Chapter 9: Estimating the Value of a Parameter

Section 9.1: Estimating a Population Proportion

sample ↳ statistic vs population ↳ parameter

· proportion $\hat{p}$ · mean $\bar{x}$ · st dev s · Var s^2

· proportion p · mean μ · st. dev σ · variance σ^2

INTRO

DESCRIPTIVE STATISTICS (• 5 # Summary • Scatter Plot • Box Plot) + **PROBABILITY** *Find a level of RELIABILITY* (Ch5, 6, 7 ← sampling; pure prob, normal) = **INFERENTIAL STATISTICS** (use sample data & prob theory to draw inferences about population)

What if we wanted to know the answer to the following poll question:

Poll: Do you (GCC student) prefer fully online courses instead of regular, face-to-face courses?

Yes = 3 No = 11 total n = 14

We are going to try and estimate the percentage of all GCC students who answer 'yes' to the above poll.

In the last section (8.2) we asked this very same question. Based on our class results, we got that the proportion of students who answers yes was: $\hat{p}_{last} = 0.2$ $\hat{p}_{now} = \frac{3}{14} = 0.214$.

This is called a **point estimate**.

a. How could we find out how this compares to the proportion of ALL GCC students?

could ask all GCC Q ⇝ unlikely.

b. Based on our sample, I am going to guess that that the true population proportion of GCC students who prefer online course is between 0.9 and 1.00. That's a percentage of between 90% and 100%.

What level of agreement do you have with this guess? (*Circle one*)

Totally agree | Somewhat agree (6) | Somewhat disagree (3) | No Way!

c. Your instructor has revised their guess! Now they believe that there is between 0.5 and 0.6 .

That's a percentage of between 50% and 60%.

What level of agreement do you now have with the instructor's new guess? (*Circle one*)

Totally agree | Somewhat agree (7) | Somewhat disagree | No Way!

d. One last revision! Now they believe that there is between 10 % and 30 % of ~~black beans in the tray~~.
That's a proportion of between 0.1 and 0.3.

What level of agreement do you now have with the instructor's new guess? (*Circle one*)

Totally agree (11) | Somewhat agree | Somewhat disagree | No Way!

e. Discuss: → key sampling variability (ch 8)

- What would happen if we took another sample? Would we get the same sample proportion?

new poll ⟶ would we get exactly $\hat{p} = 0.2$ or $\hat{p} = 0.214$? NO!

- Would you like to know the true proportion of GCC students who prefer online classes?

yes, too bad cant know :(

Def ***Point Estimate*** A single value used to approximate a population parameter.

Note: *The sample proportion* $\hat{p}$ *is the point estimate of the population proportion* p.

SAMPLE PROPORTION: $\boxed{\hat{p} = \dfrac{\textbf{successes}}{\textbf{total}} = \dfrac{x}{n}}$

Ex1: I took a sample of 40 GCC students and asked them "Do you think Sir Pugsly Farnsworth Esquire III is cute or not?" The results were as follows: **24** said **YES** and **16** said **NO**.

a. What is the variable? Is it qualitative or quantitative?

variable = answer to Q "yes" or "no" QL → key turn into QN by looking @ proportion

b. This is a BINOMIAL problem since there are only two outcomes.

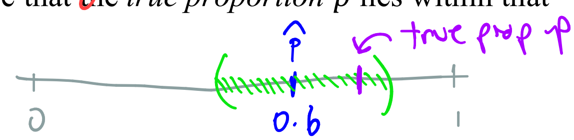
c. What would $\hat{p}$ be? $\hat{p} = \dfrac{24}{40} = 0.6$ (point estimate)

d. Which statement do you think we can make from this point estimate?

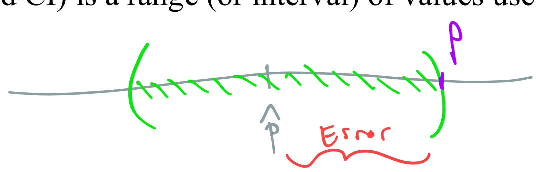
1. ***Exactly*** 60% of ALL GCC students think Sir Pugsly Farnsworth Esquire III is cute. ← unlikely due to sampling variability
2. ***About*** 60% of ALL GCC students think Sir Pugsly Farnsworth Esquire III is cute.

From this example, we can imagine that there may be some variability associated with taking a sample from a population. We will create an interval around $\hat{p}$ with the hope that the *true proportion* p lies within that interval and we state a certain level of confidence we have.

$\boxed{(\hat{p} \pm ERROR)}$



Def ***A confidence interval*** (abbreviated CI) is a range (or interval) of values used to estimate the true value of a population parameter.



Def ***Margin of Error*** The maximum likely difference between the observed sample proportion $\hat{p}$ and the true value of the population proportion p.

MARGIN OF ERROR FORMULA: $\boxed{E = z_{\alpha/2} \cdot \sqrt{\dfrac{\hat{p}\hat{q}}{n}}}$ $E = z_{\alpha/2} \cdot \sigma_{\hat{p}}$

$\boxed{\text{Note } \hat{q} = 1 - \hat{p}}$

Ex2: If the margin of error is 8%, then what interval would you expect the **true proportion** of PCC students to be who think Sir Pugsly Farnsworth Esquire III is cute?

$E = 0.08 = 8\%$ $\hat{p} - E = 0.6 - 0.08 = 0.52$

$\hat{p} = 0.6$ $\hat{p} + E = 0.6 + 0.08 = 0.68$

interval expect true proportion p to be in is $(0.52, 0.68)$

Effects on the Width of your Confidence Interval

1) Sample Size

- If n is small, the CI widens
- If n is large, the CI narrows

2) Confidence level

- The lower the confidence level, the CI narrows
- The higher the confidence level, the CI widens

Let's use our first $\hat{p}$ = 0.6 and then take two more samples where the $\hat{p}$'s were 0.51 ($\hat{p}_2$) and then 0.71 ($\hat{p}_3$). Using the same margin of error, if the true proportion of GCC who thinks Sir Pugsly is cute is 0.65, then which $\hat{p}$'s had confidence intervals that contained p?

- 1st ex: $\hat{p}_1 = 0.6$ (.52, 0.68) ✓ contains $p = 0.65$
- $\hat{p}_2 = 0.51$ $\hat{p}_2 \pm E = 0.51 \pm 0.08$: CI (0.43, 0.59) ✗ doesn't contain
- $\hat{p}_2 = 0.71$ $\hat{p}_3 \pm E = 0.71 \pm 0.08$: CI (0.63, 0.79) ✓ contains $p = 0.65$

It looks like our probability of success is 2/3 = 67% right now, but are there only three samples of 40 GCC students that can be taken?

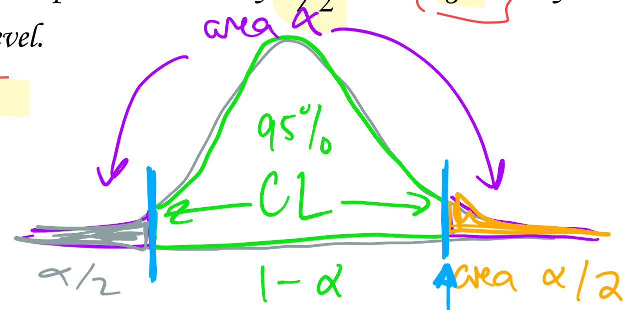
Def ***Confidence Level*** The probability equal to the proportion of times that the confidence interval actually contains the true population parameter. % of successful interval

Note: If we were to repeat the estimation process a large number of times, the confidence level is the percentage of confidence intervals that would actually contain the true population proportion.

Notation: The confidence level is denoted by $1-\alpha$

$\boxed{CL = 1-\alpha}$

Note: The ***critical value*** *(denoted $z_{\alpha/2}$) is the positive z score that separates an area of $\alpha/2$ in the right tail of the standard normal distribution. It is determined by the confidence level.*

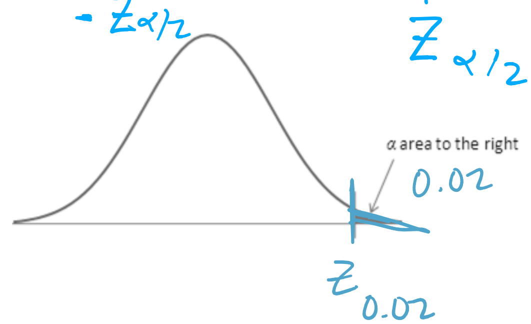


- CL = $1-\alpha$
- α = $1-CL$
- $\alpha/2$ = $(1-CL)/2$
- $z_{\alpha/2}$ = invNorm($\alpha/2$, 0, 1, RIGHT)

Ex3: Find $z_{0.02}$

$z_{0.02}$ = invNorm(0.02, 0, 1, RIGHT)

= $\boxed{2.05}$



as CL ↓, CI narrows

Ex4: Find the margin of error if $\hat{p} = 0.48$, $n = 200$, and the confidence level is as follows:

a. 96% confidence level

$CL = 0.96 = 1-\alpha$
$\alpha = 1 - CL = 0.04$
$\alpha/2 = 0.02$

$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$

• $\hat{p} = 0.48$ • $\hat{q} = 1 - \hat{p} = 1 - 0.48 = 0.52$
• $n = 200$
• $z_{\alpha/2} = z_{0.02} = \text{invNorm}(0.02, 0, 1, \text{RIGHT})$ area μ σ tail
$= 2.05$

b. 80% confidence level

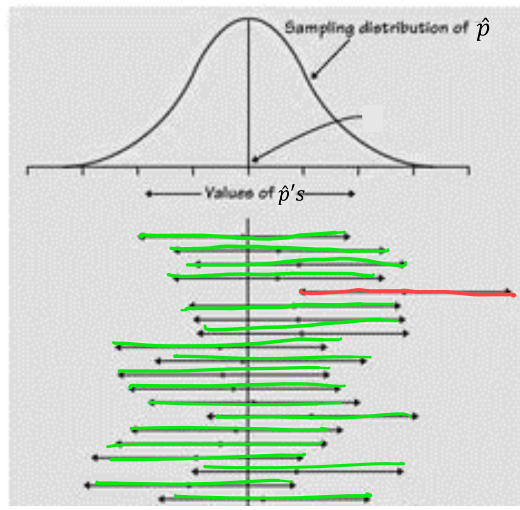
• $\text{Error} = (2.05) \cdot \sqrt{\frac{0.48 * 0.52}{200}} = \boxed{0.072}$

Part b
$CL = 0.8$
$\alpha = 1 - CL = 0.2$
$\alpha/2 = 0.1$

$z_{\alpha/2} = z_{0.1}$
$= \text{invNorm}(0.1, 0, 1, \text{RIGHT})$
$= 1.28$

$E = (1.28)\sqrt{\frac{0.48 * 0.52}{200}} = \boxed{0.045}$

Ex5: In the example to the right we took twenty different samples of size n, and found the corresponding confidence intervals with a 95% level of confidence.



A 95% confidence interval indicates that __19__ out of **20** samples from the same population will produce confidence intervals that contain p.

$0.95 * 20 = 19$

Why? The reason is **Sampling Distribution**!

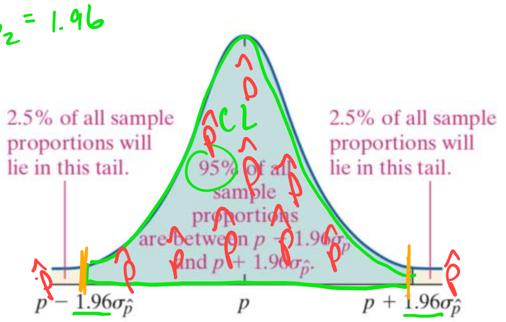
The sampling distribution for a proportion is approximately normal, and it is the theoretical distribution of ALL sample proportions, $\hat{p}$, with $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = E$. So, given a CL, we can use this distribution to estimate the number of samples that will be within $z_{\alpha/2}$ standard deviations from the true population proportion, p.

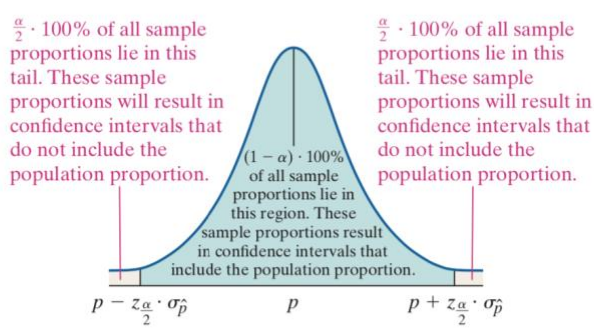
Expressed in a different way: CL% of samples ______ will be $p - (z_{\alpha/2}) \cdot \sigma_{\hat{p}} < \boxed{\hat{p}} < p + (z_{\alpha/2}) \cdot \sigma_{\hat{p}}$ sample

Crucial reordering the above (just algebra): $\underbrace{\hat{p} - \overbrace{(z_{\alpha/2}) \cdot \sigma_{\hat{p}}}^{E}}_{\text{sample \& CL}} < \boxed{p}^{\text{true}} < \underbrace{\hat{p} + \overbrace{(z_{\alpha/2}) \cdot \sigma_{\hat{p}}}^{E}}_{\text{sample \& CL}}$ or $\hat{p} - E < p < \hat{p} + E$

Key Point: this last bit says that the true population proportion, p, can be estimated from the sample proportion, $\hat{p}$, and the error, E. Moreover, the confidence interval, $CI = (\hat{p} - E, \hat{p} + E)$, will contain the true population proportion, p, CL% of the time.

$z_{0.05/2} = 1.96$





Key point: a sample statistics that is very far away from the population parameter (in the tails), will give a confidence interval that doesn't contain the population parameter.

CONFIDENCE INTERVAL FOR THE POPULATION PROPORTION

peref this one ?

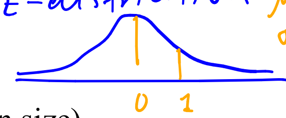
CI: (Alternative Forms) $\hat{p} - E < p < \hat{p} + E$ or $\hat{p} \pm E$ or $\boxed{(\hat{p} - E\,,\ \hat{p} + E)}$

$\cdot\ CL = 1-\alpha$ Point Estimate: $\hat{p} = x/n$

$\cdot\ \alpha = 1 - CL$ Critical Value: $z_{\alpha/2} =$ invNorm($\alpha/2$, 0, 1, RIGHT) ← distribution used for critical value:

$\cdot\ \alpha/2$ ↳ normal distribution

Error: $E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$ aka z-distribution $\mu = 0$ $\sigma = 1$



Requirements

1. The sample is a simple random sample (SRS)
2. The sample is independent or $n \leq 0.05N$ (i.e. less than 5% of population size)
3. The sample size satisfies: $n \cdot \hat{p} \cdot \hat{q} \geq 10$

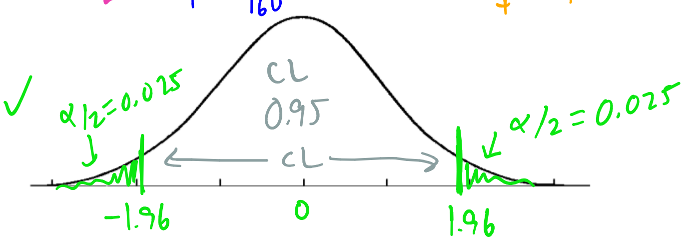
Ex6: In a 2018 random survey of 160 American Democrats, 136 said that they support Medicare-for-all, also known as single-payer healthcare. Find the 95% **confidence interval** estimate for the true proportion of Democrats in America who support Medicare-for-all. ($M \to E$)

Givens $n = 160$ $x = 136$ $CL = 0.95$

$\hat{p} = \frac{136}{160} = 0.85$ ⟹ $\hat{q} = 1 - \hat{p} = 0.15$

Check requirements

- SRS ✓
- independent ✓ $n = 160$ out of N → use 5% Rule
- $n\hat{p}\hat{q} \geq 10$: $160 * 0.85 * 0.15 = 20.4 \geq 10$ ✓



Identify point estimate (sample proportion)

$\hat{p} = \frac{136}{160} = \boxed{0.85}$

Determine critical value $z_{\alpha/2}$

- $CL = 0.95$
- $\alpha = 1 - CL = 0.05$
- $\alpha/2 = 0.025$

$z_{\alpha/2} = z_{\boxed{0.025}} =$ invNorm(0.025 (area), 0 (μ), 1 (σ), RIGHT (tail)) $= \boxed{1.96}$

Find margin of error

$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} = (1.96) * \sqrt{\frac{0.85 * 0.15}{160}} = 0.05532879$

$\boxed{E = 0.055}$

Construct confidence interval

- $\hat{p} \pm E$
- $\hat{p} - E = 0.85 - 0.055 = 0.795$ (lower)
- $\hat{p} + E = 0.85 + 0.055 = 0.905$ (upper)

Confidence Interval

$\boxed{CI: (0.795\,,\ 0.905)}$

Interpretation of CI ($M \to E$)

(template answer) *"We are* 95% *confident that the true proportion of* American Democrats who support Medicare-for-all *is between* 79.5% *and* 90.5% *."*

Does the proportion of Democrats who support Medicare-for-all appear to be substantially different than the 70% rate for the general population? If so, why do you think that is?

missing viewpoints: Republicans, Green Party, Libertarians, ...

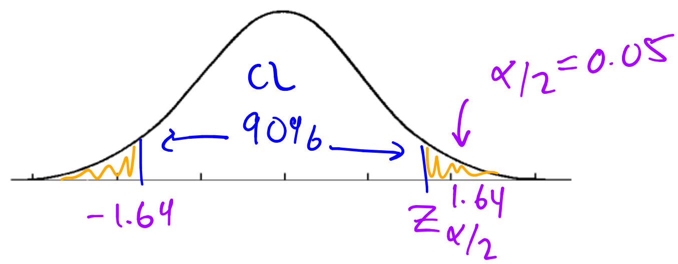
↳ because American Democrats doesn't include all different political viewpoints

SRS

EX7: In a 2018 survey on climate change, 626 out of 1278 Americans (18 years or older) were “extremely” or “very sure” it is happening. Construct a 90% confidence interval for the true percentage of Americans who are “extremely” or “very sure” climate change is happening.

Check requirements

- SRS? ✓
- indep? ✓
- $n\hat{p}\hat{q} \geq 10$? $1278 \cdot 0.49 \cdot 0.51 = 319.4$ ✓



Identify point estimate

$\hat{p} = \dfrac{626}{1278} = \boxed{0.490}$

$\hat{q} = 1 - \hat{p} = 0.51$

$CL = 0.9 = 1 - \alpha$

$\alpha = 1 - CL = 0.1$

$\alpha/2 = 0.05$

Determine critical value

$z_{\alpha/2} = z_{0.05} = \text{invNorm}(0.05, 0, 1, \text{RIGHT}) = \boxed{1.64}$

Find margin of error

$E = z_{\alpha/2} \cdot \sqrt{\dfrac{\hat{p}\hat{q}}{n}} = 1.64 \cdot \sqrt{\dfrac{0.49 * 0.51}{1278}} = \boxed{0.0229}$

Construct confidence interval

$\hat{p} \pm E$

$\hat{p} \pm E = 0.49 \pm 0.0229$

$\hat{p} - E < p < \hat{p} + E$

$(\hat{p} - E, \hat{p} + E)$

$\boxed{CI: (0.4671, 0.5129)}$ (46.7%, 51.3%)

Interpretation of CI

"We are 90% confident that the true proportion of Americans (aged 18 or older) who are "extremely" or "very sure" climate change is happening is between 46.7% and 51.3%."

GRAPHING CALCULATOR (TI-83 OR 84)

Instructions: (a) STAT $\Rightarrow$ TESTS $\Rightarrow$ 1-PropZInt

(b) Enter $\begin{cases} x = \text{number of successes} \\ n = \text{number of trials} \\ C\text{-}Level = \text{confidence level} \end{cases}$

DETERMINING SAMPLE SIZE

Point Estimate $\hat{p}$ Known	Point Estimate $\hat{p}$ Unknown
$n = \dfrac{[z_{\alpha/2}]^2 \hat{p}\hat{q}}{E^2}$	$n = \dfrac{[z_{\alpha/2}]^2}{4E^2}$

NOTE: When the sample proportion is unknown, we simply assume it is 50%, or $\hat{p} = 0.5$. Then $\hat{q} = 0.5$ as well so we get $\hat{p} \cdot \hat{q} = 0.25 = 1/4$. Which is why there's a 4 in the denominator of the second formula.

Ex8: An economist wants to know if the proportion of the children in the United States that speak a language other than English at home has changed since 2016, [when 22% of children did](). How many children would need to be surveyed if the economist wants to be within 2 percentage points of the true proportion with 92% confidence? ($M \to E$)

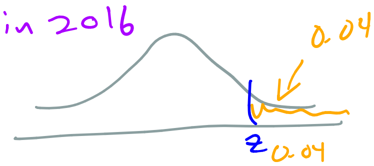
- $\hat{p} = 0.22 \to \hat{q} = 0.78$
- $E = 2\% = 0.02$
- $CL = 0.92$

$\hat{p} = 0.22$ in 2016

$CL = 0.92$

$\alpha = 1 - CL = 0.08$

$\alpha/2 = 0.04$



Want $n = ?$

Use $n = \dfrac{[z_{\alpha/2}]^2 \hat{p}\hat{q}}{E^2} = \dfrac{[1.75]^2 \cdot (0.22)(0.78)}{(0.02)^2}$

$= 1313.8$

$z_{\alpha/2} = z_{0.04} = \text{invNorm}(0.04, 0, 1, \text{Right})$

$= 1.75$

$n = 1314$ children survey

Ex9: Before US presidential elections potential candidates have their exploratory teams work to determine the percentage of people who will vote for their candidate. Kamala Harris has such a team looking into what percentage of people would vote for her. If they want to construct a 98% confidence interval with a 3% margin of error, how many people do they need to poll to achieve this? ($M \to E$)

Parameter studied: proportion (of people who will vote for Kamala Harris)

Givens
- $CL = 98\% = 0.98$
- $E = 3\% = 0.03$

Want

$n = ?$

which formula to choose for n? Select the one without $\hat{p}$ or $\hat{q}$! Don't know $\hat{p}$!

$CL = 0.98$

$\alpha = 0.02$

$\alpha/2 = 0.01$

$n = \dfrac{(z_{\alpha/2})^2}{4E^2} = \dfrac{(2.33)^2}{4 * (0.03)^2} = 1508.027$ (round up)

$z_{\alpha/2} = \text{invNorm}(\overset{\alpha/2}{0.01}, \overset{\mu}{0}, \overset{\sigma}{1}) = 2.33$

"Mrs. Harris team would need to randomly poll 1509 people to achieve a 98% confidence interval with a 3% margin of error."