TWO SAMPLE CONFIDENCE INTERVALS

Lecture 15

Instead of investigating one population, lets consider studying two opposed populations:

Men and Women, Democrats and Republicans, Graduates and Non-graduates,

Rich and Poor, Brand X and Brand Y, Children and Adults, etc.

When the characteristic being compared is numerical (height, weight, income, age…) then the object of interest is the amount of difference in the averages (means) for the two populations.

We want to know $\mu_1 - \mu_2$ (the difference in the two population means).
We estimate that parameter with the statistic $\bar{x}_1 - \bar{x}_2$ (the difference in two sample means) plus or minus a sample error.

The result is a *confidence interval for the difference between two population means*.

Statistical Inference

Goal:
To find a CI for the difference between the mean of two populations, or
To get the CI for the difference between the proportions of two populations.

Difference between two population means:
Collect samples from both populations and perform inferences on both samples to make conclusions about $\mu_1 - \mu_2$.

Basic Assumptions:

1] $X_1, X_2, \dots, X_m$ is a random sample from distribution 1 with mean μ_1 and SD σ_1.

2] $Y_1, Y_2, \dots, Y_n$ is a random sample from distribution 2 with mean μ_2 and SD σ_2.

The X and Y samples are independent of one another.

When we are after the population mean difference, $\mu_1 - \mu_2$, we use $\bar{x}_1 - \bar{x}_2$.

How good of an estimator is $\bar{x}_1 - \bar{x}_2$?

It is considered an unbiased estimation for $\bar{X} - \bar{Y}$ because:

$$E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}] = \mu_1 - \mu_2$$

Similarly, the SD of $\bar{X} - \bar{Y}$ is:

$$SD = \sqrt{Var(\bar{X} - \bar{Y})} = \sqrt{Var(\bar{X}) + Var(\bar{Y})} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Notice this is the just the pooled standard deviation from two single random variables variances.

confidence interval for the difference between two population means

If both of the population standard deviations are known, then the formula for a CI for the difference between two population means (averages) is:

$$\bar{X} - \bar{Y} \pm Z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Recall the CI for μ of one population: $\bar{X} \pm Z_{\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}} \quad = \quad \bar{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Normal populations with known standard deviations

If both populations are normal, then $\bar{X}$ and $\bar{Y}$ are both normally distributed. Independence of the two samples implies that the samples means are also independent.

Conclusion:
The difference in sample means is normally distributed, for any sample size, with:

$$(\bar{X} - \bar{Y}) \sim N\left(\mu_1 \; - \; \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

And as before, a standardized $\bar{X} - \bar{Y}$ gives a standard normal random variable:

$$Z = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}$$

If we don't know σ_1 or σ_2 and we have "enough" samples $(n \geq 30)$ then we estimate using s_1 and s_2.

Essentially, having normally distributed differences in means implies we can do all the stuff we were doing before: standardizing, confidence intervals, z-scores,…

Suppose we want to know which variety of daisy stems grow longer:

Those from Duke Daisy brand daisies, or those from Daisy Chain Brand daisies.

Let's create a 95% CI for the for the difference between the mean length of the daisy's stems.

Suppose your random sample of 100 stems of the Dukes variety averages 8.5 inches, and your random sample of 110 stems of Daisy Chain averages 7.5 inches.

From prior research we know that the population standard deviation for Duke Brand Daisy's is 0.35 inches, and the population standard deviation for Daisy Chain Brand daisy's is 0.45 inches.

So, the information you have is:

$\bar{x}_1 = 8.5, \quad \sigma_1 = 0.35, \quad n_1 = 100$ and $\bar{x}_2 = 7.5, \quad \sigma_1 = 0.45, \quad n_1 = 110$

$$Z\alpha_{/_2} = Z_{.05/_2} = Z_{.025} = 1.96$$

$\bar{X} - \bar{Y} \pm Z\alpha_{/_2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ Leaves us with: $0.8915 < \mu_1 - \mu_2 < 1.1085$

Interpretation: we can say with 100% confidence that 95% of intervals generated in this manner show that the Duke brand daisy's stem is longer, on average, than the Daisy Chain variety, by somewhere between 0.8915 and 1.1085 inches based on this particular sample.

The temptation is to say, "Well, I knew Dukes daisies were longer because their sample mean was 8.5 inches and Daisy Chain was only 7.5 inches on average.

Why do I even need a confidence interval?"

The only thing those two numbers really tell you is something about the 210 daisy stems in the sample.
You also need to factor in variation using the margin of error to be able to say something about the entire populations of daisy stems.

For example, what if Dukes daisies had a sample with an average of 7.6 inches?
That is still longer than Daisy Chains sample mean of 7.5 inches, however…

Suppose your random sample of 100 stems of the Dukes variety averages 7.6 inches, and your random sample of 110 stems of Daisy Chain averages 7.5 inches.

From prior research we know that the population standard deviation for Duke Brand Daisy's is 0.35 inches, and the population standard deviation for Daisy Chain Brand daisy's is 0.45 inches.

So, the information you have is:

$\bar{x}_1 = 7.6, \quad \sigma_1 = 0.35, \quad n_1 = 100$ and $\bar{x}_2 = 7.5, \quad \sigma_1 = 0.45, \quad n_1 = 110$

$$Z_{\alpha/2} = Z_{.05/2} = Z_{.025} = 1.96$$

$\bar{X} - \bar{Y} \pm Z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$ Leaves us with: $-0.0085 < \mu_1 - \mu_2 < 0.2085$

Interpretation: 95% of intervals generated in this manner show that the Duke brand daisy's stem is <u>**no different**</u>, on average, than the Daisy Chain variety, based on this particular sample.

So, Duke brand is longer by 0.2085 or Daisy Chain is longer by 0.0085 inches…

Suppose we run an email ad campaign. Let's call this $\text{Ad}1$.
We send, on 50 different days an ad that generates an average of 2 million page views per day with SD of 1 million page views.

Find a 95% confidence interval for the average page views.

Suppose we run an email ad campaign. Let's call this Ad1.
We send, on 50 different days an ad that generates an average of 2 million page views per day with SD of 1 million page views.
Find a 95% confidence interval for the average page views.

ANS:

$$\bar{X}_1 = 2, \qquad SD_1 = 1, \qquad n_1 = 50$$

$$\bar{X}_1 \pm Z_{\alpha/_2} \cdot \frac{s_1}{n_1} = 2 \pm 1.96 \cdot \frac{1}{\sqrt{50}} = [1.723, 2.277]$$

Interpretation: We are 100% confident that 95% of intervals constructed in this manner should contain the population mean number of page views.
This is one such interval; it was created from this particular sample's statistics.

BTW, this is not new to this lecture. This is what we did last time (one sample CI)

Suppose we run an email ad campaign. Let's call this Ad2.
We send, on 40 different days an ad that generates an average of 2.25 million page views per day with SD of ½ million page views.

Find a 95% confidence interval for the average page views.

(again, this is just another one sample CI)

Suppose we run an email ad campaign. Let's call this Ad2.
We send, on 40 different days an ad that generates an average of 2.25 million page views per day with SD of ½ million page views.
Find a 95% confidence interval for the average page views.

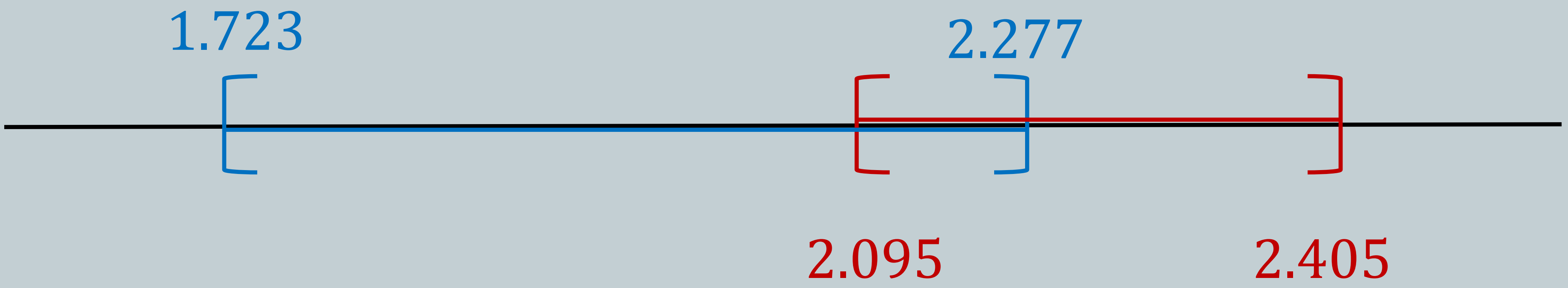
ANS:

$$\bar{X}_2 = 2.25, \qquad SD_2 = 0.5, \qquad n_2 = 40$$

$$\bar{X}_2 \pm Z\alpha_{/2} \cdot \frac{s_2}{n_2} = 2.25 \pm 1.96 \cdot \frac{0.5}{\sqrt{40}} = [2.095, 2.405]$$

Interpretation: We are 100% confident that 95% of intervals constructed in this manner should contain the population mean number of page views.
This is one such interval, created from this particular sample's statistics.

Compare Ad1 and Ad2: $[1.723, 2.277]$ and $[2.095, 2.405]$



The two intervals overlap!

So, Ad2 could be better… or they could be the same… or Ad1 could be better…

Let's do the same example again by finding a 95% CI for the difference in average page views:

$\bar{X}_1 = 2,\ SD_1 = 1,\ n_1 = 50$
$\bar{X}_2 = 2.25,\ SD_2 = 0.5,\ n_2 = 40$

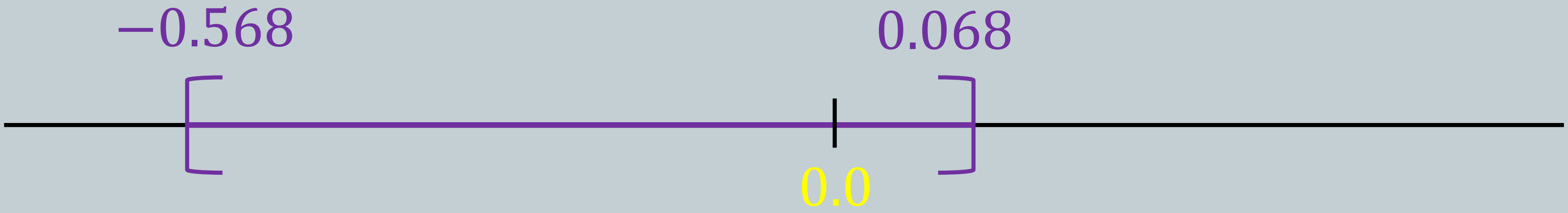
$$\bar{X}_1 - \bar{X}_2 \pm Z_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

This produces the interval: $[-0.568, 0.068]$.

Notice this interval contains 0 !

The mean difference in page views is a number somewhere in the interval below which contains 0. That is, a 0 difference!

95% of intervals constructed in this manner should contain the difference in population means for number of page views. And this one contains a difference of 0.



Now, instead of finding confidence intervals for a difference in the means of two populations, let’s use similar ideas on a proportion of interest in a population.

Consider two ranchers, Bill and Bob.
Both Bill and Bob are buffalo ranchers, and they both have large herds.

Bill's herd has M buffalo in it, with x of these buffalo being female.
Bob's herd has N buffalo in it, with y of these buffalo being female.
Which herd has more female buffalo?

Population 1 (Bill's buffalo) have a proportion of female bison: $p_1 = \frac{x}{M}$
Population 2 (Bob's buffalo) have a proportion of female bison: $p_2 = \frac{y}{N}$

When we don't know the proportion of interest, we can take samples from these herds to estimate the population proportion.

Comparing population proportions (instead of means)

Suppose a sample of size m is selected from population 1.
Suppose a sample of size n is selected from population 2.

Let X denote the number of units with the characteristic of interest in population 1.
Let Y denote the number of units with the characteristic of interest in population 2.

For each population, we can reasonably estimate the proportion, p, with the sample proportion $\hat{p}$.

Therefore, a natural estimator for the difference in population proportions $p_1 - p_2$, would be the sample proportions difference, $\hat{p}_1 - \hat{p}_2$

Where $\hat{p}_1 = \frac{X}{m}$ and $\hat{p}_2 = \frac{Y}{n}$

Suppose $\hat{p}_1 = \frac{X}{m}$ and $\hat{p}_2 = \frac{Y}{n}$.

Then $E[\hat{p}_1 - \hat{p}_2] = E[\hat{p}_1] - E[\hat{p}_2]$

$$= E\left[\frac{X}{m}\right] - E\left[\frac{Y}{n}\right]$$

$$= \frac{1}{m}E[X] - \frac{1}{n}E[Y]$$

$$= \frac{1}{m}(mp_1) - \frac{1}{n}(np_2)$$

$$= p_1 - p_2$$

In other words:

A natural estimator for the difference in population proportions $p_1 - p_2$, would be the difference in sample proportions $\hat{p}_1 - \hat{p}_2$

Similarly,

$$\begin{aligned} Var(\hat{p}_1 - \hat{p}_2) &= Var(\hat{p}_1) + Var(-\hat{p}_2) \\ &= Var(\hat{p}_1) + Var(\hat{p}_2) \\ &= Var\left(\frac{X}{m}\right) + Var\left(\frac{Y}{n}\right) \\ &= \frac{1}{m^2} Var(X) + \frac{1}{n^2} Var(Y) \\ &= \frac{1}{m^2} mp_1(1 - p_1) + \frac{1}{n^2} np_2(1 - p_2) \\ &= \frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n} \end{aligned}$$

Therefore, $SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}} \approx \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$

We generally don’t know p_1 or p_2 so we estimate using $\hat{p}_1$ and $\hat{p}_2$.

These calculations allow us to use the confidence interval for a proportion:

$$\hat{p}_1 - \hat{p}_2 \pm Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$$

When comparing two independent population proportions, the following characteristics should be present:

1] The two independent samples are simple random samples that are independent.
2] The number of successes is at least five, and the number of failures is at least five, for each of the samples.
3] Growing literature states that the population must be at least 10 or 20 times the size of the sample. This keeps each population from being over-sampled and causing incorrect results.

Comparing two proportions, like comparing two means, is common.
If two estimated proportions are different, it may be due to a difference in the populations, or it may be due to chance.

A hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the population proportions.

The difference of two proportions follows an approximate normal distribution.

Two types of medication for hiccups are being tested to determine if there is a **difference in the proportions of adult patient reactions**.

Twenty out of a random **sample of 200** adults given medication A (peanut butter) still had hiccups 15 minutes after taking the medication.

Twelve out of another **random sample of 200 adults** given medication B (standing on head) still had hiccups 15 minutes after taking the ‘medication’.

What is the 99% CI for this difference in proportions?

Is eating standing on your head better than eating peanut butter for getting rid of hiccups?

Medication A: $\hat{p}_1 = \frac{20}{200}$ Medication B: $\hat{p}_2 = \frac{12}{200}$

$$\hat{p}_1 - \hat{p}_2 \pm Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$$

$$Z_{\alpha/2} = Z_{.01/2} = Z_{.005} = 2.575$$

$$\hat{p}_1 = \frac{20}{200} \Longrightarrow (1-\hat{p}_1) = \frac{180}{200}$$

$$\hat{p}_2 = \frac{12}{200} \Longrightarrow (1-\hat{p}_2) = \frac{188}{200}$$

Therefore, we get: $[-0.02969,\ 0.10969]$

The 99% CI infers, from the sample data, there is not sufficient evidence to conclude that there is a difference in the proportions of adult patients who did not react after 15 minutes to medication *A* and medication *B*.

The New England Journal of medicine published a study on Cancer treatment:

Chemotherapy only: 154 patients, 76 survived

Chemotherapy plus radiation: 165 patients, 98 survived

Is one of these treatments better than the other?

$$\frac{76}{154} = 0.4935 = 49.4\% \text{ versus } \frac{98}{165} = .5939 = 59.4\%$$

What is the 99% CI for this difference in proportions?

$\hat{p}_1 = \frac{76}{154}$ and $\hat{p}_2 = \frac{98}{165}$ with $n_1 = 154$ and $n_2 = 165$

$$\hat{p}_1 - \hat{p}_2 \pm Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Two-Tail test

$Z_{\alpha/2} = Z_{.01/2} = Z_{.005} = 2.575$ or `ppf(0.99+0.005) = 2.576`

$$\hat{p}_1 = \frac{76}{154} \Longrightarrow (1-\hat{p}_1) = \frac{78}{154}$$

$$\hat{p}_2 = \frac{98}{165} \Longrightarrow (1-\hat{p}_2) = \frac{67}{165}$$

Therefore, we get: $[-0.247,\ 0.039]$ which contains 0!

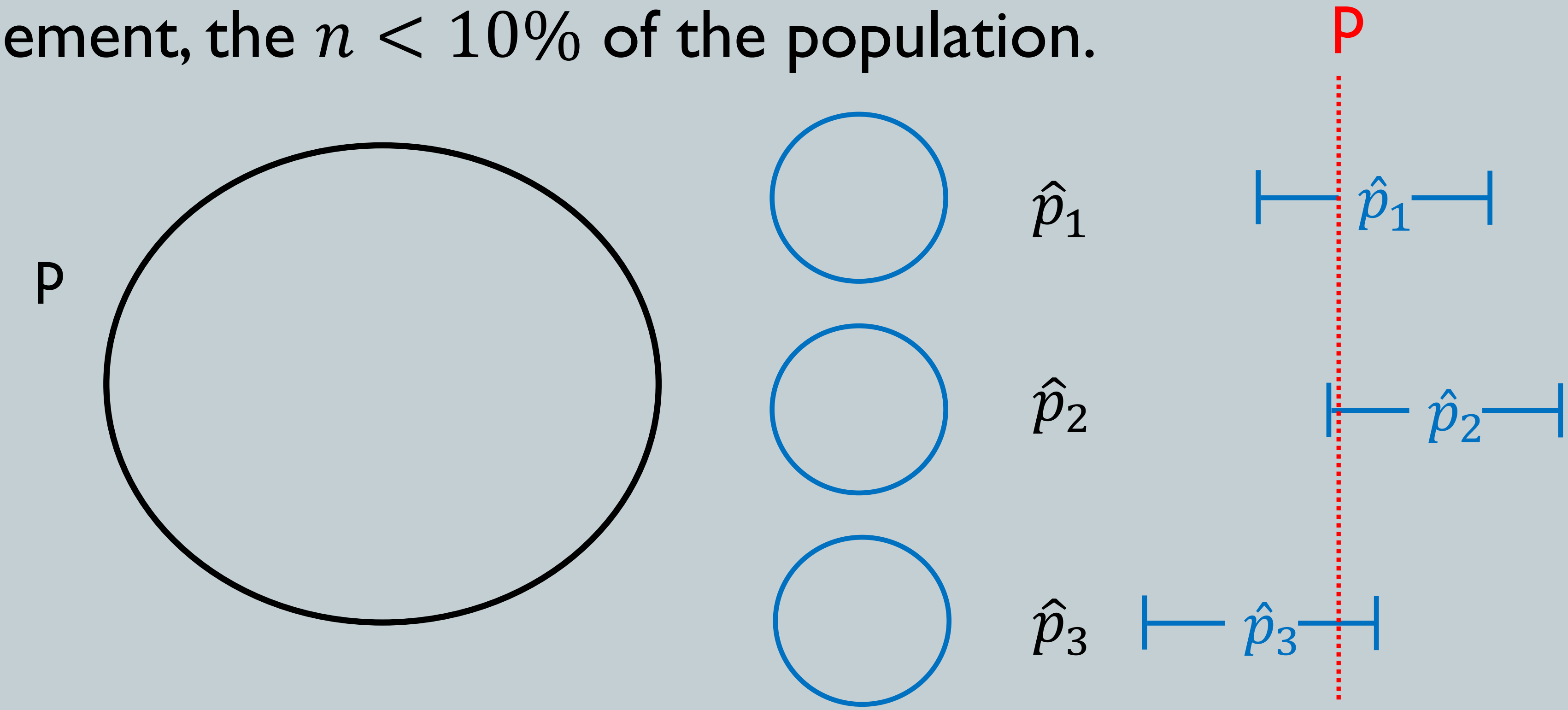
Neither treatment can be considered to better the other.

Conditions for creating a confidence Interval:

1] You must be drawing random sample

2] “Normal Condition” Expect at least 5 success and failures

3] The samples are independent. This might require replacement. If there is no replacement, the $n < 10\%$ of the population.



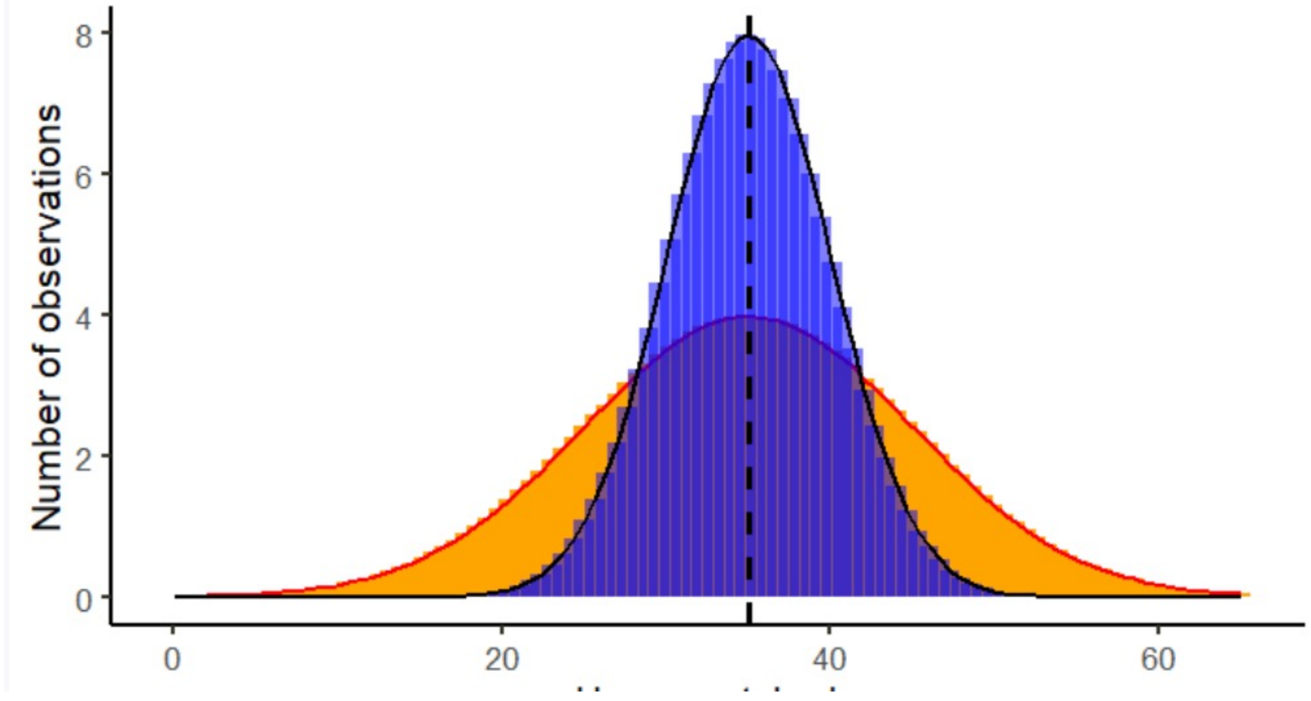
We have built CI’s for the mean, the difference in means, proportions, and the difference in proportions.

Confidence is just another word for probability.
We are ‘confident’ that 95 out of 100 times the statistic is in the interval.

A confidence interval is the mean of your estimate plus or minus the variation in the estimate.

Variation makes all the difference in interval size.

- Suppose you survey 100 workers in town A and from this sample you find the mean hours worked per week is 35.
- Suppose you survey 100 workers in town B and from this sample you find the mean hours worked per week is 35.
- This is not enough information to create a confidence interval for the actual number of hours worked by the population of either town A or town B.
- They can have different variations in their distributions.

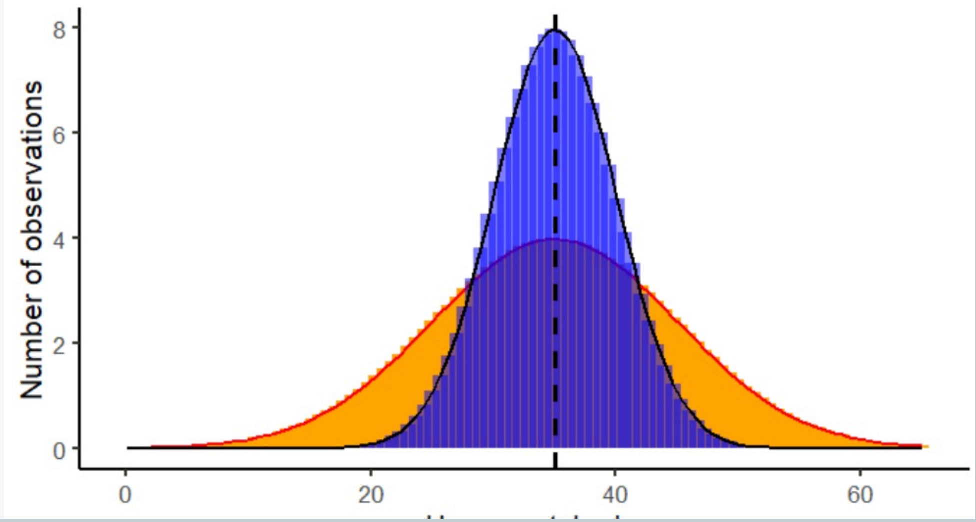


From this same survey you might find:
Town A has a standard deviation of 5, while
Town B has a standard deviation of 10.

To create a confidence interval, you need a point estimate, a critical value for the test statistic, a standard deviation, and a sample size.

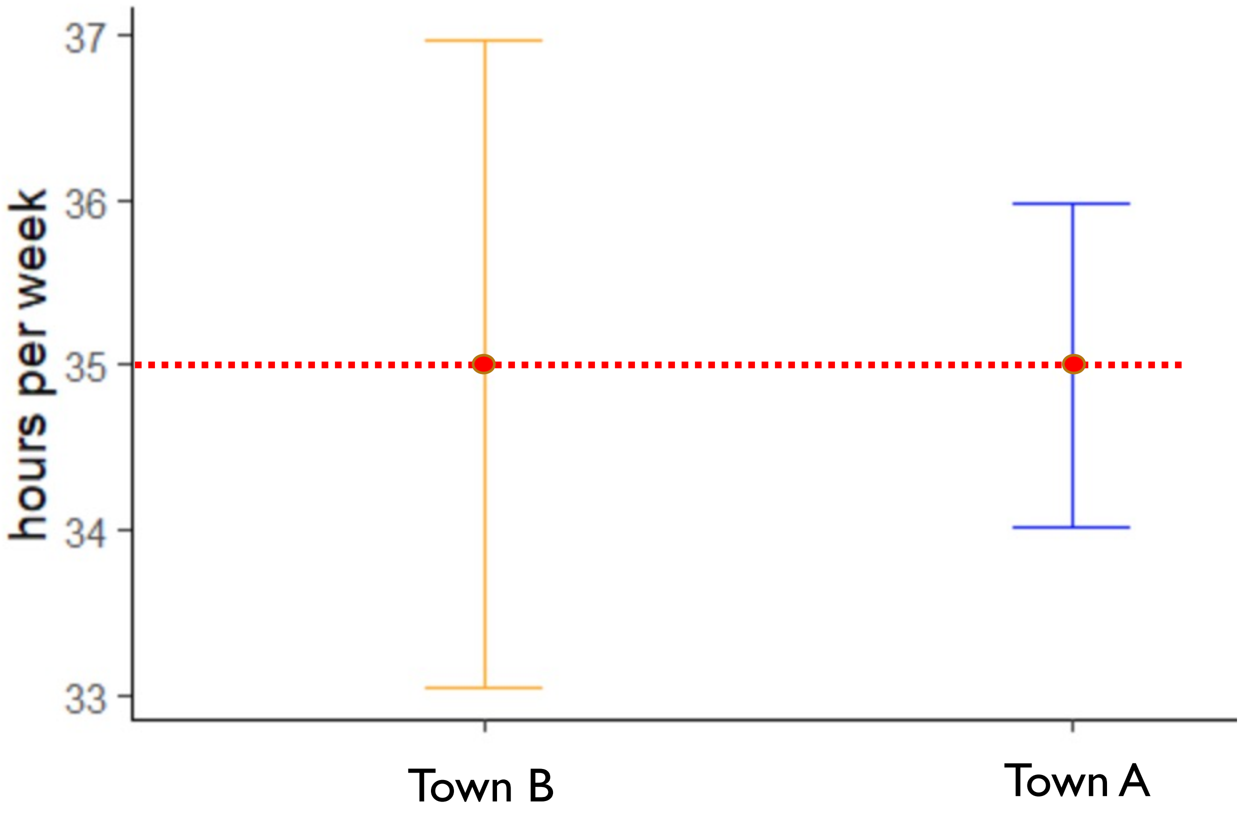
To find the critical value you need to choose α and determine a one-tail or two-tail test.

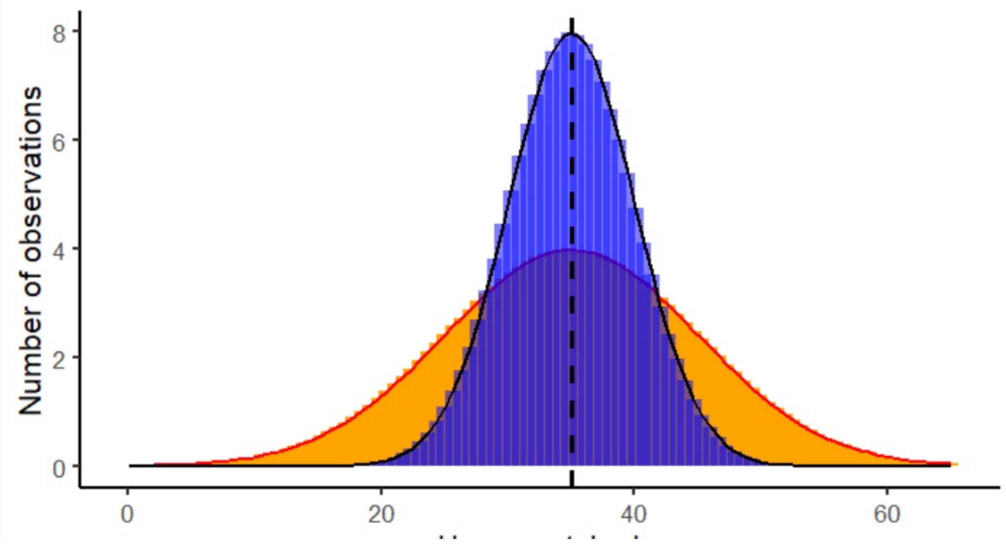
For this example, lets choose $\alpha = 0.05$ and it is a 2-tail test.
$\Phi(.975) = 1.96$ or `norm.ppf(.975) = 1.96`



Town A: $35 \pm 1.96\frac{5}{\sqrt{100}} = [34.02, 35.98]$

Town B: $35 \pm 1.96\frac{10}{\sqrt{100}} = [33.04, 36.96]$





TOWN A: $$35 \pm 1.96 \frac{5}{\sqrt{100}} = [34.02, 35.98]$$

TOWN B: $$35 \pm 1.96 \frac{10}{\sqrt{100}} = [33.04, 36.96]$$

Finding confidence intervals for proportions is very nearly the same as finding confidence intervals for means. The error or SD will be different though.

Suppose you discover from a sample of medical records that 244 out of 1792 patients require medication after procedure X.

Can you find a 99% CI for the prevalence of medication requirement after procedure X ?

$\Phi(.995) = 2.5758$ or `norm.ppf(.975) = 2.5758`

$$\hat{p} \pm Z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{244}{1792} \pm 2.5758 \cdot \sqrt{\frac{\frac{244}{1792} \cdot \frac{1548}{1792}}{1792}} = [0.1153, 0.1570]$$

Here is what we know so far:

A $100 \cdot (1-\alpha)\%$ confidence interval for the mean μ, and the value of σ is known:

$$\left[\bar{x} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \qquad \bar{x} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

A $100 \cdot (1-\alpha)\%$ confidence interval for the difference between means is given by:

$$(\bar{X} - \bar{Y}) \pm Z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

A $100 \cdot (1-\alpha)\%$ confidence interval for the difference between proportions is:

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$$

Two-Sample Confidence Intervals

Sometimes we want to compare two groups to see if there is some significant difference between them

Two overlapping one-sample CI's is tough to interpret. One two-sample CI works better! Look for 0 in the interval.

Next time: Hypothesis testing