Chapter 7: Interval Estimation

MATH 361: Probability & Statistics I

Prof. Katie Fitzgerald, PhD

Spring 2022

Contents

7.1 Confidence Intervals for Means

How good of an estimate is $\bar{x}$?

Setting the stage: we are interested in some mean in the population, μ, but we can't observe the whole population, so we estimate the population mean by the sample mean, $\bar{x}$.

μ is called a **population parameter** (some true value we want to know in the population), and $\bar{x}$ is called a **sample statistic**. Another term for sample statistic is **point estimate**.

An important question is: *How good of an estimate is $\bar{x}$?* We know it will vary from sample to sample, but how much? How close can we expect $\bar{x}$ to be to μ?

We use the **sampling distribution of $\bar{x}$** (e.g. $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ by the CLT) to quantify the uncertainty and create a range of plausible values for μ called a **confidence interval**.

Confidence interval analogy

A confidence interval provides a plausible range of values for μ. Why is this useful?

- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.
- We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.
- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Constructing a confidence interval

$E(X) = \mu$
$V(X) = \sigma^2$

Recall that Z-scores quantify distance from the mean, in standard deviation units

By the CLT, we know $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$, so $E(\bar{x}) = \mu_{\bar{x}} = \mu$ and $SD(\bar{x}) = \sigma_{\bar{x}} = \sigma/\sqrt{n}$

$SD(\bar{x}) = \sqrt{V(\bar{x})}$
$= \sqrt{\frac{\sigma^2}{n}}$

Therefore we can calculate a Z-score for the sample mean by

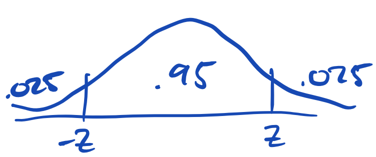
$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}},$$

$Z = \frac{X - \mu}{\sigma}$
$= \frac{\text{r.v} - \text{mean}}{\text{sd}}$

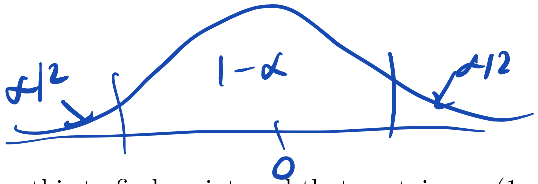
and we know that $Z \sim N(0, 1)$

Therefore $P(-z_{\alpha/2} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha$

Exercise What value of z gives $P(-z < Z < z) = 0.95$?



qnorm(.975) = 1.96
$P(-1.96 < Z < 1.96) = 0.95$
$1 - \alpha = .95$
$\alpha = .05$



We can use this to find an interval that contains μ, $(1-\alpha)100\%$ of the time.

The probability that the random interval

$$\left[\underbrace{\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)}_{lb}, \quad \underbrace{\bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)}_{ub}\right]$$

includes the unknown μ is $1-\alpha$.
We call the interval the $(1-\alpha)100\%$ **confidence interval**.
We say we are $(1-\alpha)100\%$ confident that our interval contains the true mean.

$\alpha = .05 \Rightarrow 95\%$ CI

WHY?:

$$P\left(-z_{\alpha/2} \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le \bar{x}-\mu \le z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1-\alpha$$

$$P\left(-\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} \le -\mu \le -\bar{x} + z_{\alpha/2}\sigma/\sqrt{n}\right) = 1-\alpha$$

$$P\left(\bar{x} + z_{\alpha/2}\sigma/\sqrt{n} \ge \mu \ge \bar{x} - z_{\alpha/2}\sigma/\sqrt{n}\right) = 1-\alpha$$

$$P\left(\underbrace{\bar{x} - z_{\alpha/2}\sigma/\sqrt{n}}_{lb} \le \mu \le \underbrace{\bar{x} + z_{\alpha/2}\sigma/\sqrt{n}}_{ub}\right) = 1-\alpha$$

What do we mean by "95% confident"?

Let's say we're trying to estimate the average amount of money parents spend per child on childcare in the US (in thousands of dollars). Suppose the true $\mu = 8.5$ and $\sigma = 1$. What would happen if we tried to estimate this from a random sample of 100 households (with children)?

Let's simulate! https://www.rossmanchance.com/applets/2021/confsim/ConfSim.html

Interpretation of a 95% confidence interval

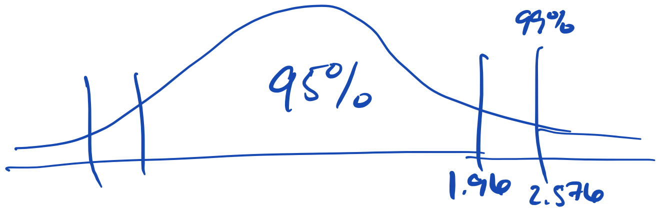
Correct interpretations:

- "We are 95% confident that the interval based on this sample contains the true mean."
- "We are 95% confident that the true mean falls between [lower bound] and [upper bound]."
- 95% of intervals constructed using this method will contain the true mean.

Incorrect interpretations:

- There is a 95% chance the parameter μ will be in the interval
  - This incorrectly implies μ is random
- "We are 95% confident that the mean in our sample is between [lower bound] and [upper bound]."
  - There is no uncertainty about our sample - the uncertainty is in making inference about the population)
- "95% of X values will fall between [lower bound] and [upper bound]"
  - This is an interval about the population mean μ, not the observed data points

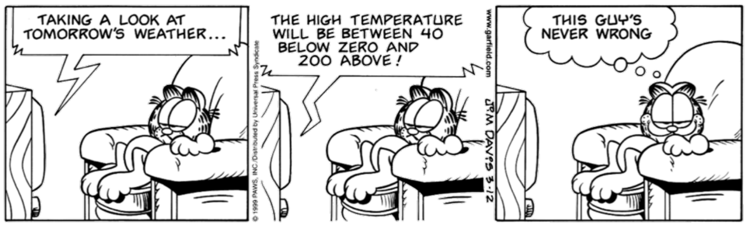
$\bar{x} \pm z_{\alpha/2} \, \sigma/\sqrt{n}$



Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

What are the drawbacks to using a wider interval?



$s^2 = \dfrac{\sum (x_i - \bar{x})^2}{n-1}$ $\quad s = \sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n-1}}$

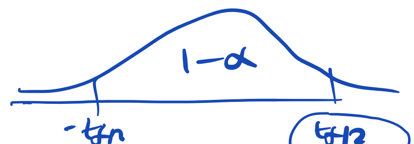
When σ is unknown

In practice, σ is also unknown (in addition to μ), and we have to estimate it by s. This introduces additional uncertainty into the estimation of our confidence interval.

We use the t-distribution to accommodate this extra uncertainty (since the t-distribution has fatter tails).

Recall that

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(df = n-1),$$



qt(.975, df=n-1)

so

$$P\left(-t_{\alpha/2}(n-1) < \frac{\bar{x} - \mu}{s/\sqrt{n}} < t_{\alpha/2}(n-1)\right) = 1 - \alpha$$

and we can construct a $(1-\alpha)100\%$ confidence interval by

$$\left[\underbrace{\bar{X} - \overset{z_{\alpha/2}}{t_{\alpha/2}(n-1)}\left(\frac{s}{\sqrt{n}}\right)}, \quad \underbrace{\bar{X} + \overset{z_{\alpha/2}}{t_{\alpha/2}(n-1)}\left(\frac{s}{\sqrt{n}}\right)}\right]$$

Not multiplication

$$\left[\bar{x} - z_{\alpha/2}\,\sigma/\sqrt{n}, \quad \bar{x} + z_{\alpha/2}\,\sigma/\sqrt{n}\right]$$

$\bar{x} \pm t_{\alpha/2}\, s/\sqrt{n}$

$\dfrac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$

$\bar{x} \pm z_{\alpha/2}\, \sigma/\sqrt{n}$

$\dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$\dfrac{x - E(x)}{\sqrt{V(x)}} \sim N(0,1)$

qnorm( )
qt( )

point estimate $\pm$ critical value $*$ $\sqrt{V(\text{pointest})}$

7.3 Confidence Intervals for Proportions

Sometimes our parameter of interest is a **proportion** rather than a mean. For example, perhaps we want to know the proportion of all US adults who support the death penalty.

A proportion comes from a categorical variable, where there are two categories: success or failure. In the population, we denote the proportion by p, and it is defined as

$$p = \frac{\text{number of successes in the population}}{\text{total number of units in the population}}$$

In the sample, we denote the proportion by $\hat{p}$, and it is defined as

$$\hat{p} = \frac{\text{number of successes in the sample}}{\text{total number of units in the sample}} = \frac{Y}{n}$$

Note that Y is the number of successes in n bernoulli trials, so $Y \sim binomial(n, p)$

$\bar{X} \sim N(\mu, \sigma^2/n)$

$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

How good of an estimate is $\hat{p}$?

Just like the sample mean $\bar{x}$ is a point estimate of the population parameter μ, the sample proportion $\hat{p}$ is a point estimate of the population parameter p.

How good of an estimate is $\hat{p}$? We know it will vary from sample to sample, but how much? How close can we expect $\hat{p}$ to be to p?

We again need to consider the **sampling distribution** (this time of of $\hat{p}$), and use it to quantify the uncertainty and create a range of plausible values (confidence interval) for p.

Sampling distribution of $\hat{p}$

Recall the mean and variance of the binomial distribution. If $Y \sim binomial(n, p)$, then

$\hat{p} = \frac{Y}{n}$

$$E(Y) = np$$
$$V(Y) = np(1-p)$$

Chapter 2

We can use this to find $E(\hat{p})$ and $V(\hat{p})$

$E(\hat{p}) = E\left(\frac{Y}{n}\right) = \frac{1}{n}E(Y) = \frac{1}{n}np = \underline{p}$ unbiased

$V(\hat{p}) = V\left(\frac{Y}{n}\right) = \frac{1}{n^2}V(Y) = \frac{1}{n^2}np(1-p) = \underline{\frac{p(1-p)}{n}}$

CLT: $\frac{X - E(X)}{\sqrt{V(X)}} \sim N(0,1)$ $\qquad \frac{\hat{p} - E(\hat{p})}{\sqrt{V(\hat{p})}}$

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

n is "large enough"

By the Central Limit Theorem,

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

as $n \to \infty$. It turns out that the normal distribution is a good approximation so long as $np \geq 5$ and $np(1-p) \geq 5$. That is, as long as there are at least 5 successes and at least 5 failures. Note, if you are trying to estimate a rare event (i.e. a small proportion), you will need a larger sample size.

Constructing a confidence interval

$$\bar{x} \pm z_{\alpha/2}\sqrt{V(\bar{x})}$$

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

In the same way as before,

$$P(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}) = 1 - \alpha$$

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \quad \leq \quad p \quad \leq \quad \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

$V(\hat{p})$
$\parallel$

So, the probability that the random interval

$$\left[\hat{p} - z_{\alpha/2}\left(\sqrt{\frac{p(1-p)}{n}}\right), \quad \hat{p} + z_{\alpha/2}\left(\sqrt{\frac{p(1-p)}{n}}\right)\right]$$

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

point estimate ± critical value $\sqrt{V(\ )}$

includes the unknown p is $1 - \alpha$.

Problem: we don't know p

It turns out that substituting $\hat{p}$ in the formula for the confidence interval still gives a pretty good approximation. That is,

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \leq \quad p \quad \leq \quad \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1 - \alpha$$

Software uses a somewhat more complicated approximation that is more accurate, but for anything we do "by hand" this approximation is sufficient.

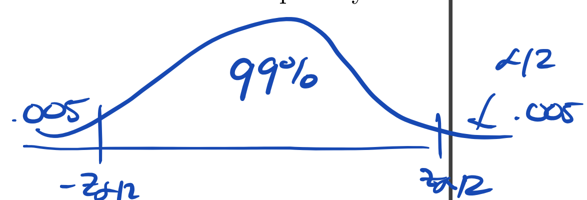
Example: General Social Survey
According to Wikipedia, the General Social Survey (GSS) is a sociological survey created and regularly collected since 1972 by the National Opinion Research Center (NORC) at the University of Chicago. It is funded by the National Science Foundation. The GSS collects information and keeps a historical record of the concerns, experiences, attitudes, and practices of residents of the United States.

Since 1972, the GSS has been monitoring societal change and studying the growing complexity of American society. It is one of the most influential studies in the social sciences, and is frequently referenced in leading publications, including The New York Times, The Wall Street Journal, and the Associated Press.

In 2018, one of the questions asked was “Do you favor or oppose the death penalty for persons convicted of murder?” 1385 responded “Favor” and 808 responded “Oppose”. The GSS is a random and representative sample of all adults (18+) living in households in the US, so we can use this data to make inference about the beliefs of that entire population.

Construct a 99% confidence interval for the proportion of US adults who favor the death penalty for persons convicted of murder.

$$\hat{p} \pm z_{\alpha/2}\underbrace{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$



$n = 1385 + 808 = 2193$

$\frac{x}{n}$ $\hat{p} = \frac{1385}{1385+808} = 0.6316$

$z_{\alpha/2}$ = qnorm(.995) = 2.576

$$0.6316 \pm 2.576\sqrt{\frac{.6316(1-.6316)}{2193}}$$

$[0.605 \; , \; 0.658]$

We are 99% confident that the true proportion of US adults who favor the death penalty is between 0.605 and 0.658

Differences in proportions

Often our parameter of interest is not a single proportion but rather a difference in proportions, $p_1 - p_2$.

Use the Central Limit Theorem to construct an approximate $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$.
Hint: first define a relevant point estimate, find its expected value and variance, and use those results to define a standardized estimate $Z \sim N(0,1)$. You can assume that the two proportions are independent.

$\hat{p}_1 = \frac{Y_1}{n_1}$

$\hat{p}_2 = \frac{Y_2}{n_2}$

$\underline{\hat{p}_1 - \hat{p}_2}$ point estimate $(\bar{x}, \hat{p})$

$E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = \underline{p_1 - p_2}$

$Y_1 \sim \text{binomial}(n_1, p_1) \qquad E(Y_1) = n_1 p_1$

$Y_2 \sim \text{binomial}(n_2, p_2) \qquad V(Y_1) = n_1 p_1(1-p_1)$

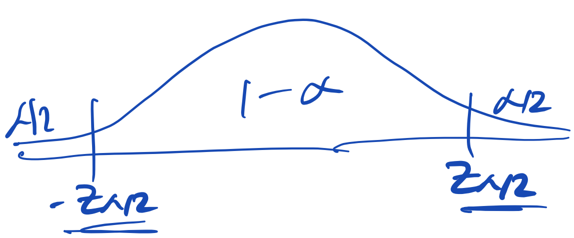
$\frac{X - E(X)}{\sqrt{V(X)}}$

$V(\hat{p}_1 - \hat{p}_2) = V(\hat{p}_1) + V(\hat{p}_2) \cancel{+ 2\,\text{cov}(\ )}$

$= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim \underline{N(0,1)}$$

$$P\left(-Z_{\alpha/2} \le \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{V(\hat{p}_1 - \hat{p}_2)}} \le Z_{\alpha/2}\right) = 1 - \alpha$$



point est $\pm$ C.V $*\sqrt{V(\text{point est})}$

$$\hat{p}_1 - \hat{p}_2 \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

You are doing the work of a statistician!

In the real world, new estimation scenarios often come up when trying to answer social or science related questions. Statisticians have to develop new estimators, determine their properties and how well they can estimate an unknown parameter of interest.

An example from my research: in education policy, there have been over 400 government-funded research studies to evaluate what works in education. Many of these have used the same outcome measure to measure student learning, so we now have many estimates $s_1^2, s_2^2, ...s_n^2$ that estimate the same σ^2. In any single study, s_i^2 is not a great estimate of σ^2; it's expensive and logistically difficult to recruit more schools into education research studies!

Can we synthesize the s_i^2s across studies to improve our estimates of σ^2, and therefore achieve narrower (more informative) confidence intervals for our estimates of how well an intervention works?

This involves developing appropriate point estimates, determining their distributions, and investigating which ones are optimal under what scenarios.

A more important (life-saving) example