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 $x_1, x_2, \ldots, x_n$ 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 $\pm 5\%$.

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 $52\% \pm 5\%$.

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: $\overline{x} \pm \frac{1}{\sqrt{n}} = 0.55 \pm \frac{1}{\sqrt{40}} = 0.55 \pm 0.16 = 55\% \pm 16\%.$

For poll 2 we have
Point estimate: $\overline{x} = 190/400 = 0.475$

Confidence interval: $\overline{x} \pm \frac{1}{\sqrt{n}} = 0.475 \pm \frac{1}{\sqrt{400}} = 0.475 \pm 0.05 = 47.5\% \pm 5\%.$

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 $\mathrm{N}(0,1)$. 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.

n	nominal conf. $1-\alpha$	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)

n	nominal conf. $1-\alpha$	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 $\mathrm{N}(0,1)$.

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 $t_{0.025} = 2.09$.