MSDS 660
Week 4

Interval Estimations and Hypothesis Testing

Recap

- Properties of a Normal Probability Distribution:
  - Probability is given by the area under the curve.
  - Mean (μ) and standard deviation (σ) determine shape and location of curve.
  - Mean can be any number, determines the center of the distribution.

- Point estimates and sampling variability
  - Sample Statistics becomes **point estimators** in statistical inferences.
  - Typically, samples result in a range of mean values, but they are close enough to the population mean

Calculation of Standard Error

The observations are independent, and the sample size is sufficiently large, the sample proportion will follow a normal distribution with the following mean and standard error:

$$\mu_{\hat{p}} = p \qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

In order for the Central Limit Theorem to hold, the sample size is typically considered sufficiently large when $np \geq 10$ and $n(1-p) \geq 10$, which is called the **success-failure condition**.

In selecting random samples of size *n* from a population, the sampling distribution of the sample mean $\bar{x}$ can be approximated by **a normal distribution** as the sample size becomes **large**.

Agenda for the Week

- Understand and compute Confidence Intervals
- Develop Null and Alternative Hypotheses
- Type I and Type II errors
- Compute Mean differences and one sample t-tests

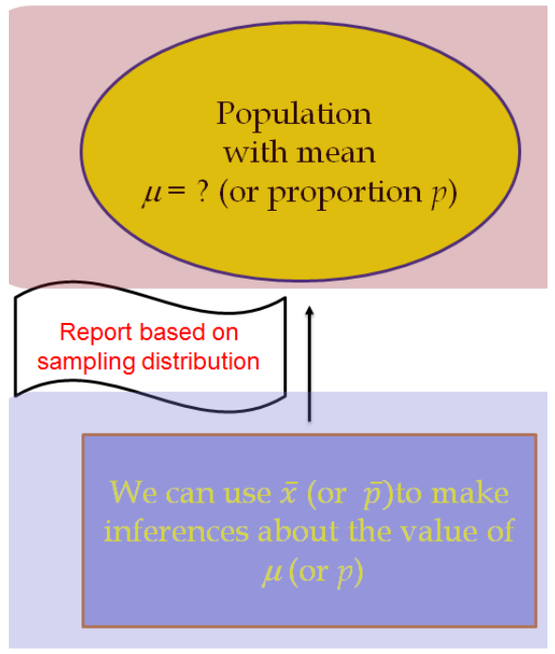
Confidence Interval

As we know, a point estimator (such as a sample mean) cannot be expected to provide the exact value of the population parameter (such as μ).

A confidence interval can be computed that extends 1.96 standard errors of the sample mean to be 95% confident that the interval captures the population mean:

$$\text{point estimate} \pm 1.96 \times SE$$

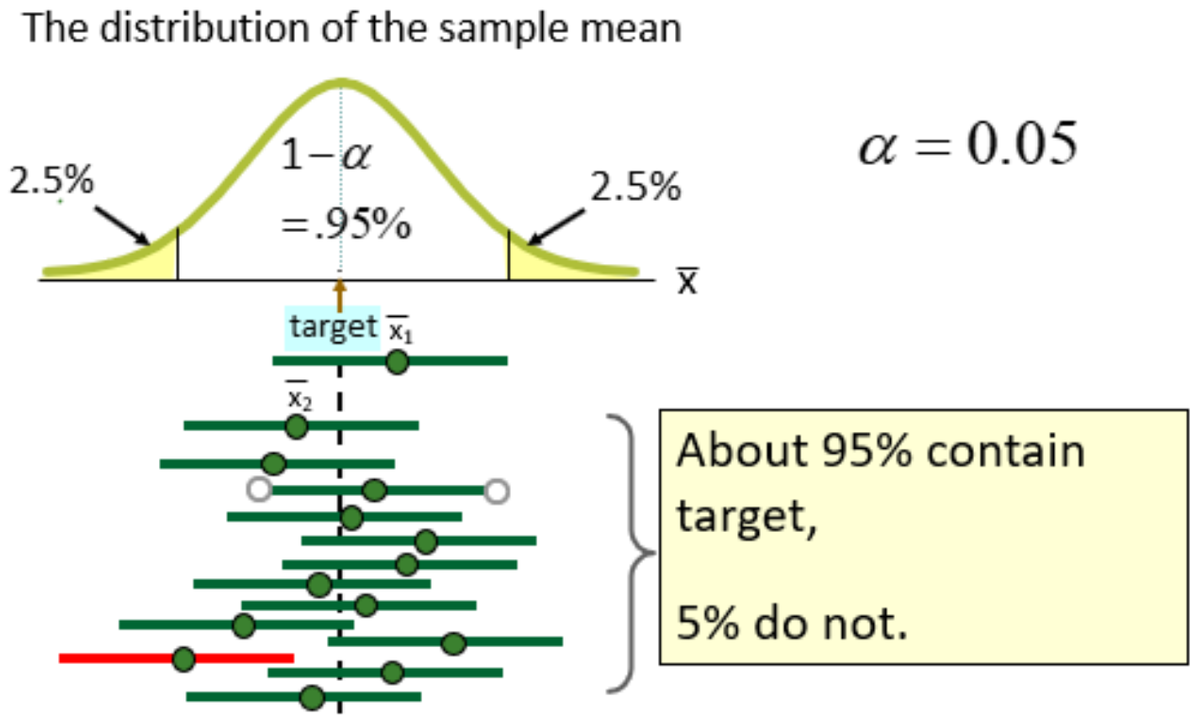
where SE corresponds to standard error



Confidence Interval

A confidence interval:

- A range of values that tells us how likely the interval contains the true population value
- In a normal distribution, 95 % of the data is within 1.96 SDs of the mean.
- Alpha (α) = level of significance



$$\alpha = 0.05$$

If you set your alpha =0.05, you ask the question: Do you want to reject your hypothesis when your sample mean falls in the most extreme 5% of the distribution?

Confidence Interval using any Confidence Level

$$\textbf{point estimate} \pm \text{z}* \times SE$$

where z* corresponds to the confidence level elected and SE to standard error

In a confidence interval, $\text{z} \times SE$ is called <u>the margin of error</u>

What is the point?

- $\bar{x}$ can be obtained directly from your sample
- Once you provide the margin of error based on your choice of confidence level (α), you can provide an estimate of μ with a level of confidence.

Computing Confidence Interval

A Pew poll in 2021 asked: Do you think widespread use of driverless passenger vehicles would be a good idea for society?

26% of respondents said it would be a good idea.

We can compute a 95% confidence interval for the proportion of American adults who think the use of driverless passenger cars is a good idea.

$$sample\ proportion \pm 1.96 \times \boldsymbol{SE}$$

$$0.26 \pm 1.96 \times 0.006 \longrightarrow (0.2482, 0.2717)$$

We are 95% confident that the actual proportion of American adults who support the use of self-driving cars is between 24% and 27%.

An Analogy- Catching a fish!

Not every net in the ocean will catch this fish, but 95 percent of the right sized nets will.

The Confidence Interval is the net.

Not all nets cast into the sea catch fish.

95% of nets catch fish.

For a given interval, the probability that the interval contains for a true mean μ is either 0 or 1 (not 95%). Once thrown into the ocean, the net either has caught a fish or it has not.

You need to cast a bigger net to be more confident you have caught the fish, i.e. use a 99% CI.

What is Hypothesis Testing?

Deciding between two possibilities based on data

- E.g., “Is it real? Or is it just due to chance?”

Hypothesis: A statement about the population

- E.g., more than 30% of customers recognize our product
- You will win the election
- Strategy Z will make you rich in the stock market

Note: a hypothesis is either TRUE or FALSE

- Even with data, you may never know for sure, because of *randomness*

- *How would you test claims like these?*

Constructing a Hypothesis Test

Step 1. State the null hypothesis: A statement about the true parameter μ

- Ho: The widespread use of driverless passenger vehicles is a good idea for society (Even better if there is a quantity associated with it).
- Ho: It's a good idea if 85% of passenger vehicles become driverless.
  - Ho: $\mu = \mu_0$, where $\mu_0 = 1.96$.

Step 2. Estimate the true parameter

- Take the sample. 26% of respondents said driverless passenger vehicles is a good idea for society

Step 3. Compute the standard error of the estimate

Step 4. Compute the test statistic (Z-score). This refers to the z-score associated with the sampling distribution of $\bar{x}$

$$z = \frac{\bar{x} - \mu}{\sigma}$$

Step 5. Make a decision. If |Z| >1.96, reject the null hypothesis. Otherwise, fail to reject.

- Alternatively: Use the value of the test statistic to compute the p-value.
- If p-value is below α= 0.05, reject. Otherwise, fail to reject.

Two-sided vs. One-sided Hypothesis Test

If only one direction of the evidence matters, it may make sense to construct a one-sided test.

- The null hypothesis is the same, but the **alternative hypothesis** is different (not just “not the null”).
- One-sided hypothesis statement:

  H0 : $\mu = \mu_0$ (note : could be ≥)

  H1 : $\mu < \mu_0$

- Only difference in the test is that we use less stringent cutoffs, i.e. only probability in one tail “counts” toward p-value.

Using CIs for Two-sided Hypothesis Tests

If you have constructed a 95 percent confidence interval, this can be used for a wide array of hypothesis tests.

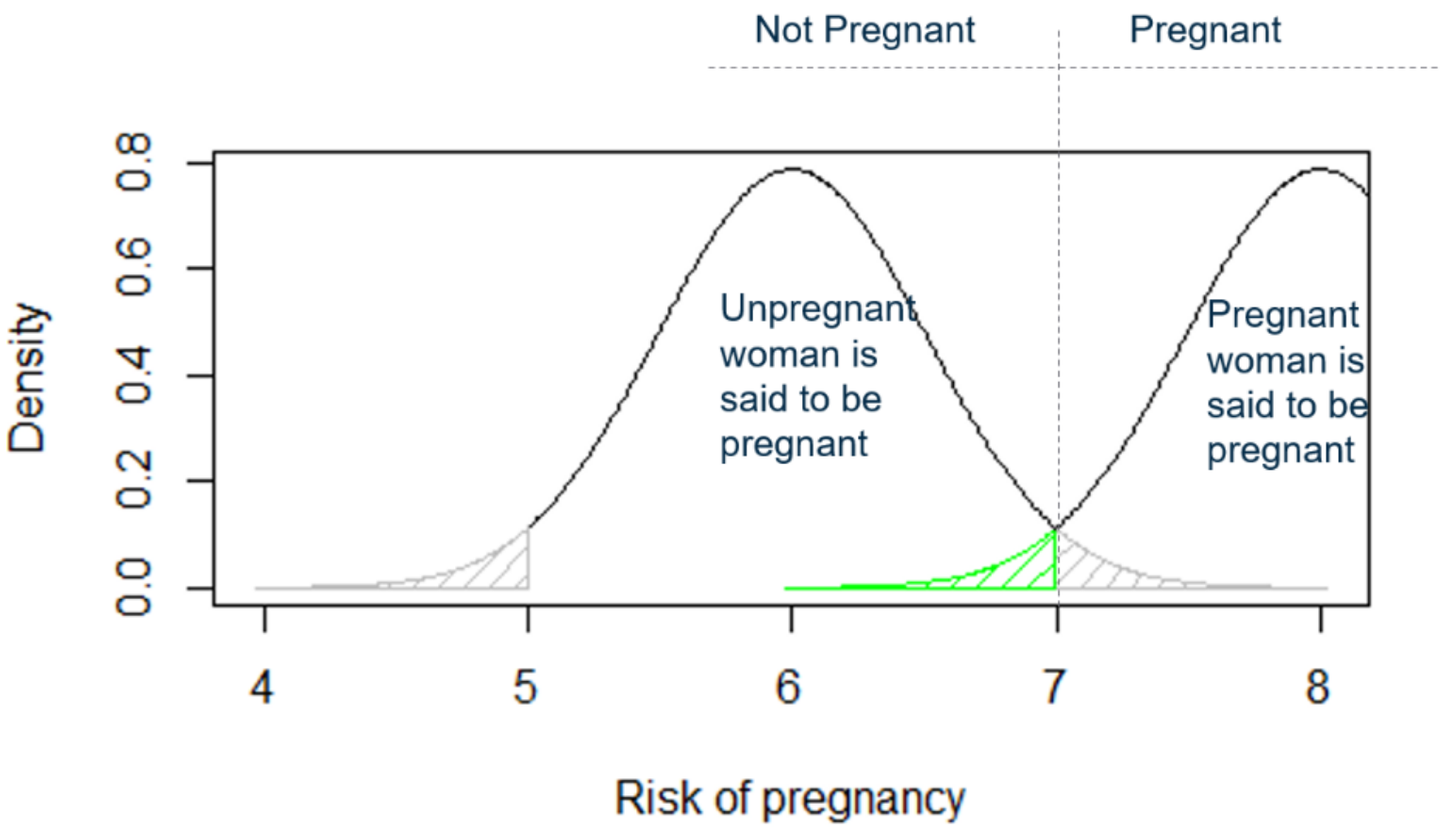
The interval contains plausible values.

If the null hypothesis is a value in the interval, fail to reject.

- Otherwise, reject.

Type I and Type II Errors

- If we apply the α= 0.05 rejection rule, we control for the false positive rate.
  - When we reject the null hypothesis and it is true, we commit a Type I Error.
  - Say that an unpregnant woman is pregnant…oh oh!
  - At α= 0.05, we make Type I Errors only 5% of the time (gray region in the plot)

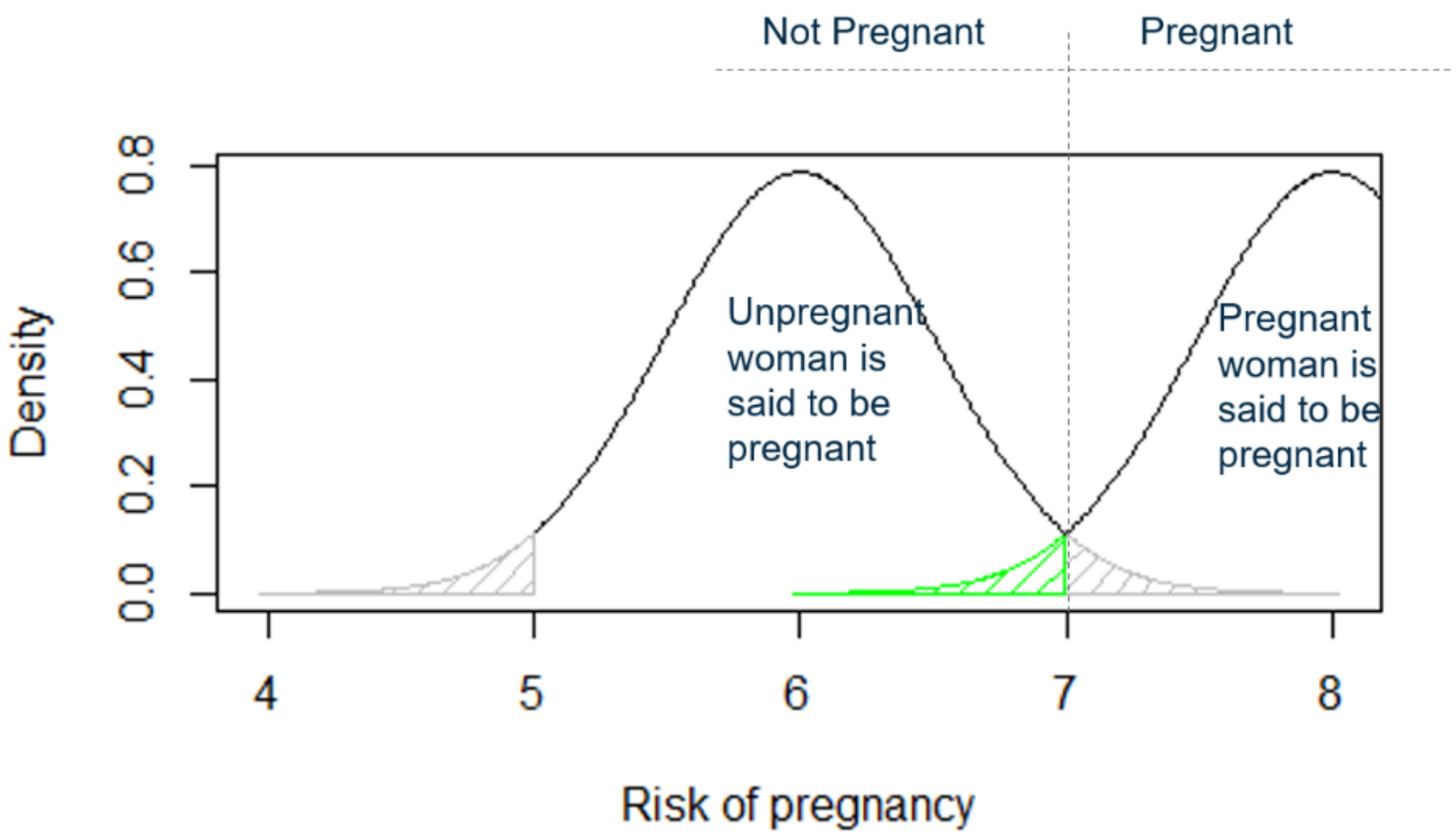


Another Type of Error

Saying a pregnant woman is not pregnant…also, oh oh!

If we apply the α= 0.05 rejection rule, we control for the false negative rate.

- When we fail to reject the null hypothesis and the alternative hypothesis is true, we commit a Type II Error (green region in plot)
- The way hypothesis tests are set up, Type II error is not controlled as well.



Tradeoff between Type I and Type II Errors

- Typically, decreasing one error increases the other error
- For a given sample size, one way to reduce the rate of false negatives (failing to reject when we should) is to relax standards for rejecting.
  - We could use a 90% CI (or $\alpha = 0.10$) instead of 95% (or $\alpha = 0.05$).
- If we reduce Type II error, we increase Type I error.
- The only way to reduce both errors is to increase the sample size.
  - Expensive: The recommendation is to increase sample size until Type II error is sufficiently low.
  - This is a power analysis: power= 1- Type II error

One- sample T-test

- Performs a test on a population mean when a population standard deviation is unknown.
- We apply the same logic as did to compute a *z*-test, only the test statistics changes and we use a *t-statistic*
- One new piece is the *degrees of freedom:*
  - Describes the shape of the *t*-distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.
  - When modeling $\bar{X}$ using the *t*-distribution, use *df* = n-1
  - Assume we have our *n* is independent and we have nearly normal observations.

So, the equation to compute a *t* – Confidence Interval for the Mean is as follows:

$$\text{point estimate} \pm t_{df} \times SE \longrightarrow \bar{x} \pm t^*_{df} \times \frac{s}{\sqrt{n}}$$

where $\bar{x}$ is the sample mean, t^*_{df} corresponds to the confidence interval and degrees of freedom *df,* and *SE* is the standard error as estimated by the sample. Openintro (pg. 256)

CI for Difference in Sample Means

- Sometimes it may be useful to test whether 2 populations have different means.
- A *t*-distribution can be used for inference when working with 2 populations and the σ is not known (this is typically the case) and the following two assumptions are met:
  - Independence: The data come from independently drawn random samples, i.e. they are not related.
  - Normality: We check that there are no outliers within each group

- Take two samples (*X* and *Y,* and compute the sample means).
  - (Openintro pg. 267)

$$Sample\ difference = \overline{X} - \overline{Y}$$

  - Compute the standard error of the difference in means.
    - If the samples are independent

$$SEmean\ difference = \sqrt{SE\frac{2}{x} + SE\frac{2}{y}}$$

Book Readings

and there is much more background material available

Confidence Intervals for a Proportion

- Chapter 5.2

Hypothesis Testing for a Proportion

- Chapter 5.3

One-sample means with the t – distribution

- Chapter 7.1

Difference of two means

- Chapter 7.3

Discussion Activity

1. a. Focus on the collisions with one and two engine planes.

   b. Remove outliers using one the methods shown in the demo.

   c. Did you decide if you want to impute values? Tell us what you decided on.

   d. Compute the average and standard deviation of the distance variable.

2. Compute a 95 percent confidence interval for the difference in mean speed at collision between one-engine and two-engine airplanes.

3. Conduct a one sample t-test for the average speed of all bird-airplane collisions.

   a. What is the conclusion of the one sample t-test?

Assignment

Using the marketing data set or a data set of your choice:

- A hypothesis of which variables might be correlated.
- Be sure to deal with outliers and impute missing data (if appropriate). Justify your decision.
- Compute a CI for a mean difference and a one sample t-test on a numerical variable of interest to you.
- A brief explanation ***1 paragraph !*** of your analysis process and your interpretation of the data. (What is the data, what did you do with the data, what do the results mean?)
- Your Rmd and knitted pdf file