Exam 2 Practice Questions –solutions, 18.05, Spring 2022

1 Topics

- Statistics: data, MLE
- Bayesian inference: prior, likelihood, posterior, predictive probability, probability intervals
- Frequentist inference: NHST

2 Using the probability tables

You should become familiar with the probability tables at the end of these notes.

Problem 1. *Use the standard normal table to find the following values. In all the problems Z is a standard normal random variable.*

(a) *(i)* $P(Z < 1.5)$ *(ii)* $P(Z > 1.5)$ *(iii)* $P(-1.5 < Z < 1.5)$ *(iv)* $P(Z \leq 1.625)$

Solution: (i) The table gives this value as $P(Z < 1.5) = 0.9332$.

(ii) This is the complement of the answer in (i): $P(Z > 1.5) = 1 - 0.9332 = 0.0668$. Or by symmetry we could use the table for -1.5.

(iii) We want $P(Z < 1.5) - P(Z < -1.5) = P(Z < 1.5) - P(Z > 1.5)$. This is the difference of the answers in (i) and (ii): 0.8664.

(iv) A rough estimate is the average of $P(Z < 1.6)$ and $P(Z < 1.65)$. That is,

$$P(Z < 1.625) \approx \frac{P(Z < 1.6) + P(Z < 1.65)}{2} = \frac{0.9452 + 0.9505}{2} = 0.9479.$$

(b) *(i) The right-tail with probability $\alpha = 0.05$.*

(ii) The two-sided rejection region with probability $\alpha = 0.2$.

(iii) Find the range for the middle 50% of probability.

Solution: (i) We are looking for the table entry with probability 0.95. This is between the table entries for $z = 1.65$ and $z = 1.60$ and very close to that of $z = 1.65$. Answer: the region is $[1.64, \infty)$. (R gives the 'exact' lower limit as 1.644854.)

(ii) We want the table entry with probability 0.1. The table probabilities for $z = -1.25$ and $z = -1.30$ are 0.1056 and 0.0968. Since 0.1 is about 1/2 way from the first to the second we take the left critical value as -1.275. Our region is

$$(-\infty, -1.275) \cup (1.275, \infty).$$

(R gives `qnorm(0.1, 0, 1) = -1.2816`.)

(iii) This is the range from $q_{0.25}$ to $q_{0.75}$. With the table we estimate $q_{0.25}$ is about 1/2 of the way from -0.65 to -0.70, i.e. ≈ -0.675. So, the range is $[-0.675, 0.675]$.

Problem 2. *The t-tables are different. They give the right critical values corresponding to probabilities. To save space we only give critical values for $p \leq 0.5$. You need to use the symmetry of the t-distribution to get them for $p < 0.5$. That is, $t_{df,\,p} = -t_{df,\,1-p}$, e.g. $t_{5,\,0.975} = -t_{5,\,0.025}$.*

Use the t-table to estimate the following values. In all the problems T is a random variable drawn from a t-distribution with the indicated number of degrees of freedom.

(a) *(i) $P(T > 1.6)$, with $df = 3$*

(ii) $P(T < 1.6)$ with $df = 3$

(iii) $P(-1.68 < T < 1.68)$ with $df = 49$

(iv) $P(-1.6 < T < 1.6)$ with $df = 49$

Solution: (i) The question asks to find which p-value goes with $t = 1.6$ when $df = 3$. We look in the $df = 3$ row of the table and find 1.64 goes with $p = 0.100$ So, $P(T > 1.6 \,|\, df = 3) \approx 0.1$. (The true value is a little bit greater.)

(ii) $P(T < 1.6 \,|\, df = 3) = 1 - P(T > 1.6 \,|\, df = 3) \approx 0.9$.

(iii) Using the $df = 49$ row of the t-table we find $P(T > 1.68 \,|\, df = 49) = 0.05$.

Now, by symmetry $P(T < -1.68 \,|\, df = 49) = 0.05$ and $\boxed{P(-1.68 < T < 1.68 \,|\, df = 49) = 0.9}$.

(iv) Using the $df = 49$ row of the t-table we find $P(T > 1.68 \,|\, df = 49) = 0.05$ and $P(T > 1.30 \,|\, df = 49) = 0.1$. We can do a rough interpolation: $P(T > 1.6 \,|\, df = 49) \approx 0.06$.

Now, by symmetry $P(T < -1.6 \,|\, df = 49) \approx 0.06$ and $\boxed{P(-1.6 < T < 1.6 \,|\, df = 49) \approx 0.88}$. (R gives 0.8839727.)

(b) *(i) The critical value for probability $\alpha = 0.05$ for 8 degrees of freedom.*

(ii) The two-sided rejection region with probability $\alpha = 0.2$ for 16 degrees of freedom.

(iii) Find the range for the middle 50% of probability with $df = 20$.

Solution: (i) This is a straightforward lookup: The $p = 0.05$, $df = 8$ entry is $\boxed{1.86}$.

(ii) For a two-sided rejection region we need 0.1 probability in each tail. The critical value at $p = 0.1$, $df = 16$ is 1.34. So (by symmetry) the rejection region is

$$(-\infty, -1.34) \cup (1.34, \infty).$$

(iii) This is the range from $q_{0.25}$ to $q_{0.75}$, i.e. from critical values $t_{0.75}$ to $t_{0.25}$. The table only gives critical for 0.2 and 0.3 For $df = 20$ these are 0.86 and 0.53. We average these to estimate the 0.25 critical value as 0.7. Answer: the middle 50% of probability is approximately between t-values -0.7 and 0.7.

(If we took into account the bell shape of the t-distribution we would estimate the 0.25 critical value as slightly closer to 0.53 than 0.86. Indeed R gives the value 0.687.)

Problem 3. *The chi-square tables are different. They give the right critical values corresponding to probabilities.*

Use the chi-square tables table to find the following values. In all the problems X^2 is a random variable drawn from a χ^2-distribution with the indicated number of degrees of freedom.

(a) *(i)* $P(X^2 > 1.6)$*, with* $df = 3$

(ii) $P(X^2 > 20)$ *with* $df = 16$

Solution: (i) Looking in the $df = 3$ row of the chi-square table we see that 1.6 is about 1/5 of the way between the values for $p = 0.7$ and $p = 0.5$. So we approximate $P(X^2 > 1.6) \approx 0.66$. (The true value is 0.6594.)

(ii) Looking in the $df = 16$ row of the chi-square table we see that 20 is about 1/4 of the way between the values for $p = 0.2$ and $p = 0.3$. We estimate $P(X^2 > 20) = 0.25$. (The true value is 0.220)

(b) *(i) The right critical value for probability* $\alpha = 0.05$ *for 8 degrees of freedom.*

(ii) The two-sided rejection region with probability $\alpha = 0.2$ *for 16 degrees of freedom.*

Solution: (i) This is in the table in the $df = 8$ row under $p = 0.05$. Answer: 15.51

(ii) We want the critical values for $p = 0.9$ and $p = 0.1$ from the $df = 16$ row of the table.

$$[0, 9.31] \cup [23.54, \infty).$$

3 Data

Problem 4. *The following data is from a random sample: 5, 1, 3, 3, 8.*

Compute the sample mean, sample standard deviation and sample median.

Solution: Sample mean $20/5 = 4$.

Sample variance $= \dfrac{1^2 + (-3)^2 + (-1)^2 + (-1)^2 + 4^2}{5-1} = 7.$

Sample standard deviation $= \sqrt{7}$.

Sample median $= 3$.

4 MLE

Problem 5. **(a)** *A coin is tossed 100 times and lands heads 62 times. Find the maximum likelihood estimate for the probability* θ *of heads.*

Solution: The likelihood function is

$$p(\text{data}|\theta) = \binom{100}{62}\theta^{62}(1-\theta)^{38} = c\theta^{62}(1-\theta)^{38}.$$

To find the MLE we find the derivative of the log-likelihood and set it to 0.

$$\ln(p(\text{data}|\theta)) = \ln(c) + 62\ln(\theta) + 38\ln(1-\theta).$$

$$\frac{d\ \ln(p(\text{data}|\theta))}{d\theta} = \frac{62}{\theta} - \frac{38}{1-\theta} = 0.$$

The algebra leads to the MLE $\boxed{\theta = 62/100}$.

(b) *A coin is tossed n times and lands heads k times. Find the maximum likelihood estimate for the probability θ of heads.*

Solution: The computation is identical to part (a). The likelihood function is

$$p(\text{data}|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} = c\theta^k(1-\theta)^{n-k}.$$

To find the MLE we set the derivative of the log-likelihood and set it to 0.

$$\ln(p(\text{data}|\theta)) = \ln(c) + k\ln(\theta) + (n-k)\ln(1-\theta).$$

$$\frac{d\,\ln(p(\text{data}|\theta))}{d\theta} = \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0.$$

The algebra leads to the MLE $\boxed{\theta = k/n}$.

Problem 6. *Suppose the data set $y_1, \ldots, y_n$ is a drawn from a random sample consisting of i.i.d. discrete uniform distributions with range 1 to N. Find the maximum likelihood estimate of N.*

Solution: If $N < \max(y_i)$ then the likelihood $p(y_1, \ldots, y_n|N) = 0$. So the likelihood function is

$$p(y_1, \ldots, y_n|N) = \begin{cases} 0 & \text{if } N < \max(y_i) \\ \left(\frac{1}{N}\right)^n & \text{if } N \geq \max(y_i) \end{cases}$$

This is maximized when N is as small as possible. Since $N \geq \max(y_i)$ the MLE is $\boxed{N = \max(y_i).}$

Problem 7. *Suppose data $x_1, \ldots, x_n$ is drawn from an exponential distribution exp(λ). Find the maximum likelihood for λ.*

Solution: The pdf of $\exp(\lambda)$ is $p(x|\lambda) = \lambda e^{-\lambda x}$. So the likelihood and log-likelihood functions are

$$p(\text{data}|\lambda) = \lambda^n e^{-\lambda(x_1+\cdots+x_n)}, \qquad \ln(p(\text{data}|\lambda)) = n\ln(\lambda) - \lambda\sum x_i.$$

Taking a derivative with respect to λ and setting it equal to 0:

$$\frac{d\,\ln(p(\text{data}|\lambda))}{d\lambda} = \frac{n}{\lambda} - \sum x_i = 0 \qquad\qquad \Rightarrow \frac{1}{\lambda} = \frac{\sum x_i}{n} = \bar{x}.$$

So the MLE is $\boxed{\lambda = 1/\bar{x}}$.

Problem 8. *Suppose $x_1, \ldots, x_n$ is a data set drawn from a geometric$(1/a)$ distribution. Find the maximum likelihood estimate of a. Here, geometric(p) means the probability of success is p and we run trials until the first success and report the total number of trials, including the success. For example, the sequence $FFFFS$ is 4 failures followed by a success, which produces $x = 5$.*

Solution: $P(x_i|a) = \left(1 - \frac{1}{a}\right)^{x_i-1}\frac{1}{a}. = \left(\frac{a-1}{a}\right)^{x_i-1}\frac{1}{a}.$

So, the likelihood function is

$$P(\text{data}|a) = \left(\frac{a-1}{a}\right)^{\sum x_i - n} \left(\frac{1}{a}\right)^n$$

The log likelihood is

$$\ln(P(\text{data}|a)) = \left(\sum x_i - n\right)(\ln(a-1) - \ln(a)) - n\ln(a).$$

Taking the derivative

$$\frac{d\,\ln(P(\text{data}|a))}{d\,a} = \left(\sum x_i - n\right)\left(\frac{1}{a-1} - \frac{1}{a}\right) - \frac{n}{a} = 0 \;\Rightarrow\; \frac{\sum x_i}{n} = a.$$

The maximum likelihood estimate is $\boxed{a = \bar{x}}$.

Problem 9. *You want to estimate the size of an MIT class that is closed to visitors. You know that the students are numbered from 1 to n, where n is the number of students. You call three random students out of the classroom and ask for their numbers, which turn out to be 1, 3, 7. Find the maximum likelihood estimate for n. (Hint: the student #'s are drawn from a discrete uniform distribution.)*

Solution: If there are n students in the room then for the data 1, 3, 7 (occuring in any order) the likelihood is

$$p(\text{data}\,|\,n) = \begin{cases} 0 & \text{for } n < 7 \\ 1/\binom{n}{3} = \frac{3!}{n(n-1)(n-2)} & \text{for } n \geq 7 \end{cases}$$

Maximizing this does not require calculus. It clearly has a maximum when n is as small as possible. Answer: $\boxed{n = 7}$.

5 Bayesian updating: discrete prior, discrete likelihood

Problem 10. Twins

(a) *Suppose 1/4 of twins are identical and 3/4 of twins are fraternal. If you are pregnant with twins of the same sex, what is the probability that they are identical?*

Solution: This is a Bayes' theorem problem. The likelihoods are

$$\begin{array}{llll} \text{P(same sex} \mid \text{identical)} = 1 & & \text{P(different sex} \mid \text{identical)} = 0 \\ \text{P(same sex} \mid \text{fraternal)} = 1/2 & & \text{P(different sex} \mid \text{fraternal)} = 1/2 \end{array}$$

The data is 'the twins are the same sex'. We find the answer with an update table

hyp.	prior	likelihood	Bayes numer.	posterior
identical	1/4	1	1/4	2/5
fraternal	3/4	1/2	3/8	3/5
Tot.	1		5/8	1

So $\boxed{\text{P(identical} \mid \text{same sex)} = 2/5 = 0.4}$.

(b) *Find the posterior odds the twins are identical. Do this by multiplying the prior odds by the Bayes factor (likelihood ratio). Check this by computing the odds directly from your answer to part (a).*

Solution: The prior odds $O(\text{identical}) = \dfrac{P(\text{identical})}{P(\text{fraternal})} = \dfrac{1/4}{3/4} = 1/3$.

The Bayes factor $BF = \dfrac{P(\text{same sex}\,|\,\text{identical})}{P(\text{same sex}|\text{fraternal})} = \dfrac{1}{1/2} = 2$.

So, posterior odds $O(\text{identical}\,|\,\text{same sex}) = O(\text{identical}) \cdot BF = \dfrac{1}{3} \cdot 2 = \boxed{\dfrac{2}{3}.}$

In part (a) we found the posterior odds $O(\text{identical}\,|\,\text{same sex}) = \dfrac{2/5}{3/5} = \dfrac{2}{3}$. This is the same as above.

Problem 11. Dice.

You have a drawer full of 4, 6, 8, 12 and 20-sided dice. You suspect that they are in proportion 1:2:10:2:1. Your friend picks one at random and rolls it twice getting 5 both times.

(a) *What is the probability your friend picked the 8-sided die?*

Solution: The data is 5. Let H_n be the hypothesis the die is n-sided. Here is the update table.

hyp.	prior	likelihood	Bayes numer.	posterior
H_4	1	0	0	0
H_6	2	$(1/6)^2$	2/36	0.243457
H_8	10	$(1/8)^2$	10/64	0.684723
H_{12}	2	$(1/12)^2$	2/144	0.060864
H_{20}	1	$(1/20)^2$	1/400	0.010956
Tot.	16		0.22819	1

So $P(H_8|\text{data}) = 0.685$.

(b) *(i) What is the probability the next roll will be a 5?*

(ii) What is the probability the next roll will be a 15?

Solution: We are asked for posterior predictive probabilities. Let x be the value of the next roll. We have to compute the total probability

$$p(x|\text{data}) = \sum p(x|H)p(H|\text{data}) = \sum \text{likelihood} \times \text{posterior}.$$

The sum is over all hypotheses. We can organize the calculation in a table where we multiply the posterior column by the appropriate likelihood column. The total posterior predictive probability is the sum of the product column.

hyp.	posterior to data	likelihood (i) $x = 5$	post. to (i)	likelihood (ii) $x = 15$	post. to (ii)
H_4	0	0	0	0	0
H_6	0.243457	1/6	0.04058	0	0
H_8	0.684723	1/8	0.08559	0	0
H_{12}	0.060864	1/12	0.00507	0	0
H_{20}	0.010956	1/20	0.00055	1/20	0.00055
Tot.	0.22819		0.13179		0.00055

So, (i) $p(x = 5|\text{data}) = 0.132$ and (ii) $p(x = 15|\text{data}) = 0.00055$.

Problem 12.
Sameer has two coins: one fair coin and one biased coin which lands heads with probability 3/4. He picks one coin at random (50-50) and flips it repeatedly until he gets a tails.

Assume that he observes 3 heads before the first tails.

(a) *What are the prior and posterior odds for the fair coin?*

Solution: The prior odds are $O(\text{fair}) = \dfrac{P(\text{fair})}{P(\text{not fair})} = \dfrac{1/2}{1/2} = 1.$

The posterior odds are the product of the prior odds and the Bayes factor (likelihood ratio). The data is $HHHT$. So, the Bayes factor is

$$BF = \frac{P(\text{data}\,|\,\text{fair})}{P(\text{data}\,|\,\text{unfair})} = \frac{(1/2)^3(1/2)}{(3/4)^3(1/4)} = \frac{16}{27}.$$

So the posterior odds $O(\text{fair}\,|\,\text{not fair}) = O(\text{fair}) \cdot BF = 1 \cdot \dfrac{16}{27} = \dfrac{16}{27} \approx 0.593.$

(b) *What are the prior and posterior predictive probabilities of heads on the next flip? Here prior predictive means prior to considering the data of the first four flips.*

Solution: The prior predictive probability of heads is

$$P(\text{fair})P(\text{heads}\,|\,\text{fair}) + P(3/4 \text{ coin})P(\text{heads}\,|\,3/4 \text{ coin}) = 0.5 \cdot 0.5 + 0.5 \cdot 0.75 = 0.625$$

The posterior predictive probability of heads is

$$P(\text{fair}\,|\,\text{data})P(\text{heads}\,|\,\text{fair}) + P(3/4 \text{ coin}\,|\,\text{data})P(\text{heads}\,|\,3/4 \text{ coin})$$

From part (a), we have the posterior odds the coin is fair are 16/27. This tells us the posterior probabilities are

$$P(\text{fair}\,|\,\text{data}) = \frac{16}{43}, \qquad P(\text{unfair}\,|\,\text{data}) = \frac{27}{43}$$

In this problem, the coin is unfair is the same as the coin has a 0.75 probability of heads. So, the posterior predictive probability of heads is $\dfrac{16}{43} \cdot 0.5 + \dfrac{27}{43} \cdot 0.75 \approx 0.657$

Method 2: We can also find the posterior probabilities needed to make the prediction using a Bayesian update table.

Let θ be the probability of the selected coin landing on heads. We have two hypothese: $\theta = 1/2$ and $\theta = 3/4$

Hyp.	Prior	Likelihood	Bayes numer.	Posterior
$\theta = 1/2$	$1/2$	$(1/2)^3(1/2)$	$1/2^5$	$16/43$
$\theta = 3/4$	$1/2$	$(3/4)^3(1/4)$	$3^3/(2 \cdot 4^4)$	$27/43$
Total	1	–	$43/512$	1

These are the same probabilities we got using odds. So, they will give the same posterior prediction for heads on the next toss.

6 Bayesian Updating: continuous prior, discrete likelihood

Problem 13. *Peter and Jerry disagree over whether 18.05 students prefer Bayesian or frequentist statistics. They decide to pick a random sample of 10 students from the class and get Shelby to ask each student which they prefer. They agree to start with a prior $f(\theta) \sim \text{Beta}(2,2)$, where θ is the percent that prefer Bayesian.*

(a) *Let x_1 be the number of people in the sample who prefer Bayesian statistics. What is the pmf of x_1?*

Solution: $x_1 \sim \text{Bin}(10, \theta)$.

(b) *Compute the posterior distribution of θ given $x_1 = 6$.*

Solution: We have prior:

$$f(\theta) = c_1\theta(1-\theta)$$

and likelihood:

$$p(x_1 = 6 \,|\, \theta) = c_2\theta^6(1-\theta)^4, \qquad \text{where} \qquad c_2 = \binom{10}{6}.$$

The Bayes numerator is $f(\theta)p(x_1|\theta) = c_1c_2\theta^7(1-\theta)^5$. So the normalized posterior is

$$f(\theta|x_1) = c_3\theta^7(1-\theta)^5$$

Since the posterior has the form of a Beta(8, 6) distribution it must be a Beta(8, 6) distribution. We can look up the normalizing coefficient $c_3 = \frac{13!}{7!\,5!}$.

(c) *Use R to compute 50% and 90% probability intervals for θ. Center the intervals so that the leftover probability in both tails is the same.*

Solution: The 50% interval is

```
[qbeta(0.25,8,6),  qbeta(0.75,8,6)] = [0.48330,  0.66319]
```

The 90% interval is

```
[qbeta(0.05,8,6),  qbeta(0.95,8,6)] = [0.35480,  0.77604]
```

(d) *The maximum a posteriori (MAP) estimate of θ (the peak of the posterior) is given by $\hat{\theta} = 7/12$, leading Jerry to concede that a majority of students are Bayesians. In light of your answer to part (c) does Jerry have a strong case?*

Solution: If the majority prefer Bayes then $\theta > 0.5$. Since the 50% interval includes $\theta < 0.5$ and the 90% interval covers a lot of $\theta < 0.5$ we don't have a strong case that $\theta > 0.5$.

As a further test we compute $P(\theta < 0.5|x_1) =$ `pbeta(0.5,8,6) = 0.29053`. So there is still a 29% posterior probability that the majority prefers frequentist statistics.

(e) *They decide to get another sample of 10 students and ask Neil to poll them. Write down in detail the expression for the posterior predictive probability that the majority of the second sample prefer Bayesian statistics. The result will be an integral with several terms. Don't bother computing the integral.*

Solution: Let x_2 be the result of the second poll. We want $p(x_2 > 5|x_1)$. We can compute this using the law of total probability:

$$p(x_2 > 5|x_1) = \int_0^1 p(x_2 > 5|\theta)p(\theta|x_1)\,d\theta.$$

The two factors in the integral are:

$$\begin{aligned} p(x_2 > 5|\theta) &= \binom{10}{6}\theta^6(1-\theta)^4 + \binom{10}{7}\theta^7(1-\theta)^3 + \binom{10}{8}\theta^8(1-\theta)^2 \\ &\quad + \binom{10}{9}\theta^9(1-\theta)^1 + \binom{10}{10}\theta^{10}(1-\theta)^0 \\ p(\theta|x_1) &= \frac{13!}{7!5!}\theta^7(1-\theta)^5 \end{aligned}$$

This can be computed exactly or numerically in R using the `integrate()` function. The answer is $P(x_2 > 5\,|x_1 = 6) = 0.5521$.

Problem 14. Coins

We have a 'bent' coin with an unknown probability θ of heads. Assume the following:

- *Prior for the value of θ: $f(\theta) = 2(1-\theta)$ on $[0,1]$.*
- *Data: toss once and get tails.*

(a) *Find the posterior pdf to this data.*

Solution: Here's the update table.

hypoth.	prior	likelihood	Bayes numerator	posterior
θ	$2(1-\theta)\,d\theta$	$1-\theta$	$2(1-\theta)^2\,d\theta$	$3(1-\theta)^2\,d\theta$
Total	1		$T = \int_0^1 2(1-\theta)^2\,d\theta = 2/3$	1

Posterior pdf: $f(\theta|x) = 3(1-\theta)^2$. (Graph below.)

Note: We don't really need to compute T. Once we know the posterior density is of the form $c\theta^2$ we only have to find the value of c which makes it have total probability 1.

(b) *Suppose you toss again and get tails. Update your posterior from part (a) using this data.*

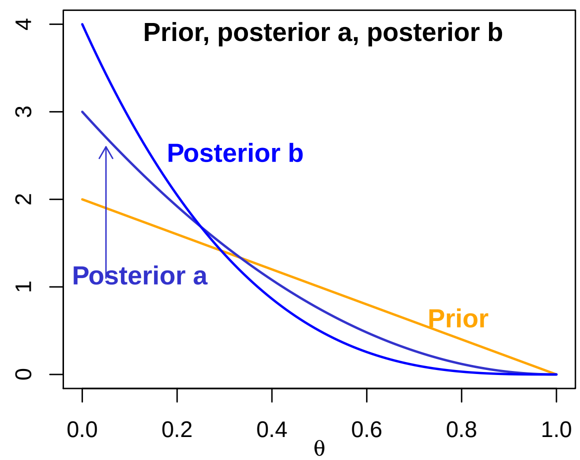
Solution: We use the posterior from part (a) as the prior for this part. Here's the table.

hypoth.	prior	likelihood	Bayes numerator	posterior
θ	$3(1-\theta)^2\,d\theta$	$1-\theta$	$3(1-\theta)^3, d\theta$	$4(1-\theta)^3\,d\theta$
Total	1		$\int_0^1 3(1-\theta)^3\,d\theta = 3/4$	1

Posterior pdf: $f(\theta|x) = (1-\theta)^3$.

(c) *On one set of axes graph the prior and the posteriors from parts (a) and (b).*

(c) Solution: Here is the plot of the prior and the two posteriors.



Problem 15. Take your medicine

A lab has an experimental treatment for a disease. The treatment will cure an unknown fraction θ of the patients it's used on. Because it is brand new, they have no idea what θ is, so they use a flat prior $f(\theta) = 1$.

In a small preliminary study, the treatment cured 16 out of 20 patients.

Use this data to find the posterior pdf for θ.

Write an integral formula for the normalizing factor (total probability of the data), but do not compute it. Call its value T and give the posterior pdf in terms of T.

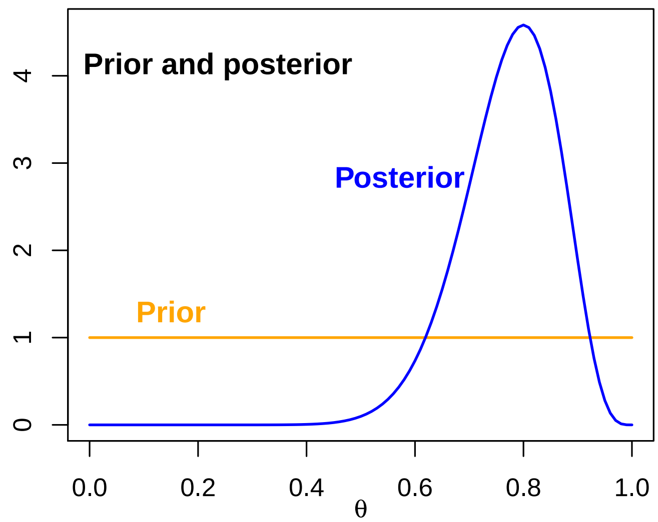
Solution: Here's the update table.

hypoth.	prior	likelihood	Bayes numerator	posterior
θ	$1 \cdot d\theta$	$\binom{20}{16}\theta^{16}(1-\theta)^4$	$\binom{20}{16}\theta^{16}(1-\theta)^{14}\,d\theta$	$c\theta^{16}(1-\theta)^4\,d\theta$
Total	1		$T = \int_0^1 \binom{20}{16}\theta^{16}(1-\theta)^4\,d\theta$	1

So, $f(\theta|x) = c\theta^{16}(1-\theta)^4$, where $c = \dfrac{\binom{20}{16}}{T}$.

A computation (or Wikipedia) would show $c = \dfrac{21}{16!\,4!}$

(This is called a Beta distribution.) Here is a plot of the prior and posterior



7 Bayesian Updating: normal-normal conjugate pairs

Problem 16. *Suppose that you have a cable whose exact length is θ. You have a ruler with known error normally distributed with mean 0 and variance 10^{-4}. Using this ruler, you measure your cable, and the resulting measurement x is distributed as $N(\theta, 10^{-4})$.*

(a) *Suppose your prior on the length of the cable is $\theta \sim N(9,1)$. If you then measure $x = 10$, what is your posterior pdf for θ?*

Solution: We have $\mu_{\text{prior}} = 9, \sigma^2_{\text{prior}} = 1$ and $\sigma^2 = 10^{-4}$. The normal-normal updating formulas are

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

So we compute $a = 1/1, b = 10000$, $\sigma^2_{post} = 1/(a+b) = 1/10001 \approx 9.999 \times 10^{-5}$ and

$$\mu_{\text{post}} = \frac{a\mu_{\text{prior}} + bx}{a+b} = \frac{100009}{10001} \approx 9.9999$$

So we have posterior distribution $f(\theta|x = 10) \sim N(9.9999, 9.999 \times 10^{-5})$.

(b) *With the same prior as in part (a), compute the total number of measurements needed so that the posterior variance of θ is less than 10^{-6}.*

Solution: We have $\sigma^2_{\text{prior}} = 1$ and $\sigma^2 = 10^{-4}$. The posterior variance of θ given observations $x_1, \ldots, x_n$ is given by

$$\frac{1}{\frac{1}{\sigma^2_{\text{prior}}} + \frac{n}{\sigma^2}} = \frac{1}{1 + n \cdot 10^4}$$

We wish to find n such that the above quantity is less than 10^{-6}. It is not hard to see that $n = 100$ is the smallest value such that this is true.

8 NHST

Problem 17. ***z-test***
Suppose we have 49 data points with sample mean 6.25 and sample variance 12. We want to test the following hypotheses

H_0: the data is drawn from a $N(4, 10^2)$ distribution.
H_A: the data is drawn from $N(\mu, 10^2)$ where $\mu \neq 4$.

(a) *Test for significance at the $\alpha = 0.05$ level. Use the tables at the end of this file to compute p-values.*

Solution: Our z-statistic is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{6.25 - 4}{10/7} = 1.575$$

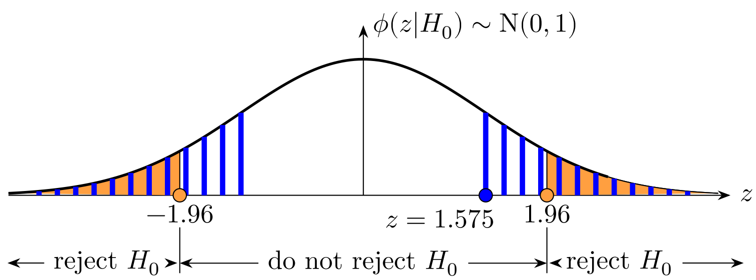
Under the null hypothesis $z \sim \mathrm{N}(0,1)$ The two-sided p-value is

$$p = 2 \times P(Z > 1.575) = 2 \times 0.0576 = 0.1152$$

The probability was computed from the z-table. We interpolated between $z = 1.57$ and $z = 1.58$ Because $p > \alpha$ we do not reject H_0.

(b) *Draw a picture showing the null pdf, the rejection region and the area used to compute the p-value.*

Solution: The null pdf is standard normal as shown. The orange shaded area is over the rejection region. The area used to compute significance is shown in orange. The area used to compute the p-value is shown with blue stripes. Note, the z-statistic outside the rejection region corresponds to the blue completely covering the orange.



Problem 18. ***t-test***
Suppose we have 49 data points with sample mean 6.25 and sample variance 36. We want to test the following hypotheses:

(a) *H_0: the data is drawn from $N(4, \sigma^2)$, where σ is unknown.*
H_A: the data is drawn from $N(\mu, \sigma^2)$ where $\mu \neq 4$.

Test for significance at the $\alpha = 0.05$ level. Use the t-table to find the p value.

Solution: Our t-statistic is

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{6.25 - 4}{6/7} = 2.625$$

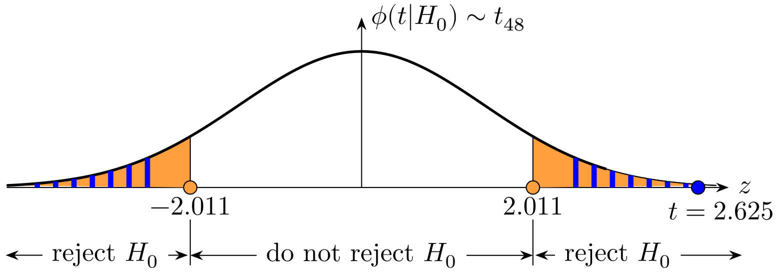
Under the null hypothesis $t \sim t_{48}$. Using the t-table we find the two-sided p-value is

$$p = 2 \times P(t > 2.625) < 2 \times 0.005 = 0.01$$

Because $p < \alpha$ we reject H_0.

(b) *Draw a picture showing the null pdf, the rejection region and the area used to compute the p-value for part (a).*

Solution: The null pdf is a t-distribution as shown. The rejection region is shown. The area used to compute significance is shown in orange. The area used to compute the p-value is shown with blue stripes. Note, the t-statistic is inside the rejection region corresponds. This corresponds to the orange completely covering the blue. The critical values for t_{48} we're looked up in the table.



Problem 19. *There are lots of good NHST problems in psets 7 and 8, the reading and in-class problems, including two-sample t test, chi-square, ANOVA, and F-test for equal variance.*

Solution: See the psets 7 and 8.

Problem 20. Probability, MLE, goodness of fit

There was a multicenter test of the rate of success for a certain medical procedure. At each of the 60 centers the researchers tested 12 subjects and reported the number of successes.

(a) *Assume that θ is the probability of success for one patient and let x be the data from one center. What is the probability mass function of x?*

Solution: This is a binomial distribution. Let θ be the Bernoulli probability of success in one test.

$$p(x = k) = \binom{12}{k} \theta^k (1-\theta)^{12-k}, \text{ for } k = 0, 1, \ldots, 12.$$

(b) *Assume that the probability of success θ is the same at each center and the 60 centers produced data: $x_1, x_2, \ldots, x_{60}$. Find the MLE for θ. Write your answer in terms of $\bar{x}$*

Solution: The likelihood function for the combined data from all 60 centers is

$$\begin{aligned} p(x_1, x_2, \ldots, x_{60} \,|\, \theta) &= \binom{12}{x_1} \theta^{x_1} (1-\theta)^{12-x_1} \binom{12}{x_2} \theta^{x_2} (1-\theta)^{12-x_2} \cdots \binom{12}{x_{60}} \theta^{x_{60}} (1-\theta)^{12-x_{60}} \\ &= c\, \theta^{\sum x_i} (1-\theta)^{\sum 12 - x_i} \end{aligned}$$

To find the maximum we use the log likelihood. At the same time we make the substitution $60 \cdot \bar{x}$ for $\sum x_i$.

$$\ln(p(\text{data} \,|\, \theta)) = \ln(c) + 60\bar{x}\ln(\theta) + 60(12 - \bar{x})\ln(1 - \theta).$$

Now we set the derivative to 0:

$$\frac{d\,\ln(p(\text{data} \,|\, \theta))}{d\,\theta} = \frac{60\bar{x}}{\theta} - \frac{60(12 - \bar{x})}{1 - \theta} = 0.$$

Solving for θ we get

$$\hat{\theta} = \frac{\bar{x}}{12}.$$

Parts (c-e) use the following table which gives counts from 60 centers, e.g. $x = 2$ occurred in 17 out of 60 centers.

x	*0*	*1*	*2*	*3*	*4*	*5*
counts	*4*	*15*	*17*	*10*	*8*	*6*

Note, the possible values of x are 0 to 12. The table shows that $x > 5$ never occurred.

(c) *Compute $\bar{x}$ the average number of successes over the 60 centers.*

Solution: The sample mean is

$$\begin{aligned} \bar{x} &= \frac{\sum(\text{count} \times x)}{\sum \text{counts}} \\ &= \frac{4 \cdot 0 + 15 \cdot 1 + 17 \cdot 2 + 10 \cdot 3 + 8 \cdot 4 + 6 \cdot 5}{60} \\ &= 2.35 \end{aligned}$$

(d) *Assuming the probability of success at each center is the same, show that the MLE for θ is $\hat{\theta} = 0.1958$.*

Solution: Just plug $\bar{x} = 2.35$ into the formula from part (b): $\hat{\theta} = \bar{x}/12 = 2.35/12 = 0.1958$

(e) *Do a χ^2 goodness of fit to test the assumption that the probability of success is the same at each center. Find the p-value and use a significance level of 0.05.*

In this test the number of degrees of freedom is the number of bins - 2.

Solution: There were 60 trials in all. Our hypotheses are:

H_0 = the probability of success is the same at all centers. (This determines the probabilities of the counts in each cell of our table.)

H_A = the probabilities for the cell counts can be anything as long as they sum to 1, i.e. x follows an arbitrary multinomial distribution.

Using the the value for $\hat{\theta}$ in part (d) we have the table below. Here are some details of the computation

Since, in principle, x can take any values between 0 and 12, the last cell in the counts is really for $x \geq 5$. Thus, the probabilities for $x = 0, 1, 2, 3, 4$ are computed using R by $p(x) =$ `dbinom(x, 12, 0.1958)`. The last probability for $x \geq 5$ is computed by `sum(dbinom(5:12, 12, 0.1058))`. (It could have been computed using pbinom.)

The expected counts are just the probabilities times the number of centers, 60. The components of X^2 are computed using the formula $X_i^2 = (E_i - O_i)^2/E_i$.

x	0	1	2	3	4	≥ 5
prob	0.0731	0.2137	0.2863	0.2324	0.1273	0.0671
Observed	4	15	17	10	8	6
Expected	4.39	12.82	17.18	13.94	7.64	4.03
X_i^2	0.0344	0.3692	0.0018	1.1149	0.0170	0.9643

The χ^2 statistic is $X^2 = \sum X_i^2 = 2.502$. There are 6 cells, so 4 degrees of freedom. The p-value is

$p =$ `1-pchisq(2.502, 4) = 0.644`

With this p-value we do not reject H_0.

The reason the degrees of freedom is two less than the number of cells is that there are two constraints on assigning cell counts assuming H_A but consistent with the statistics used to compute the expected counts. They are the total number of observations = 60, and the grand mean $\overline{x} = 2.35$.