Static formulas

Static formulas

Relative Frequency = Frequency/sum of all frequencies

Mean = sum of numbers / total numeric count

Median = if even = N+1/2. If odd N/2

Mode = Highest frequency variable.

Measure a Dispersion:

1. **Range = largest – smallest**
2. **STD (sigma) = ROOT (sum. square (x – mean) /N)**
3.
4. **Variance = Square (STD)**
5. **Z score = x observation – x\mean sample devide s (sigma) sample (x – mean/s)**

   Z score = x – x\sigma sample

$$z = \frac{x - \bar{x}}{s} = \frac{475 - 550}{75} = -1.0$$

STD <Mean -1.0

DEFINITION

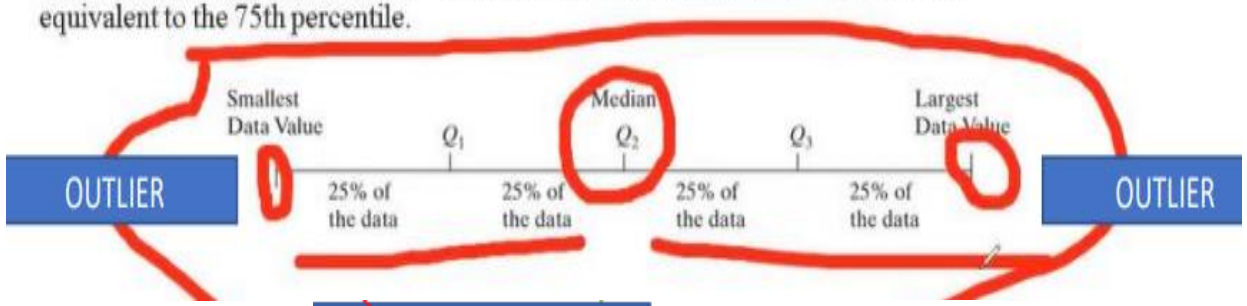
Quartiles are measures of location, denoted Q_1, Q_2, and Q_3, which divide a set of data into four groups with about 25% of the values in each group.

Quartiles Benefits :
1- Identify the exact data location
2- Identify the outlier position.

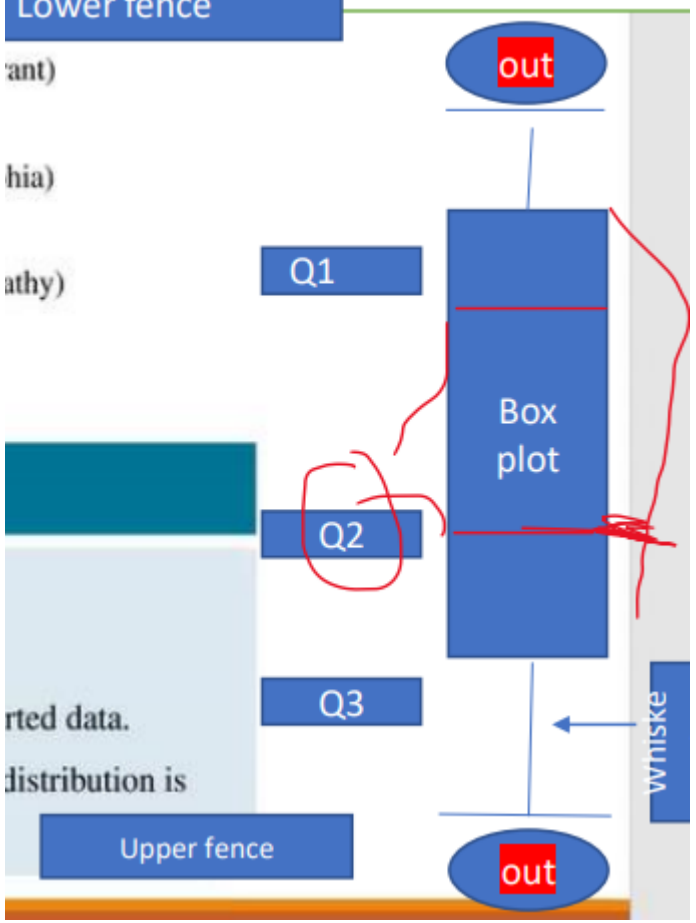
Duncan	Charlie		Grant	Aidan		Sophia	Seamus		Cathy	Drew
48	48		53	53.5		54	60		62	71
		Q1			Q2			Q3		
		50.50			53.75			61.00		

Start with Q2 then Q3 at right
Then Q1 at left

equivalent to the 75th percentile.



Lower fence



Checking for Outliers by Using Quartiles

Step 1 Determine the first and third quartiles of the data.

Step 2 Compute the interquartile range.

Step 3 Determine the fences. **Fences** serve as cutoff points for determining outliers.

$$\text{Lower fence} = Q_1 - 1.5(\text{IQR})$$
$$\text{Upper fence} = Q_3 + 1.5(\text{IQR})$$

Step 4 If a data value is less than the lower fence or greater than the upper fence, it is considered an outlier.

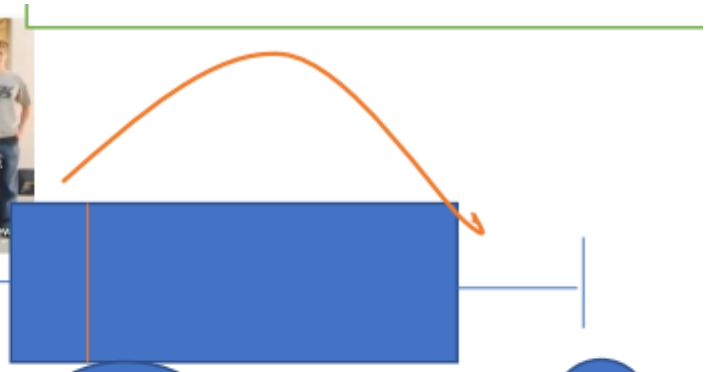
Duncan	Charlie		Grant	Aidan		Sophia	Seamus		Cathy	Drew
48	48		53	53.5		54	60		62	71
		Q1			Q2			Q3		
		50.50			53.75			61.00		

The last step is to subtract:

$$\text{IQR} = \text{Q3} - \text{Q1} = 61.00 - 50.50 = 10.50$$

$$\text{Lower fence} = Q_1 - 1.5(\text{IQR})$$
$$\text{Upper fence} = Q_3 + 1.5(\text{IQR})$$

Threshold 1.5 fixed static formula



In the previous example
LF= 50.50 - 1.5*10.5 = 34.75
UF= 61.00 + 1.5*10.5 = 76.75

Outlier

Not normal distribution skewed left
No OUTLIER

$$\text{Lower fence} = Q_1 - 1.5(\text{IQR})$$
$$\text{Upper fence} = Q_3 + 1.5(\text{IQR})$$

Threshold 1.5 fixed static formula

SNAPSHOT ▸ The Interquartile Range

WHAT IS IT? ▸ A numerical summary.

WHAT DOES IT DO? ▸ It measures the spread of the distribution of a data set.

HOW DOES IT DO IT? ▸ It computes the distance taken up by the middle half of the sorted data.

HOW IS IT USED? ▸ To measure the variability in a sample, particularly when the distribution is skewed.

Upper f

Five-Number Summary

MINIMUM Q_1 M Q_3 MAXIMUM

7 Draw and Interpret Boxplots

Ready in Excel

Drawing a Boxplot

Step 1 Determine the lower and upper fences:

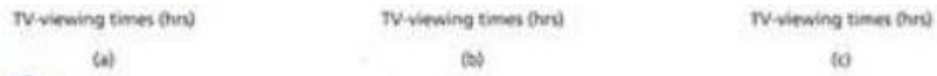
$$\text{Lower fence} = Q_1 - 1.5(\text{IQR})$$
$$\text{Upper fence} = Q_3 + 1.5(\text{IQR})$$
$$\text{where IQR} = Q_3 - Q_1$$

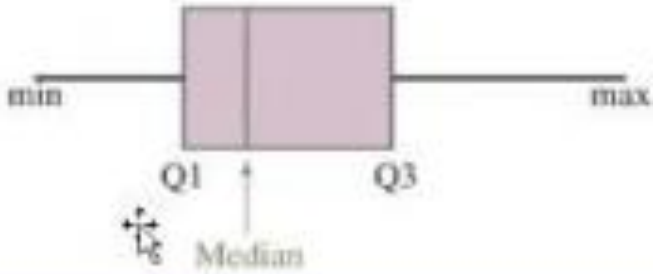
Step 2 Draw a number line long enough to include the maximum and minimum values. Insert vertical lines at Q_1, M, and Q_3. Enclose these vertical lines in a box.

Step 3 Label the lower and upper fences.

Step 4 Draw a line from Q_1 to the smallest data value that is larger than the lower fence. Draw a line from Q_3 to the largest data value that is smaller than the upper fence. These lines are called **whiskers**.

Step 5 Any data values less than the lower fence or greater than the upper fence are outliers and are marked with an asterisk (*).





Other Uses of Boxplots

Boxplots are especially suited for comparing two or more data sets. In doing so, the same scale should be used for all the boxplots.

2 Draw and Interpret Boxplots

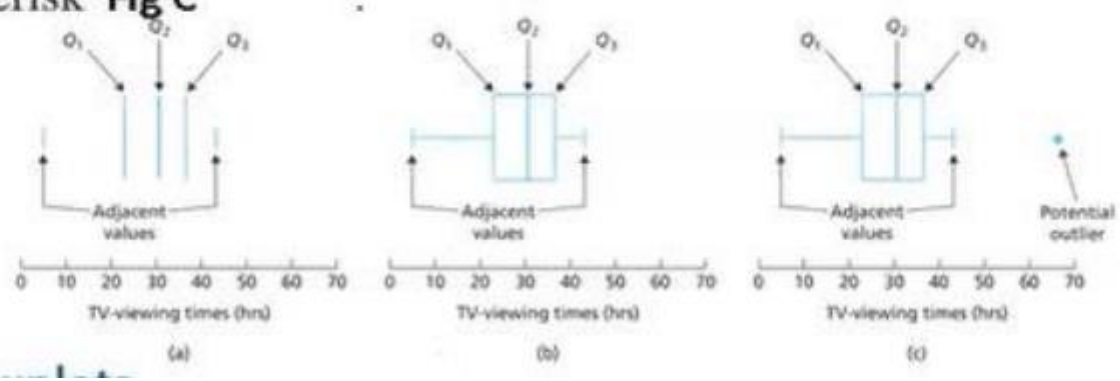
The potential outlier = 66
The adjacent values are 5 and 43

Step 3 **Draw a horizontal axis on which the numbers obtained in Steps 1 and 2 can be located. Above this axis, mark the quartiles and the adjacent values with vertical lines.**

Step 4 **Connect the quartiles to make a box, and then connect the box to the adjacent values with lines.**

Step 5 **Plot each potential outlier with an asterisk.**

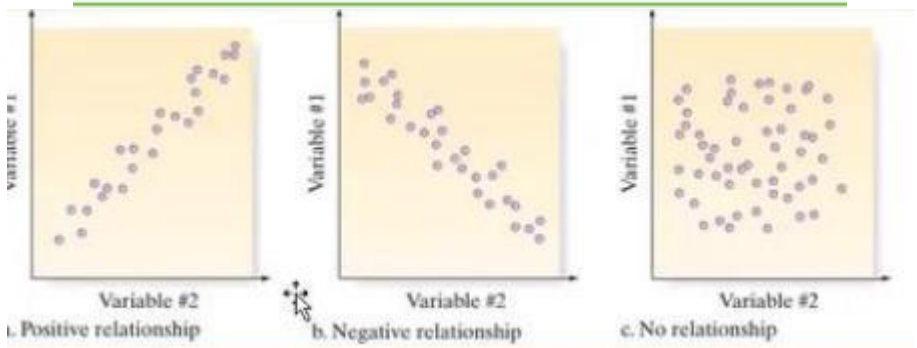
As we noted in Step 2, this data set contains one potential outlier—namely, 66. It is plotted with an asterisk **Fig C**



3-Association: Contingency, Correlation, and Regression

3.1 The Association Between Two Categorical Variables
3.2 Describe the Properties of the Linear Correlation Coefficient
3-3Least-Squares Regression

معامل الارتباط Correlation Principals:
1- Not a causality
2- Range from -1 to 1. is, $-1 \leq r \leq 1.$
Example :
If the number is .78 strong correlation
Or -1 it is perfect negative correlation.



+ - //\\

- The closer r is to ± 1 the closer the data points fall to a straight line, and the stronger the linear association is. The closer r is to 0, the weaker the linear association is.

Benefits:

1-Gives a strong indicator between the two variables In analysis .
2- found strong correlations between 2 columns you can drop one.
EX: temp impact the wheat productivity
Work our has a strong impact on the product.

Sample Linear Correlation Coefficient*

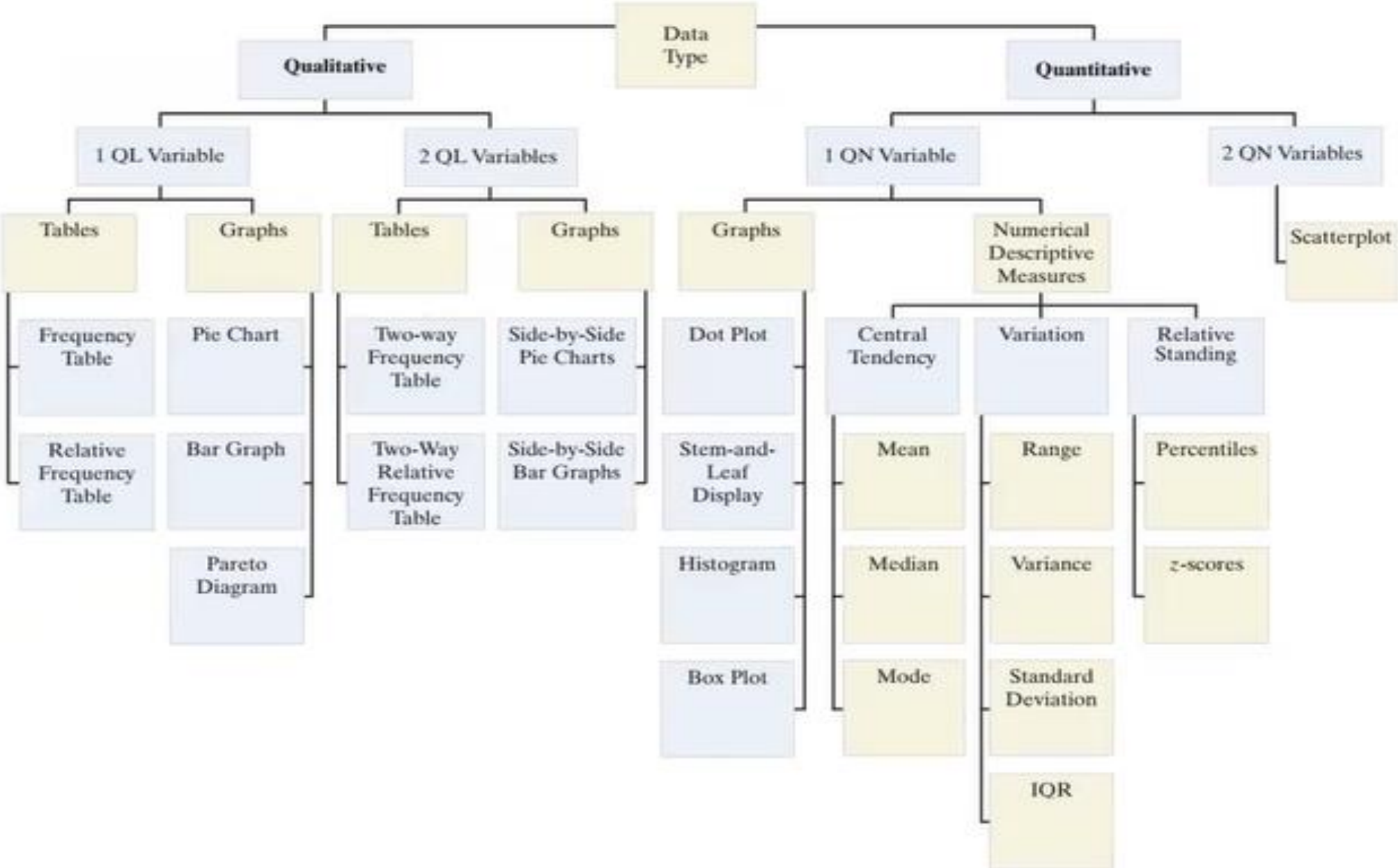
[redacted]

Z-score

$$r = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)}{n - 1}$$

First var x = mean sample
Second var y mean sample
sum (Z-score x*z-score y)/n-1 sample

Guide to Selecting the Data Description Method



Regression Line: An Equation for Predicting the Response Outcome

The **regression line** predicts the value for the response variable y as a straight-line function of the value x of the explanatory variable. Let $\hat{y}$ denote the **predicted value** of y. The equation for the regression line has the form

$$\hat{y} = a + bx.$$

In this formula, a denotes the **y-intercept** and b denotes the **slope**.

3-3 Least-Squares Regression

Example 1: Find the least-squares regression line for the data in below table:

Coefficient*

$$r = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n - 1}$$

Z-score

x	y
1	18
3	13
3	9
6	6
7	4

Solution

Mean sample (x)

Mean sample (y)

$r = -0.946$, $\bar{x} = 4$, $s_x = 2.44949$, $\bar{y} = 10$, and $s_y = 5.612486$

$$b_1 = r \cdot \frac{s_y}{s_x} = -0.946 \cdot \frac{5.612486}{2.44949} = -2.1676$$

Substitute $\bar{x} = 4$, $\bar{y} = 10$, and $b_1 = -2.1676$ into Formula (3):

$$b_0 = \bar{y} - b_1\bar{x} = 10 - (-2.1676)(4) = 18.6704$$

The least-squares regression line is

$$\hat{y} = -2.1676x + 18.6704$$

3-3 Least-Squares Regression

Example 2: Consider again the three-point data set shown in below table. Determine the regression equation for the data. Graph the regression equation and the data points.

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	3	−1	0	1	0
2	1	0	−2	0	0
3	5	1	2	1	2
				2	2

Solution

We first note that $\bar{x} = 2$ and $\bar{y} = 3$.

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{2}{2} = 1$$

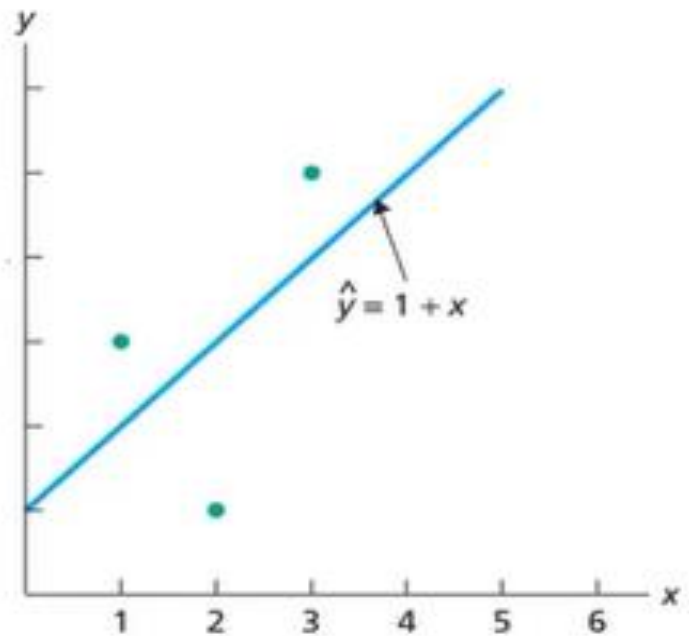
and

$$b_0 = \bar{y} - b_1\bar{x} = 3 - 1 \cdot 2 = 1.$$

Hence, the regression equation is

$$\hat{y} = b_0 + b_1 x = 1 + 1 \cdot x,$$

that is, $\hat{y} = 1 + x$.



Rules of Probabilities

1. The probability of any event E, $P(E)$, must be greater than or equal to 0 and less than or equal to 1. That is, $0 \le P(E) \le 1$.
2. The sum of the probabilities of all outcomes must equal 1. That is, if the sample space $S = \{e_1, e_2, \cdots, e_n\}$, then

$$P(e_1) + P(e_2) + \cdots + P(e_n) = 1$$

The Law of Large Numbers

As the number of repetitions of a probability experiment increases, the proportion with which a certain outcome is observed gets closer to the probability of the outcome.

Empirical Method

$$P(E) \approx \text{relative frequency of } E = \frac{\text{frequency of } E}{\text{number of trials of experiment}} \quad (1)$$

A **probability model** lists the possible outcomes of a probability experiment and each outcome's probability. A probability model must satisfy Rules 1 and 2 of the rules of probabilities.

Computing Probability Using the Classical Method

If an experiment has n equally likely outcomes and if the number of ways that an event E can occur is m, then the probability of E, $P(E)$, is

$$P(E) = \frac{\text{number of ways that } E \text{ can occur}}{\text{number of possible outcomes}} = \frac{m}{n} \quad (2)$$

So, if S is the sample space of this experiment,

$$P(E) = \frac{N(E)}{N(S)} \quad (3)$$

where $N(E)$ is the number of outcomes in E and $N(S)$ is the number of outcomes

Complementary events: The *complement of an event A*, $\overline{A}$, is the set of all sample points in the sample space that do not belong to event A.

$$P(A) + P(\overline{A}) = 1.0 \text{ for any event A}$$

Complement Rule

In words: *probability of A complement = one − probability of A*

In algebra: $P(\overline{A}) = 1 - P(A)$

Table 2

Means of Travel	Frequency
Drive alone	153
Carpool	22
Public transportation	10
Walk	5
Other means	3
Work at home	7

Problem The data in Table 2 represent the results of a survey in which 200 people were asked their means of travel to work.

(a) Use the survey data to build a probability model for means of travel to work.

(b) Estimate the probability that a randomly selected individual carpools to work. Interpret this result.

(c) Would it be unusual to randomly select an individual who walks to work?

Table 3

Means of Travel	Probability
Drive alone	0.765
Carpool	0.11
Public transportation	0.05
Walk	0.025
Other means	0.015
Work at home	0.035

Solution

(a) There are $153 + 22 + \cdots + 7 = 200$ individuals in the survey. The individuals can be thought of as trials of the probability experiment. The relative frequency for "drive alone" is $\frac{153}{200} = 0.765$. We compute the relative frequency of the other outcomes similarly and obtain the probability model in Table 3.

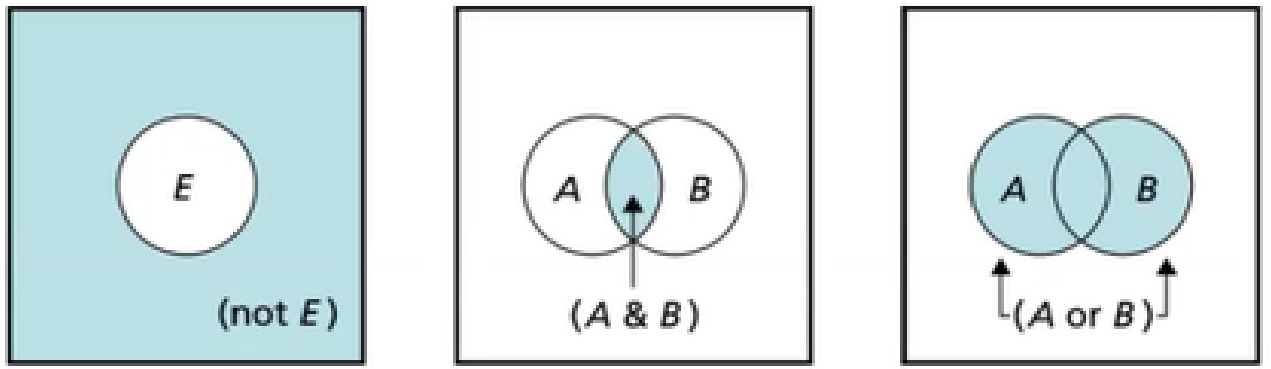
(b) From Table 3, we estimate the probability to be 0.11 that a randomly selected individual carpools to work. We interpret this result by saying, "If we were to survey 1000 individuals, we would expect about 110 to carpool to work."

(c) The probability that an individual walks to work is approximately 0.025. This means if we survey 1000 individuals, we would expect about 25 to walk to work. Therefore, it is unusual to randomly choose a person who walks to work. •

Relationships Among Events

In Words

Intersection means A *and* B (the "overlap" of the events). **Union** means A *or* B or both.



(a) (b) (c)

Relationships Among Events

(not *E*): The event "*E* does not occur"

(*A* & *B*): The event "both *A* and *B* occur"

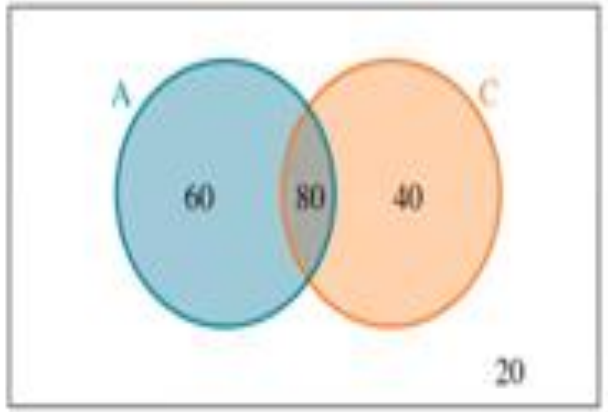
(*A* or *B*): The event "either *A* or *B* or both occur"

A **conditional probability** of an event is a probability obtained with the additional information that some other event has already occurred. $P(B|A)$ denotes the conditional probability of event B occurring, given that event A has already occurred. $P(B|A)$ can be found by dividing the probability of events A and B both occurring by the probability of event A:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

Example 2

One student is selected at random from a group of 200 known to consist of 140 full-time (80 female and 60 male) students and 60 part-time (40 female and 20 male) students. Event A is "the student selected is full-time," and event C is "the student selected is female."



(a) Are events A and C independent?

(b) Find the probability $P(\text{A and C})$ using the multiplication rule.

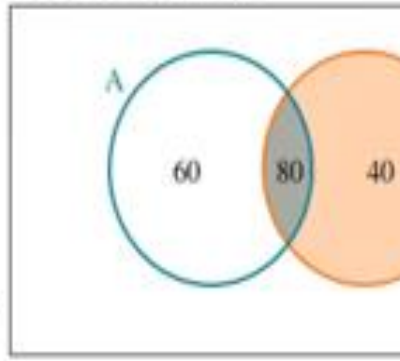
SOLUTION 1

(a) First find the probabilities $P(\text{A})$, $P(\text{C})$, and $P(\text{A}|\text{C})$:

$$P(\text{A}) = \frac{n(\text{A})}{n(S)} = \frac{140}{200} = 0.7$$

$$P(\text{C}) = \frac{n(\text{C})}{n(S)} = \frac{120}{200} = 0.6$$

$$P(\text{A}|\text{C}) = \frac{n(\text{A and C})}{n(C)} = \frac{80}{120} = 0.67$$



A and C are dependent events because $P(\text{A}) \neq P(\text{A}|\text{C})$.

(b) $P(\text{A and C}) = P(\text{C}) \cdot P(\text{A}|\text{C}) = \frac{120}{200} \cdot \frac{80}{120} = \frac{80}{200} = \mathbf{0.4}$