STA2023 Notes

Tuyen Truong

April 9, 2025

Note that these notes do not span the all the topics covered in STA2023, just some of the concepts that I think are most important to have a firm grasp on. I will update these notes. You can find these updates on my website https://maitiennn.github.io/

§1 Background

Remember that we need to distinguish between a parameter and a statistic.

Definition 1.1. We call a measure concerning the population a **parameter**. The population is the entire set of possible cases. The population is who we want to make statistical inferences about.

Definition 1.2. We call a measure concerning the sample a **statistic**. Recall that a sample is a subset of the population.

Example

Think of what we do in our labs. We are trying to draw inferences about a population (for example, ALL UF students), but the sample is taken from our lab sections which is a subset of the population.

	Parameter	Sample Statistic
Mean	μ	$\bar{x}$
Difference in Two Means	$\mu_1 - \mu_2$	$\bar{x_1} - \bar{x_2}$
Proportions	p	$\hat{p}$
Difference in Proportions	$p_1 - p_2$	$\hat{p_1} - \hat{p_2}$

Without getting too much into probability concepts, I want to talk about distributions (I will make some remarks on this later to clarify what I mean). The one frequently used in this class is a normal distribution and a standard normal distribution. In our labs, we have what is called a sampling distribution. Sample statistics are random variables because they vary from sample to sample. As a result, sample statistics have a distribution called the sampling distribution.

Definition 1.3. A sample distribution is a distribution of sample statistics with a mean approximately equal to the mean in the original distribution and a standard deviation known as the standard error.

Definition 1.4. The standard error is the standard deviation of a sampling distribution. (formulas for this are usually provided on the exams). Make sure you use the correct formulas for means/proportions.

When we have certain types of data sets, such as quantitative or categorical data, certain assumptions must be made when using them.

Definition 1.5. Data concerning one categorical variable can be summarized using a proportion. $p = \frac{\#in\ the\ category}{total\ number}$

Definition 1.6. Let n denote the sample size and p_0 the number of successes in our investigation. Assumptions for categorical variables are as follows:
(1) $np_0 \geq 15$
(2) $n(1 - p_0) \geq 15$
(3) Simple Random Sample (SRS)

Note that for quantitative data, we often summarize our data with means, medians, modes, etc. In labs, we typically work with means.

Definition 1.7. Let n denote the sample size Assumptions for quantitative variables are as follows:
(1) $n \geq 30$ or original population is normally distributed for Central Limit Theorem to apply.
(2) Simple Random Sample (SRS)

Definition 1.8. Central Limit Theorem: Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then if n is "large enough", $\overline{X}$ has approximately a normal distribution with

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The larger the value of n, the better the approximation.

§2 Confidence Intervals

Definition 2.1. Think of a confidence interval as a range (with a lower and upper limit) computed using sample statistics to estimate an unknown population parameter with a stated level of confidence (for example, 90%, 95%, ...)

Now, let's build some intuition for constructing a confidence interval (CI). At the center of a confidence interval is the sample statistic, such as a sample mean or sample proportion (the thing we want). This is known as the point estimate or estimator (if you would like to call it as such). The width of the confidence interval is determined by the margin of error. The margin of error is the amount that is subtracted from and added to the point estimate to construct the confidence interval. Here's a general formula to remember when you are asked to construct a CI:

estimator $\pm$ **margin of error** The margin of error can be further broken down into a multiplier (the level of confidence that you are given determine for the multiplier. For example, at 95% confidence, I have a multiplier of 1.96) times the standard error as follows:

Let k be the multiplier and SE be shorthand for the standard error. Then, the
margin of error = k $\cdot$ SE

§2.1 Interpreting Confidence Intervals

Template: We are 95% confident that the [input interval here] captures true mean/proportion of [insert parameter here in words] (based on what the problem deals with) for [insert population here] [add some context].

It is very important you include all the necessary components in a confidence interval: parameter, population, confidence level, interval, and a statement whether your data deals with means or proportions.
There will be further interpretations that can be done with confidence intervals later when we talk about hypothesis testing.

§2.2 Effect of Sample Size on Confidence Intervals

Definition 2.2. As n (sample size) increases, the width of the confidence interval gets narrower because the standard error decreases. Recall that standard error is proportional to $1/\sqrt{n}$.

Definition 2.3. As the confidence level increases (i.e 90, 95, 99), the width of the confidence interval gets wider because a higher confidence level requires a larger critical value (from the z- or t-distribution). While you're more confident that the interval contains the true parameter, you're also allowing for a broader range of values to ensure that higher level of certainty.

§2.3 Bootstrapping/Bootstrap Confidence Intervals

Definition 2.4. Bootstrapping is a resampling procedure that uses data from one sample to generate a sampling distribution by repeatedly taking random samples from the known sample.

Once we have a bootstrap sampling distribution there are two methods for constructing a confidence interval:
The standard deviation of the bootstrap distribution is the standard error which we can use to construct a bootstrap confidence interval. Recall that for a 95% confidence interval, given that the sampling distribution is approximately normal, the 95% confidence interval will be:

$$\text{sample statistic} \pm 2 \cdot \text{standard error}$$

For a 95% confidence interval we can find the middle 95% bootstrap statistics. This is known as the percentile method. This is the preferred method because it works regardless of the shape of the sampling distribution.

§3 Hypothesis Testing

Previously we used confidence intervals to estimate unknown population parameters. We compared confidence intervals to specified parameter values and when the specific value was contained in the interval, we concluded that there was not sufficient evidence of a difference between the population parameter and the specified value. In other words, any values within the confidence intervals were reasonable estimates of the population parameter and any values outside of the confidence intervals were not reasonable estimates. Here, we are going to look at a more formal method for testing whether a given value is a reasonable value of a population parameter. To do this we need to have a hypothesized value of the population parameter.

Definition 3.1. Given that the null hypothesis is true, the probability of obtaining a sample statistic as extreme or more extreme than the one in the observed sample, in the direction of the alternative hypothesis

Definition 3.2. A test is considered to be statistically significant when the p-value is less than or equal to the level of significance, also known as the alpha (α) level. Our usuall alpha levels will be $\alpha = 0.1, 0.05, 0.01$

Definition 3.3. The null hypothesis (H_0) is the the statement that there is not a difference in the population(s).

Definition 3.4. The alternative hypothesis (H_A) is the statement that there is some difference in the population(s).

Example

When writing hypotheses, there are three things that we need to know:

1. The parameter that we are testing (non-directional, right-tailed, or left-tailed), and
2. The direction of the test (non-directional, right-tailed, or left-tailed), and
3. The value of the hypothesized parameter. If it is non-directional (i.e. two sided), use the $\neq$ symbol, otherwise use $>$ or $<$.

At this point, we can write hypotheses for a single mean (μ), paired means ($\mu_1 - \mu_2$), a single proportion (p), the difference between two independent means ($\mu_1 - \mu_2$), the difference between two proportions ($p_1 - p_2$), a simple linear regression slope (β), and a correlation (p).
The research question will give us the information necessary to determine if the test is two-tailed (e.g., "different from," "not equal to"), right-tailed (e.g., "greater than," "more than"), or left-tailed (e.g., "less than," "fewer than").
The research question will also give us the hypothesized parameter value. This is the number that goes in the hypothesis statements (i.e., μ_0 and π_0). For the difference between two groups, regression, and correlation, this value is typically 0.
Hypotheses are always written in terms of population parameters (e.g., p and μ). The tables below display all of the possible hypotheses for the parameters that we have learned thus far. Note that the null hypothesis always includes the equality (i.e., $=$).

§3.1 Interpreting p-values

Example

If $p > \alpha$, then we **"fail to reject the null hypothesis"** and conclude that there is not enough evidence of a difference in the population. This does not mean that the null hypothesis is true; it only means that we do not have sufficient evidence to say that it is likely false. These results are **not statistically significant**.

If $p \leq \alpha$, then we **"reject the null hypothesis"** and conclude that there is a difference in the population. These results are **statistically significant**.

§3.2 Relating back to Confidence Intervals

A secondary use of confidence intervals is to support decisions in hypothesis testing, especially when the test is two-tailed (two-sided). The essence of this method is to compare the hypothesized value to the confidence interval. If the hypothesized value falls within the interval, we fail to reject the null hypothesis. Otherwise, if it is not in the interval, we have significant evidence, and so we can reject the null hypothesis.

§4 Just for Lab10

- Frame your question based on your variable to be a yes or no question.
- Formulating the hypotheses: Ask yourself do we have means or proportions? $H_0 : p_1 - p2 = 0$ vs $H_A : p_1 - p_2 \neq 0$ OR
  $H_0 : \mu_1 - \mu 2 = 0$ vs $H_A : \mu_1 - \mu_2$
- Some theory for intuition:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}$$

Recall that $\hat{p_1} - \hat{p_2}$ is approximately normally distributed with mean $p_1 - p2$ and variance

$$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

But, if we assume that the null hypothesis is true, then the population proportions equal some common value p, say, that is,

$$p_1 = p_2 = p.$$

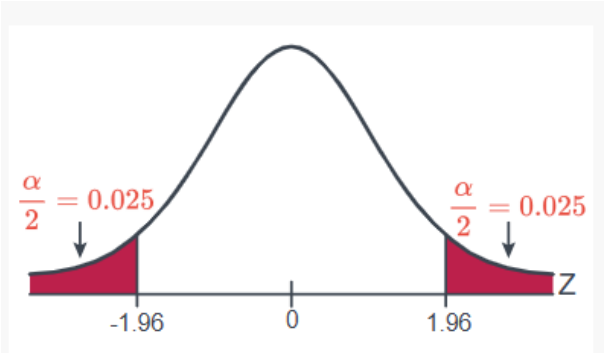
In that case, the variance becomes:

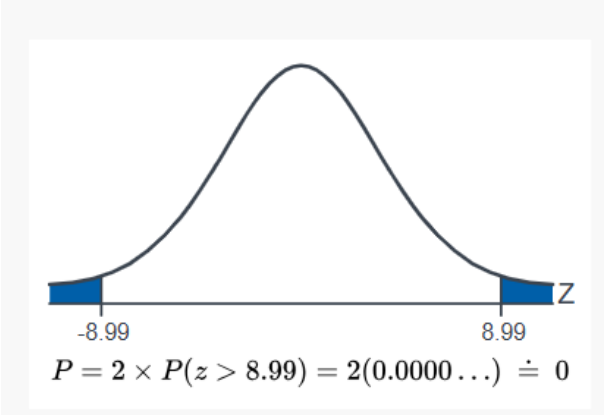
$$p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

So, under the assumption that the null hypothesis is true, we have that:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - \underbrace{(p_1 - p_2)}_{\underline{0}}}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

This is what you should be thinking of when you think of p-value:





Interpreting the p-value: With a p-value of [ ], we do not OR do have significant evidence to say that there is a difference in [insert your variable context here].

Interpreting confidence interval: We are 95% confident that [interval here] contains the true difference of [context here] in proportion of [insert your context here].