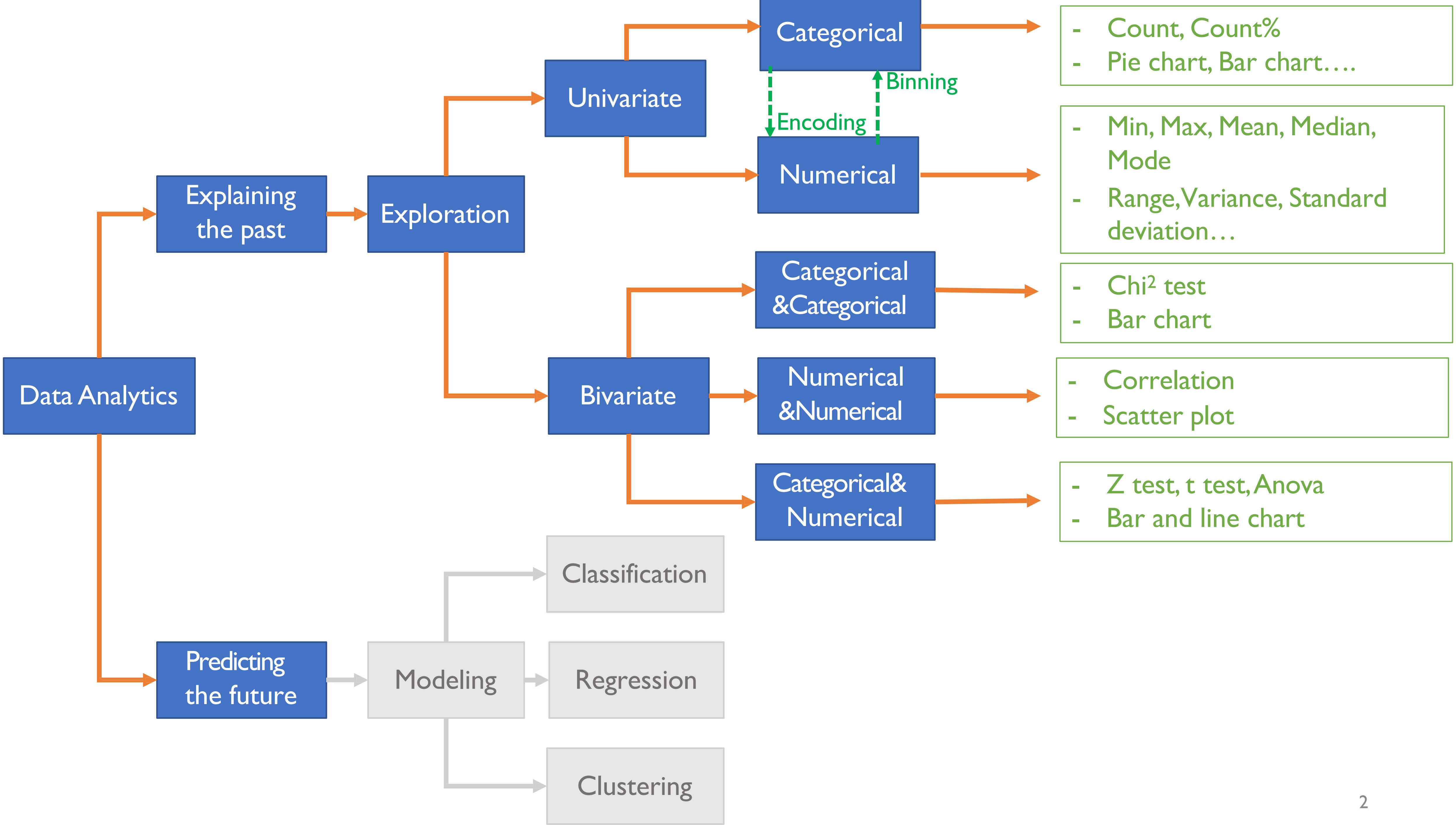
Data Exploration (part 2)

Dr. Teena Sharma

Ph.D., University of Quebec at Chicoutimi, Canada

(tsharma@etu.uqac.ca)

Summer 2024



Data Analytics

Explaining the past

Exploration

Univariate

Categorical

- Count, Count%
- Pie chart, Bar chart….

Binning

Encoding

Numerical

- Min, Max, Mean, Median, Mode
- Range,Variance, Standard deviation…

Bivariate

Categorical &Categorical

- Chi² test
- Bar chart

Numerical &Numerical

- Correlation
- Scatter plot

Categorical& Numerical

- Z test, t test,Anova
- Bar and line chart

Predicting the future

Modeling

Classification

Regression

Clustering

BIVARIATE ANALYSIS

Introduction

- Bivariate analysis is the simultaneous analysis of two variables (attributes).

- It explores the concept of relationship between two variables,
  - whether there *exists an association* and the *strength* of this association, or
  - whether there are *differences between two variables* and the *significance* of these differences.

Introduction

- There are three types of bivariate analysis

1. Numerical & Numerical
2. Categorical & Categorical
3. Numerical & Categorical

Numerical and Numerical

Scatter Plot

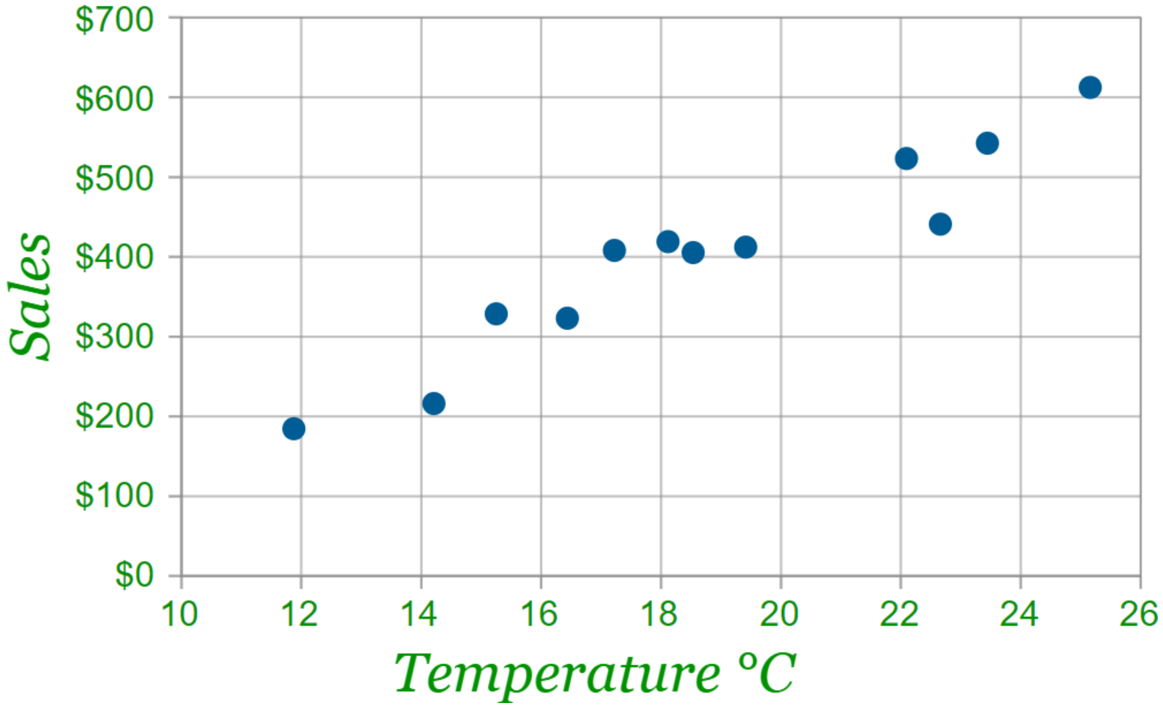
- A scatter plot is a useful visual representation of the relationship between two numerical variables (attributes).
- It is usually drawn before working out a linear correlation (?) or fitting a regression line (?).

- The resulting pattern indicates the type (linear or non-linear) and strength of the relationship between two variables.

Example

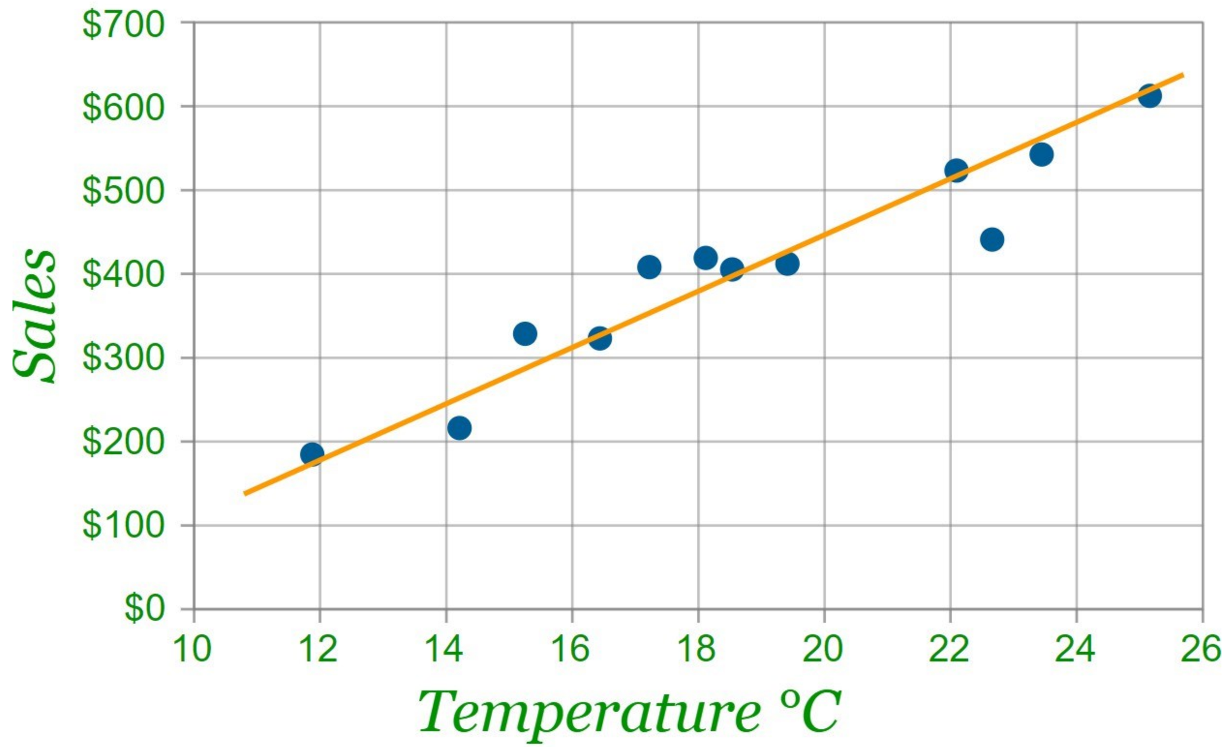
Ice Cream Sales vs Temperature

Temperature ℃	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408



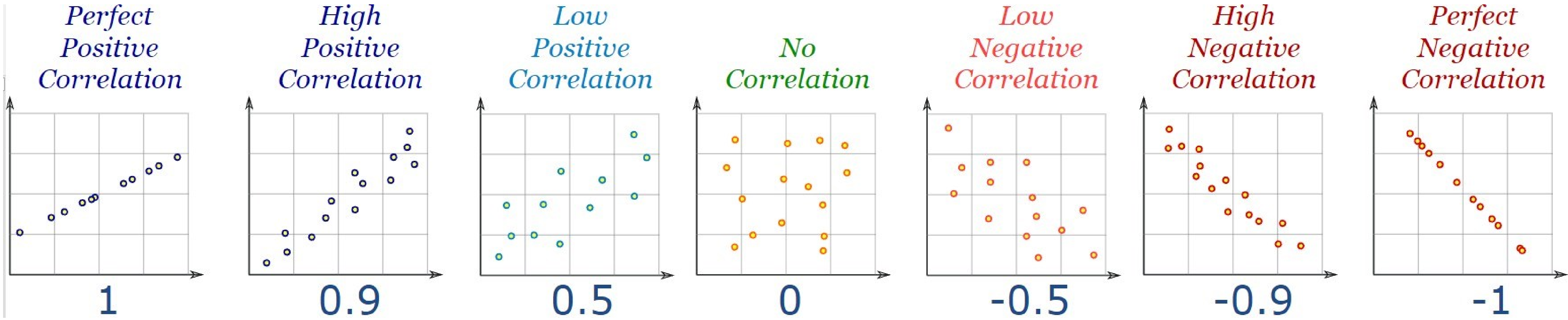
Example - Cont.,

- We can also draw a "Line of Best Fit" (also called a "Trend Line") on our scatter plot:
    - Try to have the line **as close as possible to all points**, and as many points above the line as below.



Correlation

- When the two sets of data are strongly linked together we say they have a **High Correlation**.
  - Correlation is **Positive** when the values **increase** together, and
  - Correlation is **Negative** when one value **decreases** as the other increases



Correlation – Cont.,

Correlation can have a value:

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation

Formally speaking: (Linear) correlation quantifies the strength of a linear relationship between two numerical variables.

- When there is no correlation between two variables, there is no tendency for the values of one quantity to increase or decrease with the values of the second quantity.

Correlation

$$r = \frac{Covar(x,y)}{\sqrt{Var(x)Var(y)}}$$

$r : Linear\ Correlation$

$$Covar(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n}$$

$Covar : Covariance$

$Var : Variance$

$$Var(x) = \frac{\sum(x-\bar{x})^2}{n}$$

$$Var(y) = \frac{\sum(y-\bar{y})^2}{n}$$

r only measures the **strength of a linear relationship** and is always *between -1 and 1* where

- -1 means perfect negative linear correlation and
- +1 means perfect positive linear correlation and
- 0 means no linear correlation.

Example

Temperature	83	64	72	81	70	68	65	75	71	85	80	72	69	75
Humidity	86	65	90	75	96	80	70	80	91	85	90	95	70	70

	Variance	Covariance	Correlation
Temperature	40.10	19.78	0.32
Humidity	98.23		

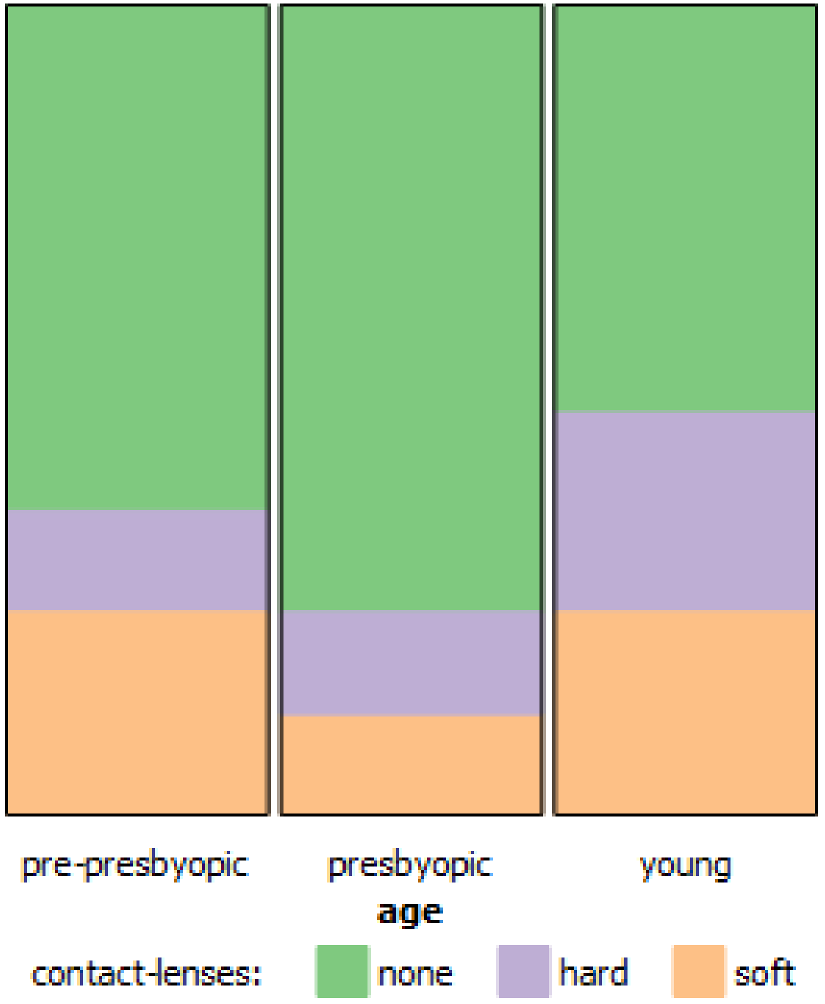
There is a *weak linear correlation* between Temperature and Humidity.

Categorical & Categorical

Stacked Column Chart

- Stacked Column chart is a useful graph to visualize the relationship between two categorical variables.

- It compares the percentage that each category from one variable contributes to a total across categories of the second variable.



Aside 1: Statistical Significance (p-value)

- Statistical significance is a determination that a relationship between two or more variables is caused by something other than chance.

- Statistical significance is used to provide evidence concerning the plausibility of the null hypothesis, which hypothesizes that there is nothing more than random chance at work in the data.

- A p-value is *a measure of the probability* that an observed difference could have occurred just by random chance.

- Generally, a p-value of 5% or lower is considered statistically significant (i.e. *there is less than 5% chance that the relationship is random or happens by chance*).

Example

- Suppose Alex, a financial analyst, is curious as to whether some investors had advance knowledge of a company's sudden failure.
- Alex decides to compare the average of daily market returns prior to the company's failure with those after to see if there is a statistically significant difference between the two averages.

- The study's p-value was 28% (>5%), indicating that a difference as large as the observed is not unusual under the chance-only explanation.
- Thus, the data did not provide compelling evidence of advance knowledge of the failure

Aside 2: Degree of Freedom

Think of it this way:

- the number of values that are free to vary in a data set.

Q. Pick a set of numbers that have a mean (average) of 10.
A. Some sets of numbers you might pick: 9, 10, 11 or 8, 10, 12 or 5, 10, 15.

Once you have chosen the first two numbers in the set, the third is fixed. In other words, you can't choose the third item in the set. The only numbers that are free to vary are the first two.

You can pick 9 + 10 or 5 + 15, but once you've made that decision
you must choose a particular number that will give you the mean you are looking for. So degrees of freedom for a set of three numbers is TWO.

Aside 2: Degree of Freedom

More generally,

- Degrees of freedom of an estimate is the **number of independent pieces of information that went into calculating the estimate.**

- It’s not quite the same as the number of items in the sample.
- In order to get the df for the estimate, you have to subtract 1 from the number of items.
- Let’s say you were finding the mean weight loss for a low-carb diet. You could use 4 people, giving 3 degrees of freedom (4 – 1 = 3), or you could use one hundred people with df = 99.

Aside 2: Degree of Freedom

- If we are dealing with one sample:

$$\textit{Degrees of Freedom} = n - 1$$

- The degree of freedom will be calculated differently if dealing with more than one sample, or if used in the context of statistical test, as seen in the next slides.

Chi-square Test for Independence

- The chi-square test can be used to determine the association between categorical variables.

- It is based on the difference between the expected frequencies (*e*) and the observed frequencies (*n*) in one or more categories in the frequency table.

Chi-square Test for Independence

- *The chi-square distribution* returns a probability for the computed chi-square and the degree of freedom.
- A probability of zero shows a complete dependency between two categorical variables and a probability of one means that two categorical variables are completely independent.

- *Tchouproff Contingency Coefficient* measures the amount of dependency between two categorical variables.

Chi-square Test

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{n_{i.} n_{.j}}{n}$$

$$df = (r-1)(c-1)$$

Tchouproff Contingency Coefficient

$$\rho_\tau = \sqrt{\frac{\chi^2}{n\sqrt{(c-1)(r-1)}}}$$

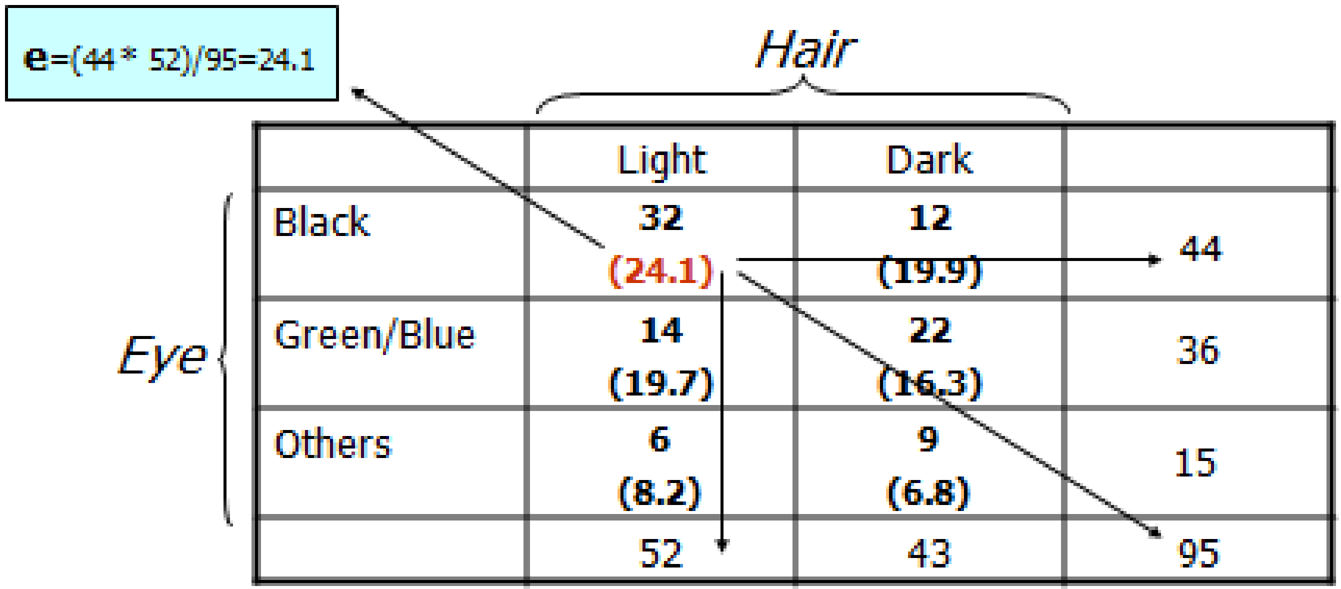
e - the expected frequencies

n - the observed frequencies

r – number of rows in contingency table

c – number of columns in contingency table

Chi-square Test



e=(44* 52)/95=24.1

Hair

Eye

	Light	Dark	
Black	32 (24.1)	12 (19.9)	44
Green/Blue	14 (19.7)	22 (16.3)	36
Others	6 (8.2)	9 (6.8)	15
	52	43	95

$$\chi^2 = 10.67$$

$$df = (r-1)(c-1) = (3-1)(2-1) = 2$$

$$p = 0.005$$

$$\rho_z = \sqrt{\frac{10.67}{95\sqrt{(3-1)(2-1)}}} = 0.28$$

Categorical & Numerical

Line Chart with Error Bars

- A line chart with error bars displays information as a series of data points connected by straight line segments.
- Each data point is average of the numerical data for the corresponding category of the categorical variable with error bar showing standard error SE.
- It is a way to summarize how pieces of information are related and how they vary depending on one another

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

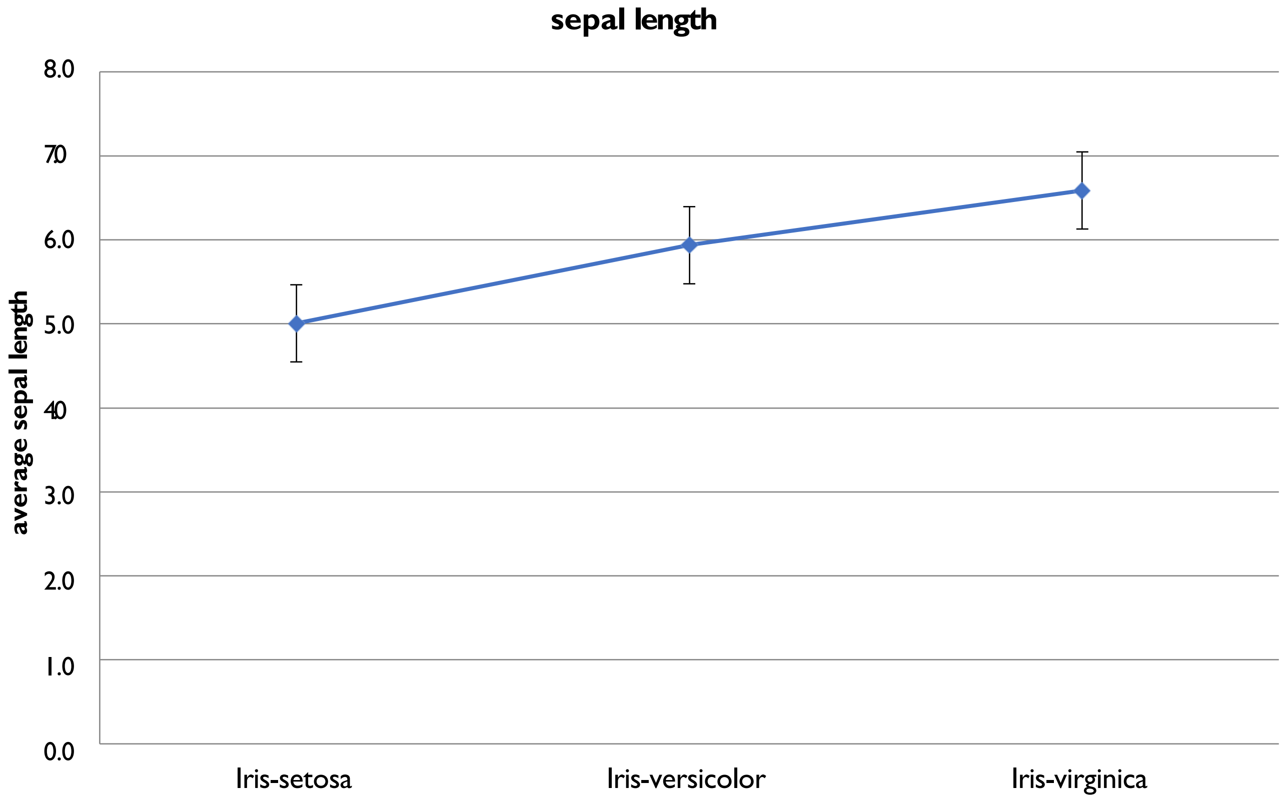
SE = standard error of the sample

σ = sample standard deviation

n = number of samples

Line Chart with Error Bars

sepal length		
Iris-setosa	Iris-versicolor	Iris-virginica
5.1	7	6.3
4.9	6.4	5.8
4.7	6.9	7.1
4.6	5.5	6.3
5	6.5	6.5
5.4	5.7	7.6
4.6	6.3	4.9
5	4.9	7.3
4.4	6.6	6.7
4.9	5.2	7.2
...	...	...



Z-test and t-test

- Z-test and t-test are basically the same.

- They assess whether the averages of two groups are statistically different from each other.

- This analysis is appropriate for comparing the averages of a numerical variable for two categories of a categorical variable.

Z-test

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

where:

- $\overline{X}_1, \overline{X}_2$: *Averages*
- S_1^2, S_2^2: *Variances*
- N_1, N_2: *Counts*
- $\boldsymbol{Z}$: *Standard Normal Distribution*

If the probability of Z is small, the difference between two averages is more significant.

T-test

- When the n_1 or n_2 is less than 30 we use the t-test instead of the Z-test

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

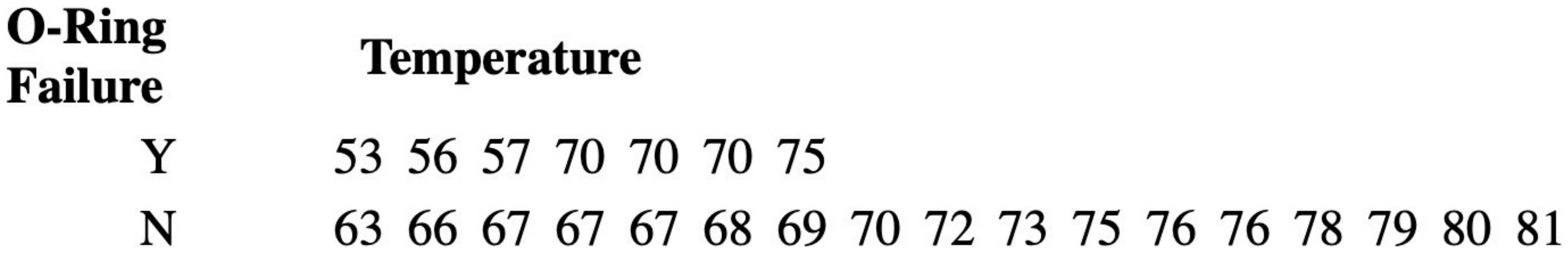
$$S^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2}$$

where:

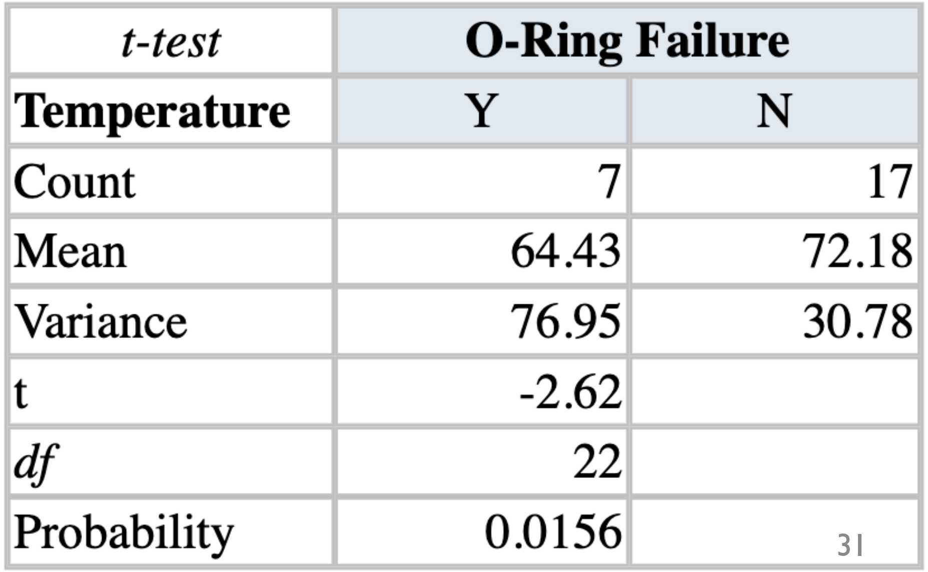
- $\overline{X}_1, \overline{X}_2$: *Averages*
- S_1^2, S_2^2 : *Variances*
- N_1, N_2 : *Counts*
- t : has t distribution with $N_1 + N_2 - 2$ degree of freedom

Example

- Is there a significant difference between the means (averages) of the numerical variable (Temperature) in two different categories of the categorical variable (O-Ring Failure)?

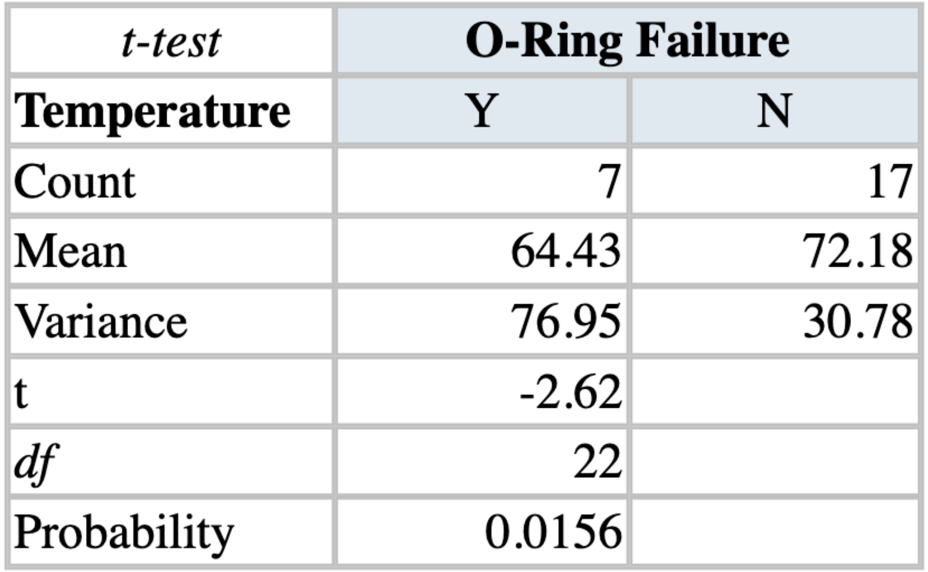


O-Ring Failure	Temperature
Y	53 56 57 70 70 70 75
N	63 66 67 67 67 68 69 70 72 73 75 76 76 78 79 80 81



t-test	O-Ring Failure	
Temperature	Y	N
Count	7	17
Mean	64.43	72.18
Variance	76.95	30.78
t	-2.62	
df	22	
Probability	0.0156	

Example

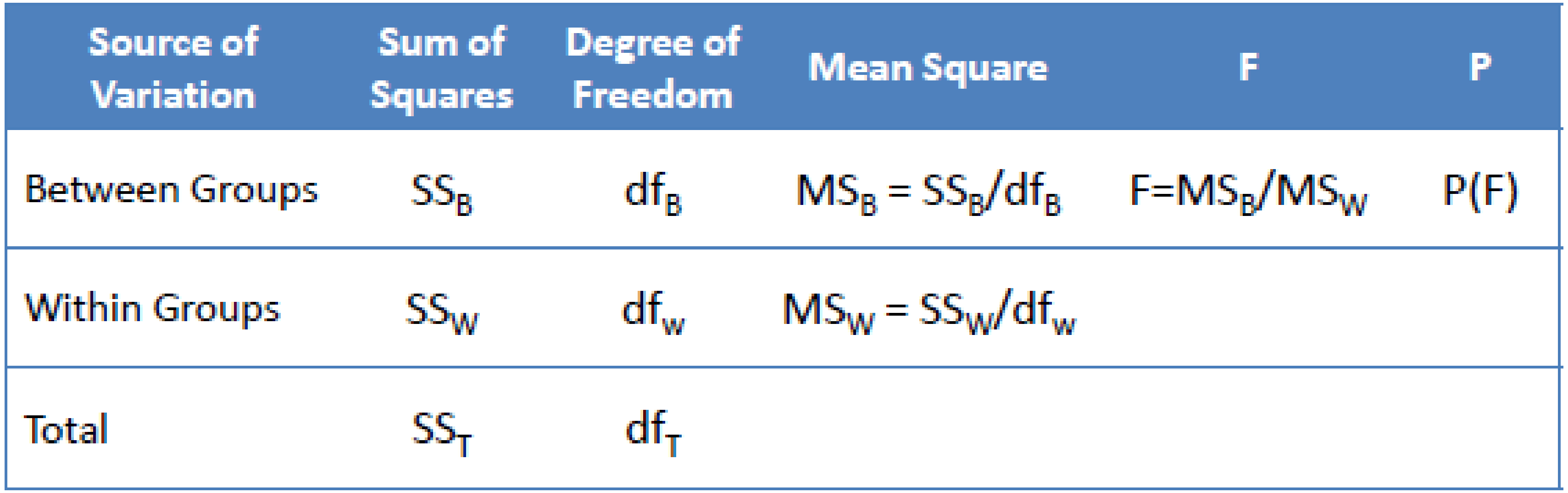


t-test	**O-Ring Failure**	
Temperature	Y	N
Count	7	17
Mean	64.43	72.18
Variance	76.95	30.78
t	-2.62	
df	22	
Probability	0.0156	

The low probability (0.0156) means that the difference between the average temperature for failed O-Ring and the average temperature for intact O-Ring is significant

Analysis of Variance (ANOVA)

- The ANOVA test assesses whether the averages of more than two groups are statistically different from each other.
- This analysis is appropriate for comparing the averages of a numerical variable for more than two categories of a categorical variable.



Source of Variation	Sum of Squares	Degree of Freedom	Mean Square	F	P
Between Groups	SS_B	df_B	$MS_B = SS_B/df_B$	$F=MS_B/MS_W$	P(F)
Within Groups	SS_W	df_w	$MS_W = SS_W/df_w$		
Total	SS_T	df_T			

Analysis of Variance (ANOVA)

$$SS_B = \sum_{i=1}^{k} \frac{(\sum X)_i^2}{N_i} - \frac{\left(\sum_{i=1}^{k} \left(\sum X\right)_i\right)^2}{\sum_{i=1}^{k} N_i}$$

$$SS_W = \sum_{i=1}^{k} \left(\sum X^2\right)_i - \sum_{i=1}^{k} \frac{(\sum X)_i^2}{N_i}$$

$$SS_T = \sum_{i=1}^{k} \left(\sum X^2\right)_i - \frac{\left(\sum_{i=1}^{k} \left(\sum X\right)_i\right)^2}{\sum_{i=1}^{k} N_i}$$

$$df_W = \sum N_i - k$$

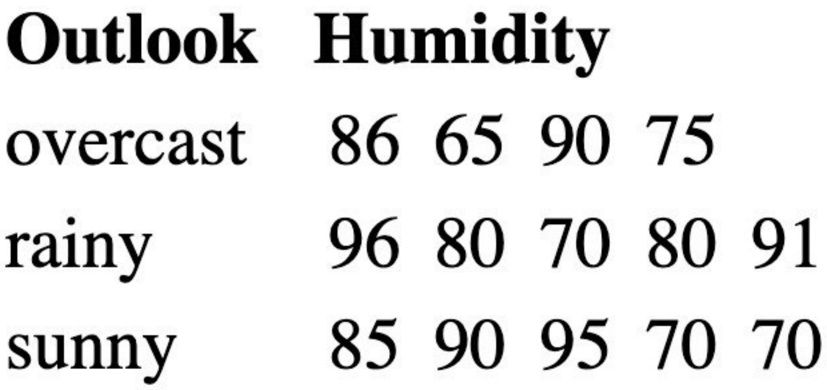
$$df_B = k - 1$$

$$df_T = \sum N_i - 1$$

F : has F distribution with df_B and df_w degree of freedom

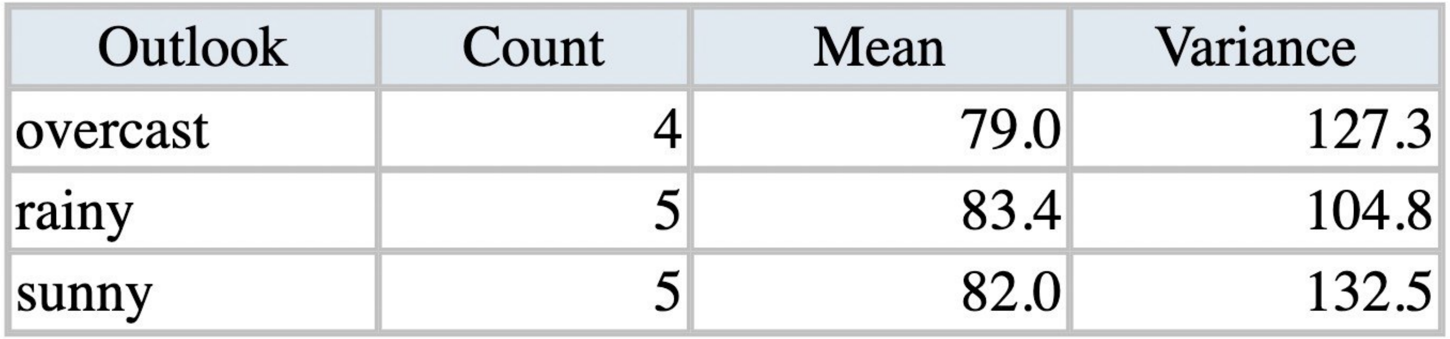
Example

- Is there a significant difference between the averages of the numerical variable (Humidity) in the three categories of the categorical variable (Outlook)?

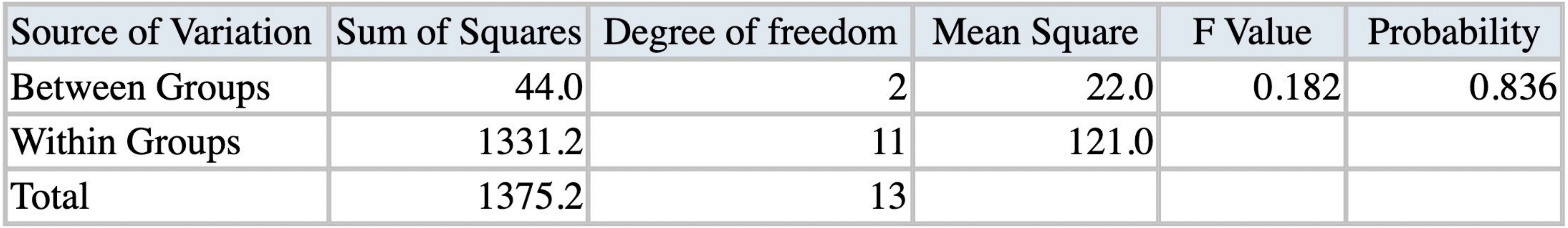


Outlook	Humidity				
overcast	86	65	90	75	
rainy	96	80	70	80	91
sunny	85	90	95	70	70

Example



Outlook	Count	Mean	Variance
overcast	4	79.0	127.3
rainy	5	83.4	104.8
sunny	5	82.0	132.5



Source of Variation	Sum of Squares	Degree of freedom	Mean Square	F Value	Probability
Between Groups	44.0	2	22.0	0.182	0.836
Within Groups	1331.2	11	121.0		
Total	1375.2	13			

There is no significant difference between the averages of Humidity in the three categories of Outlook.

Practice