Introduction to Statistics I

Instructor: Jodin Morey   moreyj@lemoyne.edu

Previous Lecture

- Sampling distr of the sample proportions $\widehat{p}$
- Spread (SD) of sampling distr: $\widehat{s} = \sqrt{\frac{\pi(1-\pi)}{n}}$
- CLT: shape/center/$\widehat{s}$. Holds if $n\pi \geq 10$ & $n(1-\pi) \geq 10$.

Topic 16: Confidence Intervals for Proportions

One can't always do repeated sampling. How much do you trust a single sample?

Example: In an August poll, Mike Lawler was leading Mondaire Jones for representative of NY's 17th district. The survey asked 433 likely voters: "If the election were today, who would you vote for?"

43% Lawler   38% Jones

How accurate was this poll? After all, it only surveyed 433 voters.

RQ: Could it be that Jones *actually* had 50% favorability? (could he win?)

Population/Parameter π/Sample/Statistic $\widehat{p}$?

Population: District 17 voters.

Parameter: Proportion of District 17 voters who would vote for Jones.

Sample: 433 Likely Voters.

Statistic: Proportion of the 433 NY voters in 17th district who say they would vote for Jones: $\widehat{p} = 0.38$.

Simulating Polls

If Jones actually has 50% favorability, is it possible to get a sample proportion $\widehat{p}$ of 0.38 w/a sample size of 433?

Let's simulate it. Go to link:

"Edit Proportion" → 0.5

Set "$n =$" 433

"Generate 1000 Samples"

Set "left tail" to 0.38.

bit.ly/introstatsdata

Applets: Sampling Distr

Population Parameter

The simulation showed that if the true population parameter π is 50%,

it's not reasonable to think Jones would get 38% from a sample of 433 people.

We conclude from this poll that Jones very likely did not have 50% of vote.

What parameter numbers π could more reasonably give us results like we saw in this poll?

Can we get a range of reasonable values for π just from $\widehat{p} = 0.38$?

(Back to the applet for simulation. Calculate area under curve for $\widehat{p}$ or more extreme. Must choose either left/right tail.)

Not 50%. But maybe 40%. And maybe 35%. But not 30%.

The range of **likely values for** π, based on observed $\widehat{p}$, is called a **Confidence Interval** (CI).

Cl Theory: For any $\widehat{p}$, we can generate a list of likely π's for which it's reasonable to get the statistic $\hat{p}$ we observed.

How to do this w/out running simulations for every possible π?

Recall **CLT** describes relationship between π and $\hat{p}$ (if $n\pi \geq 10$ & $n(1-\pi) \geq 10$):

$\hat{p}$ distr is approx. normal, mean is at π, SD is: $\widehat{s} = \sqrt{\frac{\pi(1-\pi)}{n}}$.

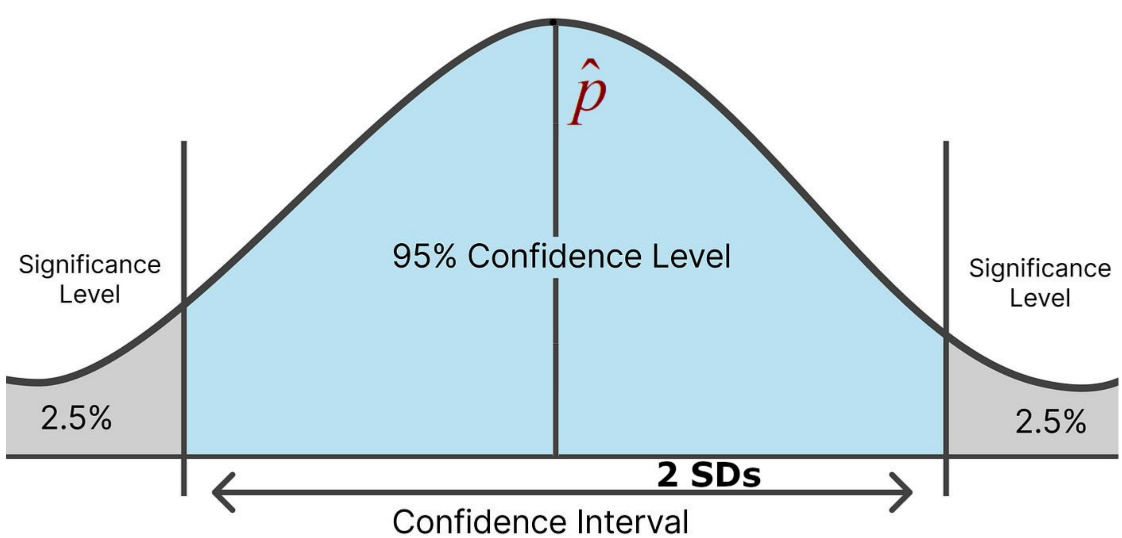
In particular, SD from CLT tells us average distance of the $\hat{p}$ from π.

Also recall the **empirical rule**: 68% of data pts are within 1 SD of mean, 95% within 2 SDs, nearly all within 3 SDs.

So, in 95% of samples, the statistic $\hat{p}$ is at most 2 SDs away from π.

Thus, for each sample $\hat{p}$, if we add/subtract 2 SDs, then for about 95% of samples this interval will contain π.

This is a 95% **confidence interval** (CI).



Minor Obstacle

Formula for SD is: $\widehat{s} = \sqrt{\frac{\pi(1-\pi)}{n}}$. But this relies on unknown parameter π (!?!).

So if we want a CI, we need another way to calculate $\widehat{s}$.

Standard Error (se): Is an approximation of $\widehat{s}$ given in CLT.

Replace the unknown parameter π with known statistic $\hat{p}$. So, $se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Similarly, in the technical requirements. So $n\hat{p} \geq 10$ & $n(1-\hat{p}) \geq 10$.

Therefore, the **CI Formula** is: $\hat{p} \pm z^*(se)$ where $\hat{p}$ is sample proportion,

z^* is the desired # of SDs (called the **critical value**), and $se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

So, CI is $\left(\hat{p} - z^*(se),\ \hat{p} + z^*(se)\right)$.

Recall our Example:

$n = 433, \qquad \hat{p} = 0.38.$

So SD is: $se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.38(1-0.38)}{433}} \approx 0.023.$

Thus the 95% **CI** is $\left(\hat{p} - z^*(se),\ \hat{p} + z^*(se)\right) = \left(0.38 - 1.960(0.023),\ 0.38 + 1.960(0.023)\right) = (0.335, 0.425).$

The 99% **CI** is $\left(\hat{p} - z^*(se),\ \hat{p} + z^*(se)\right) = \left(0.38 - 2.567(0.023),\ 0.38 + 2.567(0.023)\right) = (0.321, 0.440).$

Actual election results: Lawler 50%, Jones 44%.

Critical Values z^*

We can't usefully create CIs with 100% guarantee that CI contains π, because $\hat{p}$ varies randomly.

However, using CLT and the normal dist, we know 95% of $\hat{p}$'s are within 1.96 (not exactly 2) SDs from π.

If we build CIs using $z^* = 1.96$ SDs, then 95% of CIs will contain π. So, 1.96 is called the **critical value** for 95%.

Confidence Levels

We can increase the % of CIs that contain π by changing critical value z^*.

If wider, it'll contain π more often. If narrower, it'll contain π less often.

Percent of time CIs contains π is called the **confidence level**.

Confidence Levels and Critical Values:

Confidence Level	Critical Value (z^*)
80%	1.282
90%	1.645
95%	1.960
99%	2.567

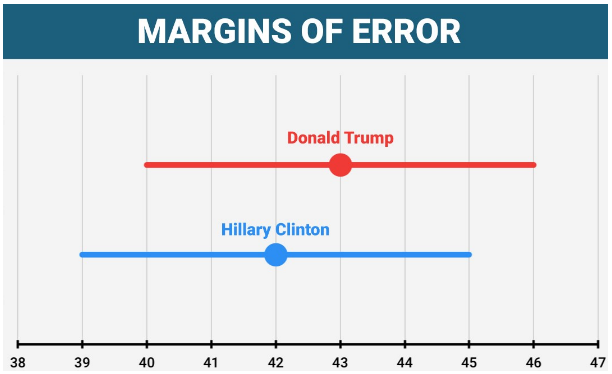
"95% CI" means that 95% of CIs we make using this procedure for different samples contain π.

Confidence level (95%) is our accuracy rate. For any given CI, we've no way of knowing if that particular CI contains π. But we have an accuracy rate of 95%.

Margin-of-Error (*moe*): Max distance we expect $\hat{p}$ to be from π is known as margin-of-error. $moe = z^*(se)$.

It's also the **half-width** of the CI.

Many polls report their results w/statistic $\hat{p}$ and *moe*.



2016 POTUS Race

Activities: 16-2

Day 2 - Topic 16: Confidence Intervals for Proportions

Return to Lawler/Jones Poll Example: Change Research poll reports Jones with 38% w/*moe* of 4.5 percentage points. What's the 95% CI?

Calculate CI as: $\hat{p} \pm moe$. So, $0.38 \pm 0.045 = (0.335, 0.425)$.

So, we're 95% **confident** that between 33.5% and 42.5% of voters will vote for Jones.

95% **confidence** means that "if we ran this poll many times, we belive 95% of the resulting CIs would contain π."

Let's make 95% CI for **Lawler**'s statistic.

Recall: 43% of likely voters said they planned to vote for Lawler.

$n = 433, \quad \hat{p} = 0.43, \quad z^* = 1.96$

$se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.43(1-0.43)}{433}} \approx 0.024.$ (standard error) ...

$moe = z^*(se) = 1.96(0.024) \approx 0.047.$

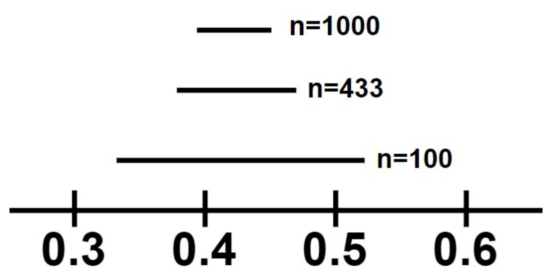
95% CI: $\hat{p} \pm moe = 0.43 \pm 0.047 = (0.383, 0.477).$

We're 95% confident the % of voters who'll vote for Lawler is between 38.3% and 47.7%.

Effect of Sample Size on CI

Let's try different sample sizes with $se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and $\hat{p} = 0.48$, realling that half-width is: $moe = z^*(se)$.

Sample Size (n)	Standard Error (se)
1000	0.016
433	0.024
100	0.046



CIs for $\hat{p} = 0.48$ and various sample sizes

As sample size increases, se decreases. So CIs will be narrower.

(Why? Imagine the sample size were nearly entire population. What would each $\hat{p}$ be? Would they vary much?)

Narrower CIs are more useful, so larger sample sizes are beneficial, because they increase accuracy.

Another way to change the half-width $z^*(se)$ is to change confidence level, and thus change z^*.

This decreased confidence narrows the CIs.

Confidence Levels and Critical Values:

Confidence Level	Critical Value (z^*)
80%	1.282
90%	1.645
95%	1.960
99%	2.567

Demanding higher confidence results in wider CI.

So if we want more confidence, we must be less precise (or increase sample size).

Activities: 16-X

bit.ly/introstatsdata

Applets: Simulating Confidence Intervals

What did we learn?

- Confidence intervals (CI)
- Standard error, *se*
- Critical values
- Confidence levels
- Margins of error, *moe*