Introduction to Statistics I

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

Previous Lecture

- 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, $moe = z^*(\widehat{se})$

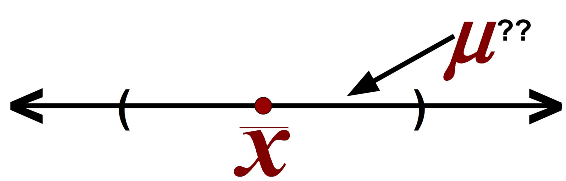
§8.1 and §8.2: Confidence Intervals for Population Means

As we did for qualitative vars, we now develop CIs for quantitative vars. Notice the similarities!

How do we estimate the average μ of some value in a population (temperature, size, weight, 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 calculate $\bar{x}$. But, this is probably *not a completely accurate estimate*.
It's another *point estimate*: a single number used to estimate the value of an unknown parameter.

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



Below, we discuss how to create this interval.

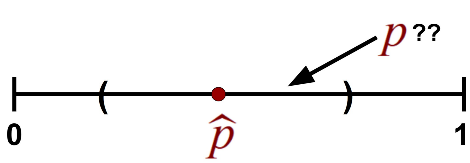
Recall Confidence Intervals (CIs)

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

CIs for Portions: $\hat{p} \pm z^*(\widehat{se})$ where $\hat{p}$ is sample proportion,

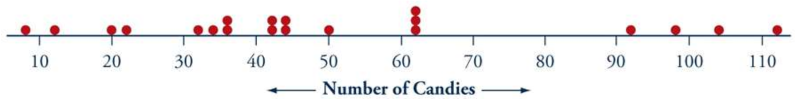
z^* is the desired # of SDs (called the **critical value**), and $\widehat{se} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

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



So the CI consisted of three parts: A statistic $\hat{p}$, a critical value, and a standard error.
Can we do something similar for quantitative vars?

M&Ms Example. Research was done into factors that motivate students to eat unthinkingly. One study involved a SRS of 20 students, inviting them to take as many M&Ms from a large bowl as they wanted.



They found the mean to be 50.15, and the SD to be 30.13.

But how accurate is the mean as an estimate of μ?

RQ: Could it be that students actually take $\mu = 60$ M&Ms on average?

Recall that **CLT** tells us (if tech conds are met) about the shape/center/spread of the sampling distr.

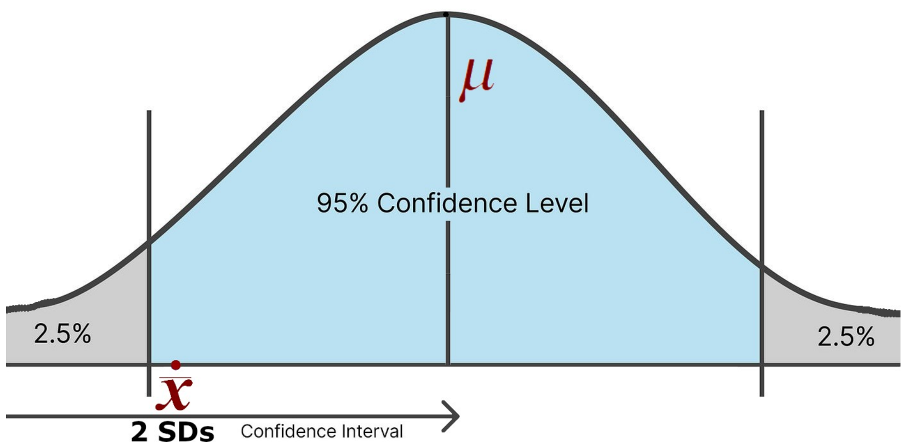
In particular that the $\bar{x}$ distr is ≈normal.

Also recall the **empirical rule**: 68% of the $\bar{x}$ are within 1 SD of μ, 95% within 2 SDs, nearly all within 3 SDs.

So, in 95% of samples, the statistic $\bar{x}$ is at most 2 SDs away from μ.

Thus, using the empirical rule, for each sample $\bar{x}$, if we add/subtract 2 SDs ($\bar{x} \pm 2se$), then for about 95% of samples, the interval $(\bar{x} - 2se,\ \bar{x} + 2se)$ will contain μ.

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



95% of $\bar{x}$s (which are in the blue region) will capture μ with a 2SD Conf Interv

Therefore, the **CI Formula** is: $\bar{x} \pm z^*(se)$ where $\bar{x}$ is sample mean,

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

So, in interval notation CI is $\left(\bar{x} - z^*(se),\ \bar{x} + z^*(se)\right)$.

Recall Confidence Levels

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

Back to our Example. Recall: $n = 20$, and $\bar{x} = 50.15$. Now assume we know the population SD $\sigma = 32.35$.
How do we calculate the 95% CI?

The 95% **CI** is $\left(\bar{x} - z^*(se),\ \bar{x} + z^*(se)\right) = \left(50.15 - 1.960(\frac{32.35}{\sqrt{20}}),\ 50.15 + 1.960(\frac{32.35}{\sqrt{20}})\right)$
$\approx (35.97,\ 64.33)$. In context?

“We’re 95% confident that the average number of M&Ms taken by all students is between 35.97 and 64.33."
What’s the 99% **CI**?

The 99% **CI** is $\left(\hat{p} - z^*(se),\ \hat{p} + z^*(se)\right) = \left(50.15 - 2.567(\frac{32.35}{\sqrt{20}}),\ 50.15 + 2.567(\frac{32.35}{\sqrt{20}})\right) \approx (31.58, 68.72)$.

What do we mean by "95% confident"? Why does it make sense that the 99% CI is bigger?

---

§8.2: CIs for Population Means

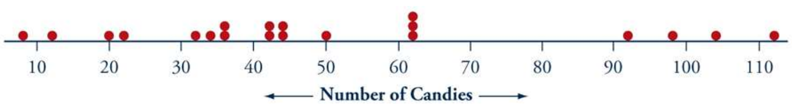
Minor Obstacle

Formula for *se* is: $\frac{\sigma}{\sqrt{n}}$. But this relies on unknown parameter σ (!?!).

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

Instead, we look at the sample SD, which measures variation we *actually* observe. Symbol for sample SD?

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



So we have the *estimated* standard error: $\overline{se} = \frac{s}{\sqrt{n}}$.

But can we now calculate our *moe* as $z^*\overline{se}$? Unfortunately, there’s still a bit of a glitch!

Critical Values (t^*)

Recall that with CLT and SD $\frac{\sigma}{\sqrt{n}}$, the sampling distr of $\bar{x}$ was *normal* and the critical values were z^*.

Now that we’re estimating w/$\overline{se} = \frac{s}{\sqrt{n}}$, does this work w/the normal distr?
Can we use the same critical values?

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

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

Set: Pop mean (μ) to 60, pop SD to 32.35.

Sample size to 20, confidence level to 95%.

Ask for 10,000 intervals, then click "Sample."

bit.ly/introstatsdata

Applets: Simulating Confidence Intervals

Under "Results": what % of the 10,000 CIs captured μ?

Now click "Reset" and repeat with "Method" called "z with s." $(\frac{s}{\sqrt{n}})$

What % of the 1000 intervals capture μ?

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

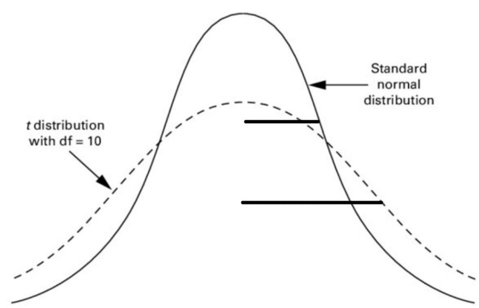
Take-away: CIs are innacurate (too short) when we use s instead of σ (especially when n is small)!! What to do?

We could make bigger z^* values to compensate. This leads to ...

---

t-Distr

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



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

They compensate for estimating se from a sample.

How much do we need to compensate? As we saw in our simulation, it depends on sample size n.

Degrees of Freedom

Higher sample sizes give us an $\overline{se}$ that is a more accurate estimate of se.

So, w/higher sample size, t^* should be smaller (closer to z^*) since we need to compensate less.

Thus, there's a t-distr for every sample size. They are distinguished by their **degrees of freedom** (df).

$df = n - 1$.

As a result, critical values are denoted t^*_{n-1}, with the degrees of freedom in the subscript.

Back to the M&Ms CI

Recall for M&Ms, w/a sample size of 20, we found the sample mean to be 50.15 M&Ms, and the SD to be 30.13 M&Ms. What *df* should we use?

bit.ly/introstatsdata

Applet: *t*-Dist Calculator

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

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

So, what is t^*_{19} for a 95% CI?

$t^*_{19} = 2.093$ How do we calculate the 95% CI?

CIs for Means and s

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

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

So, to create a 95% CI for M&Ms:

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

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

Critical value: $t^*_{19} = 2.093$.

So 95% CI for μ is:

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

$(35.96,\ 64.34)$. In context?

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

Activity: **8.2b**

Optional Activity: **8.2a**

Sample Size Considerations

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

- Standard error is larger, and
- Critical value t^*_{n-1} is larger.

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

What did we learn?

- Intuition for confidence Intervals for quantitative vars
- Estimated standard error when we don't know σ: $\overline{se} = \frac{s}{\sqrt{n}}$
- t-distribution, degrees of freedom (df), critical values t^*_{n-1}
- CI is: $\overline{x} \pm t^*_{n-1}\left(\frac{s}{\sqrt{n}}\right)$ (if tech conds are met)