STA238 Notes

Statistics for real

https://github.com/ICPRplshelp/

Last updated March 16, 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 μ (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 ith person, we may have this space:

$$\boxed{Y_1,\, Y_2,\, \ldots,\, Y_{10},\, \ldots,\, Y_N} \leftarrow N = 50000$$

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:

$$\boxed{y_1,\, y_2,\, \ldots,\, y_5,\, \ldots,\, y_n} \leftarrow n = 200$$

Usually, μ and σ, 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 σ^2.

1.1 Linking Population and Samples

Population:

$$\boxed{\mu,\, \sigma^2}^{N} \quad \text{UNKNOWN}$$

$$\boxed{\bar{y},\, s^2}^{n} \quad \text{KNOWN}$$

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:**

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

All the samples are identically distributed.

For the **statistic:**

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

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

CASE 1. σ is known, AND population is normal.

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

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

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

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

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

CASE 4. σ is unknown, population is normal, and $n < 30$.

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

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

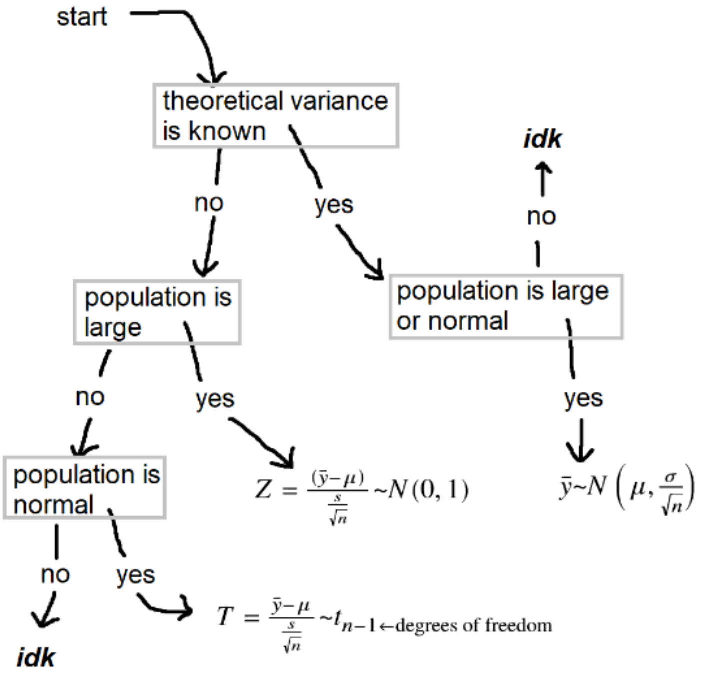
start
theoretical variance is known
no
yes
population is large
population is large or normal
idk
no
yes
no
yes
population is normal
$Z = \frac{(\bar{y}-\mu)}{\frac{s}{\sqrt{n}}} \sim N(0,1)$
$\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
no
yes
idk
$T = \frac{\bar{y}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$ ←degrees of freedom

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:

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

The total number of successes divided by sample size.

The sampling distribution for sample proportions:

$$\begin{aligned} E(\widehat{p}) &= E(\bar{y}) = \mu = p \\ &\Rightarrow \boxed{E(\widehat{p}) = p} \\ V(\widehat{p}) &= V(\bar{y}) = \frac{\sigma^2}{n} = \frac{p(1-p)}{n} \\ &\Rightarrow \boxed{V(\widehat{p}) = \frac{p(1-p)}{n}} \\ SD(\widehat{p}) &= \sqrt{\frac{p(1-p)}{n}} \end{aligned}$$

CONDITION FOR NORMAL APPROXIMATION:

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

- Expected number of successes and expected number of failures both are ≥ 10.

1.3 T-Distribution

With condition in which we don't know σ, 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

$$\begin{aligned}
&\mu \to \bar{y} \to N \text{ or } t \\
&p \to \widehat{p} \to N \\
&\sigma^2 \to s^2 \to \chi^2
\end{aligned}$$

We want to conduct statistical inferences for μ. $\bar{y}$ 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 $\widehat{p}$, 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 σ^2 is s^2.

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

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

Which is the definition of the sample variance. We can use this to make statistical inferences of σ^2. We can't find σ^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:

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

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 σ^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:

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

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 σ^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 $\sigma^2 = 1$:

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

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

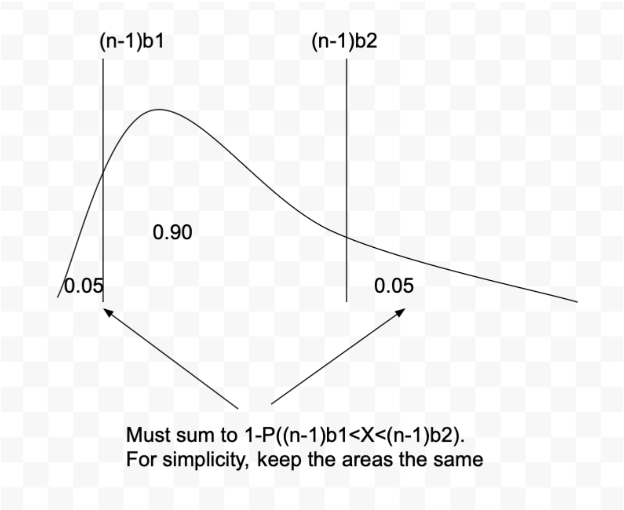
(n-1)b1
(n-1)b2
0.90
0.05
0.05
Must sum to 1-P((n-1)b1<X<(n-1)b2).
For simplicity, keep the areas the same

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 σ^2 is, then our confidence interval will likely contain σ^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:

$$\begin{aligned} W_1 &\sim \chi^2_{\nu_1 \leftarrow df,\ \text{not necessairly int}} \\ W_2 &\sim \chi^2_{\nu_2} \end{aligned}$$

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 ν_1 is the degrees of freedom):

$$F$$

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

Here, F is said to have an F distribtuon 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.

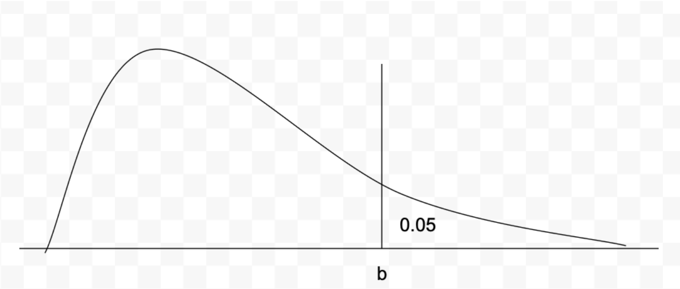

Figure 3: F-distributed density function

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

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

2 Point Estimation

Expect a question on this

ESTIMATORS

Expectation means population mean. So:

$$E\left(\overline{y}\right) = \mu$$

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

$$\underbrace{\widehat{\mu}}_{\text{estimator of } \mu} = \overline{y}$$

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.

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

$\overline{y}$ is the unbiased estimate of μ.

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

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

So, the sample variance is an unbiased estimator for σ. So:

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

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 σ^2.

2.1 The General Parameter θ

Now, we've learned about the following statistics:

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

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

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

$\widehat{\theta}$ is an unbiased estimator if $E\left(\widehat{\theta}\right) = \theta$. Otherwise, it is said to be a biased estimator. Now go ahead and replace θ and $\widehat{\theta}$ with $(\mu,\ \bar{y})$, $\left(s^2,\ \sigma^2\right)$, $(\widehat{p},\ p)$ respectively.

2.2 Bias of the Point Estimator

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

$$B\left(\widehat{\theta}\right) = E\left(\widehat{\theta}\right) - \theta$$

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

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

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

2.3 Mean Squared Error

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

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

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

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

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

2.4 Point Estimator vs. Point Estimate

An estimator is a random variable with a distribution. An estimate is like something we've actually observed.

3 Confidence Intervals

We can estimate $\widehat{\mu}$ using $\bar{y} = \underbrace{5.5}_{\text{example}}$. We want an interval estimate such as a 95% confidence interval for $\mu = \underbrace{(4.8,\ 5.9)}_{\text{example}}$. 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** α.

$$\alpha = 0.05$$

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 μ. Consider a large sample, the sample size ≥ 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

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

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

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

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

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

Multiply all sides by -1:

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

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

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

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 σ is unknown, but we have a large sample, replace σ with s.

3.1.1 Confidence Intervals for Population Proportion

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

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

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

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

3.2 CIs for Population Proportions, Again

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

$$\widehat{p} \pm Z_{\frac{\alpha}{2}} SE\left(\widehat{p}\right)$$

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

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 size of $\widehat{p}$ is the margin of error (how much we can be off, or half the size of the CI):

$$ME = Z_{\frac{\alpha}{2}} SE\left(\widehat{p}\right)$$

This is how much error we can allow based on the sample proportion. $Z_{\frac{\alpha}{2}}$ is the critical value; α 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 σ is unknown, use s instead of σ. If it is known, use σ instead of s, that is for $SE(\bar{y}) = \frac{\sigma \text{ or } s}{\sqrt{n}}$.

3.3.1 Small Sample

In cases where σ 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 α 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 σ is known **and** the sample comes from a normal distribution.

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

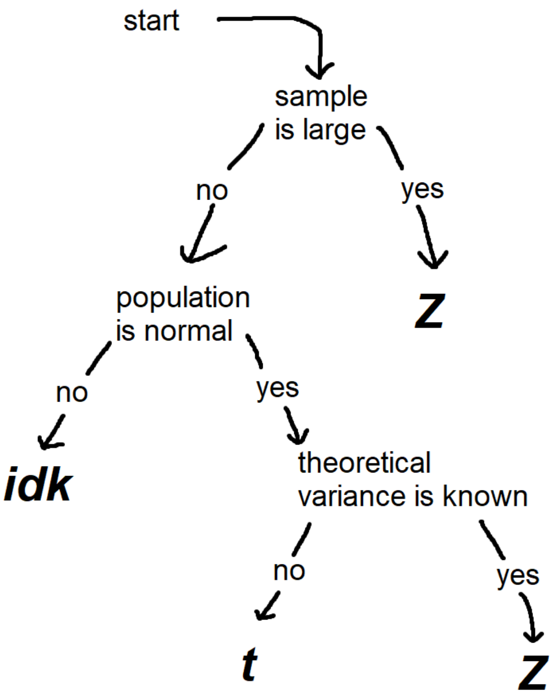

Figure 4: Flow chart.

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 μ, which is:

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

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

Where $SE\left(\bar{y}\right)=\frac{\sigma}{\sqrt{n}}$ 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, $ME=Z_{\frac{\alpha}{2}}SE\left(\bar{y}\right)$.

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.

$$R = \max\left(Y_1, \ldots, Y_N\right) - \min\left(Y_1, \ \ldots, \ Y_N\right)$$

3.6 Confidence Intervals For The Variance (Welcome Back, χ^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 σ^2. We know that

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

We assume the level of significance α. We **want** to find:

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

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

We will end up with

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

Flip all of them:

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

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

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

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.

3.7 Upper and Lower Confidence Bounds

You can create confidence intervals that go in one direction from your observed value. In the left direction, that would be your lower confidence bound, and in the right direction, that would be the upper confidence bound. In that case, the significance level would not be cut in half, and you would use ˘ or $+$ instead of $\pm$.

4 Hypothesis Testing

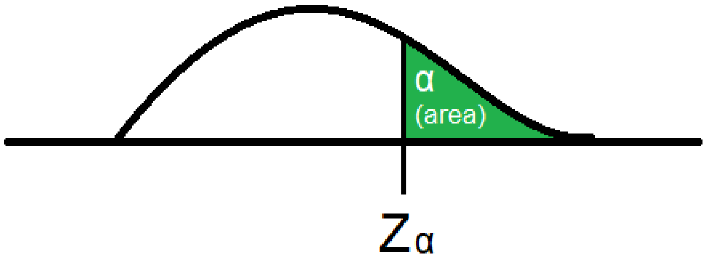

Figure 5: A reminder what Z subscript means

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 α chance of occurring. $P\left(\text{Reject } H_0 | H_0 \text{ is true}\right)$. The lower you set α, 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 $P\left(\text{Fail to reject } H_0 | H_0 \text{ is false}\right)$

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 Sample Tests for the POPULATION MEAN

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_A &: \theta \neq \theta_0 \quad \text{two tailed} \end{aligned}$$

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

If you end up in case 1 and the question doesn't tell you the samples are normally distributed, you can assume it but state that you did.

4.3.1 Case 1: t

Unknown σ, small sample size, and samples are normally distributed

The test statistic is:

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

4.3.2 Case 2: Z

Any of the two points must hold to proceed with this case:

- Large sample ($n \geq 30$), or
- The sample comes from a normal distribution and σ is known:

The test statistic is:

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

Replace S with σ 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: $\left\{|Z_c| > Z_{\frac{\alpha}{2}} \leftarrow \text{PAY ATTENTION TO THIS}\right\}$ I mean $\left(-\infty,\, -Z_{\frac{\alpha}{2}}\right] \cup \left[Z_{\frac{\alpha}{2}},\, \infty\right)$

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

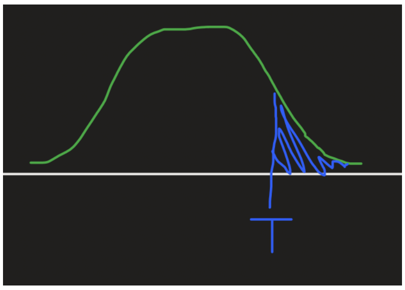

Figure 6: Upper tail calculations

Lower tail: for smaller test statistics

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

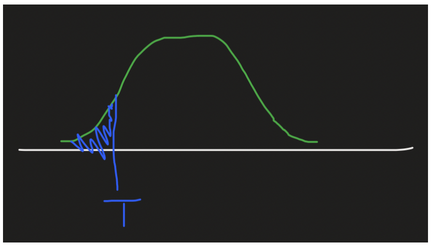

Figure 7: Lower tail calculations

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.

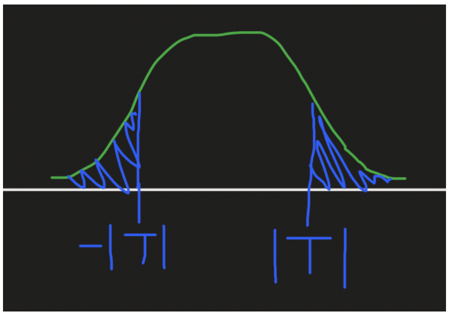

Figure 8: Two-tailed

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 α, or your threshold to reject the null hypothesis.

If we have a confidence interval of μ 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 α is the same.

4.8 Hypothesis Testing for Variance

ASSUMPTIONS: Samples are IID and the population is normally distributed.

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 α 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:

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

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

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

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

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

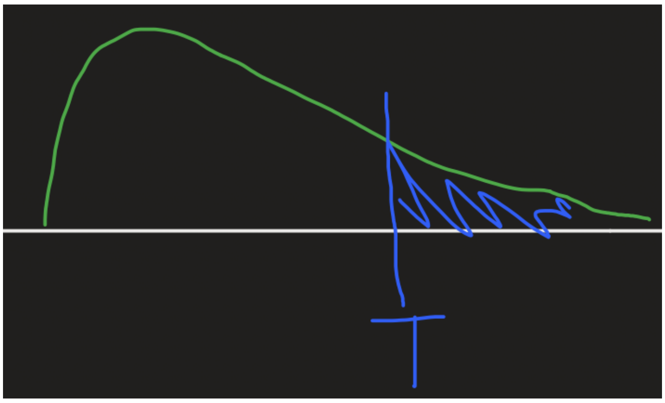

Figure 9: Right side p-value. Use this if your sample variance is larger than what was said in the null hypothesis.

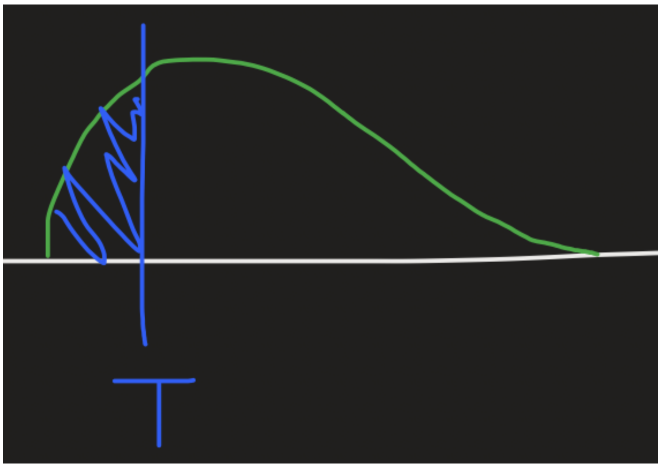

Figure 10: Left side p-value. Use this if your sample variance is smaller than what was said in the null hypothesis.

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)

4.10 How To Communicate Rejection or Acceptance Of The Null

If no evidence, say:

- Fail / do not reject H_0. We do not have sufficient evidence to indicate what was said in H_A at $\alpha = \ldots$.

If there is evidence, say:

- There is sufficient evidence to reject H_0 and indicate H_A.

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.**

$$\mu_1,\, \sigma_1^2 \overset{\text{sample}}{\Rightarrow} n_1,\, \bar{y}_1,\, \sigma_1^2$$
$$\mu_2,\, \sigma_2^2 \overset{\text{sample}}{\Rightarrow} n_2,\, \bar{y}_2,\, \sigma_2^2$$

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 $V(X-Y) = V(X) + V(Y)$):

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

When you take the square roof 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 Difference Between Means

5.1.1 BOTH SAMPLES ARE LARGE

Theoretical variance is known:

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

I don't know the theoretical variance

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

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
  - Don't want to do this? Then, you better get me a large sample (you can't)
- 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.

5.2 Confidence Intervals for Difference Between Population Proportions

To do statistical inferences between two populations, first we need to assume that both populations are independent of each other, **and we need to choose randomly**.

If we have population 1 and population 2, we can do statistical inference on $p_1 - p_2$.

- We collect a sample of n_1 from population 1 and ask if they like something. We get $\widehat{p}_1 = \frac{y_1}{n_1}$
- We collect a sample of n_2 from population 2 and ask if they like something. We get $\widehat{p}_2 = \frac{y_2}{n_2}$

The point estimator for $p_1 - p_2$ is $\widehat{p}_1 - \widehat{p}_2$. $E(\widehat{p}_1 - \widehat{p}_2) = p_1 - p_2$, so we have an unbiased estimator. The variance is

$$\begin{aligned} V\left(\widehat{p}_1-\widehat{p}_2\right) &= V\left(\widehat{p}_1\right)+V\left(\widehat{p}_2\right) \\ &= \frac{\widehat{p}_1\left(1-\widehat{p}_1\right)}{n_1}+\frac{\widehat{p}_2\left(1-\widehat{p}_2\right)}{n_2} \end{aligned}$$

The standard error is the standard deviation in this context:

$$\begin{aligned} & SE\left(\widehat{p}_1-\widehat{p}_2\right) \\ &= \sqrt{\frac{\widehat{p}_1\left(1-\widehat{p}_1\right)}{n_1}+\frac{\widehat{p}_2\left(1-\widehat{p}_2\right)}{n_2}} \end{aligned}$$

Then we can construct a confidence interval.

❗ The following steps will work only if ALL of the following applies:

- $n_1\widehat{p}_1 \geq 10$
- $n_1\left(1-\widehat{p}_1\right) \geq 10$
- $n_2\widehat{p}_2 \geq 10$
- $n_2\left(1-\widehat{p}_2\right) \geq 10$
- The number of people sampled must be less than 10% of the population
  - To ensure independence

In other words, BOTH samples must have a success and fail count of at least 10.
The normal distribution assumption cannot apply if these conditions are not met.

The $100(1-\alpha)\%$ confidence interval for p_1-p_2 is:

$$\widehat{p}_1 - \widehat{p}_2 \pm Z_{\frac{\alpha}{2}} \cdot SE\left(\widehat{p}_1 - \widehat{p}_2\right)$$

5.3 Hypothesis Testing with Difference Between Means

$$H_0 : \mu_1 - \mu_2 = D_0$$

$$H_A = \begin{cases} \mu_1 - \mu_2 > D_0 & \text{upper tail} \\ \mu_1 - \mu_2 < D_0 & \text{lower tail} \\ \mu_1 - \mu_2 \neq D_0 & \text{two tailed} \end{cases}$$

Any number can be D_0. It may depend on context, and it is not always D_0. $D_0 = 0$ makes sense the most when I'm trying to see whether there is a difference after all.

5.3.1 Small Sample

Occurs when $n_1 < 30$ OR $n_2 < 30$

Remember: $\sigma_1^2 = \sigma_2^2$ must hold or be assumed for the small sample test to be valid, **otherwise WE CANNOT TEST THIS (we need this assumption for the pooled standard deviation to work)**.

Our test statistic, under the null hypothesis, is:

$$T = \frac{\overline{y}_1 - \overline{y}_2 - D_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{df=n_1+n_2-2}$$

5.3.2 Large Sample

Use σ in place of s if it is known.

$$Z_c = \frac{\overline{y}_1 - \overline{y}_2 - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,\ 1)$$

5.4 Hypothesis Testing for Population Proportions

Similar to the two-sample variance, under $H_0:\ p_1 - p_2 = 0$, we assume that $\sigma_1^2 = \sigma_2^2$.

The **pooled** proportion (them “averaged” out) is:

$$\widehat{p} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2}$$

Then, our test statistic is:

$$Z_c = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,\ 1)$$

5.5 When solving a question like this

Follow these steps even if you could write them faster otherwise.

5.5.1 NULL AND ALTERNATE HYPOTHESIS

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_A : \mu_1 - \mu_2 \neq 0$$

5.5.2 TEST STATISTIC

$$z = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim Z(0,\ 1) \text{ or something else}$$

5.5.3 OBSERVED VALUE

$$z_c = \text{plug in the values}$$

5.5.4 REJECTION REGION

The rejection region is all $|z_c| > 1.96 = Z_{\frac{\alpha}{2}}$. You should draw out the graph. z_c is in the rejection region, so we reject the null hypothesis.

5.5.5 CONCLUSIONS

If REJECT:

- z_c is in the rejection region.
- Reject H_0.
- We have sufficient evidence to conclude that $\mu_1 \neq \mu_2$ (the alternative hypothesis) at $\alpha = TBA$

If FAIL TO REJECT:

- z_c is not in the rejection region.
- Do not reject H_0.
- We have insufficient evidence to support the fact that $\mu_1 - \mu_2 \neq 0$ at $\alpha = TBA$

Follow this exactly otherwise for some reason you'll lose marks. For the concluding sentence, you can only use symbols if you have defined them at the start explicitly. Otherwise, express it in the context of the question.

5.5.6 Give me a range of P-values

$$\begin{array}{ccc} 0.2 & > p > & 0.1 \\ \uparrow & \uparrow & \uparrow \\ 1.337 & < 1.65 < & 1.747 \end{array}$$

Is the way you should write it. Know that you may write something different if this were a one-sided test.

6 Paired Data

Let's test how effective a piece of medicine is.

We collect n patients.

Before each patient takes the medicine, we record their blood pressure.

Patient number	Before treatment	After treatment
1	$y_{1,1}$	$y_{2,1}$
2	$y_{1,2}$	$y_{2,2}$
3	$y_{1,3}$	$y_{2,3}$
n	$y_{1,n}$	$y_{2,n}$

We have n individuals but $2n$ expeirments. What makes this type of experiment different is that **the results after are dependent on the measurement before the treatment (are correlated).**

We are going to have to transform the entire dataset we have above. How about we measure the difference (denoted by d_i).

Patient number	Difference d_i
1	$d_1 = y_{1,\ 1} - y_{2,1}$
2	$d_2 = y_{1,2} - y_{2,\ 2}$
3	$d_3 = y_{1,3} - y_{2,3}$
n	$d_n = y_{1,n} - y_{2,n}$

To do hypothesis testing, we'll consider this parameter:

$$d_1,\ d_2,\ \ldots,\ d_n$$

$$\mu_d = \mu_1 - \mu_2$$

The sample mean $\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$

The sample variance $S_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \overline{d}^2\right)$

When we compute these values, we can do hypothesis testing:

$$H_0 : \mu_d = 0$$
$$H_A : \mu_d \neq 0 \text{ or } < \text{ or } >$$

If introducing something you want to see if it increases the value, test $\mu_d < 0$. And vice versa.

6.1 Paired T-Test: The Test Statistic

When we assume the null hypothesis is true and the sample is small:

$$T = \frac{\overline{d}}{\frac{S_d}{\sqrt{n}}} \sim t_{n-1}$$

Assumptions:

- Population differences are normal
- Sample differences are randomly selected from the population differences (at least we have to sample randomly)

6.2 Paired Confidence Interval

For large samples

$$\overline{d} \pm Z_{\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n_d}}$$

Or for small samples where each sample is normally distributed

$$\overline{d} \pm \frac{t_{\frac{\alpha}{2},\, df=n-1} s_d}{\sqrt{n_d}}$$

7 Moments and MLEs

Given two unbiased estimators $\widehat{\theta}_1,\ \widehat{\theta}_2$, which of them are efficient? The one with lower variance.

The efficiency of $\widehat{\theta}_1$ relative to $\widehat{\theta}_2$:

$$\text{eff}\left(\widehat{\theta}_1,\ \widehat{\theta}_2\right) = \frac{V\left(\widehat{\theta}_2\right)}{V\left(\widehat{\theta}_1\right)}$$

7.1 Consistency

AN ESTIMATOR CAN ONLY BE CONSISTENT IF IT IS UNBIASED. AN ESTIMATOR IS CONSISTENT IF THE VARIANCE CONVERGES TO 0 AS n GROWS LARGE

Consider a population parameter μ. $\overline{y}$ is an estimator for that, which is $\widehat{\theta}_n = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ (where n is the no. of samples)

This means that $\overline{y}$ depends on n.

$$E\left(\widehat{\theta}_n\right) = \mu = E\left(\overline{y}\right)$$

We know that Y has a μ and σ. $V(Y) = \sigma^2$. From the sampling distribution, $V\left(\widehat{\theta}_n\right) = V\left(\overline{y}\right) = \frac{\sigma^2}{n}$

When $n \to \infty$, $\frac{\sigma^2}{n} \to 0$.

So we can claim that $\overline{Y}$ is a consistent estimator for μ. A **consistent** estimator means that if we increase the sample size, it should converge to the true value (μ here). To verify it, show that:

$$\lim_{n\to\infty} V\left(\widehat{\theta}_n\right) = 0$$

7.2 Method of Moments

The kth moment of a distribution is $E\left(X^k\right)$. The kth sample moment is $\frac{1}{n}\sum_{i=1}^{n} X_i^k$.

$$\begin{aligned} &Y \sim (\mu,\, \sigma) \\ &Y_1,\, Y_2,\, \ldots,\, Y_n \text{ are the random samples} \end{aligned}$$

We do not know their values, yet $E(Y) = \mu$.

We'll find the method of moments estimator. There's a procedure to find it:

$$\widehat{E(Y)} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Above is how we get the method of moments estimator. This means that the method of moments estimator is $\overline{Y}$, as $\mu = \overline{Y}$.

If we want σ^2, we need the second moment: $E\left(Y^2\right)$. The variance formula is $\sigma^2 = V(Y) = E\left(Y^2\right) - E(Y)^2$. From there, we are able to compute $E\left(Y^2\right) = \sigma^2 + \mu^2$.

The sample version of $E\left(Y^2\right)$ is:

$$\begin{aligned} \widehat{E\left(Y^2\right)} &= \frac{1}{n}\sum_{i=1}^{n} (Y_i)^2 \\ \widehat{\sigma}^2 + \widehat{\mu}^2 &= \frac{1}{n}\sum_{i=1}^{n} (Y_i)^2 \\ \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} (Y_i)^2 - \overline{Y}^2 \\ \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 \end{aligned}$$

This is a biased estimator for σ^2.

7.2.1 Want to say that again?

The method of moments is used to find any unknown parameter that can be found in the expected value, based on sample moments.

A pattern for solving this problem: we want to figure out an estimate for the unknown parameter θ. We know that X follows a distribution that involves the parameter θ. And by calculating the expected value of that distribution, we attain, where g is some function:

$$E(X) = g(\theta)$$

Simultaneously, we want to assign the theoretical expected value to the random sample mean:

$$E(X) = \overline{X} = g(\theta)$$

Then the method of moments estimator $\widehat{\theta}$ is

$$\widehat{\theta} = g^{-1}\left(\overline{X}\right)$$

DO BEWARE THAT $V\left(\overline{X}\right) = \frac{\sigma^2}{n}$

IF YOU ARE REQUIRED TO FIND MULTIPLE ESTIMATORS, YOU'LL GET A SYSTEM OF EQUATIONS. There is no concrete way to solve them, just like doing LaGrange multipliers.

7.3 Maximum Likelihood Estimation

Assume that $Y_1, Y_2, \ldots, Y_n$ are random samples. We're going to take a random sample from the exponential distribution:

$$f(y) = \frac{1}{\theta} e^{-\frac{y}{\theta}}, y > 0$$

The definition of the likelihood function denoted by L is a joint probability density function:

$$L(y_1, y_2, \ldots, y_n \mid \theta) = f(y_1, y_2, \ldots, y_n)$$

Because all samples are IID, we can product them:

$$= \prod_{i=1}^{n} f(y_i)$$

For the exponential density function, it is:

$$\mathscr{L}(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{y_i}{\theta}} = \frac{1}{\theta^n} e^{-\frac{1}{\theta}\sum_{i=1}^{n} y_i}$$

And that is the likelihood function for the exponential distribution.

To find the MAXIMUM LIKELIHOOD ESTIMATOR (MLE), OPTIMIZE THE LIKELIHOOD FUNCTION (take the log of it if you need to as log is injective AND strictly increasing, and has some nice properties). Moreover, $\widehat{\theta} = \underset{\theta}{\text{argmax}} \mathscr{L}(\theta)$

If you need multiple estimators, optimize with respect to each variable separately

The log likelihood function is denoted with l. You may see $l(\theta)$ be used if we're trying to find the MLE for θ.

If you can't optimize that function due to it breaking the extreme value theorem, good luck making θ depend on Y_i, which will MOST likely be $Y_{(1)}$ or $Y_{(n)}$. You'll really have to look at the context. Remember that the MLE is the θ that is most likely to have occurred with the samples I've gotten $Y_1 \ldots Y_n$.

One hint is to assign some value to y_i for all i, and see the behavior of the likelihood function. You'll all on your own by that point.

8 Regression

- $x \to$ independent variable.
- $y \to$ dependent variable.

Deterministic model: $y = mx$. This implies that y can be found **exactly** when we know x. There is no allowance for error. In real situations, we cannot find deterministic (ideal) models. Hence, in statistics, we allow error terms, known as random error.

This is a probabilistic model:

$$y = mx + \text{Random error}$$

It includes both deterministic components and random components. For example, if we hypothesis that response time y is related to percentage BAC x, then

$$y = 1.5x + \text{Random error}$$

8.1 Sampling and Scatter plots

If we collect samples $(x_1, y_1), (x_2, y_2), \ldots$, we can plot them on a scatter plot.

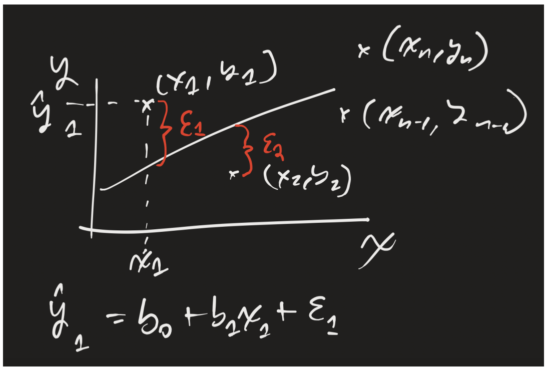

Figure 11: A scatter plot

We have $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Here, $\beta_0,\ \beta_1$ are unknown parameters. In probabilistic models, we allow some error ε_i., By assumption, $\varepsilon_i \sim N(0,\ \sigma)$.

The error:

$$\varepsilon_i = y_i - b_0 - b_1 x_i$$

Normally, in our model, we want to minimize the average error squared:

$$\begin{aligned} g(b_0,\ b_1) &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \end{aligned}$$

To find the best line of fit, we want to minimize $\sum_{i=1}^{n} \varepsilon_i^2$.

When we minimize this, we are able to get the estimate for the unknown parameters $\beta_0,\ \beta_1$.

We minimize just like any other optimization problem: by differentiating.

$$\frac{\partial g(b_0,\ b_1)}{\partial b_0} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$$

And I also need to differentiate with respect to b_1:

$$\frac{\partial g(b_0,\ b_1)}{\partial b_1} = -2\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i)$$

Equate both to 0 and find b_0 and b_1.

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0 \\
&\Rightarrow \sum_{i=1}^{n} y_i - nb_0 - b_1 \sum_{i=1}^{n} x_i = 0 \\
&\Rightarrow nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad (1)
\end{aligned}$$

From the other equation, which is from $\frac{\partial g}{\partial b_1}$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \quad (2)$$

It follows that

$$\begin{aligned}
&(1) \times \sum_{i=1}^{n} x_i - (2) \times n \\
&= b_1 \left(\left(\sum_{i=1}^{n} x_i \right)^2 - n \sum_{i=1}^{n} x_i^2 \right) \\
&= \left(\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i \right) - n \sum_{i=1}^{n} x_i y_i \\
\Rightarrow b_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{S_{xy}}{S_{xx}}
\end{aligned}$$

(See below where S_{xy} and S_{xx} comes from)

For b_0L

$$\begin{aligned}
nb_0 + b_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\
b_0 + b_1\overline{x} = \overline{y} &\quad \text{div both sides by } n \\
\Rightarrow b_0 &= \overline{y} - b_1\overline{x}
\end{aligned}$$

This means from the dataset; we are able to find the slope and intercept values.

$$b_0 = \widehat{\beta_0},\, b_1 = \widehat{\beta_1}$$

8.2 Notations of Simple Linear Regression

- y is dep
- x is independent variable or predictor.
  - You CANNOT use categorical variables
- $\beta_0 + \beta_1 x$ is the deterministic component
- ε is the random error component, and is assumed to follow a $N(0,\, \sigma)$ distribution
- β_0 is the y-intercept of the line
- β_1 is the slope of the line

8.3 Least Squares

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \\ &= \sum_{i=1}^{n} (x_i y_i - \overline{x} y_i - \overline{y} x_i + \overline{xy}) \\ &= \sum_{i=1}^{n} x_i y_i - \overline{x} \sum_{i=1}^{n} y_i - \overline{y} \sum_{i=1}^{n} x_i + n\overline{xy} \end{aligned}$$

Note that $\sum_{i=1}^{n} x_i = n\overline{x}$, so:

$$\begin{aligned} &= \sum_{i=1}^{n} x_i y_i - \overline{xy} n - \overline{y} n \overline{x} + n\overline{xy} \\ &= \sum_{i=1}^{n} x_i y_i + n\overline{xy} \end{aligned}$$

This means that

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\overline{xy}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

8.4 Residuals and Estimating σ

The ith fitted value is:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

$$= \overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 x_i$$
$$= \overline{y} + \widehat{\beta}_1 (x_i - \overline{x})$$

Residual values: how do we estimate it?

$$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{\varepsilon}_i$$

$\widehat{\varepsilon}_i = e_i$ is the estimated residual. Find it using this equation:

$$\widehat{\varepsilon}_i = e_i = \widehat{y}_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$$

RESIDUALS MEAN ESTIMATED ERROR.

ε_i is the original error; $\widehat{\varepsilon}_i$ is the estimated error.

8.5 Error Sum of Squares

The error sum of squared (or residual sum of squares), denoted by SSE, is:

$$\text{SSE} = \sum_{i=1}^{n} (e_i - \overline{e})^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

And the least squares estimate of σ^2 is:

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-2}$$

The **residual** standard deviation is:

$$\widehat{\sigma} = \sqrt{\frac{\text{SSE}}{n-2}}$$

$$\text{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

8.6 Decomposition of Total Sum of Squares

$$\begin{aligned} y_i - \overline{y} &= (y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y}) \\ \underbrace{\sum_{i=1}^{n} (y_i - \overline{y})^2}_{\text{SST}} &= \underbrace{\sum_{i=1}^{n} (y_i - \overline{y}_i)}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})}_{\text{SSReg}} \end{aligned}$$

8.7 Interpretation

If

$$\widehat{y} = b_1 + b_0 x$$

Then we estimate that as x increases by 1, y increases by b_0, we expect it to. That is the meaning of the slope.

If $x = 0$, then $\widehat{y} = b_1$.

8.8 Coefficient of Determination

R^2 measures the extent of the relationship. It is defined by

$$R^2 = 1 - \frac{SSE}{SST}$$

Interpreted as the proportion of observed y variation that can be explained with the linear regression model.

8.9 Hypothesis Testing With The Slope

We can find that, based on $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

We can do the following hypothesis tests:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

H_0 means no relationship between x and y.

We need some assumptions:

- The scatter plot should be linear, not curved
- ε_i should be independent
- Equal variance for all points in the scatter plot. This means no cone shape
- Error component follows normal distribution: $\varepsilon_i \sim N(0,\ \sigma)$

Based on these assumptions, we can conduct hypothesis testing for $\beta_1 = 0$ or $\beta_1 \neq 0$.

To find the test statistic:

$$\widehat{\beta}_1 \sim N\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right)$$

The standardized version of $\widehat{\beta}_1$ is:

$$T = \frac{\widehat{\beta}_1 - \underset{\text{under } H_0}{\beta_1}}{\frac{\widehat{\sigma}}{\sqrt{S_{xx}}}} \sim N(0,\ 1)$$

Unfortunately, σ is unknown, so we will consider $\widehat{\sigma}$. So, it is:

$$T = \frac{\widehat{\beta}_1 - \underset{\text{under } H_0}{\beta_1}}{\frac{\widehat{\sigma}}{\sqrt{S_{xx}}}} \sim t_{n-2}$$

(If n is large, you may use normal distribution)

Here, we have two parameters, so we use $t-2$.

$$T = \frac{\widehat{\beta}_1 - \beta_1}{S_{\widehat{\beta}_1}} \sim t_{n-2}$$

Hypothesis testing:

$$H_0 : \beta_1 = \beta_{1_0}$$
$$H_A : \beta_1 <, >, \neq \beta_{1_0}$$

$$SE\left(\widehat{\beta}_1\right) = \frac{\widehat{\sigma}}{\sqrt{S_{xx}}}$$