UNDERSTANDING RELATIONSHIPS.

Chapter 4

What about the relationship between variables?

- A critical step in making sense of data is **an understanding of the relationships** between different variables.
- For example: Is there a relationship between
  - Interest rates and inflation or education level and income?
- The existence of an **association between variables does not imply** that one variable **causes** another.
- These **relationships or associations** can be established through an examination of different **summary tables** and **data visualizations** as well as **calculations** that **measure the strength and confidence** in the relationship.
- Ways to **understand relationships** between **pairs of variables** through data visualizations, tables that summarize the data, and specific calculated metrics
- **Each approach** is driven by **how the variables** have been **categorized** such as the scale on which they are measured.

VISUALIZING RELATIONSHIPS BETWEEN VARIABLES

Scatterplots

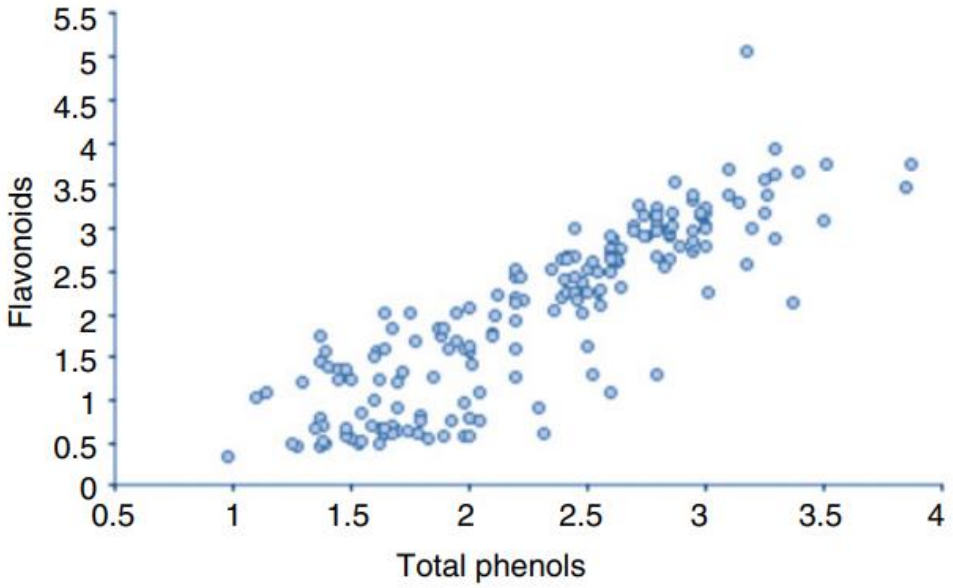


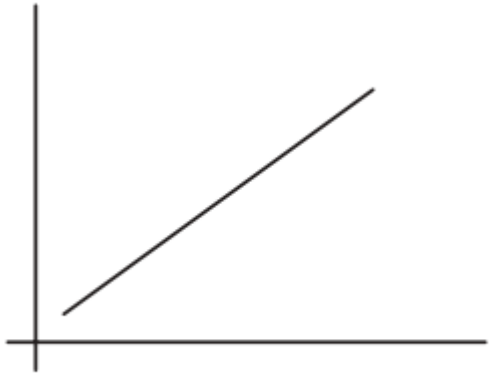
FIGURE 4.1 Example of a scatterplot where each point corresponds to an observation.

- Used to **identify whether a relationship exists** between two **continuous variables** measured on the **ratio or interval scales**.
- **Two variables** are plotted on **the x-and y-axis**.
- **Each poin**t displayed on the scatterplot is **a single observation.**
- **Position** of the point is determined by **value of the two variables**

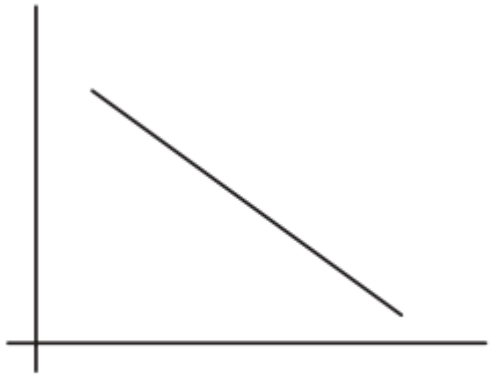
VISUALIZING RELATIONSHIPS BETWEEN VARIABLES

- **Relationship between variables can be complex**; however, a number of **characteristics of the relationship** can be measured:
  1. **Direction:**
     - In comparing two variables, **a positive relationship** results when **higher values** in the **first variable coincide** with **higher values** in the **second variable** and vice versa.
     - **Negative relationships** result when **higher values** in the **first variable coincide** with **lower values** in the **second** variable and vice versa.
     - Possible situations where relationship between variables is more complex, having c**ombination of positive and negative relationships** at various points.
  2. **Shape:**
     - A **relationship** is **linear** when it is **drawn as a straight line:** As values for one variable change, the second variable changes proportionally.
     - A **non linear relationship** is drawn as a curve indicating that as the first variable changes, the change in the second variable is not proportional.

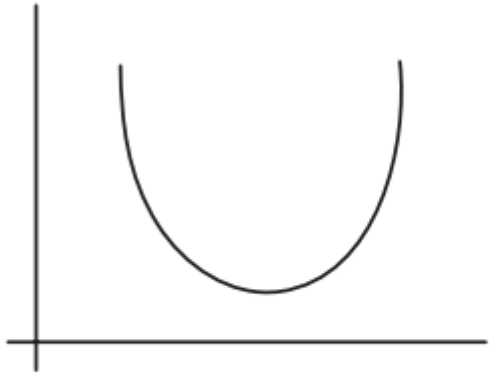
VISUALIZING RELATIONSHIPS BETWEEN VARIABLES



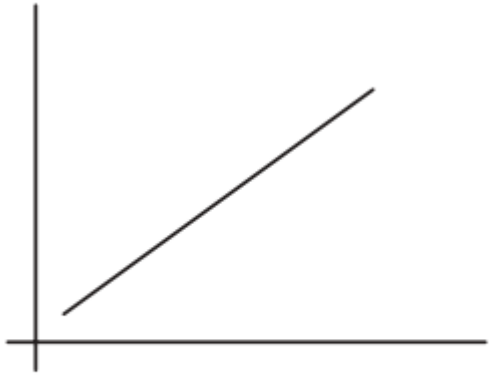
Positive



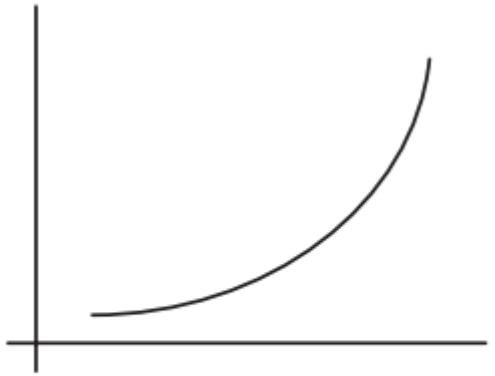
Negative



Both positive and negative at various points



Linear



Nonlinear

VISUALIZING RELATIONSHIPS BETWEEN VARIABLES

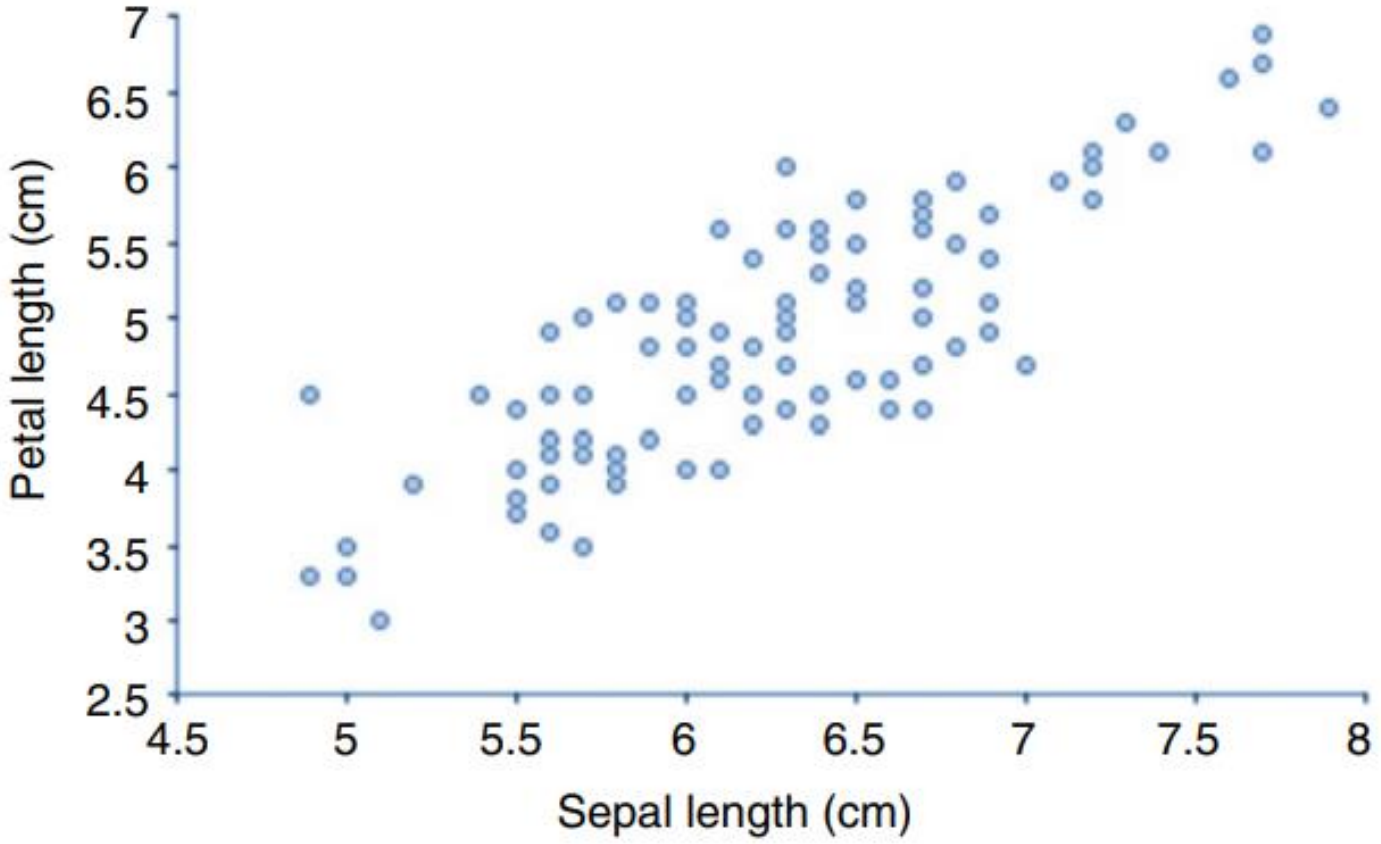


FIGURE 4.2 A scatterplot showing a positive relationship.

Iris flower data set or Fisher's Iris data set

Relationship is primarily **linear:**

As **sepal length (cm)** increases,
petal length (cm) increases proportionally

VISUALIZING RELATIONSHIPS BETWEEN VARIABLES

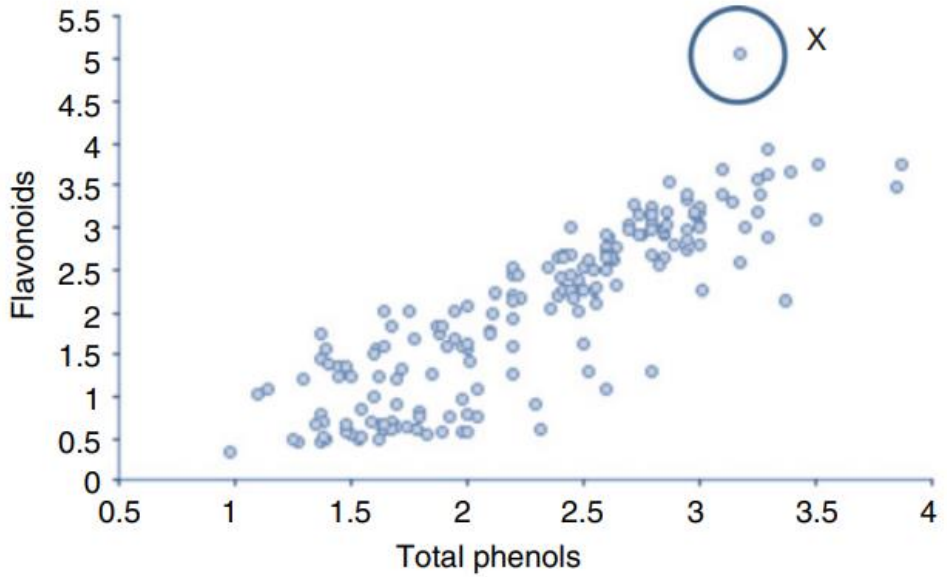


FIGURE 4.3 Observation (marked as X) that does not follow the relationship.

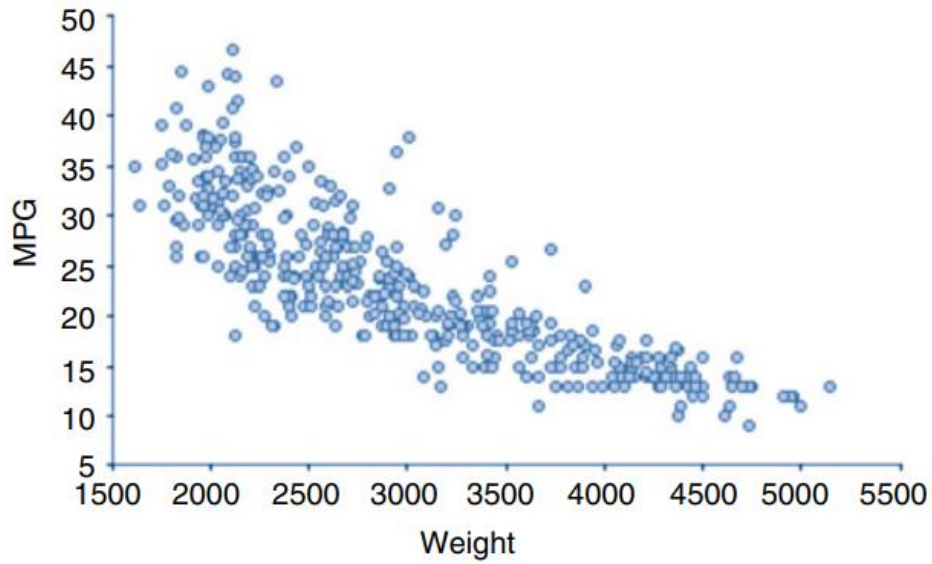


FIGURE 4.4 Scatterplot showing a negative nonlinear relationship.

Points that **do not follow this linear relationship.**
Outliers

A curve indicating that there is also a **nonlinear relationship** between the two variables
As **weight increases** MPG decreases, but the **rate of decrease** is **not proportional**.

VISUALIZING RELATIONSHIPS BETWEEN VARIABLES

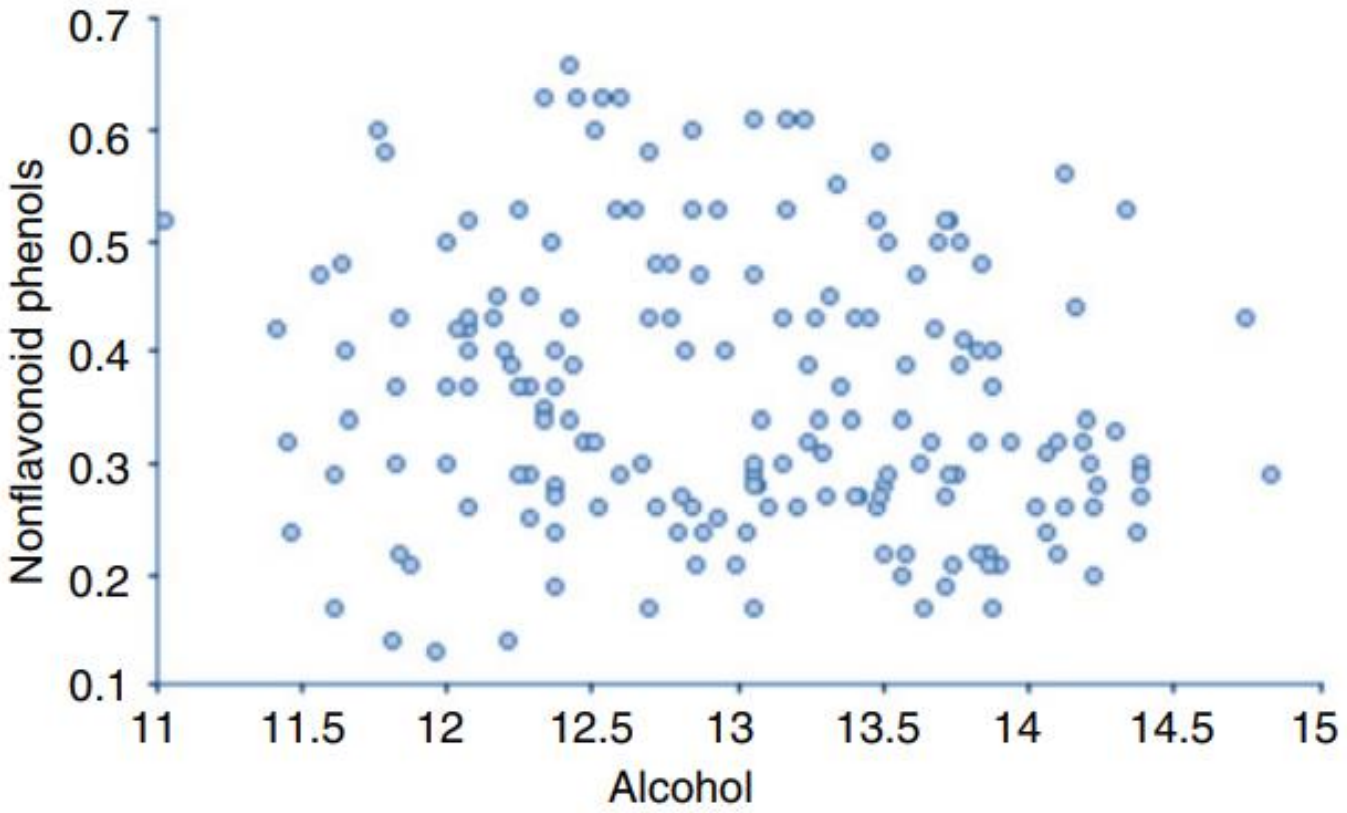
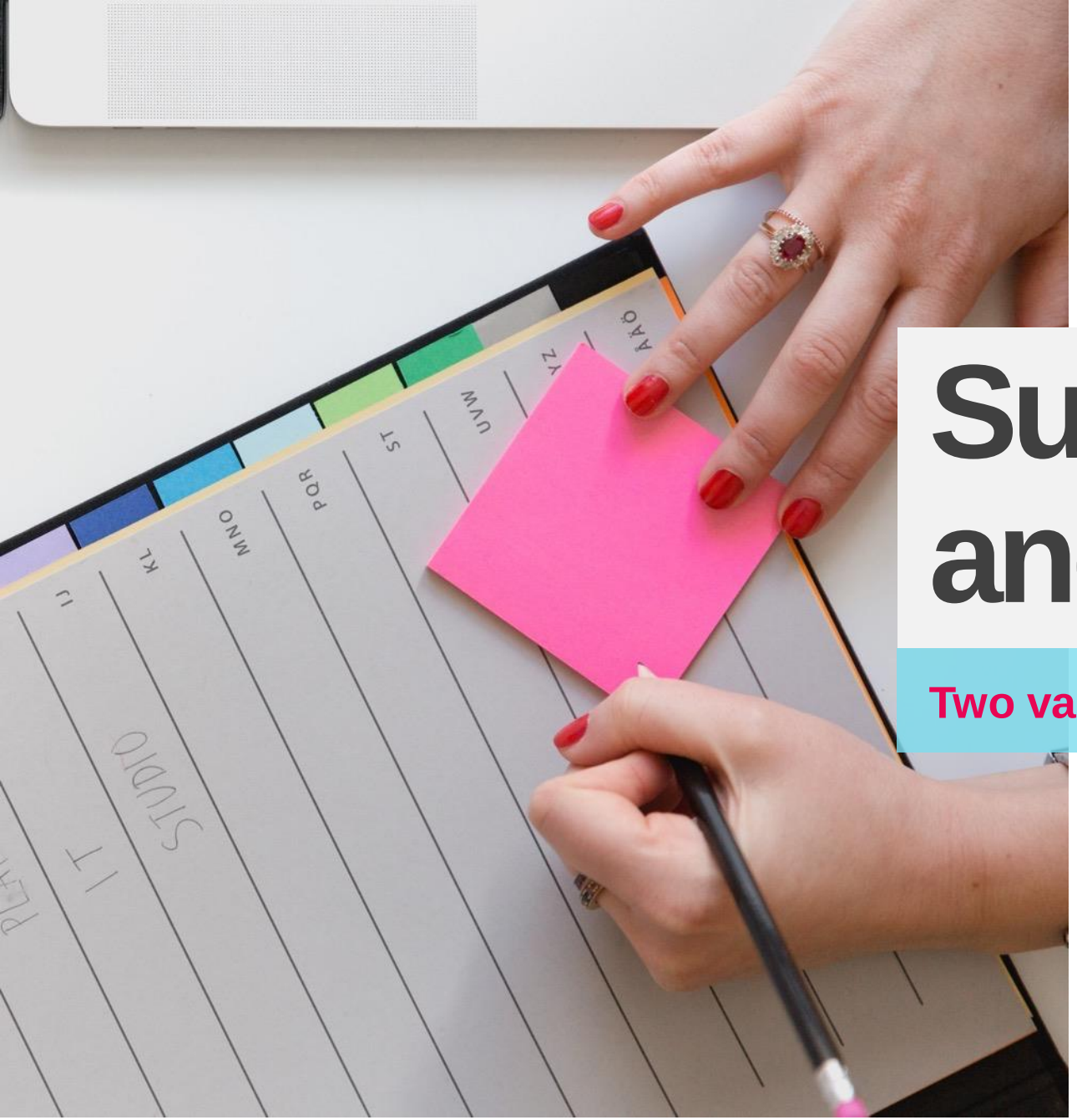


FIGURE 4.5 Scatterplot showing no relationships.

Indicates that there is **no obvious relationship** between Alcohol and Nonflavonoid phenols in this data set



Summary Tables and Charts

Two variables | **At least one of the variables is discrete**

AFTER A BREAK.

Next class.

Summary table.

- Common way of **understanding the relationship** between **two variables** where **at least one** of the variables is **discrete.**

For example:
A national retail company collected information on the **sale of individual products** for every store.

To **summarize** performance of these stores, they wish to generate a **summary table** to communicate the **average sales per store.**

TABLE 2.6 Retail Transaction Data Set

Customer	Store	Product Category	Product Description	Sale Price ($)	Profit ($)
B. March	New York, NY	Laptop	DR2984	950	190
B. March	New York, NY	Printer	FW288	350	105
B. March	New York, NY	Scanner	BW9338	400	100
J. Bain	New York, NY	Scanner	BW9443	500	125
T. Goss	Washington, DC	Printer	FW199	200	60
T. Goss	Washington, DC	Scanner	BW39339	550	140
L. Nye	New York, NY	Desktop	LR21	600	60
L. Nye	New York, NY	Printer	FW299	300	90
S. Cann	Washington, DC	Desktop	LR21	600	60
E. Sims	Washington, DC	Laptop	DR2983	700	140
P. Judd	New York, NY	Desktop	LR22	700	70
P. Judd	New York, NY	Scanner	FJ3999	200	50
G. Hinton	Washington, DC	Laptop	DR2983	700	140
G. Hinton	Washington, DC	Desktop	LR21	600	60
G. Hinton	Washington, DC	Printer	FW288	350	105
G. Hinton	Washington, DC	Scanner	BW9443	500	125
H. Fu	New York, NY	Desktop	ZX88	450	45
H. Taylor	New York, NY	Scanner	BW9338	400	100

Summary table.

Class	Count	Minimum (petal width (cm))	Maximum (petal width (cm))	Mean (petal width (cm))	Median (petal width (cm))	Standard deviation (petal width (cm))
Iris-setosa	50	0.1	0.6	0.244	0.2	0.107
Iris-versicolor	50	1	1.8	1.33	1.3	0.198
Iris-virginica	50	1.4	2.5	2.03	2	0.275

FIGURE 4.6 Example of a summary table.

- A **single categorical variable** (or a ***continuous variable converted into categories***)
  - is **used to group the observations**, and
  - **each row** of the table **represents a single group.**
- **Summary tables** will often show:
  - **Count** of the number of observations (or percentage) that has the particular value (or range).
  - **Descriptive statistics** that summarize a set of observations can be used including **mean, median, mode, sum, minimum, maximum, variance**, and **standard deviation**.
- Summary tables provide a way to **visualize data:**
  - Yes, it's a table, but by aggregating and summarizing information from a large data set, summary tables allow you to see things in the data you might otherwise not see.
- **Summary tables** allow you to **manipulate and create new data.**

https://archive.ics.uci.edu/ml/datasets/iris

Iris Data Set

Download: Data Folder, Data Set Description

Abstract: Famous database; from Fisher, 1936

Source:

Creator:

R.A. Fisher

Data Set Information:

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Predicted attribute: class of iris plant.

```
4.8,3.0,1.4,0.3,Iris-setosa
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
```

```
Summary Statistics:
                 Min  Max   Mean    SD   Class Correlation
   sepal length: 4.3  7.9   5.84  0.83    0.7826
    sepal width: 2.0  4.4   3.05  0.43   -0.4194
   petal length: 1.0  6.9   3.76  1.76    0.9490  (high!)
    petal width: 0.1  2.5   1.20  0.76    0.9565  (high!)
```

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

Class	Count	Minimum (petal width (cm))	Maximum (petal width (cm))	Mean (petal width (cm))	Median (petal width (cm))	Standard deviation (petal width (cm))
Iris-setosa	50	0.1	0.6	0.244	0.2	0.107
Iris-versicolor	50	1	1.8	1.33	1.3	0.198
Iris-virginica	50	1.4	2.5	2.03	2	0.275

FIGURE 4.6 Example of a summary table.

- Figure 4.6: **Relationship between two variables class** and **petal width (cm)** (Fisher, 1936).
- The **class** variable is a **discrete variable (nominal)** (values: “Iris-setosa,” “Iris-versicolor,” and “Iris-virginica”; first column)
- **50 observations** corresponding to each of these values
- **Each row** of the table **describes** the corresponding **50 observations**.
- **Each row** is populated with **summary information** about the **second variable (petal width (cm))** for the set of 50 observations.
- Example:
  - Minimum and maximum values are shown alongside the mean, median, and standard deviation.
  - Notice that **class “Iris-setosa”** is **associated** with the **smallest petal width with a mean of 0.2**.
  - Set of 50 observations for the **class “Iris-versicolor”** has a **mean of 1.33**, and
  - **class “Iris-virginica”** has the **highest mean of 2.03**.

View as a graph.

Class	Count	Minimum (petal width (cm))	Maximum (petal width (cm))	Mean (petal width (cm))	Median (petal width (cm))	Standard deviation (petal width (cm))
Iris-setosa	50	0.1	0.6	0.244	0.2	0.107
Iris-versicolor	50	1	1.8	1.33	1.3	0.198
Iris-virginica	50	1.4	2.5	2.03	2	0.275

FIGURE 4.6 Example of a summary table.

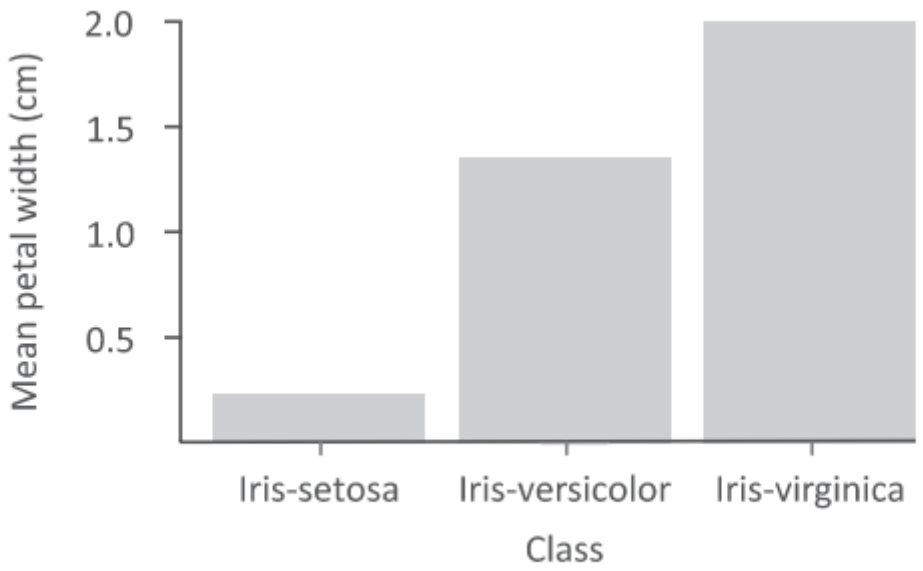


FIGURE 4.7 Example of a bar graph.

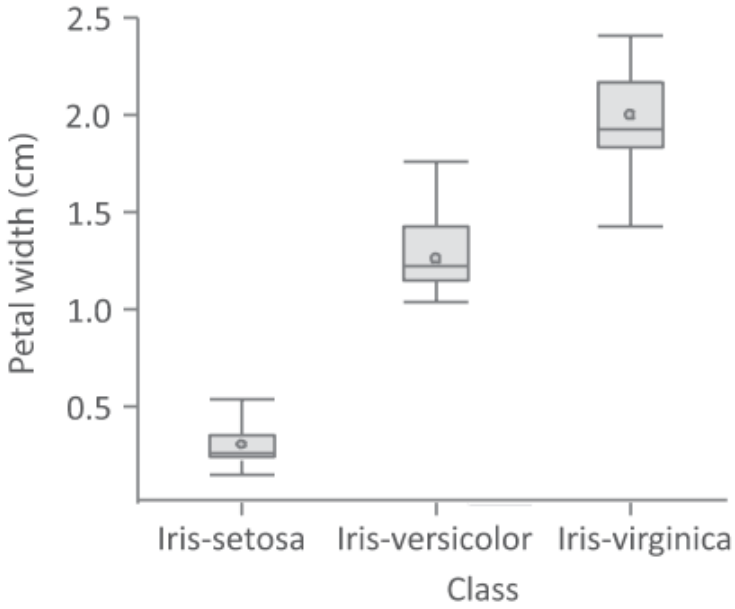


FIGURE 4.8 Example of a multiple box plot graph.

Ordinal variables.

MPG categories	Count	Mean (weight)
0–20 MPG	151	3832
20–30 MPG	151	2612
30–50 MPG	90	2157

FIGURE 4.9 Example of summary table where the categorical variable is ordinal.

- **Summary tables** also used to **show the relationship** between **ordinal variables and another variable**.
- Look at Figure 4.9:
  - **Three ordered categories** are used to group the observations.
  - Since the *categories can be ordered*, it is **possible to see** how **mean weight changes** as the **MPG category increases**.
  - It is clear from this table that as ***MPG categories increases*** the ***mean weight decreases.***

Ordinal variables.

MPG categories	Count	Mean (weight)
0–20 MPG	151	3832
20–30 MPG	151	2612
30–50 MPG	90	2157

FIGURE 4.9 Example of summary table where the categorical variable is ordinal.

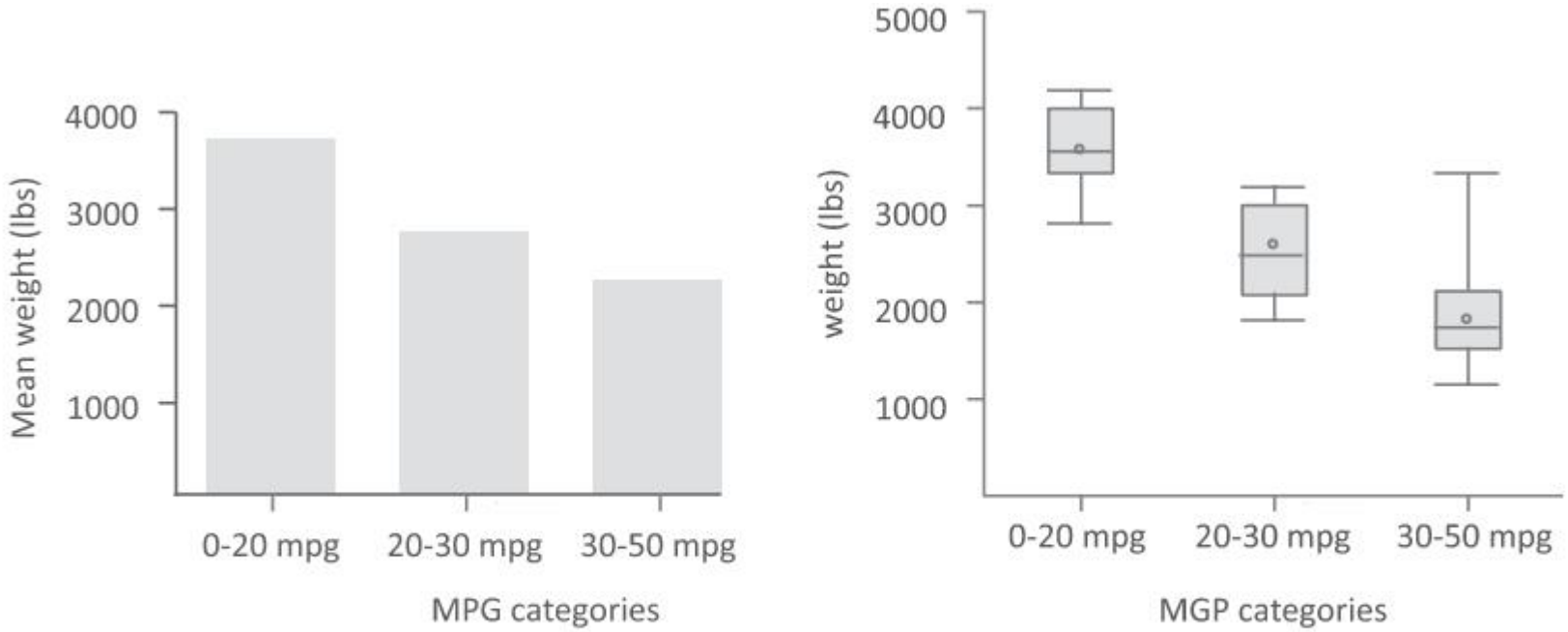
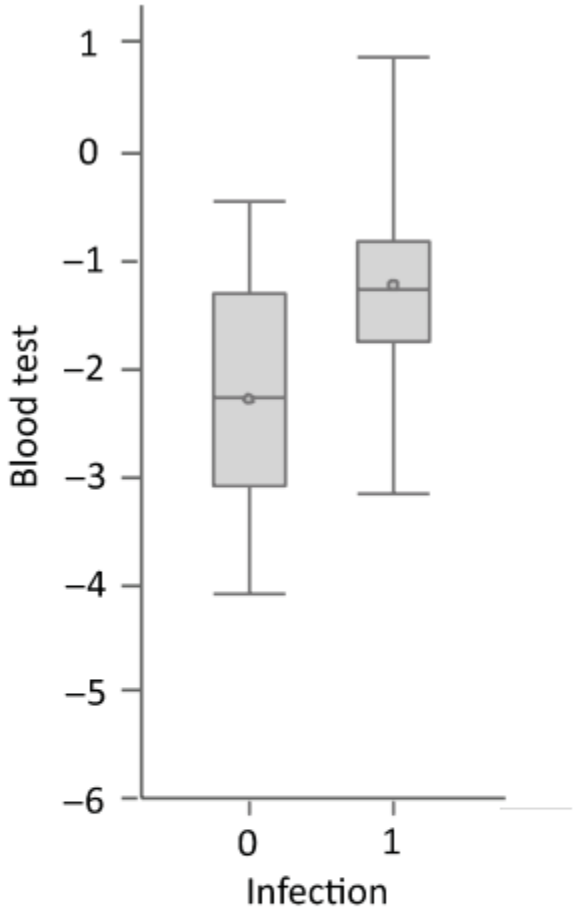


FIGURE 4.10 Graph of summary data for an ordinal variable against a continuous (weight).

Type of infection.

- **Binary variable** is used to represent a variable with two possible values, with **0 representing one value** and **1 the other.**
- For example,
  - **0:** case where a patient has a specific infection, and
  - 1: case where the patient does not.



Problem: New blood test is being investigated to predict whether a patient has a specific type of infection

Infection	Count	Minimum (blood test)	Maximum (blood test)	Mean (blood test)	Standard deviation (blood test)
0	23	−4.13	−0.393	−2.18	1.04
1	32	−3.05	0.893	−1.09	0.834

FIGURE 4.11 Summary table and corresponding box plot chart to summarizing the results of a trial for a new blood test to predict an infection.

Type of infection.

Blood test ranges	Count	Mean (Infection)
–5 – –4	2	0.02
–4 – –3	6	0.06
–3 – –2	9	0.19
–2 – –1	15	0.55
–1 – 0	11	0.85
0 – 1	9	0.95
1 – 2	3	1.0

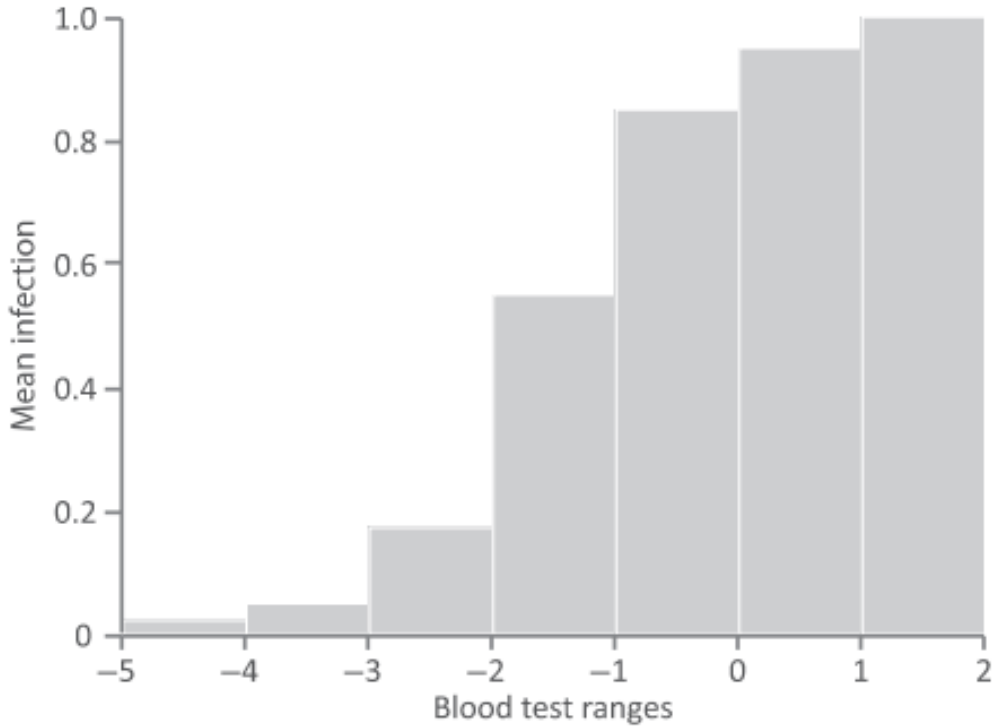


FIGURE 4.12 Summary table and histogram using continuous data that has been binned to generate the group summarized using the binary variable.

Cross-Classification Tables.

Provide insight into the **relationship between two categorical variables** or **non-categorical variables transformed** to **categorical variables.**

Infection Class

Test Results	Infection negative	Infection positive	Totals
Blood test negative	17	10	27
Blood test positive	6	22	28
Totals	23	32	55

FIGURE 4.13 Contingency table showing the relationship between two dichotomous variables.

- A **variable** is often **dichotomous**; however, a contingency table can represent **variables with more than two values.**
- Look at Figure 4.13: Example of a contingency table for two variables over a series of patients:
  - **Test Results** and **Infection Class.**
  - Variable **Infection Class** identifies whether a patient has the specific infection (two possible values (**"Infection negative"** and **"Infection positive"**).
  - Corresponding variable **Test Results** identified whether the blood test results were **positive** ("Blood test positive") or **negative** ("Blood test negative").

Contingency table for two Categorical variables over a series of patients: Test Results and Infection Class

Infection Class

	Infection negative	Infection positive	Totals
Blood test negative	17	10	27
Blood test positive	6	22	28
Totals	23	32	55

Test Results

FIGURE 4.13 Contingency table showing the relationship between two dichotomous variables.

Totals for the **Test Results** value

Data set had 55 observations

Totals for **Infection Class**

Shows **number of patients** that **correspond** to **pairs of values.**

Contingency tables: View of **relationship between two categorical variables**.

- New blood test **did not perfectly identify** presence or absence of the infection.
  - **Correctly classified** presence of the infection for **39 patients** (22 + 17)
    - **Incorrectly classified** the infection in **16 patients** (10 + 6).

Contingency tables used to understand relationships between categorical (both nominal and ordinal) variables where there are more than two possible values.

Gender

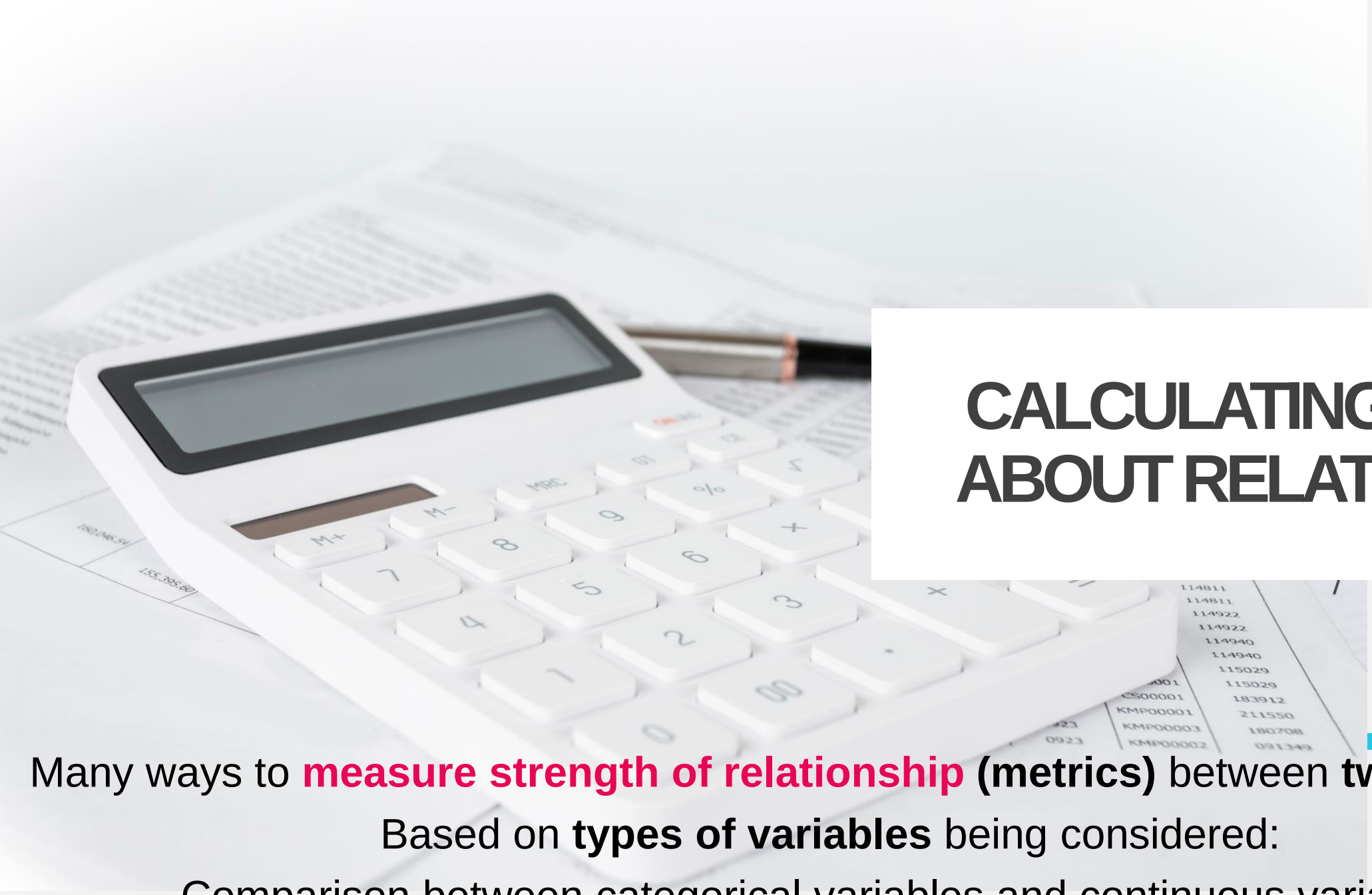
Dichotomous

	Male	Female	Totals
10–19	847	810	1657
20–29	4878	3176	8054
30–39	6037	2576	8613
40–49	5014	2161	7175
50–59	3191	1227	4418
60–69	1403	612	2015
70–79	337	171	508
80–89	54	24	78
90–99	29	14	43
Total	21,790	10,771	32,561

Age-group

Nine categories

Data set had **32,561** observations



CALCULATING METRICS ABOUT RELATIONSHIPS.

Many ways to **measure strength of relationship** **(metrics)** between **two variables**.

Based on **types of variables** being considered:

Comparison between categorical variables and continuous variables.

Correlation Coefficients.

- For **pairs of variables** measured on an interval or ratio scale, **a correlation coefficient (r)** can be calculated.
- This value **quantifies** the **linear relationship** between the variables by generating values from **−1.0 to +1.0.**
- If the **optimal straight line is drawn** through the **points on a scatterplot**, the value of **r** reflects **how closely** the points **lie to this line**.

"r positive": indicates a positive correlation between pair of variables, and
"r negative": indicates a negative correlation.
"r close to 0": indicates little or no relationship between the variables.

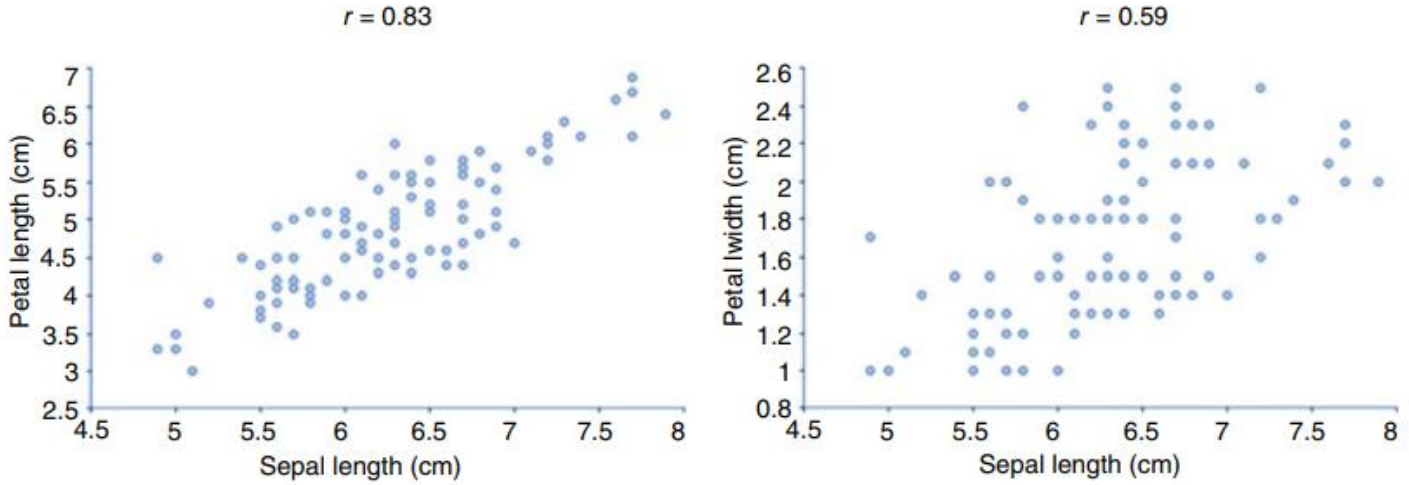


FIGURE 4.15 Scatterplots illustrate values for the correlation coefficient (*r*).

strong positive correlation

Weaker correlation

TABLE 4.2 Table Showing the Calculation of the Correlation Coefficient

x_i	y_i
92	6.3
145	7.8
30	3
70	5.5
75	6.5
105	5.5
110	6.5
108	8
45	4
50	5
160	7.5
155	9
180	8.6
190	10
63	4.2
85	4.9
130	6
132	7

The formula used to calculate r is shown here:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

where **x** and **y** are variables,
$\mathbf{x_i}$ are the individual values of x,
y_i are the individual values of y
$\bar{x}$ is the mean of the x variable,
$\bar{\mathbf{y}}$ is the mean of the y variable
s_x and s_y are the standard deviations of x and y, and
n is the number of observations.

TABLE 4.2 Table Showing the Calculation of the Correlation Coefficient

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
92	6.3	−14.94	−0.11	1.64
145	7.8	38.06	1.39	52.90
30	3	−76.94	−3.41	262.37
70	5.5	−36.94	−0.91	33.62
75	6.5	−31.94	0.09	−2.87
105	5.5	−1.94	−0.91	1.77
110	6.5	3.06	0.09	0.28
108	8	1.06	1.59	1.69
45	4	−61.94	−2.41	149.28
50	5	−56.94	−1.41	80.04
160	7.5	53.06	1.09	58.07
155	9	48.06	2.59	124.68
180	8.6	73.06	2.19	160.00
190	10	83.06	3.59	298.19
63	4.2	−43.94	−2.21	97.11
85	4.9	−21.94	−1.51	33.13
130	6	23.06	−0.41	−9.45
132	7	25.06	0.59	14.79

Sum = 1,357.06

$\bar{x} = 106.94$

$\bar{y} = 6.41$

$s_x = 47.28 \quad s_y = 1.86$

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

$$r = \frac{1357.06}{(18-1)(47.28)(1.86)}$$

$$r = 0.91$$

Pearson's Correlation Coefficient

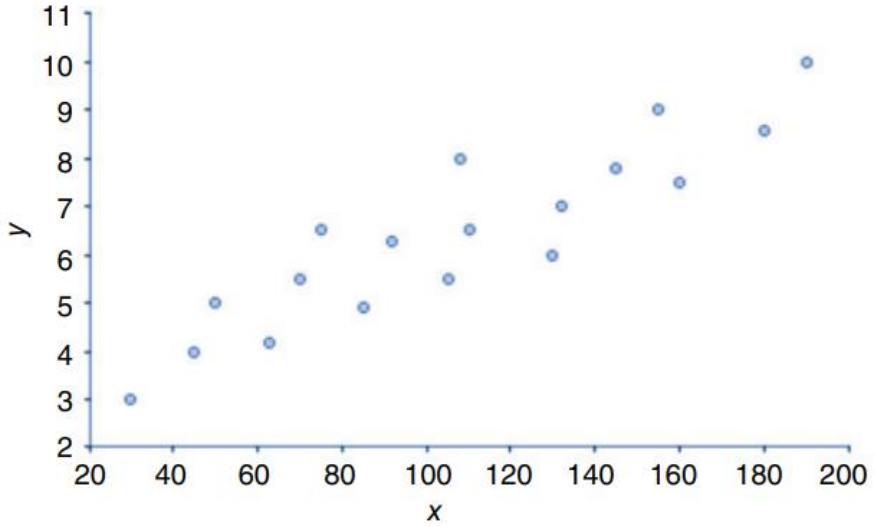


FIGURE 4.16 Scatterplot to illustrate the relationship between the x and variables.

After the BREAK.

Next class: Exercises

Exercises.

1. Generate a **contingency table** summarizing the variables **Store** and **Product** category.
2. Generate the following **summary tables:**
   a. **Grouping by Customer** with a count of the number of observations and the sum of **Sale price ($)** for each row.
   b. **Grouping by Store** with a count of the number of observations and the mean **Sale price ($)** for each row.
   c. **Grouping by Product** category with a count of the number of observations and the sum of the **Profit ($)** for each row.
3. Create a **scatterplot** showing Sales **price ($)** against **Profit ($)**

TABLE 2.6 Retail Transaction Data Set

Customer	Store	Product Category	Product Description	Sale Price ($)	Profit ($)
B. March	New York, NY	Laptop	DR2984	950	190
B. March	New York, NY	Printer	FW288	350	105
B. March	New York, NY	Scanner	BW9338	400	100
J. Bain	New York, NY	Scanner	BW9443	500	125
T. Goss	Washington, DC	Printer	FW199	200	60
T. Goss	Washington, DC	Scanner	BW39339	550	140
L. Nye	New York, NY	Desktop	LR21	600	60
L. Nye	New York, NY	Printer	FW299	300	90
S. Cann	Washington, DC	Desktop	LR21	600	60
E. Sims	Washington, DC	Laptop	DR2983	700	140
P. Judd	New York, NY	Desktop	LR22	700	70
P. Judd	New York, NY	Scanner	FJ3999	200	50
G. Hinton	Washington, DC	Laptop	DR2983	700	140
G. Hinton	Washington, DC	Desktop	LR21	600	60
G. Hinton	Washington, DC	Printer	FW288	350	105
G. Hinton	Washington, DC	Scanner	BW9443	500	125
H. Fu	New York, NY	Desktop	ZX88	450	45
H. Taylor	New York, NY	Scanner	BW9338	400	100

Solutions.

1. Contingency table summarizing the variables Store and Product category.

Product category	Store: NY	Store: DC	Total
Laptop	1	2	3
Printer	2	2	4
Scanner	4	2	6
Desktop	3	2	5
Total	10	8	18

TABLE 2.6 Retail Transaction Data Set

Customer	Store	Product Category	Product Description	Sale Price ($)	Profit ($)
B. March	New York, NY	Laptop	DR2984	950	190
B. March	New York, NY	Printer	FW288	350	105
B. March	New York, NY	Scanner	BW9338	400	100
J. Bain	New York, NY	Scanner	BW9443	500	125
T. Goss	Washington, DC	Printer	FW199	200	60
T. Goss	Washington, DC	Scanner	BW39339	550	140
L. Nye	New York, NY	Desktop	LR21	600	60
L. Nye	New York, NY	Printer	FW299	300	90
S. Cann	Washington, DC	Desktop	LR21	600	60
E. Sims	Washington, DC	Laptop	DR2983	700	140
P. Judd	New York, NY	Desktop	LR22	700	70
P. Judd	New York, NY	Scanner	FJ3999	200	50
G. Hinton	Washington, DC	Laptop	DR2983	700	140
G. Hinton	Washington, DC	Desktop	LR21	600	60
G. Hinton	Washington, DC	Printer	FW288	350	105
G. Hinton	Washington, DC	Scanner	BW9443	500	125
H. Fu	New York, NY	Desktop	ZX88	450	45
H. Taylor	New York, NY	Scanner	BW9338	400	100

Solutions.

2.a. Summary tables: **Grouping by Customer** with a count of the number of observations and the sum of Sale price ($) for each row

Customer	Number of Observations	Sum of Sales Price ($)
B. March	3	1700
J. Bain	1	500
T. Goss	2	750
L. Nye	2	900
S. Cann	1	600
E. Sims	1	700
P. Judd	2	900
G. Hinton	4	2150
H. Fu	1	450
H. Taylor	1	400

2.b. Summary tables: **Grouping by Store** with a count of the number of observations and the mean Sale price ($) for each row

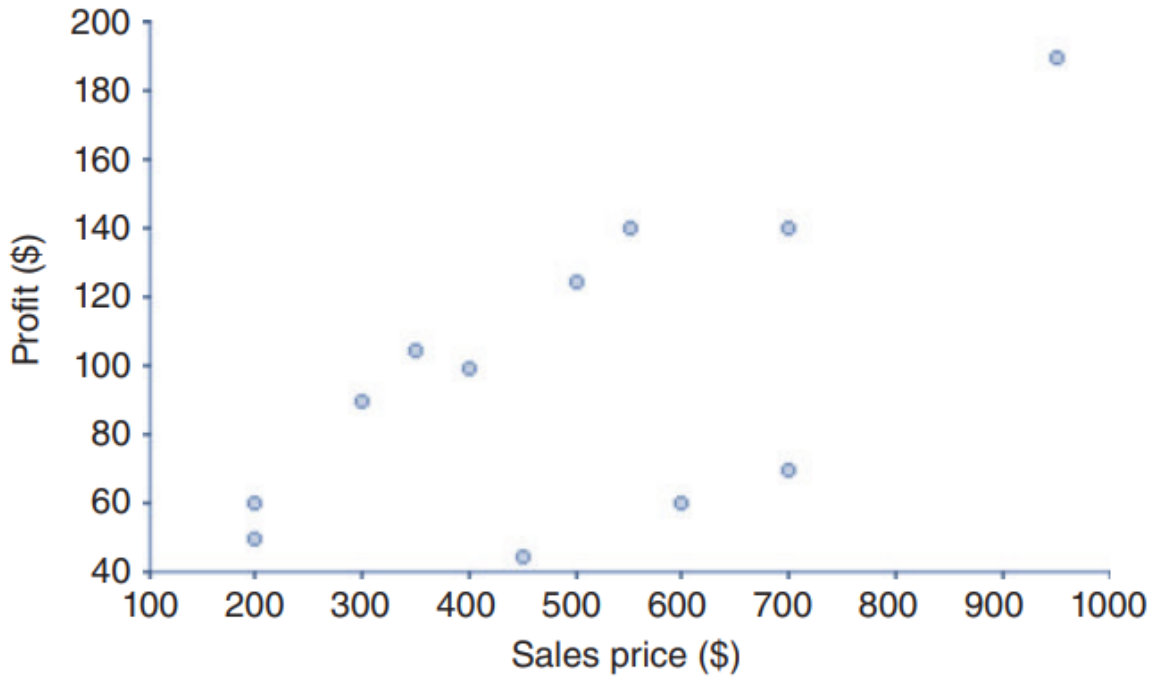
Store	Number of Observations	Mean Sale Price ($)
New York, NY	10	485
Washington, DC	8	525

Solutions.

2.c. Summary tables: **Grouping by Product category** with a count of the number of observations and the sum of the Profit ($) for each row.

Product Category	Number of Observations	Sum of Profit ($)
Laptop	3	470
Printer	4	360
Scanner	6	640
Desktop	5	295

3. Create a **scatterplot** showing Sales price ($) against Profit ($)



Shall we hypothesize?

t-Tests Comparing Two Groups

t-Tests Comparing Two Groups (Hypothesis Testing)

- Concept can be extended to compare the **mean values of two subsets:** Explore if **means of two groups are different enough** to say the **difference is significant** or conclude that a difference is simply **due to chance.**
- Look at difference between two groups: consider **mean values of the two groups**, and **deviation** of the data for the **two groups**.
- Formula:
  - Assumes that **value being assessed across the two groups** is both **independently** and **normally distributed** and **variances between the two groups** are either **equal or similar**.
  - Considers **difference between the two groups** and information concerning the **distribution of the two groups:**

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\bar{x}_1$ is the mean value of the first group

$\bar{x}_2$ is the mean value of the second group, and

n_1, n_2 are the number of observations in the first and second group respectively, and

s_p is an estimate of the standard deviation (pooled estimate).

t-Tests Comparing Two Groups.

The s_p is calculated using the following formula: $$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- where n1 and n2 are number of observations in group 1 and group 2, and
  s_1^2 1, s_2^2 are the calculated variances for group 1 and group 2.
- This **formula** follows **a t-distribution**, with the number of degrees of freedom (df) calculated as $$df = n_1 + n_2 - 2$$

  where it ***cannot be assumed that the variances across the two groups are equal***
- Another formula is used: $$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
- where $\bar{x}_1$ and $\bar{x}_2$ are the average values for the two groups, s_1^2 and s_2^2 are the calculated variances for two groups, and n_1 and n_2 are number of observations in the two groups.

t-Tests Comparing Two Groups.

- Again, it follows a **t-distribution** and the **number of degrees of freedom (df)** is calculated using the following formula:

$$\mathrm{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

- **t-values:**
  - **Positive** if **mean of group 1** is **larger than** **mean of group 2**
  - **Negative** if **mean of group 2** is **larger than** **mean of group 1**.
- These t-values can be used in a hypothesis test where the null hypothesis states that the two means are equal and the alternative hypothesis states that the two means are not equal. This t-value can be used to accept or reject the null hypothesis as well as calculate a p-value.

Chi-square.

Hypothesis test for **use nominal or ordinal scale** variables.
Allows analysis of whether there is a **relationship between** two categorical variables.

A **hypothesis test**, necessary to **state a null and alternative hypothesis.**

H_0: There is no relationship.
H_a: There is a relationship

TABLE 4.7 Calculation of Chi-Square

k	Category	Observed (O)	Expected (E)	$(O-E)^2/E$
1	r = Brand X, c = 43221	5,521	4,923	72.6
2	r = Brand Y, c = 43221	4,597	4,913	20.3
3	r = Brand Z, c = 43221	4,642	4,925	16.3
4	r = Brand X, c = 43026	4,522	4,764	12.3
5	r = Brand Y, c = 43026	4,716	4,754	0.3
6	r = Brand Z, c = 43026	5,047	4,766	16.6
7	r = Brand X, c = 43212	4,424	4,780	26.5
8	r = Brand Y, c = 43212	5,124	4,770	26.3
9	r = Brand Z, c = 43212	4,784	4,782	0.0008
				Sum = 191.2

- Using Table 4.7, look at whether a **relationship exists** between where a **consumer lives** (represented by a zip code) and the **brand of washing powder** they buy (brand X, brand Y, and brand Z). “r” and “c” refer to the row (r) and column (c) in a contingency table.
- **Chi-Square test compares** the **observed frequencies** with the **expected frequencies.**
- The expected frequencies are calculated using the following formula: $E_{r,c}$:

$$E_{r,c} = \frac{r \times c}{n}$$

where $E_{r,c}$ is the expected frequency for a particular cell in a contingency table, r is the row count, c is the column count and n is the total number of observations in the sample.

Chi-square.

Test for independence

- To calculate **expected frequency** for the table cell where the washing powder is brand X and the zip code is 43221 would be:

$$E_{\text{Brand\_X,43221}} = \frac{14{,}467 \times 14{,}760}{43{,}377}$$

$$E_{\text{Brand\_X,43221}} = 4{,}923$$

TABLE 4.7 Calculation of Chi-Square

k	Category	Observed (O)	Expected (E)	$(O-E)^2/E$
1	r = Brand X, c = 43221	5,521	4,923	72.6
2	r = Brand Y, c = 43221	4,597	4,913	20.3
3	r = Brand Z, c = 43221	4,642	4,925	16.3
4	r = Brand X, c = 43026	4,522	4,764	12.3
5	r = Brand Y, c = 43026	4,716	4,754	0.3
6	r = Brand Z, c = 43026	5,047	4,766	16.6
7	r = Brand X, c = 43212	4,424	4,780	26.5
8	r = Brand Y, c = 43212	5,124	4,770	26.3
9	r = Brand Z, c = 43212	4,784	4,782	0.0008
				Sum = 191.2

Computed $\chi 2$

The Chi-Square test (χ^2) is computed with the following equation:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where **k** is the number of all categories,
O_i is the observed cell frequency, and
E_i is the expected cell frequencyTable 4.7 shows the for this example.

The test is usually performed when all observed cell frequencies are greater than 10.

Chi-square.

- There is a **critical value at which the null hypothesis is rejected (χ2 c)**
- **Critical value** is found in a **standard Chi-Square table**

TABLE 4.7 Calculation of Chi-Square

k	Category	Observed (O)	Expected (E)	$(O-E)^2/E$
1	r = Brand X, c = 43221	5,521	4,923	72.6
2	r = Brand Y, c = 43221	4,597	4,913	20.3
3	r = Brand Z, c = 43221	4,642	4,925	16.3
4	r = Brand X, c = 43026	4,522	4,764	12.3
5	r = Brand Y, c = 43026	4,716	4,754	0.3
6	r = Brand Z, c = 43026	5,047	4,766	16.6
7	r = Brand X, c = 43212	4,424	4,780	26.5
8	r = Brand Y, c = 43212	5,124	4,770	26.3
9	r = Brand Z, c = 43212	4,784	4,782	0.0008
				Sum = 191.2

The value is dependent on the degrees of freedom (df), which is calculated: $df = (r-1) \times (c-1)$

For example, Number of degrees of freedom for the above example is (3 − 1) × (3 − 1) = 4.

Looking up the **critical value for df = 4 and α = 0.05**, the **critical value** is **9.488**.

Conclusion: 9.488 is less than the **calculated chi-square value of 191.2**,
We **reject the null hypothesis**
We state: **There is a relationship between zip codes and brands of washing powder.**

		Washing powder brand			
		Brand X	Brand Y	Brand Z	
Zip code	**43221**	5,521	4,597	4,642	**14,760**
	43029	4,522	4,716	5,047	**14,285**
	43212	4,424	5,124	4,784	**14,332**
		14,467	**14,437**	**14,473**	**43,377**

		Washing powder brand			
		Brand X	Brand Y	Brand Z	
Zip code	**43221**	4,923	4,913	4,925	**14,760**
	43026	4,764	4,754	4,766	**14,285**
	43212	4,780	4,770	4,782	**14,332**
		14,467	**14,437**	**14,473**	**43,377**

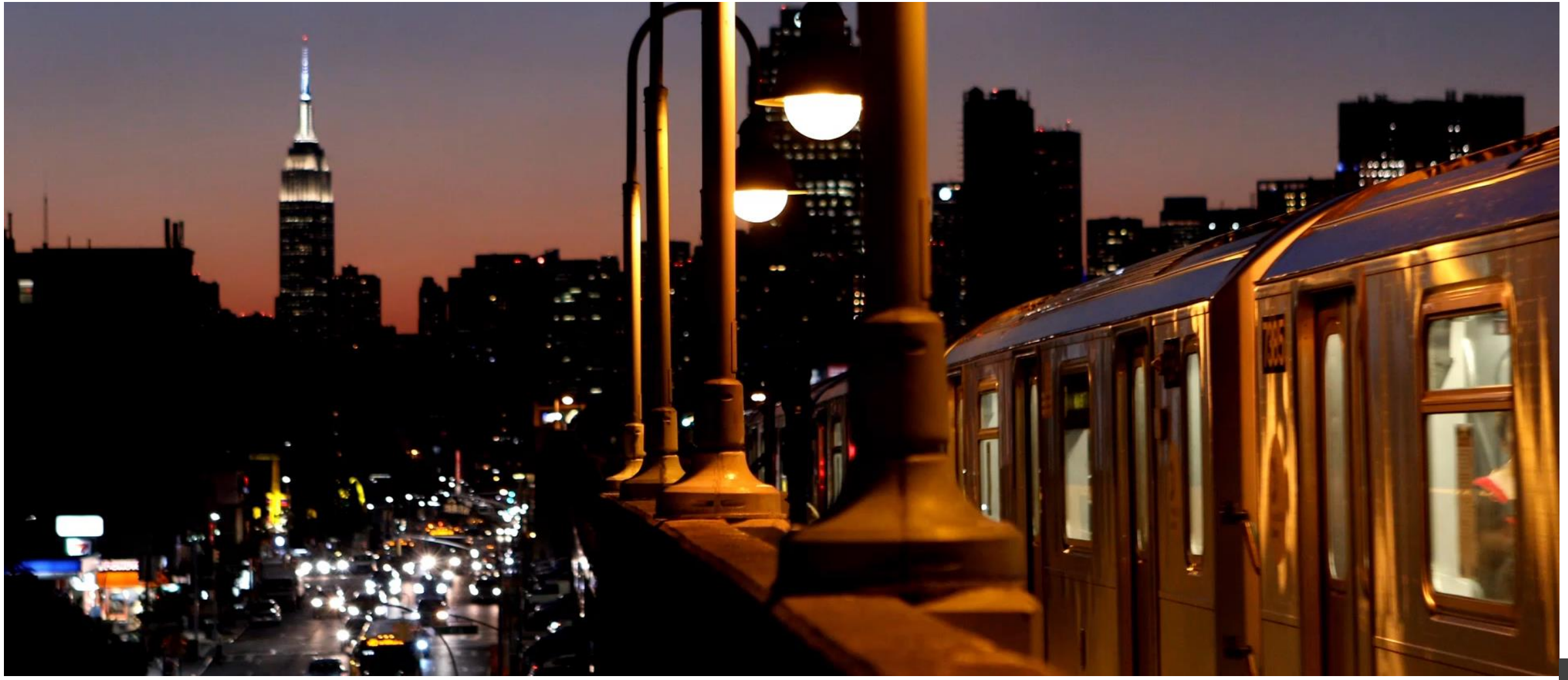
	Probability													
df	**0.99**	**0.98**	**0.95**	**0.90**	**0.80**	**0.70**	**0.50**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.001**
1	0.0^3157	0.0^3628	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345	16.266
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.467
5	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.515
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697

Looking up critical chi-square value, for $df = 4$ and $\alpha = 0.05$

Adapted from Table IV of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, sixth Edition, Pearson Education Limited, © 1963 R. A. Fisher and F. Yates

1963 R. A. Fisher and F. Yates

BACK from the BREAK



Is gender independent of education level? A random sample of 395 people was surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey is summarized in the following table:

Question: Are gender and education level dependent at **5% level** of significance?

Gender	Pre-Grad	Graduation	Post Graduation	PhD	Total
Male	60	54	46	41	**201**
Female	40	44	53	57	**194**
Total	**100**	**98**	**99**	**98**	**395**

df=(r-1)(c-1) = (1)(3)=4

Observed Values

Expected Values Values

Gender	Pre-Grad	Graduation	Post Graduation	PhD	Total
Male	50.63?				201
Female					194
Total	100	98	99	98	395

$$E_{r,c} = \frac{r \times c}{n}$$

$E_{1,1}$ = r (total) x c(total) / n = 50.63?

ANOVA.

ONE WAY

Completely randomized one-way Analysis of Variance (ANOVA) that compares the means of three or more different groups.

Determines whether **there is a difference** between the groups.

TABLE 4.5 Calls Processed by Different Call Centers

Call Center A	Call Center B	Call Center C	Call Center D
136	124	142	149
145	131	145	157
139	128	139	154
132	130	145	155
141	129	143	151
143	135	141	156
138	132	138	
139		146	

- This method can be applied to cases where the **groups are independent** and **random,** the **distributions are normal** and the **populations have similar variances.**
- Example: Table 4.5 Call centers are approximately the **same size** and handle a certain **number of calls** each day.
- An **analysis** of the **different call centers** based on the **average number of calls** processed each day is required to **understand whether one or more of the call centers** are **under- or over-performing**.

ANOVA.

ONE WAY

H_0: Means are equal.
H_a: Means are not equal.

This test looks at both the **variation** **within the groups** and **between the groups.**

TABLE 4.6 Calculating Means and Variances

	Call Center A	Call Center B	Call Center C	Call Center D	Groups ($k = 4$)
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		
Count	8	7	8	6	*Total count* $N = 29$
Mean	139.1	129.9	142.4	153.7	
Variance	16.4	11.8	8.6	9.5	

The test performs the **following steps:**

1. Calculate group means and variance.
2. Determine the within-group variation.
3. Determine the between-group variation.
4. Determine the F-statistic, which is based on the between-group and within group ratio.
5. Test the significance of the F-statistic

ANOVA.

ONE WAY

1. Calculate group means and variances.

In Table 4.6:

- A **count** along with the **mean and variance** for **each call center** has been calculated.
- **Total number of groups** (k = 4) and **total number of observations** (N = 29) listed.

TABLE 4.6 Calculating Means and Variances

	Call Center A	Call Center B	Call Center C	Call Center D	Groups ($k = 4$)
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		
Count	8	7	8	6	*Total count* $N = 29$
Mean	139.1	129.9	142.4	153.7	
Variance	16.4	11.8	8.6	9.5	

- An **average of all values (x = 140.8)** is calculated by summing all values and dividing it by the number of observations:

$$\bar{\bar{x}} = \frac{136 + 145 + \ldots + 151 + 156}{29} = 140.8$$

ANOVA.

ONE WAY

1. Calculate group means and variances.

$$\bar{\bar{x}} = \frac{136 + 145 + \ldots + 151 + 156}{29} = 140.8$$

2. Determine the within-group variation.

TABLE 4.6 Calculating Means and Variances

	Call Center A	Call Center B	Call Center C	Call Center D	Groups ($k = 4$)
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		
Count	8	7	8	6	*Total count* $N = 29$
Mean	139.1	129.9	142.4	153.7	
Variance	16.4	11.8	8.6	9.5	

- The variation within groups is **defined as the within-group variance or mean square within (MSW).**
- To calculate this value, weighted sum of the variance for the individual groups is used.
- The weights are based on the number of observations in each group.
- This sum is divided by the number of degrees of freedom calculated by subtracting the number of groups (k) from the total number of observations (N):

$$MSW = \frac{\sum_{i=1}^{k} (n_i - 1)s_i^2}{N - k}$$

ANOVA.

ONE WAY

1. Calculate group means and variances.

$$\bar{\bar{x}} = \frac{136 + 145 + \ldots + 151 + 156}{29} = 140.8$$

2. Determine the within-group variation.

TABLE 4.6 Calculating Means and Variances

	Call Center A	Call Center B	Call Center C	Call Center D	Groups ($k = 4$)
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		
Count	8	7	8	6	*Total count* $N = 29$
Mean	139.1	129.9	142.4	153.7	
Variance	16.4	11.8	8.6	9.5	

$$MSW = \frac{(8-1)\times 16.4 + (7-1)\times 11.8 + (8-1)\times 8.6 + (6-1)\times 9.5}{(29-4)}$$

$$MSW = 11.73$$

$$MSW = \frac{\sum_{i=1}^{k} (n_i - 1)s_i^2}{N - k}$$

ANOVA.

ONE WAY

1. Calculate group means and variances.

2. Determine the within-group variation.

3. Determine the between-group variation

TABLE 4.6 Calculating Means and Variances

	Call Center A	Call Center B	Call Center C	Call Center D	Groups ($k = 4$)
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		
Count	8	7	8	6	*Total count* $N = 29$
Mean	139.1	129.9	142.4	153.7	
Variance	16.4	11.8	8.6	9.5	

Next, the **between-group variation** or **mean square between (MSB)** is calculated.

MSB is the **variance between the group means:** Calculated using a weighted sum of the squared difference between the group mean (x̄i) and the average of all observations $(\bar{\bar{x}})$. This sum is divided by the number of degrees of freedom. This is calculated by subtracting one from the number of groups (k). The following formula is used to calculate the MSB:

$$MSB = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2}{k - 1}$$

where n_i is no. for each group and x̄i is the avg for each group

ANOVA.

ONE WAY

1. Calculate group means and variances.

2. Determine the within-group variation.

3. Determine the between-group variation

TABLE 4.6 Calculating Means and Variances

	Call Center A	Call Center B	Call Center C	Call Center D	Groups ($k = 4$)
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		
Count	8	7	8	6	*Total count* $N = 29$
Mean	139.1	129.9	142.4	153.7	
Variance	16.4	11.8	8.6	9.5	

$$MSB = \frac{\begin{array}{c}(8 \times (139.1 - 140.8)^2) + (7 \times (129.9 - 140.8)^2) \\ +(8 \times (142.4 - 140.8)^2) + (6 \times (153.7 - 140.8)^2)\end{array}}{4 - 1}$$

$$MSB = 624.58$$

ANOVA.

ONE WAY

1. Calculate group means and variances.

2. Determine the within-group variation.

3. Determine the between-group variation

4. Determine the F-statistic

TABLE 4.6 Calculating Means and Variances

	Call Center A	Call Center B	Call Center C	Call Center D	Groups ($k = 4$)
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		
Count	8	7	8	6	*Total count* $N = 29$
Mean	139.1	129.9	142.4	153.7	
Variance	16.4	11.8	8.6	9.5	

The F-statistic is the ratio of the MSB and the MSW:

$$F = \frac{MSB}{MSW}$$

ANOVA.

ONE WAY

1. Calculate group means and variances.

2. Determine the within-group variation.

3. Determine the between-group variation

4. Determine the F-statistic

TABLE 4.6 Calculating Means and Variances

	Call Center A	Call Center B	Call Center C	Call Center D	Groups ($k = 4$)
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		
Count	8	7	8	6	*Total count* $N = 29$
Mean	139.1	129.9	142.4	153.7	
Variance	16.4	11.8	8.6	9.5	

The F-statistic is the ratio of the MSB and the MSW:

$$F = \frac{MSB}{MSW}$$

$$F = \frac{624.58}{11.73}$$

$$F = 53.25$$

ANOVA.

ONE WAY

1. Calculate group means and variances.

2. Determine the within-group variation.

3. Determine the between-group variation

4. Determine the F-statistic

4. Test the significance of the F-statistic

TABLE 4.6 Calculating Means and Variances

	Call Center A	Call Center B	Call Center C	Call Center D	Groups ($k = 4$)
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		
Count	8	7	8	6	*Total count* $N = 29$
Mean	139.1	129.9	142.4	153.7	
Variance	16.4	11.8	8.6	9.5	

Determine the degrees of freedom **(df)** for the two mean squares (within and between). The degrees of freedom for the MSW (df_{within}) is calculated using the following formula:

$$df_{\text{within}} = N - k$$

where N is the total number of observations in all groups and k is the number of groups.

ANOVA.

ONE WAY

1. Calculate group means and variances.

2. Determine the within-group variation.

3. Determine the between-group variation

4. Determine the F-statistic

4. Test the significance of the F-statistic

TABLE 4.6 Calculating Means and Variances

	Call Center A	Call Center B	Call Center C	Call Center D	Groups ($k = 4$)
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		
Count	8	7	8	6	*Total count* $N = 29$
Mean	139.1	129.9	142.4	153.7	
Variance	16.4	11.8	8.6	9.5	

The degrees of freedom for the MSB ($df_{between}$) is calculated using the following formula:

$$df_{\text{between}} = k - 1$$

where k is the number of groups.

ANOVA.

ONE WAY

1. Calculate group means and variances.

2. Determine the within-group variation.

3. Determine the between-group variation

4. Determine the F-statistic

4. Test the significance of the F-statistic

TABLE 4.6 Calculating Means and Variances

	Call Center A	Call Center B	Call Center C	Call Center D	Groups ($k = 4$)
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		
Count	8	7	8	6	*Total count* $N = 29$
Mean	139.1	129.9	142.4	153.7	
Variance	16.4	11.8	8.6	9.5	

$$df_{\text{between}} = 4 - 1 = 3$$

$$df_{\text{within}} = 29 - 4 = 25$$

ANOVA.

ONE WAY

1. Calculate group means and variances.

2. Determine the within-group variation.

3. Determine the between-group variation

4. Determine the F-statistic

4. Test the significance of the F-statistic

TABLE 4.6 Calculating Means and Variances

	Call Center A	Call Center B	Call Center C	Call Center D	Groups ($k = 4$)
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		
Count	8	7	8	6	*Total count* $N = 29$
Mean	139.1	129.9	142.4	153.7	
Variance	16.4	11.8	8.6	9.5	

$$df_{\text{between}} = 4 - 1 = 3$$

$$df_{\text{within}} = 29 - 4 = 25$$

ANOVA.

ONE WAY

1. Calculate group means and variances.

2. Determine the within-group variation.

3. Determine the between-group variation

4. Determine the F-statistic

4. Test the significance of the F-statistic

TABLE 4.6 Calculating Means and Variances

	Call Center A	Call Center B	Call Center C	Call Center D	Groups ($k = 4$)
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		
Count	8	7	8	6	*Total count* $N = 29$
Mean	139.1	129.9	142.4	153.7	
Variance	16.4	11.8	8.6	9.5	

We already calculated the **F-statistic to be 53.39**.
Indicates that the **mean variation between groups is much greater than the mean variation within group**s due to errors.
To test this, look up the critical F-statistic from an F-table. To find this critical value we need α (confidence level), v1 (df between), and v2 (df within).
The critical value for the **F-statistic is 3.01** (**when α is 0.05**). Since the **calculated F-statistic is greater than the critical value, we reject the null hypothesis.** The means for **the different call centers are not equal**

Take a break. We go to the Next chapter.