Intermediate Statistics – Analysis of Variance

Analysis of Variance (ANOVA)
(General Linear Model)

Analysis of Variance (ANOVA)

Introduction

- **ANalysis Of VAriance (ANOVA)** is a statistical technique for testing if **3(+)** population **means are all equal**.
  - **one-way ANOVA** for comparing **3(+) groups on 1 variable**.
    For example, do all children from school A, B and C have equal mean IQ scores?
  - **two-way ANOVA** for comparing **3(+) groups on 2 variables**.
    For example, a company wants to compare worker productivity based on two independent variables, say salary and skill set. It is utilized to observe the interaction between the two factors. It tests the effect of two factors at the same time.

- **Basic Framework of ANOVA**
  - Want to study the **effect** of one or more **qualitative** variables on a **quantitative** outcome variable.
  - **Qualitative** variables are referred to as **factors**.
  - **Characteristics** that differentiates factors are referred to as **levels**.

Analysis of Variance (ANOVA)

One way Analysis of variance

- ❑ An **ANOVA** test is a way to find out if a survey or experiment results are **significant**. In other words, they help you to figure out **if you need to reject the null hypothesis** and consequently **accept the alternate hypothesis**.

- ❑ Basically, you're **testing groups** to see if there's a **difference between them**.

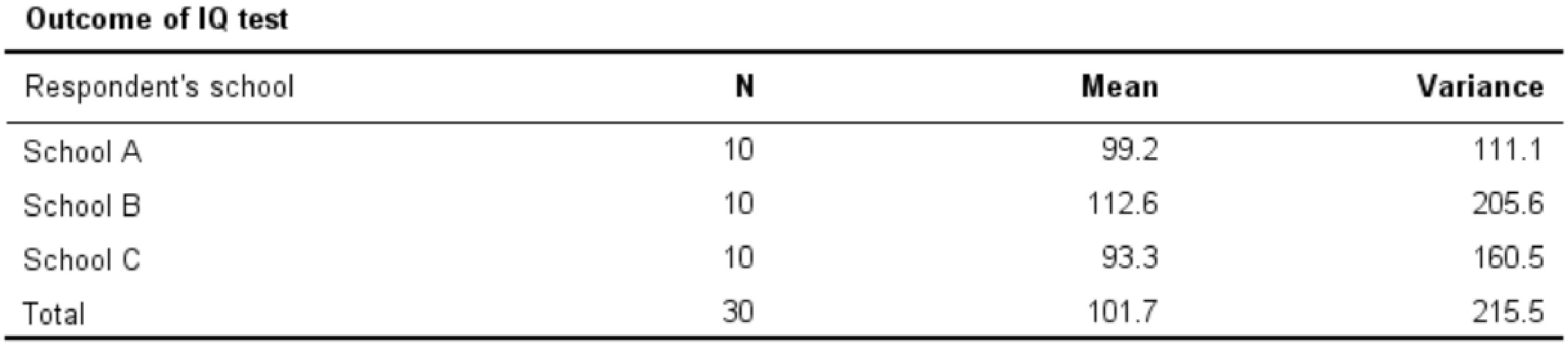
Examples of when you might want to test different groups:

- ➢ A group of psychiatric patients are trying **three different therapies**: counseling, medication and biofeedback. You want to see if one therapy is better than the others.
- ➢ A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
- ➢ Students from different colleges take the same exam. You want to see if one college outperforms the other.

Analysis of Variance (ANOVA)

One way Analysis of variance – Simple Example

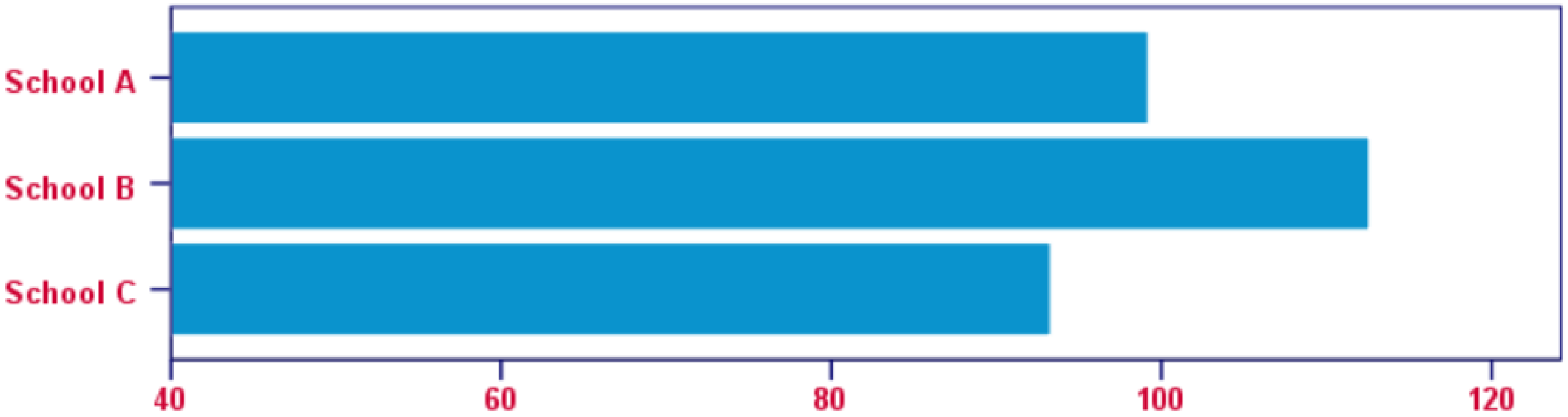
A scientist wants to know if all children from schools A, B and C have **equal mean** IQ scores. Each school has 1,000 children. It takes too much time and money to test all 3,000 children. So, a simple random sample of $N = 10$ children from each school is tested.



Outcome of IQ test

Respondent's school	N	Mean	Variance
School A	10	99.2	111.1
School B	10	112.6	205.6
School C	10	93.3	160.5
Total	30	101.7	215.5

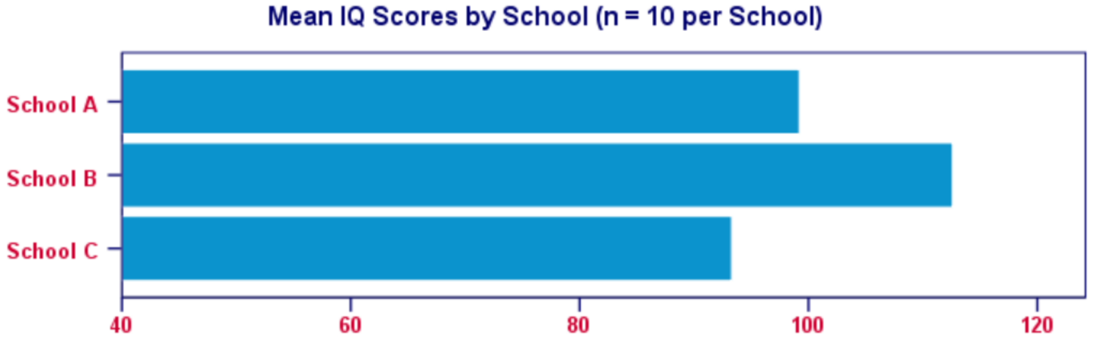
Mean IQ Scores by School (n = 10 per School)



Analysis of Variance (ANOVA)

One way Analysis of variance

Clearly, our sample from school B has the highest mean IQ ~113 points. The lowest mean IQ ~93 points is seen for school C.



Now, here's the problem: our mean IQ scores are only based on tiny samples of 10 children per school.

So couldn't it be that **all** 1,000 children per school have the same **mean** IQ?

Perhaps we just happened to sample the smartest children from school B and the dumbest children from school C? Is that realistic?

We'll try and show that this statement -our **null hypothesis**- is not credible given our data.

Analysis of Variance (ANOVA)

ANOVA - Null Hypothesis

- The null hypothesis for (any) **ANOVA** is that **all population means are exactly equal**.

- If this holds, then our **sample means** will probably **differ** a bit.

- After all, **samples** **always differ a bit from the** **populations** they represent.

- However, the sample **means probably** shouldn't differ **too much**.

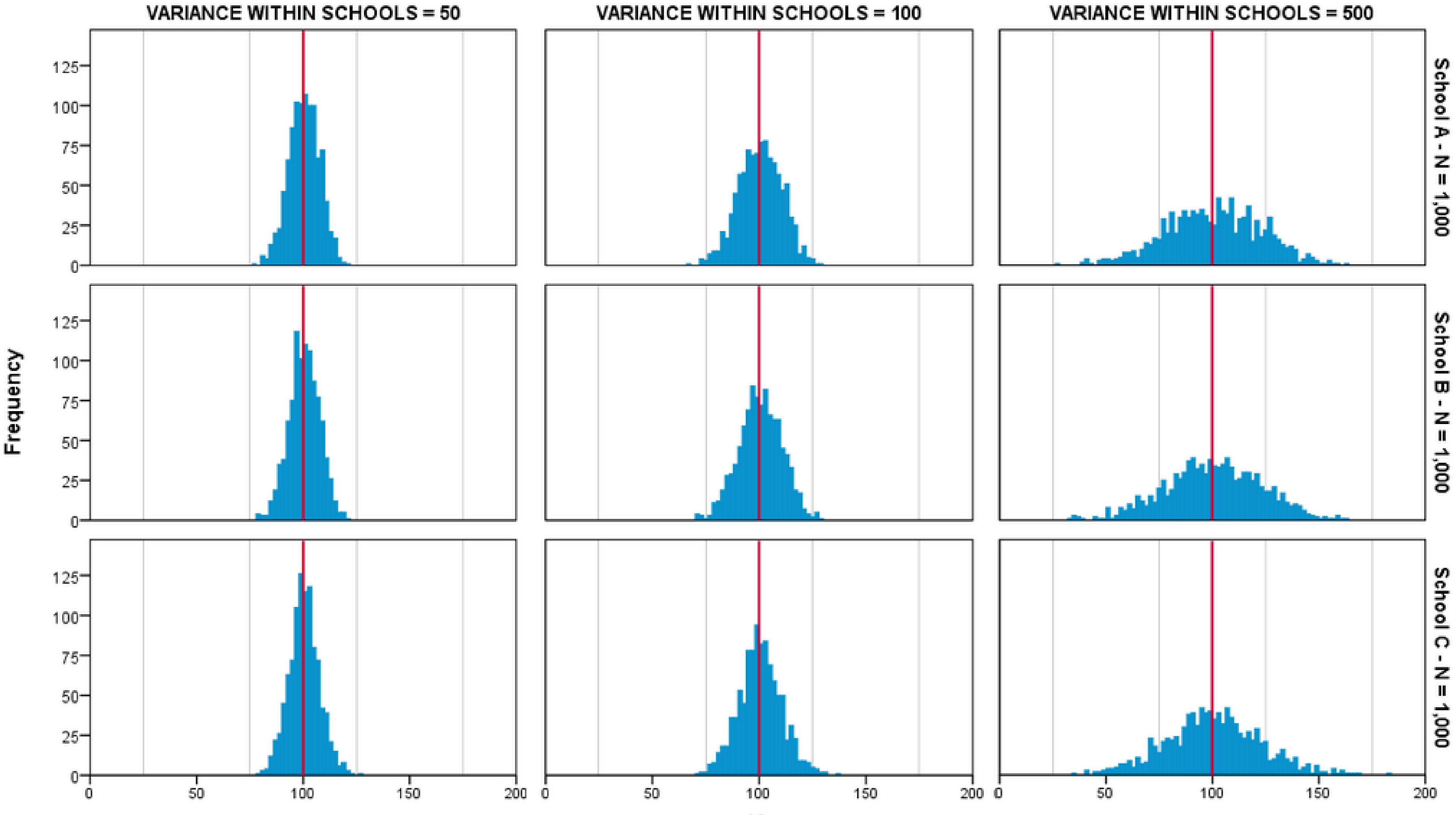
- Such an outcome would be unlikely under our **null hypothesis** of **equal population means**.

- So if we **do** find this, we'll **probably no longer believe that our population means were really equal**.

Analysis of Variance (ANOVA)

ANOVA - Null Hypothesis

3 theoretical scenarios



Analysis of Variance (ANOVA)

One way Analysis of variance

- The 3 leftmost histograms show population distributions for IQ in schools A, B and C. Their **narrowness** indicates a **small variance within each school**.
- If we would sample $n = 10$ students from each school, should we expect very different sample means? Probably not.
- **Why?** Well, due to the small variance within each school, the sample means will be close to the (equal) population means. These narrow histograms don't leave a lot of room for their sample means to fluctuate and -hence- differ.

- The 3 rightmost histograms show the **opposite scenario**: the histograms are wide, indicating a **large variance within each school**. If we'd sample $n = 10$ students from each school, the means in these samples may easily differ quite a lot.

- In short, **larger variances within schools probably result in a larger variance between sample means per school**.

Analysis of Variance (ANOVA)

One-Way ANOVA

A **one-way ANOVA** is a type of statistical test that **compares the variance in the group means** within a sample whilst considering only **one independent variable or factor**.

What are the hypotheses of a One-Way ANOVA?

- In a one-way ANOVA there are two possible hypotheses.
- The **null hypothesis (H_0)** is that there is **no difference** between the groups and equality between means.
- The **alternative hypothesis (H_1)** is that **there is a difference** between the means and groups.

What is the rationale of a One-Way ANOVA?

- Basic idea → partition total variation of the data into two sources:
  1. Variation within levels (groups)
  2. Variation between levels (groups)

What are the assumptions of a One-Way ANOVA?

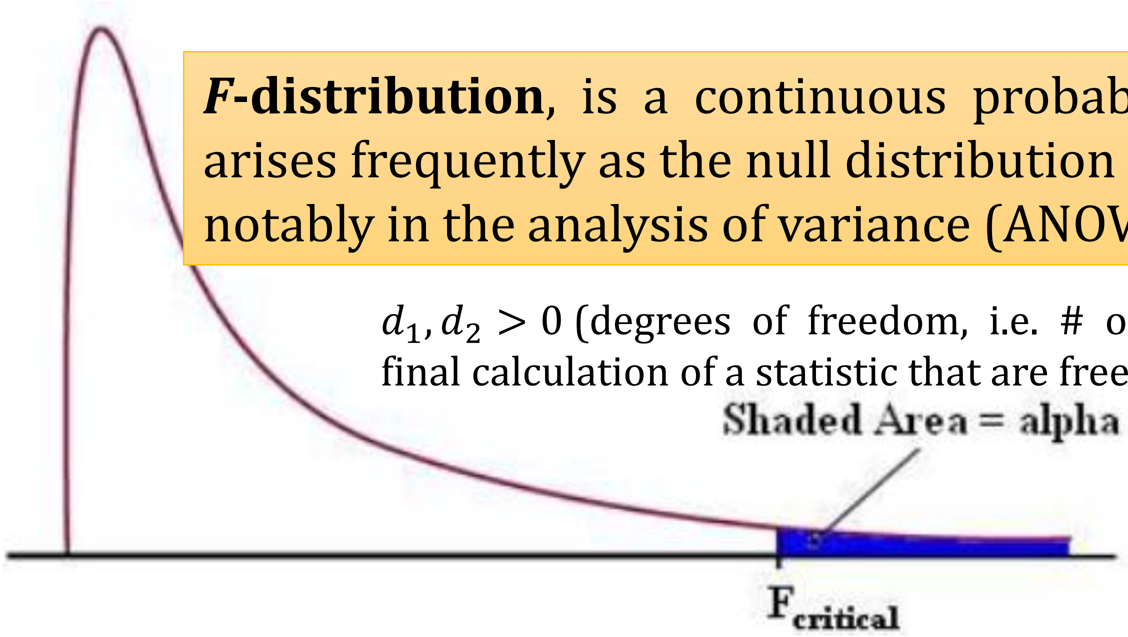
- Normality – That each sample is taken from a normally distributed population
- Sample independence – that each sample has been drawn independently of the other samples
- Variance Equality – That the variance of data in the different groups should be the same
- The dependent variable should be continuous

Analysis of Variance (ANOVA)

F-test/ F-statistic / F-distribution: Definition

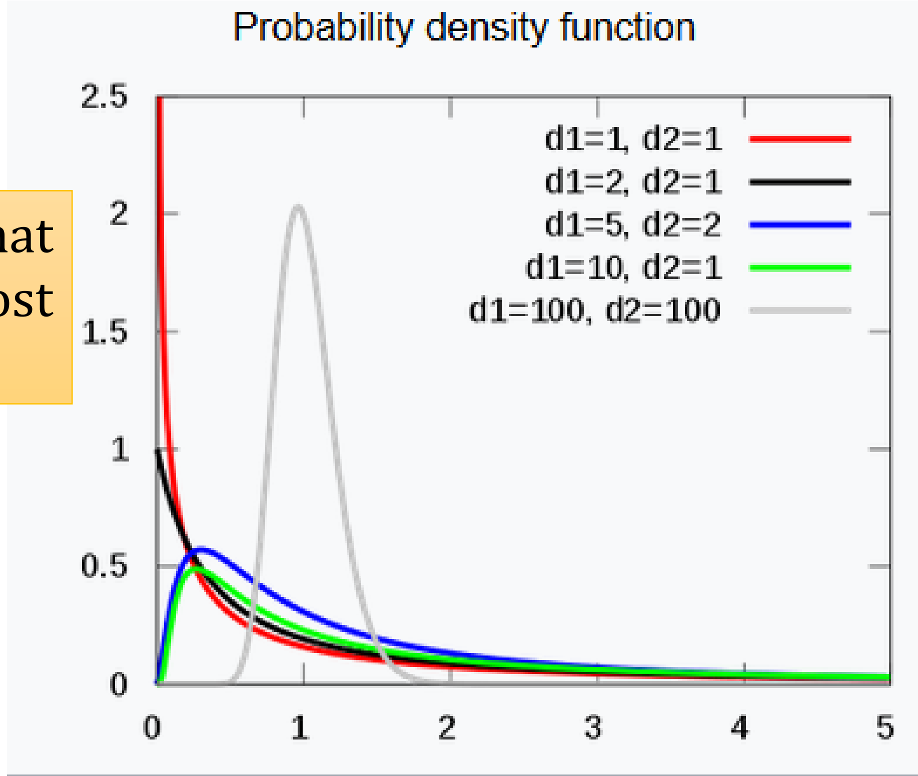
An F −statistic is a value you get when you run an **analysis of variance (ANOVA)** test **or** a **regression analysis** to find out if the means between two populations are significantly different. It's similar to a t −statistic from a t −test;

- ***t* −test** will tell you if a ***single*** variable is statistically significant and
- ***F* −test** will tell you if a ***group*** of variables are **<u>jointly</u>** significant.



***F*-distribution**, is a continuous probability distribution that arises frequently as the null distribution of a test statistic, most notably in the analysis of variance (ANOVA), e.g. *F*-test.

$d_1, d_2 > 0$ (degrees of freedom, i.e. # of values in the final calculation of a statistic that are free to vary)



Analysis of Variance (ANOVA)

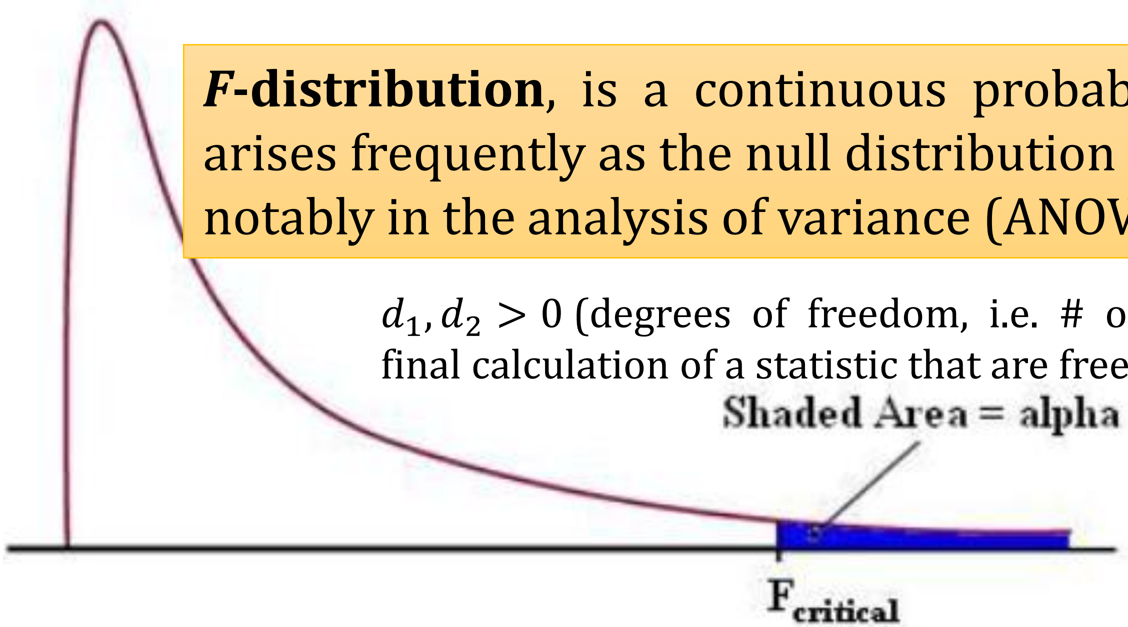
F-test/ F-statistic / F-distribution: Definition

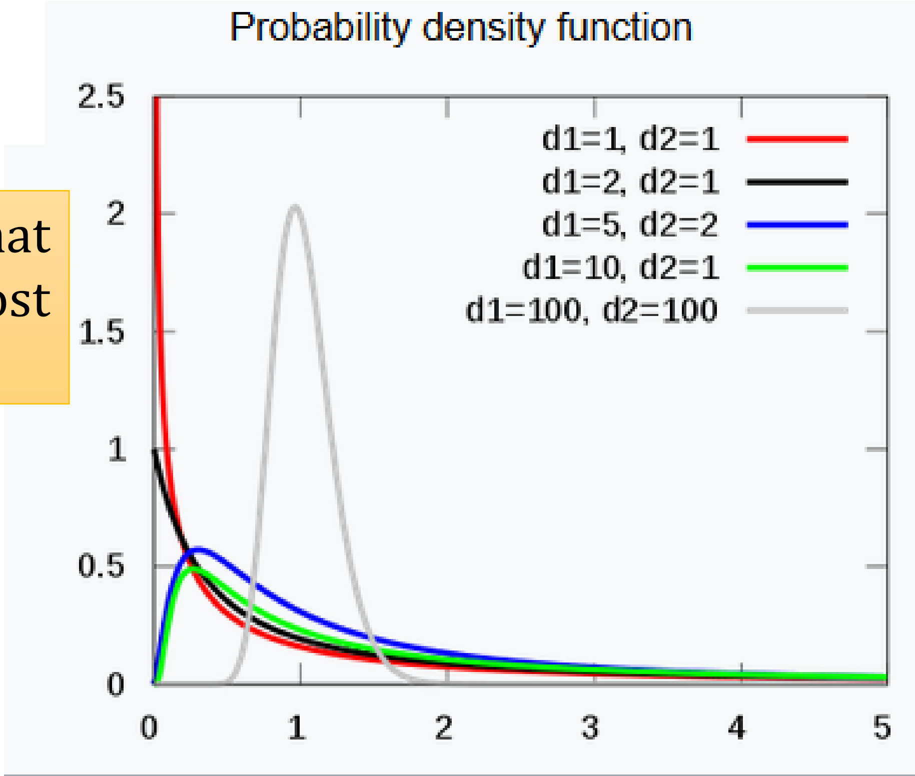
$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1}\, d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}}}{x\, \mathrm{B}\!\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

$$= \frac{1}{\mathrm{B}\!\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}\, x\right)^{-\frac{d_1+d_2}{2}}$$

***F*-distribution**, is a continuous probability distribution that arises frequently as the null distribution of a test statistic, most notably in the analysis of variance (ANOVA), e.g. *F*-test.

$d_1, d_2 > 0$ (degrees of freedom, i.e. # of values in the final calculation of a statistic that are free to vary)





Analysis of Variance (ANOVA)

F-test/ F-statistic / F-distribution: Definition

The F −value in one way **ANOVA** is a tool to help you answer questions like: ***“Is the variance between the means of two or more populations significantly different?***”

The $F -$ value in the **ANOVA** test also determines the p −value; The $p -$ value is the probability of getting a result at least as extreme as the one that was actually observed, given that the null hypothesis is true.

The $p -$ value is a probability, while the f −ratio is a test statistic, calculated as:

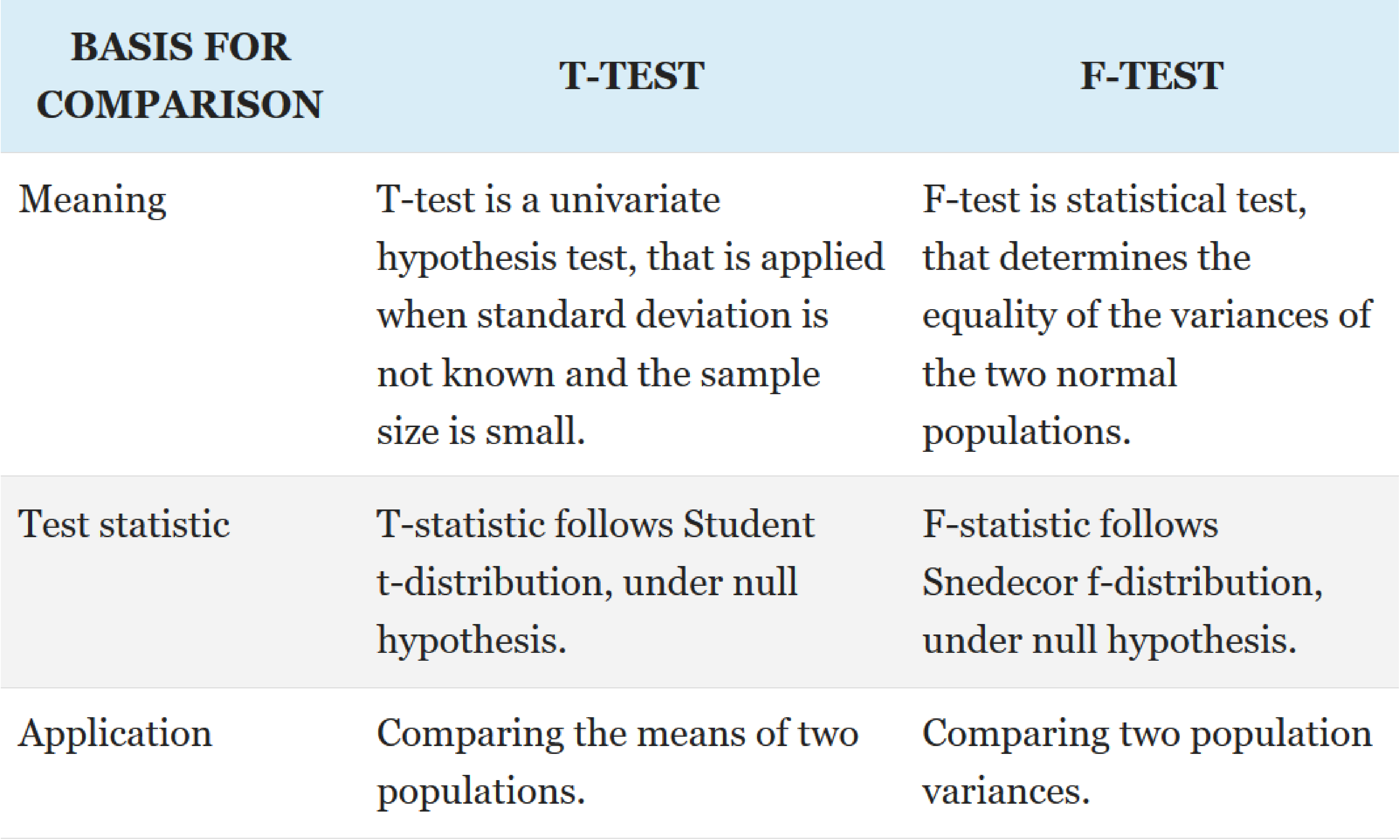
$$F\ value = \frac{\text{variance of the group means } (Mean\ Square\ Between)}{\text{mean of the within group variances } (Mean\ Squared\ Error)}$$

When do I reject the null hypothesis?

Reject the null when your p −value is smaller than your alpha level. You should not reject the null if your critical f value is smaller than your $F -$ value, unless you also have a small p −value.

Analysis of Variance (ANOVA)

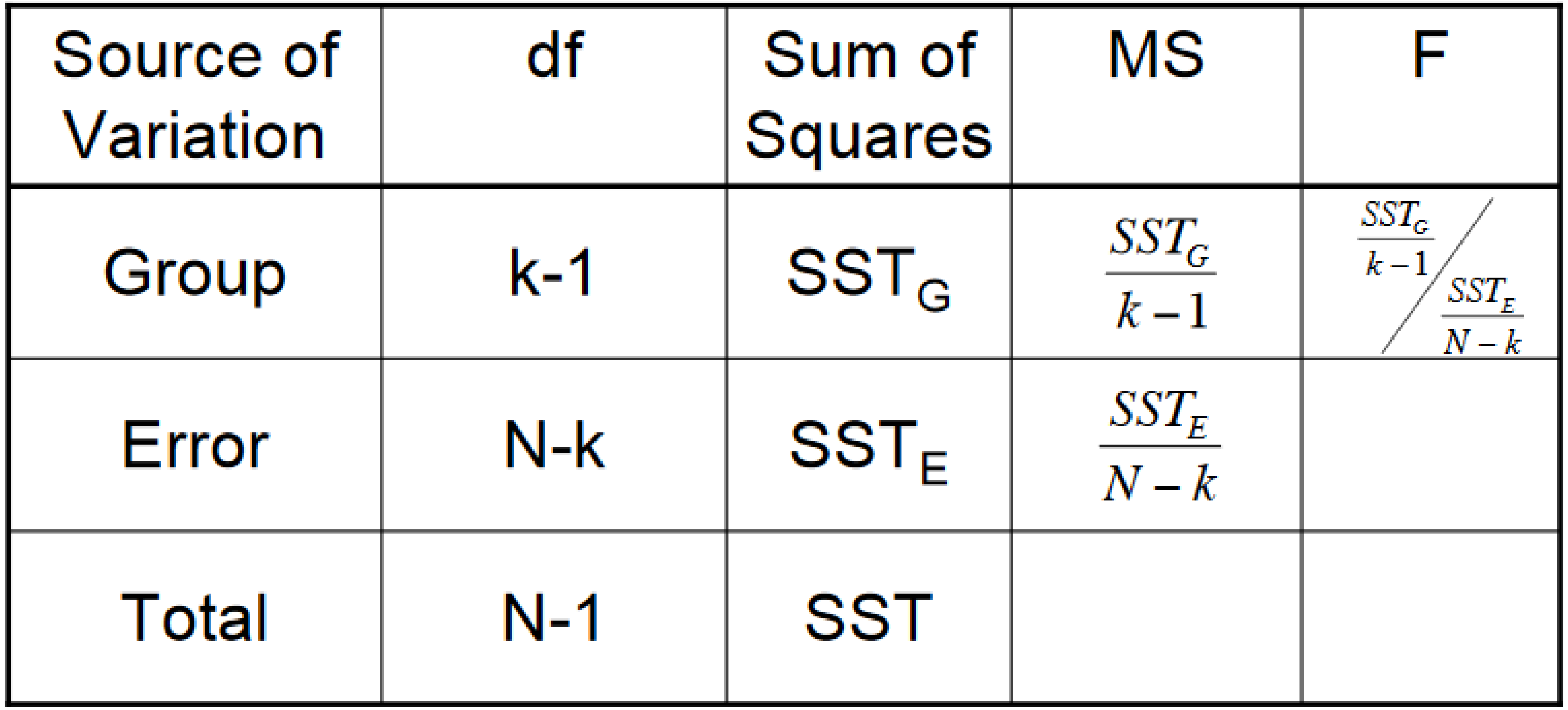
Difference Between T-test and F-test



BASIS FOR COMPARISON	T-TEST	F-TEST
Meaning	T-test is a univariate hypothesis test, that is applied when standard deviation is not known and the sample size is small.	F-test is statistical test, that determines the equality of the variances of the two normal populations.
Test statistic	T-statistic follows Student t-distribution, under null hypothesis.	F-statistic follows Snedecor f-distribution, under null hypothesis.
Application	Comparing the means of two populations.	Comparing two population variances.

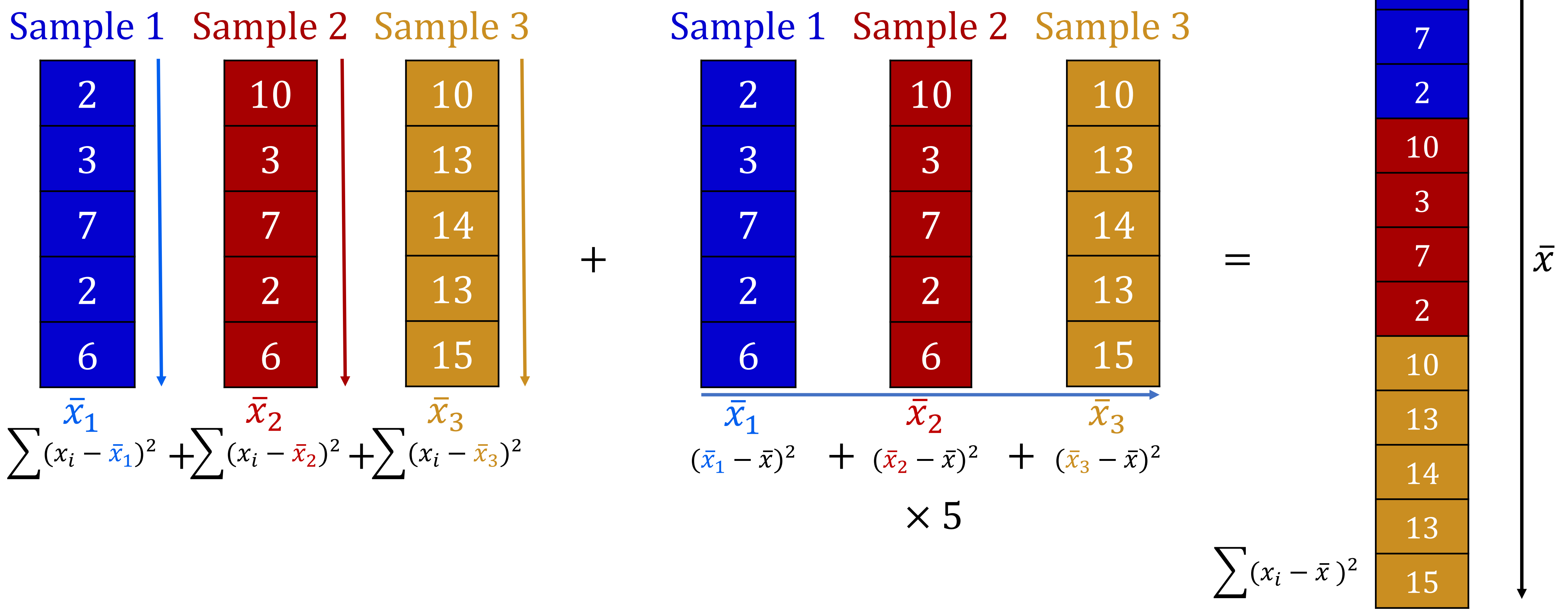
Analysis of Variance (ANOVA)

ANOVA Table



Source of Variation	df	Sum of Squares	MS	F
Group	k-1	SST_G	$\frac{SST_G}{k-1}$	$\frac{SST_G}{k-1} \Big/ \frac{SST_E}{N-k}$
Error	N-k	SST_E	$\frac{SST_E}{N-k}$	
Total	N-1	SST		

Intermediate Statistics – ANOVA- Example 1a



Sample 1	Sample 2	Sample 3
2	10	10
3	3	13
7	7	14
2	2	13
6	6	15
$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$

$$\sum(x_i-\bar{x}_1)^2+\sum(x_i-\bar{x}_2)^2+\sum(x_i-\bar{x}_3)^2$$

$+$

Sample 1	Sample 2	Sample 3
2	10	10
3	3	13
7	7	14
2	2	13
6	6	15
$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$

$$(\bar{x}_1-\bar{x})^2 + (\bar{x}_2-\bar{x})^2 + (\bar{x}_3-\bar{x})^2$$

× 5

$=$

2
3
7
2
10
3
7
2
10
13
14
13
15

$\bar{x}$

$$\sum(x_i-\bar{x})^2$$

Sum of **S**quares
Within samples **(SSW)** +
54

Sum of **S**quares
Between samples **(SSB)** =
203.3

(SST) Total
Sum of **S**quares
257.3

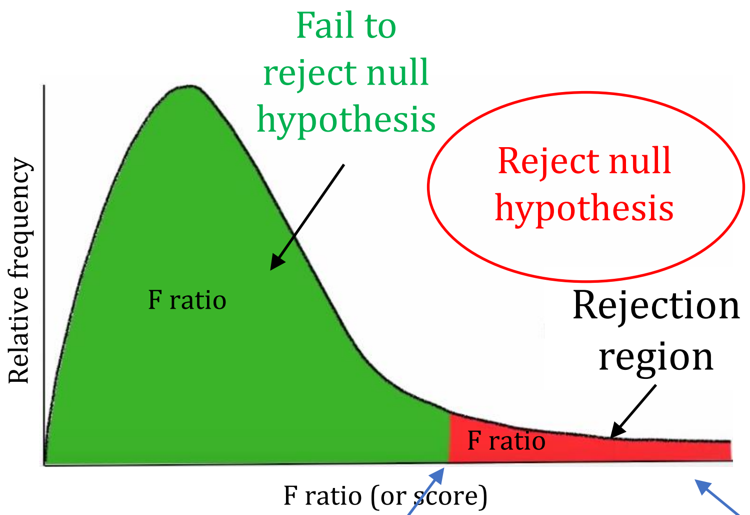
Intermediate Statistics – ANOVA- Example 1b

$$F = \frac{\text{between} - \text{groups variance}}{\text{within} - \text{group variance}}$$

$$\frac{SSB}{\mathbf{d.o.f}} = \frac{SSB}{(\#\textbf{ of samples}) - \mathbf{1}} = \frac{203.3}{\mathbf{2}} = 101.667$$

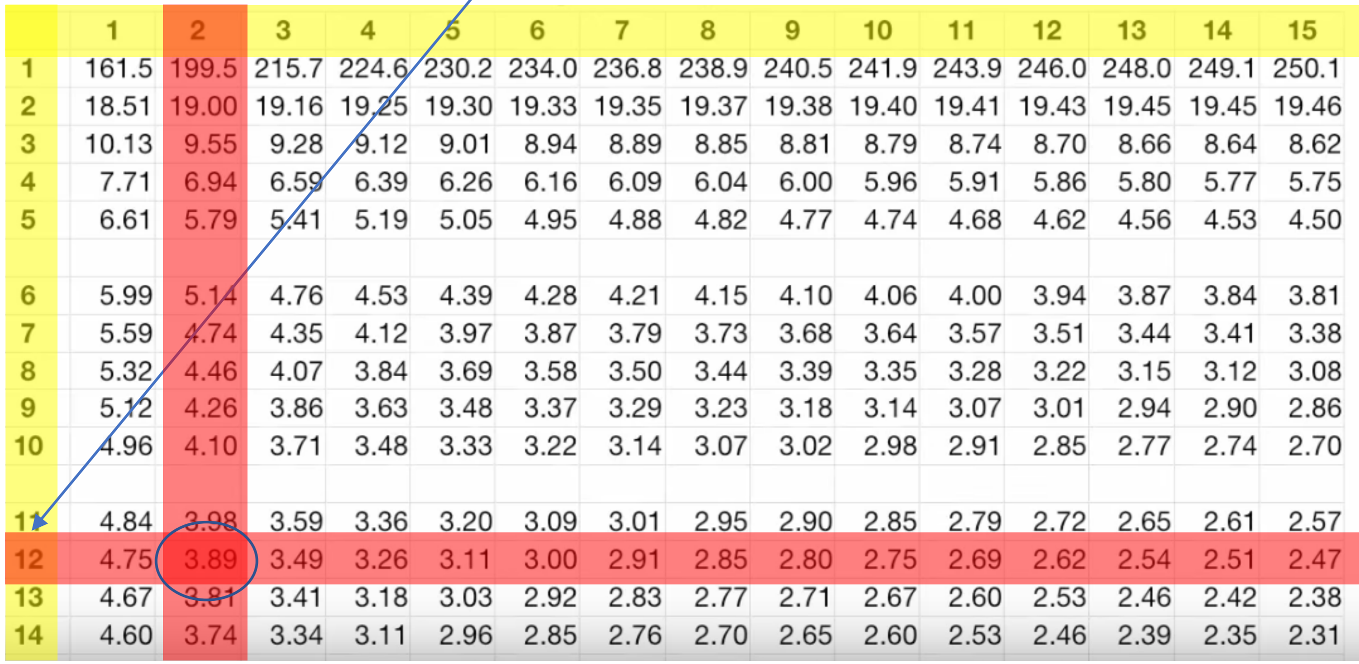
$$\frac{SSW}{\mathbf{d.o.f}} = \frac{SSW}{(\#\textbf{ of observations}) - (\#\textbf{ of samples})} = \frac{54}{\mathbf{15-3}} = \frac{54}{\mathbf{12}} = 4.5$$

$$\mathrm{F} = \mathrm{F}(2, 12) = \frac{101.667}{4.5} = 22.59, p > 0.05$$



Critical value = 3.89

F = 22.59

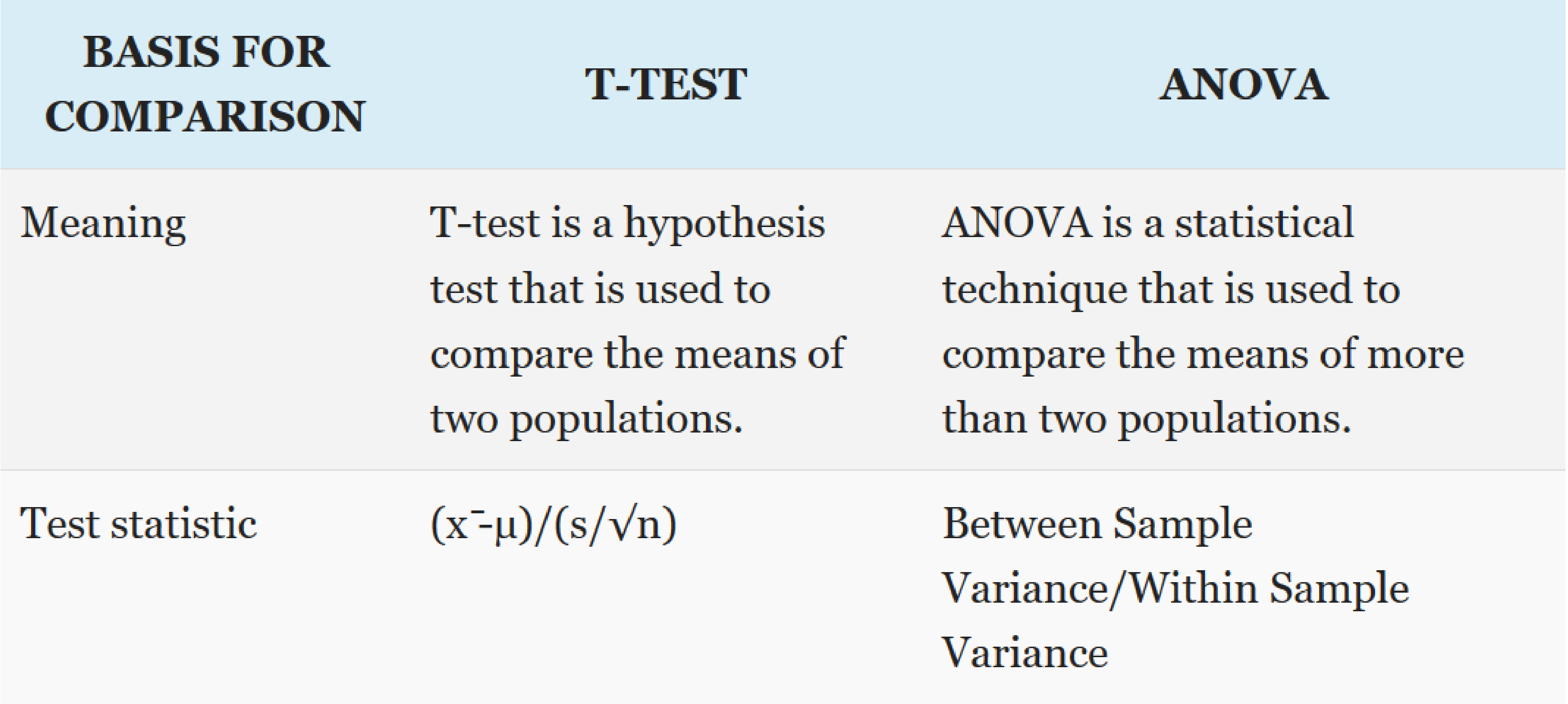


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	161.5	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	246.0	248.0	249.1	250.1
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31

Analysis of Variance (ANOVA)

ANOVA vs. t-test

Student’s t-test will tell you if there is a significant variation between groups. A t-test compares means while the **ANOVA** compares variances between populations.



BASIS FOR COMPARISON	T-TEST	ANOVA
Meaning	T-test is a hypothesis test that is used to compare the means of two populations.	ANOVA is a statistical technique that is used to compare the means of more than two populations.
Test statistic	$(\bar{x}-\mu)/(s/\sqrt{n})$	Between Sample Variance/Within Sample Variance

You *could* technically perform a series of t-tests on your data. However, as the groups grow in number, you may end up with a *lot* of pair comparisons that you need to run. **ANOVA** will give you a single number (the F-statistic) and one p –value to help you support or reject the null hypothesis.

https://keydifferences.com/difference-between-t-test-and-anova.html

Analysis of Variance (ANOVA)

Two-Way ANOVA

A two-way ANOVA is, like a one-way ANOVA, a hypothesis-based test. However, in the two-way ANOVA each sample is defined in two ways, and resultingly put into two categorical groups.

What are the hypotheses of a Two-Way ANOVA?

Because the two-way ANOVA consider the effect of two categorical factors, and the effect of the categorical factors on each other, there are three pairs of null or alternative hypotheses for the two-way ANOVA.

What are the assumptions of a Two-Way ANOVA?

- Your dependent variable should be continuous
- The two independent variables should be in categorical, independent groups.
- Sample independence – that each sample has been drawn independently of the other samples
- Variance Equality – That the variance of data in the different groups should be the same
- Normality – That each sample is taken from a normally distributed population

Analysis of Variance (ANOVA)

One vs Two-Way ANOVA

	One-Way ANOVA	**Two-Way ANOVA**
Definition	A test that allows one to make comparisons between the means of three or more groups of data.	A test that allows one to make comparisons between the means of three or more groups of data, where two independent variables are considered.
Number of Independent Variables	One.	Two.
What is Being Compared?	The means of three or more groups of an independent variable on a dependent variable.	The effect of multiple groups of two independent variables on a dependent variable and on each other.
Number of Groups of Samples	Three or more.	Each variable should have multiple samples.

Analysis of Variance (ANOVA)

Simple Linear Regression & ANOVA

- ❑ A not-so-obvious fact is that simple (ordinary least square) **linear regression** under standard conditions is **a special case of ANOVA**. The variation of the y_i is conventionally measured in terms of the deviations:

$$y_i - \bar{y}$$

- ❑ The measure of total variation, denoted by SST, is the sum of the squared deviations:

$$SST = \sum (y_i - \bar{y})^2$$

- ❑ If SST = 0, all observations are the same (no variability).

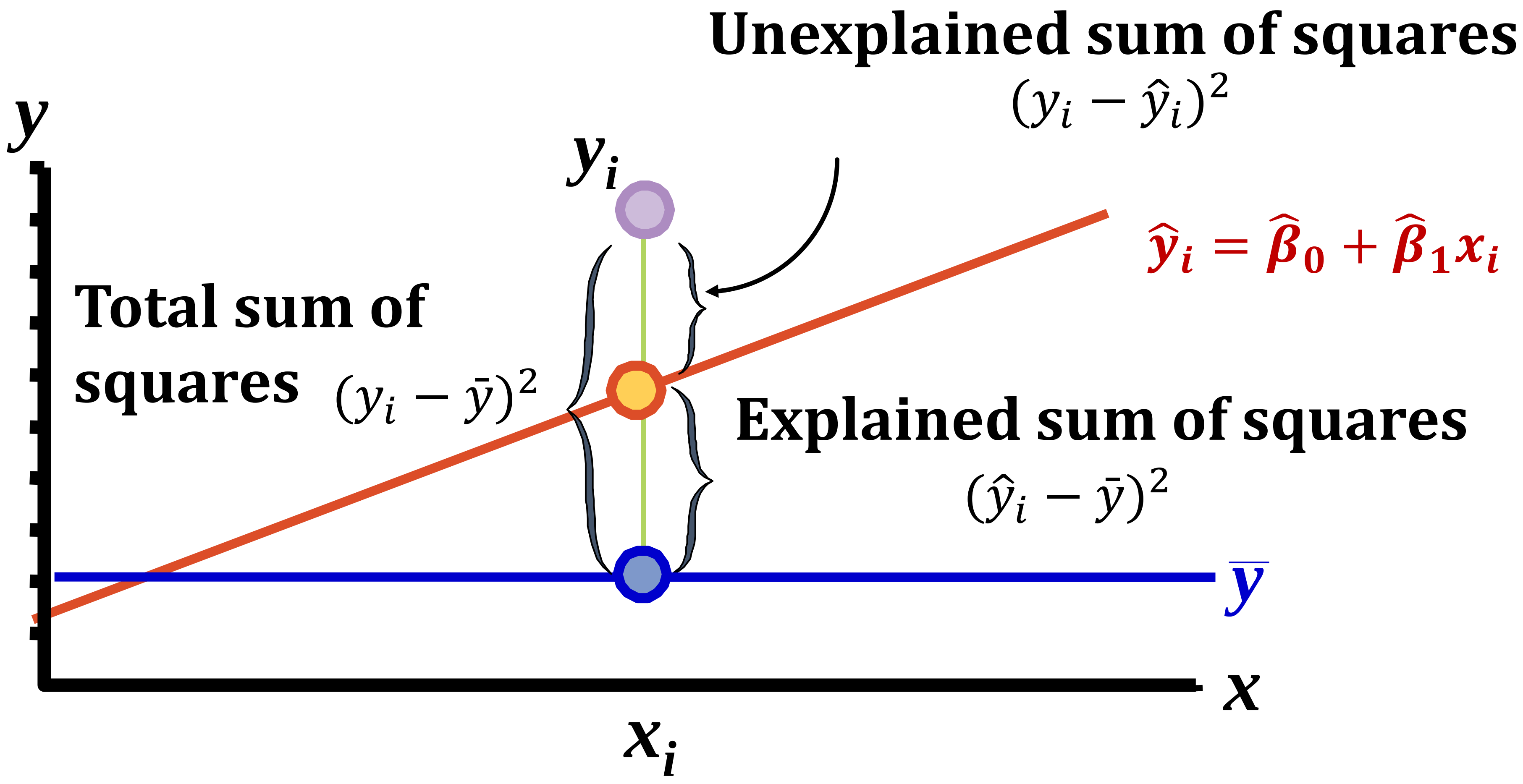
- ❑ The greater is SST, the greater is the variation among the y values.

- ❑ When we use the regression model, the measure of variation is that of the y observations variability around the fitted line:

$$y_i - \hat{y}_i$$

PIB2: Data Analysis – Simple Linear Regression

Linear Regression Assumptions: Variation Measures



Unexplained sum of squares

$(y_i - \hat{y}_i)^2$

Total sum of squares $(y_i - \bar{y})^2$

Explained sum of squares

$(\hat{y}_i - \bar{y})^2$

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

Analysis of Variance (ANOVA)

Simple Linear Regression & ANOVA

$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$$

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \quad (1)$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x \quad (2)$$

❑ Substituting equation (2) in equation (1), … and skipping a bit of algebra …, we get:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

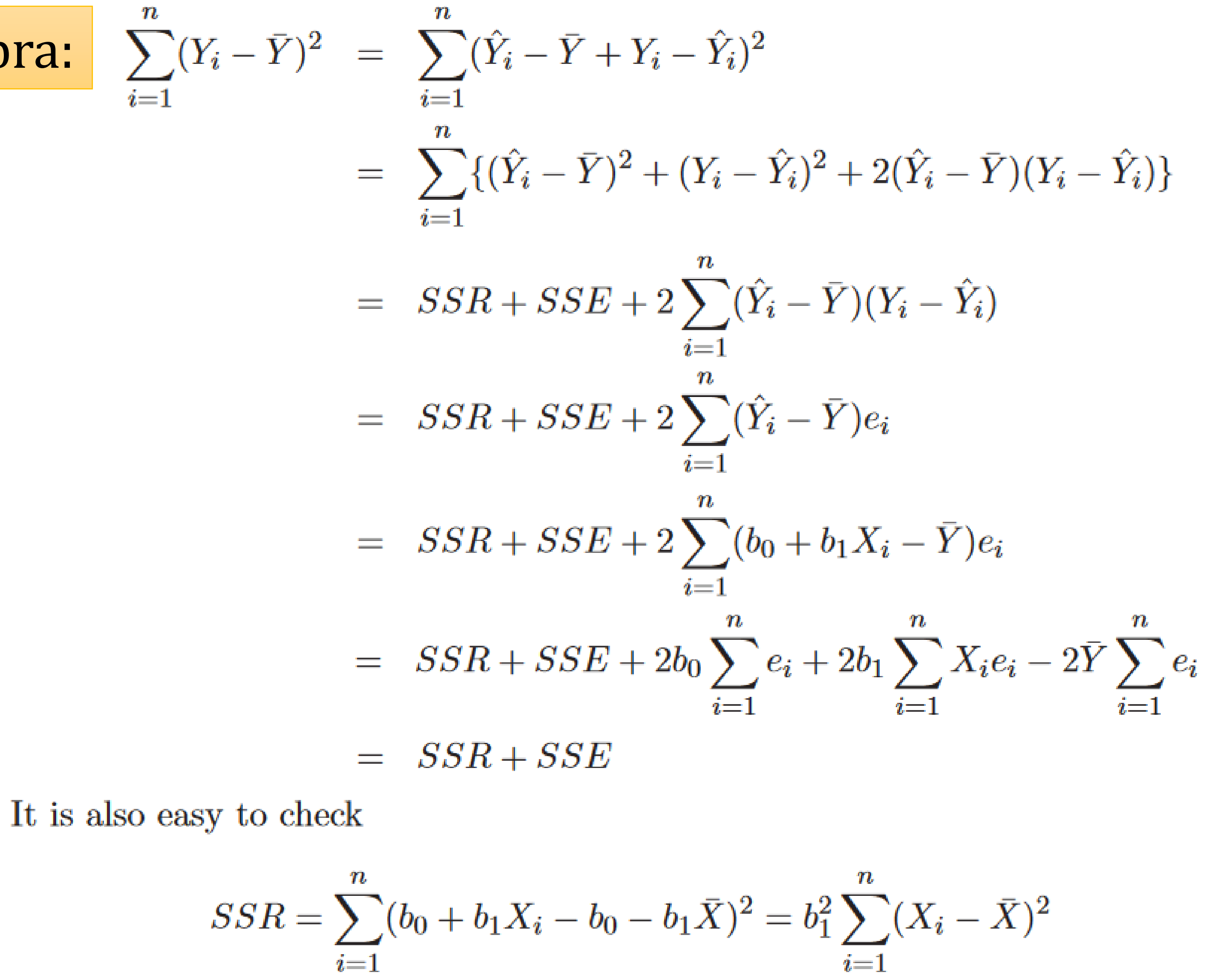
$$\text{SST} \quad = \quad \text{SSR} \quad + \quad \text{SSE}$$

→ We obtain the mean squares by dividing these numbers by the corresponding degrees of freedom. This is the relationship between **simple linear regression** **and** **ANOVA**.

Analysis of Variance (ANOVA)

Simple Linear Regression & ANOVA (… derivation …)

Here is the previously skipped algebra:



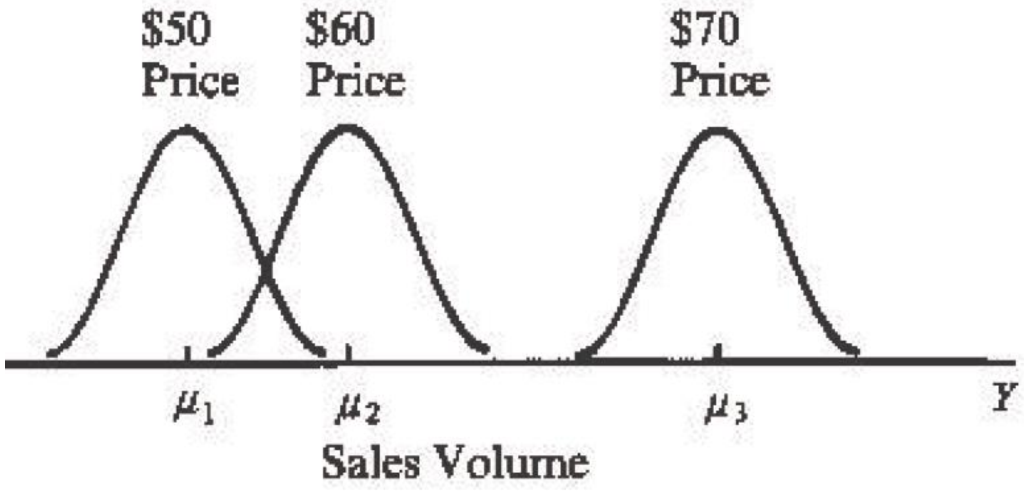
$$
\begin{aligned}
\sum_{i=1}^{n}(Y_i-\bar{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y}+Y_i-\hat{Y}_i)^2 \\
&= \sum_{i=1}^{n}\{(\hat{Y}_i-\bar{Y})^2+(Y_i-\hat{Y}_i)^2+2(\hat{Y}_i-\bar{Y})(Y_i-\hat{Y}_i)\} \\
&= SSR+SSE+2\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})(Y_i-\hat{Y}_i) \\
&= SSR+SSE+2\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})e_i \\
&= SSR+SSE+2\sum_{i=1}^{n}(b_0+b_1X_i-\bar{Y})e_i \\
&= SSR+SSE+2b_0\sum_{i=1}^{n}e_i+2b_1\sum_{i=1}^{n}X_ie_i-2\bar{Y}\sum_{i=1}^{n}e_i \\
&= SSR+SSE
\end{aligned}
$$

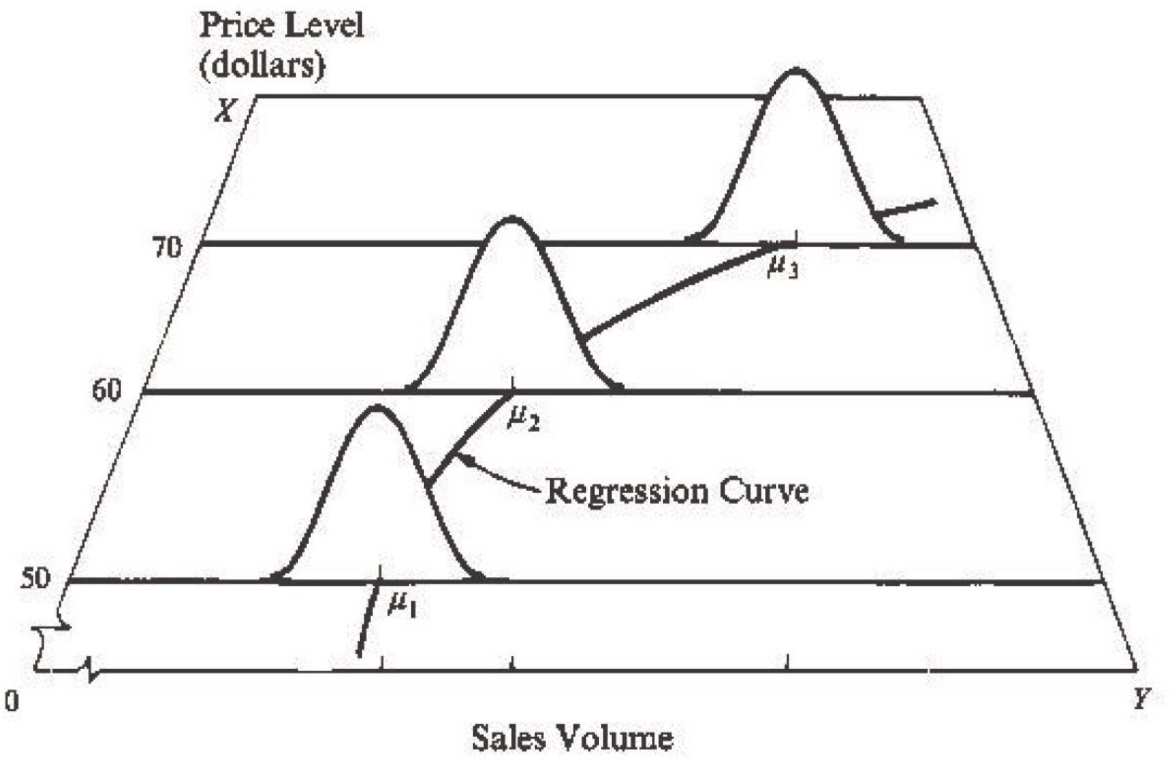
It is also easy to check

$$SSR=\sum_{i=1}^{n}(b_0+b_1X_i-b_0-b_1\bar{X})^2=b_1^2\sum_{i=1}^{n}(X_i-\bar{X})^2$$

Analysis of Variance (ANOVA)

ANOVA | **Regression analysis**





KEY DIFFERENCE:

In **ANOVA** → NO assumption is made about the manner in which Price and Sales Volume are related

Intermediate Statistics – ANOVA

Hands-on Activity – R session

Practice Exercises

ANOVA – **HWA** – Group Exercise Sheet: a,b

`HWA_a_ANOVA.Rmd`
`HWA_b_ANOVA.Rmd`

Correlation / Linear Regression /ANOVA

Blog article: **Introduction to Linear Regression and Polynomial Regression**
https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb

Supporting material: Video Session

- **P Values, clearly explained:** https://www.youtube.com/watch?v=5Z9OIYA8He8
- **Covariance:** https://www.youtube.com/watch?v=qtaqvPAeEJY
- **Pearson's Correlation:** https://www.youtube.com/watch?v=xZ_z8KWkhXE
- **R-squared explained:** https://www.youtube.com/watch?v=2AQKmw14mHM

- **Fitting a line to data,:** https://www.youtube.com/watch?v=PaFPbb66DxQ

- **Linear Models Pt. 1 - Linear Regression:** https://www.youtube.com/watch?v=nk2CQITm_eo
- **Linear Models Pt. 2 - t-tests and ANOVA:** https://www.youtube.com/watch?v=NF5_btOaCig
  (Pt. 2 → also for next lecture where ANOVA is introduced)
- **Linear Models Pt.3 - Design Matrices:** https://www.youtube.com/watch?v=CqLGvwi-5Pc
- **Linear Models Pt.4 - Design Matrix in R :** https://www.youtube.com/watch?v=Hrr2anyK_5s