Lecture 10

Point Estimates and Confidence Intervals

Today’s Updates / Reminders

- Exam grading underway. The grades will appear in Moodle first. Exams to be published after that, likely very soon after that.
- PRWA #1 (Graph) – part 1 – closes at the end of the day next class.
  - There are **TWO** due dates for a PRWA… submission day and grading day. Submission is due at the end of the day next time we meet… at 11:59pm.
  - If you do not make a submission, you get a zero for the entire thing. As part of the peer-grading, your assignment gets immediately assigned to another person to assess. If there is no submission, you are excluded from both due dates.
- There is no lab due Friday, but lab 5 using Excel will open this week.
- HW 6 opens today.

Chapter 6

Point Estimates

Scenario

- On Election Night, have you ever noticed they predict the winner of an election when only 1% of precincts are reporting? They do this by taking exit polls, which means they ask people who they voted for as they are walking out. They do this all over Wake County, the state of NC, and every state in the United States. As we will see, you don't have to ask many people to be able to accurately predict who will win. For the millions of people in NC, they will probably ask a few thousand people and get the answers they need. They do this using confidence intervals. You take a sample (trying to be as varied as possible) and then you calculate the percentage of voters who voted for a candidate. You decide how confident you want to be (think the newscaster saying, "The margin of error is +/- 3 points.") and you can predict the winner!

Summary of confidence interval (no calculations) - https://youtu.be/tFWsuO9f74o

Point Estimates

- A point estimate is a single number that can be regarded as a sensible value for a given parameter θ. Sample data is obtained and the statistic desired is calculated and we put a hat over it to suggest it is an estimate… $\hat{\theta}$
- It is said to be unbiased if the expected value of the estimator… $E(\hat{\theta}) = \theta$.
- Biased estimators arise in situation such as an uncalibrated measurement device, such as a scale that measures 5 kg above or below the actual measurement. Or it could be any bias we talked about in lecture 1.

Common Point Estimates

- For a binomial random variable with parameters n and p, the sample proportion is:

  $\hat{p} = \frac{X}{n}$, which is an unbiased estimator of p.

- For a random sample from a distribution with mean μ and variance, σ^2:

  $s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$, which is an unbiased estimator of σ^2

  $\bar{X} = \frac{\sum X_i}{n}$, which is an unbiased estimator of μ

The Standard Error

- The standard error of an estimator $\hat{\theta}$ is its standard deviation, $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$.

  This is the magnitude or representative deviation between an estimate and the value of θ. If this involves unknown parameters (as it often does), we call this estimated standard error and we label it $\hat{\sigma}_{\theta}$ or by $s_{\hat{\theta}}$.

- For $\hat{p}$, we can show that the standard error is $\sqrt{\frac{pq}{n}}$.
- For $\bar{x}$, we can show that the standard error is $\frac{\sigma}{\sqrt{n}}$.
- Further, for data that is approximately normal, we can be reasonably confident that the true value of any θ lies within two standard errors (or standard deviations) of the mean.

Chapter 7

Confidence Intervals (CIs)

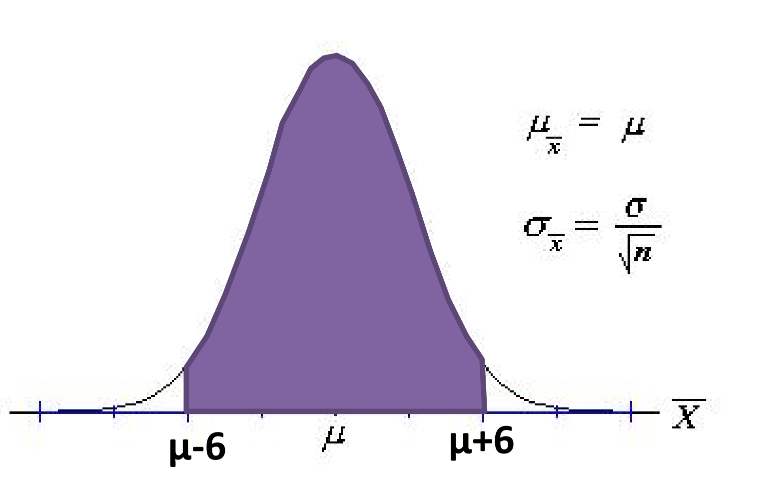
Confidence Intervals

- When we wish to make estimates of a population parameter, a confidence interval is an established way to give a range of values where we think the parameter will lie.
- The point estimate by itself is not good enough. As we saw from sampling, there is variation in the world. In the dice rolling example we did in class,
    - Grace rolled an average of 1.4 (1,1,2,2,2).
    - But then Peter R rolled 6,6,4,2,6 and Peter C rolled 3,6,3,6,6, which averages to 5.8.
- The expected value, E(X), should be 3.5. Those two rolls were not typical. But we did see that with much sampling (all your dice rolls), the average of your averages worked out to be 3.46.
- And the standard error for rolling 5 dice should be 0.76 and we're easily within just one standard error (0.07 < 0.76).
- We knew a theoretical E(X) or μ in this case, but **in most real-life scenarios**, **we do not**. So it's hard to know if 3.46 is good.

How can we estimate the parameter μ with just one value of $\overline{X}$ calculated from one SRS. And how confident could we be of the accuracy of the estimation?

Consider the example:
Suppose $\sigma = 30$ and the size of the SRS is 100.
Applying the Central Limit Theorem:

$$\overline{X} \text{ is approx } N(\mu, \frac{30}{\sqrt{100}}) = N(\mu, 3)$$



Looking at the distribution of $\overline{X}$, the 68-95-99.7 rule tells us that there is a 0.95 probability that the measured value $\overline{X}$ will be within 6 points (two standard deviations(2x3)) of the population mean μ.
Or similarly: There is a 0.95 probability that the population mean μ is within 6 points of $\bar{x}$, i.e. between $\bar{x}$-6 and $\bar{x}$+6.

So, if we measure $\bar{x} = 182 \rightarrow \bar{x} \pm 6 = [176, 188]$

Since 2σ = 6 in the example, we have 95% chance of having μ in this interval.

In this example, the interval of numbers between the values $[176, 188]$ is called a **<u>95% confidence interval for μ.</u>**

There are two possibilities:

- $[176, 188]$ contains the true μ → This happens 95% of the time
- $[176, 188]$ does not contain the true μ

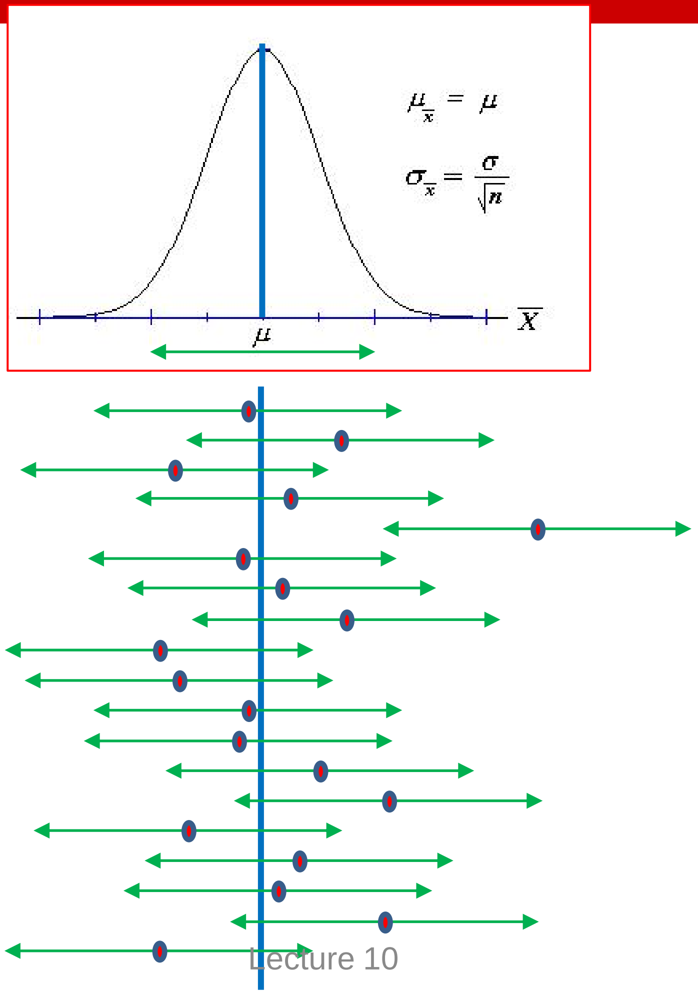
CI for Means (σ is known)

This is a very unlikely scenario. How would you NOT know the mean, μ, but somehow be clued in to σ?
But derivation of the formulas is simplified doing it this way first. 😉

- We want to know μ. So let's rearrange our favorite formula to solve for it.
- If we know $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ and we rearrange to solve for μ, we get $\mu = \bar{X} - Z * \frac{\sigma}{\sqrt{n}}$
- So assuming we know σ, we'll get $\bar{X}$ and n from the sample we take. So what do we use for Z? This is where we decide how certain we want to be.
- Empirical rule says that ± 2 standard deviations (Z = ± 2) is 95% (but actually it's 95.45%). If we use tables, we see that ± 1.96 standard deviations (Z = ± 1.96) is much closer to being exactly 95%.
- So using two values for Z (-1.96 and +1.96), we should get a lower and upper limit of what constitutes a good point estimate. This is a **confidence interval**.

What Does It Mean?

- So we just created the rules for a 95% confidence interval. How do you interpret this?
  - Recall that from my dice rolling experiment, I knew the standard error of the mean was 0.76. If I take ±1.96 of those standard errors, I should have 95% of all my samples fall within those bounds.
  - Said another way, 3.5 ± 1.96*0.76 = (2.003, 4.997) gives an interval where, if I did this experiment many times over and over (let's say 59 times), 95% of my sample means would fall between those values. FYI, y'all had 56 of 59 sample dice roll means between this… 94.9%
  - We *could* say that if my point estimate is between those values, it is a "good one."
- Turning that wording and thinking around to express μ (which is my stated goal), we would place our 59 point estimates where we currently see 3.5 and say that 95% of those intervals would contain the actual unknown value of μ.
- Let's look at a picture. I've gone cross-eyed. 😵



In the long run, 95% of all samples give an interval that contains μ

Each ⟷ corresponds to 95% CI calculated for a sample of size n

● indicates the $\bar{x}$ value for a specific sample

Interpretation

Interpreting a Confidence Interval

We must be careful to interpret confidence intervals correctly. There is a correct interpretation and many different and creative incorrect interpretations of the confidence interval $0.828 < p < 0.872$.

Correct: "We are 95% confident that the interval from 0.828 to 0.872 actually does contain the true value of the population proportion p." This means that if we were to select many different samples of size 1007 and construct the corresponding confidence intervals, 95% of them would actually contain the value of the population proportion p. (In this correct interpretation, the confidence level of 95% refers to the *success rate of the process* used to estimate the population proportion.)

Incorrect: "There is a 95% chance that the true value of p will fall between 0.828 and 0.872."

Incorrect: "95% of sample proportions will fall between 0.828 and 0.872."

Definition

Any **confidence interval** has two parts: an **interval** computed from the data and a **confidence level**.

- **Interval:** $\text{estimate} \pm \text{margin of error}$

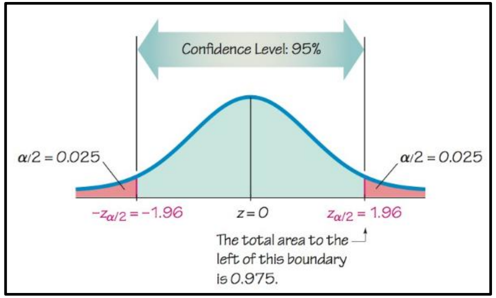
- **Confidence level, C:** states the probability that the method will give a correct answer (90%, 95%, etc.).
- We chose 95% initially because it made the math easy because of the empirical rule, but we are free to choose any level C.

Confidence Interval: A level C confidence interval for a parameter is an interval computed from sample data by a method that has probability C of producing an interval containing the true value of the parameter.

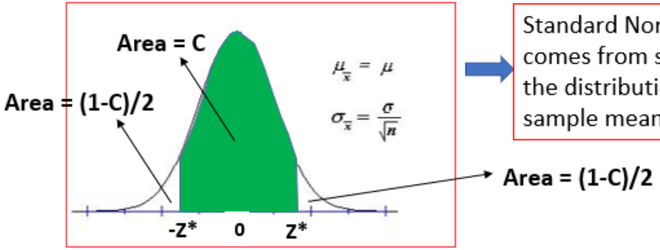
Confidence Interval for μ (σ is known)

- A 100(1 – α)% confidence interval for the mean μ of a normal population when the value of σ is known is given by:

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$



- Or equivalently, $\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
- The symbol, α, indicates the likelihood of error. In a 95% CI, there is a 5% chance of error… a 2.5% chance we're too low and a 2.5% chance we're too high. This is why z has a subscript of α/2. The area in the tails at each end is 0.025.



Standard Normal Curve that comes from standardizing the distribution of the sample mean!

Other Confidence Intervals

- The most commonly used CIs are 90%, 95%, and 99%. If we reverse lookup the area in the tails for α/2:

Confidence, C	90%	95%	99%
Alpha, α	10%, or 0.10	5%, or 0.05	1%, or 0.01
α/2	0.05	0.025	0.005
$z_{\alpha/2}$	1.645	1.960	2.576

- The confidence level is exactly C if the population distribution is Normal. When the population distribution is not Normal, the confidence level is approximately C when the sample size, n, is large.

Margin of Error

- If we combine all the terms after the ± symbol, we call this "margin of error."

$$m = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Relationship between Confidence level & Margin of error

(keeping σ, and n constant)

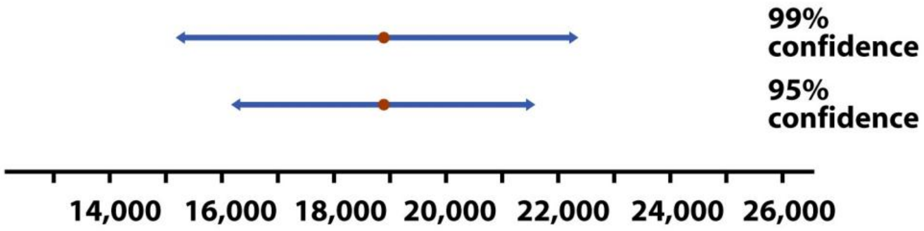


Figure 6-6
Introduction to the Practice of Statistics, Fifth Edition

As confidence level ↑, margin of error ↑.

Relationship between σ and the margin of error

(at fixed confidence level, i.e. $z_{\alpha/2}$=constant, and n constant)

$$m = z^{*}\frac{\sigma}{\sqrt{n}}$$

From this formula we can also see:

As the population standard deviation $\uparrow$, margin of error $\uparrow$.

Relationship between sample size & margin of error
(at fixed confidence level, and fixed σ)

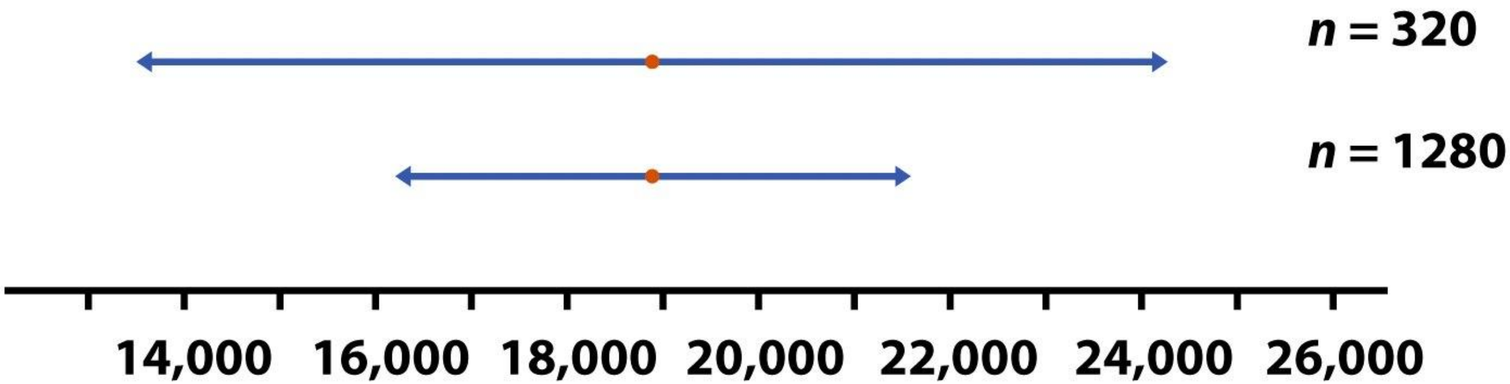


Figure 6-5
Introduction to the Practice of Statistics, Fifth Edition

$$m = z^* \frac{\sigma}{\sqrt{n}}$$

As sample size ↑, margin of error ↓.

Sample Size Determination

- We can choose the sample size to have a certain margin of error:
- Solving margin of error, m, for the variable n:

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

We need to <u>round up</u> n, if we want the margin of error (m) to be less or equal than a certain value!

Example

- A specialty tea shop is studying the possibility of opening up at a Hillsborough Street location. Before taking such a decision they want to study the market. Out of the entire population of Raleigh residents that are 18 years old and older, they select a SRS of 100. Each individual in the SRS is asked the question, “How many boxes of tea have you bought in the last three months?” The sample mean turns out to be 2.7. Let’s assume it is known that the standard deviation of the population is 2.3.

Example

- Calculate the 90% confidence interval for the mean number of tea boxes bought by the entire population that is 18 years old and older in Raleigh, in the last three months.

C-Level	$z_{\alpha/2}$
0.90	1.645
0.95	1.96
0.99	2.576

Example

- What should be the sample size if we want a margin of error of 0.1 for a 90% confidence interval?

Large Sample Interval for μ

- Recall the central limit theorem… it says even if X is not normally distributed, $\bar{X}$ is normally distributed given that the sample size is greater than 30.
- So extending that to a confidence interval, what this means is that we can use S instead of σ, since we rarely know σ anyway.
- If n is sufficiently large, the standardized variable

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has approximately a standard normal distribution.

Large Sample Interval for μ

- Therefore, we can say:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

is a large-sample confidence interval for μ with confidence level approximately 100(1 – α)%. This formula is valid regardless of the shape of the population distribution.

- We use a more conservative rule of $n>40$ for this scenario (as opposed to CLT where we said $n>30$) because of the additional variability introduced by using S in place of σ.

Summary

- A **level C confidence interval for a parameter** is an interval computed from data obtained from one SRS sample by a method that has probability C of producing an interval containing the true value of the parameter. Any **confidence interval** has two parts: an **interval** computed from the data and a **confidence level**. The interval is given by the **estimate** (+/-) a **margin of error**. The confidence level, given by the area under the sample mean distribution, states the probability that the method will give a correct answer. The margin of error increases as the confidence level increases and as the population standard deviation increases. The margin of error decreases as the sample size increases. It is possible to choose the sample size to have a certain margin of error with a certain confidence level.