Chapter 4 – Statistical Testing and Modelling

Sampling

Sampling

- One of the great benefits of statistical models is that a reasonably sized (>30) random sample will almost always reflect the population.
- The challenge becomes, how do we select members randomly, and avoid bias?

Sampling Bias

- There are several forms of bias:

  Selection Bias

  Perhaps the most common, this type of bias favors those members of a population who are more inclined and able to answer polls.

Sampling Bias

Selection Bias

Undercoverage Bias: making too few observations or omitting entire segments of a population

Sampling Bias

Selection Bias

Self-selection Bias: people who volunteer may differ significantly from those in the population who don't

Sampling Bias

- Selection Bias
  - Healthy-user Bias: the sample may come from a healthier segment of the overall population –people who walk/jog, work outside, follow healthier behaviors, etc.

Undercoverage Bias

- A hospital survey of employees conducted during daytime hours
- Neglects to poll people who work the night shift.

Self-Selection Bias

- An online survey about a sports team
- Only people who feel strongly about the team will answer the survey.

Sampling Bias

- Survivorship Bias
  - If a population improves over time,
  - it may be due to lesser members leaving the population due to death, expulsion, relocation, etc.

A Classic Puzzle

- At the start of World War I, British soldiers wore cloth caps.
- The war office became alarmed at the high number of **head injuries**, so they issued metal helmets to all soldiers.

A Classic Puzzle

- They were surprised to find that the number of head injuries *increased* with the use of metal helmets.
- If the intensity of fighting was the same before and after the change, why should the number of head injuries increase?

A Classic Puzzle

- Answer: You have to consider *all* of the data
- Before the switch, many things that gave head injuries to soldiers wearing metal helmets would have caused fatalities for those wearing cloth caps!

Another Survivorship Example

- In World War II, statistician Abraham Wald worked for America’s Statistical Research Group (SRG)

Adapted from https://en.wikipedia.org/wiki/Abraham_Wald

Another Survivorship Example

- One problem the SRG worked on was to examine the distribution of damage to aircraft by enemy fire and to advise the best placement of additional armor.

Another Survivorship Example

- Common logic was to provide greater protection to parts that received more damage.

Another Survivorship Example

- Wald saw it differently – he felt that damage must be more uniformly distributed and that aircraft that could return had been hit in less vulnerable parts.

Another Survivorship Example

- Wald proposed that the Navy reinforce the areas where returning aircraft were undamaged, since those were areas that, if hit, would cause the plane to be lost!

Sampling Distribution

Sampling Distribution

There are three distinct types of distribution of data which are –

1.Population Distribution, characterizes the distribution of elements of a population

2.Sample Distribution, characterizes the distribution of elements of a sample drawn from a population

3.Sampling Distribution, describes the expected behavior of a large number of simple random samples drawn from the same population.

Sampling distributions constitute the theoretical basis of statistical inference and are of considerable importance in business decision-making. **Sampling distributions** are important in statistics because they provide a major simplification on the route to statistical inference.

Definition

- A sampling distribution is a theoretical probability distribution of a statistic obtained through a large number of samples drawn from a specific population
- A sampling distribution is a graph of a statistics(i.e. mean, mean absolute value of the deviation from the mean,range,standard deviation of the sample, unbiased estimate of variance, variance of the sample) for sample data.

CHARACTERISTICS

- Usually a univariate distribution.
- Closely approximate a normal distribution.
- Sample statistic is a random variable – sample mean , sample & proportion
- A theoretical probability distribution
- The form of a sampling distribution refers to the shape of the particular curve that describes the distribution.

Functions of sampling distribution

SAMPLING DISTRIBUTION IS A GRAPH WHICH PERFORM SEVERAL DUTIES TO SHOW DATA GRAPHICALLY.

SAMPLING DISTRIBUTION WORKS FOR :

- MEAN
- MEAN ABSOLUTE VALUE OF THE DEVIATION FROM THE MEAN
- RANGE
- STANDARD DEVIATION OF THE SAMPLE
- UNBIASED ESTIMATE OF THE SAMPLE
- VARIANCE OF THE SAMPLE

WHY SAMPLING DISTRIBUTION IS IMPORTANT????

PROPERTIES OF STATISTICS

SELECTION OF DISTRIBUTIO TYPE TO MODEL SCORE

HYPOTHESIS TESTING

<u>i)Properties of Statistic :</u> Statistic have different properties as estimators of a population parameters. The sampling distribution of a statistic provides a window into some of the important properties. For example if the expected value of a statistic is equal to the expected value of the corresponding population parameter, the statistic is said to be unbiased

Consistency is another valuable property to have in the estimation of a population parameter, as the statistic with the smallest standard error is preferred as an estimator estimator A statistic used to estimate a model parameter.of the corresponding population parameter, everything else being equa.l

ii) Selection of distribution type to model scores :

The sampling distribution provides the theoretical foundation to select a distribution for many useful measures. For example, the central limit theorem describes why a measure, such as intelligence, that may be considered a summation of a number of independent quantities would necessarily be distributed as a normal (Gaussian) curve.

iii) Hypothesis Testing :

The sampling distribution is integral to the hypothesis testing procedure. The sampling distribution is used in hypothesis testing to create a model of what the world would look like given the null hypothesis was true and a statistic was collected an infinite number of times. A single sample is taken, the sample statistic is calculated, and then it is compared to the model created by the sampling distribution of that statistic when the null hypothesis is true. If the sample statistic is unlikely given the model, then the model is rejected and a model with real effects is more likely.

Types of sampling distribution

The types of sampling distribution are as follows:

1) Sampling Distribution of the Mean:
Sampling distribution of means of a population data is defined as the theoretical probability distribution of the sample means which are obtained by extracting all the possible samples having the same size from the given population.

Given a finite population with mean (m) and variance (s^2). When sampling from a normally distributed population, it can be shown that the distribution of the sample mean will have the following properties -

Properties of the sampling distribution

1. The distribution of $\bar{X}$ will be normal.
2. The mean $\mu_{\bar{x}}$ of the distribution of the values of $\bar{X}$ will be the same as the mean of the population from which the samples were drawn; $\mu_{\bar{x}} = \mu$.
3. The variance, $\sigma^2_{\bar{x}}$, of the distribution of $\bar{X}$ will be equal to the variance of the population divided by the sample size; $\sigma^2_{\bar{x}} = \sigma^2 / n$

8

2) Sampling Distribution of the Proportion :

SAMPLING DISTRIBUTION OF THE PROPORTION IS FOUND WHEN THE SAMPLE PROPORTION AND PROPORTION OF SUCCESSES ARE GIVEN.

PROPERTIES :

- SAMPLE PROPORTION TEND TO TARGET THE VALUE OF PROPORTION.

- UNDER CERTAIN CONDITIONS, THE DISTRIBUTION OF SAMPLE PROPORTION CAN BE APPROXIMATED BY A NORMAL DISTRIBUTION.

Example:

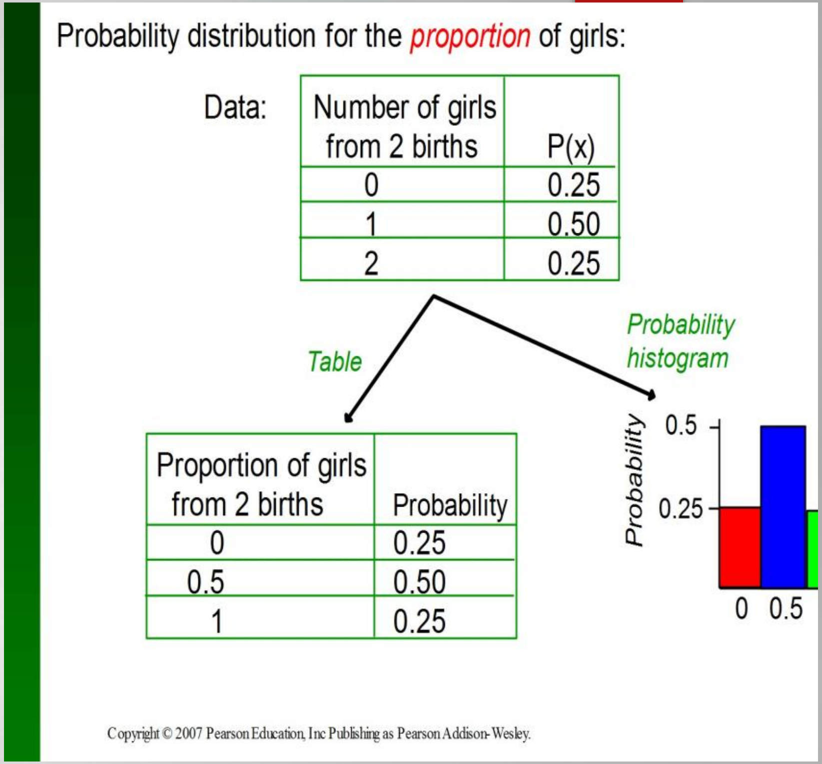
Sample distribution of the proportion of the girls from sample space for two randomly selected births:bb,bg,gb,gg
All four outcomes are equally likely:

Probabilities:

P(0 girls)=0.25

P(1 girl)=0.50

P(2girls)=0.75



Central Limit Theorem

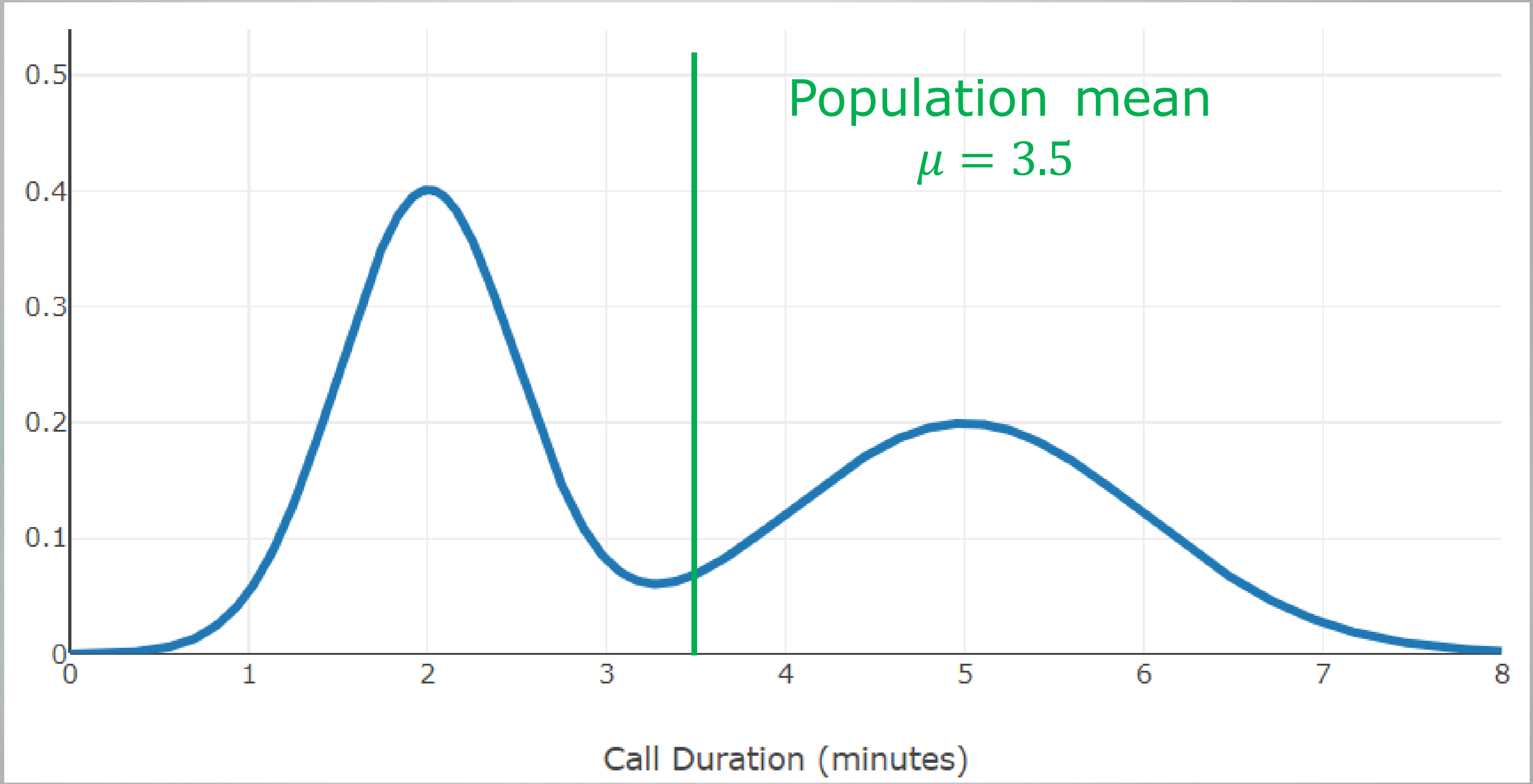
Central Limit Theorem

- What makes sampling such a good statistical tool is the Central Limit Theorem
- Recall that a sample mean often varies from the population mean.
- The CLT considers a large number of random sample tests.

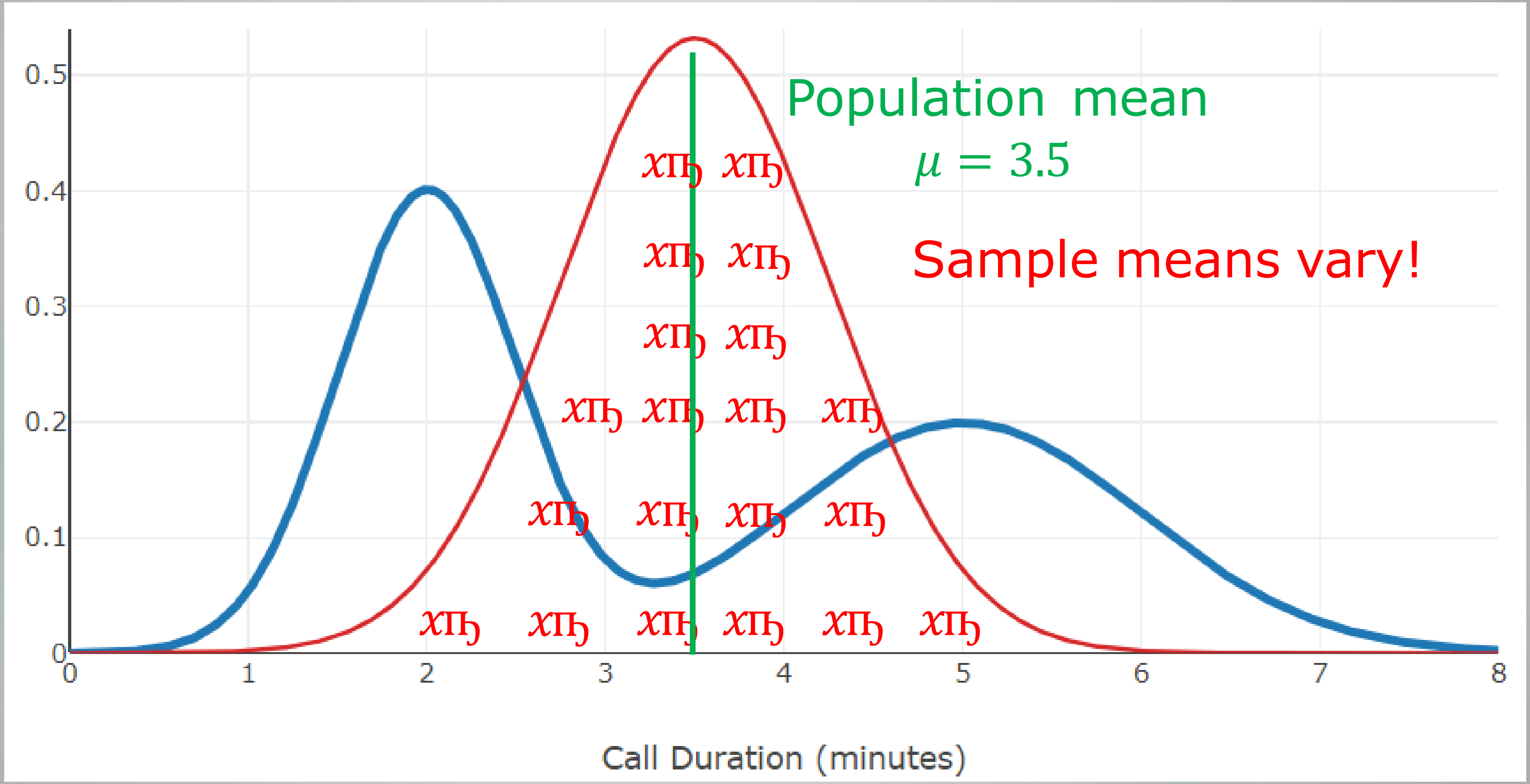
Central Limit Theorem

- The CLT states that the mean values from a group of samples will be *normally distributed* about the population mean, even if the population itself is not normally distributed.
- That is, 95% of all sample means should fall within 2σ of the population mean

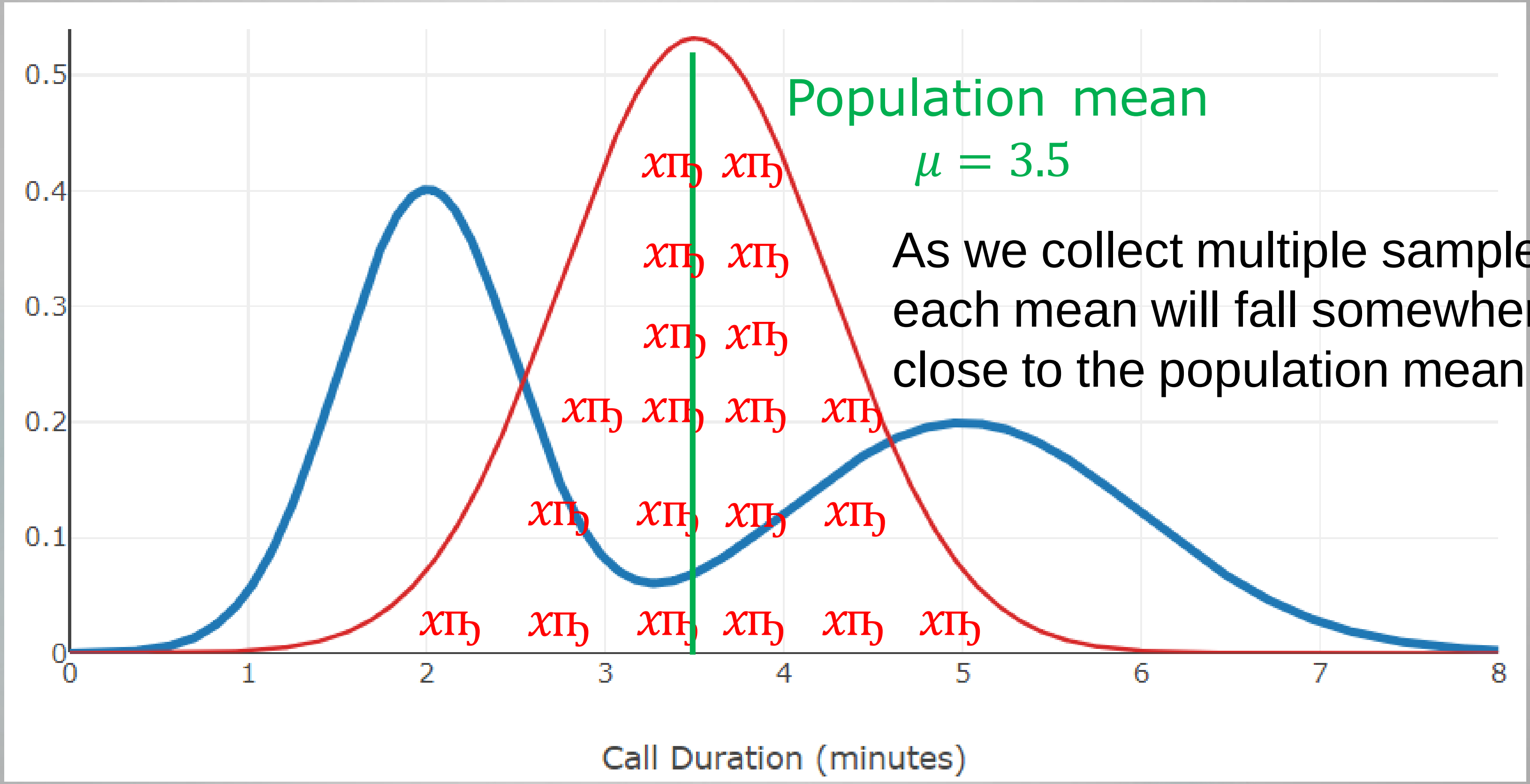
Central Limit Theorem



Central Limit Theorem



Central Limit Theorem

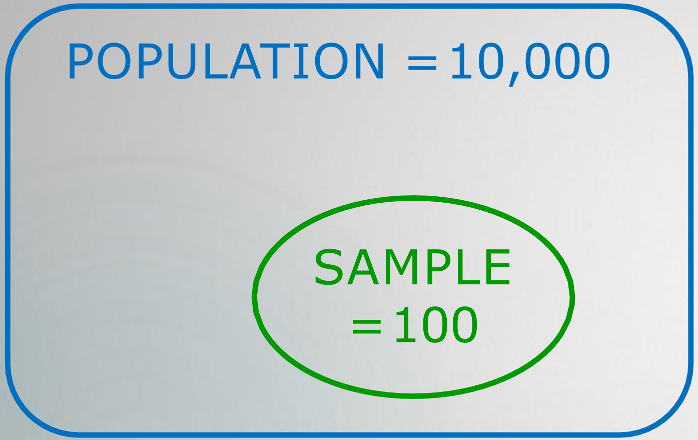


Standard Error

Standard Error

- Let’s quickly review terminology
- Let’s say we have a population of voters
- It is unrealistic to poll the entire population, so we poll a sample
- We calculate a statistic from that sample that lets us estimate a parameter of the population

Standard Error



N = # population members

P = population parameter

σ = pop. standard deviation

n = # sample members

$\hat{p}$ = sample statistic

$SE_{\hat{p}}$ = standard error of the sample

Standard Error

POPULATION = 10,000

SAMPLE = 100

- If for the population of Australia, the mean height is 5′9″, and for our 100-person sample the mean height is 5′10″, then

$$P = 5'9''$$

$$\hat{p} = 5'10''$$

$$SE_{\hat{p}} = Standard\ Error\ of\ the\ Mean$$

Standard Error of the Mean

- Where the population standard deviation describes how wide individual values stray from the population mean,
  the Standard Error of the Mean describes how far a sample mean may stray from the population mean.

Standard Error of the Mean

- If the population standard deviation σ is known, then the sample standard error of the mean can be calculated as:

$$SE = \frac{\sigma}{\sqrt{n}}$$

Standard Error Exercise

- An IQ Test is designed to have a mean score of 100 with a standard deviation of 15 points.
- If a sample of 10 scores has a mean of 104, can we assume they come from the general population?

Standard Error Exercise

- Sample of 10 IQ Test scores:

$$n = 10 \quad \bar{x} = 104 \quad \sigma = 15$$

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{10}} = 4.743$$

- 68% of 10-item sample means are expected to fall between 95.257 and 104.743

Hypothesis Testing

Hypothesis Testing

- Hypothesis Testing is the application of statistical methods to real-world questions.
- We start with an assumption, called the null hypothesis
- We run an experiment to test this null hypothesis

Hypothesis Testing

- Based on the results of the experiment, we either reject or fail to reject the null hypothesis
- If the null hypothesis is rejected, then we say the data supports another, mutually exclusive alternate hypothesis
- We never "PROVE" a hypothesis!

Framing the Hypothesis

- How do we frame the question that forms our null hypothesis?
- At the start of the experiment, the null hypothesis is assumed to be true.
- If the data fails to support the null hypothesis, only then can we look to an alternative hypothesis

Framing the Hypothesis

If testing something assumed to be true,
the null hypothesis can reflect the assumption:

Claim: *"Our product has an average shipping weight of 3.5kg."*

Null hypothesis: average weight = 3.5kg

Alternate hypothesis: average weight ≠ 3.5kg

Framing the Hypothesis

If testing a claim we *want* to be true,
but can't assume, we test its opposite:

Claim: *"This prep course improves test scores."*

Null hypothesis: old scores ≥ new scores

Alternate hypothesis: old scores < new scores

Framing the Hypothesis

The null hypothesis should contain an equality (=, ≤ ,≥):

average shipping weight = 3.5kg $\boxed{H_0\colon\ \mu = 3.5}$

The alternate hypothesis should not have an equality (≠,<,>):

average shipping weight ≠ 3.5kg $\boxed{H_1\colon\ \mu \neq 3.5}$

Framing the Hypothesis

The null hypothesis should contain an equality (=, ≤ ,≥):

old scores ≥ new scores

$$H_0\colon \mu_0 \geq \mu_1$$

The alternate hypothesis should not have an equality (≠,<,>):

old scores < new scores

$$H_1\colon \mu_0 < \mu_1$$

Hypothesis Testing

- So what lets us reject or fail to reject the null hypothesis?

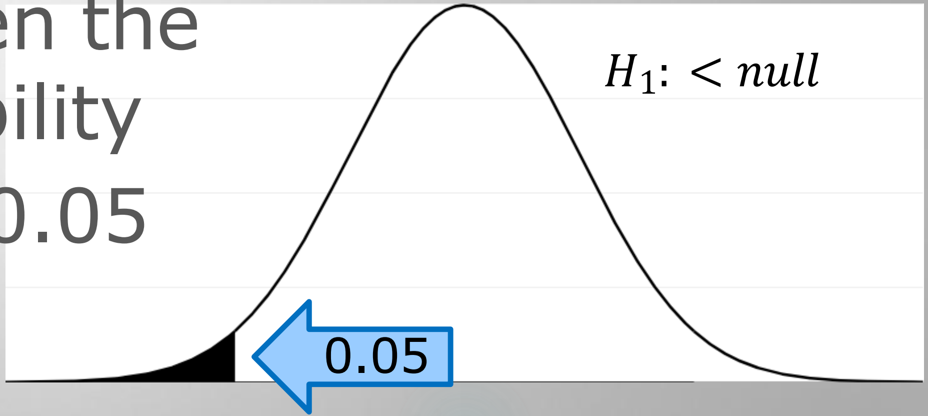
Hypothesis Testing

- We run an experiment and record the result.
- Assuming our null hypothesis is valid, if the probability of observing these results is very small (inside of 0.05) then we reject the null hypothesis.
- Here 0.05 is our level of significance

$$\alpha = 0.05$$

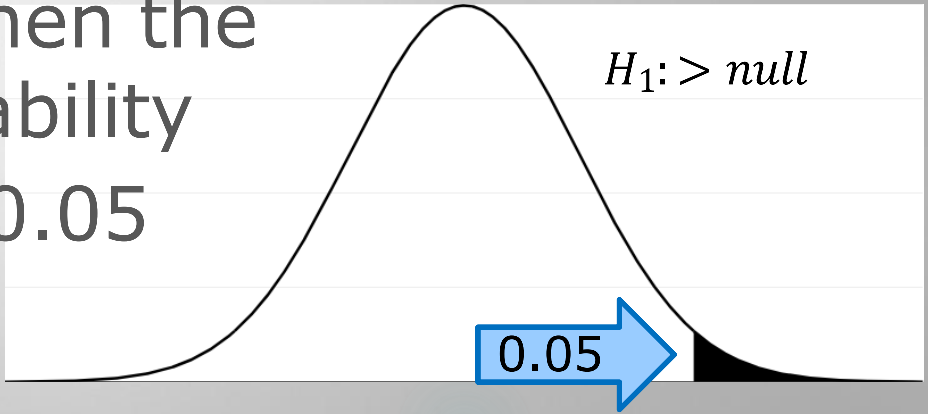
Hypothesis Testing - Tails

- The level of significance α is the area inside the *tail(s)* of our null hypothesis.
- If $\alpha = 0.05$ and the alternative hypothesis is *less than* the null, then the left-tail of our probability curve has an area of 0.05



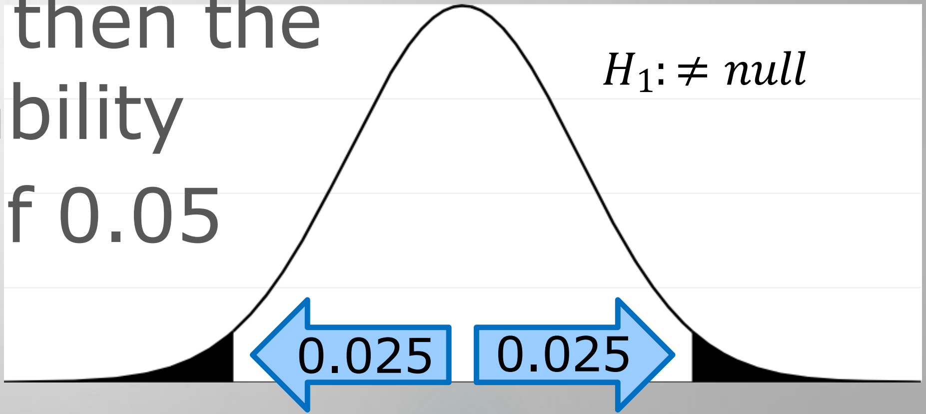
Hypothesis Testing - Tails

- The level of significance α is the area inside the *tail(s)* of our null hypothesis.
- If $\alpha = 0.05$ and the alternative hypothesis is *more than* the null, then the right-tail of our probability curve has an area of 0.05



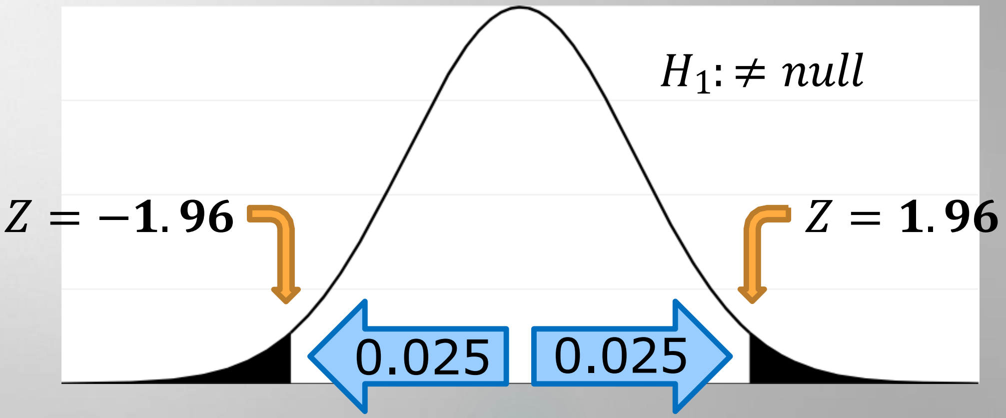
Hypothesis Testing - Tails

- The level of significance α is the area inside the *tail(s)* of our null hypothesis.
- If $\alpha = 0.05$ and the alternative hypothesis is *not equal to* the null, then the two tails of our probability curve *share* an area of 0.05



Hypothesis Testing - Tails

- These areas establish our critical values or Z-scores:



Tests of Mean vs. Proportion

- we'll work through full examples of Hypothesis Testing.
- There are two main types of tests:
- Test of Means
- Test of Proportions

Tests of Mean vs. Proportion

- Each of these two types of tests has their own test statistic to calculate.
- Let's review the situation for each test before we work through some examples in the upcoming lectures.

Tests of Mean vs. Proportion

- Mean
  when we look to find an average, or specific value in a population we are dealing with means
- Proportion
  whenever we say something like "35%" or "most" we are dealing with proportions

Test Statistics

- When working with means:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

assumes we know the population standard deviation

- When working with proportions:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}}$$

Hypothesis Testing – P-value Test

In a traditional test:

- take the level of significance α
- use it to determine the critical value
- compare the test statistic to the critical value

In a P-value test:

- take the test statistic
- use it to determine the P-value
- compare the P-value to the level of significance α

Hypothesis Testing – P-value Test

"If the P-value is low,
the null must go!"
reject H_0

"If the P-value is high,
the null must fly!"
fail to reject H_0

Testing Example Exercise #1

Testing Exercise #1- Mean

- For this next example we'll work in the left-hand side of the probability distribution, with negative z-scores
- We'll show how to run the hypothesis test using the traditional method, and then with the P-value method

Testing Exercise #1- Mean

- A company is looking to improve their website performance.
- Currently pages have a mean load time of 3.125 seconds, with a standard deviation of 0.700 seconds.
- They hire a consulting firm to improve load times.

$$\boxed{\begin{aligned}\mu &= 3.125\\ \sigma &= 0.700\end{aligned}}$$

Testing Exercise #1- Mean

- Management wants a 99% confidence level
- A sample run of 40 of the new pages has a mean load time of 2.875 seconds.
- Are these results statistically faster than before?

$$\mu = 3.125$$
$$\sigma = 0.700$$
$$\alpha = 0.01$$
$$n = 40$$
$$\bar{x} = 2.875$$

Testing Solution #1- Mean

1. State the null hypothesis:

$$H_0\colon \mu \geq 3.125$$

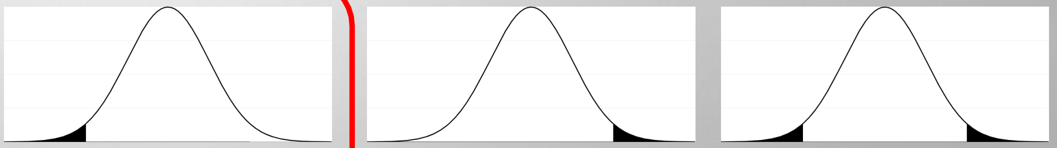
2. State the alternative hypothesis:

$$H_1\colon \mu < 3.125$$

3. Set a level of significance:

$$\alpha = 0.01$$

4. Determine the test type:



left tail | right tail | two tail

Testing Solution #1- Mean

TRADITIONAL METHOD:

5. Test Statistic:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{2.875 - 3.125}{0.7/\sqrt{40}} = -2.259$$

6. Critical Value:

z-table lookup on 0.01 $z = -2.325$

$\mu = 3.125$
$\sigma = 0.700$
$\alpha = 0.01$
$n = 40$
$\bar{x} = 2.875$

$Z = -2.259$
$z = -2.325$

Testing Solution #1- Mean

TRADITIONAL METHOD:

7. Fail to Reject the Null Hypothesis

Since $-2.259 > -2.325$, the test statistic falls outside the rejection region

We can't say that the new web pages are statistically faster.

$$\mu = 3.125$$
$$\sigma = 0.700$$
$$\alpha = 0.01$$
$$n = 40$$
$$\bar{x} = 2.875$$

$$Z = -2.259$$
$$z = -2.325$$

Testing Solution #1- Mean

P-VALUE METHOD:

5. Test Statistic:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{2.875 - 3.125}{0.7/\sqrt{40}} = -2.259$$

6. P-Value:

z-table lookup on -2.26 $P = 0.0119$

$$\mu = 3.125$$
$$\sigma = 0.700$$
$$\alpha = 0.01$$
$$n = 40$$
$$\bar{x} = 2.875$$

$$Z = -2.259$$
$$P = 0.0119$$

Testing Solution #1- Mean

P-VALUE METHOD:

7. Fail to Reject the Null Hypothesis

Since $0.0119 > 0.01$, the P-value is greater than the level of significance α

We can't say that the new web pages are statistically faster.

$$\mu = 3.125$$
$$\sigma = 0.700$$
$$\alpha = 0.01$$
$$n = 40$$
$$\bar{x} = 2.875$$

$$Z = -2.259$$
$$P = 0.0119$$

Testing Example Exercise #2

Testing Exercise #2 - Proportion

- A video game company surveys 400 of their customers and finds that 58% of the sample are teenagers.
- Is it fair to say that most of the company's customers are teenagers?

Testing Solution #2 - Proportion

1. Set the null hypothesis: $H_0: P \leq 0.50$
2. Set the alternative hypothesis: $H_1: P > 0.50$
3. Calculate the test statistic:

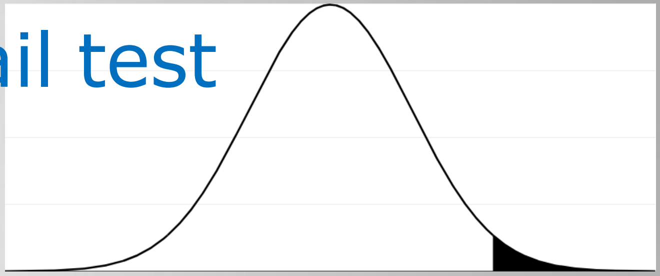
$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{0.58 - 0.50}{\sqrt{\frac{0.50(1 - 0.50)}{400}}} = \frac{0.08}{0.025} = \mathbf{3.2}$$

Testing Solution #2 - Proportion

4. Set a significance level: $\alpha = 0.05$
5. Decide what type of tail is involved:

   $H_1\colon P > 0.50$ means a right-tail test
6. Look up the critical value:

   $$\mathbf{Z = 1.645}$$

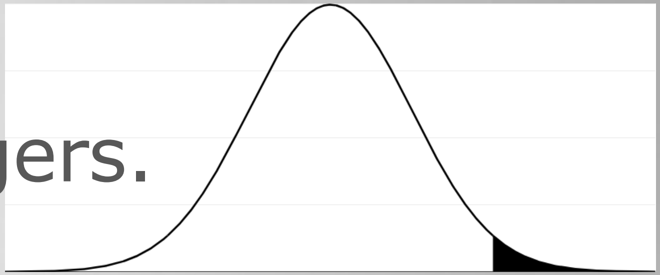


Critical Value = 1.645

Test Statistic = 3.2

Testing Solution #2 - Proportion

7. Based on the sample,
   we reject the null hypothesis,
   and support the claim that
   most customers are teenagers.

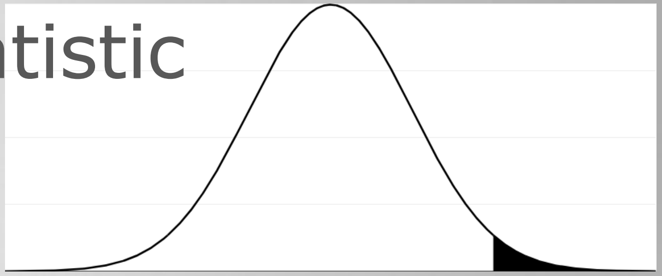


Critical Value = 1.645

Test Statistic = 3.2

Testing Solution #2 - Proportion

NOTE: The size of the sample matters! If we had started with a sample size of 40 instead of 400, our test statistic would have been only 1.01, and we would fail to reject the null hypothesis.



Critical Value = 1.645

Test Statistic = 3.2

Type 1and Type 2 Errors

Type I and Type II Errors

- Often in medical fields (and other scientific fields) hypothesis testing is used to test against results where the "truth" is already known.
- For example, testing a new diagnostic test for cancer for patients you have already succesfully diagnosed by other means.

Type I and Type II Errors

- In this situation, you already know if the Null Hypothesis is True or False.
- In these situations where you already know the "truth", then you would know its possible to commit an error with your results .

Type I and Type II Errors

- This type of analysis is common enough that these errors already have specific names:
- Type I Error
- Type II Error

Type I and Type II Errors

- If we reject a null hypothesis that should have been supported, we've committed a **Type I Error**

H_0: *There is no fire*

Pull the fire alarm, only to find out there really was no fire.

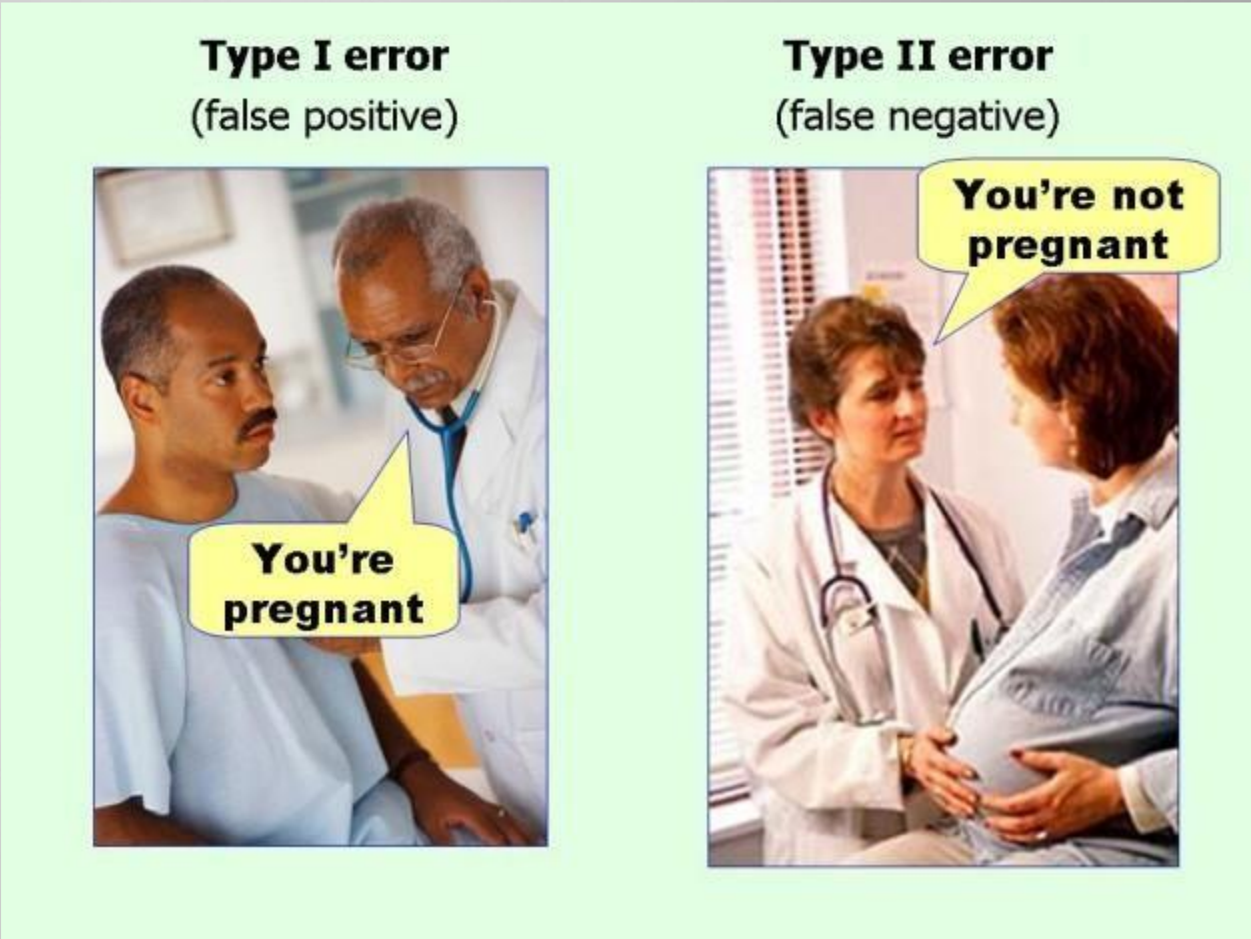
Type I and Type II Errors

- If we fail to reject a null hypothesis that should have been rejected we've committed a Type II Error

H_0: *There is no fire*

Don't pull the fire alarm, only to find there really is a fire.

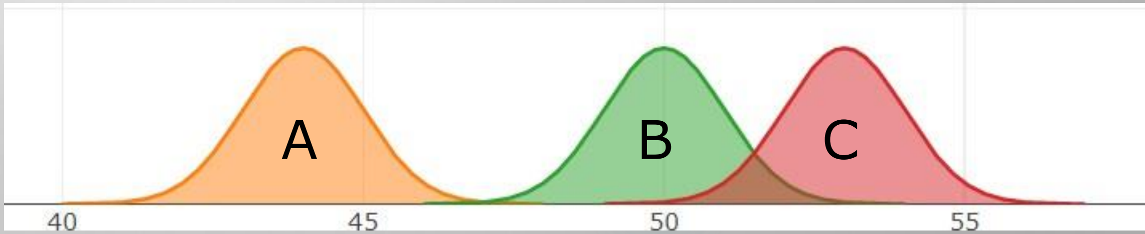
H_0: $Not\ pregnant$
H_1: $Are\ pregnant$



ANOVA
Analysis of Variance

ANOVA

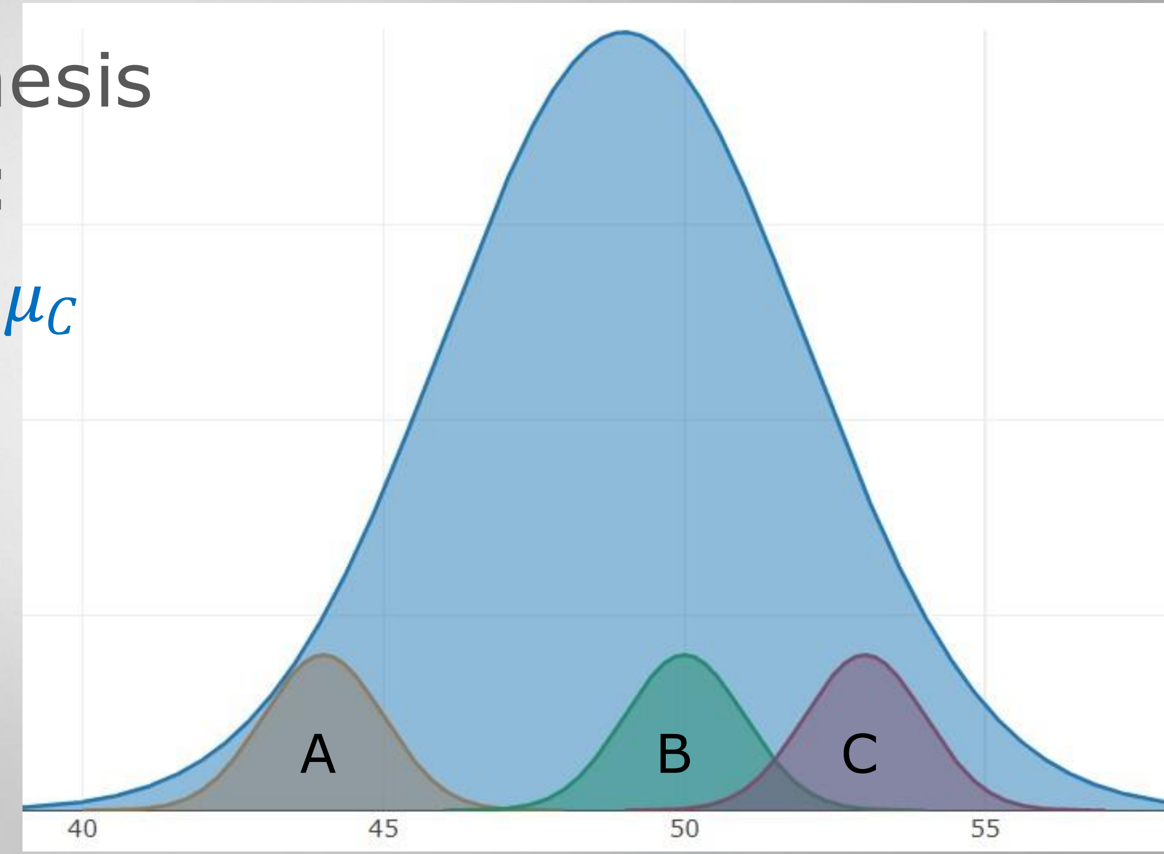
- In the previous section we tested two samples to see if they likely came from the same parent population.
- What if we had three (or more) samples?
- Could we do the same thing?



ANOVA

- Our null hypothesis would look like:

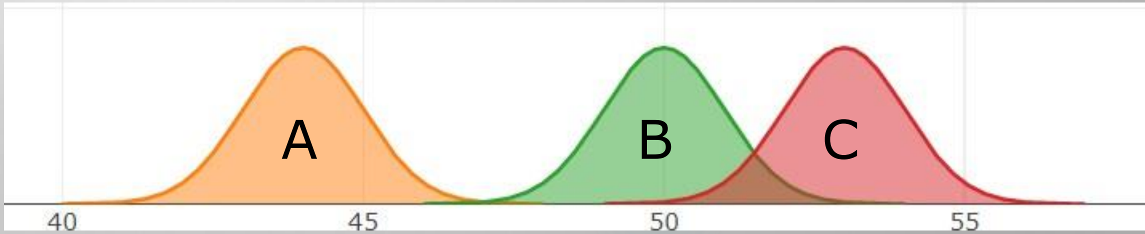
$$H_0\colon \mu_A = \mu_B = \mu_C$$



ANOVA

- We *could* test each pair:

$$H_0\colon \mu_A = \mu_B \quad \alpha = 0.05$$
$$H_0\colon \mu_A = \mu_C \quad \alpha = 0.05$$
$$H_0\colon \mu_B = \mu_C \quad \alpha = 0.05$$



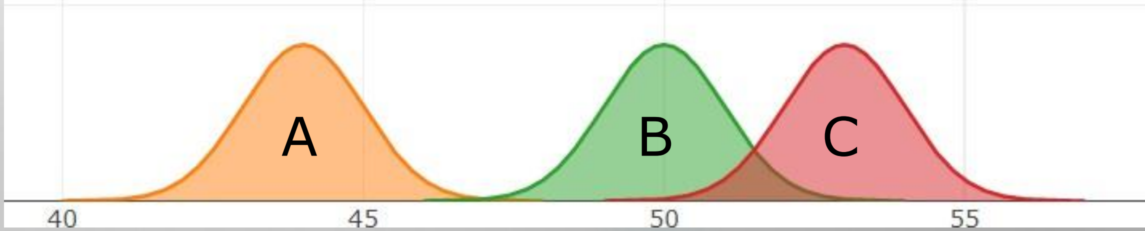
ANOVA

- The problem is, our overall confidence drops:

$$H_0\!: \mu_A = \mu_B \quad \alpha = 0.05$$
$$H_0\!: \mu_A = \mu_C \quad \alpha = 0.05$$
$$H_0\!: \mu_B = \mu_C \quad \alpha = 0.05$$

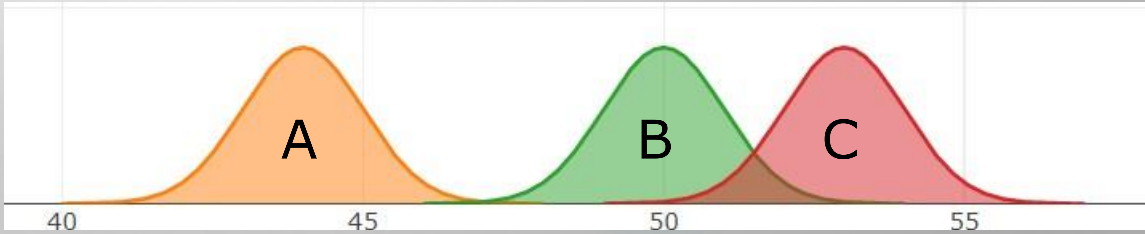
$$.95 \times .95 \times .95 = .857$$

$$85.7\%\ confidence\ level$$



ANOVA

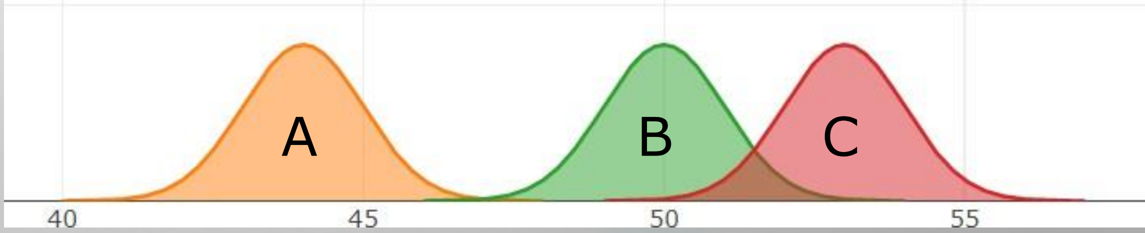
- This is where ANOVA comes in!
- We compute an F value, and compare it to a critical value determined by our degrees of freedom (the number of groups, and the number of items in each group)



ANOVA

Let's work with some data:

GroupA	GroupB	GroupC
37	62	50
60	27	63
52	69	58
43	64	54
40	43	49
52	54	52
55	44	53
39	31	43
39	49	65
23	57	43



ANOVA

First calculate the sample means

Next calculate the overall mean

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	44	50	53
μ_{TOT}	49		

ANOVA

ANOVA considers two types of variance:

- Between Groups
  - how far group means stray from the total mean
- Within Groups
  - how far individual values stray from their respective group mean

ANOVA

The F value we’re trying to calculate is simply the ratio between these two variances!

$$F = \frac{Variance\ Between\ Groups}{Variance\ Within\ Groups}$$

ANOVA

Recall the equation for variance:

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{SS}{df}$$

Here $\Sigma(x-\bar{x})^2$ is the "sum of squares" SS

and $n-1$ is the "degrees of freedom" df

ANOVA

So the formula for the F value becomes:

$$F = \frac{Variance\ Between\ Groups}{Variance\ Within\ Groups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

SSG = Sum of Squares Groups $\quad df_{groups}$ = degrees of freedom (groups)

SSE = Sum of Squares Error $\quad df_{error}$ = degrees of freedom (error)

ANOVA

$$SSG = 420$$

Sum of Squares Groups

$$(\mu_A - \mu_{TOT})^2 = (44 - 49)^2 = 25$$

$$(\mu_B - \mu_{TOT})^2 = (50 - 49)^2 = 1$$

$$(\mu_C - \mu_{TOT})^2 = (53 - 49)^2 = 16$$

$$42$$

Multiply by the number of items in each group:

$$42 \times 10 = 420$$

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

ANOVA

$$SSG = 420$$
$$df_{groups} = 2$$

GroupA	GroupB	GroupC
37	62	50
60	27	63
52	69	58
43	64	54
40	43	49
52	54	52
55	44	53
39	31	43
39	49	65
23	57	43

$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

Degrees of Freedom Groups

$$df_{groups} = n_{groups} - 1$$
$$= 3 - 1$$
$$= 2$$

ANOVA

$$SSG = 420$$
$$df_{groups} = 2$$
$$SSE = 3300$$

Sum of Squares Error

$(x_A-\mu_A)^2$	$(x_A-\mu_A)^2$	$(x_B-\mu_B)^2$	$(x_B-\mu_B)^2$	$(x_C-\mu_C)^2$	$(x_C-\mu_C)^2$
49	64	144	16	9	1
256	121	529	36	100	0
64	25	361	361	25	100
1	25	196	1	1	144
16	441	49	49	16	100
	1062		**1742**		**496**

TOTAL	**3300**

$(37-44)^2$
$=(-7)^2$
$=49$

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

ANOVA

$$SSG = 420$$
$$df_{groups} = 2$$
$$SSE = 3300$$
$$df_{error} = 27$$

GroupA	GroupB	GroupC
37	62	50
60	27	63
52	69	58
43	64	54
40	43	49
52	54	52
55	44	53
39	31	43
39	49	65
23	57	43

$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

Degrees of Freedom Error

$$df_{error} = (n_{rows} - 1) * n_{groups}$$
$$= (10 - 1) * 3$$
$$= 27$$

ANOVA

$$SSG = 420$$
$$df_{groups} = 2$$
$$SSE = 3300$$
$$df_{error} = 27$$

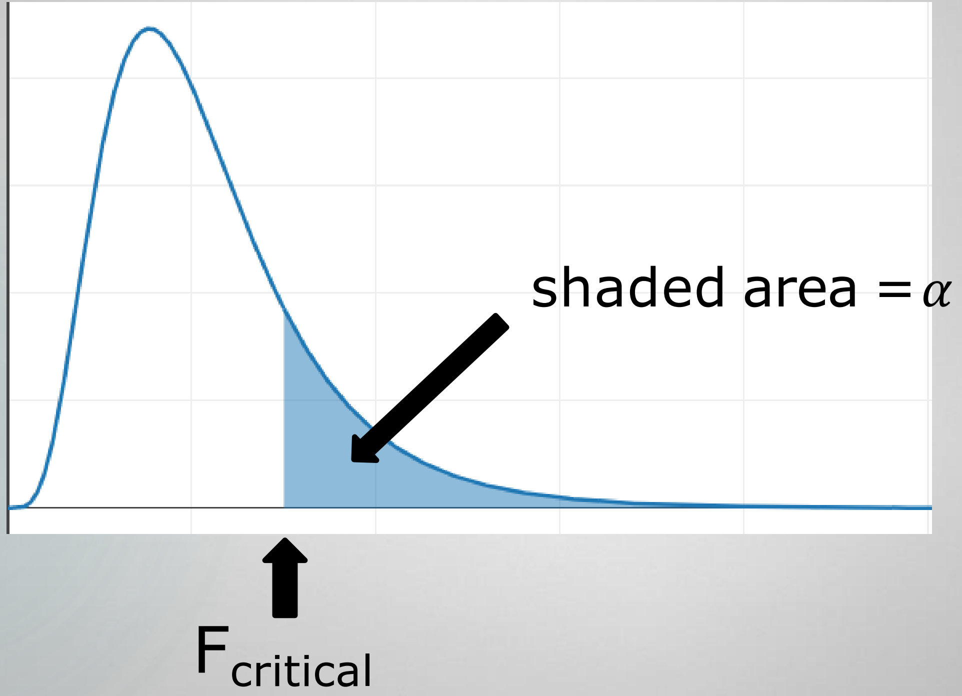
Plug these into our formula:

$$F = \frac{\frac{SS}{df_{group}}}{\frac{SS}{df_{error}}} = \frac{\frac{420}{2}}{\frac{3300}{27}} = \frac{210}{122.22} = \mathbf{1.718}$$

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

F Distribution

F-Distribution



F-Distribution

Look up our critical value from an F-table

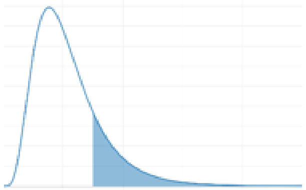
use a table set for

95% confidence

find numerator df

find denominator df

critical value = 3.35



F-Table Upper Tail Area of 0.05

denominator df	Numerator df 1	2	3	4	5
25	4.24	3.39	2.99	2.76	2.60
26	4.23	3.37	2.98	2.74	2.59
27	4.21	3.35	2.96	2.73	2.57
28	4.20	3.34	2.95	2.71	2.56
29	4.18	3.33	2.93	2.70	2.55
30	4.17	3.32	2.92	2.69	2.53

ANOVA

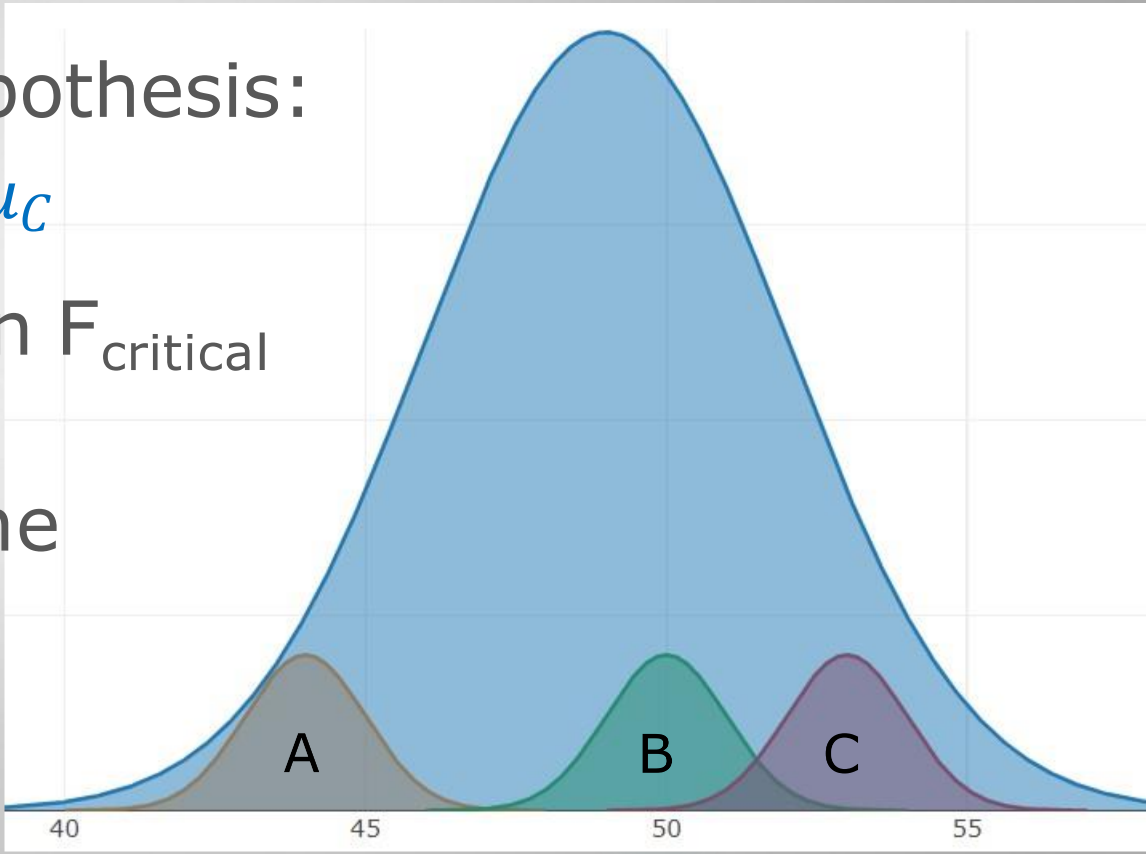
Recall our null hypothesis:

$$H_0\colon \mu_A = \mu_B = \mu_C$$

Since F is less than $F_{critical}$

$$1.718 < 3.354$$

we fail to reject the null hypothesis!



ANOVA Exercise #1

- In an effort to receive faster payment of invoices, a company introduces two discount plans
- One set of customers is given a 2%discount if they pay their invoice early
- Another set is offered a 1%discount
- A third set is not offered any incentive

ANOVA Exercise #1

- The results are as follows:
- Using ANOVA, can we say that the offers result in faster payments?

2% disc	1% disc	no disc
11	21	14
16	15	11
9	23	18
14	10	16
10	16	21

ANOVA Exercise #1

1 Calculate the means

	2% disc	1% disc	no disc
	11	21	14
	16	15	11
	9	23	18
	14	10	16
	10	16	21
$\mu_{2,1,0}$	**12**	**17**	**16**
μ_{TOT}	**15**		

$SSG = 70$

ANOVA Exercise #1

2. Find Sum of Squares Groups

$$(\mu_2 - \mu_{TOT})^2 = (12 - 15)^2 = 9$$
$$(\mu_1 - \mu_{TOT})^2 = (17 - 15)^2 = 4$$
$$(\mu_0 - \mu_{TOT})^2 = (16 - 15)^2 = 1$$
$$14$$

Multiply by the number of items in each group:

$$14 \times 5 = 70$$

	2% disc	1% disc	no disc
	11	21	14
	16	15	11
	9	23	18
	14	10	16
	10	16	21
$\mu_{2,1,0}$	12	17	16
μ_{TOT}	15		

ANOVA Exercise #1

$$SSG = 70$$
$$df_{groups} = 2$$

3. Degrees of Freedom Groups

$$df_{groups} = n_{groups} - 1$$
$$= 3 - 1$$
$$= 2$$

	2% disc	1% disc	no disc
	11	21	14
	16	15	11
	9	23	18
	14	10	16
	10	16	21
$\mu_{2,1,0}$	12	17	16
μ_{TOT}	15		

ANOVA Exercise #1

$$SSG = 70$$
$$df_{groups} = 2$$
$$SSE = 198$$

4. Sum of Squares Error

$(x_2-\mu_2)^2$	$(x_1-\mu_1)^2$	$(x_0-\mu_0)^2$
1	16	4
16	4	25
9	36	4
4	49	0
4	1	25
34	**106**	**58**

TOTAL	

	2% disc	1% disc	no disc
	11	21	14
	16	15	11
	9	23	18
	14	10	16
	10	16	21
$\mu_{2,1,0}$	**12**	**17**	**16**
μ_{TOT}	**15**		

ANOVA Exercise #1

$$SSG = 70$$
$$df_{groups} = 2$$
$$SSE = 198$$
$$df_{error} = 12$$

5. Degrees of Freedom Error

$$df_{error} = (n_{rows} - 1) * n_{groups}$$
$$= (5 - 1) * 3$$
$$= 12$$

2% disc	1% disc	no disc
11	21	14
16	15	11
9	23	18
14	10	16
10	16	21

$\mu_{2,1,0}$	12	17	16
μ_{TOT}	15		

ANOVA Exercise #1

$$SSG = 70$$
$$df_{groups} = 2$$
$$SSE = 198$$
$$df_{error} = 12$$

2% disc	1% disc	no disc
11	21	14
16	15	11
9	23	18
14	10	16
10	16	21

$\mu_{2,1,0}$	12	17	16
μ_{TOT}	15		

6. Calculate F value:

$$F = \frac{\frac{SS}{df_{groups}}}{\frac{SS}{df_{error}}} = \frac{\frac{70}{2}}{\frac{198}{12}} = \frac{35}{16.5} = \mathbf{2.121}$$

7. Look up $F_{critical}$: $\mathbf{3.885}$

ANOVA Exercise #1

$$SSG = 70$$
$$df_{groups} = 2$$
$$SSE = 198$$
$$df_{error} = 12$$

Since F falls to the left of $F_{critical}$

$$2.121 < 3.885$$

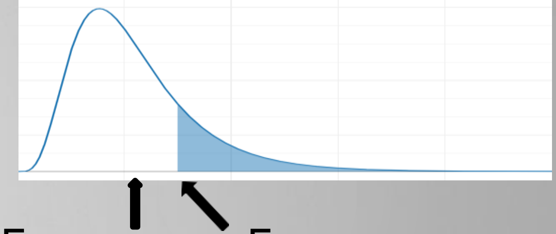
we fail to reject the null hypothesis!

$F_{calculated}$ =2.121

$F_{critical}$ =3.885

ANOVA Exercise #1

We don’t have enough to support the idea that our offers changed the average number of days that customers took to pay their invoices!



Two-Way ANOVA

One-Way vs Two-Way ANOVA

- In the previous examples we used one-way ANOVA to test one independent variable.
- For the invoice problem, the independent variable was the incentive offered.
- The dependent variable was the time it took to receive payment.

One-Way vs Two-Way ANOVA

- Two-Way ANOVA lets us test two independent variables at the same time
- For the invoice example, we might also consider the amount due
- We would have 3 invoices for $50, 3 for $100, etc. and offer different incentives at each dollar amount.

One-Way vs Two-Way ANOVA

- The resulting data might look like this:
- Here, each row or dollar amount is called a block.
- Essentially, we want to isolate and remove any variance contributed by the blocks, to better understand the variance in the groups.

	2% disc	1% disc	no disc
$50	16	23	21
$100	14	21	16
$150	11	16	18
$200	10	15	14
$250	9	10	11

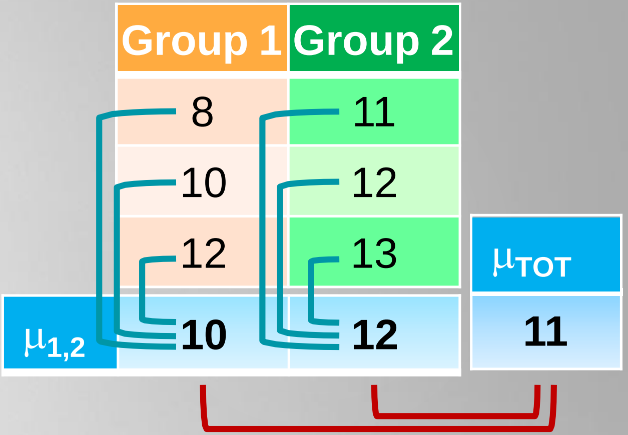
One-Way vs Two-Way ANOVA

- So how do we do that?

	2% disc	1% disc	no disc
$50	16	23	21
$100	14	21	16
$150	11	16	18
$200	10	15	14
$250	9	10	11

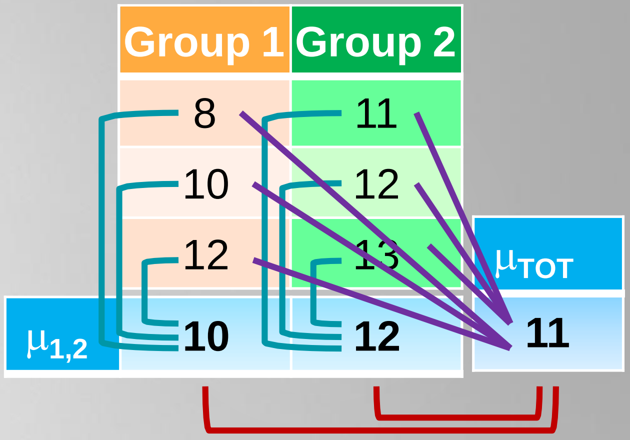
Two-Way ANOVA

- The goal of ANOVA is to separate different aspects of the total variance.
- In the previous examples we had only
  Sum of Squares Groups (SSG) » between groups
  and Sum of Squares Error (SSE) » within groups



Two-Way ANOVA

- These two variances SSG and SSE add up to our total variance
  Sum of Squares Total (SST)



Sum of Squares Groups (SSG) » between groups

and Sum of Squares Error (SSE) » within groups

Two-Way ANOVA

- Now we'll look at variance between rows, or blocks

	Group 1	Group 2	
Block A	8	11	
Block B	10	12	
Block C	12	13	μ_{TOT}
$\mu_{1,2}$	**10**	**12**	

Sum of Squares Groups (SSG) » between groups

and Sum of Squares Error (SSE) » within groups

Two-Way ANOVA

- First calculate the block means
- Then calculate the Sum of Squares Blocks (SSB) Sum of Squares Groups (SSG) and Sum of Squares Error (SSE)
  - » between blocks
  - » between groups
  - » within groups

	Group 1	Group 2	$\mu_{A,B,C}$
Block A	8	11	
Block B	10	12	
Block C	12	13	
$\mu_{1,2}$			11

Two-Way ANOVA

$$F = \frac{Var.Between\ Groups}{Var.Within\ Groups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

- ANOVA still considers the relationship between the SSG and the SSE

	Group 1	Group 2	$\mu_{A,B,C}$
Block A	8	11	
Block B	10	12	
Block C	12	13	
$\mu_{1,2}$			11

Sum of Squares Blocks (SSB) » between blocks

Sum of Squares Groups (SSG) » between groups

and Sum of Squares Error (SSE) » within groups

Two-Way ANOVA

$$F = \frac{Var.Between\ Groups}{Var.Within\ Groups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

	Group 1	Group 2	$\mu_{A,B,C}$
Block A	8	11	
Block B	10	12	
Block C	12	13	
$\mu_{1,2}$			11

- By calculating the SSB, we remove some of the variance in SSE

Sum of Squares Blocks (SSB) » between blocks

Sum of Squares Groups (SSG) » between groups

and Sum of Squares Error (SSE) » within groups

Two-Way ANOVA

$$F = \frac{Var.Between\ Groups}{Var.Within\ Groups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

Sum of Squares Groups (SSG)

$$(\mu_1 - \mu_{TOT})^2 = (10 - 11)^2 = 1$$
$$(\mu_2 - \mu_{TOT})^2 = (12 - 11)^2 = 1$$
$$\overline{\quad\quad 2 \quad\quad}$$

multiply by the number of items in each group: $2 \times 3 = 6$

	Group 1	Group 2	$\mu_{A,B,C}$
Block A	8	11	
Block B	10	12	
Block C	12	13	
$\mu_{1,2}$			11

$$\boxed{SSG = 6}$$

Two-Way ANOVA

$$F = \frac{Var.Between\ Groups}{Var.Within\ Groups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

Sum of Squares Blocks (SSB)

	Group 1	Group 2	$\mu_{A,B,C}$
Block A	8	11	
Block B	10	12	
Block C	12	13	
$\mu_{1,2}$			11

$$(\mu_A - \mu_{TOT})^2 = (9.5 - 11)^2 = 2.25$$

$$(\mu_B - \mu_{TOT})^2 = (11 - 11)^2 = 0$$

$$(\mu_C - \mu_{TOT})^2 = (12.5 - 11)^2 = 2.25$$

$$4.5$$

$SSG = 6$

$SSB = 9$

multiply by the number of items in each block: $4.5 \times 2 = 9$

Two-Way ANOVA

$$F = \frac{Var.Between\ Groups}{Var.Within\ Groups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

Sum of Squares Total (SST)

$$(8-11)^2+(11-11)^2+$$
$$(10-11)^2+(12-11)^2+$$
$$(12-11)^2+(13-11)^2 = 16$$

no need to multiply since every item is represented

	Group 1	Group 2	$\mu_{A,B,C}$
Block A	8	11	
Block B	10	12	
Block C	12	13	
$\mu_{1,2}$			11

$SSG = 6$
$SSB = 9$
$SST = 16$

Two-Way ANOVA

$$F = \frac{Var.Between\ Groups}{Var.Within\ Groups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

Sum of Squares Error (SSE)

$$SSE = SST - SSG - SSB$$

$$= 16 - 6 - 9 = 1$$

	Group 1	Group 2	$\mu_{A,B,C}$
Block A	8	11	
Block B	10	12	
Block C	12	13	
$\mu_{1,2}$			11

no need to multiply since we’re working with totals already

$SSG = 6$
$SSB = 9$
$SST = 16$
$SSE = 1$

Two-Way ANOVA

$$F = \frac{Var.Between\,Groups}{Var.Within\,Groups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

	Group 1	Group 2	$\mu_{A,B,C}$
Block A	8	11	
Block B	10	12	
Block C	12	13	
$\mu_{1,2}$			11

So how do we calculate F?

Degrees of Freedom Groups is unchanged:

$$df_{groups} = n_{groups} - 1$$
$$= 2 - 1$$
$$= 1$$

$SSG = 6$
$SSB = 9$
$SST = 16$
$SSE = 1$
$df_{groups} = 1$

Two-Way ANOVA

$$F = \frac{Var.Between\ Groups}{Var.Within\ Groups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

	Group 1	Group 2	$\mu_{A,B,C}$
Block A	8	11	
Block B	10	12	
Block C	12	13	
$\mu_{1,2}$			11

So how do we calculate F?

Degrees of Freedom Error has changed:

$$\begin{aligned} df_{error} &= (n_{blocks} - 1)(n_{groups} - 1) \\ &= (3-1)(2-1) \\ &= 2 \end{aligned}$$

$SSG = 6$
$SSB = 9$
$SST = 16$
$SSE = 1$
$df_{groups} = 1$
$df_{error} = 2$

Two-Way ANOVA

$$F = \frac{Var.Between\ Groups}{Var.Within\ Groups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

	Group 1	Group 2	$\mu_{A,B,C}$
Block A	8	11	
Block B	10	12	
Block C	12	13	
$\mu_{1,2}$			11

So how do we calculate F?

$$F = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}} = \frac{\frac{6}{1}}{\frac{1}{2}} = \mathbf{12}$$

$SSG = 6$

$SSB = 9$

$SST = 16$

$SSE = 1$

$df_{groups} = 1$

$df_{error} = 2$

Two-Way ANOVA

$$F = \frac{Var.Between\ Groups}{Var.Within\ Groups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

$F_{groups} = \mathbf{12}$ feels like a high value.

	Group 1	Group 2	$\mu_{A,B,C}$
Block A	8	11	
Block B	10	12	
Block C	12	13	
$\mu_{1,2}$			**11**

However, in a two-way ANOVA, $F_{critical}$ is found for groups and blocks separately!

$SSG = 6$
$SSB = 9$
$SST = 16$
$SSE = 1$
$df_{groups} = 1$
$df_{error} = 2$

Two-Way ANOVA

$$F = \frac{Var.Between\ Groups}{Var.Within\ Groups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

$F_{groups} = \mathbf{12}$ feels like a high value.

	Group 1	Group 2	$\mu_{A,B,C}$
Block A	8	11	
Block B	10	12	
Block C	12	13	
$\mu_{1,2}$			11

For groups, with 1df in the numerator and 2 df in the denominator,

$$F_{critical} = \mathbf{18.5}$$

$SSG = 6$
$SSB = 9$
$SST = 16$
$SSE = 1$
$df_{groups} = 1$
$df_{error} = 2$

ANOVA Exercise #2

- Let’s go back to the invoice problem, and add a new independent variable
- Here each block represents an invoice amount
- The dependent variable is still days elapsed until payment

	2% disc	1% disc	no disc
$50	16	23	21
$100	14	21	16
$150	11	16	18
$200	10	15	14
$250	9	10	11

ANOVA Exercise #2

1. Calculate the group means, the block means, and the total mean

	2% disc	1% disc	no disc	μ_{block}
$50	16	23	21	
$100	14	21	16	
$150	11	16	18	
$200	10	15	14	
$250	9	10	11	
μ_{col}				15

ANOVA Exercise #2

2. Sum of Squares Groups

$$(\mu_2 - \mu_{TOT})^2 = (12 - 15)^2 = 9$$

$$(\mu_1 - \mu_{TOT})^2 = (17 - 15)^2 = 4$$

$$(\mu_0 - \mu_{TOT})^2 = (16 - 15)^2 = 1$$

$$14$$

Multiply by the number of items in each group:

$$14 \times 5 = 70$$

	2% disc	1% disc	no disc	μ_{block}
$50	16	23	21	
$100	14	21	16	
$150	11	16	18	
$200	10	15	14	
$250	9	10	11	
μ_{col}				15

$$SSG = 70$$

ANOVA Exercise #2

3. Degrees of Freedom Groups

$$df_{groups} = n_{groups} - 1$$
$$= 3 - 1$$
$$= 2$$

	2% disc	1% disc	no disc	μ_{block}
$50	16	23	21	
$100	14	21	16	
$150	11	16	18	
$200	10	15	14	
$250	9	10	11	
μ_{col}				15

$SSG = 70$ $df_{groups} = 2$

ANOVA Exercise #2

4. Sum of Squares Blocks

$$(\mu_{50} - \mu_{TOT})^2 = (20 - 15)^2 = 25$$
$$(\mu_{100} - \mu_{TOT})^2 = (17 - 15)^2 = 4$$
$$(\mu_{200} - \mu_{TOT})^2 = (15 - 15)^2 = 0$$
$$(\mu_{200} - \mu_{TOT})^2 = (13 - 15)^2 = 4$$
$$(\mu_{250} - \mu_{TOT})^2 = (10 - 15)^2 = 25$$

58

$$58 \times 3 = 174$$

	2% disc	1% disc	no disc	μ_{block}
$50	16	23	21	
$100	14	21	16	
$150	11	16	18	
$200	10	15	14	
$250	9	10	11	
μ_{col}				15

$SSG = 70$
$SSB = 174$

$df_{groups} = 2$

ANOVA Exercise #2

5. Sum of Squares Total

$(x_2-\mu_{tot})^2$	$(x_1-\mu_{tot})^2$	$(x_0-\mu_{tot})^2$
1	64	36
1	36	1
16	1	9
25	0	1
36	25	16
79	**126**	**63**

TOTAL	

	2% disc	1% disc	no disc	μ_{block}
$50	16	23	21	
$100	14	21	16	
$150	11	16	18	
$200	10	15	14	
$250	9	10	11	
μ_{col}				**15**

$$SSG = 70$$
$$SSB = 174$$
$$SST = 268$$

$$df_{groups} = 2$$

ANOVA Exercise #2

6. Sum of Squares Error

$$SSE = SST - SSG - SSB$$
$$= 268 - 70 - 174 = 24$$

	2% disc	1% disc	no disc	μ_{block}
\$50	16	23	21	
\$100	14	21	16	
\$150	11	16	18	
\$200	10	15	14	
\$250	9	10	11	
μ_{col}				15

$SSG = 70$
$SSB = 174$
$SST = 268$
$SSE = 24$

$df_{groups} = 2$

ANOVA Exercise #2

7. Degrees of Freedom Error

$$df_{error} = (n_{blocks} - 1)(n_{groups} - 1)$$
$$= (5 - 1)(3 - 1)$$
$$= 8$$

	2% disc	1% disc	no disc	μ_{block}
$50	16	23	21	
$100	14	21	16	
$150	11	16	18	
$200	10	15	14	
$250	9	10	11	
μ_{col}				15

$SSG = 70$

$SSB = 174$

$SST = 268$

$SSE = 24$

$df_{groups} = 2$

$df_{error} = 8$

ANOVA Exercise #2

8. Calculate F

$$F = \frac{\dfrac{SSG}{df_{groups}}}{\dfrac{SSE}{df_{error}}} = \frac{\dfrac{70}{2}}{\dfrac{24}{8}} = \frac{35}{3} = \mathbf{11.67}$$

	2% disc	1% disc	no disc	μ_{block}
$50	16	23	21	
$100	14	21	16	
$150	11	16	18	
$200	10	15	14	
$250	9	10	11	
μ_{col}				15

$SSG = 70$

$SSB = 174$

$SST = 268$

$SSE = 24$

$df_{groups} = 2$

$df_{error} = 8$

F = 11.67

ANOVA Exercise #2

9. Find $F_{critical}$

$$\alpha = 0.05$$

$$df_{numerator} = 2$$

$$df_{denominator} = 8$$

$$\boldsymbol{F_{critical} = 4.46}$$

	2% disc	1% disc	no disc	μ_{block}
$50	16	23	21	
$100	14	21	16	
$150	11	16	18	
$200	10	15	14	
$250	9	10	11	
μ_{col}				15

$SSG = 70$

$SSB = 174$

$SST = 268$

$SSE = 24$

$df_{grouupps} = 2$

$df_{error} = 8$

F = 11.67

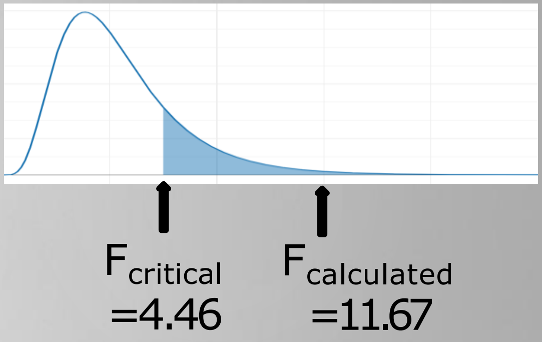
$F_{critical} = 4.46$

ANOVA Exercise #2

Since F falls to the right of $F_{critical}$

$$4.46 < 11.67$$

we reject the null hypothesis!



$SSG = 70$

$SSB = 174$

$SST = 268$

$SSE = 24$

$df_{groups} = 2$

$df_{error} = 8$

$\mathrm{F} = 11.67$

$\mathrm{F}_{\mathrm{critical}} = 4.46$