Introduction to Statistics I

Instructor: Jodin Morey   moreyj@lemoyne.edu

Previous Lecture

- Distributions of Quantitative Vars
- CLT: Shape/Center/SD, if pop. is normal or $n \geq 30$.
- Symbols: π, μ, σ, $\hat{p}$, $\bar{x}$, s

Topic 19: Confidence Intervals for Means

CLT Comparison

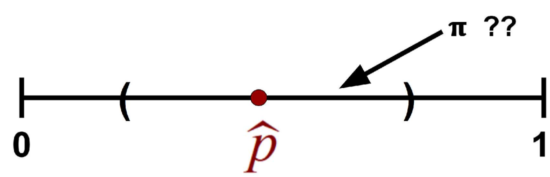
	Categorical Var	Quantitative Var
Sample Statistic	$\hat{p}$ (proportion)	$\bar{x}$ (mean)
 **Tech Cond: SRS and:**	When $n\pi \geq 10$ and $n(1-\pi) \geq 10$ or $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$	When population is normal or $n \geq 30$
Shape of Sampling Distr	Approximately Normal	Normal or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

Recall Confidence Intervals (CIs)

For proportions, we used CLT to develop CIs to capture π, near $\hat{p}$.

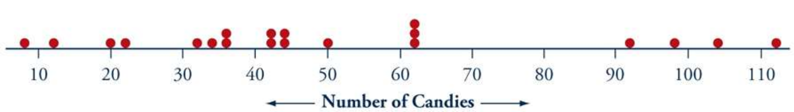
CIs for Proportions: $\hat{p} \pm z^* se$, where $\hat{p}$ is statistic, $se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$,

and z^* is the critical value, which depends on **confidence level**.



So a CI consists of three parts: A statistic $\hat{p}$, standard error, and a critical value z^*.
Can we do something similar for quantitative variables?

M&Ms Example: Research into factors that motivate people to eat unthinkingly. One study involved 20 students, and invited them to take as many M&Ms from a large bowl as they wanted.



They got the following data:

n	Mean	SD	Min	Q_L	Median	Q_U	Max
20	50.15	30.31	7.00	32.35	42.50	62.00	111.00

Suppose we plan a party. It might be helpful to know:

RQ: What is average # of M&Ms students take from a large bowl?

Population, Var, Var Type, Parameter, Symbol ?

- **Population**: All students
- **Var**: How many M&Ms did they take?
- **Var Type**: Quantitative
- **Parameter**: Average # of M&Ms eaten by all students
- **Symbol**: μ

To create CI for μ, we need three things: (similar to catagorical vars)

- Statistic,
- Standard error (*se*), and
- Critical value.

Statistic

Since we're looking at a quantitative var and an average, we need a sample mean. Symbol?

$\bar{x}$

In the example, the sample mean is 50.15, so the statistic for our CI is $\bar{x} = 50.15$.

Standard Error (se)

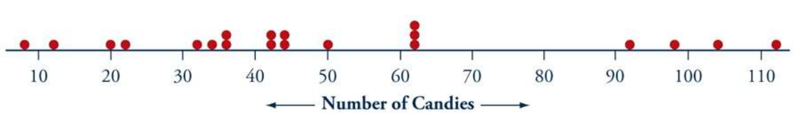
In CLT for means, the SD of $\bar{x}$ is: $\frac{\sigma}{\sqrt{n}}$, where σ is the population SD.

However, σ is a parameter, and is thus **unknown**. How can we calculate $\frac{\sigma}{\sqrt{n}}$?

Instead, we look at sample SD, which measures variation we *actually* observe.

Symbol for the sample SD?

$s = 30.31$. This measures the spread in these 20 data pts.



If sample is representative, sample SD should be similar to population SD σ.

If we replace unknown param σ with observed SD s, we have **standard error**: $se = \frac{s}{\sqrt{n}}$.

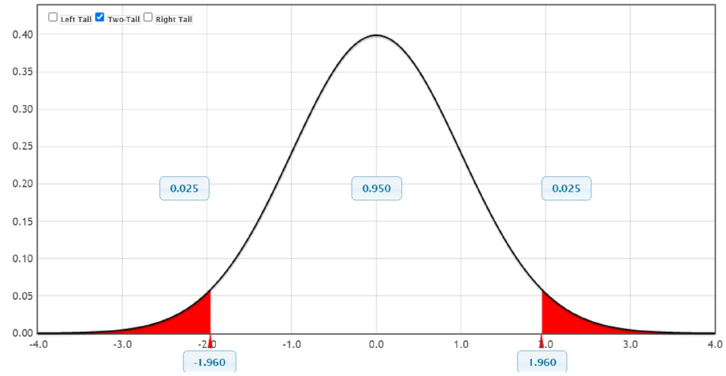
Recall: to create CI, we need three things:

- Statistic: $\overline{x} = 50.15$
- Standard error: $se = \frac{s}{\sqrt{n}} = \frac{30.31}{\sqrt{20}} \approx 6.78$.
- And a critical value (was called z^* for categorical).

Critical Values (t^*)

Recall For *proportions,* critical values (z^*) come from the Normal dist, and the CLT stated that the sampling distr of $\widehat{p}$ is normal.

Eg: we use 1.96 for 95% CI, because 95% of pts are within 1.96 SDs of π w/normal distr.



For *means*, the CLT states that the sampling distr of $\overline{x}$ is

- Normal, w/the
- Mean at μ, and a
- SD of $\frac{\sigma}{\sqrt{n}}$.

Can we use the same critical values? They tell us the # of SDs to get different confidence levels.

SD is $\frac{\sigma}{\sqrt{n}}$, but we don't know actual spread of the population (σ).

Instead, we estimate by substituting in s for σ instead: $se = \frac{s}{\sqrt{n}}$. This has consequences for quantitative critical values.

Let's Experiment: At this link, switch "Statistic" to "Means."

Select "Method" called "z with σ." $\left(\frac{\sigma}{\sqrt{n}}\right)$

Set: Pop mean (μ) to 50, pop SD to 30.

Sample size to 20, confidence level to 95%.

Ask for 1000 intervals, click "Sample."

bit.ly/introstatsdata

Simulating Confidence Intervals

What % of the 1000 CIs captured μ?

Now repeat with "Method" called "z with s." $\left(\frac{s}{\sqrt{n}}\right)$

What % of the 1000 intervals capture μ?

Repeat w/$n = 10$, then $n = 30$.

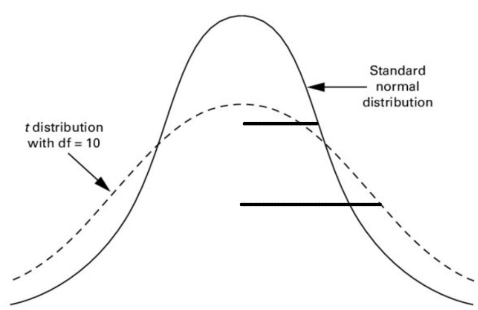
Take-away: CIs are innacurate (too short) when n is small and when we use s instead of σ !! What to do?

We could make bigger z^* values. See the following...

---

t-Distr

t-distr is similar to normal distr, but shorter & fatter.



t-distr critical values (t^*) are slightly bigger than normal ones (z^*).

Acts as penalty for estimating SD from sample.

How much penalty do we need? Depends on the sample size.

Degrees of Freedom

Higher sample size $\Rightarrow$ More accurate estimate of SD.

So, w/higher sample size, we need less penalty than w/smaller sample size.

There's a t-distr for every sample size. They are labeled by their **degrees of freedom** (df).

$df = n - 1$.

Critical value is denoted t^* or t^*_{n-1}, with df in subscript.

Back to M&Ms CI

Recall for M&Ms data set, we found:

n	Mean	SD	Min	Q_L	Median	Q_U	Max
20	50.15	30.31	7.00	32.35	42.50	62.00	111.00

What *df* should we use?

$n = 20$, so $df = 20 - 1 = 19$.

If we want 95% CI, we find:

$t^*_{19} = 2.093$.

bit.ly/introstatsdata

t-dist Calculator

Statdistributions.com/t/

two tails, and p-value $= 1 - 0.95 = 0.05$

CIs for Means

Recall, CIs have form: statistic ± (critical value) × (standard error).

So, CI is: $\bar{x} \pm t^*_{n-1}\left(\frac{s}{\sqrt{n}}\right)$

So, to create CI for M&Ms:

Statistic: $\bar{x} = 50.15$,

SD: $se = \frac{s}{\sqrt{n}} = \frac{30.31}{\sqrt{20}} \approx 6.778$,

Critical value: $t^* = 2.093$.

So 95% CI for μ is:

$\bar{x} \pm t^*_{n-1}\left(\frac{s}{\sqrt{n}}\right) = 50.15 \pm 2.093(6.778) = 50.15 \pm 14.19$

$(35.96, 64.34)$. Meaning (in context)?

We're 95% confident that mean # of candies chosen by students is between 35.96 & 64.34.

Tech. **Conds for CI**:

- Simple Random Sample
- Normal population *OR* $n \geq 30$.

Activity 19-2

Another Example: Take a sample of students, ask them how many hours of sleep they got last night. You find a sample mean of 6.6 hrs and sample SD of 0.825.

Task: make 95% CI for mean # of hrs of sleep.

Case 1: $n = 10, \quad \bar{x} = 6.6, \quad s = 0.825.$

se, df, CI ??

$$se = \frac{s}{\sqrt{n}} = \frac{0.825}{\sqrt{10}} \approx 0.2609.$$

$$df = 9$$

$$\bar{x} \pm t^*\left(\frac{s}{\sqrt{n}}\right) = 6.6 \pm 2.262(0.2609)$$

So, CI is $(6.010, \ 7.190)$.

Case 2: $n = 30, \quad \bar{x} = 6.6, \quad s = 0.825$

se, df, CI ??

$$se = \frac{s}{\sqrt{n}} = \frac{0.825}{\sqrt{30}} \approx 0.1506.$$

$$df = 30 - 1 = 29$$

$$t^* = 2.045$$

$$\bar{x} \pm t^*\left(\frac{s}{\sqrt{n}}\right) = 6.6 \pm 2.045(0.1506)$$

So, CI is $(6.292, \ 6.908)$.

Sample Size Considerations

When sample size is *smaller*, the resulting CI is *wider* for two reasons:

- Standard error is larger, and
- Critical value t^* is higher.

So *larger* samples sizes *reduce* the width of CI for both reasons.

What did we learn?

- CIs for quantitative vars
- Standard error (quant): $se = \frac{s}{\sqrt{n}}$
- Degrees of freedom (df), Critical values t^* or t^*_{n-1}
- t-distribution