

---

# STA238 Notes

Statistics for real

<https://github.com/ICPRplshelp/>

Last updated February 7, 2023

# 1 Sampling Distribution and the Central Limit Theorem

**IDEA** – The sampling distribution is meant to give us a probability distribution for the sample mean and variance (being called statistic) if we took random samples out of a random variable. Because there is a chance we could take a sample and somehow end up with a sample mean being very far from  $\mu$  (but it occurs very rarely – and the sampling distribution can tell us how likely that is to happen).

- We can view random samples as **unbiased samples**. In a perfect world, statistical methods like these work perfectly, but garbage in, garbage out.
- The main goal is that as long as our sample is completely random, we can guess the theoretical mean and variance, and ALSO calculate how confident we should be with our guess (that is, confidence intervals).

An **EXPERIMENTAL UNIT** is someone or something which we might collect data from.

A **POPULATION** is the set of ALL units we're interested in.

A **VARIABLE** is a characteristic or property of an individual unit from the population. Each person (or thing) has a property, right?

If we look at a population of people and  $Y_i$  represents the age of the  $i$ th person, we may have this space:

$$[Y_1, Y_2, \dots, Y_{10}, \dots, Y_N] \leftarrow N = 50000$$

If we have measurements for EVERYONE, then we can calculate the population mean:

$$\mu = \frac{1}{N} \sum_{i=1}^N Y_i$$

And the variance:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \mu)^2$$

Not feasible to get everyone. Right? Let's collect a random sample (no, we will never feasibly get that) of 200. With our sample, we can measure their age:

$$[y_1, y_2, \dots, y_5, \dots, y_n] \leftarrow n = 200$$

Usually,  $\mu$  and  $\sigma$ , the population parameters, are unknown and are too difficult to measure, and that value can fluctuate. Good thing is that we can estimate close to that parameter.

After collecting the sample, we can measure some statistic(s): a function of the sample observation

- For example, the sample mean:  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$
- And the sample standard deviation:  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$ 
  - The denominator contains a  $n - 1$  to get an unbiased estimator for  $\sigma^2$ .

## 1.1 Linking Population and Samples

Population:

$\mu, \sigma^2$	UNKNOWN
$\bar{y}, s^2$	KNOWN

Sampling distribution helps bridge the gap between the unknown and the known.

We assume that  $\bar{y}$  and  $s^2$  are random variables. What are the sampling distributions for them? For the **samples**:

$$\begin{aligned}y_1, y_2, \dots, y_n &\Leftarrow (\mu, \sigma) \\E(y_i) &= \mu \\V(y_i) &= \sigma^2 \\i &= 1, 2, \dots, n\end{aligned}$$

**All the samples are identically distributed.**

For the **statistic**:

$$\begin{aligned}E(\bar{y}) &= E\left(\frac{1}{n} \sum_{i=1}^n y_i\right) = \frac{1}{n} \sum_{i=1}^n E(y_i) = \frac{1}{n} \cdot \mu \cdot n = \mu \\V(\bar{y}) &= V\left(\frac{1}{n} \sum_{i=1}^n y_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(y_i) = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n} \\SD(\bar{y}) &= \sqrt{V(\bar{y})} = \frac{\sigma}{\sqrt{n}} \quad \text{standard error}\end{aligned}$$

So, IF  $Y \sim (\mu, \sigma)$ , THEN  $\bar{y} \sim ? \left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ ?

**CASE 1.**  $\sigma$  is known, AND population is normal.

- Then,  $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

**CASE 2.**  $\sigma$  is known, AND  $n \geq 30$  ( $n$  is large).

- Then, by CLT,  $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ , irrespective of  $y_i$ 's distribution.

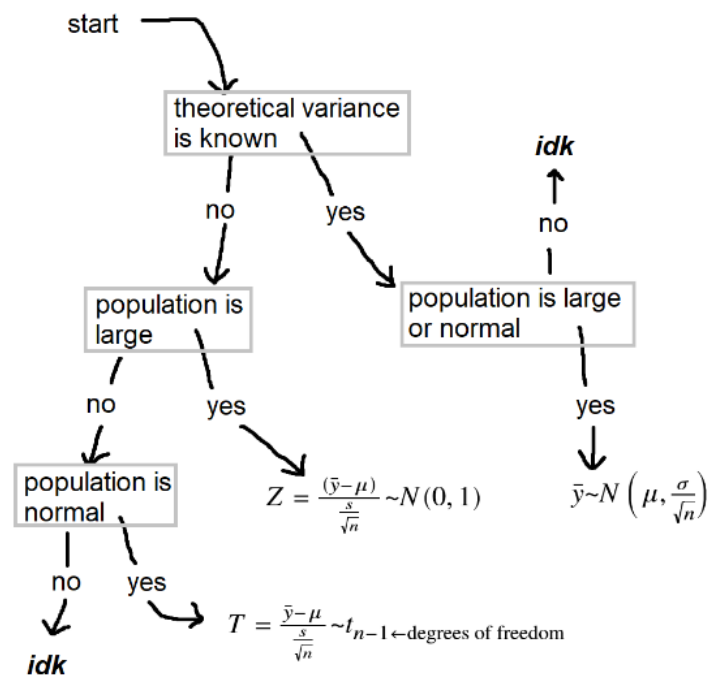
**CASE 3.**  $\sigma$  is unknown, AND  $n \geq 30$ .

- $Z = \frac{(\bar{y}-\mu)}{\frac{s}{\sqrt{n}}} \sim N(0, 1)$
- Questions using case 3 must be given such that doing this is possible.

**CASE 4.**  $\sigma$  is unknown, population is normal, and  $n < 30$ .

- $T = \frac{\bar{y}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1} \leftarrow \text{degrees of freedom}$ 
  - The  $n - 1$  stands for degrees of freedom

If you know these four cases, it's easier to make statistical inferences.



**Figure 1:** This is what the four points above told me

## 1.2 Sample Proportion

A **binary variable** takes only two outcomes, such as tossing a coin. When tossing a coin, we're able to get success or failure:

$$y_i = 1 \text{ if success otherwise } 0$$

Success will be denoted as  $S$  and failure will be denoted as  $F$ . Then,  $Y \sim \text{Ber}(p)$ , where  $p$  is the probability of success.

$$P(S) = p \quad P(F) = 1 - p$$

Hence  $P(Y = y) = p^y(1 - p)^{1-y}$ . Then, for a Bernoulli distribution:

$$\mu = E(Y) = p$$

$$\sigma^2 = V(Y) = p(1 - p)$$

**Definition:** Consider sample proportions:

$$\hat{p}_{\text{sample proportion}} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}$$

The total number of successes divided by sample size.

The sampling distribution for sample proportions:

$$E(\hat{p}) = E(\bar{y}) = \mu = p$$

$$\Rightarrow \boxed{E(\hat{p}) = p}$$

$$V(\hat{p}) = V(\bar{y}) = \frac{\sigma^2}{n} = \frac{p(1-p)}{n}$$

$$\Rightarrow \boxed{V(\hat{p}) = \frac{p(1-p)}{n}}$$

$$SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

**CONDITION FOR NORMAL APPROXIMATION:**

$$(np \geq 10 \text{ and } n(1-p) \geq 10) \Rightarrow \hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

- Expected number of successes and expected number of failures both are  $\geq 10$ .

### 1.3 T-Distribution

With condition in which we don't know  $\sigma$ , population is normal,  $n < 30$ :

$$T = \frac{\bar{y} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

You can find out how to solve T-distribution-type questions in the STA237 notes document. In practice, you'll always be calculating that using a program.

By the way,  $t_\infty \sim N(0, 1)$ .

The  $t$ -distribution is symmetric around 0. To read the  $t$ -distribution table, focus on the degrees of freedom that you want to target. If we happen to know the area of one side of the tail, we can find the critical points  $t_{\text{corresponding tail area}}$ . Look for the **one-tailed probability** equal to the area under the tail of the curve. That is the critical value; value on the  $x$ -axis.

Hence,  $t_{0.025, 11} = 2.201$ .

The  $t$ -curve is symmetric, so the area below the left and the right tails should match. When we add the area below both tails, we get the area of one tail multiplied by two.

## 1.4 Chi-Squared Distribution

$$\mu \rightarrow \bar{y} \rightarrow N \text{ or } t$$

$$p \rightarrow \hat{p} \rightarrow N$$

$$\sigma^2 \rightarrow s^2 \rightarrow \chi^2$$

We want to conduct statistical inferences for  $\mu$ .  $\bar{y}$  is the candidate. We have two options:

- Normal distribution
- $t$ -distribution

Which depends on the sample size. If we are interested on the population proportion, we use sample proportion  $\hat{p}$ , which follows the normal distribution **under certain conditions (if it fails, we can't answer anything, not even use the  $t$ -distribution)**.

The candidate for population variance  $\sigma^2$  is  $s^2$ .

$$y_1, y_2, \dots, y_n \sim N(\mu, \sigma)$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$



Which is the definition of the sample variance. We can use this to make statistical inferences of  $\sigma^2$ . We can't find  $\sigma^2$  directly, and we **cannot** find the distribution for  $s^2$  directly, but we can find the distribution for:

$$\frac{(n-1)s^2}{\sigma^2} \quad \text{this quantity}$$

And it follows this distribution:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1=df}^2$$

We can prove this result using the moment generating function, but we're not going to use it here.

For this distribution, what we are sampling from **must be normally distributed**.

We'll use this result to conduct statistical inferences for  $\sigma^2$ .

### 1.4.1 Examples of Chi-2

You can't pour exactly 500mL of water into a bottle each time and get it perfectly. So, each of your water buckets, which you aimed to pour water in, will hold water. The quantity of water held for each bucket will be a random variable  $Y$ .

Better to use a confidence interval, right?

You might get a question looking like this, where you want to find  $b_1$  and  $b_2$ :

$$P(b_1 \leq s^2 \leq b_2) = 0.9$$

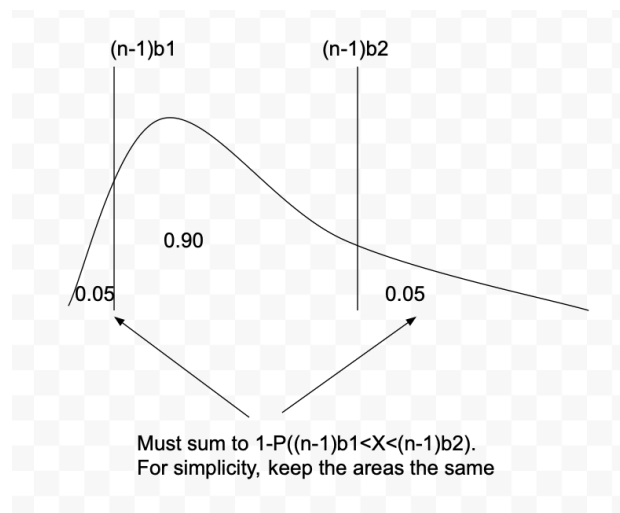
To find  $b_1$  and  $b_2$ , we must know the distribution of  $s^2$ . But we don't know, so we'll set up for that quantity. Multiply both sides by  $(n-1)$  and divide it by  $\sigma^2$ , so the probability does not change.

$$P\left(\frac{(n-1)b_1}{\sigma^2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \frac{(n-1)b_2}{\sigma^2}\right) = 0.9$$

If we let  $n = 10$  and  $\sigma^2 = 1$ :

$$P(9b_1 \leq \chi_9^2 \leq 9b_2) = 0.9$$

The  $\chi^2$  distribution is skewed to the right, so it is not symmetric. We have  $df = 9$ .



**Figure 2:** Chi-2 density visualization

If we focus on  $df = 9$ , at the **row** of the chi-squared table:

$$(n-1)b_2 = 16.919 = \chi_{0.05}^2$$

$$(n-1)b_1 = 3.325 = \chi_{0.95}^2$$

From this, we can conclude that (recall  $n = 10$ ):

$$b_1 = 0.369$$

$$b_2 = 1.88$$

And we've built a 90% confidence interval on  $s^2$ . If we don't know what  $\sigma^2$  is, then our confidence interval will likely contain  $\sigma^2$  as a term.

### 1.4.2 (Optional) Why the above works

A bit of a proof for the above

$$\begin{aligned}
 (n-1)s^2 &= \sum_{i=1}^n (y_i - \bar{y})^2 \\
 \frac{(n-1)s^2}{\sigma^2} &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sigma^2} \\
 &= \sum_{i=1}^n \frac{(y_i - \mu)^2}{\sigma^2} - \frac{(\bar{y} - \mu)^2}{\frac{\sigma^2}{n}}
 \end{aligned}$$

$$\begin{aligned}
 \frac{y_i - \mu}{\sigma} &\sim N(0, 1) \\
 \left( \frac{y_i - \mu}{\sigma} \right)^2 &\sim \chi_1^2 \quad \text{squaring both sides, not proving this}
 \end{aligned}$$

You won't be tested on this, though.

## 1.5 F-Distribution (Statistical Inferences for Two Populations)

If we have a bunch of students and we have  $Y \rightarrow$  height for each student. We can't say that all students are the exact same. We'll have to separate gender apart (one of the covariate information), age, and so on.

If we split students into

- Population 1: F student population
- Population 2: M student population

This means we can divide students into groups. In statistics, all assumptions are ideally wrong, but when we use some assumptions, we can make some real conclusions for usage.

If we assume that all F and M students have the same characteristics (i.e., all have the same age), and

- Population 1 has mean  $\mu_1$ ,  $\sigma_1^2$
- Population 2 has mean  $\mu_2$ ,  $\sigma_2^2$  (population mean and variance respectively, for M and F student height)

What if we want to calculate statistical inference for  $\mu_1 - \mu_2$ , or statistical inference for  $\frac{\sigma_1^2}{\sigma_2^2}$ ?

In hypothesis testing:

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad H_A : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

If I want to do statistical inference for  $\frac{\sigma_1^2}{\sigma_2^2}$ , this is an unknown ratio. But the corresponding statistics is

$$\frac{s_1^2}{s_2^2}$$

It means that we can collect samples from the first population and the second population. So:

- For the F group, we have  $n_1, \bar{y}_1, s_1^2$
- For the M group, we have  $n_2, \bar{y}_2, s_2^2$

If we know the distribution of  $\frac{s_1^2}{s_2^2}$ , then we can make statistical inferences with  $\frac{\sigma_1^2}{\sigma_2^2}$ .

We'll need to start with the chi-squared distribution first. Firstly, we assume that:

$$W_1 \sim \chi_{v_1 \leftarrow df}^2, \text{ not necessarily int}$$

$$W_2 \sim \chi_{v_2}^2$$

Both are random variables, so they take some values in a particular range. More importantly,  $W_1$  and  $W_2$  are **independent**. Under these assumptions, we can do statistical inference. If that is the case (note that  $v_1$  is the degrees of freedom):

$$F = \frac{W_1/v_1}{W_2/v_2} \sim F_{\substack{v_1 \leftarrow \text{numerator} \\ v_2 \leftarrow \text{denominator df}}} \text{ or } F_{v_1, v_2}$$

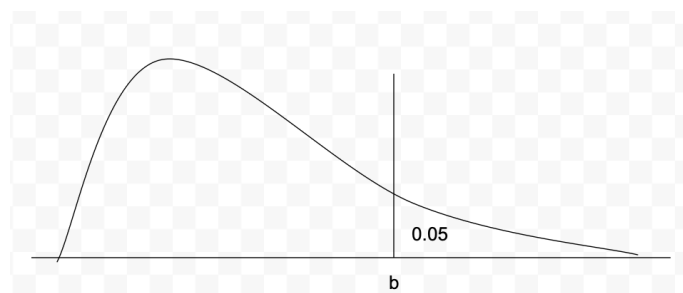
Here,  $F$  is said to have an  $F$  distribution with  $v_1$  numerator degrees of freedom and  $v_2$  denominator degrees of freedom.

The reason why  $W$  is divided by  $v$  because  $W$  already is multiplied by  $v$  and we need to cancel out its effects.

This is related to ANOVA. ANOVA (analysis of variance) has applications in agriculture when we consider which fertilizer is good, which ones aren't.

### 1.5.1 Examples

The  $F$  distribution looks like it is skewed to the right.



**Figure 3:** F-distributed density function

Your questions will probably look like

$$P(F \leq b) = 0.95$$

Good luck using the table here. Consult the STA237 notes on how to use the table. Again, you'll be comparing sample variances.

## 1.6 A Quick Way To Calculate The Sample Variance

$$\frac{1}{n-1} \left( \left( \sum_{i=1}^n y_i^2 \right) - n \cdot \bar{y}^2 \right)$$

## 2 Point Estimation

Expect a question on this

### ESTIMATORS

Expectation means population mean. So:

$$E(\bar{y}) = \mu$$

What I'm going to do: estimate the mean.

$$\underset{\text{estimator of } \mu}{\hat{\mu}} = \bar{y}$$

For example, what is the average height of the Canadian population? From a sample, we might estimate that it is 5.5ft from a 1000-person random sample.

$\mu$  is still unknown, and it won't be equal to exactly 5.5ft.

$\bar{y}$  is the unbiased estimate of  $\mu$ .

When we define  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$ , we can prove that

$$E(s^2) = \sigma^2 \quad \text{what did you expect?}$$

So, the sample variance is an unbiased estimator for  $\sigma$ . So:

$$\underset{\text{estimator for variance}}{\hat{\sigma}^2} = s^2$$

**BIASED ESTIMATOR (the evil version of the unbiased estimator)**

$$s_*^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

THE DIFFERENCE IS THAT IT ISN'T AN UNBIASED ESTIMATOR. If you define sample variance using this formula. So,  $s_*^2$  is a **biased** estimator for  $\sigma^2$ .

**2.1 The General Parameter  $\theta$** 

Now, we've learned about the following statistics:

- $\bar{y} \leftrightarrow \mu$
- $\hat{p} \leftrightarrow p$ , the population proportions
- $s^2 \leftrightarrow \sigma^2$

The general parameter  $\theta$  MIGHT be any of  $\mu$ ,  $p$ ,  $\sigma^2$  (think of interfaces, abstraction, polymorphism... could be any, and I don't know which one yet).

$\hat{\theta}$  is the point estimator for the parameter  $\theta$ .

$\hat{\theta}$  is an unbiased estimator if  $E(\hat{\theta}) = \theta$ . Otherwise, it is said to be a biased estimator.

Now go ahead and replace  $\theta$  and  $\hat{\theta}$  with  $(\mu, \bar{y})$ ,  $(s^2, \sigma^2)$ ,  $(\hat{p}, p)$  respectively.

**2.2 Bias of the Point Estimator**

The bias of a point estimator  $\hat{\theta}$  is:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

When considering  $\bar{y}$  and  $\mu$ , what is the bias?

$$B(\bar{y}) = E(\bar{y}) - \mu = \mu - \mu = 0$$

And there, we can say that  $\hat{y}$  is an **unbiased estimator**.

## 2.3 Mean Squared Error

$$\begin{aligned} MSE(\hat{\theta}) &= E\left((\hat{\theta} - \theta)^2\right) \\ &= V(\hat{\theta}) + \left(B(\hat{\theta})\right)^2 \end{aligned}$$

No need to prove this. However, the variance for  $\hat{\theta}$  is...

Well, look at  $V(\bar{y}) = E\left((\bar{y} - E(\bar{y}))^2\right)$ . So:

$$E\left(\underbrace{(\hat{\theta} - E(\hat{\theta}))}_{\text{variance}} + \underbrace{(E(\hat{\theta}) - \theta)}_{\text{bias}}\right)^2$$

If we expand this, we get  $MSE(\hat{\theta}) = V(\hat{\theta}) + B(\hat{\theta})^2$ . Don't prove it.

## 3 Confidence Intervals

We can estimate  $\hat{\mu}$  using  $\bar{y} = \underset{\text{example}}{5.5}$ . We want an interval estimate such as a 95% confidence interval for  $\mu = \underset{\text{example}}{(4.8, 5.9)}$ . So, there is a 5% possibility that we are out of the interval (such as a 5% error).

An interval estimator is a rule specifying for using the sample measurements to calculate two number that form the endpoints of an interval.

The interval should capture the true population parameter (be within), such as the population mean.

An interval that is too wide gives us no information. A 100% confidence interval doesn't give us any meaning, so we'll consider 95%, 99%, or 90% confidence intervals. We can allow some error in the CI, and the **error is called the level of significance  $\alpha$** .



$$\alpha = 0.05$$

The confidence level is  $1 - \alpha$ . The confidence the unknown population parameter will be in that interval.

- Lower bound is lower confidence limit
- Upper bound is upper confidence limit

### 3.1 Large-Sample CIs

Take  $\bar{y}$  and  $\mu$ . Consider a large sample, the sample size  $\geq 30$ . By CLT, we can say that  $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ . We can standardize:  $Z = \frac{\bar{y} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$ .

We want to construct a  $1 - \alpha$  confidence interval, and we'll going to make a two-sided confidence interval, so we're going to distribute error equally on the tails of the normal curve.

#### BUILDING THE CONFIDENCE INTERVAL

Then, we get critical values which we would put on the  $x$ -axis:

- RIGHT:  $Z_{\frac{\alpha}{2}}$
- LEFT:  $-Z_{\frac{\alpha}{2}}$

Because we have  $-Z_{\frac{\alpha}{2}}, Z_{\frac{\alpha}{2}}$

We want the random variable in the middle to be set to

$$-Z_{\frac{\alpha}{2}} < \frac{\bar{y} - \mu}{\frac{\sigma}{\sqrt{n}}} < Z_{\frac{\alpha}{2}}$$

Multiply both sides by  $\frac{\sigma}{\sqrt{n}}$

$$-Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \bar{y} - \mu < Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

And move  $\bar{y}$  to adjacent sides:

$$-Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} - \bar{y} < -\mu < Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} - \bar{y}$$

Multiply all sides by  $-1$ :

$$\bar{y} - \frac{Z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}} < \mu < \bar{y} + \frac{Z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}$$

And there, we've got our  $100(1 - \alpha)\%$  confidence interval for  $\mu = \bar{y} \pm \frac{Z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}$  (that is the uncertainty). Which is:

$$\bar{y} \pm \left( Z_{\frac{\alpha}{2}} SE(\bar{y}) \right)$$

This is the general result. If we want to do this for population proportion, just take the mean and variance and plug in stuff from there.

If you need to use a  $t$ -distribution yet, replace  $Z$  with  $t$ . If  $\sigma$  is unknown, but we have a large sample, replace  $\sigma$  with  $s$ .

### 3.1.1 Confidence Intervals for Population Proportion

To form a  $100(1 - \alpha)\%$  CI for an unknown proportion  $p$  ( $\hat{p}$  can either be normal or unknown):

- Is  $\hat{p} \pm Z_{\frac{\alpha}{2}} SE(\hat{p})$ , which is the CI
- $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$

We can't make a confidence interval if we need an unknown parameter. We're going to estimate it:

$$\widehat{SE}(\hat{p}) = \hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

### 3.2 CIs for Population Proportions, Again

To construct confidence intervals for an unknown population proportion  $p$ , it is:

$$\hat{p} \pm Z_{\frac{\alpha}{2}} SE(\hat{p})$$

$$SE(\hat{p}) = \hat{\sigma}_{\hat{p}} = \widehat{SD}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

When we estimate the SD of a sampling distribution, we call it a standard error.



The standard error helps us tell the variability of a statistic obtained from a sample.

The extent of the interval on either side of  $\hat{p}$  is the margin of error (how much we can be off, or half the size of the CI):

$$ME = Z_{\frac{\alpha}{2}} SE(\hat{p})$$

This is how much error we can allow based on the sample proportion.  $Z_{\frac{\alpha}{2}}$  is the critical value;  $\alpha$  is the level of significance.



$Z_p$  is  $q$  such that  $P(Z < q) = 1 - p$ .



If any question **requires** me to know  $p$ , the theoretical sample proportion:

1. If given exactly (“in the neighborhood of”), use that value.
2. If given a range, take the midpoint of the range.
3. If not given, take  $p = 0.5$ . There is a good reason to do that.

For questions that want an integer value, and the higher the value, the “better” it is according to context (even if it takes more physical effort, such as required sample size), when calculating it **TAKE THE CEILING OF YOUR CALCULATIONS**.

### 3.3 CIs For The Population Mean

If  $\sigma$  is unknown, use  $s$  instead of  $\sigma$ . If it is known, use  $\sigma$  instead of  $s$ , that is for  $SE(\bar{y}) = \frac{\sigma \text{ or } s}{\sqrt{n}}$ .

#### 3.3.1 Small Sample

In cases where  $\sigma$  is unknown and  $n < 30$ , and the population is normal, then the CI is:

$$\bar{y} \pm t_{n-1, \frac{\alpha}{2}} SE(\bar{y}) \quad SE(\bar{y}) = \frac{s}{\sqrt{n}}$$

And  $\alpha$  is the level of significance. In short, replace  $Z_{\frac{\alpha}{2}}$  with  $t_{n-1, \frac{\alpha}{2}}$

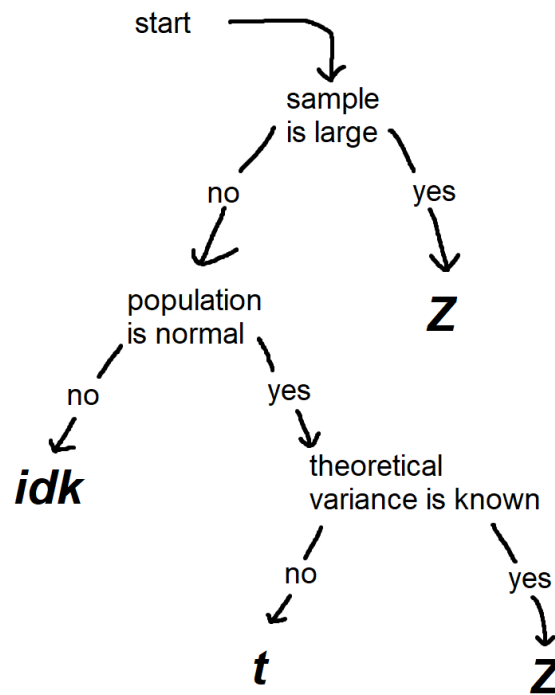
#### 3.3.2 Large Sample

This is the confidence interval if:

- $n \geq 30$

- Or BOTH  $\sigma$  is known **and** the sample comes from a normal distribution.

$$\bar{y} \pm Z_{\frac{\alpha}{2}} SE(\bar{y})$$



**Figure 4:** Flow chart.

### 3.4 Determining Required Sample Size

To determine the sample size (at least, it must be large) to form a  $100(1 - \alpha)\%$  CI for  $\mu$ , which is:

$$= \bar{y} \pm \underbrace{Z_{\frac{\alpha}{2}} \cdot SE(\bar{y})}_{\text{MARGIN OF ERROR}}$$



The margin of error is a function of **the level of significance** and the **standard error** of the statistic.

Where  $SE(\bar{y}) = \frac{\sigma}{\sqrt{n}}$  or  $\frac{s}{\sqrt{n}}$ , depending on which one we can use.

This means that  $ME = \frac{Z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}$ . Squaring it, we get

$$ME^2 = \frac{Z_{\frac{\alpha}{2}}^2 \sigma^2}{n}$$

Isolating  $n$ , we get:

$$n = \left\lceil \frac{Z_{\frac{\alpha}{2}}^2 \sigma^2}{ME^2} \right\rceil = \left\lceil \left( \frac{Z_{\frac{\alpha}{2}} \sigma}{ME} \right)^2 \right\rceil$$

You'll see this formula in the formula sheet. As a reminder,  $ME = Z_{\frac{\alpha}{2}} SE(\bar{y})$ .

### 3.5 Approximating the Population Variance

**We can use  $s$  from a previous sample if we already have it.** Otherwise:

We may approximate (guess completely and hopefully with context) the range of observations in the population and estimate that

$$\sigma = \frac{R}{4}$$

Then, you can use that in your calculations. Best used in the context of the equation above, in the previous sub-section.

$$R = \max(Y_1, \dots, Y_N) - \min(Y_1, \dots, Y_N)$$

### 3.6 Confidence Intervals For The Variance (Welcome Back, $\chi^2$ )

If  $n$  samples come from a normal distribution  $N(\mu, \sigma)$ , then the test statistic is  $S^2$ , which is the sample variance. It is the point estimator for  $\sigma^2$ . We know that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

We assume the level of significance  $\alpha$ . We **want** to find:

$$\chi_{1-\frac{\alpha}{2}}^2 \quad \text{AND} \quad \chi_{\frac{\alpha}{2}}^2$$

For a confidence level of  $100(1 - \alpha)\%$

We will end up with

$$\chi_{1-\frac{\alpha}{2}}^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_{\frac{\alpha}{2}}^2$$

Flip all of them:

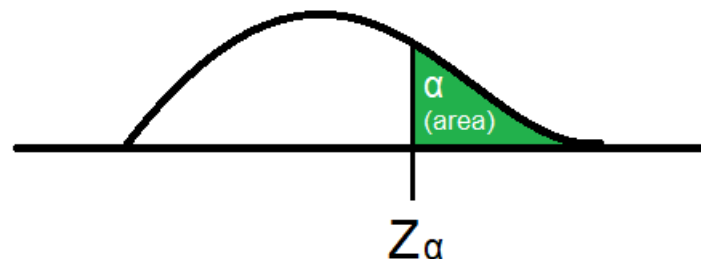
$$\begin{aligned} \frac{1}{\chi_{1-\frac{\alpha}{2}}^2} &> \frac{\sigma^2}{(n-1)S^2} > \frac{1}{\chi_{\frac{\alpha}{2}}^2} \\ \Leftrightarrow \frac{1}{\chi_{\frac{\alpha}{2}}^2} &< \frac{\sigma^2}{(n-1)S^2} < \frac{1}{\chi_{1-\frac{\alpha}{2}}^2} \end{aligned}$$

Then, multiply all sides by  $(n-1)S^2$ :

$$\frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}}^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}}^2}$$

You'll find this in the formula sheet. You can use this result to construct the confidence interval for the variance. There will be 1 or 2 questions in the midterm for this section.

## 4 Hypothesis Testing



**Figure 5:** A reminder what Z subscript means

$H_0$  is the null hypothesis, and  $H_A$  is the alternative hypothesis. We can do the following to  $H_0$ :

- Reject it
- Fail to reject it

Do NOT accept it, as it cannot be accepted.

### 4.1 Single Sample Hypothesis Testing

A statistical hypothesis is a statement about the numerical value of a population parameter. For example:

$$H_0 : \mu = 24$$

$$H_A : \mu > 24$$

The null hypothesis ( $H_0$ ) is some claim about the population parameter that the researcher wants to test, and the alternative hypothesis ( $H_A$ ) is the value of the population parameter for which the researcher wants to gather evidence to support.





Conventionally, the null hypothesis will **always** contain the equal sign:  $\mu = 24$ , even if the context of the question suggests otherwise.

The test statistic is a sample statistic, computed from the information provided in the sample. This is used to decide between the null and alternative hypothesis.

## 4.2 Errors

Remember that  $p$ -values are calculated GIVEN  $H_0$  is true.

A **Type I** error occurs when you wrongly reject  $H_0$ . It has an  $\alpha$  chance of occurring.  $P(\text{Reject } H_0 | H_0 \text{ is true})$ . The lower you set  $\alpha$ , the less likely you're going to make an error.

A **Type II** error occurs when you fail to reject  $H_0$ , when you shouldn't. The probability it occurs is  $P(\text{Fail to reject } H_0 | H_0 \text{ is false})$

The **rejection region** of a statistical test is the set of possible values for the test statistic that results in  $H_0$  being rejected.



Mnemonic: If you fail **2** reject, you might be making a type **2** error.

## 4.3 Large Sample Tests for the POPULATION MEAN

$$H_0 : \theta = \theta_0$$

$$H_A : \theta \neq \theta_0 \quad \text{two tailed}$$

You may have to use different methods depending on what you know.

### 4.3.1 Case 1

Unknown  $\sigma$ , small sample size

The test statistic is:

$$T = \frac{\bar{Y} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

#### 4.3.2 Case 2

Large sample ( $n \geq 30$ ) or the sample comes from a normal distribution:

The test statistic is:

$$Z_C = \frac{\bar{Y} - \mu_0}{\frac{S}{\sqrt{n}}} \sim N(0, 1)$$

Replace  $S$  with  $\sigma$  if the sample variance is known, whenever possible.

### 4.4 Rejection Regions

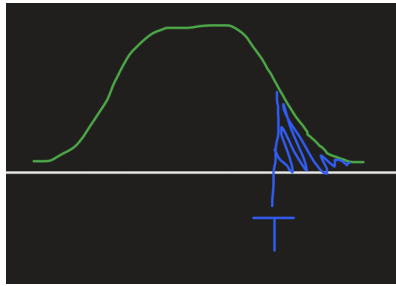
- Upper tail:  $\{Z_c > Z_\alpha\}$  I mean  $[Z_\alpha, \infty)$
- Lower tail:  $\{Z_c < -Z_\alpha\}$  I mean  $(-\infty, -Z_\alpha]$
- Two-tailed:  $\{|Z_c| > Z_{\frac{\alpha}{2}} \leftarrow \text{PAY ATTENTION TO THIS}\}$  I mean  $(-\infty, -Z_{\frac{\alpha}{2}}] \cup [Z_{\frac{\alpha}{2}}, \infty)$

Two-tailed rejection regions are stronger.

### 4.5 Calculating P-Values For The Population Mean

**Upper tail:** p-value if the alternative hypothesis looks like  $H_A : \mu > c$ , and your test statistic is large.

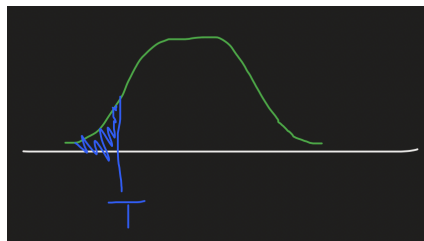
- For case 1, calculate  $P(T > t_{n-1})$ .
- For case 2, calculate  $P(Z_C > Z)$ .



**Figure 6:** Upper tail calculations

**Lower tail:** for smaller test statistics

- For case 1, calculate  $P(T < t_{n-1})$ .
- For case 2, calculate  $P(Z_C < Z)$ .

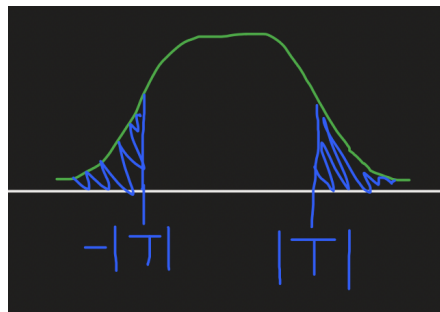


**Figure 7:** Lower tail calculations

**Both:**

Go with the approach that fits the most (depending on whether your normalized observation is negative or positive), then multiply the result by two. This means that you'll likely temporarily rewrite  $H_A : \mu \neq c$  to  $H_A : \mu > c$  ( $<$  works fine).

The two-tailed p-value is **always** double the one-tailed p-value regardless of the distribution.



**Figure 8:** Two-tailed

## 4.6 One tail vs. Two Tails

A two tailed p-value will be double the 1-tail p-value.

## 4.7 CIs and Hypothesis Testing



If a confidence interval of a population parameter captures what the null hypothesis proposes the parameter should be, then do not reject the null hypothesis.

The significance level of your confidence interval must match  $\alpha$ , or your threshold to reject the null hypothesis.

If we have a confidence interval of  $\mu$  that is 95% (remember that no % means default 95%):

And you get asked to conduct hypothesis testing,  $\mu = 100$ . Alternative hypothesis:  $\mu \neq 100$ , with  $\alpha = 0.05$ .

What is your conclusion? We don't need to conduct hypothesis testing again: if the CI captures  $\mu = 100$ , then do not reject the null hypothesis.

This avoids the need to do hypothesis testing twice, but make sure  $\alpha$  is the same.

## 4.8 Hypothesis Testing for Variance

The test statistic is:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

To calculate p-values, you may have to find a range of what the p-value could be, rather than calculate it directly as the chi-squared table isn't that precise. Only reject if  $\alpha$  is way above even the maximum of the range. The same applies for calculating the p-value using a t-distribution.

The rejection region for a  $H_A$  that uses the  $\neq$  sign is:

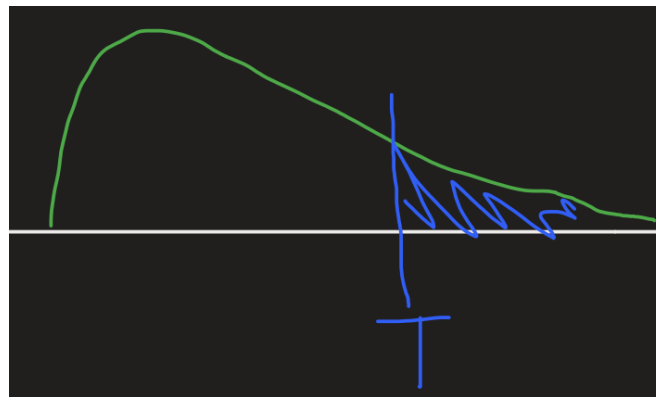
$$\left[0, \chi_{1-\frac{\alpha}{2}}^2\right] \cup \left[\chi_{\frac{\alpha}{2}}^2, \infty\right)$$

The rejection region for a  $H_A$  that uses  $\sigma \leq x$  (lower tail):

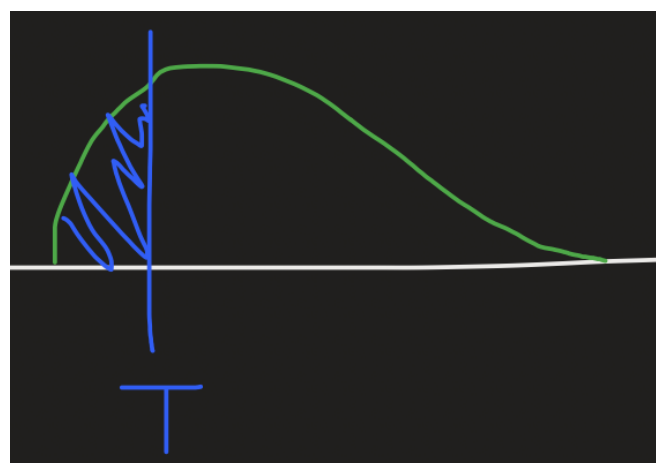
$$\left[0, \chi_{1-\frac{\alpha}{2}}^2\right]$$

The rejection region that uses  $\sigma \geq x$ :

$$\left[\chi_{\frac{\alpha}{2}}^2, \infty\right)$$



**Figure 9:** Right side p-value. Use this if your sample variance is larger than what was said in the null hypothesis.



**Figure 10:** Left side p-value. Use this if your sample variance is smaller than what was said in the null hypothesis.

For double sided, just multiply the resulting  $p$ -value by two.

#### 4.9 Steps for hypothesis testing without a calculator

1. Determine  $H_0$  and  $H_A$
2. Get the test statistic

3. Get the observed value  $Z_C$
4. Get its p-value OR compare with the rejection region
  - a. You can choose either method UNLESS you are specifically asked on a test to use a specific one from the two.
5. Make your conclusions (sufficient or insufficient evidence to reject  $H_0$ )

## 5 Multi-Sample Hypothesis Testing

We have two populations and population parameters for both of them. If we want to conduct statistical inference, we need to collect samples. **We assume the two populations are independent and our samples from each are IID.**

$$\begin{aligned} \mu_1, \sigma_1^2 &\xRightarrow{\text{sample}} n_1, \bar{y}_1, \sigma_1^2 \\ \mu_2, \sigma_2^2 &\xRightarrow{\text{sample}} n_2, \bar{y}_2, \sigma_2^2 \end{aligned}$$

We can conduct inferences on how well the population parameters match:

$$\mu_1 - \mu_2$$

Firstly, what is the point estimator for  $\mu_1 - \mu_2$ ? A simple guess:

$$\bar{y}_1 - \bar{y}_2$$

If we take the expected value, we can confirm that this is an unbiased estimator (remember that  $V(X - Y) = V(X) + V(Y)$ ):

$$\begin{aligned} E(\bar{y}_1 - \bar{y}_2) &= \mu_1 - \mu_2 \\ V(\bar{y}_1 - \bar{y}_2) &= V(\bar{y}_1) + V(\bar{y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \end{aligned}$$

When you take the square root of this, you can get the standard deviation:  $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ , which is the **standard error of the test statistic** of  $\bar{y}_1 - \bar{y}_2$ .

## 5.1 Confidence Intervals For The Mean

### 5.1.1 BOTH SAMPLES ARE LARGE

Theoretical variance is known:

$$\mu_1 - \mu_2 = \bar{y}_1 - \bar{y}_2 \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

I don't know the theoretical variance

$$\mu_1 - \mu_2 = \bar{y}_1 - \bar{y}_2 \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

### 5.1.2 EITHER OF THE SAMPLES ARE SMALL, AND...

ADDITIONAL REQUIREMENTS: Both populations must be

- Normally distributed
- Have equal variances ( $\sigma_1^2 = \sigma_2^2$ ), meaning there's no differences among the variances
- The random samples must be selected independently.
- The two populations are independent of each other

Then, the confidence interval is:



$$\begin{aligned} & (\mu_1 - \mu_2) \\ &= (\bar{y}_1 - \bar{y}_2) \\ & \pm t_{\frac{\alpha}{2}, df=n_1+n_2-2} \sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \end{aligned}$$

Where the pooled sample variance  $S_p^2$  is (essentially, we combine its effects of  $S_1$  and  $S_2$  into one):

$$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$$

By the way:

$$\frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1+n_2-2}$$

Pooled sample variance requires the population variances to match.