.

The main purpose of significance testing is to use sample statistics to draw conlusions about population parameters. For example, we might test if the average height of men in a given population is greater than 70 inches.

4 A gallery of common significance tests related to the normal distribution

We will show a number of tests that all assume normal data. For completeness we will include the z and t tests we've already explored.

You shouldn't try to memorize these tests. It is a hopeless task to memorize the tests given here and even more hopeless to memorize all the tests we've left out. Rather, your goal should be to be able to find the correct test when you need it. Pay attention to the types of hypotheses the tests are designed to distinguish and the assumptions about the data needed for the test to be valid. We will work through the details of these tests in class and on homework.

The null distributions for all of these tests are all related to the normal distribution by explicit formulas. We will not go into the details of these distributions or the arguments showing how they arise as the null distributions in our significance tests. However, the arguments are accessible to anyone who knows calculus and is interested in understanding them. Given the name of any distribution, you can easily look up the details of its construction and properties online. You can also use R to explore the distribution numerically and graphically.

When analyzing data with any of these tests one thing of key importance is to verify that the assumptions are true or at least approximately true. For example, you shouldn't use a test that assumes the data is normal unless you've checked that the data is approximately normal.

The script class19.r contains examples of using R to run some of these tests. It is posted in our usual place for R code.

4.1 z-test

- Use: Test if the population mean equals a hypothesized mean.
- Data: $x_1, x_2, \ldots, x_n$.
- Assumptions: The data are independent normal samples:
  $x_i \sim N(\mu, \sigma^2)$ where μ is unknown, but σ is known.
- H_0: For a specified μ_0, $\mu = \mu_0$.

- H_A:

Two-sided:	$\mu \neq \mu_0$
one-sided-greater:	$\mu > \mu_0$
one-sided-less:	$\mu < \mu_0$

- Test statistic: $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$
- Null distribution: $\phi(z \mid H_0)$ is the pdf of $Z \sim N(0, 1)$.
- p-value:

Two-sided:	$p = P(\lvert Z\rvert > z)$	=	`2*(1-pnorm(abs(z), 0, 1))`
one-sided-greater:	$p = P(Z > z)$	=	`1 - pnorm(z, 0, 1)`
one-sided-less:	$p = P(Z < z)$	=	`pnorm(z, 0, 1)`

- R code: There does not seem to be a single R function in the base R packages that runs a z-test. There are other packages you can install that have a z.test function. Of course, it is easy enough to get R to compute the z score and p-value. There is an example of this in class19.r.

Example 2. We quickly reprise our example from the class 17 notes.

IQ is normally distributed in the population according to a N$(100, 15^2)$ distribution. We suspect that most MIT students have above average IQ so we frame the following hypotheses.

H_0 = MIT student IQs are distributed identically to the general population
= MIT IQ's follow a N$(100, 15^2)$ distribution.
H_A = MIT student IQs tend to be higher than those of the general population
= the average MIT student IQ is greater than 100.

Notice that H_A is one-sided.

Suppose we test 9 students and find they have an average IQ of $\bar{x} = 112$. Can we reject H_0 at a significance level $\alpha = 0.05$?

Solution: Our test statistic is

$$z = \frac{\bar{x} - 100}{15/\sqrt{9}} = \frac{36}{15} = 2.4.$$

The right-sided p-value is therefore

$$p = P(Z \geq 2.4) = \texttt{1- pnorm(2.4,0,1) = 0.0081975.}$$

Since $p \leq \alpha$ we reject the null hypothesis in favor of the alternative hypothesis that MIT students have higher IQs on average.

4.2 One-sample t-test of the mean

- Use: Test if the population mean equals a hypothesized mean.
- Data: $x_1, x_2, \ldots, x_n$.
- Assumptions: The data are independent normal samples:

  $x_i \sim N(\mu, \sigma^2)$ where both μ and σ are unknown.

- H_0: For a specified μ_0, $\mu = \mu_0$
- H_A:
  Two-sided: $\mu \neq \mu_0$
  one-sided-greater: $\mu > \mu_0$
  one-sided-less: $\mu < \mu_0$
- Test statistic: $t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$,

  where s^2 is the sample variance: $s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$
- Null distribution: $\phi(t \,|\, H_0)$ is the pdf of $T \sim t(n-1)$.
  (Student t-distribution with $n-1$ degrees of freedom)
- p-value:
  Two-sided: $p = P(|T| > t)$ = `2*(1-pt(abs(t), n-1))`
  one-sided-greater: $p = P(T > t)$ = `1 - pt(t, n-1)`
  one-sided-less: $p = P(T < t)$ = `pt(t, n-1)`
- R code example: For data $x = 1, 3, 5, 7, 2$ we can run a one-sample t-test with H_0: $\mu_0 = 2.5$ using the R command:

  ```
  t.test(x, mu = 2.5, alternative="two.sided")
  ```

  This will return a several pieces of information including the mean of the data, t-value and the two-sided p-value. See the help for this function for other argument settings.

Example 3. Look in the class 18 notes or slides for an example of this test. The class 19 example R code also gives an example.

4.3 Two-sample t-test for comparing means

4.3.1 The case of equal variances

We start by describing the test assuming equal variances.

- Use: Test if the population means from two populations differ by a hypothesized amount.
- Data: $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$.
- Assumptions: Both groups of data are independent normal samples:

  $$x_i \sim N(\mu_x, \sigma^2)$$
  $$y_j \sim N(\mu_y, \sigma^2)$$

  where both μ_x and μ_y are unknown and possibly different. The variance σ^2 is unknown, but the same for both groups.
- H_0: For a specified $\Delta\mu$ the difference of means $\mu_x - \mu_y = \Delta\mu$
- H_A:
  Two-sided: $\mu_x - \mu_y \neq \Delta\mu$
  one-sided-greater: $\mu_x - \mu_y > \Delta\mu$
  one-sided-less: $\mu_x - \mu_y < \Delta\mu$

- Test statistic: $t = \dfrac{\overline{x} - \bar{y} - \Delta\mu}{s_P}$,
  where s_x^2 and s_y^2 are the sample variances of the x and y data respectively, and s_P^2 is (sometimes called) the pooled sample variance:
  $$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right) \quad \text{and} \quad df = n+m-2$$
- Null distribution: $\phi(t \,|\, H_0)$ is the pdf of $T \sim t(df)$, the t-distribution with $df = n+m-2$ degrees of freedom.
- p-value:

Two-sided:	$p = P(\lvert T\rvert > t)$	=	`2*(1-pt(abs(t), df))`
one-sided-greater:	$p = P(T > t)$	=	`1 - pt(t, df)`
one-sided-less:	$p = P(T < t)$	=	`pt(t, df)`

- R code: The R function `t.test` will run a two-sample t-test. See the example code in class19.r. In `t.test` the argument `mu` is used for what we have called $\Delta\mu$.

Notes: 1. Most often the test is done with $\Delta\mu = 0$. That is, the null hypothesis is the the means are equal, i.e. $\mu_x - \mu_y = 0$.

2. If the x and y data have the same length $n = m$, then the formula for s_p^2 becomes simpler:
$$s_p^2 = \frac{s_x^2 + s_y^2}{n}$$

Example 4. Look in the class 18 notes or slides for an example of the two-sample t-test.

4.3.2 The case of unequal variances

There is a form of the t-test for when the variances are not assumed equal. It is sometimes called Welch's t-test.

This looks exactly the same as the case of equal except for a small change in the assumptions and the formula for the pooled variance:

- Use: Test if the population means from two populations differ by a hypothesized amount.
- Data: $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$.
- Assumptions: Both groups of data are independent normal samples:
  $$x_i \sim N(\mu_x, \sigma_x^2)$$
  $$y_j \sim N(\mu_y, \sigma_y^2)$$
  where both μ_x and μ_y are unknown and possibly different. The variances σ_x^2 and σ_y^2 are unknown and not assumed to be equal.
- H_0, H_A: Exactly the same as the case of equal variances.

- Test statistic: $t = \dfrac{\overline{x} - \bar{y} - \Delta\mu}{s_P}$,
  where s_x^2 and s_y^2 are the sample variances of the x and y data respectively, and s_P^2 is (sometimes called) the pooled sample variance:
  $$s_p^2 = \frac{s_x^2}{n} + \frac{s_y^2}{m} \quad \text{and} \quad df = \frac{(s_x^2/n + s_y^2/m)^2}{(s_x^2/n)^2/(n-1) + (s_y^2/m)^2/(m-1)}$$
- Null distribution: $\phi(t \,|\, H_0)$ is the pdf of $T \sim t(df)$, the t distribution with df degrees of freedom.
- p-value: Exactly the same as the case of equal variances.
- R code: The function `t.test` also handles this case if you set the argument `var.equal=FALSE`.

Notes. 1. In truth, the null distribution given above only approximates the exact null distribution.

2. Notice that the degrees of freedom are unlikely to be a whole number.

3. Some people recommend always using Welch's t-test, even if the variances are believed to be equal. This avoids making the assumption that the variances are equal and has very little downside if they are equal.

4.3.3 The paired two-sample t-test

When the data naturally comes in pairs (x_i, y_i), we can use the paired two-sample t-test. (After checking the assumptions are valid!)

Example 5. To measure the effectiveness of a cholesterol lowering medication we might test each subject before and after treatment with the drug. So for each subject we have a pair of measurements:

$$\begin{aligned} x_i &= \text{cholesterol level before treatment} \\ y_i &= \text{cholesterol level after treatment.} \end{aligned}$$

Example 6. To measure the effectiveness of a cancer treatment we might pair each subject who received the treatment with one who did not. In this case we would want to pair subjects who are similar in terms of stage of the disease, age, sex, etc.

- Use: Test if the average difference between paired values in a population equals a hypothesized value.
- Data: $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ must have the same length.
- Assumptions: The differences $w_i = x_i - y_i$ between the paired samples are independent draws from a normal distribution $\mathrm{N}(\mu, \sigma^2)$, where μ and σ are unknown.
- H_0: For a specified μ_0, $\mu = \mu_0$.
- H_A:

Two-sided:	$\mu \neq \mu_0$
one-sided-greater:	$\mu > \mu_0$
one-sided-less:	$\mu < \mu_0$

- Test statistic: $t = \dfrac{\overline{w} - \mu_0}{s/\sqrt{n}}$,

  where s^2 is the sample variance: $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(w_i - \overline{w})^2$
- Null distribution: $\phi(t \mid H_0)$ is the pdf of $T \sim t(n-1)$.
  (Student t-distribution with $n-1$ degrees of freedom)
- p-value:

  Two-sided: $p = P(|T| > t) =$ `2*(1-pt(abs(t), n-1))`

  one-sided-greater: $p = P(T > t) =$ `1 - pt(t, n-1)`

  one-sided-less: $p = P(T < t) =$ `pt(t, n-1)`
- R code: The R function `t.test` will do a paired two-sample test if you set the argument `paired=TRUE`. You can also run a one-sample t-test on $x - y$. There are examples of both of these in class19.r

Notes. 1. This is just a one-sample t-test using w_i.

2. Another way to write the assumption is that we assume a relation between x_i and y_i of the form $y_i = x_i + \mu + e$. Here μ is some (unknown) constant, and e is random error (noise) of mean 0 and (unknown) variance σ^2.

Example 7. The following example is taken from Rice [1]

To study the effect of cigarette smoking on platelet aggregation Levine (1973) drew blood samples from 11 subjects before and after they smoked a cigarette and measured the extent to which platelets aggregated. Here is the data:

Before	25	25	27	44	30	67	53	53	52	60	28
After	27	29	37	56	46	82	57	80	61	59	43
Difference	2	4	10	12	16	15	4	27	9	-1	15

The null hypothesis is that smoking had no effect on platelet aggregation, i.e. that the difference between before and after should have mean $\mu_0 = 0$. We ran a paired two-sample t-test to test this hypothesis. Here is the R code: (It's also in class19.r.)

```
before.cig = c(25,25,27,44,30,67,53,53,52,60,28)
after.cig = c(27,29,37,56,46,82,57,80,61,59,43)
mu0 = 0
result = t.test(after.cig, before.cig, alternative="two.sided", mu=mu0, paired=TRUE)
print(result)
```

Here is the output:

```
    Paired t-test
data: after.cig and before.cig
t = 4.2716, df = 10, p-value = 0.001633
alternative hypothesis: true difference in means is not equal to 0
mean of the differences: 10.27273
```

We got the same results with the one-sample t-test:

```
t.test(after.cig - before.cig, mu=0)
```

[1] John Rice, *Mathematical Statistics and Data Analysis*, 2nd edition, p. 412. This example references P.H Levine (1973) An acute effect of cigarette smoking on platelet function. *Circulation, 48, 619-623.*

4.4 One-way ANOVA (F-test for equal means)

- Use: Test if the population means from n groups are all the same.
- Data: (n groups, m samples from each group)

  $$\begin{array}{llll} x_{1,1}, & x_{1,2}, & \dots, & x_{1,m} \\ x_{2,1}, & x_{2,2}, & \dots, & x_{2,m} \\ & & \dots & \\ x_{n,1}, & x_{n,2}, & \dots, & x_{n,m} \end{array}$$

- Assumptions: Data for each group is an independent normal sample drawn from distributions with (possibly) different means but the same variance:

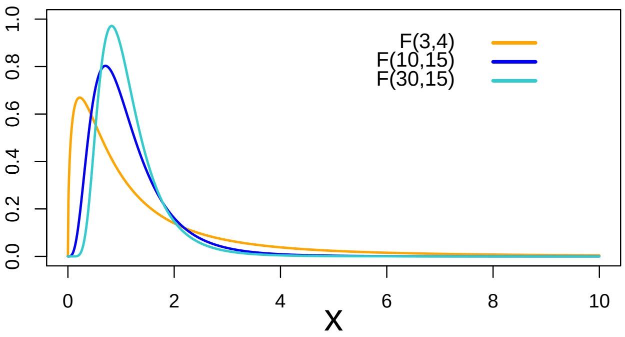
  $$\begin{aligned} x_{1,j} &\sim N(\mu_1, \sigma^2) \\ x_{2,j} &\sim N(\mu_2, \sigma^2) \\ &\dots \\ x_{n,j} &\sim N(\mu_n, \sigma^2) \end{aligned}$$

  The group means μ_i are unknown and possibly different. The variance σ is unknown, but the same for all groups.

- H_0: All the means are identical $\mu_1 = \mu_2 = \dots = \mu_n$.
- H_A: Not all the means are the same.
- Test statistic: $f = \frac{\text{MS}_B}{\text{MS}_W}$, where

  $$\begin{aligned} \bar{x}_i &= \text{mean of group } i \\ &= \frac{x_{i,1} + x_{i,2} + \dots + x_{i,m}}{m}. \\ \overline{x} &= \text{grand mean of all the data.} \\ s_i^2 &= \text{sample variance of group } i \\ &= \frac{1}{m-1}\sum_{j=1}^{m}(x_{i,j} - \bar{x}_i)^2. \\ \text{MS}_B &= \text{between group variance} \\ &= m \times \text{sample variance of group means} \\ &= \frac{m}{n-1}\sum_{i=1}^{n}(\bar{x}_i - \overline{x})^2. \\ \text{MS}_W &= \text{average within group variance} \\ &= \text{sample mean of } s_1^2, \dots, s_n^2 \\ &= \frac{s_1^2 + s_2^2 + \dots + s_n^2}{n} \end{aligned}$$

- Idea: If the μ_i are all equal, test statistic f, which is a ratio, should be near 1. If they are not equal then MS_B should be larger while MS_W should remain about the same, so f should be larger. We won't give a proof of this.
- Null distribution: $\phi(f \mid H_0)$ is the pdf of $F \sim F(n-1, n(m-1))$.
  This is the F-distribution with $(n-1)$ and $n(m-1)$ degrees of freedom. Several F-distributions are plotted below.
- p-value: $p = P(F > f) =$ `1- pf(f, n-1, n*(m-1)))`



Notes: 1. ANOVA tests whether all the means are the same. It does not test whether some subset of the means are the same.
2. There is a test where the variances are not assumed equal.
3. There is a test where the groups don't all have the same number of samples.
4. R has a function `aov()` to run ANOVA tests.
5. See: https://en.wikipedia.org/wiki/F-test

Example 8. The table shows patients' perceived level of pain (on a scale of 1 to 6) after 3 different medical procedures.

T_1	T_2	T_3
2	3	2
4	4	1
1	6	3
5	1	3
3	4	5

(1) Set up and run an F-test comparing the means of these 3 treatments.

(2) Based on the test, what might you conclude about the treatments?

Solution: Using the code below, the F statistic is 0.325 and the p-value is 0.729 At any reasonable significance level we will fail to reject the null hypothesis that the average pain level is the same for all three treatments..

Note, it is not reasonable to conclude the the null hypothesis is true. With just 5 data points per procedure we might simply lack the power to distinguish different means.

R code to perform the test

```
# DATA ----
T1 = c(2,4,1,5,3)
T2 = c(3,4,6,1,4)
T3 = c(2,1,3,3,5)

procedure = c(rep('T1',length(T1)),rep('T2',length(T2)),rep('T3',length(T3)))
pain = c(T1,T2,T3)
data.pain = data.frame(procedure,pain)
aov.data = aov(pain~procedure,data=data.pain) # do the analysis of variance
print(summary(aov.data)) # show the summary table

# class19.r also shows code to compute the ANOVA by hand.
```

The summary shows a p-value (shown as `Pr(>F)`) of 0.729. Therefore we do not reject the null hypothesis that all three group population means are the same.

4.5 Chi-square test for goodness of fit

This is a test of how well a hypothesized probability distribution fits a set of data. The test statistic is called a chi-square statistic and the null distribution associated to the chi-square statistic is the chi-square distribution. It is denoted by $\chi^2(df)$ where the parameter df is called the degrees of freedom.

Suppose we have an unknown probability mass function given by the following table.

Outcomes	ω_1	ω_2	...	ω_n
Probabilities	p_1	p_2	...	p_n

In the chi-square test for goodness of fit we hypothesize a set of values for the probabilities. Typically we will hypothesize that the probabilities follow a known distribution with certain parameters, e.g. binomial, Poisson, multinomial. The test then tries to determine if this set of probabilities could reasonably have generated the data we collected.

- Use: Test whether discrete data fits a specific finite probability mass function.
- Data: An observed count O_i for each possible outcome ω_i.
- Assumptions: None
- H_0: The data was drawn from a specific discrete distribution.
- H_A: The data was drawn from a different distribution.
- Test statistic: The data consists of observed counts O_i for each ω_i. From the null hypothesis probability table we get a set of expected counts E_i. There are two statistics that we can use:

$$\text{Likelihood ratio statistic } G = 2 * \sum O_i \ln\left(\frac{O_i}{E_i}\right)$$

$$\text{Pearson's chi-square statistic } X^2 = \sum \frac{(O_i - E_i)^2}{E_i}.$$

  It is a theorem that under the null hypothesis $X^2 \approx G$ and both are approximately chi-square. Before computers, X^2 was used because it was easier to compute. Now, it is better to use G although you will still see X^2 used quite often.

- Degrees of freedom df: For chi-square tests the number of degrees of freedom can be a bit tricky. In this case $df = n - 1$. It is computed as the number of cell counts that can be freely set under H_A consistent with the statistics needed to compute the expected cell counts assuming H_0.
- Null distribution: Assuming H_0, both statistics (approximately) follow a chi-square distribution with df degrees of freedom. That is both $\phi(G \mid H_0)$ and $\phi(X^2 \mid H_0)$ have (approximately) the same pdf as $Y \sim \chi^2(df)$.
- p-value: Extreme data means large values of X^2, i.e. large differences between the observed and expected counts. So,

  $p = P(Y > G) =$ `1 - pchisq(G, df)`

  $p = P(Y > X^2) =$ `1 - pchisq(`X^2`, df)`

- R code: The R function `chisq.test` can be used to do the computations for a chi-square test use X^2. For G you either have to do it by hand or find a package that has a function. (It will probably be called `likelihood.test` or `G.test`.

Notes. 1. When the likelihood ratio statistic G is used the test is also called a G-test or a likelihood ratio test.

Example 9. First chi-square example. Suppose we have an experiment that produces numerical data. For this experiment the possible outcomes are 0, 1, 2, 3, 4, 5 or more. We run 51 trials and count the frequency of each outcome, getting the following data:

Outcomes	0	1	2	3	4	≥ 5
Observed counts	3	10	15	13	7	3

Suppose our null hypothesis H_0 is that the data is drawn from 51 trials of a binomial(8, 0.5) distribution and our alternative hypothesis H_A is that the data is drawn from some other distribution. Do all of the following:

1. Make a table of the observed and expected counts.
2. Compute both the likelihood ratio statistic G and Pearson's chi-square statistic X^2.
3. Compute the degrees of freedom of the null distribution.
4. Compute the p-values corresponding to G and X^2.

Solution: All of the R code used for this example is in class19.r.

1. Assuming H_0 the data truly comes from a binomial(8, 0.5) distribution. We have 51 total observations, so the expected count for each outcome is just 51 times its probability. We computed the binomial(8, 0.5) probabilities and expected counts in R:

Outcomes	0	1	2	3	4	≥ 5
Observed counts	3	10	15	13	7	3
H_0 probabilities	0.0039	0.0313	0.1094	0.2188	0.2734	0.3633
Expected counts	0.20	1.59	5.58	11.16	13.95	18.53

2. Using the formulas above we compute that $X^2 = 116.41$ and $G = 66.08$

3. The only statistic used in computing the expected counts was the total number of observations 51. So, the degrees of freedom is 5, i.e we can set 5 of the cell counts freely and the last is determined by requiring that the total number is 51.

4. The p-values are pG =`1 - pchisq(G, 5)` and $pX2$ = `1 - pchisq(`X^2`, 5)`. Both p-values are effectively 0. For almost any significance level we would reject H_0 in favor of H_A.

4.5.1 Degrees of freedom in chi-square tests

We alreay gave a quick definition of degrees of freedom for a chi-square test. Here we will try to go a little slower in showing how to compute degrees of freedom.

To start, recall that in a chi-square test, our table has n observed counts. Then, we use observed counts and the null hypothesis to compute n expected counts. This is typically done by computing some statistics and using them estimate the parameters needed to compute the expected counts. For example, in the previous example the statistic computed was the total number of counts.

Now, imagine that we are allowed to fabricate the n observed counts, but we demand that our made up observations produce the same statistics as the true observed counts. That is, our imaginary observed counts need to produce the same expected counts as the true data. The degrees of freedom is the number of fake observed counts we can freely choose. The rest will be determined by our constraint that they produce the same statistics.

Example 10. (Degrees of freedom.) Suppose we have the following observed counts

Outcomes	0	1	2	3	4	5
Observed counts	6	9	13	12	7	3

Suppose our null hypothesis H_0 is that the data producing these counts was drawn from 50 trials of a binomial(5, θ) distribution. Our alternative hypothesis H_A is that the data is drawn from some other distribution. Run a chi-square test of these hypotheses with significance 0.05.

Solution: To compute expected counts we need a value of θ. Since it is not known, we have to use the data to estimate it.

The total number of observations is 50. So, the mean of the data is

$$m = \frac{6 \cdot 0 + 9 \cdot 1 + 13 \cdot 2 + 12 \cdot 3 + 7 \cdot 4 + 3 \cdot 5}{50} = 2.28.$$

The expected value of a binom(5,θ) distribution, is 5θ. The maximum likelihood estimate for θ is $\hat{\theta} = m/5 = 0.456$.

Now, just like in the previous example, we can compute expected counts for each possible outcome. The expected count of outcome k is $50 \cdot p(k)$. In R this is `50*dbinom(k, 50,` $\hat{\theta}$`)`. We have the following table

Outcomes	0	1	2	3	4	5
Observed counts	6	9	13	12	7	3
Expected counts	2.38	9.98	16.74	14.03	5.88	0.99

To determine the degrees of freedom:
(i) We have 6 observed counts.

(ii) To compute the expected counts we need the total number of counts = 50 and our estimate of $\hat{\theta} = m/5$ That is, we have two constraints: the total number of counts is 50, and mean $m = 2.28$.

So, to get the same expected counts, we could choose 4 of the observed counts freely and then set last two counts so the constraints are met. Thus, there are 4 degrees of freedom.

More briefly: 6 observed counts - 2 constraints = 4 degrees of freedom.

Using R we can compute $G = 8.01$, with p-value 0.09. Thus, at significance level 0.05 we would not reject the null hypothesis.

Note, the X^2 statistic is 11.05 with p-value 0.026. Clearly this example is a borderline case, since we reject H_0 when using X^2 and we don't when using G.

It bears repeating, for reasons like this, we never say H_0 is false. The most we can say is that the data does not support rejecting H_0.

4.5.2 More examples

Example 11. Mendel's genetic experiments (Adapted from Rice *Mathematical Statistics and Data Analysis, 2nd ed.*, example C, p.314)

In one of his experiments on peas Mendel looked at 2 genetic trait pairs: smooth/wrinkled and yellow/green. Symbolically we label a smooth gene S and a wrinkled gene s. Likewise we use Y and y for yellow and green respectively.

Mendel started by selecting a parent generation of homozygous plants. They were either smooth/yellow (genes $SSYY$) and wrinkled/green (genes $ssyy$). He crossed the smooth/yellow with the wrinkled/green peas creating the, so called, F_1 generation consisting of plants with genes $SsYy$. Since smooth (S) and yellow Y are both dominant traits, all these plants were smooth/yellow.

He then crossed 556 pairs of the F_1 generation to create the F_2 generation. We would expect 1/4 of the F_2 generation to have two smooth genes (SS), 1/4 to have two wrinkled genes (ss), and the remaining 1/2 to be heterozygous (Ss). We also expect these fractions for yellow (Y) and green (y) genes. If the color and smoothness genes are inherited independently and smooth and yellow are both dominant we'd expect the following table of frequencies for phenotypes.

	Yellow	Green	
Smooth	9/16	3/16	3/4
Wrinkled	3/16	1/16	1/4
	3/4	1/4	1

Probability table for the null hypothesis

So from the 556 crosses the expected number of smooth yellow peas is $556 \times 9/16 = 312.75$. Likewise for the other possibilities. Here is a table giving the observed and expected counts from Mendel's experiments.

	Observed count	Expected count
Smooth yellow	315	312.75
Smooth green	108	104.25
Wrinkled yellow	102	104.25
Wrinkled green	31	34.75

The null hypothesis is that the observed counts are random samples distributed according to the frequency table given above. We use the counts to compute our statistics

The likelihood ratio statistic is

$$\begin{aligned} G &= 2 * \sum O_i \ln\left(\frac{O_i}{E_i}\right) \\ &= 2 * \left(315 \ln\left(\frac{315}{412.75}\right) + 108 \ln\left(\frac{108}{104.25}\right) + 102 \ln\left(\frac{102}{104.25}\right) + 31 \ln\left(\frac{31}{34.75}\right)\right) \\ &= 0.618 \end{aligned}$$

Pearson's chi-square statistic is

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{2.75^2}{312.75} + \frac{3.75^2}{104.25} + \frac{2.25^2}{104.25} + \frac{3.75^2}{34.75} = 0.604$$

You can see that the two statistics are very close. This is usually the case. In general the likelihood ratio statistic is more robust and should be preferred.

The degrees of freedom is 3, because there are 4 observed quantities and one relation between them, i.e. they sum to 556. So, under the null hypothesis G follows a $\chi^2(3)$ distribution. Using R to compute the p-value we get

$$p = \texttt{1- pchisq(0.618, 3) = 0.8923}$$

Assuming the null hypothesis we would see data at least this extreme almost 90% of the time. We would not reject the null hypothesis for any reasonable significance level.

The p-value using Pearson's statistic is 0.8955 –nearly identical.

The script class19.r shows these calculations and also how to use `chisq.test` to run a chi-square test directly.

4.6 Chi-square test for homogeneity

This is a test to see if several independent sets of random data are all drawn from the same distribution. (The meaning of homogeneity in this case is that all the distributions are the same.)

- Use: Test whether m different independent sets of discrete data are drawn from the same distribution.
- Outcomes: ω_1, ω_2, ..., ω_n are the possible outcomes. These are the same for each set of data.
- Data: We assume m independent sets of data giving counts for each of the possible outcomes. That is, for data set i we have an observed count $O_{i,j}$ for each possible outcome ω_j.
- Assumptions: None
- H_0: Each data set is drawn from the same distribution. (We don't specify what this distribution is.)
- H_A: The data sets are not all drawn from the same distribution.
- Test statistic: See the example below. There are mn cells containing counts for each outcome for each data set. Using the null distribution we can estimate expected counts for each of the data sets. The statistics X^2 and G are computed exactly as above.
- Degrees of freedom df: $(m-1)(n-1)$. (See the example below.)
- The null distribution $\chi^2(df)$. The p-values are computed just as in the chi-square test for goodness of fit.
- R code: The R function `chisq.test` can be used to do the computations for a chi-square test use X^2. For G you either have to do it by hand or find a package that has a function. (It will probably be called `likelihood.test` or `G.test`.

Example 12. Someone claims to have found a long lost work by William Shakespeare. They ask you to test whether or not the play was actually written by Shakespeare .

You go to `https://www.opensourceshakespeare.org` and pick a random 12 pages from *King Lear* and count the use of common words. You do the same thing for the 'long lost work'. You get the following table of counts.

Word	a	an	this	that
King Lear	150	30	30	90
Long lost work	90	20	10	80

Using this data, set up and evaluate a significance test of the claim that the long lost book is by William Shakespeare. Use a significance level of 0.1.

Solution: The null hypothesis H_0: For the 4 words counted the long lost book has the same relative frequencies as the counts taken from *King Lear*.

The total word count of both books combined is 500, so the the maximum likelihood estimate of the relative frequencies assuming H_0 is simply the total count for each word divided by the total word count.

Word	a	an	this	that	Total count
King Lear	150	30	30	90	300
Long lost work	90	20	10	80	200
totals	240	50	40	170	500
rel. frequencies under H_0	240/500	50/500	40/500	170/500	500/500

Now the expected counts for each book under H_0 are the total count for that book times the relative frequencies in the above table. The following table gives the counts: (observed, expected) for each book.

Word	a	an	this	that	Totals
King Lear	(150, 144)	(30, 30)	(30, 24)	(90, 102)	(300, 300)
Long lost work	(90, 96)	(20, 20)	(10, 16)	(80, 68)	(200, 200)
Totals	(249, 240)	(50, 50)	(40, 40)	(170, 170)	(500, 500)

The chi-square statistic is

$$\begin{aligned} X^2 &= \sum \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{6^2}{144} + \frac{0^2}{30} + \frac{6^2}{24} + \frac{12^2}{102} + \frac{6^2}{96} + \frac{0^2}{20} + \frac{6^2}{16} + \frac{12^2}{68} \\ &\approx 7.9 \end{aligned}$$

There are 8 cells and all the marginal counts are fixed because they were needed to determine the expected counts. To be consistent with these statistics we could freely set the values in 3 cells in the table, e.g. the 3 blue cells, then the rest of the cells are determined in order to make the marginal totals correct. Thus $df = 3$. (Or we could recall that $df = (m-1)(n-1) = (3)(1) = 3$, where m is the number of columns and n is the number of rows.)

Using R we find `p = 1-pchisq(7.9,3) = 0.048`. Since this is less than our significance level of 0.1 we reject the null hypothesis that the relative frequencies of the words are the same in both books.

If we make the further assumption that all of Shakespeare's plays have similar word frequencies (which is something we could check) we conclude that the book is probably not by Shakespeare.

4.7 Other tests

There are far too many other tests to even make a dent. We will see some of them in class and on psets. Again, we urge you to master the paradigm of NHST and recognize the importance of choosing a test statistic with a known null distribution.

Comparison of frequentist and Bayesian inference. Class 20, 18.05 Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to explain the difference between the p-value and a posterior probability to a doctor.

2 Introduction

We have now learned about two schools of statistical inference: Bayesian and frequentist. Both approaches allow one to evaluate evidence about competing hypotheses. In these notes we will review and compare the two approaches, starting from Bayes' formula.

3 Bayes' formula as touchstone

In our first unit (probability) we learned Bayes' formula, a perfectly abstract statement about conditional probabilities of events:

$$P(A \,|\, B) = \frac{P(B \,|\, A)P(A)}{P(B)}.$$

We began our second unit (Bayesian inference) by reinterpreting the events in Bayes' formula:

$$P(\mathcal{H} \,|\, \mathcal{D}) = \frac{P(\mathcal{D} \,|\, \mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}.$$

Now $\mathcal{H}$ is a hypothesis and $\mathcal{D}$ is data which may give evidence for or against $\mathcal{H}$. Each term in Bayes' formula has a name and a role.

- The prior $P(\mathcal{H})$ is the probability that $\mathcal{H}$ is true before the data is considered.
- The posterior $P(\mathcal{H} \,|\, \mathcal{D})$ is the probability that $\mathcal{H}$ is true after the data is considered.
- The likelihood $P(\mathcal{D} \,|\, \mathcal{H})$ is the evidence about $\mathcal{H}$ provided by the data $\mathcal{D}$.
- $P(\mathcal{D})$ is the total probability of the data taking into account all possible hypotheses.

If the prior and likelihood are known for all hypotheses, then Bayes' formula computes the posterior exactly. Such was the case when we rolled a die randomly selected from a cup whose contents you knew. We call this the deductive logic of probability theory, and it gives a direct way to compare hypotheses, draw conclusions, and make decisions.

In most experiments, the prior probabilities on hypotheses are not known. In this case, our recourse is the art of statistical inference: we either make up a prior (Bayesian) or do our best using only the likelihood (frequentist).

The Bayesian school models uncertainty by a probability distribution over hypotheses. One's ability to make inferences depends on one's degree of confidence in the chosen prior, and the robustness of the findings to alternate prior distributions may be relevant and important.

The frequentist school only uses conditional distributions of data given specific hypotheses. The presumption is that some hypothesis (parameter specifying the conditional distribution of the data) is true and that the observed data is sampled from that distribution. In particular, the frequentist approach does not depend on a subjective prior that may vary from one investigator to another.

These two schools may be further contrasted as follows:

Bayesian inference

- uses probabilities for both hypotheses and data.
- depends on the prior and likelihood of observed data.
- requires one to know or construct a 'subjective prior'.
- dominated statistical practice before the 20th century.
- may be computationally intensive due to integration over many parameters.

Frequentist inference (NHST)

- never uses or gives the probability of a hypothesis (no prior or posterior).
- depends on the likelihood $P(\mathcal{D} \mid \mathcal{H}))$ for both observed and unobserved data.
- does not require a prior.
- dominated statistical practice during the 20th century.
- tends to be less computationally intensive.

Frequentist measures like p-values and confidence intervals continue to dominate research, especially in the life sciences. However, in the current era of powerful computers and big data, Bayesian methods have undergone an enormous renaissance in fields like machine learning and genetics. There are now a number of large, ongoing clinical trials using Bayesian protocols, something that would have been hard to imagine a generation ago. While professional divisions remain, the consensus forming among top statisticians is that the most effective approaches to complex problems often draw on the best insights from both schools working in concert.

4 Critiques and defenses

4.1 Critique of Bayesian inference

1. The main critique of Bayesian inference is that a subjective prior is, well, subjective. There is no single method for choosing a prior, so different people will produce different priors and may therefore arrive at different posteriors and conclusions.

2. Furthermore, there are philosophical objections to assigning probabilities to hypotheses, as hypotheses do not constitute outcomes of repeatable experiments in which one can measure long-term frequency. Rather, a hypothesis is either true or false, regardless of whether one knows which is the case. A coin is either fair or unfair; treatment 1 is either better or worse than treatment 2; the sun will or will not come up tomorrow.

4.2 Defense of Bayesian inference

1. The probability of hypotheses is exactly what we need to make decisions. When the doctor tells me a screening test came back positive I want to know what is the probability this means I'm sick. That is, I want to know the probability of the hypothesis "I'm sick".

2. Using Bayes' theorem is logically rigorous. Once we have a prior all our calculations have the certainty of deductive logic.

3. By trying different priors we can see how sensitive our results are to the choice of prior.

4. It is easy to communicate a result framed in terms of probabilities of hypotheses.

5. Even though the prior may be subjective, one can specify the assumptions used to arrive at it, which allows other people to challenge it or try other priors.

6. The evidence derived from the data is independent of notions about 'data more extreme' that depend on the exact experimental setup (see the "Stopping rules" section below).

7. Data can be used as it comes in. There is no requirement that every contingency be planned for ahead of time.

4.3 Critique of frequentist inference

1. It is ad-hoc and does not carry the force of deductive logic. Notions like 'data more extreme' are not well defined. The p-value depends on the exact experimental setup (see the "Stopping rules" section below).

2. Experiments must be fully specified ahead of time. This can lead to paradoxical seeming results. See the 'voltmeter story' in:
`https://en.wikipedia.org/wiki/Likelihood_principle`

3. The p-value and significance level are notoriously prone to misinterpretation. Careful statisticians know that a significance level of 0.05 means the probability of a type I error is 5%. That is, if the null hypothesis is true then 5% of the time it will be rejected due to randomness. Many (most) other people erroneously think a p-value of 0.05 means that the probability of the null hypothesis is 5%.

Strictly speaking you could argue that this is not a critique of frequentist inference but, rather, a critique of popular ignorance. Still, the subtlety of the ideas certainly contributes to the problem. (see "Mind your p's" below).

4.4 Defense of frequentist inference

1. It is objective: all statisticians will agree on the p-value. Any individual can then decide if the p-value warrants rejecting the null hypothesis.

2. Hypothesis testing using frequentist significance testing is applied in the statistical analysis of scientific investigations, evaluating the strength of evidence against a null hypothesis with data. The interpretation of the results is left to the user of the tests. Different users may apply different significance levels for determining statistical significance. Frequentist statistics does not pretend to provide a way to choose the significance level; rather it explicitly describes the trade-off between type I and type II errors.

3. Frequentist experimental design demands a careful description of the experiment and methods of analysis before starting. This helps control for experimenter bias.

4. The frequentist approach has been used for over 100 years and we have seen tremendous scientific progress. Although the frequentist themself would not put a probability on the belief that frequentist methods are valuable, shouldn't this history give the Bayesian a strong prior belief in the utility of frequentist methods?

5 Mind your p's.

We run a two-sample t-test for equal means, with $\alpha = 0.05$, and obtain a p-value of 0.04. What are the odds that the two samples are drawn from distributions with the same mean?

(a) 19/1 (b) 1/19 (c) 1/20 (d) 1/24 (e) unknown

Solution: (e) unknown. Frequentist methods only give probabilities of statistics conditioned on hypotheses. They do not give probabilities of hypotheses.

6 Stopping rules

When running a series of trials we need a rule on when to stop. Two common rules are:
1. Run exactly n trials and stop.
2. Run trials until you see a certain result and then stop.

In this example we'll consider two coin tossing experiments.
Experiment 1: Toss the coin exactly 6 times and report the number of heads.
Experiment 2: Toss the coin until the first tails and report the number of heads.

Jon is worried that his coin is biased towards heads, so before using it in class he tests it for fairness. He runs an experiment and reports to Jerry that his sequence of tosses was $HHHHHT$. But Jerry is only half-listening, and he forgets which experiment Jon ran to produce the data.

Frequentist approach.
Since he's forgotten which experiment Jon ran, Jerry the frequentist decides to compute the p-values for both experiments given Jon's data.

Let θ be the probability of heads. We have the null and one-sided alternative hypotheses

$$H_0 : \theta = 0.5, \qquad H_A : \theta > 0.5.$$

Experiment 1: The null distribution is binomial(6, 0.5) so, the one sided p-value is the probability of 5 or 6 heads in 6 tosses. Using R we get

$$p = \texttt{1 - pbinom(4, 6, 0.5)} = 0.1094.$$

Experiment 2: The null distribution is geometric(0.5) so, the one sided p-value is the probability of 5 or more heads before the first tails. Using R we get

$$p = \texttt{1 - pgeom(4, 0.5)} = 0.0313.$$

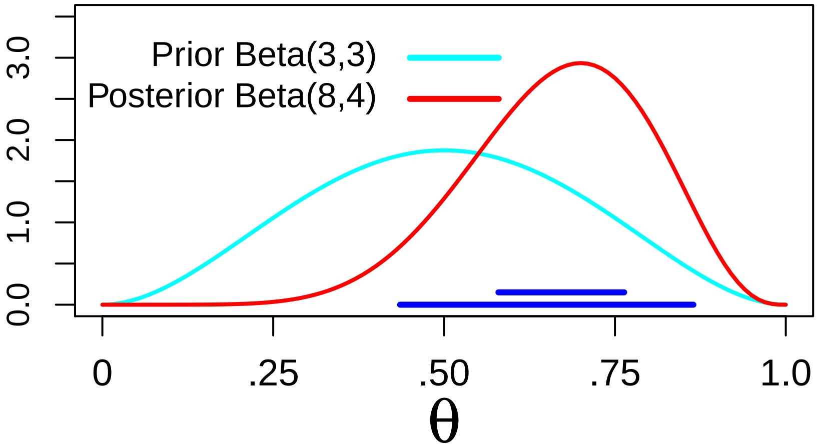
Using the typical significance level of 0.05, the same data leads to opposite conclusions! We would reject H_0 in experiment 2, but not in experiment 1.

The frequentist is fine with this. The set of possible outcomes is different for the different experiments so the notion of extreme data, and therefore p-value, is different. For example, in experiment 1 we would consider $THHHHH$ to be as extreme as $HHHHHT$. In experiment 2 we would never see $THHHHH$ since the experiment would end after the first tails.

Bayesian approach.
Jerry the Bayesian knows it doesn't matter which of the two experiments Jon ran, since the binomial and geometric likelihood functions (columns) for the data $HHHHHT$ are proportional. In either case, he must make up a prior, and he chooses Beta(3,3). This is a relatively flat prior concentrated over the interval $0.25 \le \theta \le 0.75$.

See https://mathlets.org/mathlets/beta-distribution/

Since the beta and binomial (or geometric) distributions form a conjugate pair the Bayesian update is simple. Data of 5 heads and 1 tails gives a posterior distribution Beta(8,4). Here is a graph of the prior and the posterior. The blue lines at the bottom are 50% and 90% probability intervals for the posterior.



Prior and posterior distributions with 0.5 and 0.9 probability intervals

Here are the relevant computations in R:

Posterior 50% probability interval: `qbeta(c(0.25, 0.75), 8, 4)` = [0.58 0.76]
Posterior 90% probability interval: `qbeta(c(0.05, 0.95), 8, 4)` = [0.44 0.86]
$P(\theta > 0.50 \,|\, \text{data})$ = `1- pbeta(0.5, posterior.a, posterior.b)` = 0.89

Starting from the prior Beta(3,3), the posterior probability that the coin is biased toward heads is 0.89.

7 Making decisions

Quite often the goal of statistical inference is to help with making a decision, e.g. whether or not to undergo surgery, how much to invest in a stock, whether or not to go to graduate school, etc.

In statistical decision theory, consequences of taking actions are measured by a utility function. The utility function assigns a weight to each possible outcome; in the language of probability, it is simply a random variable.

For example, in my investments I could assign a utility of d to the outcome of a gain of d dollars per share of a stock (if $d < 0$ my utility is negative). On the other hand, if my tolerance for risk is low, I will assign a more negative utility to losses than to gains (say, $-d^2$ if $d < 0$ and d if $d \geq 0$).

A decision rule combines the expected utility with evidence for each hypothesis given by the data (e.g., p-values or posterior distributions) into a formal statistical framework for making decisions.

In this setting, the frequentist will consider the expected utility given a hypothesis

$$E[U \mid \mathcal{H}]$$

where U is the random variable representing utility. There are frequentist methods for combining the expected utility with p-values of hypotheses to guide decisions.

The Bayesian can combine $E[U \mid \mathcal{H}]$ with the posterior (or prior if it's before data is collected) to create a Bayesian decision rule.

In either framework, two people considering the same investment may have different utility functions and make different decisions. For example, a riskier stock (with higher potential upside and downside) will be more appealing with respect to the first utility function above than with respect to the second (loss-averse) one.

A significant theoretical result is that for any decision rule there is a Bayesian decision rule which is, in a precise sense, at least as good a rule.

8 The likelihood principle

We briefly mention the likelihood principle. It can be stated succinctly as

> 'All of the evidence from data is contained in its likelihood function'

- Controversial
- Consistent with Bayesian updating.
  It only uses the column in the likelihood table that is for the data we actually saw.
- Inconsistent with NHST.
  Computing significance and p-values uses the entire likelihood table. That is, it relies on the probabilities of both observed and unobserved data (the full experimental design).

- See https://en.wikipedia.org/wiki/Likelihood_principle

Confidence intervals based on normal data
Class 22, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to determine whether an expression defines a valid interval statistic.
2. Be able to compute z and t confidence intervals for the mean given normal data.
3. Be able to compute the χ^2 confidence interval for the variance given normal data.
4. Be able to define the confidence level of a confidence interval.
5. Be able to explain the relationship between the z confidence interval (and confidence level) and the z non-rejection region (and significance level) in NHST.

2 Introduction

We continue to survey the tools of frequentist statistics. Suppose we have a model (probability distribution) for observed data with an unknown parameter. We have seen how NHST uses data to test the hypothesis that the unknown parameter has a particular value.

We have also seen how point estimates like the MLE use data to provide an estimate of the unknown parameter. On its own, a point estimate like $\bar{x} = 2.2$ carries no information about its accuracy; it's just a single number, regardless of whether its based on ten data points or one million data points.

For this reason, statisticians augment point estimates with confidence intervals. For example, to estimate an unknown mean μ we might be able to say that our best estimate of the mean is $\overline{x} = 2.2$ with a 95% confidence interval $[1.2, 3.2]$. Another way to describe the interval is: $\overline{x} \pm 1$.

We will leave to later the explanation of exactly what the 95% confidence level means. For now, we'll note that taken together the width of the interval and the confidence level provide a measure on the strength of the evidence supporting the hypothesis that the μ is close to our estimate $\overline{x}$. You should think of the confidence level of an interval as analogous to the significance level of a NHST. As explained below, it is no accident that we often see significance level $\alpha = 0.05$ and confidence level $0.95 = 1 - \alpha$.

We will first explore confidence intervals in situations where you will easily be able to compute by hand: z and t confidence intervals for the mean and χ^2 confidence intervals for the variance. We will use R to handle all the computations in more complicated cases. Indeed, the challenge with confidence intervals is not their computation, but rather interpreting them correctly and knowing how to use them in practice.

3 Interval statistics

Recall that our definition of a statistic is anything that can be computed from data. In particular, the **formula for a statistic cannot include unknown quantities**.

Example 1. Suppose $x_1, \ldots, x_n$ is drawn from $\mathrm{N}(\mu, \sigma^2)$ where μ and σ are unknown.

(i) $\overline{x}$ and $\overline{x} - 5$ are statistics.

(ii) $\overline{x} - \mu$ is **not** a statistic since μ is unknown.

(iii) If μ_0 a known value, then $\overline{x} - \mu_0$ is a statistic. This case arises when we consider the null hypothesis $\mu = \mu_0$. For example, if the null hypothesis is $\mu = 5$, then the statistic $\overline{x} - \mu_0$ is just $\overline{x} - 5$ from (i).

We can play the same game with intervals to define interval statistics

Example 2. Suppose $x_1, \ldots, x_n$ is drawn from $\mathrm{N}(\mu, \sigma^2)$ where μ is unknown.

(i) The interval $[\overline{x} - 2.2,\ \overline{x} + 2.2] = \overline{x} \pm 2.2$ is an interval statistic.

(ii) If σ is known, then $\left[\overline{x} - \frac{2\sigma}{\sqrt{n}},\ \overline{x} + \frac{2\sigma}{\sqrt{n}}\right]$ is an interval statistic.

(iii) On the other hand, if σ is unknown then $\left[\overline{x} - \frac{2\sigma}{\sqrt{n}},\ \overline{x} + \frac{2\sigma}{\sqrt{n}}\right]$ is **not** an interval statistic.

(iv) If s^2 is the sample variance, then $\left[\overline{x} - \frac{2s}{\sqrt{n}},\ \overline{x} + \frac{2s}{\sqrt{n}}\right]$ is an interval statistic because s^2 is computed from the data.

We will return to (ii) and (iv), as these are respectively the z and t confidence intervals for estimating μ.

Technically an interval statistic is nothing more than a pair of point statistics giving the lower and upper bounds of the interval. Our reason for emphasizing that the interval is a statistic is to highlight the following:

1. The interval is random – new random data will produce a new interval.

2. As frequentists, we are perfectly happy using it because it doesn't depend on the value of an unknown parameter or hypothesis.

3. As usual with frequentist statistics we have to assume a certain hypothesis, e.g. value of μ, before we can compute probabilities about the interval.

   Example 3. Suppose we draw n samples $x_1, \ldots, x_n$ from a $\mathrm{N}(\mu, 1)$ distribution, where μ is unknown. Suppose we wish to know the probability that 0 is in the interval $[\overline{x} - 2, \overline{x} + 2]$. Without knowing the value of μ this is impossible. However, we can compute this probability for any given (hypothesized) value of μ.

4. A warning which will be repeated: Be careful in your thinking about these probabilities. Confidence intervals are a frequentist notion. Since frequentists do not compute probabilities of hypotheses, the **confidence level is never a probability that the unknown parameter is in the specific confidence interval computed from the given data**.

4 z confidence intervals for the mean

Throughout this section we will assume that we have normally distributed data:

$$x_1,\, x_2,\, \ldots,\, x_n \sim \mathrm{N}(\mu, \sigma^2).$$

As we often do, we will introduce the main ideas through examples, building on what we know about rejection and non-rejection regions in NHST until we have constructed a confidence interval.

4.1 Definition of z confidence intervals for the mean

We start with z confidence intervals for the mean. First we'll give the formula. Then we'll walk through the derivation in one entirely numerical example. This will give us the basic idea. Then we'll repeat this example, replacing the explicit numbers by symbols. Finally we'll work through a computational example.

Definition: Suppose the data $x_1, \ldots, x_n \sim \mathrm{N}(\mu, \sigma^2)$, with unknown mean μ and known variance σ^2. The $(1-\alpha)$ confidence interval for μ is

$$\left[\overline{x} \;-\; \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}},\;\; \overline{x} \;+\; \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}\right], \tag{1}$$

where $z_{\alpha/2}$ is the right critical value $P(Z > z_{\alpha/2}) = \alpha/2$.

For example, if $\alpha = 0.05$ then $z_{\alpha/2} = 1.96$ so the 0.95 (or 95%) confidence interval is

$$\left[\overline{x} - \frac{1.96\sigma}{\sqrt{n}},\, \overline{x} + \frac{1.96\sigma}{\sqrt{n}}\right].$$

We've created an applet that generates normal data and displays the corresponding z confidence interval for the mean. It also shows the t-confidence interval, as discussed in the next section. Play around to get a sense for random intervals!

https://mathlets.org/mathlets/confidence-intervals/

Example 4. Suppose we collect 100 data points from a $\mathrm{N}(\mu, 3^2)$ distribution and the sample mean is $\overline{x} = 12$. Give the 95 % confidence interval for μ.

Solution: Using formula 1, this is trivial to compute: the 95% confidence interval for μ is

$$\left[\overline{x} - \frac{1.96\sigma}{\sqrt{n}},\, \overline{x} + \frac{1.96\sigma}{\sqrt{n}}\right] = \left[12 - \frac{1.96\cdot 3}{10},\, 12 + \frac{1.96\cdot 3}{10}\right]$$

4.2 Explaining the definition part 1: non-rejection regions

Our next goal is to explain the definition 1 starting from our knowledge of rejection/non-rejection regions. The phrase 'non-rejection region' is not pretty, but we will discipline ourselves to use it instead of the inacurate phrase 'acceptance region'.

Example 5. Suppose that $n = 12$ data points are drawn from $\mathrm{N}(\mu, 5^2)$ where μ is unknown. As usual, call the average of the data $\overline{x}$. Set up a two-sided z-test of $H_0 : \mu = 2.71$ at significance level $\alpha = 0.05$. Describe the rejection and non-rejection regions.

Solution: Under the null hypothesis ($\mu = 2.71$) we have

$$z = \frac{\overline{x} - 2.71}{5/\sqrt{12}} \sim \mathrm{N}(0,1)$$

We know that, for $\alpha = 0.05$, the non-rejection region for z is

$$[-1.96, 1.96].$$

That is, we do not reject if, assuming H_0, z is within two standard deviations of the standardized mean. By definition, this means

$$P(-1.96 \le z \le 1.96 \,|\, \mu = 2.71) = 0.95.$$

And, the rejection region is

$$(-\infty, -1.96) \cup (1.96, \infty).$$

For confidence intervals, we will want to unwind the definition of z and write the regions in terms of $\overline{x}$. This allows us to directly use the natural statistic $\overline{x}$.

Example 6. Redo the previous example using $\overline{x}$ as the test statistic.

Solution: Under the null hypothesis ($\mu = 2.71$) we have $x_i \sim \mathrm{N}(2.71, 5^2)$ and thus

$$\overline{x} \sim \mathrm{N}(2.71, 5^2/12)$$

where $5^2/12$ is the variance $\overline{x}$. We know that for normal data, significance $\alpha = 0.05$ corresponds to a rejection region starting 1.96 standard deviations from the hypothesized mean. That is,

Non-rejection region: We do not reject H_0 if $\overline{x}$ is in the interval

$$\left[2.71 \;-\; \frac{1.96 \cdot 5}{\sqrt{12}},\;\; 2.71 \;+\; \frac{1.96 \cdot 5}{\sqrt{12}}\right] \;=\; [-0.12,\; 5.54].$$

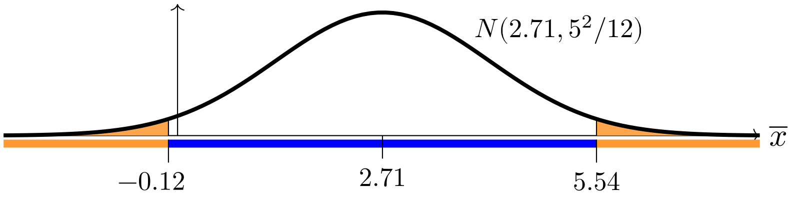
That is, we do not reject if, assuming H_0, $\overline{x}$ is within two standard deviations of the hypothesized mean. By definition, this means

$$P(-0.12 \le \overline{x} \le 5.54 \,|\, \mu = 2.71) = 0.95.$$

Rejection region:

$$\left(-\infty,\; 2.71 \;-\; \frac{1.96 \cdot 5}{\sqrt{12}}\right] \;\cup\; \left[2.71 \;+\; \frac{1.96 \cdot 5}{\sqrt{12}},\; \infty\right) \;=\; (-\infty,\; -0.12] \;\cup\; [5.54,\; \infty).$$

The following figure shows the rejection and non-rejection regions for $\overline{x}$. The regions represent ranges of $\overline{x}$ so they are represented by the colored bars on the $\overline{x}$ axis. The area of the shaded region in the tails is the significance level.



The rejection (orange) and non-rejection (blue) regions for $\overline{x}$.

Now, what about different data or null hypotheses. This is straight-forward, let's redo the previous example using symbols for all quantities.

Example 7. Suppose that n data points are drawn from $\mathrm{N}(\mu, \sigma^2)$ where μ is unknown and σ is known. Set up a two-sided significance test of $H_0 : \mu = \mu_0$ using the statistic $\overline{x}$ at significance level α. Describe the rejection and non-rejection regions.

Solution: Under the null hypothesis $\mu = \mu_0$ we have $x_i \sim \mathrm{N}(\mu_0, \sigma^2)$ and thus

$$\overline{x} \sim \mathrm{N}(\mu_0, \sigma^2/n),$$

where σ^2/n is the variance $(\sigma_{\overline{x}})^2$ of $\overline{x}$ and μ_0, σ and n are all known values.

Let $z_{\alpha/2}$ be the critical value: $P(Z > z_{\alpha/2}) = \alpha/2$. Then the non-rejection and rejection regions are separated by the values of $\overline{x}$ that are $z_{\alpha/2} \cdot \sigma_{\overline{x}}$ from the hypothesized mean. Since $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$ we have

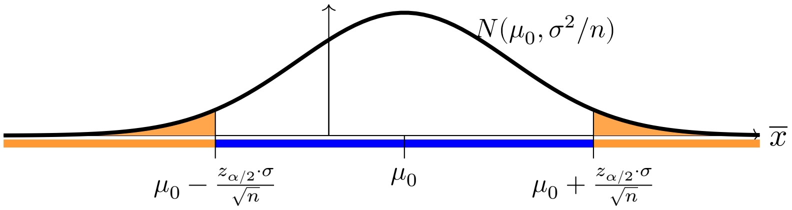
Non-rejection region: we do not reject H_0 if $\overline{x}$ is in the interval

$$\left[\mu_0 - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}},\ \mu_0 + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}\right] \tag{2}$$

Rejection region:

$$\left(-\infty,\ \mu_0 - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}\right] \cup \left[\mu_0 + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}},\ \infty\right).$$

We get the same figure as above, with the explicit numbers replaced by symbolic values.



The rejection (orange) and non-rejection (blue) regions for $\overline{x}$.

4.3 Manipulating intervals: algebraic pivoting

We need to get comfortable manipulating intervals. In general, we will make use of the type of 'obvious' statements that can be hard to get across. First is the notion of pivoting. Stripping away the statistical terms, pivoting is the following algebraic maneuver.

Example 8. Algebraic pivoting. Suppose we have two variables a and b. Suppose also that a is in the interval $[b-4, b+6]$. Show that b is in the interval $[a-6, a+4]$.

Solution: We are given, $b-4 \leq a \leq b+6$. Therefore,

$$-4 \leq a-b \leq 6 \;\; \Rightarrow \;\; 4 \geq b-a \geq -6 \;\; \Rightarrow \;\; a+4 \geq b \geq a-6. \quad \text{QED}$$

This is called pivoting because the roles of a and b are reversed along with the direction of the inequalities.

In the example above, the ranges on either side of b are different. Quite often they will be the same. Here are some simple numerical examples of pivoting for symmetric intervals.

Example 9. (i) 1.5 is in the interval $[0-2.3, 0+2.3]$, so 0 is in the interval $[1.5-2.3, 1.5+2.3]$

(ii) Likewise 1.5 is not in the interval $[0-1, 0+1]$, so 0 is not in the interval $[1.5-1, 1.5+1]$.

4.4 Pivoting non-rejection intervals to confidence intervals

For normal data, the non-rejection region for $\overline{x}$ is an interval centered on μ_0. By pivoting, we get the confidence interval for μ centered on $\overline{x}$.

Example 10. Suppose we have n data points with a sample mean $\overline{x}$ and hypothesized mean $\mu_0 = 2.71$. Suppose also that the null distribution is $x_i \sim \mathrm{N}(\mu_0, 3^2)$. Then with a significance level of 0.05 we have:

(1a) The non-rejection region is centered on $\mu_0 = 2.71$. That is, we don't reject H_0 if $\overline{x}$ is in the interval

$$\left[\mu_0 - \frac{1.96\sigma}{\sqrt{n}},\, \mu_0 + \frac{1.96\sigma}{\sqrt{n}}\right]$$

(1b) Assuming the null hypothesis we have

$$P(\overline{x} \text{ is in the non-rejection region} \,|\, H_0) = 1 - \alpha = 0.95.$$

That is,

$$P\left(\mu_0 - \frac{1.96\sigma}{\sqrt{n}} \;\leq\; \overline{x} \;\leq\; \mu_0 + \frac{1.96\sigma}{\sqrt{n}} \,|\, H_0\right) = 0.95$$

(2a) Pivoting (1a) gives: we don't reject H_0 if μ_0 is in the interval

$$\left[\overline{x} - \frac{1.96\sigma}{\sqrt{n}},\, \overline{x} + \frac{1.96\sigma}{\sqrt{n}}\right]$$

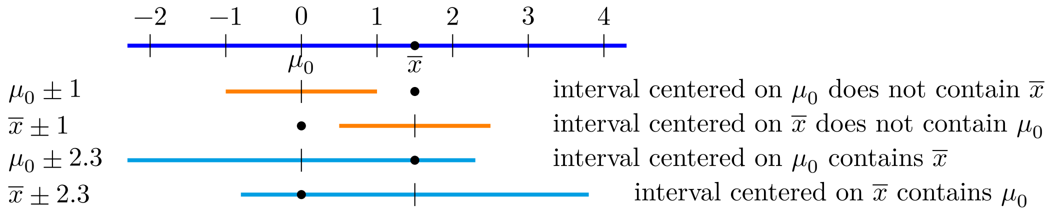
(2b) Pivoting (1b) gives: assuming the null hypothesis we have

$$P\left(\overline{x} - \frac{1.96\sigma}{\sqrt{n}} \;\leq\; \mu_0 \leq \overline{x} + \frac{1.96\sigma}{\sqrt{n}} \,|\, H_0\right) = 0.95$$

The interval in (2a) is called the 0.95 confidence interval for μ. It is centered on $\overline{x}$, it has the same width as the non-rejection region.

Again, notice the symmetry: the statement '$\overline{x}$ is in the non-rejection interval around μ_0' is equivalent to 'μ_0 is in the confidence interval $[\overline{x}-1.96\sigma, \overline{x}+1.96\sigma]$ around $\overline{x}$'.

Here is a visualization of *pivoting* from intervals around μ_0 to intervals around $\overline{x}$. In the figures, $\mu_0 = 1$ and $\overline{x} = 1.5$. The first pair of intervals have width 2 and the second pair have width 4.6.


-2
-1
0
1
2
3
4
μ_0
$\overline{x}$
$\mu_0 \pm 1$
interval centered on μ_0 does not contain $\overline{x}$
$\overline{x} \pm 1$
interval centered on $\overline{x}$ does not contain μ_0
$\mu_0 \pm 2.3$
interval centered on μ_0 contains $\overline{x}$
$\overline{x} \pm 2.3$
interval centered on $\overline{x}$ contains μ_0

The first pair of intervals shows the interval $\mu_0 \pm 1$ pivoted to the interval $\overline{x} \pm 1$. Since $\overline{x}$ is not in the first interval, μ_0 is not in the pivoted interval. In the second pair of intervals, since $\overline{x}$ is in the interval $\mu_0 \pm 2.3$, we see μ_0 is in the pivoted interval $\overline{x} \pm 2.3$.

4.5 Summary of normal confidence intervals: definition and properties

Suppose $x_1, x_2, \ldots, x_n$ are independent data from a $\mathrm{N}(\mu, \sigma^2)$ distribution. We assume μ is unknown, but σ is known.

- **Definition.** The $1-\alpha$ confidence interval for μ is
$$\left[\overline{x} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}},\ \overline{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right],$$
where $z_{\alpha/2}$ is standard normal $\alpha/2$ critical value.
- The confidence interval only depends on $\overline{x}$ and known values, so it is a statistic.
- The confidence interval is random: different data generate different intervals.
- If the null hypothesis is $\mu = \mu_0$, then the confidence interval is found by pivoting the non-rejection region. If μ_0 is in the $1-\alpha$ confidence interval, then we do not reject H_0 at significance level α. Likewise, we do reject H_0 at significance level α if μ_0 is *not* in the $1-\alpha$ confidence interval.
- Assuming H_0, then in 95% of random trials the 95% confidence interval will contain μ_0.

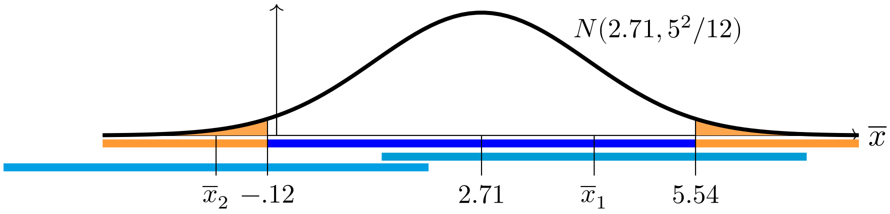
The following figure illustrates how we don't reject H_0 if the confidence interval around $\overline{x}$ contains μ_0 and we reject H_0 if the confidence interval doesn't contain μ_0. There is a lot in the figure so we will list carefully what you are seeing:

1. We started with the figure from Example 6 which shows the null distribution for $\mu_0 = 2.71$ and the rejection and non-rejection regions.

2. We added two possible values of the statistic $\overline{x}$, i.e. $\overline{x}_1$ and $\overline{x}_2$, and their confidence intervals. Note that the width of each interval is exactly the same as the width of the non-rejection region since both use $\pm\dfrac{1.96 \cdot 5}{\sqrt{12}}$.

The first value, $\overline{x}_1$, is in the non-rejection region and its interval includes the null hypothesis $\mu_0 = 2.71$. This illustrates that not rejecting H_0 corresponds to the confidence interval containing μ_0.

The second value, $\overline{x}_2$, is in the rejection region and its interval does not contain μ_0. This illustrates that rejecting H_0 corresponds to the confidence interval not containing μ_0.



The non-rejection region (blue) and two confidence intervals (light blue).

We can still wring one more essential observation out of this example. Our choice of null hypothesis $\mu = 2.71$ was completely arbitrary. If we replace $\mu = 2.71$ by any other hypothesis $\mu = \mu_0$ then the confidence interval is the same, i.e. it does not depend on any hypothesis.

4.6 Explaining the definition part 3: translating a general non-rejection region to a confidence interval

Note that the specific values of σ and n in the preceding example were of no particular consequence, so they can be replaced by their symbols. In this way we can take Example 7 quickly through the same steps as Example6.

In words, Equation 2 and the corresponding figure say that we don't reject if

$$\overline{x} \text{ is in the interval } \mu_0 \pm \frac{z_{\alpha/2}\sigma}{\sqrt{n}}.$$

This is exactly equivalent to saying that we don't reject if

$$\mu_0 \text{ is in the interval } \overline{x} \pm \frac{z_{\alpha/2}\sigma}{\sqrt{n}}. \qquad (3)$$

We can rewrite equation 3 as: at significance level α we don't reject if

$$\text{the interval } \left[\overline{x} - \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}},\ \ \overline{x} + \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}\right] \text{ contains } \mu_0. \qquad (4)$$

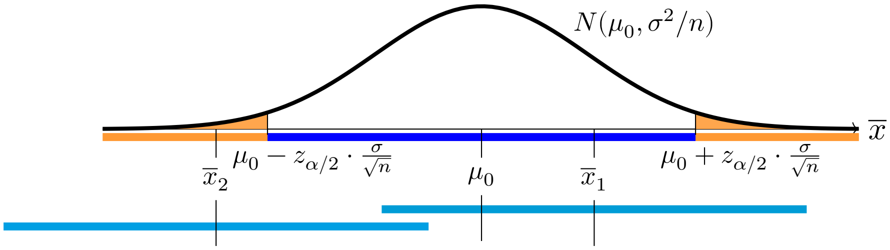
We call the interval 4 a $(1-\alpha)$ confidence interval because, assuming $\mu = \mu_0$, on average it will contain μ_0 in the fraction $(1-\alpha)$ of random trials.

The following figure illustrates the point that μ_0 is in the $(1-\alpha)$ confidence interval around $\overline{x}$ is equivalent to $\overline{x}$ is in the non-rejection region (at significance level α) for $H_0 : \mu_0 = \mu$.

The figure shows $\overline{x}_1$ is in the non-rejection region for μ_0, so the confidence interval around $\overline{x}_1$ contains μ_0.

Similarly, $\overline{x}_2$ is not in the non-rejection region for μ_0, so the confidence interval around $\overline{x}_2$ does not contain μ_0.

Note, that the confidence intervals and the non-rejection region all have the same width!


$N(\mu_0, \sigma^2/n)$
$\overline{x}$
$\overline{x}_2$
$\mu_0 - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
μ_0
$\overline{x}_1$
$\mu_0 + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

4.7 Computational example

Example 11. Suppose the data 2.5, 5.5, 8.5, 11.5 was drawn from a $\mathrm{N}(\mu, 10^2)$ distribution with unknown mean μ.

(a) Compute the point estimate $\overline{x}$ for μ and the corresponding 50%, 80% and 95% confidence intervals.

(b) Consider the null hypothesis $\mu = 1$. Would you reject H_0 at $\alpha = 0.05$? $\alpha = 0.20$? $\alpha = 0.50$? Do these two ways: first by checking if the hypothesized value of μ is in the relevant confidence interval and second by constructing a rejection region.

Solution: (a) We compute that $\overline{x} = 7.0$. The critical points are
$z_{0.025} = \texttt{qnorm(0.975)} = 1.96$, $z_{0.1} = \texttt{qnorm(0.9)} = 1.28$, $z_{0.25} = \texttt{qnorm(0.75)} = 0.67$.

Since $n = 4$ we have $\overline{x} \sim \mathrm{N}(\mu, 10^2/4)$, i.e. $\sigma_{\overline{x}} = 5$. So we have:

$$\begin{array}{lclclcl}
95\% \text{ conf. interval} & = & [\overline{x} - z_{0.025}\sigma_{\overline{x}}, \quad \overline{x} + z_{0.025}\sigma_{\overline{x}}] & = & [7 - 1.96 \cdot 5, \quad 7 + 1.96 \cdot 5] & = & [-2.8, \quad 16.8] \\
80\% \text{ conf. interval} & = & [\overline{x} - z_{0.1}\sigma_{\overline{x}}, \quad \overline{x} + z_{0.1}\sigma_{\overline{x}}] & = & [7 - 1.28 \cdot 5, \quad 7 + 1.28 \cdot 5] & = & [\ \ 0.6, \quad 13.4] \\
50\% \text{ conf. interval} & = & [\overline{x} - z_{0.75}\sigma_{\overline{x}}, \quad \overline{x} + z_{0.75}\sigma_{\overline{x}}] & = & [7 - 0.67 \cdot 5, \quad 7 + 0.67 \cdot 5] & = & [3.65, \quad 10.35]
\end{array}$$

Each of these intervals is a range estimate of μ. Notice that the higher the confidence level, the wider the interval needs to be.

(b) Since $\mu = 1$ is in the 95% and 80% confidence intervals, we would not reject the null hypothesis at the $\alpha = 0.05$ or $\alpha = 0.20$ levels. Since $\mu = 1$ is not in the 50% confidence interval, we would reject H_0 at the $\alpha = 0.5$ level.

We construct the rejection regions using the same critical values as in part (a). The difference is that rejection regions are intervals centered on the hypothesized value for μ: $\mu_0 = 1$ and confidence intervals are centered on $\overline{x}$. Here are the rejection regions.

$$\begin{array}{lclclclcl}
\alpha = 0.05 & \Rightarrow & (-\infty, \mu_0 - z_{0.025}\sigma_{\overline{x}}] & \cup & [\mu_0 + z_{0.025}\sigma_{\overline{x}}, \infty) & = & (-\infty, -8.8] & \cup & [10.8, \infty) \\
\alpha = 0.20 & \Rightarrow & (-\infty, \mu_0 - z_{0.1}\sigma_{\overline{x}}] & \cup & [\mu_0 + z_{0.1}\sigma_{\overline{x}}, \infty) & = & (-\infty, -5.4] & \cup & [7.4, \infty) \\
\alpha = 0.25 & \Rightarrow & (-\infty, \mu_0 - z_{0.25}\sigma_{\overline{x}}] & \cup & [\mu_0 + z_{0.25}\sigma_{\overline{x}}, \infty) & = & (-\infty, -2.35] & \cup & [4.35, \infty)
\end{array}$$

To to do the NHST we must check whether or not $\overline{x} = 7$ is in the rejection region.

$\alpha = 0.05$: $7 < 10.8$ is not in the rejection region.
We do not reject the hypothesis that $\mu = 1$ at a significance level of 0.05.
$\alpha = 0.2$: $7 < 7.4$ is not in the rejection region.
We do not reject the hypothesis that $\mu = 1$ at a significance level of 0.2.
$\alpha = 0.5$: $7 > 4.35$ is in the rejection region.
We reject the hypothesis that $\mu = 1$ at a significance level 0.5.

We get the same answers using either method.

5 t-confidence intervals for the mean

This will be nearly identical to normal confidence intervals. In this setting σ is not known, so we have to make the following replacements.

1. Use $s_{\overline{x}} = \dfrac{s}{\sqrt{n}}$ instead of $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$. Here s is the sample variance we used before in t-tests
2. Use t-critical values instead of z-critical values.

5.1 Definition of t-confidence intervals for the mean

Definition: Suppose that $x_1, \ldots, x_n \sim \mathrm{N}(\mu, \sigma^2)$, where the values of the mean μ and the standard deviation σ are both unknown. . The $(1 - \alpha)$ confidence interval for μ is

$$\left[\overline{x} - \frac{t_{\alpha/2} \cdot s}{\sqrt{n}}, \ \overline{x} + \frac{t_{\alpha/2} \cdot s}{\sqrt{n}}\right], \qquad (5)$$

here $t_{\alpha/2}$ is the right critical value $P(T > t_{\alpha/2}) = \alpha/2$ for $T \sim t(n-1)$ and s^2 is the sample variance of the data.

5.2 Construction of t confidence intervals

For t confidence intervals we repeat the construction of normal confidence intervals with σ replaced by its estimate s.

Suppose that n data points are drawn from $\mathrm{N}(\mu, \sigma^2)$ where μ and σ are unknown. We'll derive the t confidence interval following the same pattern as for the z confidence interval.

Under the null hypothesis $\mu = \mu_0$, we have $x_i \sim \mathrm{N}(\mu_0, \sigma^2)$. So the studentized mean follows a Student t distribution with $n - 1$ degrees of freedom:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \sim t(n-1).$$

Let $t_{\alpha/2}$ be the critical value: $P(T > t_{\alpha/2}) = \alpha/2$, where $T \sim t(n-1)$. We know from running one-sample t-tests that the non-rejection region is given by

$$|t| \leq t_{\alpha/2}$$

Using the definition of the t-statistic to write the rejection region in terms of $\overline{x}$ we get: at significance level α we don't reject if

$$\frac{|\overline{x}-\mu_0|}{s/\sqrt{n}} \le t_{\alpha/2} \quad \Leftrightarrow \quad |\overline{x}-\mu_0| \le t_{\alpha/2}\cdot\frac{s}{\sqrt{n}}.$$

Geometrically, the right hand side says that we don't reject if

$$\mu_0 \ \text{ is within } \ t_{\alpha/2}\cdot\frac{s}{\sqrt{n}} \text{ of } \overline{x}.$$

This is exactly equivalent to saying that we don't reject if

$$\text{the interval } \left[\overline{x} \ - \ \frac{t_{\alpha/2}\cdot s}{\sqrt{n}}, \ \ \overline{x} \ + \ \frac{t_{\alpha/2}\cdot s}{\sqrt{n}}\right] \ \text{ contains } \mu_0.$$

This interval is the confidence interval defined in 5.

Example 12. Suppose the data 2.5, 5.5, 8.5, 11.5 was drawn from a $\mathrm{N}(\mu,\sigma^2)$ distribution with μ and σ both unknown.

Give interval estimates for μ by finding the 95%, 80% and 50% confidence intervals.

Solution: By direct computation we have $\overline{x}=7$ and $s^2=15$. The critical points are $t_{0.025} = \texttt{qt(0.975)} = 3.18$, $t_{0.1} = \texttt{qt(0.9)} = 1.64$, and $t_{0.25} = \texttt{qt(0.75)} = 0.76$.

$$\begin{aligned}
\text{95\% conf. interval} &= \left[\overline{x}-t_{0.025}\cdot\frac{s}{\sqrt{n}}, \quad \overline{x}+t_{0.025}\cdot\frac{s}{\sqrt{n}}\right] &&= [0.84, \quad 13.16]\\
\text{80\% conf. interval} &= \left[\overline{x}-t_{0.1}\cdot\frac{s}{\sqrt{n}}, \quad \overline{x}+t_{0.1}\cdot\frac{s}{\sqrt{n}}\right] &&= [3.82, \quad 10.18]\\
\text{50\% conf. interval} &= \left[\overline{x}-t_{0.25}\cdot\frac{s}{\sqrt{n}}, \quad \overline{x}+t_{0.25}\cdot\frac{s}{\sqrt{n}}\right] &&= [5.53, \quad 8.47]
\end{aligned}$$

All of these confidence intervals give interval estimates for the value of μ. Again, notice that the higher the confidence level, the wider the corresponding interval.

6 Chi-square confidence intervals for the variance

We now turn to an interval estimate for the unknown variance.

Definition: Suppose the data $x_1,\dots,x_n$ is drawn from $\mathrm{N}(\mu,\sigma^2)$ with mean μ and standard deviation σ both unknown. The $(1-\alpha)$ confidence interval for the variance σ^2 is

$$\left[\frac{(n-1)s^2}{c_{\alpha/2}}, \ \ \frac{(n-1)s^2}{c_{1-\alpha/2}}\right]. \tag{6}$$

Here $c_{\alpha/2}$ is the right critical value $P(X^2 > c_{\alpha/2}) = \alpha/2$ for $X^2 \sim \chi^2(n-1)$ and s^2 is the sample variance of the data.

The derivation of this interval is nearly identical to that of the previous derivations, now starting from the chi-square test for variance. The basic fact we need is that, for data drawn from $\mathrm{N}(\mu,\sigma^2)$, the statistic

$$\frac{(n-1)s^2}{\sigma^2}$$

follows a chi-square distribution with $n-1$ degrees of freedom. So given the null hypothesis $H_0 : \sigma = \sigma_0$, the test statistic is $(n-1)s^2/\sigma_0^2$ and the non-rejection region at significance level α is

$$c_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma_0^2} < c_{\alpha/2}.$$

Pivoting algebra converts this to

$$\frac{(n-1)s^2}{c_{1-\alpha/2}} > \sigma_0^2 > \frac{(n-1)s^2}{c_{\alpha/2}}.$$

This says we don't reject if

$$\text{the interval } \left[\frac{(n-1)s^2}{c_{\alpha/2}}, \frac{(n-1)s^2}{c_{1-\alpha/2}}\right] \text{ contains } \sigma_0^2$$

This is our $(1-\alpha)$ confidence interval.

A difference from the z and t confidence intervals is that this chi-square confidence intervals are not exactly symmetric around the estimator s^2. The reason is that the chi-square distribution (with $n-1$ degrees of freedom) is not symmetric around its mean $n-1$.

We will continue our exploration of confidence intervals next class. In the meantime, truly the best way is to internalize the meaning of the confidence level is to experiment with the confidence interval applet:

https://mathlets.org/mathlets/confidence-intervals/

Confidence Intervals: Three Views
Class 23, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to find z, t and χ^2 confidence intervals using the corresponding standardized statistics.

2. Be able to use a hypothesis test to find a confidence interval for an unknown parameter.

3. Refuse to answer questions that ask, in essence, 'given a confidence interval what is the probability or odds that it contains the true value of the unknown parameter?'

2 Introduction

Our approach to confidence intervals in the previous reading was a combination of standardized statistics and hypothesis testing. Today we will consider each of these perspectives separately, as well as introduce a third formal viewpoint. Each provides its own insight.

1. **Standardized statistic.** Most confidence intervals are based on standardized statistics with known distributions like z, t or χ^2. This provides a straightforward way to construct and interpret confidence intervals as a point estimate plus or minus some error.

2. **Hypothesis testing.** Confidence intervals may also be constructed from hypothesis tests. In cases where we don't have a standardized statistic this method will still work. It agrees with the standardized statistic approach in cases where they both apply.

This view connects the notions of significance level α for hypothesis testing and confidence level $1 - \alpha$ for confidence intervals; we will see that in both cases α is the probability of making a 'type 1' error. This gives some insight into the use of the word confidence. This view also helps to emphasize the frequentist nature of confidence intervals.

3. **Formal.** The formal definition of confidence intervals is perfectly precise and general. In a mathematical sense it gives insight into the inner workings of confidence intervals. However, because it is so general it sometimes leads to confidence intervals without useful properties. We will not dwell on this approach. We offer it mainly for those who are interested.

3 Confidence intervals via standardized statistics

The strategy here is essentially the same as in the previous reading. Assuming normal data we have what we called standardized statistics like the standardized mean, Studentized mean, and standardized variance. These statistics have well known distributions which depend on hypothesized values of μ and σ. We then use algebra to produce confidence intervals for μ or σ.

Don't let the algebraic details distract you from the essentially simple idea underlying confidence intervals: we start with a standardized statistic (e.g., z, t or χ^2) and use some algebra to get an interval that depends only on the data and known parameters.

3.1 z-confidence intervals for the mean: normal data with known standard deviation

z-confidence intervals for the mean of normal data are based on the standardized mean, i.e. the z-statistic. We start with n independent normal samples

$$x_1, x_2, \ldots, x_n \sim \mathrm{N}(\mu, \sigma^2).$$

We assume that μ is the unknown parameter of interest and σ is known. Notationally, let's write the (unknown) true value of μ as μ_0

We know that the standardized mean is standard normal:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1).$$

For the standard normal critical value $z_{\alpha/2}$ we have: $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$. Thus,

$$P\left(-z_{\alpha/2} < \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \mid \mu = \mu_0\right) = 1 - \alpha$$

A little bit of algebra puts this in the form of an interval around μ:

$$P\left(\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_0\right) = 1 - \alpha$$

We can emphasize that the interval depends only on the statistic $\overline{x}$ and the known value σ by writing this as

$$P\left(\left[\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \ \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] \text{ contains } \mu \mid \mu = \mu_0\right) = 1 - \alpha.$$

This is the $(1 - \alpha)$ z-confidence interval for μ. We often write it using the shorthand

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Think of it as $\overline{x} \pm$ error.

Make sure you notice that the probabilities are conditioned on $\mu = \mu_0$. As with all frequentist statistics, we have to fix hypothesized values of the parameters in order to compute probabilities.

3.2 t-confidence intervals for the mean: normal data with unknown mean and standard deviation

t-confidence intervals for the mean of normal data are based on the Studentized mean, i.e. the t-statistic.

Again we have $x_1, x_2, \ldots, x_n \sim \mathrm{N}(\mu, \sigma^2)$, but now we assume both μ and σ are unknown. As we did above, let's write the (unknown) true value of μ as μ_0. We know that the Studentized mean follows a Student t distribution with $n-1$ degrees of freedom. That is,

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \sim t(n-1),$$

where s^2 is the sample variance.

Now all we have to do is replace the standardized mean by the Studentized mean and the same logic we used for z gives us the t-confidence interval: start with

$$P\left(-t_{\alpha/2} < \frac{\overline{x} - \mu}{s/\sqrt{n}} < t_{\alpha/2} \mid \mu = \mu_0\right) = 1 - \alpha.$$

A little bit of algebra isolates μ in the middle of an interval:

$$P\left(\overline{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \mid \mu = \mu_0\right) = 1 - \alpha$$

We can emphasize that the interval depends only on the statistics $\overline{x}$ and s by writing this as

$$P\left(\left[\overline{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}},\ \overline{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right] \text{ contains } \mu \mid \mu = \mu_0\right) = 1 - \alpha.$$

This is the $(1-\alpha)$ t-confidence interval for μ. We often write it using the shorthand

$$\overline{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Think of it as $\overline{x} \pm$ error.

3.3 Chi-square confidence intervals for variance: normal data with unknown mean and standard deviation

You guessed it: χ^2-confidence intervals for the variance of normal data are based on the standardized variance, i.e. the χ^2-statistic.

We follow the same logic as above to get a χ^2-confidence interval for σ^2. Because this is the third time through it we'll move a little more quickly.

We assume we have n independent normal samples: $x_1, x_2, \ldots, x_n \sim \mathrm{N}(\mu, \sigma^2)$. We assume that μ and σ are both unknown and write the (unknown) true value of σ as σ_0. The standardized variance is

$$X^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(n-1).$$

We know that the X^2 statistic follows a χ^2 distribution with $n-1$ degrees of freedom.

For Z and t we used, without comment, the symmetry of the distributions to replace $z_{1-\alpha/2}$ by $-z_{\alpha/2}$ and $t_{1-\alpha/2}$ by $-t_{\alpha/2}$. Because the χ^2 distribution is not symmetric we need to be explicit about the critical values on both the left and the right. That is,

$$P(c_{1-\alpha/2} < X^2 < c_{\alpha/2}) = 1 - \alpha,$$

where $c_{\alpha/2}$ and $c_{1-\alpha/2}$ are right tail critical values. Thus,

$$P\left(c_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < c_{\alpha/2} \mid \sigma = \sigma_0\right) = 1-\alpha$$

A little bit of algebra puts this in the form of an interval around σ^2:

$$P\left(\frac{(n-1)s^2}{c_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{c_{1-\alpha/2}} \mid \sigma = \sigma_0\right) = 1-\alpha$$

We can emphasize that the interval depends only on the statistic s^2 by writing this as

$$P\left(\left[\frac{(n-1)s^2}{c_{\alpha/2}},\ \frac{(n-1)s^2}{c_{1-\alpha/2}}\right] \text{ contains } \sigma^2 \mid \sigma = \sigma_0\right) = 1-\alpha.$$

This is the $(1-\alpha)$ χ^2-confidence interval for σ^2.

4 Confidence intervals via hypothesis testing

Suppose we have data drawn from a distribution with a parameter θ whose value is unknown. A significance test for the value θ has the following short description.

1. Set the null hypothesis $H_0 : \theta = \theta_0$ for some special value θ_0, e.g. we often have $H_0 : \theta = 0$.

2. Use the data to compute the value of a test statistic, call it x.

3. If x is far enough into the tail of the null distribution (the distribution assuming the null hypothesis) then we reject H_0.

In the case where there is no special value to test we may still want to estimate θ. This is the reverse of significance testing; rather than seeing if we should reject a specific value of θ because it doesn't fit the data we want to find the range of values of θ that do, in some sense, fit the data. This gives us the following definitions.

Definition. Given a value x of the test statistic, the $(1-\alpha)$ confidence interval contains all values θ_0 which are not rejected (at significance level α) when they are the null hypothesis.

Definition. A type 1 CI error occurs when the confidence interval does not contain the true value of θ.

For a $(1-\alpha)$ confidence interval the type 1 CI error rate is α.

Example 1. Here is an example relating confidence intervals and hypothesis tests. Suppose data x is drawn from a binomial(12, θ) distribution with θ unknown. Let $\alpha = 0.1$ and create the $(1-\alpha) = 90\%$ confidence interval for each possible value of x.

Solution: Our strategy is to look at one possible value of θ at a time and choose rejection regions for a significance test with $\alpha = 0.1$. Once this is done, we will know, for each value of x, which values of θ are not rejected, i.e. the confidence interval associated with x.

To start we set up a likelihood table for binomial(12, θ) in Table 1. Each row shows the probabilities $p(x|\theta)$ for one value of θ. To keep the size manageable we only show θ in increments of 0.1.

$\theta\backslash x$	0	1	2	3	4	5	6	7	8	9	10	11	12
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.23	0.38	0.28
0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.13	0.24	0.28	0.21	0.07
0.7	0.00	0.00	0.00	0.00	0.01	0.03	0.08	0.16	0.23	0.24	0.17	0.07	0.01
0.6	0.00	0.00	0.00	0.01	0.04	0.10	0.18	0.23	0.21	0.14	0.06	0.02	0.00
0.5	0.00	0.00	0.02	0.05	0.12	0.19	0.23	0.19	0.12	0.05	0.02	0.00	0.00
0.4	0.00	0.02	0.06	0.14	0.21	0.23	0.18	0.10	0.04	0.01	0.00	0.00	0.00
0.3	0.01	0.07	0.17	0.24	0.23	0.16	0.08	0.03	0.01	0.00	0.00	0.00	0.00
0.2	0.07	0.21	0.28	0.24	0.13	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0.28	0.38	0.23	0.09	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 1. Likelihood table for Binomial(12, θ)

Tables 2-4 below show the rejection region (in orange) and non-rejection region (in blue) for the various values of θ. To emphasize the row-by-row nature of the process the Table 2 just shows these regions for $\theta = 1.0$, then Table 3 adds in regions for $\theta = 0.9$ and Table 4 shows them for all the values of θ.

Immediately following the tables we give a detailed explanation of how the rejection/non-rejection regions were chosen.

$\theta\backslash x$	0	1	2	3	4	5	6	7	8	9	10	11	12	significance
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.000
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.23	0.38	0.28	
0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.13	0.24	0.28	0.21	0.07	
0.7	0.00	0.00	0.00	0.00	0.01	0.03	0.08	0.16	0.23	0.24	0.17	0.07	0.01	
0.6	0.00	0.00	0.00	0.01	0.04	0.10	0.18	0.23	0.21	0.14	0.06	0.02	0.00	
0.5	0.00	0.00	0.02	0.05	0.12	0.19	0.23	0.19	0.12	0.05	0.02	0.00	0.00	
0.4	0.00	0.02	0.06	0.14	0.21	0.23	0.18	0.10	0.04	0.01	0.00	0.00	0.00	
0.3	0.01	0.07	0.17	0.24	0.23	0.16	0.08	0.03	0.01	0.00	0.00	0.00	0.00	
0.2	0.07	0.21	0.28	0.24	0.13	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	
0.1	0.28	0.38	0.23	0.09	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
0.0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Table 2. Likelihood table for binomial(12, θ) with rejection (orange)/non-rejection (blue) regions for $\theta = 1.0$

$\theta\backslash x$	0	1	2	3	4	5	6	7	8	9	10	11	12	significance
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.000
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.23	0.38	0.28	0.026
0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.13	0.24	0.28	0.21	0.07	
0.7	0.00	0.00	0.00	0.00	0.01	0.03	0.08	0.16	0.23	0.24	0.17	0.07	0.01	
0.6	0.00	0.00	0.00	0.01	0.04	0.10	0.18	0.23	0.21	0.14	0.06	0.02	0.00	
0.5	0.00	0.00	0.02	0.05	0.12	0.19	0.23	0.19	0.12	0.05	0.02	0.00	0.00	
0.4	0.00	0.02	0.06	0.14	0.21	0.23	0.18	0.10	0.04	0.01	0.00	0.00	0.00	
0.3	0.01	0.07	0.17	0.24	0.23	0.16	0.08	0.03	0.01	0.00	0.00	0.00	0.00	
0.2	0.07	0.21	0.28	0.24	0.13	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	
0.1	0.28	0.38	0.23	0.09	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
0.0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Table 3. Likelihood table: rejection (orange)/non-rejection (blue) regions for $\theta = 1.0$ and 0.9

$\theta\backslash x$	0	1	2	3	4	5	6	7	8	9	10	11	12	significance
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.000
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.23	0.38	0.28	0.026
0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	**0.13**	0.24	0.28	0.21	0.07	0.073
0.7	0.00	0.00	0.00	0.00	0.01	0.03	0.08	0.16	**0.23**	0.24	0.17	0.07	0.01	0.052
0.6	0.00	0.00	0.00	0.01	0.04	0.10	0.18	0.23	**0.21**	0.14	0.06	0.02	0.00	0.077
0.5	0.00	0.00	0.02	0.05	0.12	0.19	0.23	0.19	**0.12**	0.05	0.02	0.00	0.00	0.092
0.4	0.00	0.02	0.06	0.14	0.21	0.23	0.18	0.10	0.04	0.01	0.00	0.00	0.00	0.077
0.3	0.01	0.07	0.17	0.24	0.23	0.16	0.08	0.03	0.01	0.00	0.00	0.00	0.00	0.052
0.2	0.07	0.21	0.28	0.24	0.13	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.073
0.1	0.28	0.38	0.23	0.09	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.026
0.0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.000

Table 4. Likelihood table: rejection (orange)/non-rejection (blue) regions for $\theta = 0.0$ to 1.0

Choosing the rejection and non-rejection regions in the tables

The first problem we confront is how exactly to choose the rejection region. We used two rules:

1. The total probabilitiy of the rejection region, i.e. the significance, should be less than or equal to 0.1. (Since we have a discrete distribution it is impossible to make the significance exactly 0.1.)

2. We build the rejection region by choosing values of x one at a time, always picking the unused value with the smallest probability. We stop when the next value would make the significance more that 0.1.

There are other ways to choose the rejection region which would result in slight differences. Our method is one reasonable way.

Table 2 shows the rejection (orange) and non-rejection (blue) regions for $\theta = 1.0$. This is a special case because most of the probabilities in this row are 0.0. We'll move right on to the next table and step through the process for that.

In Table 3, let's walk through the steps used to find these regions for $\theta = 0.9$.

- The smallest probability is when $x = 0$, so $x = 0$ is in the rejection region.
- The next smallest is when $x = 1$, so $x = 1$ is in the rejection region.
- We continue with $x = 2, \dots, 8$. At this point the total probability in the rejection region is 0.026.
- The next smallest probability is when $x = 9$. Adding this probability (0.09) to 0.026 would put the total probability over 0.1. So we leave $x = 9$ out of the rejection region and stop the process.

Note three things for the $\theta = 0.9$ row:

1. None of the probabilities in this row are truly zero, though some are small enough that they equal 0 to 2 decimal places.

2. We show the significance for this value of θ in the right hand margin. More precisely, we show the significance level of the NHST with null hypothesis $\theta = 0.9$ and the given rejection region.

3. The rejection region consists of values of x. When we say the rejection region is shown in orange we really mean the rejection region contains the values of x corresponding to the probabilities highlighted in orange.

Think: Look back at the $\theta = 1.0$ row and make sure you understand why the rejection region is $x = 0, \dots, 11$ and the significance is 0.000.

Example 2. Using Table 4 determine the 0.90 confidence interval when $x = 8$.

Solution: The 90% confidence interval consists of all those θ that would not be rejected by an $\alpha = 0.1$ hypothesis test when $x = 8$. Looking at the table, the blue (non-rejected) entries in the column $x = 8$ correspond to $0.5 \leq \theta \leq 0.8$: the confidence interval is $[0.5, 0.8]$.

Remark: The point of this example is to show how confidence intervals and hypothesis tests are related. Since Table 4 has only finitely many values of θ, our answer is close but not exact. Using a computer we could look at many more values of θ. For this problem we used R to find that, correct to 2 decimal places, the confidence interval is $[0.42, 0.85]$.

Example 3. Explain why the expected type one CI error rate will be at most 0.092, provided that the true value of θ is in the table.

Solution: The short answer is that this is the maximum significance for any θ in Table 4. Expanding on that slightly: we make a type one CI error if the confidence interval does not contain the true value of θ, call it θ_{true}. This happens exactly when the data x is in the rejection region for θ_{true}. The probability of this happening is the significance for θ_{true} and this is at most 0.092.

Remark: The point of this example is to show how confidence level, type one CI error rate and significance for each hypothesis are related. As in the previous example, we can use R to compute the significance for many more values of θ. When we do this we find that the maximum significance for any θ is 0.1 ocurring when $\theta \approx 0.0452$.

Summary notes:

1. We start with a test statistic x. The confidence interval is random because it depends on x.

2. For each hypothesized value of θ we make a significance test with significance level α by choosing rejection regions.

3. For a specific value of x the associated confidence interval for θ consists of all θ that aren't rejected for that value, i.e. all θ that have x in their non-rejection regions.

4. Because the distribution is discrete we can't always achieve the exact significance level, so our confidence interval is really an 'at least 90% confidence interval'.

Example 4. Open the applet `https://mathlets.org/mathlets/confidence-intervals/`. We want you to play with the applet to understand the random nature of confidence intervals and the meaning of confidence as (1 - type I CI error rate).

(a) Read the help. It is short and will help orient you in the applet. Play with different settings of the parameters to see how they affect the size of the confidence intervals.

(b) Set the number of trials to $N = 1$. Click the 'Run N trials' button repeatedly and see that each time data is generated the confidence intervals jump around.

(c) Now set the confidence level to $c = 0.5$. As you click the 'Run N trials' button you should see that about 50% of the confidence intervals include the true value of μ. The 'Z correct' and 't correct' values should change accordingly.

(d) Now set the number of trials to $N = 100$. With $c = 0.8$. The 'Run N trials' button will now run 100 trials at a time. Only the last confidence interval will be shown in the graph, but the trials all run and the 'percent correct' statistics will be updated based on all 100 trials.

Click the run trials button repeatedly. Watch the correct rates start to converge to the confidence level. To converge even faster, set $N = 1000$.

5 Formal view of confidence intervals

Recall: An interval statistic is an interval I_x computed from data x. An interval is determined by its lower and upper bounds, and these are random because x is random.

We suppose that x is drawn from a distribution with pdf $f(x|\theta)$ where the parameter θ is unknown.

Definition: A $(1-\alpha)$ confidence interval for θ is an interval statistic I_x such that

$$P(I_x \text{ contains } \theta_0 \mid \theta = \theta_0) \;=\; 1-\alpha$$

for all possible values of θ_0.

We wish this was simpler, but a definition is a definition and this definition is one way to weigh the evidence provided by the data x. Let's unpack it a bit.

The confidence level of an interval statistic is a probability concerning a random interval and a hypothesized value θ_0 for the unknown parameter. Precisely, it is the probability that the random interval I_x (computed from random data x) contains the value θ_0, given that the model parameter truly is θ_0. Since the true value of θ is unknown, the frequentist statistician defines 95% confidence intervals so that the 0.95 probability is valid no matter which hypothesized value of the parameter is actually true.

6 Comparison with Bayesian probability intervals

Confidence intervals are a frequentist notion, and as we've repeated many times, frequentists don't assign probabilities to hypotheses, e.g., to the value of an unknown parameter. Rather they compute likelihoods; that is, probabilities about data or associated statistics given a hypothesis (note the condition $\theta = \theta_0$ in the formal view of confidence intervals). Note that the construction of confidence intervals proceeds entirely from the full likelihood table.

In contrast Bayesian posterior probability intervals are truly the probability that the value of the unknown parameter lies in the reported range. We add the usual caveat that this depends on the specific choice of a (possibly subjective) Bayesian prior.

This distinction between the two is subtle because Bayesian posterior probability intervals and frequentist confidence intervals share the following properties:

1. They start from a model $f(x|\theta)$ for observed data x with unknown parameter θ.

2. Given data x, they give an interval $I(x)$ specifying a range of values for θ.

3. They come with a number (say 0.95) that is the probability of something.

In practice, many people misinterpret confidence intervals as Bayesian probability intervals, forgetting that frequentists **never** place probabilities on hypotheses (this is analogous to mistaking the p-value in NHST for the probability that H_0 is false). The next section explores this mistake in some detail. The harm of this misinterpretation is somewhat mitigated by that fact that, given enough data and a reasonable prior, Bayesian and frequentist intervals often work out to be quite similar.

For an amusing example illustrating how they can be quite different, see the first answer in the link just below (involving chocolate chip cookies!). This example uses the formal definitions and is really about confidence sets instead of confidence intervals.

https://stats.stackexchange.com/questions/2272/whats-the-difference-between-a-confidence-interval-and-a-credible-interval

7 Misinterpreting confidence intervals

It is very tempting to think that given a 95% confidence interval for, say, the mean, the probability that the true mean is in the confidence interval is 95%.

We know this can't be true because the value of the mean can only be hypothesized and Frequentists don't assign probabilities to hypotheses. To be more concrete, if the mean is θ and the confidence interval is $[45, 55]$ then the statement $45 \leq \theta \leq 55$ is a hypothesis, so asking for the probability that $45 \leq \theta \leq 55$, is asking for the probability of a hypothesis.

The mistake is subtle and hard to wrap your mind around. It boils down to a question of what is being randomly sampled. Here is an attempt to explain the issue.

First, consider a test for a disease. Assume a person is given the test. Let T^+ be a positive test and D^+ be that they have the disease. Assume the test is 95% accurate, i.e. $P(T^+|D^+) = 0.95$. We know (base rate fallacy) that this does not imply that $P(D^+|T^+) = 0.95$.

Let's look at this from a different angle: Implicit in $P(T^+|D^+) = 0.95$ is the following experiment: Draw a random person from the set of all people with the disease and give them the test. Then 95% will test positive. That is, the population sampled is all people with the disease and the event considered is that the chosen person in that population tests positive.

For $P(D^+|T^+)$, the experiment is to draw a random person from the set of all people who tested positive. The probability is the fraction who have the disease. That is, the population sampled is all people who test positive and the event considered is that the chosen person in that population has the disease.

We can't expect $P(T^+|D^+)$ and $P(D^+|T^+)$ to be the same, since we're sampling from different populations and looking at different events. The probability $P(D^+|T^+)$ can be computed using Bayes' theorem from the (prior) probability $P(D^+)$ and the likelihoods $P(T^+|D+)$, $P(T^+|D^-)$.

Confidence intervals are a little more abstract, but the analysis is similar. Just as in testing for a disease, the populations sampled will be different. One source of difficulty is that the events are essentially the same in both cases.

Let's assume we have a distribution with unknown mean θ_0. We generate some data and compute a 95% confidence interval for the mean. The value of 95% comes from the following implied experiment: imagine having run many trials and created a confidence interval for each one. Then, 95% of confidence intervals contain the true mean. In notation,

$$P(\text{random interval contains } \theta_0) = 0.95.$$

That is, the random sample is drawn from the set of all confidence intervals generated by our trials. The event in question is that the chosen interval contains the true mean.

What if we run one experiment and generate the 95% confidence interval, call it I. To a Frequentist, the true mean is not random and we have a fixed interval. So, to the Frequentist, it makes no sense to ask about the probability θ_0 is in I.

To a Bayesian, it is fine to consider θ_0 as randomly drawn from a probability distribution –they often interpret it as a description of the uncertainty of our knowledge. So, they can ask for the probability

$$P(\text{random } \theta_0 \text{ is in a given } I).$$

So, here, the random sample is drawn from the set of possible means and the event considered is that the chosen mean is contained in the given interval.

As in the disease testing example, what population is being randomly sampled is different in the two cases, i.e in the first we have a random interval, in the second we have a random value for the true mean. As noted above, in both cases the event is that the true mean is in the interval.

We finish by noting that $P(\text{random } \theta_0 \text{ is in } I)$ can be computed using Bayes' theorem and depends on the prior distribution for the true mean and the likelihoods that each mean will generate the given confidence interval. The formula is a little unwieldy. Here it is.

- Call the interval I and the true mean θ_0.
- Call the data I. This is a shorthand for the data that the interval I is based on.
- Let $p(\theta)$ be the prior probability that $\theta_0 = \theta$.
- The likelihood $f(I|\theta)$ is the probability (or density) that the experiment would produce the interval I given $\theta_0 = \theta$.
- Let $p(\theta|I)$ be the posterior probability that $\theta_0 = \theta$. This is the updated probability found using the Bayes' theorem.

Bayes' theorem gives us

$$p(\theta|I) = \frac{f(I|\theta)p(\theta)}{f(I)}, \text{ where } f(I) = \sum_{\theta} f(I|\theta)p(\theta).$$

So we have, $P(\theta_0 \in I|I) = \sum_{\theta \text{ in } I} p(\theta|I)$. (As usual, if θ has a continuous range of values, then the sums will be replaced by integrals.)

Confidence Intervals for the Mean of Non-normal Data
Class 23, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to derive the formula for conservative normal confidence intervals for the proportion θ in Bernoulli data.
2. Be able to find rule-of-thumb 95% confidence intervals for the proportion θ of a Bernoulli distribution.
3. Be able to find large sample confidence intervals for the mean of a general distribution.

2 Introduction

So far, we have focused on constructing confidence intervals for data drawn from a normal distribution. We'll now switch gears and learn about confidence intervals for the mean when the data is not necessarily normal.

We will first look carefully at estimating the probability θ of success when the data is drawn from a Bernoulli(θ) distribution – recall that θ is also the mean of the Bernoulli distribution.

Then we will consider the case of a large sample from an unknown distribution. In this case we can appeal to the central limit theorem to justify the use z-confidence intervals.

3 Bernoulli data and polling

One common use of confidence intervals is for estimating the proportion θ in a Bernoulli(θ) distribution. For example, suppose we want to use a political poll to estimate the proportion of the population that supports candidate A, or equivalent the probability θ that a random person supports candidate A. In this case we have a simple rule-of-thumb that allows us to quickly compute a confidence interval.

3.1 Conservative normal confidence intervals

Suppose we have i.i.d. data $x_1, x_2, \dots, x_n$ all drawn from a Bernoulli(θ) distribution. then a conservative normal $(1-\alpha)$ confidence interval for θ is given by

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{1}{2\sqrt{n}}. \tag{1}$$

The proof given below uses the central limit theorem and the observation that $\sigma = \sqrt{\theta(1-\theta)} \leq 1/2$.

You will also see in the derivation below that this formula is conservative, providing an 'at least $(1-\alpha)$' confidence interval.

Example 1. A pollster asks 196 people if they prefer candidate A to candidate B and finds that 120 prefer A and 76 prefer B. Find the 95% conservative normal confidence interval for θ, the proportion of the population that prefers A.

Solution: We have $\overline{x} = 120/196 = 0.612$, $\alpha = 0.05$ and $z_{0.025} = 1.96$. The formula says a 95% confidence interval is

$$I \approx 0.612 \pm \frac{1.96}{2 \cdot 14} = 0.612 \pm 0.007.$$

3.2 Proof of Formula 1

The proof of Formula 1 will rely on the following fact.

Fact. The standard deviation of a Bernoulli(θ) distribution is at most 0.5.

Proof of fact: Let's denote this standard deviation by σ_θ to emphasize its dependence on θ. The variance is then $\sigma_\theta^2 = \theta(1-\theta)$. It's easy to see using calculus or by graphing this parabola that the maximum occurs when $\theta = 1/2$. Therefore the maximum variance is $1/4$, which implies that the standard deviation σ_p is less the $\sqrt{1/4} = 1/2$.

Proof of formula (1). The proof relies on the central limit theorem which says that (for large n) the distribution of $\overline{x}$ is approximately normal with mean θ and standard deviation $\sigma_\theta/\sqrt{n}$. For normal data we have the $(1-\alpha)$ z-confidence interval

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma_\theta}{\sqrt{n}}$$

The trick now is to replace σ_θ by $\frac{1}{2}$: since $\sigma_\theta \le \frac{1}{2}$ the resulting interval around $\overline{x}$

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{1}{2\sqrt{n}}$$

is always at least as wide as the interval using $\pm\sigma_\theta/\sqrt{n}$. A wider interval is more likely to contain the true value of θ so we have a 'conservative' $(1-\alpha)$ confidence interval for θ.

Again, we call this conservative because $\frac{1}{2\sqrt{n}}$ overestimates the standard deviation of $\bar{x}$, resulting in a wider interval than is necessary to achieve a $(1-\alpha)$ confidence level.

3.3 How political polls are reported

Political polls are often reported as a value with a margin-of-error. For example you might hear

52% favor candidate A with a margin-of-error of ±5%.

The actual precise meaning of this is

> if θ is the proportion of the population that supports A then the point estimate for θ is 52% and the 95% confidence interval is 52% ± 5%.

Notice that reporters of polls in the news do not mention the 95% confidence. You just have to know that that's what pollsters do.

The 95% rule-of-thumb confidence interval.
Recall that the $(1-\alpha)$ conservative normal confidence interval is

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{1}{2\sqrt{n}}.$$

If we use the standard approximation $z_{0.025} = 2$ (instead of 1.96) we get the rule-of thumb 95% confidence interval for θ:

$$\overline{x} \pm \frac{1}{\sqrt{n}}.$$

Example 2. Polling. Suppose there will soon be a local election between candidate A and candidate B. Suppose that the fraction of the voting population that supports A is θ.

Two polling organizations ask voters who they prefer.

1. The firm of *Fast and First* polls 40 random voters and finds 22 support A.
2. The firm of *Quick but Cautious* polls 400 random voters and finds 190 support A.

Find the point estimates and 95% rule-of-thumb confidence intervals for each poll. Explain how the statistics reflect the intuition that the poll of 400 voters is more accurate.

Solution: For poll 1 we have

Point estimate: $\overline{x} = 22/40 = 0.55$

Confidence interval: $\overline{x} \pm \dfrac{1}{\sqrt{n}} = 0.55 \pm \dfrac{1}{\sqrt{40}} = 0.55 \pm 0.16 = 55\% \pm 16\%.$

For poll 2 we have

Point estimate: $\overline{x} = 190/400 = 0.475$

Confidence interval: $\overline{x} \pm \dfrac{1}{\sqrt{n}} = 0.475 \pm \dfrac{1}{\sqrt{400}} = 0.475 \pm 0.05 = 47.5\% \pm 5\%.$

The greater accuracy of the poll of 400 voters is reflected in the smaller margin of error, i.e. 5% for the poll of 400 voters vs. 16% for the poll of 40 voters.

Other binomial proportion confidence intervals
There are many methods of producing confidence intervals for the proportion p of a binomial(n, p) distribution. For a number of other common approaches, see:

https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval

4 Large sample confidence intervals

One typical goal in statistics is to estimate the mean of a distribution. When the data follows a normal distribution we could use confidence intervals based on standardized statistics to estimate the mean.

But suppose the data $x_1, x_2, \ldots, x_n$ is drawn from a distribution with pmf or pdf $f(x)$ that may not be normal or even parametric. If the distribution has finite mean and variance and if n is sufficiently large, then the following version of the central limit theorem shows we can still use a standardized statistic.

Central Limit Theorem: For large n, the sampling distribution of the studentized mean is approximately standard normal: $\dfrac{\bar{x}-\mu}{s/\sqrt{n}} \approx \mathrm{N}(0,1).$

So for large n the $(1-\alpha)$ confidence interval for μ is approximately

$$\left[\bar{x} - \frac{s}{\sqrt{n}} \cdot z_{\alpha/2},\ \ \bar{x} + \frac{s}{\sqrt{n}} \cdot z_{\alpha/2}\right]$$

where $z_{\alpha/2}$ is the $\alpha/2$ critical value for $\text{N}(0,1)$. This is called the large sample confidence interval.

Example 3. How large must n be?

Recall that a type 1 CI error occurs when the confidence interval does not contain the true value of the parameter, in this case the mean. Let's call the value $(1-\alpha)$ the *nominal* confidence level. We say nominal because unless n is large we shouldn't expect the true type 1 CI error rate to be α.

We can run numerical simulations to approximate of the true confidence level. We expect that as n gets larger the true confidence level of the large sample confidence interval will converge to the nominal value.

We ran such simulations for x drawn from the exponential distribution exp(1) (which is far from normal). For several values of n and nominal confidence level c we ran 100,000 trials. Each trial consisted of the following steps:

1. draw n samples from exp(1).
2. compute the sample mean $\bar{x}$ and sample standard deviation s.
3. construct the large sample c confidence interval: $\overline{x} \pm z_{\alpha/2} \cdot \dfrac{s}{\sqrt{n}}$.
4. check for a type 1 CI error, i.e. see if the true mean $\mu = 1$ is not in the interval.

With 100,000 trials, the empirical confidence level should closely approximate the true level. For comparison we ran the same tests on data drawn from a standard normal distribution. Here are the results.

	nominal conf.	
n	$1-\alpha$	simulated conf.
20	0.95	0.905
20	0.90	0.856
20	0.80	0.762
50	0.95	0.930
50	0.90	0.879
50	0.80	0.784
100	0.95	0.938
100	0.90	0.889
100	0.80	0.792
400	0.95	0.947
400	0.90	0.897
400	0.80	0.798

Simulations for exp(1)

	nominal conf.	
n	$1-\alpha$	simulated conf.
20	0.95	0.936
20	0.90	0.885
20	0.80	0.785
50	0.95	0.944
50	0.90	0.894
50	0.80	0.796
100	0.95	0.947
100	0.900	0.896
100	0.800	0.797
400	0.950	0.949
400	0.900	0.898
400	0.800	0.798

Simulations for $\text{N}(0,1)$.

For the exp(1) distribution we see that for $n = 20$ the simulated confidence of the large sample confidence interval is less than the nominal confidence $1-\alpha$. But for $n = 100$ the simulated confidence and nominal confidence are quite close. So for exp(1), n somewhere between 50 and 100 is large enough for most purposes.

Think: For $n = 20$ why is the simulated confidence for the $\mathrm{N}(0,1)$ distribution is smaller than the nominal confidence?

This is because we used $z_{\alpha/2}$ instead of $t_{\alpha/2}$. For large n these are quite close, but for $n = 20$ there is a noticable difference, e.g. $z_{0.025} = 1.96$ and $t_{0.025} = 2.09$.

Bootstrap confidence intervals
Class 24, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to construct and sample from the empirical distribution of data.
2. Be able to explain the bootstrap principle.
3. Be able to design and run an empirical percentile or basic bootstrap to compute confidence intervals.
4. Be able to design and run a parametric bootstrap to compute confidence intervals.

2 Introduction

The empirical bootstrap is a statistical technique popularized by Bradley Efron in 1979. Though remarkably simple to implement, the bootstrap would not be feasible without modern computing power. The key idea is to perform computations on the data itself to estimate the variation of statistics that are themselves computed from the same data. That is, the data is 'pulling itself up by its own bootstrap.' (A google search of 'by ones own bootstraps' will give you the etymology of this metaphor.) Such techniques existed before 1979, but Efron widened their applicability and demonstrated how to implement the bootstrap effectively using computers. He also coined the term 'bootstrap' [1].

The empircal bootstrap is also known as the nonparametric bootstrap.

Our main application of the bootstrap will be to estimate the variation of point estimates; that is, to estimate confidence intervals. An example will make our goal clear.

Example 1. Suppose we have data

$$x_1, x_2, \dots, x_n$$

If we knew the data was drawn from $\mathrm{N}(\mu, \sigma^2)$ with the unknown mean μ and known variance σ^2 then we have seen that

$$\left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

is a 95% confidence interval for μ.

Now suppose the data is drawn from some completely unknown distribution. To have a name we'll call this distribution F and its (unknown) mean μ. We can still use the sample mean $\overline{x}$ as a point estimate of μ. But how can we find a confidence interval for μ around $\overline{x}$? Our answer will be to use the bootstrap!

[1] Paraphrased from Dekking et al. *A Modern Introduction to Probabilty and Statistics*, Springer, 2005, page 275.

In fact, we'll see that the bootstrap handles other statistics as easily as it handles the mean. For example: the median, other percentiles or the trimmed mean. These are statistics where, even for normal distributions, it can be difficult to compute a confidence interval from theory alone.

3 Sampling

In statistics to sample from a set is to choose elements from that set. In a random sample the elements are chosen randomly. There are two common methods for random sampling.

Sampling without replacement
Suppose we draw 10 cards at random from a deck of 52 cards without putting any of the cards back into the deck between draws. This is called sampling without replacement or simple random sampling. With this method of sampling our 10 card sample will have no duplicate cards.

Sampling with replacement
Now suppose we draw 10 cards at random from the deck, but after each draw we put the card back in the deck and shuffle the cards. This is called sampling with replacement. With this method, the 10 card sample might have duplicates. It's even possible that we would draw the 6 of hearts all 10 times.

Think: What's the probability of drawing the 6 of hearts 10 times in a row?

Example 2. We can view rolling an 8-sided die repeatedly as sampling with replacement from the set {1,2,3,4,5,6,7,8}. Since each number is equally likely, we say we are sampling uniformly from the data.

Note. In practice if we take a small sample from a very large set then it doesn't matter whether we sample with or without replacement. For example, if we randomly sample 400 out of 300 million people in the U.S., then it is so unlikely that the same person will be picked twice that there is no real difference between sampling with or without replacement.

4 The empirical distribution of data

The empirical distribution of data is simply the distribution that you see in the data. Let's illustrate this with an example.

Example 3. Suppose we roll an 8-sided die 10 times and get the following data, written in increasing order:

```
1, 1, 2, 3, 3, 3, 3, 4, 7, 7.
```

Imagine writing these values on 10 slips of paper, putting them in a hat and drawing one at random. Then, for example, the probability of drawing a 3 is 4/10 and the probability of drawing a 4 is 1/10. The full empirical distribution can be put in a probability table

value x	1	2	3	4	7
$p(x)$	2/10	1/10	4/10	1/10	2/10

Notation. If we label the true distribution the data is drawn from as F, then we'll label the empirical distribution of the data as F^*. If we have enough data then the law of large numbers tells us that F^* should be a good approximation of F.

Example 4. In the dice example just above, the true and empirical distributions are:

value x	1	2	3	4	5	6	7	8
true $p(x)$	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
empirical $p(x)$	2/10	1/10	4/10	1/10	0	0	2/10	0

The true distribution F and the empirical distribution F^* of the 8-sided die.

Because F^* is derived strictly from data we call it the empirical distribution of the data. We will also call it the resampling distribution. Notice that we always know F^* explicitly. In particular the expected value of F^* is just the sample mean $\overline{x}$.

5 Resampling

The empirical (or nonparametric) bootstrap proceeds by resampling from the data. We continue the dice example above.

Example 5. Suppose we have 10 data points, given in increasing order:

```
1, 1, 2, 3, 3, 3, 3, 4, 7, 7
```

We view this as a sample taken from some underlying distribution. To resample is to sample with replacement from the empirical distribution, e.g. put these 10 numbers in a hat and draw one at random. Then put the number back in the hat and draw again. You draw as many numbers as the desired size of the resample.

To get us a little closer to implementing this on a computer we rephrase this in the following way. Label the 10 data points $x_1, x_2, \ldots, x_{10}$. To resample is to draw a number j from the uniform distribution on $\{1, 2, \ldots, 10\}$ and take x_j as our resampled value. In this case we could do so by rolling a 10-sided die. For example, if we roll a 6 then our resampled value is $x_6 = 3$, the 6^{th} element in our list.

If we want a resampled data set of size 5, then we roll the 10-sided die 5 times and choose the corresponding elements from the list of data. If the 5 rolls are

```
5, 3, 6, 6, 1
```

then the resample is

```
3, 2, 3, 3, 1.
```

Notes: 1. Because we are sampling with replacement, the same data point can appear multiple times when we resample.

2. Also because we are sampling with replacement, we can have a resample data set of any size we want, e.g. we could resample 1000 times.

Of course, in practice one uses a software package like R to do the resampling.

5.1 Star notation

If we have sample data of size n

$$x_1, x_2, \ldots, x_n$$

then we denote a resample of size m by adding a star to the symbols

$$x_1^*, x_2^*, \ldots, x_m^*$$

Similarly, just as $\overline{x}$ is the mean of the original data, we write $\overline{x}^*$ for the mean of the resampled data.

6 The empirical bootstrap

Suppose we have n data points

$$x_1, x_2, \ldots, x_n$$

drawn from a distribution F. An empirical bootstrap sample (or nonparametric bootstrap sample) is a resample of the same size n:

$$x_1^*, x_2^*, \ldots, x_n^*.$$

You should think of the latter as a sample of size n drawn from the empirical distribution F^*. For any statistic v computed from the original sample data, we can define a bootstrap statistic v^*. They both use the same formula but v^* is computed using the resampled data. With this notation we can state the bootstrap principle.

6.1 The bootstrap principle

The bootstrap setup is as follows:

1. $x_1, x_2, \ldots, x_n$ is a data sample drawn from a distribution F.
2. u is a statistic computed from the sample.
3. F^* is the empirical distribution of the data (the resampling distribution).
4. $x_1^*, x_2^*, \ldots, x_n^*$ is a resample of the data of the same size as the original sample
5. u^* is the statistic computed from the resample.

Then the bootstrap principle says that

1. F^* is approximately equal to F.

2. The statistic u is approximated by u^*.

3. The variation of u is approximated by the variation of u^*.

Our real interest is in point 3: we can approximate the variation of u by that of u^*. It turns out that, in practice, the bootstrap gives a reasonable approximation of the variation. We will exploit this to estimate the size of confidence intervals.

6.2 Why the resample is the same size as the original sample

This is straightforward: the variation of the statistic u will depend on the size of the sample. If we want to approximate this variation we need to use resamples of the same size.

7 The empirical bootstrap confidence interval

Here we will show two methods of computing an empirical confidence interval: the percentile bootstrap confidence interval in part (c) below and the basic bootstrap confidence interval in part (d).

A search of the internet credits these names to Efron. Below, we will briefly discuss the merits of each. The basic confidence interval is also called the reverse percentile confidence interval.

We will illustrate both methods with a 'toy' example using the same data.

Example 6. Toy example. We start with a made-up set of data that is small enough to show each step explicitly. The sample data is

```
30, 37, 36, 43, 42, 43, 43, 46, 41, 42
```

(a) Use the data to give a point estimate of the mean of the underlying distribution.

(b) Resample this to get 20 bootstrap samples.

(c) Use the bootstrap samples to find the 80% bootstrap percentile confidence interval for the mean.

(d) Use the bootstrap samples to find the 80% bootstrap basic confidence interval for the mean.

Note: R code for this example is shown in the section 'R annotated scripts' below. The code is also implemented in the R script, `class24-empiricalbootstrap.r` which is posted with our other R code.

Solution: (a) The sample mean is $\overline{x} = 40.3$, this is our point estimate.

(b) Our original sample contains 10 points. We used R to generate 20 bootstrap samples, each of size 10. Each of the 20 columns in the following array is one bootstrap sample. The values under the line are the means of the columns

36	36	42	42	41	42	42	43	42	36	43	42	42	43	43	43	36	43	36	46
43	37	36	43	43	41	36	41	46	30	43	46	42	30	43	43	41	41	37	43
43	43	43	42	46	42	42	43	43	43	36	43	42	30	36	43	42	41	41	37
42	43	37	37	43	36	43	43	43	41	42	42	37	43	36	42	46	43	43	42
36	46	36	41	43	30	43	42	46	46	43	37	46	42	46	43	41	43	41	36
41	42	43	43	46	30	36	41	36	46	36	30	42	43	42	37	42	41	37	43
42	46	30	46	30	43	42	41	46	42	37	46	43	43	37	43	30	43	46	37
43	42	43	37	42	43	46	43	37	42	42	37	36	43	46	30	43	46	46	41
43	43	41	46	46	43	30	46	36	41	42	42	36	42	37	36	46	43	42	43
30	46	43	42	43	42	41	42	37	43	43	43	43	43	41	36	43	42	43	46
39.9	42.4	39.4	41.9	42.3	39.2	40.1	42.5	41.2	41.0	40.7	40.8	40.9	40.2	40.7	39.6	41.0	42.6	41.2	41.4

(c) For the percentile method, we first compute $\overline{x}^*$ from each of our bootstrap samples. Here they are, sorted in increasing order.

```
39.2 39.4 39.6 39.9 40.1 40.2 40.7 40.7 40.8 40.9 41.0 41.0 41.2 41.2 41.4 41.9
42.3 42.4 42.5 42.6
```

The percentile method says to use the distribution of $\overline{x}^*$ as an approximation to the distribution of $\overline{x}$. The 80% confidence interval stretches from the 10th to the 90th percentile.

Since we have 20 bootstrap means, these are given by the 2nd and 18th elements in our list. Therefore the boostrap 80% percentile confidence interval is $[39.4,\ 42.4]$.

To make the example readable, we only computed 20 bootstrap means. The beautiful key to the bootstrap is, that since $\overline{x}^*$ is computed by resampling the original data, we can have a computer simulate $\overline{x}^*$ as many times as we'd like. Hence, by the law of large numbers, we can estimate the distribution of $\overline{x}^*$ with high precision.

Note: In our R code we use the quantile function to find the percentile values. This is easier than figuring out the index. It also is a little more sophisticated in finding the quantiles for a discrete set of values.

(d) The basic method is quite similar to the percentile method. The difference is that uses an alegebraic 'pivot' analogous to the pivot used in finding z or t confidence intervals. As in Example 1, to make the confidence interval we want to know how much the distribution of $\overline{x}$ varies around μ. That is, we'd like to know the distribution of

$$\delta = \overline{x} - \mu.$$

If we knew this distribution, then we could use the same algebra we saw when pivoting from non-rejection regions to confidence intervals. That is, we could find $\delta_{0.1}$ and $\delta_{0.9}$, the 0.1 and 0.9 critical values of δ. Then we would have

$$P(\delta_{0.9} \leq \overline{x} - \mu \leq \delta_{0.1} \,|\, \mu) = 0.8 \;\Leftrightarrow\; P(\overline{x} - \delta_{0.9} \geq \mu \geq \overline{x} - \delta_{0.1} \,|\, \mu) = 0.8$$

which gives an 80% confidence interval of

$$[\overline{x} - \delta_{0.1},\ \overline{x} - \delta_{0.9}]\,.$$

(As always with confidence intervals, we hasten to point out that the probabilities computed above are probabilities concerning the statistic $\overline{x}$ given that the true mean is μ.)

The bootstrap principle offers a practical approach to estimating the distribution of $\delta = \overline{x} - \mu$. It says that we can approximate it by the distribution of

$$\delta^* = \overline{x}^* - \overline{x}$$

where $\overline{x}^*$ is the mean of an empirical bootstrap sample.

Here is the beautiful key: since δ^* is computed by resampling the original data, we can have a computer simulate δ^* as many times as we'd like. Hence, by the law of large numbers, we can estimate the distribution of δ^* with high precision.

Next we compute $\delta^* = \overline{x}^* - \overline{x}$ for each bootstrap sample (i.e. each column) and sort them from smallest to biggest:

-1.1 -0.9 -0.7 -0.4 -0.2 -0.1 0.4 0.4 0.5 0.6 0.7 0.7 0.9 0.9 1.1 1.6 2.0 2.1 2.2 2.3

We will approximate the critical values $\delta_{0.1}$ and $\delta_{0.9}$ by $\delta^*_{0.1}$ and $\delta^*_{0.9}$. Since $\delta^*_{0.1}$ is at the 90th percentile we choose the 18th element in the list, i.e. 2.1. Likewise, since $\delta^*_{0.9}$ is at the 10th percentile we choose the 2nd element in the list, i.e. -0.9.

Therefore our bootstrap 80% basic confidence interval for μ is

$$[\overline{x} - \delta^*_{0.1},\ \overline{x} - \delta^*_{0.9}] \;=\; [40.3 - 2.1,\ 40.3 + 0.9] \;=\; [38.2,\ 41.2]$$

In this example we only generated 20 bootstrap samples so they would fit on the page. Using R, we would generate 10000 or more bootstrap samples in order to obtain a very accurate estimate of $\delta^*_{0.1}$ and $\delta^*_{0.9}$.

8 Justification for the bootstrap principle

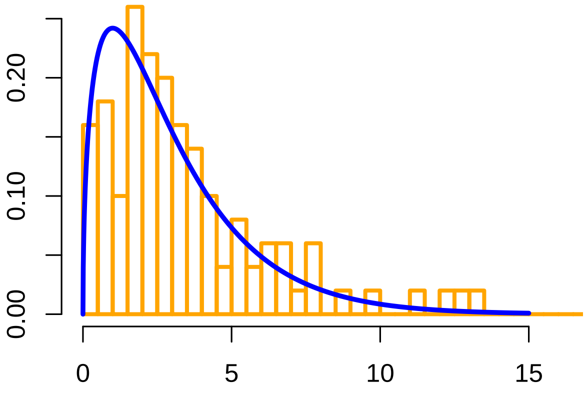
The bootstrap is remarkable because resampling gives us a decent estimate on how the point estimate might vary. We can only give you a 'hand-waving' explanation of this, but it's worth a try. The bootstrap is based roughly on the law of large numbers, which says, in short, that with enough data the empirical distribution will be a good approximation of the true distribution. Visually it says that the histogram of the data should approximate the density of the true distribution.

First let's note what resampling cannot do for us: it cannot improve our point estimate. For example, if we estimate the mean μ by $\overline{x}$ then in the bootstrap we would compute $\overline{x}^*$ for many resamples of the data. If we took the average of all the $\overline{x}^*$ we would expect it to be very close to $\overline{x}$. This wouldn't tell us anything new about the true value of μ.

Even with a fair amount of data the match between the true and empirical distributions is not perfect, so there will be error in our estimates for the mean (or any other value). But the amount of variation in the estimates is much less sensitive to differences between the true density and the data histogram: as long as they are reasonably close, the empirical and true distributions will exhibit the similar amounts of variation. So, in general the bootstrap principle is more robust when approximating the distribution of relative variation than when approximating absolute distributions.

What we have in mind is the scenario of our examples. The distribution (over different sets x of experimental data) of $\overline{x}$ is 'centered' at μ and the distribution (over different bootstrap samples x^* of x) of $\overline{x}^*$ is centered at $\overline{x}$. If there is a significant separation between $\overline{x}$ and μ then these two distributions will also differ significantly. On the other hand the distribution of $\delta = \overline{x} - \mu$ describes the variation of $\overline{x}$ about its center. Likewise the distribution of $\delta^* = \overline{x}^* - \overline{x}$ describes the variation of $\overline{x}^*$ about its center. So even if the centers are quite different the two variations about the centers can be approximately equal.

The figure below illustrates how the empirical distribution approximates the true distribution. To make the figure we generate 100 random values from a chi-square distribution with 3 degrees of freedom. The figure shows the pdf of the true distribution as a blue line and a histogram of the empirical distribution in orange.



The true and empirical distributions are approximately equal.

9 Other statistics

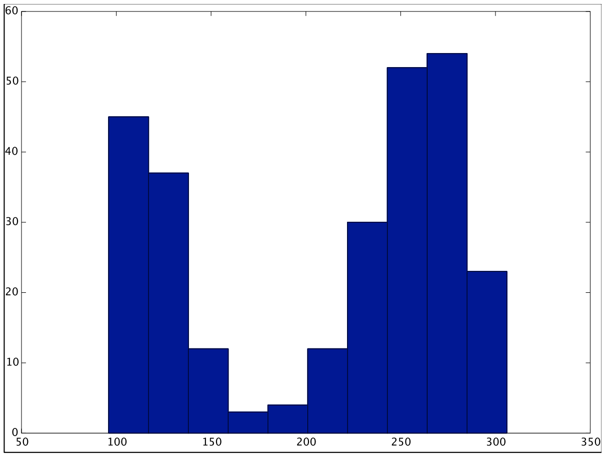
So far in this class we've avoided confidence intervals for the median and other statistics because their sample distributions are hard to describe theoretically. The bootstrap has no such problem. In fact, to handle the median all we have to do is change 'mean' to 'median' in the R code from Example 6.

Example 7. Old Faithful: confidence intervals for the median

Old Faithful is a geyser in Yellowstone National Park in Wyoming:

https://en.wikipedia.org/wiki/Old_Faithful

There is a publicly available data set which gives the durations of 272 consecutive eruptions. Here is a histogram of the data.



Question: Estimate the median length of an eruption and give a 90% percentile bootstrap confidence interval for the median.

Solution: The full answer to this question is in the R file `oldfaithful.r` and the Old Faithful data set. Both are posted on the class R code page. (Look under 'Other R code' for the old faithful script and data.)

Note: the code in `oldfaithful.r` assumes that the data `oldfaithful-data.txt` is in the current working directory.

Let's walk through a summary of the steps needed to answer the question.

1. Data: $x_1, \ldots, x_{272}$

2. Data median: $x_{\text{median}} = 240$

3. Find the median x^*_{median} of a bootstrap sample $x^*_1, \ldots, x^*_{272}$. Repeat 1000 times.

Put these 1000 values in order and pick out the 0.05 and 0.95 quantiles , i.e. the 50th and 950th biggest values. (In R we do this using the quantile() function) Call these $m^*_{0.05}$ and $m^*_{0.95}$.

5. The 90% percentile confidence interval for the medium is $[m^*_{0.05},\, m^*_{0.95}]$.

The bootstrap 90% CI we found for the Old Faithful data was $[230, 246]$. Since we used 1000 bootstrap samples a new simulation starting from the same sample data should produce a similar interval. If in Step 3 we increase the number of bootstrap samples to 10000, then the intervals produced by simulation would tend to be even more similar to each other. One common strategy is to increase the number of bootstrap samples until the resulting simulations produce intervals that vary less than some acceptable level.

Example 8. Using the Old Faithful data, estimate $P(|\overline{x} - \mu| > 5 \,|\, \mu)$.

Solution: We proceed exactly as in the previous example in finding a basic bootstrap confidence interval, except we use the mean instead of the median.

1. Data: $x_1, \ldots, x_{272}$

2. Data mean: $\overline{x} = 209.27$

3. Find the mean $\overline{x}^*$ of 1000 empirical bootstrap samples: $x_1^*, \ldots, x_{272}^*$.

4. Compute the bootstrap differences

$$\delta^* = \overline{x}^* - \overline{x}$$

5. The bootstrap principle says that we can use the distribution of δ^* as an approximation for the distribution $\delta = \overline{x} - \mu$. Thus,

$$P(|\overline{x} - \mu| > 5 \,|\, \mu) = P(|\delta| > 5 \,|\, \mu) \approx P(|\delta^*| > 5)$$

One bootstrap simulation for the Old Faithful data gave 0.230 for this probability.

10 Parametric bootstrap

Before getting started, we note that we only show the simplest algorithm for parametric bootstrap confidence intervals. In practice, more sophisticated algorithms are used. Since the bootsrap principle is the same in all cases, we feel it is worth encountering the simple algorithm in 18.05.

The examples in the previous sections all used the empirical bootstrap, which makes no assumptions at all about the underlying distribution and draws bootstrap samples by resampling the data. In this section we will look at the parametric bootstrap. The main difference between the parametric and empirical bootstrap is the source of the bootstrap sample. For the parametric bootstrap, we generate the bootstrap sample from a parametrized distribution.

Another difference, is that, for the parametric bootstrap, we will use basic bootstrap confidence intervals. It's hard to find a definitive answer on whether the percentile or basic interval is better in this case. As we noted above, the real answer seems to be that, in practice, parametric bootstrap confidence intervals are computed with more sophisticated algorithms.

Here are the elements of using the parametric bootstrap to estimate a confidence interval for a parameter.

0. Data: $x_1, \ldots, x_n$ drawn from a distribution $F(\theta)$ with unknown parameter θ.

1. A statistic $\hat{\theta}$ that estimates θ.

2. Our bootstrap samples are drawn from $F(\hat{\theta})$.

3. For each bootstrap sample

$$x_1^*, \dots, x_n^*$$

we compute $\hat{\theta}^*$ and the bootstrap difference $\delta^* = \hat{\theta}^* - \hat{\theta}$.

4. The bootstrap principle says that the distribution of δ^* approximates the distribution of $\delta = \hat{\theta} - \theta$.

5. Use the bootstrap differences to make a bootstrap confidence interval for θ.

Example 9. Suppose the data $x_1, \dots, x_{300}$ is drawn from an exp(λ) distribution. Assume also that the data mean $\overline{x} = 2$. Estimate λ and give a 95% parametric bootstrap confidence interval for λ.

Solution: This is implemented in the R script `class24-parametricbootstrap.r` which is posted with our other R code.

It's will be easiest to explain the solution using commented code.

```
# Parametric bootstrap

# Given 300 data points with mean 2.
# Assume the data is exp(lambda)

# PROBLEM: Compute a 95% parametric bootstrap confidence interval for lambda

# We are given the number of data points and mean
n = 300
xbar = 2

# The MLE for lambda is 1/xbar
lambda_hat = 1.0/xbar

# Generate the bootstrap samples
# Each column is one bootstrap sample (of 300 resampled values)
n_boot = 1000

# Here's the key difference with the empirical bootstrap:
# We draw the bootstrap sample from Exponential(lambda_hat).
x = rexp(n*n_boot, lambda_hat)
bootstrap_sample = matrix(x, nrow=n, ncol=n_boot)

# Compute the bootstrap lambda_star
lambda_star = 1.0/colMeans(bootstrap_sample)

# Compute the differences
delta_star = lambda_star - lambda_hat

# Find the 0.05 and 0.95 quantile for delta_star
d = quantile(delta_star, c(0.05,0.95))

# Calculate the 95% confidence interval for lambda.
ci = lambda_hat - c(d[2], d[1])

# This line of code is just one way to format the output text.
# sprintf is an old C function for doing this. R has many other
# ways to do the same thing.
s = sprintf("Confidence interval for lambda: [%.3f, %.3f]", ci[1], ci[2])
```

```
cat(s)
```

11 Building a better bootstrap

The first thing to say, is that we have just scratched the surface of bootstrap techniques. There are more sophisticated methods that correct for bias in the original sample or for skewness in the underlying distribution. There are also methods for when the original sample size is small.

For the nonparametric bootstrap, there are different opinions about which of the basic and percentile methods gives the most accurate results. What is clear is that, if the empirical distribution is symmetric, then the basic and percentile confidence intervals are equivalent. If the distribution is skewed then the two intervals are skewed in opposite directions.

Hesterberg in (https://www.tandfonline.com/doi/full/10.1080/00031305.2015.1089789) is fairly convincing that the percentile method performs better on skewed distributions.

On the other hand, Rice says of the percentile method, "Although this direct equation of quantiles of the bootstrap sampling distribution with confidence limits may seem initially appealing, it's rationale is somewhat obscure." [2]

We'll give Hesterberg the final word. He contributed several excellent responses to a discussion group. Unfortunately the link is no longer active, but here is one of his posts.

> Skewness is a fact of life. So what do you do when there is skewness? A bootstrap percentile interval is usually a good choice – much better than a symmetric interval like t, that pretends there is no skewness. It is fine for a beginning stats student, and for most applications in practice. Someone more advanced (or using easy-to-use software) can use a BCa or bootstrap-t interval, for better accuracy.
>
> Bootstrapping the median is a different issue – for small samples the median is quite sensitive to whether the population being sampled from is continuous or discrete. So if the population is continuous, but you use the nonparametric bootstrap that draws from a discrete distribution (the data), the bootstrap distribution won't look like the true sampling distribution. Even so, the bootstrap percentile interval is not bad in this case, close to the exact confidence interval for the median (typically the same or one order statistic different).

12 R annotated transcripts

12.1 Using R to generate a empirical bootstrap confidence intervals

This code only generates 20 bootstrap samples. In real practice we would generate many more bootstrap samples. It is making a bootstrap confidence interval for the mean. This code is implemented in the R script `class24-empiricalbootstrap.r` which is posted with our other R code.

[2] John Rice, *Mathematical Statistics and Data Analysis*, 2nd edition, p. 272.

```
# Data for the example 6
x = c(30,37,36,43,42,43,43,46,41,42)
n = length(x)

# sample mean
xbar = mean(x)

n_boot = 20
# Generate 20 bootstrap samples, i.e. an n x 20 array of
# random resamples from x
tmp_data = sample(x, n*n_boot, replace=TRUE)
bootstrap_sample = matrix(tmp_data, nrow=n, ncol=n_boot)

# Compute the means $\overline{x}^*$
xbar_star = colMeans(bootstrap_sample)

# Calculate the bounds for the 80% percentile confidence interval.
percentile_ci = quantile(xbar_star, c(0.1, 0.9))
cat('80% percentile confidence interval: ',percentile_ci, '\n')

# Compute $\delta^*$ for each bootstrap sample
delta_star = xbar_star - xbar

# Find the 0.1 and 0.9 quantiles for delta_star
d = quantile(delta_star, c(0.1, 0.9))

# Calculate the bounds for the 80% basic confidence interval.
# Note how pivoting reverses the order of d[1] and d[2]
basic_ci = xbar - c(d[2], d[1])
cat('80% basic confidence interval: ',basic_ci, '\n')

# ALTERNATIVE: the quantile() function is sophisticated about
# choosing a quantile between two data points. A less sophisticated
# approach is to pick the quantiles by sorting xbar_start and delta_star and
# choosing the index that corresponds to the desired quantiles.
# This is what we did in the text above. We show the code using this for the
percentile method below.

# Sort the results
sorted_xbar_star = sort(xbar_star)

# Find the 0.1 and 0.9 quantiles values of xbar_star
q1_alt = sorted_xbar_star[2]
q9_alt = sorted_xbar_star[18]

# Find and print the 80% percentile confidence interval for the mean
ci_alt = c(q1_alt, q9_alt)
cat('Alternative confidence interval: ',ci_alt, '\n')
```

Linear regression
Class 26, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to use the method of least squares to fit a line to bivariate data.
2. Be able to give a formula for the total squared error when fitting any type of curve to data.
3. Be able to say the words homoscedasticity and heteroscedasticity.

2 Introduction

Suppose we have collected bivariate data (x_i, y_i), $i = 1, \dots, n$. The goal of linear regression is to model the relationship between x and y by finding a function $y = f(x)$ that is a close fit to the data. The modeling assumptions we will use are that x_i is **not** random and that y_i is a function of x_i plus some random noise. With these assumptions x is called the independent or predictor variable and y is called the dependent or response variable.

Here is a series of examples showing the results of linear regression. We will discuss the details of how to do linear regression in the next section.

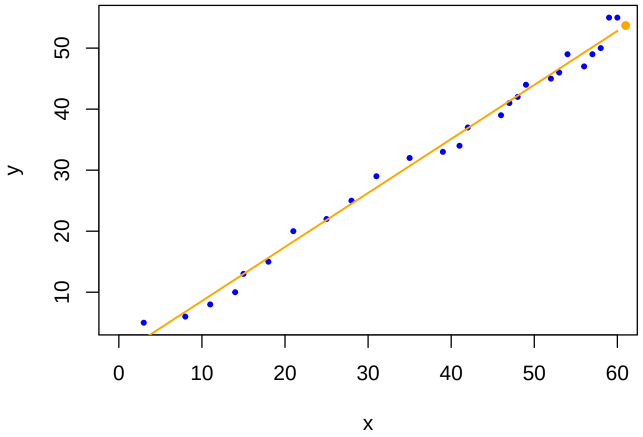
Example 1. The cost of a first class stamp in cents over time is given in the following list.
0.05 (1963) 0.06 (1968) 0.08 (1971) 0.10 (1974) 0.13 (1975) 0.15 (1978) 0.20 (1981)
0.22 (1985) 0.25 (1988) 0.29 (1991) 0.32 (1995) 0.33 (1999) 0.34 (2001) 0.37 (2002)
0.39 (2006) 0.41 (2007) 0.42 (2008) 0.44 (2009) 0.45 (2012) 0.46 (2013) 0.49 (2015)
0.49 (2017) 0.50 (2018) 0.55 (2019)

Using the R function `lm` we found the 'least squares fit' for a line to this data is

$$y = -0.21390 + 0.88203x,$$

where x is the number of years since 1960 and y is in cents.

Using this result we 'predict' that in 2021 ($x = 61$) the cost of a stamp will be 53.6 cents (since $-0.21390 + 0.88203 \cdot 61 = 53.6$).



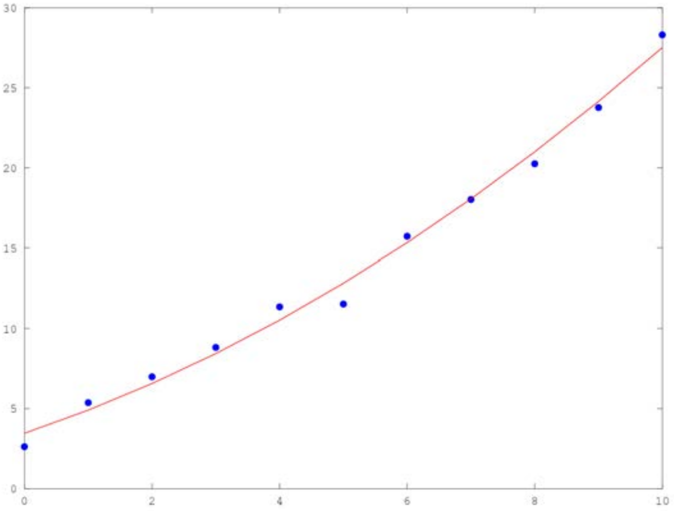
Stamp cost (cents) vs. time (years since 1960). Orange dot is predicted cost in 2021.

Note that none of the data points actually lie on the line. Rather this line has the 'best fit' with respect to all the data, with a small error for each data point.

(Note, the actual cost of a stamp dropped in January 2021 was 55 cents. See https://en.wikipedia.org/wiki/History_of_United_States_postage_rates)

Example 2. Suppose we have n pairs of fathers and adult sons. Let x_i and y_i be the heights of the i^{th} father and son, respectively. The least squares line for this data could be used to predict the adult height of a young boy from that of his father.

Example 3. We are not limited to best fit lines. For all positive d, the method of least squares may be used to find a polynomial of degree d with the 'best fit' to the data. Here's a figure showing the least squares fit of a parabola ($d = 2$).



Fitting a parabola, $ax^2 + bx + c$, to data

Example 4. In fact, we can use linear regression to fit many other types of curves to bivariate data.

3 Fitting a line using least squares

Suppose we have data (x_i, y_i) as above. Our first goal is to find the line

$$y = ax + b$$

that 'best fits' the data. Our model says that each y_i is predicted by x_i up to some error ϵ_i:

$$y_i = ax_i + b + \epsilon_i.$$

So

$$\epsilon_i = y_i - ax_i - b.$$

The method of least squares finds the values $\hat{a}$ and $\hat{b}$ of a and b that minimize the sum of the squared errors:

$$S(a,b) = \sum \epsilon_i^2 = \sum_i (y_i - ax_i - b)^2.$$

Using calculus or linear algebra (details in the appendix), we find

$$\hat{a} = \frac{s_{xy}}{s_{xx}} \qquad \hat{b} = \bar{y} - \hat{a}\,\bar{x} \tag{1}$$

where

$$\bar{x} = \frac{1}{n}\sum x_i, \quad \bar{y} = \frac{1}{n}\sum y_i, \quad s_{xx} = \frac{1}{(n-1)}\sum (x_i - \bar{x})^2, \quad s_{xy} = \frac{1}{(n-1)}\sum (x_i - \bar{x})(y_i - \bar{y}).$$

Here $\bar{x}$ is the sample mean of x, $\bar{y}$ is the sample mean of y, s_{xx} is the sample variance of x, and s_{xy} is the sample covariance of x and y.

Example 5. Use least squares to fit a line to the following data: (0,1), (2,1), (3,4).

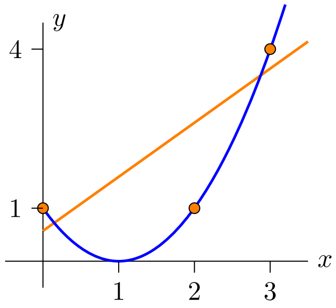
Solution: In our case, $(x_1, y_1) = (0,1)$, $(x_2, y_2) = (2,1)$ and $(x_3, y_3) = (3,4)$. So

$$\bar{x} = \frac{5}{3}, \quad \bar{y} = 2, \quad s_{xx} = \frac{7}{3}, \quad s_{xy} = 2$$

Using the above formulas we get

$$\hat{a} = \frac{6}{7}, \quad \hat{b} = \frac{4}{7}.$$

So the least squares line has equation $y = \frac{6}{7}x + \frac{4}{7}$. This is shown as the orange line in the following figure. We will discuss the blue parabola soon.



Least squares fit of a line (orange) and a parabola (blue)

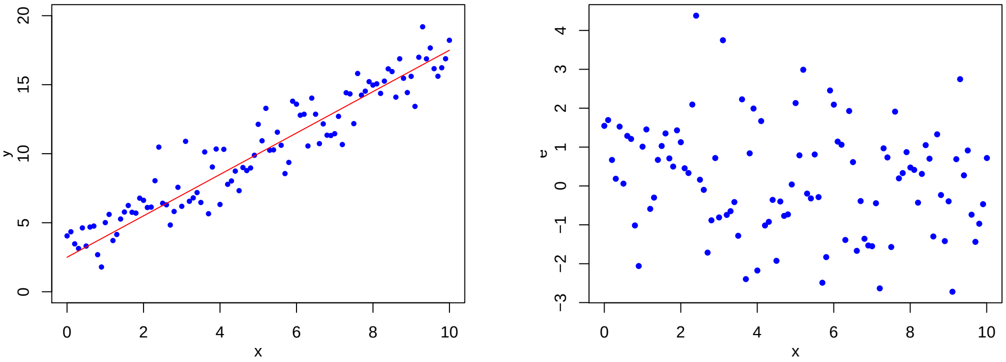
Simple linear regression: It's a little confusing, but the word linear in 'linear regression' does not refer to fitting a line. We will explain its meaning below. However, the most common curve to fit is a line. When we fit a line to bivariate data it is called simple linear regression.

3.1 Residuals

For a line the model is

$$y_i = \hat{a}x + \hat{b} + \epsilon_i.$$

We think of $\hat{a}x_i + \hat{b}$ as predicting or explaining y_i. The left-over term ϵ_i is called the residual, which we think of as random noise or measurement error. A useful visual check of the linear regression model is to plot the residuals. The data points should hover near the regression line. The residuals should look about the same across the range of x.

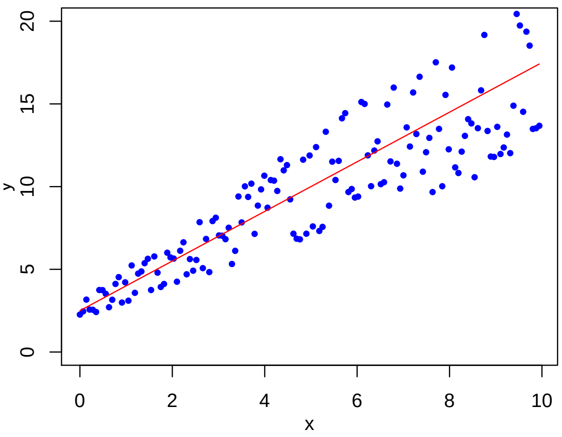


Data with regression line (left) and residuals (right). Note the homoscedasticity.

3.2 Homoscedasticity

An important assumption of the linear regression model is that the residuals ϵ_i have the same variance for all i. This is called homoscedasticity. You can see this is the case for both figures above. The data hovers in the band of fixed width around the regression line and at every x the residuals have about the same vertical spread.

Below is a figure showing heteroscedastic data. The vertical spread of the data increases as x increases. Before using least squares on this data we would need to transform the data to be homoscedastic.



Heteroscedastic Data

4 Linear regression for fitting polynomials

When we fit a line to data it is called simple linear regression. We can also use linear regression to fit polynomials to data. The use of the word linear in both cases may seem confusing. This is because the word 'linear' in linear regression does not refer to fitting a line. Rather it refers to the linear algebraic equations for the unknown parameters.

Example 6. Take the same data as in Example 5 and use least squares to find the best fitting parabola to the data.

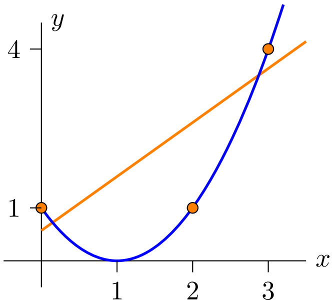
Solution: A parabola has the formula $y = ax^2 + bx + c$. The squared error is

$$S(a, b, c) = \sum (y_i - (ax_i^2 + bx_i + c))^2.$$

After substituting the given values for each x_i and y_i, we can use calculus to find the triple (a, b, c) that minimizes S. With this data, we find that the least squares parabola has equation

$$y = x^2 - 2x + 1.$$

Note that for 3 points the quadratic fit is perfect.



Least squares fit of a line (orange) and a parabola (blue)

Example 7. The pairs (x_i, y_i) may give the age and vocabulary size of a n children. Since we expect that young children acquire new words at an accelerating pace, we might guess that a higher order polynomial would best fit the data.

Example 8. (Transforming the data) Sometimes it is necessary to transform the data before using linear regression. For example, let's suppose the relationship is exponential, i.e. $y = ce^{ax}$. Then

$$\ln(y) = ax + \ln(c).$$

So we can use simple linear regression on the data $(x_i, \ln(y_i))$ to obtain a model

$$\ln(y) = \hat{a}x + \hat{b}$$

and then exponentiate to obtain the exponential model

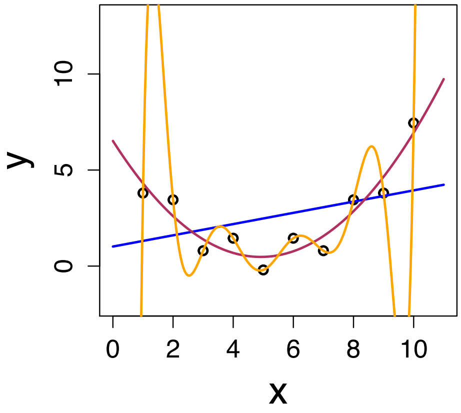
$$y = e^{\hat{b}}e^{\hat{a}x}.$$

4.1 Overfitting

You can always achieve a better fit by using a higher order polynomial. For instance, given 6 data points (with distinct x_i) one can always find a fifth order polynomial that goes through all of them. This can result in what's called overfitting. That is, fitting the noise as well as the true relationship between x and y. An overfit model will fit the original data better but perform less well on predicting y for new values of x. Indeed, a primary challenge of statistical modeling is balancing model fit against model complexity.

Example 9. In the plot below, we fit polynomials of degree 1, 2, and 9 to bivariate data consisting of 10 data points. The degree 2 model (maroon) gives a significantly better fit than the degree 1 model (blue). The degree 10 model (orange) gives fits the data exactly, but at a glance we would guess it is overfit. That is, we don't expect it to do a good job fitting the next data point we see.

In fact, we generated this data using a quadratic model, so the degree 2 model will tend to perform best fitting new data points.



4.2 R function `lm`

As you would expect we don't actually do linear regression by hand. Computationally, linear regression reduces to solving simultaneous equations, i.e. to matrix calculations. The R function `lm` can be used to fit any order polynomial to data. (`lm` stands for linear model). We will explore this in the next studio class. In fact `lm` can fit many types of functions besides polynomials, as you can explore using R help or google.

5 Multiple linear regression

Data is not always bivariate. It can be trivariate or even of some higher dimension. Suppose we have data in the form of tuples

$$(y_i,\ x_{1,i},\ x_{2,i},\ \dots\ x_{m,i})$$

We can analyze this in a manner very similar to linear regression on bivariate data. That is, we can use least squares to fit the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m.$$

Here each x_j is a predictor variable and y is the response variable. For example, we might be interested in how a fish population varies with measured levels of several pollutants, or we might want to predict the adult height of a son based on the height of the mother and the height of the father.

We don't have time in 18.05 to study multiple linear regression, but we wanted you to see the name.

6 Least squares as a statistical model

The linear regression model for fitting a line says that the value y_i in the pair (x_i, y_i) is drawn from a random variable

$$Y_i = ax_i + b + \varepsilon_i$$

where the 'error' terms ε_i are independent random variables with mean 0 and standard deviation σ. The standard assumption is that the ε_i are i.i.d. with distribution $N(0, \sigma^2)$. So, the mean of Y_i is given by:

$$E[Y_i] = ax_i + b + E[\varepsilon_i] = ax_i + b.$$

From this perspective, the least squares method chooses the values of a and b which minimize the sample variance about the line.

In fact, under the assumption that $\varepsilon_i \sim N(0, \sigma^2)$, the least square estimate $(\hat{a}, \hat{b})$ coincides with the maximum likelihood estimate for the parameters (a, b); that is, among all possible coefficients, $(\hat{a}, \hat{b})$ are the ones that make the observed data most probable.

7 Regression to the mean

The reason for the term 'regression' is that the predicted response variable y will tend to be 'closer' to (i.e., regress to) its mean than the predictor variable x is to its mean. Here closer is in quotes because we have to control for the scale (i.e. standard deviation) of each variable. The way we control for scale is to first standardize each variable.

$$u_i = \frac{x_i - \bar{x}}{\sqrt{s_{xx}}}, \quad v_i = \frac{y_i - \bar{y}}{\sqrt{s_{yy}}}.$$

Standardization changes the mean to 0 and variance to 1:

$$\bar{u} = \bar{v} = 0, \quad s_{uu} = s_{vv} = 1.$$

The algebraic properties of covariance show

$$s_{uv} = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \rho,$$

the correlation coefficient. Thus the least squares fit to $v = au + b$ has

$$\hat{a} = \frac{s_{uv}}{s_{uu}} = \rho \quad \text{and} \quad \hat{b} = \bar{v} - \hat{a}\bar{u} = 0.$$

So the least squares line is $v = \rho u$. Since ρ is the correlation coefficient, it is between -1 and 1. Let's assume it is positive and less than 1 (i.e., x and y are positively but not perfectly correlated). Then the formula $v = \rho u$ means that if u is positive then the predicted value of v is less than u. That is, v is closer to 0 than u. Equivalently,

$$\frac{y - \bar{y}}{\sqrt{s_{yy}}} < \frac{x - \bar{x}}{\sqrt{s_{xx}}}$$

i.e., y regresses to $\bar{y}$. Notice how the standardization takes care of controlling the scale.

Consider the extreme case of 0 correlation between x and y. Then, no matter what the x value, the predicted value of y is always $\bar{y}$. That is, y has regressed all the way to its mean.

Note also that the regression line always goes through the point $(\bar{x}, \bar{y})$.

Example 10. Regression to the mean is important in longitudinal studies. Rice (*Mathematical Statistics and Data Analysis*) gives the following example. Suppose children are given an IQ test at age 4 and another at age 5 we expect the results will be positively correlated. The above analysis says that, on average, those kids who do poorly on the first test will tend to show improvement (i.e. regress to the mean) on the second test. Thus, a useless intervention might be misinterpreted as useful since it seems to improve scores.

Example 11. Another example with practical consequences is reward and punishment. Imagine a school where high performance on an exam is rewarded and low performance is punished. Regression to the mean tells us that (on average) the high performing students will do slightly worse on the next exam and the low performing students will do slightly better. An unsophisticated view of the data will make it seem that punishment improved performance and reward actually hurt performance. There are real consequences if those in authority act on this idea.

8 Appendix

We collect in this appendix a few things you might find interesting. You will not be asked to know these things for exams.

8.1 Proof of the formula for least square fit of a line

The most straightforward proof is to use calculus. The sum of the squared errors is

$$S(b,a) = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Taking partial derivatives (and remembering that x_i and y_i are the data, hence constant)

$$\begin{aligned} \frac{\partial S}{\partial b} &= \sum_{i=1}^{n} -2(y_i - ax_i - b) = 0 \\ \frac{\partial S}{\partial a} &= \sum_{i=1}^{n} -2x_i(y_i - ax_i - b) = 0 \end{aligned}$$

Summing this up we get two linear equations in the unknowns b and a:

$$\begin{aligned} \left(\sum x_i\right) a + nb &= \sum y_i \\ \left(\sum x_i^2\right) a + \left(\sum x_i\right) b &= \sum x_i y_i \end{aligned}$$

Solving for a and b gives the formulas in Equation (1).

A sneakier approach which avoids calculus is to standardize the data, find the best fit line, and then unstandardize. We omit the details.

For a slew of applications across disciplines see:
https://en.wikipedia.org/wiki/Linear_regression#Applications_of_linear_regression

8.2 Measuring the fit

Once one computes the regression coefficients, it is important to check how well the regression model fits the data (i.e., how closely the best fit line tracks the data). A common but crude 'goodness of fit' measure is the coefficient of determination, denoted R^2. We'll need some notation to define it. The total sum of squares is given by:

$$\text{TSS} = \sum (y_i - \bar{y})^2.$$

The residual sum of squares is given by the sum of the squares of the residuals. When fitting a line, this is:

$$\text{RSS} = \sum (y_i - \hat{a}\, x_i - \hat{b})^2.$$

The RSS is the "unexplained" portion of the total sum of squares, i.e. unexplained by the regression equation. The difference TSS − RSS is the "explained" portion of the total sum of squares. The coefficient of determination R^2 is the ratio of the "explained" portion to the total sum of squares:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}.$$

In other words, R^2 measures the proportion of the variability of the data that is accounted for by the regression model. A value close to 1 indicates a good fit, while a value close to 0 indicates a poor fit. In the case of simple linear regression, R^2 is simply the square of the correlation coefficient between the observed values y_i and the predicted values $ax_i + b$.

Example 12. In the overfitting example (9), the values of R^2 are:

degree	R^2
1	0.3968
2	0.9455
9	1.0000

Notice the goodness of fit measure increases as n increases. The fit is better, but the model also becomes more complex, since it takes more coefficients to describe higher order polynomials.