Introduction to Statistics I

Textbook: Elementary Statistics (4th Edition, by Navidi and Monk), and Workshop Statistics (4th Edition, by Rossman and Chance).

Previous Lecture

- Intuition behind Qualitative CLT
- Qualitative CLT results - sample distr has:

  Shape is ≈normal, center $\mu_{\hat{p}}$ is at parameter μ, spread is $se = \sqrt{\frac{p(1-p)}{n}}$

- CLT technical conditions: SRS & $np \geq 10$ & $n(1-p) \geq 10$.

Recall CLT Comparison

	Quantitative Var (mean)	**Qualitative Var** (proportion)
Sample Statistic	$\bar{x}$	$\hat{p}$
Tech Cond - SRS and:	population is normal or $n \geq 30$	$np \geq 10$ and $n(1-p) \geq 10$
Shape of Sampling Distr	Normal or Approximately Normal	Approximately Normal
Center (same as pop)	μ	p
Standard Deviation	$se = \frac{\sigma}{\sqrt{n}}$	$se = \sqrt{\frac{p(1-p)}{n}}$

§8.3: Confidence Intervals for Population Proportion

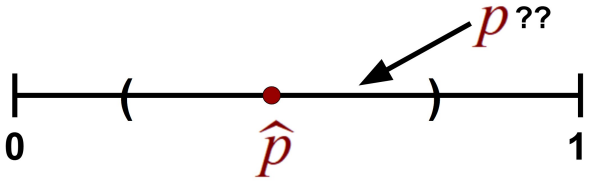
How do we estimate the proportion p of something in a population (colored candies, overweight newborns, etc.)?

We usually don't have access to the whole population, and often must settle for only *one sample*.

The obvious answer (best guess), is to find: $\hat{p}$. But, this is probably *not completely accurate*.

It's called a *point estimate*: a single number used to estimate the value of an unknown parameter.

It's better to have an interval around our point estimate, thus increasing the chance that we "capture" the true parameter p.



Below, we discuss how to create this interval.

Example: In an August 2024 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?" The poll showed:

43% Lawler 38% Jones

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

Population/Parameter p/Sample/Statistic $\hat{p}$?

Population: District 17 voters.

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

Sample: The 433 likely voters.

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

Simulating Polls

To examine this, let's simulate what happens when we sample from such a population.

If Jones actually has 50% favorability, how likely is it to get a proportion $\hat{p}$ of 0.38 (or less) w/a sample size of 433?

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 for Proportion

Population Parameter

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

it's unreasonable to think Jones would get 38% from a sample of 433 people.

We conclude that Jones very likely **did not** have 50% of the vote.

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

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

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

Not $p = 50\%$. But maybe 40%. And maybe 35%. But not 30%.

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

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

How do we do this w/out running simulations for every possible p?

Recall that **CLT** tells us (if tech conds are met) about the shape/center/spread of the $\hat{p}$ distr.

In particular that the $\hat{p}$ distr is ≈normal.

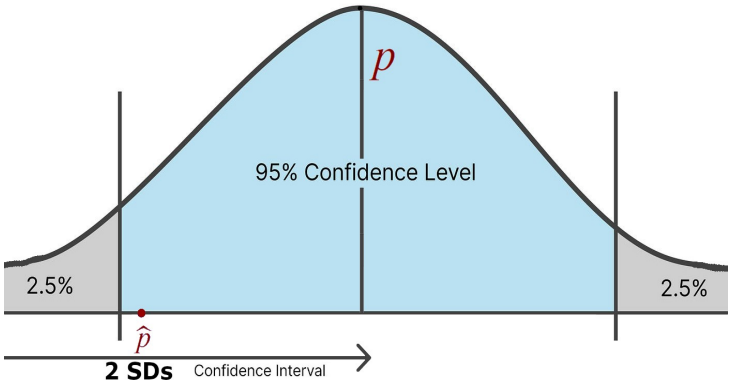
Also recall that the **empirical rule** applies to normal distr's.

Thus: 68% of $\hat{p}$ are within 1 SD of p, 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 p .

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

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



95% of $\hat{p}$s (which are in the blue region) will capture p with a 2SD CI

Minor Obstacle

Formula for SD is: $se = \sqrt{\frac{p\,(1-p\,)}{n}}$. But this relies on unknown parameter p (!?!).

So if we want a CI, we need another way.

Instead of se, we replace the unknown parameter p w/known statistic $\hat{p}$.

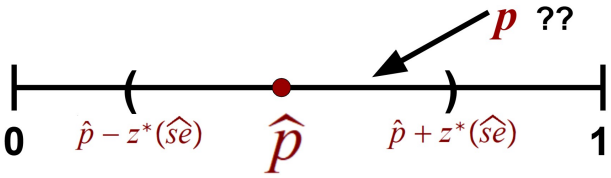
Thus, we have an *estimate for the standard error*: $\widehat{se} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Similarly for the CLT technical requirements: $n\hat{p} \geq 10$ & $n(1-\hat{p}) \geq 10$.

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

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

In interval notation, the CI is $\left(\hat{p} - z^*(\widehat{se}),\ \hat{p} + z^*(\widehat{se})\right)$.



Critical Values z^*

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

Correction: 95% of $\hat{p}$'s are within 1.96 (not exactly 2) SDs from p.

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

Confidence Levels

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

If wider (z^* bigger), the CI will contain p more often. If narrower (z^* smaller), it'll contain p less often.

The **percent of CIs** that contains p 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 p.

Recall our Example:

$n = 433, \quad \hat{p} = 0.38.$ What's the 95% **CI**?

The SD is: $\widehat{se} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.38(1-0.38)}{433}} \approx 0.02333.$

Thus the 95% **CI** is $\left(\hat{p} - z^*(\widehat{se}),\ \hat{p} + z^*(\widehat{se})\right)$

$= \left(0.38 - 1.960(0.02333),\ 0.38 + 1.960(0.02333)\right) = (0.3343, 0.4257).$ What's the 99% **CI**?

The 99% **CI** is $\left(\hat{p} - z^*(\widehat{se}),\ \hat{p} + z^*(\widehat{se})\right)$

$= \left(0.38 - 2.567(0.02333),\ 0.38 + 2.567(0.02333)\right) = (0.3201, 0.4399).$ In context?

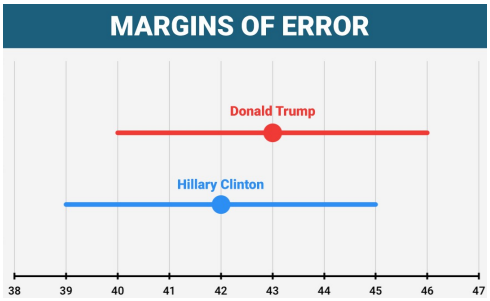
"We're 99% confident that the proportion of voters who favor Jones is between 32.1% and 44%."

Optional Activity: 8.3a

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

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

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



2016 POTUS Race

Returning to Lawler/Jones: One poll reported Jones having 38% with a *moe* of 4.5 percentage pts at the 95% confidence level. What's the 95% CI?

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

We're 95% **confident** that between 33.5% and 42.5% of voters will vote for Jones. What does 95% confidence mean?

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

Let's make a 95% CI for **Lawler's** statistic.
Recall: 43% of likely voters said they planned to vote for Lawler.

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

$\widehat{se} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.43(1-0.43)}{433}} \approx 0.02379.$ (est. standard error) ...

$moe = z^*(\widehat{se}) = 1.96(0.02379) \approx 0.04663.$

95% CI: $\hat{p} \pm moe = 0.43 \pm 0.04663 = (0.3834, 0.4766)$. In context?

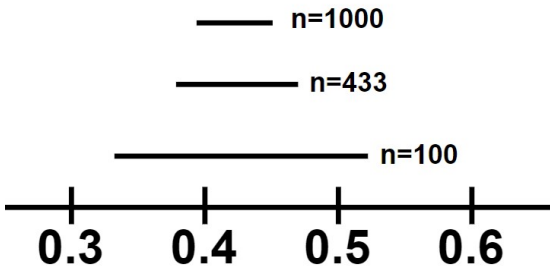
We're 95% confident the % of voters who'll vote for Lawler is between 38.3% and 47.7%.

Effect of Sample Size on CI

Since the CI half-width is: $moe = z^*(\widehat{se})$, where $\widehat{se} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, what can we predict will happen as sample size increases?

Let's try different sample sizes with $\hat{p} = 0.48$.

Sample Size (n)	Standard Error ($\widehat{se}$)
1000	0.016
433	0.024
100	0.046



95% CIs for $\hat{p} = 0.48$ and various sample sizes

As sample size increases, $\widehat{se}$ decreases. So the CIs will be narrower.

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

Narrower CIs are more useful, so larger sample sizes are desireable.

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

This decreased confidence, but narrows the CIs.

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).

Activity: 8.3b

bit.ly/introstatsdata

Applets: Simulating Confidence Intervals

What did we learn?

- Estimated standard error, $\widehat{se}$
- Critical values z^*
- Confidence intervals (CI) $\hat{p} \pm z^*(\widehat{se})$
- Confidence levels (80%, 90%, etc.)
- Margins of error, $moe = z^*(\widehat{se})$