5. Statistical Inference: Estimation

LPO.8800: Statistical Methods in Education Research

Sean P. Corcoran

Last time

- *Sampling distributions* (e.g., of $\bar{x}$, $\hat{\pi}$)
- *Standard error*: a measure of variability in a sampling distribution
- Central Limit Theorem: tells us the distribution of $\bar{x}$ when x is normally distributed or otherwise (and n is large)
- Law of Large Numbers: variation in sampling distribution shrinks to zero as sample size gets large
- $\bar{x}$ is *unbiased*
- *Efficiency* of estimators
- Simulating sampling distributions in Stata

Estimating population parameters

The goal of inferential statistics is to make inferences about **population parameters** (e.g., μ, σ^2) using **sample statistics** (e.g., $\bar{x}$, s^2).

- Estimators vs. estimates
  - The term **estimator** refers to the method, or statistic (e.g. $\bar{x}$), used to estimate a population parameter (e.g. μ). Often represented by a "hat" over the parameter being estimated, such as $\hat{\mu}$ or $\hat{\sigma}$.
  - When used as a noun, an **estimate** refers to a specific realization of a sample statistic. An *estimator* applied to data gives you an *estimate*.
- Point vs. interval estimation
  - **Point estimation** is the process of estimating a specific parameter.
  - **Interval estimation** is the process of estimating a *range* of likely values for the parameter. Takes into account variability in the sampling distribution of the estimator.

Point estimates

Lecture 4 introduced the sampling distribution of $\bar{x}$, a point estimator of the population mean μ. We saw that:

- The sampling distribution of $\bar{x}$ has a mean of μ and standard error $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- According to the Central Limit Theorem:
  - If x is distributed *normal*: the sampling distribution of $\bar{x}$ is normal.
  - If x is *not* distributed normal: the sampling distribution of $\bar{x}$ is *approximately* normal if n is sufficiently large.

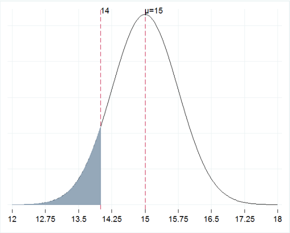
Example from Lecture 4

Suppose we plan to draw a random sample of 16 from the population of x and compute the sample mean $\bar{x}$. We know x is normally distributed with μ=15 and σ=3. What will the sampling distribution of $\bar{x}$ look like? From the CLT we know:

- $\bar{x}$ will have a normal distribution
- $\bar{x}$ will have a mean of 15
- $\bar{x}$ will have a standard error $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{16}} = 0.75$

Example from Lecture 4

Knowing its sampling distribution, we can make statements about how likely particular realized values of $\bar{x}$ are. For example, in the above scenario, what is the probability we will draw a random sample with an $\bar{x}$ of 14 or lower?

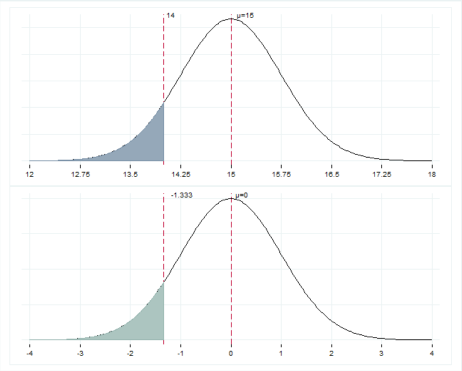


Example from Lecture 4

Putting in z-score (standard normal) terms:

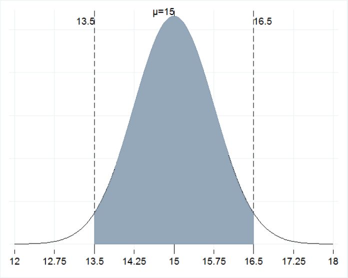
- The value $\bar{x} = 14$ is $z = \frac{14-15}{0.75} = -1.33$ standard errors below the population mean.
- Thus: $\Pr(\bar{x} \leq 14) = \Pr(z \leq -1.33)$
- Stata: `display normal(-1.33)` $= 0.092$

Example from Lecture 4



Example from Lecture 4

What is the probability we will draw a random sample with an $\bar{x}$ **between** 13.5 and 16.5?

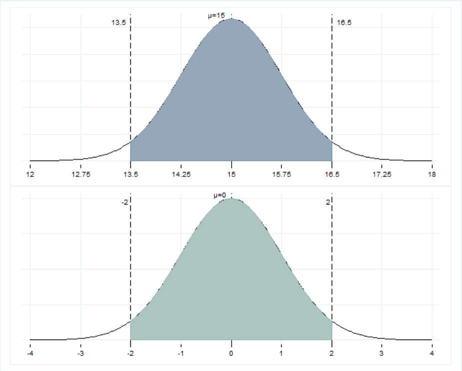


Example from Lecture 4

Putting in z-score (standard normal) terms:

- The value $\bar{x} = 13.5$ is $z = \frac{13.5-15}{0.75} = -2$ standard errors below the population mean.
- The value $\bar{x} = 16.5$ is $z = \frac{16.5-15}{0.75} = +2$ standard errors above the population mean.
- Thus $\Pr(13.5 \leq \bar{x} \leq 16.5) = \Pr(-2 \leq z \leq +2)$
- Stata: `display normal(2) - normal(-2)` $= 0.954$ (about 95%)

Example from Lecture 4



Confidence intervals

The goal of interval estimation is to construct an interval that will—*under repeated samples*—contain the population parameter $(1-\alpha)\%$ of the time

- $(1-\alpha)\%$ is the **confidence level**, the percentage of times in repeated samples that the interval will contain the population parameter:
  - 95% confidence level ($\alpha = 0.05$)
  - 99% confidence level ($\alpha = 0.01$)
  - 90% confidence level ($\alpha = 0.10$)
- α is the **error probability**, the proportion of times in repeated samples that the interval will not contain the population parameter. It is a measure of our tolerance for error (more on this later). We typically choose this.
- An interval estimate with a specified confidence level is a **confidence interval**.

Confidence intervals

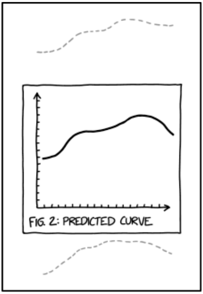
All else equal, the tradeoff of *narrower* (more **precise**) confidence intervals is a greater probability that the interval will *not* contain the true population parameter.

- A *low confidence level* is associated with narrower (more *precise*) confidence intervals, but a higher likelihood of error—i.e., computing an interval that does not contain the population parameter.
- A *high confidence level* is associated with wider (less *precise*) confidence intervals, but a lower likelihood of error.
- This terminology can be confusing, because "precise" does not mean more *accurate*, but rather a narrower range of values.

Ideally, you'd like <u>both</u> narrow CIs *and* a high confidence level.

Confidence intervals

Another approach...



SCIENCE TIP: IF YOUR MODEL IS BAD ENOUGH, THE CONFIDENCE INTERVALS WILL FALL OUTSIDE THE PRINTABLE AREA.

Confidence interval for μ

Under certain assumptions, we saw that $\bar{x}$ will fall within two standard errors of the true population mean roughly 95% of the time. More accurately, $\bar{x}$ will fall within **1.96** standard errors of the true population mean 95% of the time. If this is the case, then the interval:

$$\bar{x} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right), \bar{x} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

will include the population mean (μ) 95% of the time. This is known as a **95% confidence interval for μ.** By the same logic, this interval will *not* contain the population mean 5% of the time ($\alpha = 0.05$).

Confidence interval for μ

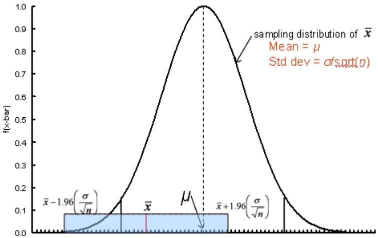


Figure: Confidence interval around $\bar{x}$

Confidence interval for μ

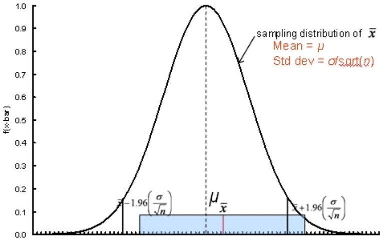


Figure: Confidence interval around $\bar{x}$

Confidence interval for μ

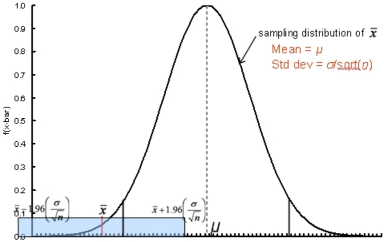


Figure: Confidence interval around $\bar{x}$

Example 1a

A sample of 121 women is taken (n=121) where the mean height is estimated to be $\bar{x} = 64$ inches. It is known that $\sigma = 3$. Construct a 95% confidence interval for the true population mean height (μ).

$$64 \pm 1.96 \left(\frac{3}{\sqrt{121}} \right) = (63.47, 64.53)$$

This is our interval estimate of the range of likely values for μ. For later reference, note the width of this interval is 1.06 inches.

Confidence interval for μ

Changing the confidence level is only a matter of changing the z value used in the interval estimator. For a confidence level of $(1 - \alpha)$%, the appropriate z value is the one for which there is a probability $\alpha/2$ of exceeding. Then the $(1 - \alpha)$% confidence interval is:

$$\bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

Confidence interval for μ

Choosing the appropriate z-values:

- There is a 0.005 probability that z falls above 2.576 and a 0.005 probability that z falls below -2.576. Thus **2.576** is the z value used for a 99% confidence interval ($\alpha = 0.01$).

- There is a 0.05 probability that z falls above 1.645 and a 0.05 probability that z falls below -1.645. Thus **1.645** is the z value used for a 90% confidence interval ($\alpha = 0.10$).

```
display (-1)*invnormal(0.005)

display (-1)*invnormal(0.05)
```


Example 1b

A sample of 121 women is taken (n=121) where the mean height is estimated to be $\bar{x} = 64$ inches. It is known that $\sigma = 3$. Construct a **99%** confidence interval for the true population mean height (μ).

$$64 \pm \mathbf{2.576} \left(\frac{3}{\sqrt{121}} \right) = (63.29, 64.71)$$

Note this is a width of 1.42 inches, a wider range of values than the 95% confidence interval.

Example 1c

A sample of 121 women is taken (n=121) where the mean height is estimated to be $\bar{x} = 64$ inches. It is known that $\sigma = 3$. Construct a **90%** confidence interval for the true population mean height (μ).

$$64 \pm \mathbf{1.645} \left(\frac{3}{\sqrt{121}} \right) = (63.55, 64.45)$$

Note this has a width of 0.9 inches, a narrower range of values than the 95% confidence interval.

Confidence intervals in Stata

Confidence intervals for the mean are straightforward in Stata with the `mean` command. (In this example we are using s as an estimator for σ, which was assumed to be known in the earlier examples. More on this later). For example, using the NELS data the 95% and 99% confidence intervals for mean 8th grade math achievement can be found using:

- `mean` *achmat08*
- `mean` *achmat08*, `level(99)`

A 95% confidence interval is the default in Stata, but the confidence level can be changed with the `level` option.

Example 2

From a sample of 3,650 days, the average daily high temperature in Richmond, VA is estimated to be 69 degrees. It is known that $\sigma = 17.5$. Construct a **95%** confidence interval for the true population mean temperature in Richmond (μ).

$$69 \pm 1.96 \left(\frac{17.5}{\sqrt{3650}} \right) = (68.43, 69.57)$$

Example 3

IQ scores are scaled to have a mean of 100 and a standard deviation of 15. Suppose in a sample of 42 people, $\bar{x} = 103$. Construct a **95%** confidence interval for mean IQ.

$$103 \pm 1.96 \left(\frac{15}{\sqrt{42}} \right) = (98.46, 107.54)$$

Assumptions

Constructing confidence intervals for μ is easy, but it is very important to keep in mind that there are underlying assumptions:

- Independent random sampling from the same distribution.
- Sampling distribution for $\bar{x}$ is normal:
  - Is the underlying x normal?
  - If x is non-normal, is n large enough that the sampling distribution for $\bar{x}$ is approximately normal?
- Examples 1-3: σ is *known*. (In practice we have to estimate σ)

To the extent these assumptions don't apply, we are limited in our ability to say the confidence interval as presented above will contain μ $(1-\alpha)$% of the time.

Confidence interval for a proportion

- Thus far we've described a confidence interval for the mean of a random variable x. A special case is a dichotomous variable that can only take on the values 1 or 0 (e.g., will vote for Biden or not; smokes cigarettes or not).
- Recall that the proportion of x that equal 1 in the population is denoted π. The proportion of x equal to 0 is $1-\pi$.
- We estimate π with the *sample* proportion $\hat{\pi}$, the mean of the dichotomous variable x. (E.g., in a sample of 900 likely voters, $\hat{\pi} = 0.65$).

Confidence interval for a proportion

- From Lectures 3-4, the population standard deviation of Bernoulli x is $\sqrt{\pi(1-\pi)}$.
- From Lecture 4: with large enough n, the Central Limit Theorem tells us our sample proportion $\hat{\pi}$ will be normally distributed with a mean of π and a standard error of $\sqrt{\pi(1-\pi)/n}$.
- Knowing this, we can construct confidence intervals for π in the same manner as for μ. For example, the following interval will contain π in 95% of samples:

$$\hat{\pi} \pm 1.96\sqrt{\frac{\pi(1-\pi)}{n}}$$

Confidence interval for a proportion

When we do not know π (which is almost always true, otherwise we would not need to estimate it), we can use our estimate of it, $\hat{\pi}$, in place of π:

$$\hat{\pi} \pm 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

In a sample of 900 likely voters in which 65% of voters say they will vote for Joe Biden, a 95% confidence interval for the population proportion is:

$$0.65 \pm 1.96\sqrt{\frac{0.65(1-0.65)}{900}}$$

$$0.65 \pm 1.96 * (0.0159) = (0.6188, 0.6812)$$

With a sample of only 900 likely voters, we are able to produce a confidence interval with a relatively small range of likely values.

Learning statistics is important!

Confidence intervals for *proportions* in Stata

Confidence intervals for a population proportion are straightforward in Stata with the `proportion` command. For example, using the NELS data the 95% confidence interval for taking advanced math in 8th grade can be found using:

- `proportion` *`advmath8`*`, citype(wald)`

A 95% confidence interval is the default in Stata, but the confidence level can be changed with `level`. The option `citype(wald)` is needed for Stata to use the standard error formula $\sqrt{\hat{\pi}(1-\hat{\pi})/n}$. (There are other options for this calculation).

Note z (not t) is generally used for confidence intervals for proportions. (Intuitively, this is because s does not have to be estimated separately. Once you have $\hat{\pi}$, you have s).

Margin of error

The product of z and the standard error is often called the **margin of error**. For example, the margin of error in the poll of voters above is:

$$1.96\sqrt{\frac{0.65(1-0.65)}{900}} = 0.031$$

or 3.1 percentage points. The margin of error is what is added/subtracted from $\hat{\pi}$ to get the confidence interval.

Assumptions

Again, our ability to say the above confidence interval for π contains the true parameter $(1-\alpha)$% of the time depends on:

- Independent random sampling from the same distribution.
- Sampling distribution for $\hat{\pi}$ is normal.

With proportions, a good rule of thumb is that the sampling distribution for $\hat{\pi}$ is approximately normal if $n\pi \geq 10$ and $n(1-\pi) \geq 10$. Can use your sample n and $\hat{\pi}$ to test the plausibility of this assumption.

Confidence interval when σ is unknown

- Our earlier application of the confidence interval formula for μ assumed that σ is *known*, which is almost never the case.
- When σ is *unknown*, we estimate it with the sample standard deviation s.
- In doing so, we introduce additional sampling variability that would not be present if we knew σ.
- The standardized version of $\bar{x}$ no longer follows a standard normal (z) distribution, but rather a **Student's *t* distribution** (see next slide).
- The confidence interval for μ then becomes:

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

The t distribution

The t-distribution is similar to the standard normal (z):

- It is symmetric and bell-shaped.
- Its mean and median are zero.
- Its standard deviation depends on its *degrees of freedom*, which in this case is $df = n - 1$.
- For small n, the tails of the t-distribution are *thicker* than those of the standard normal (i.e., greater variance).
- For large n, the t-distribution looks approximately like the standard normal.

The t distribution

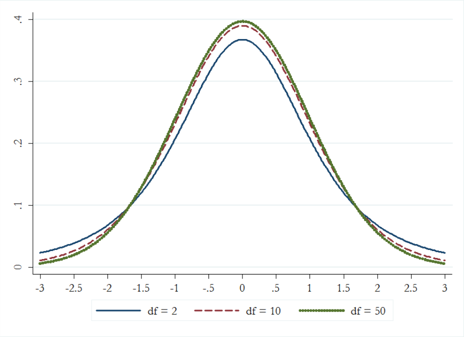


Figure: t-distribution with 2, 10, and 350 df

The t distribution

What are the implications of thicker tails? The probability of falling within 1.96 standard deviations of the mean (0) is somewhat *lower* for t than for z. Also, the likelihood of more extreme values (landing in the tails) is somewhat higher.

- Using the *df*, one can easily use Stata, statistical tables, or an online calculator to find values of t associated with particular probabilities. (Similar to what we did for z).

The t distribution—using Stata

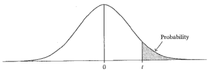
Using Stata to find probabilities and values from the t distribution:

- `display ttail(`*df*`,`*t*`)` finds the probability of exceeding a given t value. Comparable to `display normal(`*z*`)`, except that `ttail()` gives you the probability *above* a certain t, while `normal()` gives you the probability *below* a certain z.
- `display invttail(`*df*`,`*p*`)` finds the value of t for which there is a probability p of exceeding. Comparable to `display invnormal(`*p*`)`, except that `invttail()` gives you the t for which the probability *above* t is p, while `invnormal()` gives you the z for which the probability *below* z is p.
- When finding the appropriate t to use in a confidence interval, $\alpha/2$ is your p in these functions!

The t distribution—using Stata

- For example, suppose n=21. To find the t value for a 95% confidence interval, use `display invttail(20,0.025)`.
- $t = 2.086$, which you should note is *larger* than the 1.96 value used in 95% confidence intervals when σ was known.

The *t* distribution



	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.611
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467

Figure: t-distribution lookup table

Example 1, revisited

This time assume a sample of 41 women is taken (n=41) where the mean height is estimated to be $\bar{x} = 64$ inches. We estimate $s = 3$. Construct a 95% confidence interval for the true population mean height (μ).

$$\bar{x} \pm t_{0.025} \left(\frac{s}{\sqrt{n}} \right)$$

$$64 \pm 2.021 \left(\frac{3}{\sqrt{41}} \right) = (63.05, 64.95)$$

The t value of 2.021 comes from the t-distribution table with df=40 and $\alpha/2 = 0.025$. In Stata:

```
display invttail(40,0.025)
```

Example 4

From a sample of 20, suppose the mean test score is 76.1 with $s = 15.2$. Construct a **90%** confidence interval for the true population mean test score (μ).

$$\bar{x} \pm t_{0.05}\left(\frac{s}{\sqrt{n}}\right)$$

$$76.1 \pm 1.729\left(\frac{15.2}{\sqrt{20}}\right) = (70.22, 81.98)$$

The t value of 1.729 comes from the t-distribution table with df=19 and $\alpha/2 = 0.05$. In Stata:

```
display invttail(19,0.05)
```


Example 5

Nabisco, makers of Chips Ahoy, claims that there are 1000 chocolate chips in each 18 oz bag of cookies. The company offered $25,000 in scholarships to schools if students could verify the claim. The Air Force Academy collected a sample of 42 bags. $\bar{x} = 1261$ and $s = 117.6$. Construct a 95% confidence interval for the true mean number of chips (μ).

$$\bar{x} \pm t_{0.025}\left(\frac{s}{\sqrt{n}}\right)$$

$$1261 \pm 2.02\left(\frac{117.6}{\sqrt{42}}\right) = (1222.4, 1297.6)$$

The t value of 2.02 comes from the t-distribution table with df=41 and $\alpha/2 = 0.025$. In Stata:

```
display invttail(41,0.025)
```

Confidence intervals in Stata

Because σ is always estimated, Stata's `mean` command uses t values in constructing confidence intervals. These values are close to z whenever n is large. (Below, $z95 = 1.96$).

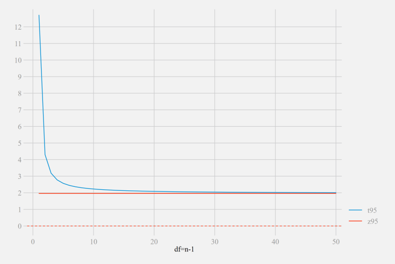


Figure: z versus t for a 95% confidence interval

Margins of error and sample size

In all of the above examples, confidence intervals have taken the form of **point estimate $\pm$ margin of error**. The margin of error was determined by three things: the desired confidence level, the underlying population standard deviation, and the sample size. For example:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

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

Margins of error and sample size

We can use the margin of error to determine how large a sample needs to be in order for point estimates to be within some distance of the true population mean $(1-\alpha)\%$ of the time. Example:

- We would like $\bar{x}$ to be within 4 of μ in 95% of samples
- Put another way, our margin of error will need to be 4
- Assume $\sigma = 25$
- How large will n need to be?

Margins of error and sample size

$$\begin{aligned}1.96\left(\frac{25}{\sqrt{n}}\right) &= 4\\ \frac{25}{\sqrt{n}} &= 2.0408\\ \frac{25}{2.0408} &= \sqrt{n}\\ 12.25 &= \sqrt{n}\\ 150 &= n\end{aligned}$$

Margins of error and sample size

Notice the minimum required sample size depends on both the confidence level (which determines the z-score, in this case 1.96) and the population standard deviation. For a given desired margin of error m we can write the minimum sample size required as:

$$n = \left(\frac{z\sigma}{m}\right)^2$$

Of course, σ is not known and can't be estimated with s before collecting data. But one may be able to use other information about the distribution of x to approximate this value (e.g., prior studies, similar data collections).

Margins of error and sample size

Note the margin of error declines in proportion to $\sqrt{n}$—there are diminishing returns to sample size in reducing the margin of error.

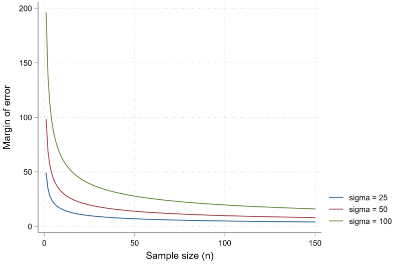


Figure: Margin of error as a function of sample size

Margins of error and sample size: proportions

The procedure for estimating minimum required sample sizes is similar for proportions. For a given desired margin of error m (between 0 and 1), we can write the minimum sample size required as:

$$n = \left(\frac{z}{m}\right)^2 \pi(1-\pi)$$

Again, π is not known before collecting data, but may be approximated given other information. (For example, in an opinion poll, the researcher could use previous polls on the same question to approximate π). The **most conservative estimate for π is 0.50**.

Margins of error and sample size: proportions

Table: Minimum sample size required for margin of error m and π

m:	0.05	0.04	0.03	0.02	0.01
π	Minimum sample size required				
0.30	323	504	896	2,017	8,067
0.35	350	546	971	2,185	8,740
0.40	369	576	1,024	2,305	9,220
0.45	380	594	1,056	2,377	9,508
0.50	384	600	1,067	2,401	9,604
0.55	380	594	1,056	2,377	9,508
0.60	369	576	1,024	2,305	9,220
0.65	350	546	971	2,185	8,740
0.70	323	504	896	2,017	8,067

Margins of error and sample size: proportions

For a given n, sampling variability is greatest when $\pi = 0.5$

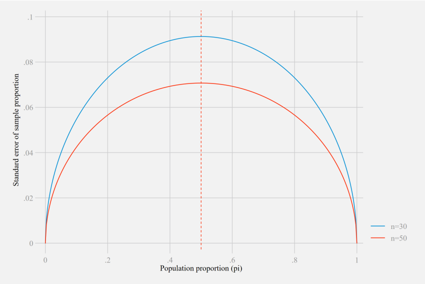


Figure: Standard error of sample proportion as a function of π

Margins of error and sample size: proportions

Why? When π is very large or very small there is little possibility of sampling error. When π is closer to 0.5, there is more opportunity for sampling error.

Romer (2020) and significance testing

In Lecture 6 we will take a different approach to inference—significance testing. Many researchers (including Romer, 2020) argue for the reporting of confidence intervals rather than significance testing alone. He finds most papers (in economics, at least) do not do this.

- Confidence intervals provide a "range of likely values" for the population parameter.
- They tell us what values of the parameter the data provide strong evidence against, and what values they provide little reason to object to.

More on this in Lectures 6-7.

Simulations

Let's try simulating $(1-\alpha)\%$ confidence intervals for $\bar{x}$ in Stata:

- Draw a sample of x of size n from a known population distribution
- Calculate $\bar{x}$ and construct the $(1-\alpha)\%$ confidence interval, first assuming we know σ, and then assuming we don't know σ and have to estimate it using s.
- In what percent of simulations does the confidence interval contain the true population μ in each case?

Simulations

Can do the same thing for proportions:

- Draw a sample of (dichotomous) x of size n from a population with known π
- Calculate $\hat{\pi}$ and calculate a $(1-\alpha)\%$ confidence interval.
- In what percent of simulations does the confidence interval contain the true population mean π?

Note: always use z (not t) when constructing confidence intervals for proportions. Do check to ensure there are 10+ cases where $x=0$ and 10+ cases where $x=1$. This is the rule of thumb for normality of $\hat{\pi}$