ANOVA

*AN*alysis *O*f *VA*riance

ANOVA

The purpose of ANOVA (Analysis of Variance) is to test for significant differences between means of different groups.

Let us say 3 groups of students were given 3 different memory pills and their scores in an exam recorded. We want to understand if the differences are due to within group differences or between group differences.

Group 1	Group 2	Group 3
3	5	5
2	3	6
1	4	7
$\bar{X}_1 = 2$	$\bar{X}_2 = 4$	$\bar{X}_3 = 6$

$$\bar{X} = \frac{3+2+1\ 5+3+4+5+6\ 7}{9} = 4$$

Total sum of square, SST

$$= (3-4)^2 + (2-4)^2 + (1-4)^2 + (5-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2$$

$$= 30$$

When there are ***m*** groups and ***n*** members in each group, the degrees of freedom are ***mn* - *1*,** since we can calculate one member knowing the overall mean.

How much of this variation is coming from within the groups and how much from between the groups?

Group 1	Group 2	Group 3
3	5	5
2	3	6
1	4	7
$\bar{X}_1 = 2$	$\bar{X}_2 = 4$	$\bar{X}_3 = 6$

$$\bar{X} = \frac{3+2+1\ 5+3+4+5+6\ 7}{9} = 4$$

Total sum of square Within, SSW

$= (3-2)^2 + (2-2)^2 + (1-2)^2 + (5-4)^2 + (3-4)^2 + (4-4)^2 + (5-6)^2 + (6-6)^2 + (7-6)^2$

$= 6$

When there are ***m*** groups and ***n*** members in each group, the degrees of freedom are ***m*(*n* - *1*),** since we can calculate one member knowing the group mean.

Total sum of square between, SSB = $3(2-4)^2 + 3(4-4)^2 + 3(6-4)^2 = 24$

When there are *m groups, the degrees of freedom are **m** - **1**.*

- **SST = SSW + SSB**
- Also, for degrees of freedom, $\boldsymbol{mn}$ - $\mathbf{1}$ = $\boldsymbol{m}$ $\boldsymbol{n}$ - $\mathbf{1}$ + ($\boldsymbol{m}$ - $\mathbf{1}$)

Group 1	Group 2	Group 3
3	5	5
2	3	6
1	4	7
$\bar{X}_1 = 2$	$\bar{X}_2 = 4$	$\bar{X}_3 = 6$

$$\bar{X} = \frac{3+2+1\ 5+3+4+5+6\ 7}{9} = 4$$

Given that mean of group 3 is highest and that of group 1 lowest, can we conclude that the pills given to group 3 had a larger impact or is it just variation within the group?

Let us have a null hypothesis that the population means of the 3 groups from which the samples were taken have the same mean, i.e., the pills do not have an impact on the performance in the exam. $\mu 1$ = $\mu 2$ = $\mu 3$. Let us also have a significance level, α = 0.10.

- What is the alternate hypothesis?
- The pills have an impact on performance.

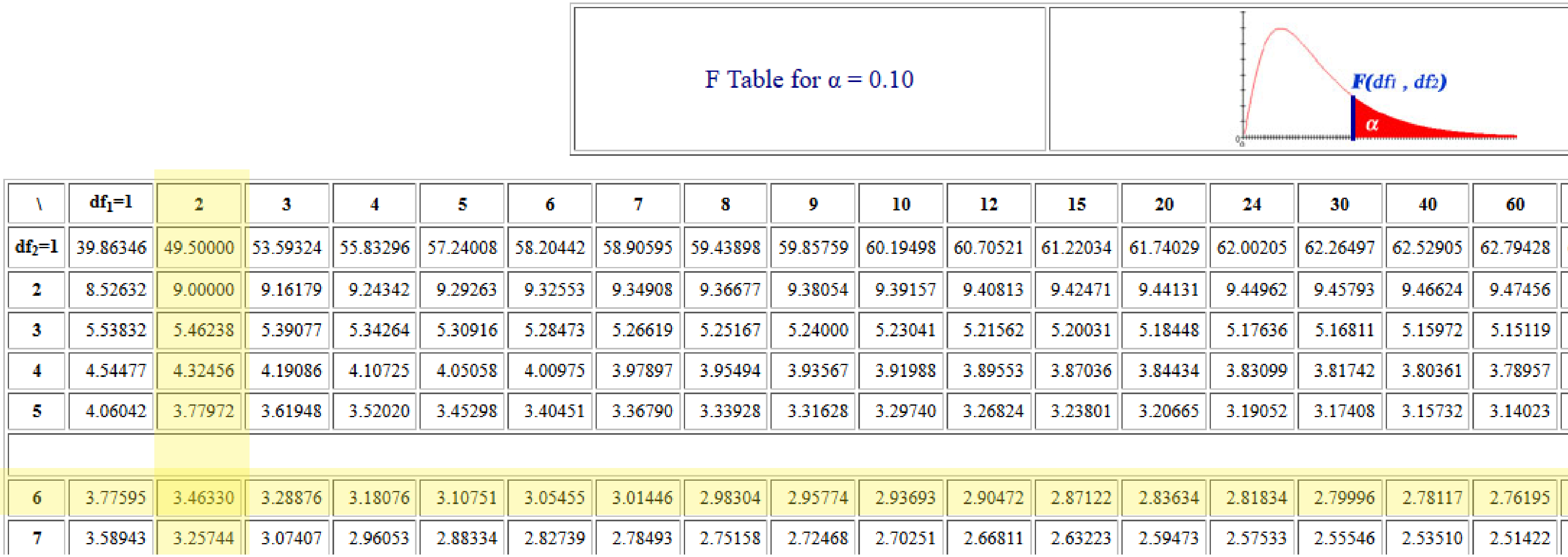
Group 1	Group 2	Group 3
3	5	5
2	3	6
1	4	7
$\bar{X}_1 = 2$	$\bar{X}_2 = 4$	$\bar{X}_3 = 6$

$$\bar{X} = \frac{3+2+1\ 5+3+4+5+6\ 7}{9} = 4$$

The test statistic used is F-statistic.

$$F - statistic = \frac{\frac{SSB}{df_{ssb}}}{\frac{SSW}{df_{ssw}}} = \frac{\frac{24}{2}}{\frac{6}{6}} = 12$$

If numerator is much bigger than the denominator, it means variation **between means has** bigger impact than variation **within, thus rejecting the null hypothesis.**



F Table for α = 0.10

\	df_1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60
df_2=1	39.86346	49.50000	53.59324	55.83296	57.24008	58.20442	58.90595	59.43898	59.85759	60.19498	60.70521	61.22034	61.74029	62.00205	62.26497	62.52905	62.79428
2	8.52632	9.00000	9.16179	9.24342	9.29263	9.32553	9.34908	9.36677	9.38054	9.39157	9.40813	9.42471	9.44131	9.44962	9.45793	9.46624	9.47456
3	5.53832	5.46238	5.39077	5.34264	5.30916	5.28473	5.26619	5.25167	5.24000	5.23041	5.21562	5.20031	5.18448	5.17636	5.16811	5.15972	5.15119
4	4.54477	4.32456	4.19086	4.10725	4.05058	4.00975	3.97897	3.95494	3.93567	3.91988	3.89553	3.87036	3.84434	3.83099	3.81742	3.80361	3.78957
5	4.06042	3.77972	3.61948	3.52020	3.45298	3.40451	3.36790	3.33928	3.31628	3.29740	3.26824	3.23801	3.20665	3.19052	3.17408	3.15732	3.14023
6	3.77595	3.46330	3.28876	3.18076	3.10751	3.05455	3.01446	2.98304	2.95774	2.93693	2.90472	2.87122	2.83634	2.81834	2.79996	2.78117	2.76195
7	3.58943	3.25744	3.07407	2.96053	2.88334	2.82739	2.78493	2.75158	2.72468	2.70251	2.66811	2.63223	2.59473	2.57533	2.55546	2.53510	2.51422

The df are 2 for numerator and 6 for denominator.

F_c, the critical F-statistic, therefore, is 3.46330. 12 is way higher than this and hence we reject the null hypothesis. That means the pills do have an impact on the performance.

INNOMATICS
TECHNOLOGY HUB

ANOTHER EXAMPLE

A wind turbine manufacturer is testing 3 different designs of the turbines. It picks 3 different sites in the same district to install each model of the turbine. The mean power output (MW) over the day is measured for 9 consecutive days in each of the sites.

We want to understand if the differences are due to within-group differences or between-group differences.

Model 1			Model 2			Model 3		
3	4	3	3	5	7	5	5	5
2	5	5	6	7	6	6	5	7
4	3	3	4	4	8	7	6	6
$\bar{X}_1 = 3.56\ MW$			$\overline{X_2} = 5.56\ MW$			$\overline{X_3} = 5.78\ MW$		

$$\bar{X} = \frac{134}{27} = 4.96\ MW$$

Total Sum of Squares, SST = 62.96

Total Sum of Squares Within, SSW = 36.00

Total Sum of Squares Between, SSB = 26.96

What is the null hypothesis?

All 3 sites from which the samples were taken have the same population-mean, i.e., the turbine design does not have an impact on the power production. That is

$$\mu_1 = \mu_2 = \mu_3$$

Let us also specify a significance level, $\alpha = 0.10$.

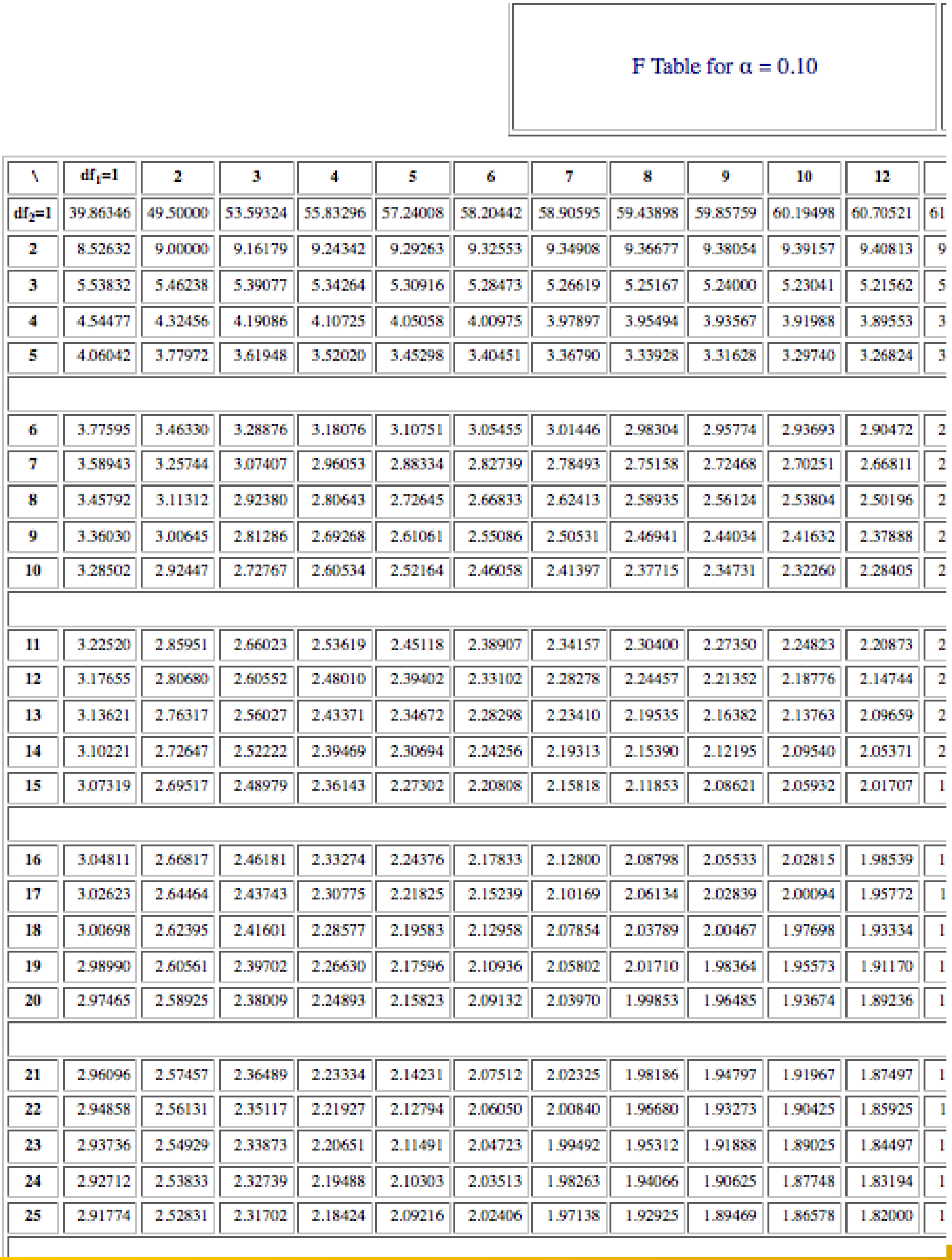
What is the alternate hypothesis?

The turbine design does impact the power output.

Compute the statistics

$$F\ - statistic = \frac{\frac{SSB}{df_{SSB}}}{\frac{SSW}{df_{SSW}}} = \frac{\frac{26.96}{2}}{\frac{36}{24}} = 8.99$$

If numerator is much bigger than the denominator, it means variation **between means has bigger impact than variation within,** thus rejecting the null hypothesis.



F Table for α = 0.10

\	df_1=1	2	3	4	5	6	7	8	9	10	12	
df_2=1	39.86346	49.50000	53.59324	55.83296	57.24008	58.20442	58.90595	59.43898	59.85759	60.19498	60.70521	61
2	8.52632	9.00000	9.16179	9.24342	9.29263	9.32553	9.34908	9.36677	9.38054	9.39157	9.40813	9
3	5.53832	5.46238	5.39077	5.34264	5.30916	5.28473	5.26619	5.25167	5.24000	5.23041	5.21562	5
4	4.54477	4.32456	4.19086	4.10725	4.05058	4.00975	3.97897	3.95494	3.93567	3.91988	3.89553	3
5	4.06042	3.77972	3.61948	3.52020	3.45298	3.40451	3.36790	3.33928	3.31628	3.29740	3.26824	3
6	3.77595	3.46330	3.28876	3.18076	3.10751	3.05455	3.01446	2.98304	2.95774	2.93693	2.90472	2
7	3.58943	3.25744	3.07407	2.96053	2.88334	2.82739	2.78493	2.75158	2.72468	2.70251	2.66811	2
8	3.45792	3.11312	2.92380	2.80643	2.72645	2.66833	2.62413	2.58935	2.56124	2.53804	2.50196	2
9	3.36030	3.00645	2.81286	2.69268	2.61061	2.55086	2.50531	2.46941	2.44034	2.41632	2.37888	2
10	3.28502	2.92447	2.72767	2.60534	2.52164	2.46058	2.41397	2.37715	2.34731	2.32260	2.28405	2
11	3.22520	2.85951	2.66023	2.53619	2.45118	2.38907	2.34157	2.30400	2.27350	2.24823	2.20873	2
12	3.17655	2.80680	2.60552	2.48010	2.39402	2.33102	2.28278	2.24457	2.21352	2.18776	2.14744	2
13	3.13621	2.76317	2.56027	2.43371	2.34672	2.28298	2.23410	2.19535	2.16382	2.13763	2.09659	2
14	3.10221	2.72647	2.52222	2.39469	2.30694	2.24256	2.19313	2.15390	2.12195	2.09540	2.05371	2
15	3.07319	2.69517	2.48979	2.36143	2.27302	2.20808	2.15818	2.11853	2.08621	2.05932	2.01707	1
16	3.04811	2.66817	2.46181	2.33274	2.24376	2.17833	2.12800	2.08798	2.05533	2.02815	1.98539	1
17	3.02623	2.64464	2.43743	2.30775	2.21825	2.15239	2.10169	2.06134	2.02839	2.00094	1.95772	1
18	3.00698	2.62395	2.41601	2.28577	2.19583	2.12958	2.07854	2.03789	2.00467	1.97698	1.93334	1
19	2.98990	2.60561	2.39702	2.26630	2.17596	2.10936	2.05802	2.01710	1.98364	1.95573	1.91170	1
20	2.97465	2.58925	2.38009	2.24893	2.15823	2.09132	2.03970	1.99853	1.96485	1.93674	1.89236	1
21	2.96096	2.57457	2.36489	2.23334	2.14231	2.07512	2.02325	1.98186	1.94797	1.91967	1.87497	1
22	2.94858	2.56131	2.35117	2.21927	2.12794	2.06050	2.00840	1.96680	1.93273	1.90425	1.85925	1
23	2.93736	2.54929	2.33873	2.20651	2.11491	2.04723	1.99492	1.95312	1.91888	1.89025	1.84497	1
24	2.92712	2.53833	2.32739	2.19488	2.10303	2.03513	1.98263	1.94066	1.90625	1.87748	1.83194	1
25	2.91774	2.52831	2.31702	2.18424	2.09216	2.02406	1.97138	1.92925	1.89469	1.86578	1.82000	1

The *df are 2 for numerator and 24 for* denominator.
Fc, the critical F-statistic, therefore, is 2.53833.

Our F=8.99 is way higher than this and hence we reject the null hypothesis. That means the turbine design does have an impact on the power production.

Anova: Single Factor

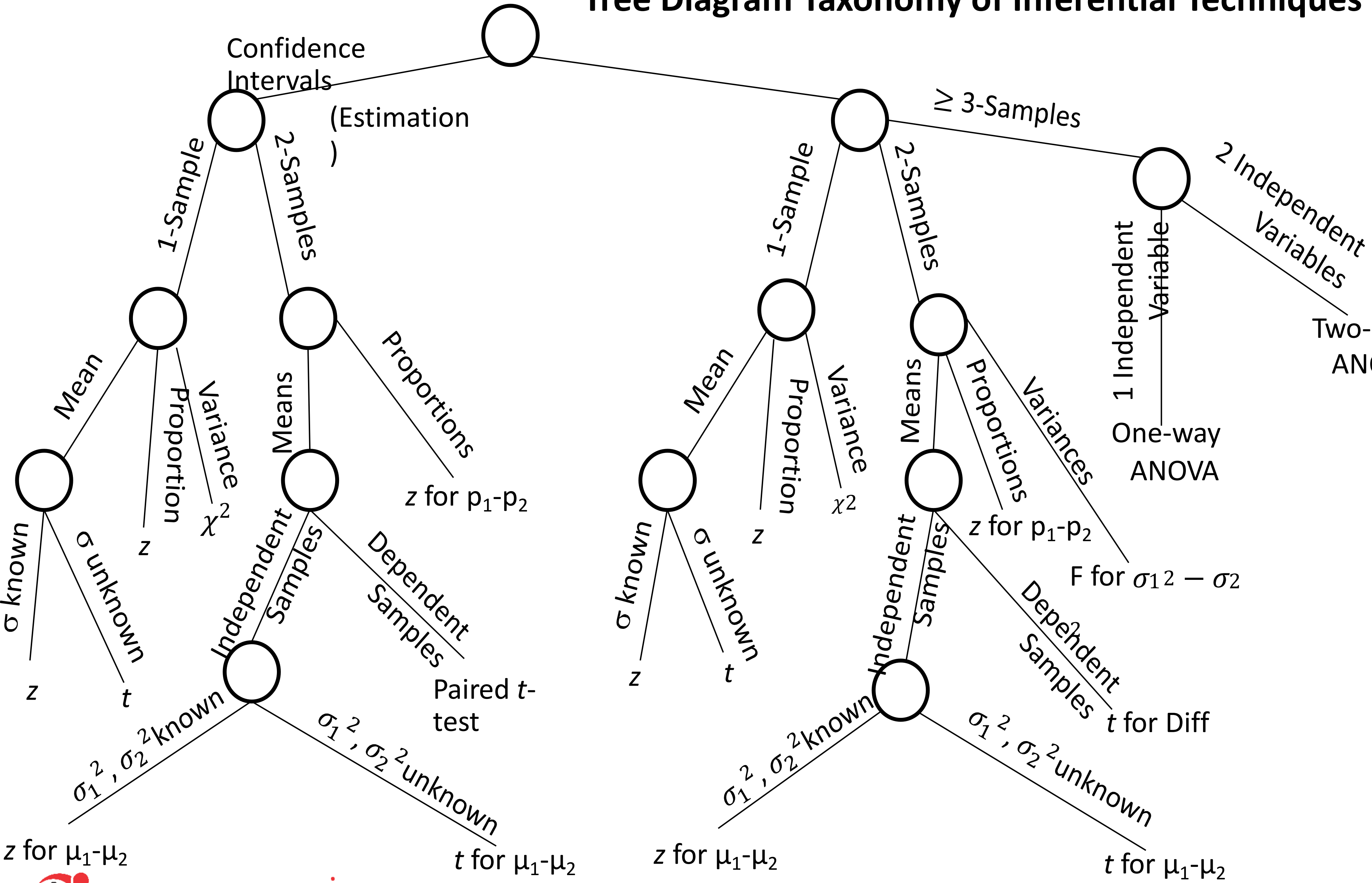
SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Group1	9	32	3.55556	1.027778
Group2	9	50	5.55556	2.777778
Group3	9	52	5.77778	0.694444

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	26.963	2	13.4815	8.987654	0.0012	2.5383
Within Groups	36	24	1.5			
Total	62.963	26				

Tree Diagram Taxonomy of Inferential Techniques

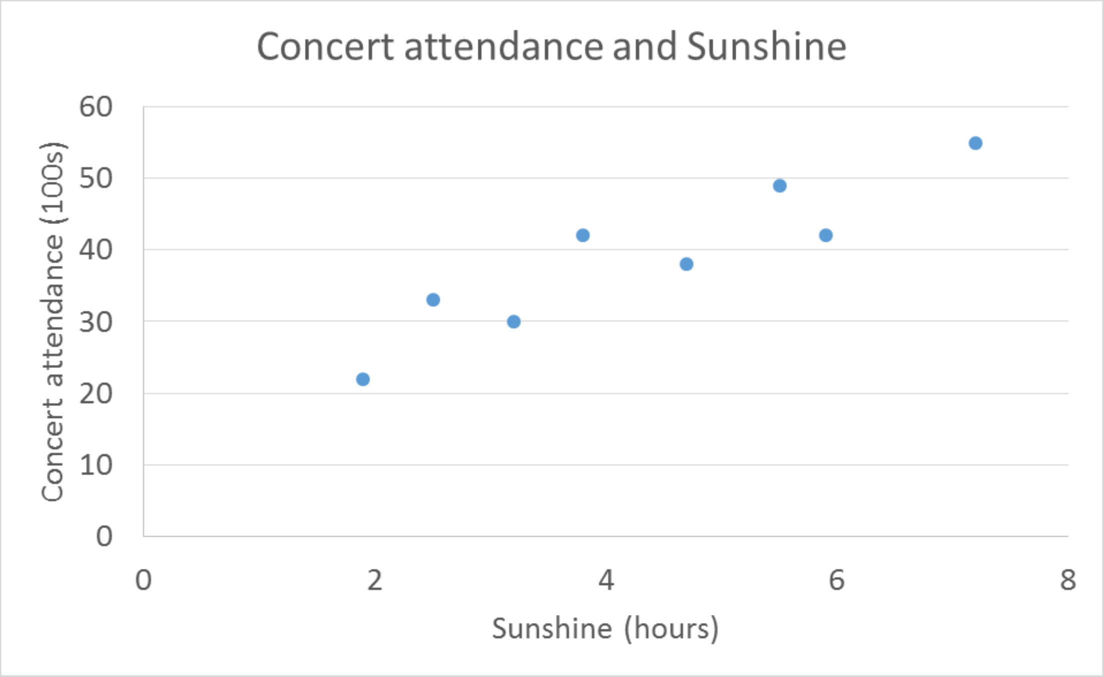


Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

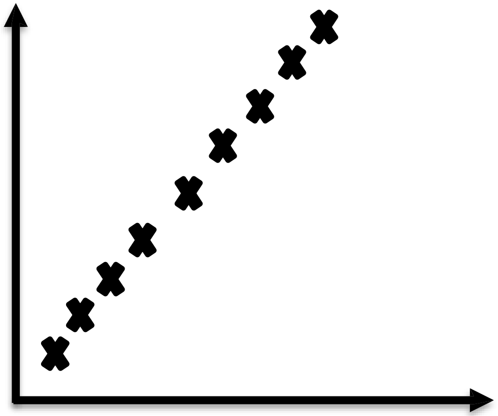
- The band makes a loss if less than 3500 people attend.
- Based on predicted hours of sunshine, can we predict ticket sales?
- Are sunshine and concert attendance correlated?

Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

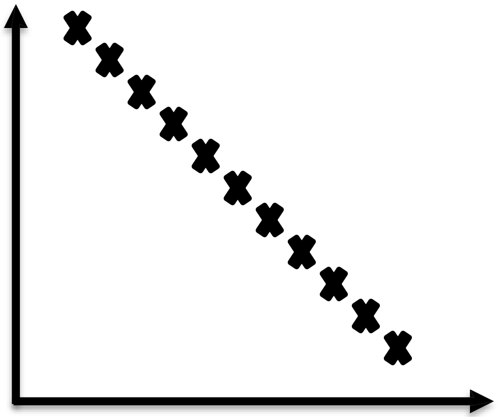
- Independent variable (explanatory) – Sunshine – Plotted on X-axis
- Dependent variable (response) – Concert attendance – Plotted on Y-axis



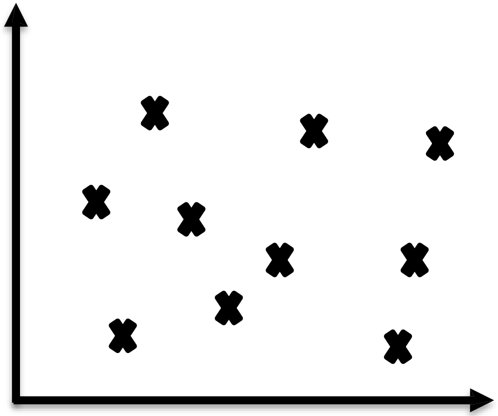
- Hours of sunshine and concert attendance are correlated, i.e., in general, longer sunshine hours indicate higher attendance.



Positive Linear Correlation

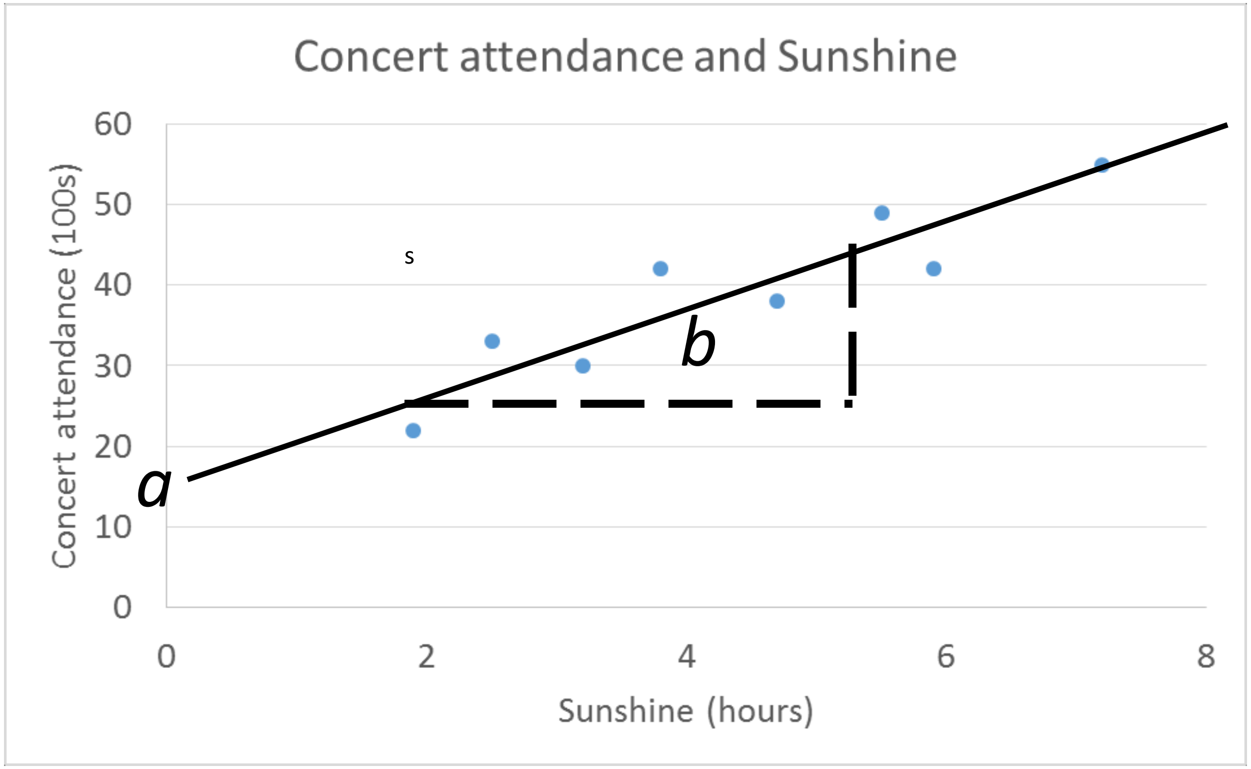


Negative Linear Correlation



No Correlation

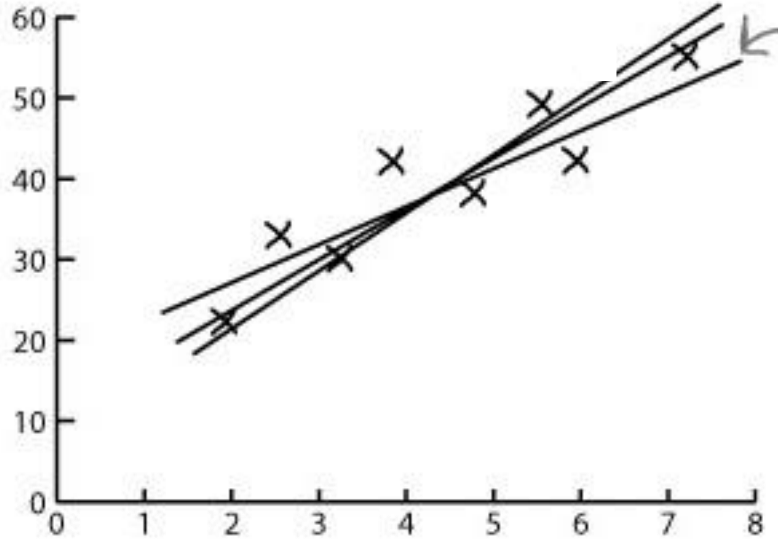
We need to find the equation of the line.

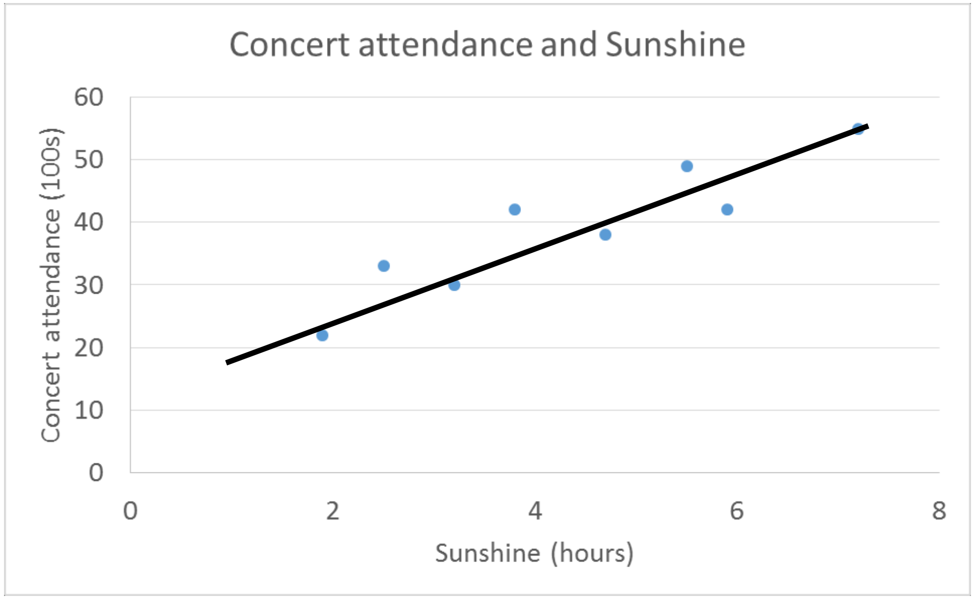


$y = a + bx$

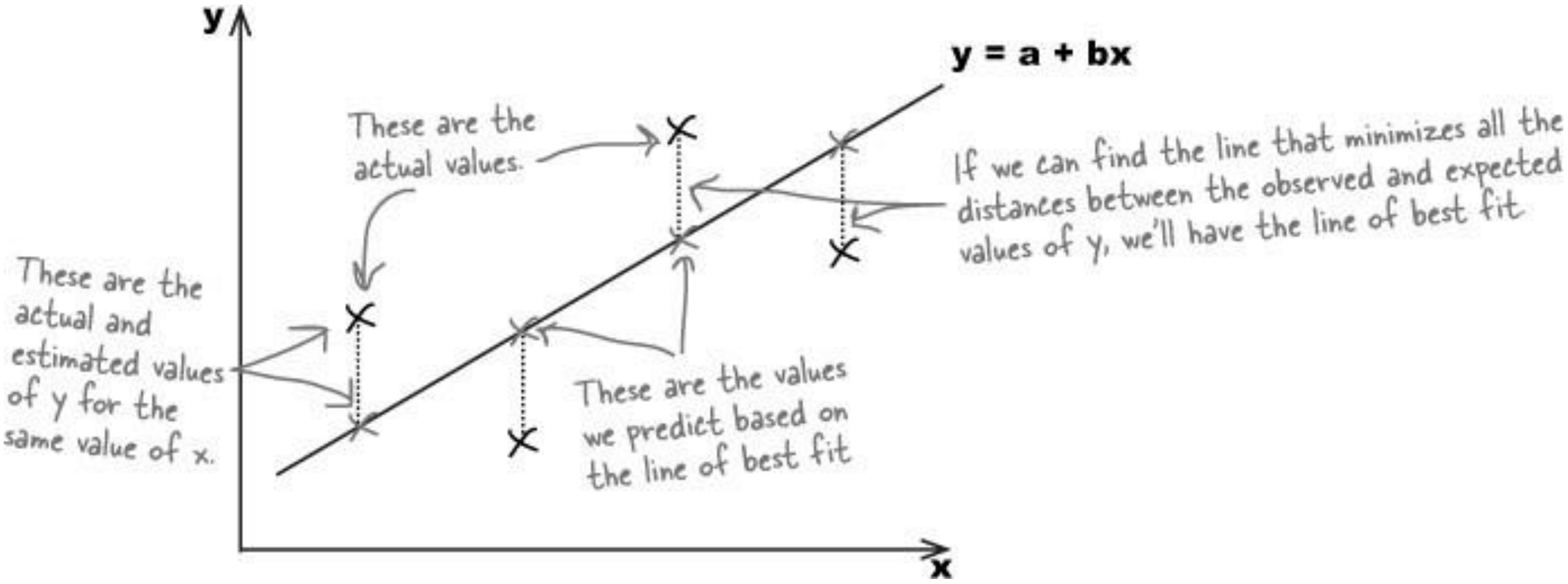
Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

- Line of the best fit





We need to minimize errors


y
y = a + bx
These are the actual values.
These are the actual and estimated values of y for the same value of x.
If we can find the line that minimizes all the distances between the observed and expected values of y, we'll have the line of best fit
These are the values we predict based on the line of best fit
x

We could do that by minimizing $\sum(y_i \ - \hat{y}_i)$, where $y_i$is the actual value and $\hat{y}_i$ its estimate. $(y_i \ - \hat{y}_i)$ is also know as residual

We need to minimize errors.

Just as we did when finding variance, we find the **sum of squared errors or SSE.** *Note in variance calculations, we subtract mean,* $\bar{y}, not\ \hat{y}_i$.

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

The value of b, the slope, that minimizes the SSE is given by

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

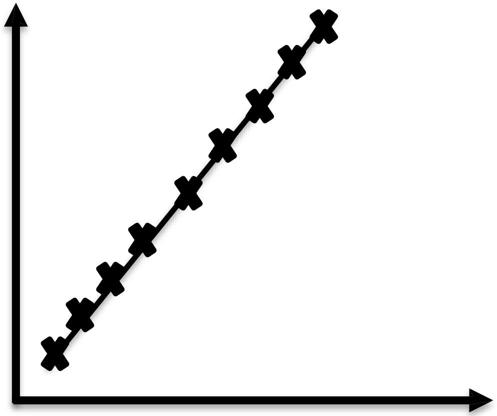
The value of *b,* the slope, that minimizes the SSE is given by

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

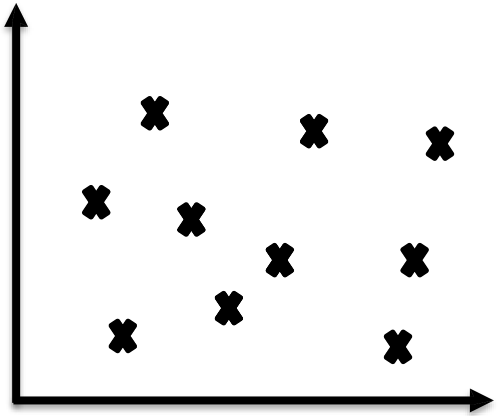
How do you calculate a ? The line of best fit must pass through $(\bar{x}, \bar{y})$. Substituting in the equation y = a+ b x, we can find a.

This method of fitting the line of best fit is called **least squares regression.**

But how do you know how good this fitted line is ?



Perfect Linear Correlation



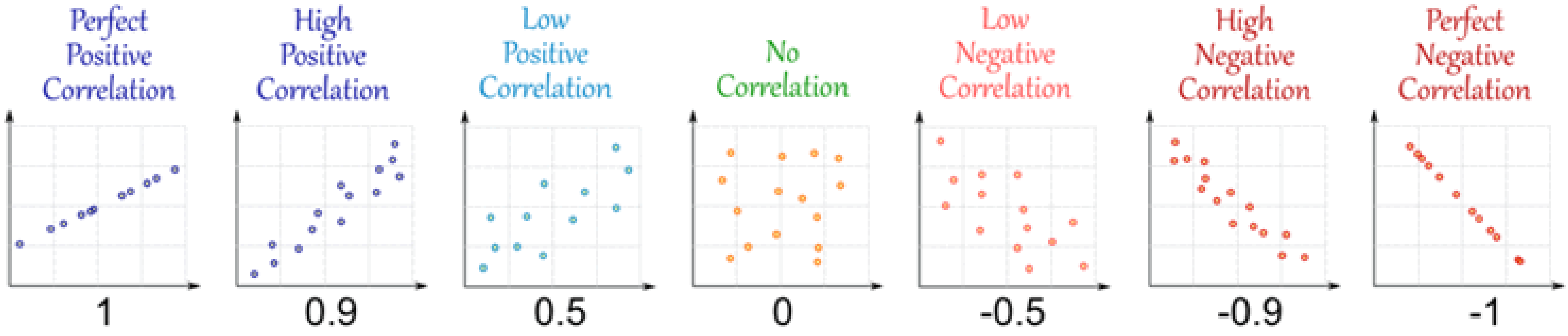
No Linear Correlation

The fit of the line is given by **correlation coefficient *r*.**

$$r = \frac{S_{xy}}{S_x S_y}$$

Correlation Coefficient

Correlation coefficient, *r, is a* number between *-1* and *1* and tells us how well a regression line fits the data.



It gives the strength and direction of the relationship between two variables.

Correlation Coefficient

$r = \frac{bs_x}{s_y}$ where *b* is the slope of the line of best fit, s_x is the standard deviation of the *x* values in the sample, and s_y is the standard deviation of the y values in the sample.

$s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ and $s_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n-1}}$

Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

Find *r* for this data

Correlation Coefficient and Covariance

$s_x^2 = \frac{\sum(x-\bar{x})^2}{(n-1)}$, $s_y^2 = \frac{\sum(y-\bar{y})^2}{(n-1)}$, $s_{xy} = \frac{\sum(x-\bar{x})\,(y-\bar{y})}{(n-1)}$, where s_x^2 is the sample variance of the x values, s_y^2 is the sample variance of the y values and s_{xy} is the covariance.

$b = \frac{s_{xy}}{s_x^2}$ and so, $r = \frac{s_{xy}}{s_x s_y}$

Covariance and Correlation

$$s_{xy} = \frac{\sum(x-\bar{x})\,(y-\bar{y})}{(n-1)}, r = \frac{s_{xy}}{s_x s_y}$$

- The value of covariance itself doesn't say much. It only shows whether the variables are moving together (positive value) or opposite to each other (negative value).

- To know the strength of how the variables move together, covariance is standardized to the dimensionless quantity, correlation.

Coefficient of Determination

The coefficient of determination is given by r^2 or R^2. It is the percentage of variation in the *y* variable that is explainable by the x variable. For example, what percentage of the variation in open-air concert attendance is explainable by the number of hours of predicted sunshine.

If r^2 = 0, it means you can't predict the *y* value from the x value.

If r^2 = 1, it means you can predict the *y* value from the x value without any errors.

Usually, r^2 is between these two extremes.

Covariance, Correlation and R^2

How do the interest rates of federal funds and the commodities futures index co-vary and correlate?

Month	Interest Rate	Futures Index
1	7.43	221
2	7.48	222
3	8.00	226
4	7.75	225
5	7.60	224
6	7.63	223
7	7.68	223
8	7.67	226
9	7.59	226
10	8.07	235
11	8.03	233
12	8.00	241

Covariance, Correlation and R^2

Month	Interest Rate	Futures Index	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})(y-\overline{y})$
1	7.43	221			
2	7.48	222			
3	8.00	226			
4	7.75	225			
5	7.60	224			
6	7.63	223			
7	7.68	223			
8	7.67	226			
9	7.59	226			
10	8.07	235			
11	8.03	233			
12	8.00	241			
Mean	7.74	227.08			
StDev	0.22	6.07			

$$Cov = \frac{12.216}{11} = 1.11$$

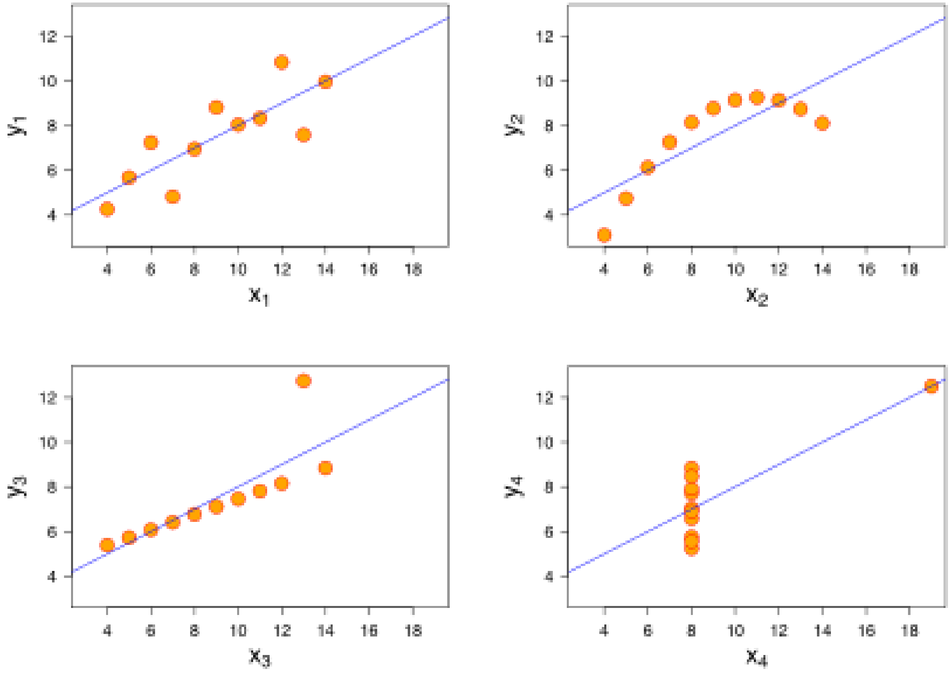
$$r = \frac{1.11}{0.22 * 6.07}$$
$$= 0.815$$

$$R^2 = 0.815^2 = 0.665$$

Anscombe's Quartet

Anscombe's quartet							
I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.1	10	7.46	8	6.6
8	6.95	8	8.1	8	6.77	8	5.8
13	7.58	13	8.7	13	12.7	8	7.7
9	8.81	9	8.8	9	7.11	8	8.8
11	8.33	11	9.3	11	7.81	8	8.5
14	9.96	14	8.1	14	8.84	8	7
6	7.24	6	6.1	6	6.08	8	5.3
4	4.26	4	3.1	4	5.39	19	13
12	10.8	12	9.1	12	8.15	8	5.6
7	4.82	7	7.3	7	6.42	8	7.9
5	5.68	5	4.7	5	5.73	8	6.9

Property	Value
Mean of x in each case	9 (exact)
Sample variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Sample variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	y = 3.00 + 0.500x (to 2 and 3 decimal places, respectively)



THE STORY

We started by looking at various types of data (Categorical/Numerical) and we sought methods to describe the central tendency of the data. We discussed Mean, Median and Mode.

We realized to understand the data better, along with the central tendency,
We need to understand the spread. We looked at Range, Interquartile Range, Variance and Standard Deviation.

We discussed Box and Whisker plots as a way to represent both central-tendency and the spread in the data.

We moved on to basic Probability theory and described rules of probabilities using Venn Diagrams. We talked about conditional probabilities and that led to Bayes Theorem.

We then looked at some of the commonly occurring Probability Distributions and their properties, and looked at the expected values, their variance and the probabilities of various possible outcomes.

Then we saw how the *Sampling Distributions of Means* tend to normal distribution irrespective of how the population is distributed and learned how to describe populations based on available sample data.

We then looked at Confidence Intervals to properly describe the conclusions about populations based on samples.

Then we studied Hypothesis Tests as an alternative inferential technique to prove our claims. Of course, there are errors in these tests too.

We then studied various statistical tests to test hypotheses. We looked at how to analyze results and find differences between what we expect and what we get, through χ^2 Distributions (goodness-of-fit).

We studied ANOVA, 2-sample t-tests and F test as a means of understanding significant differences between means and variances.

We also studied Independence, Correlation and Covariance between variables and learned about Regression basics.

Reference

Head First Statistics