Chapter 3

3.1 Measures of Location:

(1) Mean (2) Median (3) Mode (4) Percentiles (5) Quartile

(1) Mean:

The most important measure of location is the mean or average value for a variable.

Sample Mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is a sample statistics.

n: is the number of observations

x_i: is the i^{th} observation

Example one: Consider the following sample of a class marks:

46	54	42	46	32
x_1	x_2	x_3	x_4	x_5

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{46+54+42+46+32}{5} = 44$$

The sample mean class mark is 44.

Population Mean $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$ N is the total (whole) population number

Note that the sample mean $\bar{x}$ is a point estimator of the population mean μ.

[2] **Median:** is a measure of central location.

→ Arrange the data in ascending order (smaller to largest value)

(a) For an odd number of observations, the median is the middle value.

(b) For an even = = = , = = = = average of the two middle values.

In Example one: Arranging the data ascending order:

32 42 46 46 54

Because n = 5 odd, the median is the middle value = 46

Example two: If the data were ordered like:

32 42 46 46 54 60 70 80 90 96

Then n = 10 even, the median = $\frac{54+60}{2} = \frac{114}{2} = 57$

---

* Whenever a data set contains extreme values, the median is often the prefered measure of central location

---

[3] **Mode:** is another measure of location.

The mode is the value that occurs with greatest frequ

In Example one: The mode is 46

---

* If data contains exactly two modes, then the data are called bimodal.

* If = , more then = = , = = = = = multimodal.

* More than 2 modes would not be helpful in measuring location for data.

[4] Percentile (المئين) :

The p^{th} percentile is a value such that at least p percent of the observations are less than or equal to this value, and at least (100 - p) percent of the observation are greater than or equal to this value.

How to calculate the p^{th} percentile ?

step 1 : Arrange the data in ascending order (smallest to greatest value).

step 2 : Compute an index $i = \left(\frac{p}{100}\right) n$ where

p is the percentile of interest

n is the number of observations.

step 3 : [a] if i is not integer, then round up to the next integre greater than i . This tells us the position of the p^{th} percentile.

[b] if i is an integer , then the p^{th} percentile is the average of the values in positions i and i+1.

Example [a] Compute the 85^{th} percentile for the following data

46 54 42 46 32

step 1 : 32 42 46 46 54

step 2 : $i = \frac{85}{100} \times 5 = 4.25$

step 3 : Because i is not integer $\Rightarrow i = 5$ 5^{th} position $\Rightarrow$ we see that the 85^{th} percentile is the data value 54

[b] Compute the 40^{th} percentile

$i = \frac{40}{100} \times 5 = 2$ integer $\Rightarrow$ the 40^{th} percentile $= \frac{42 + 46}{2} = 44$

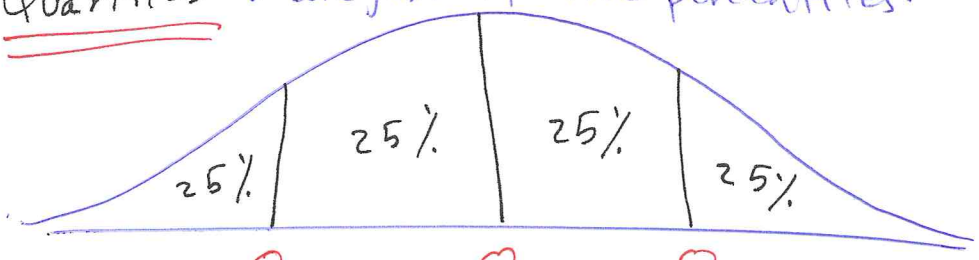
(c) Compute the 50^{th} percentile:

$$i = \left(\frac{50}{100}\right) \times 5 = 2.5 \Rightarrow i = 3$$

$\Rightarrow$ the 50^{th} percentile is 46 "Median"

Note that the 50^{th} percentile is also the median.

(5) Quartiles: are just specific percentiles.



Q_1: first Quartile (25^{th} percentile)

Q_2: Second Quartile (50^{th} percentile) (Median)

Q_3: Third Quartile (75^{th} percentile)

Example: Compute Q_1, Q_2, Q_3 for the data:

46 54 42 46 32

Step 1: 32 42 46 46 54

Step 2:

- $Q_1 = 25^{th}$ percentile

$$i = \left(\frac{25}{100}\right) \times 5 = 1.25 \Rightarrow i = 2 \text{ } 2^{nd} \text{ position}$$

$\Rightarrow Q = 42$

- $Q_2 = 50^{th}$ percentile = median = 46 see part (c)

- $Q_3 = 75^{th}$ percentile

$$i = \left(\frac{75}{100}\right) \times 5 = 3.75 \Rightarrow i = 4 \text{ } 4^{th} \text{ position}$$

$\Rightarrow Q_3 = 46$

3.2 Measures of Variability

[1] Range [2] Interquartile Range [3] Variance [4] Standard deviation [5] Coefficient of Variation.

[1] **Range:** The simplest measure of variability is the range.

Range = largest value − Smallest value.

[2] **Interquartile Range:** is a measure of variability that overcomes the dependency on extreme values.

↓ used to measure the variability when extrem values are present in the data.

Interquartile Range (IQR) = $Q_3 - Q_1$

Q_3: third quartile (75^{th} percentile)

Q_1: 1^{st} quartile (25^{th} percentile)

Note that IQR is the range for the middle 50% of the data.

[3] **Variance:** is a measure of variability that utilizes all the data.

$$\text{Population variance } \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

μ: population mean

N: population size.

estimator

$$\text{Sample variance } s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$\bar{x}$: sample mean

n: sample size.

we divid by n-1 to make estimate that is unbaised for the populat

- Note that $\sum_{i=1}^{N} (x_i - \mu)$ or $\sum_{i=1}^{n} (x_i - \bar{x})$ is called the deviation about the mean and it is always zero.
- Thus, the variance is the average of the square deviation.
- The variance is useful in comparing the variability of two or more variables.

[4] **standard deviation** : is positive square root of the variance. (measures the variability)

Sample standard deviation = $s = \sqrt{s^2}$ ⟩ estimator

Population standard deviation = $\sigma = \sqrt{\sigma^2}$

[5] **Coefficient of Variation** : To indicate how large the standard deviation is relative to the mean.

$$\text{Coefficient of variation} = \left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\%$$

Example: (Q13, Q14 page 95) Consider a sample with data values of 10, 20, 12, 17, 16 Find

[a] Range = largest value − smallest value

= 20 − 10 = 10

[b] Interquartile Range (IQR) = $Q_3 - Q_1$

First we order the data ascending

10 12 16 17 20

Q_3 (75^{th} percentile) $i = \left(\frac{75}{100}\right) \times 5 = 3.75$

i is rounded up to 4^{th} position

$Q_3 = 17$

Q_1 (25^{th} percentile) $i = \left(\frac{25}{100}\right) \times 5 = 1.25$

i is rounded up to 2^{nd} position

$Q_1 = 12$

$IQR = Q_3 - Q_1 = 17 - 12 = 5$

[c] Variance : The sample variance is given by

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}$$

. Note that $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{5} x_i}{5}$

$$= \frac{10+20+12+17+16}{5} = \frac{75}{5} = 15$$

x_i	$x_i - 15$	$(x_i - 15)^2$
10	$10-15 = -5$	$(-5)^2 = 25$
20	$20-15 = 5$	$(5)^2 = 25$
12	$12-15 = -3$	$(-3)^2 = 9$
17	$17-15 = 2$	$(2)^2 = 4$
16	$16-15 = 1$	$(1)^2 = 1$
	Total = 0	$\sum (x_i-15)^2 = 64$

$\sum (x_i - \bar{x}) = 0$

$$s^2 = \frac{\sum_{i=1}^{5}(x_i-15)^2}{5-1} = \frac{64}{4}$$

$s^2 = 16$ sample variance

[d] standard deviation: The sample standard deviation is $s = \sqrt{s^2} = \sqrt{16} = 4$

[e] Coefficient of variations :

$$\text{Coefficient of variation} = \left(\frac{\text{sample standard deviation}}{\text{sample Mean}} \times 100\right)\%$$

$$= \left(\frac{s}{\bar{x}} \times 100\right)\%$$

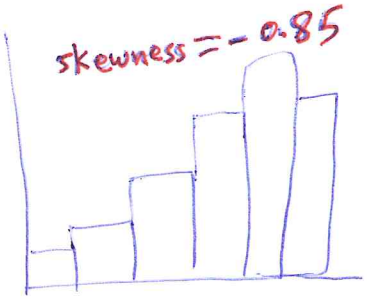
$$= \left(\frac{4}{15} \times 100\right)\%$$

$= 26.67\%$ This means that the sample standard deviation is 26.67% of the value of the sample mean.

3.3 Measures of Distribution shape

An important numerical measure of the shape of a distribution is called skewness.

- For data skewed to the left, the skewness is negative (see A)
- = = = = = right, = = = positive (see B)
- If the data are symmetric, then the skewness is zero (see C)



(A)

Moderately skewed left
(Mean < Median)

(B)

Moderately skewed right
Mean > Median

(C)

Symmetric
(Mean = Median)

(D)

Highley skewed Right

- Note that the median is the best measure of location when the data are highly skewed. Because the median is not affected by high or lower values. (very high / very low)

Z-Scores

Z-score measures the relative location that determine how far a particular value from the mean.

- Suppose that a sample of n observations with values $x_1, x_2, \cdots, x_n$. Then a **z-score** is defined by

$$z_i = \frac{x_i - \bar{x}}{s}$$

where x_i is the i^{th} observation
$\bar{x}$ is the sample mean
s is the sample standard dev.
z_i is the z-score for x_i

- Note that z-score is also called the standardized value.
  z-score is the number of standard deviations x_i from $\bar{x}$.

- $z_1 = 1.2$ means that x_1 is 1.2 standard deviation $> \bar{x}$
  $z_2 = -0.5$ mean that x_1 is 0.5 standard deviation $< \bar{x}$

- If $z_i > 0$, then $x_i > \bar{x}$
  If $z_i < 0$, then $x_i < \bar{x}$
  If $z_i = 0$, then $x_i = \bar{x}$

Example (@ 26 page 102) Consider a sample with a mean of 500 and a standard deviation of 100. What are the z-scores for the data values x_1 520, x_2 650, x_3 500, x_4 450, x_5 280.

$\bar{x} = 500$
$s = 100$

$$z_1 = \frac{x_1 - \bar{x}}{s} = \frac{520 - 500}{100} = 0.2$$

$$z_2 = \frac{x_2 - \bar{x}}{s} = \frac{650 - 500}{100} = 1.5$$

$$z_3 = \frac{x_3 - \bar{x}}{s} = \frac{500 - 500}{100} = 0$$

$$z_4 = \frac{x_4 - \bar{x}}{s} = \frac{450 - 500}{100} = -0.5$$

$$z_5 = \frac{x_5 - \bar{x}}{s} = \frac{280 - 500}{100} = -2.2$$

Chebyshev's Theorem

Chebyshev's Theorem: At least $\left(1-\frac{1}{z^2}\right)$ of the data values must be within z standard deviations of the mean, where $z>1$.

Chebyshev's Theorem enables us to make statements about the proportion of data values that must be within a specified number of standard deviations of the mean.

- At least 75% of the data values must be within $z=2$ standard deviations of the mean.
- At least 89% of the data values must be within $z=3$ standard deviations of the mean
- At least 94% of the data values must be within $z=4$ standard deviations of the mean

---

Example: Suppose for a given sample with mean = 20 and standard deviation of 2. Use Chebyshev's Theorem to determine the percentage of the data within each of the following ranges:

$\bar{x}=20$
$s=2$

[a] 16 – 24 [b] 14 – 26 [c] 10 – 30 [d] 15 – 25

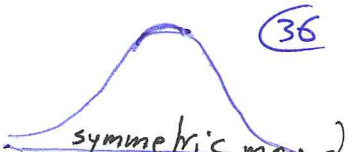
[a] $z=\frac{24-20}{2}=\frac{4}{2}=2 \Rightarrow \left(1-\frac{1}{(2)^2}\right)=1-\frac{1}{4}=0.75$ or 75% (at least)

[b] $z=\frac{26-20}{2}=\frac{6}{2}=3 \Rightarrow \left(1-\frac{1}{(3)^2}\right)=1-\frac{1}{9}=89\%$ (at least)

[c] $z=\frac{30-20}{2}=\frac{10}{2}=5 \Rightarrow \left(1-\frac{1}{(5)^2}\right)=1-\frac{1}{25}=96\%$ (at least)

[d] $z=\frac{25-20}{2}=\frac{5}{2}=2.5 \Rightarrow \left(1-\frac{1}{(2.5)^2}\right)=1-\frac{1}{6.25}=84\%$ (at least)

Empirical Rule



For data having a bell-shaped distribution

- Approximately 68% of the data values will be within 1 standard deviation of the mean.
- Approximately 95% of the data values will be within 2 standard deviations of the mean.
- Almost all of the data values will be within 3 standard deviations of the mean.

Example (Q28 page 102) Suppose the data have a bell-shaped distribution with mean 30 and standard deviation 5. Use the empirical rule to determine the percentage (or range) of the data within each of the following ranges:

$\bar{x} = 30$, $s = 5$

[a] 20 to 40 [b] 15 to 45 [c] 25 - 35

[a] $z = \frac{40-30}{5} = \frac{10}{5} = 2 \Rightarrow$ 95% "approximately"

[b] $z = \frac{45-30}{5} = \frac{15}{5} = 3 \Rightarrow$ Almost all data values

[c] $z = \frac{35-30}{5} = \frac{5}{5} = 1 \Rightarrow$ 68% "approximately"

Detecting Outliers:

Outliers: are observations that are unusually large or unusually small (extreme values)

- They may be included incorrectly in the data set.
- We use z-score to detect outliers.
- Since within $z = 3$, we almost have all the data values in a bell-shaped distribution. Then any observation with $z < -3$ or $z > 3$ is an outlier.

3.4 Exploratory Data Analysis

Stem-and-leaf display is one of the exploratory data analysis. Now we consider another exploratory data analysis (five-number summary)

Five number Summary:

1. Smallest value
2. First quartile (Q_1)
3. Median (Q_2) "second quartile"
4. Third quartile (Q_3)
5. Largest value.

Example: (Q_{36} page 106) Consider a sample with data values

27, 25, 20, 15, 30, 34, 28, 25

Provide the five number summary.

first we order data ascending:

15, 20, 25, 25, 27, 28, 30, 34

1. Smallest value 15
5. Largest value 34 "must be in the end"

2. Q_1 "25th percentile": $i = \left(\frac{25}{100}\right) 8 = 2 \Rightarrow$ we take the average of the 2nd and 3rd position

$$Q_1 = \frac{20+25}{2} = 22.5$$

3. Q_2 "50th percentile": $i = \left(\frac{50}{100}\right) \times 8 = 4 \Rightarrow$ we take the average of the 4th and 5th position

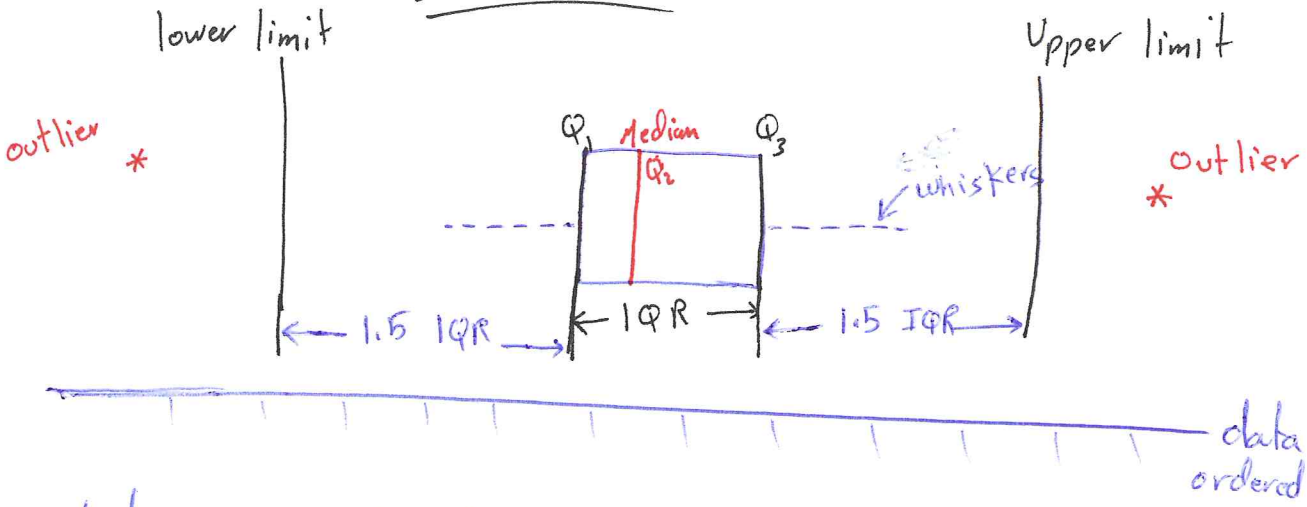
$$Q_2 = \frac{25+27}{2} = 26$$

4. Q_3 "75th percentile": $i = \left(\frac{75}{100}\right) \times 8 = 6 \Rightarrow$ we take the average of the 6th and 7th position

$$Q_3 = \frac{28+30}{2} = 29$$

The five numbers summary are 15, 22.5, 26, 29, 34

Box Plot

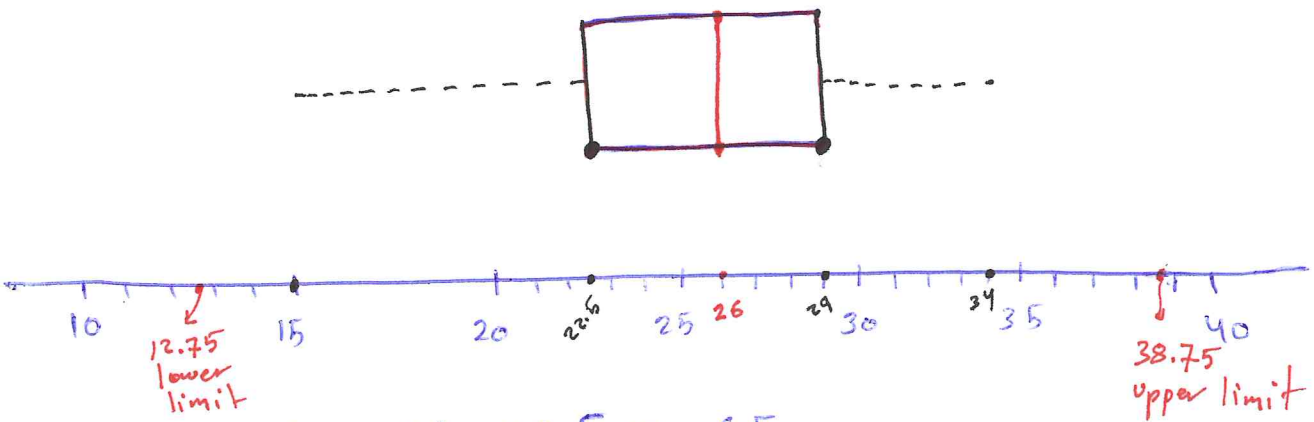


- Box plot is a graphical summary of data that is based on the five number summary.
- Box plot can be used to identify outliers.
- whiskers (dashed lines) are drawn from the ends of the box to the smallest and largest values inside the limits.

Example: (Φ_{37} page 106) show the box plot for the data in Φ_{36}.

Five numbers summary are 15, 22.5, 26, 29, 34 (Q_1, Q_2, Q_3)

smallest value: 15 — Median: 26 — largest value: 34



$IQR = Q_3 - Q_1 = 29 - 22.5 = 6.5$

To find limits $= 1.5(IQR) = 1.5(6.5) = 9.75$

Upper limit $= Q_3 + 9.75 = 38.75$

lower limit $= Q_1 - 9.75 = 12.75$

> we don't have outliers

3.5 Measures of Association Between 2 variables

[1] Covariance and [2] Correlation are descriptive measures for the relationship between two variables.

---

[1] Covariance: • For a sample of size n with observations (x_i, y_i), $i = 1, 2, \dots, n$

the **sample covariance** is $$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

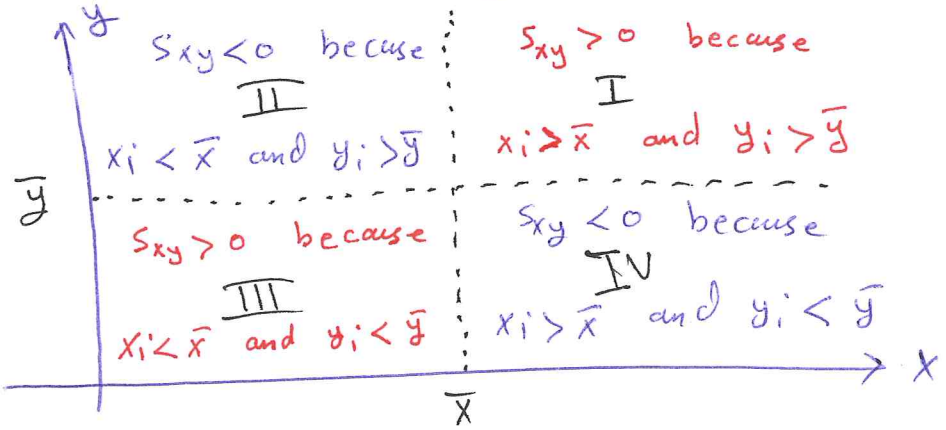
• **Population covariance** $$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

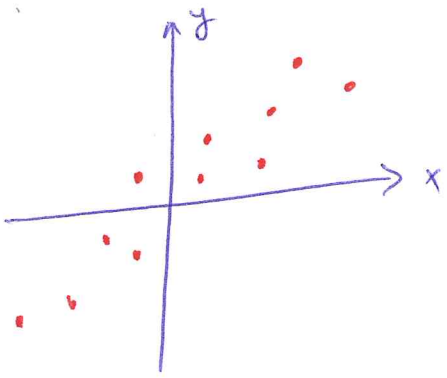
• **Interpretation of the covariance:** The covariance is a measure of the linear association between two variables.

  • If $s_{xy} > 0$, then there is a positive linear association between x and y. (If $x \uparrow$, then $y \uparrow$)

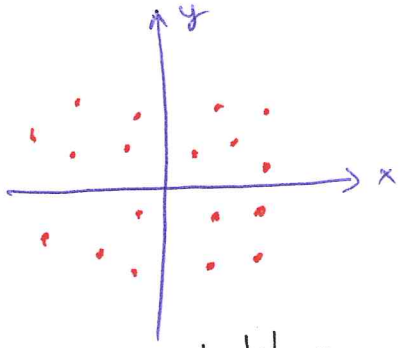
  • If $s_{xy} < 0$, then there is a negative linear association between x and y. (If $x \uparrow$, then $y \downarrow$)

  • If $s_{xy} = 0$ "or close to zero", then there is no linear association between x and y.

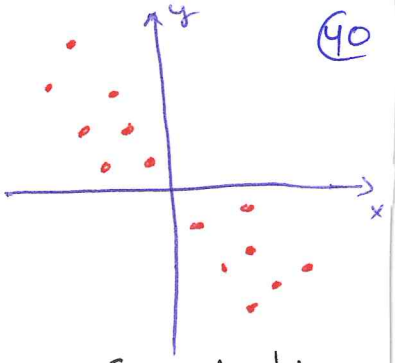




S_{xy} Positive
(x and y are positively linearly related)



S_{xy} approximately 0
(x and y are not linearly related)

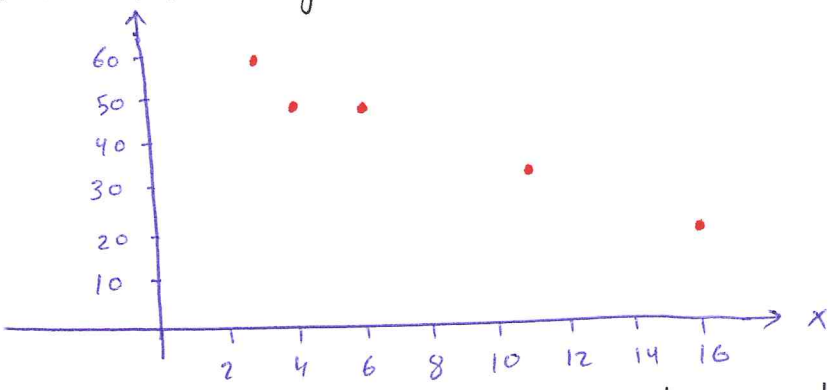


S_{xy} Negative
(x and y are negatively linearly related)

Example: (Q45) Five observations taken for two variables:

x_i	4	6	11	3	16
y_i	50	50	40	60	30

[a] Develop a scatter diagram with x on the horizontal axis.



[b] What does the scatter diagram indicate about the relationship between the two variables?

There is a negative linear relationship between x and y

[c] Compute and interpret the sample covariance.

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
4	50	-4	4	-16
6	50	-2	4	-8
11	40	3	-6	-18
3	60	-5	14	-70
16	30	8	-16	-128
40	230			-240

$$\bar{x} = \frac{\sum x_i}{n} = \frac{40}{5} = 8$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{230}{5} = 46$$

$$S_{xy} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

$$= \frac{-240}{4} = -60$$

The sample covariance indicate a negative linear association between x and y.

[2] Correlation Coefficient

$$r_{xy} = \frac{S_{xy}}{S_x S_y}, \quad \text{where}$$

r_{xy} = sample correlation coefficient "pearson product moment correlation coefficient"

S_{xy} = sample covariance

S_x = sample standard deviation of x $\qquad s_x = \sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}$

S_y = Sample standard deviation of y. $\qquad s_y = \sqrt{\frac{\Sigma(y_i-\bar{y})^2}{n-1}}$

---

* Pearson product moment correlation coefficient "Population data"

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \quad \text{where}$$

ρ_{xy} = population correlation coefficient

σ_{xy} = population covariance

σ_x = population standard deviation for x

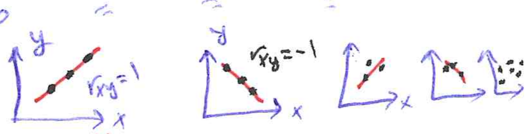
σ_y = population standard deviation for y

---

* The sample correlation coefficient r_{xy} is the estimator of the population correlation coefficient ρ_{xy}.

Interpretation of the correlation coefficient:

- The correlation coefficient ranges from -1 to 1 $\qquad r_{xy} \in [-1, 1]$
- When the 〃 〃 values close to -1, it indicate a stronge linear negative relationship between x and y.
- 〃 〃 〃 〃 〃 〃 〃 1, 〃 〃 〃 〃 positive 〃 〃 〃 〃 〃 〃
- 〃 〃 〃 〃 〃 〃 〃 0 〃 〃 a weak relaship between x and y.



scatters

⇒ Example (Q45)

[d] Compute and interpret the sample correlation coefficient.

x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
4	50	-4	4	16	16
6	50	-2	4	4	16
11	40	3	-6	9	36
3	60	-5	14	25	196
16	30	8	-16	64	256
				118	520

$$S_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{118}{5-1}} = \sqrt{29.5} = 5.43$$

$$S_y = \sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}} = \sqrt{\frac{520}{5-1}} = 11.4$$

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{-60}{(5.43)(11.4)} = -0.969$$

The sample correlation coefficient (-0.969) indicates of a strong negative linear relationship.

3.6 The Weighted Mean and Working with Grouped Data.

Weighted Mean (sample): $\bar{x}_w = \dfrac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$ where

x_i = value of observation i

w_i = weight for observation i

weighted population Mean $\mu_w = \dfrac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$

Computing a grade point average is a good example of the use of a weighted mean

---

Example (Q52 page 122) Consider the following data and crossponding weights:

x_i	Weight (w_i)	$w_i x_i$
3.2	6	19.2
2.0	3	6.0
2.5	2	5.0
5.0	8	40.0
12.7	19	70.2

(a) Compute the weighted mean

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{70.2}{19} = 3.69$$

(b) Compute the sample mean without weighting:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{12.7}{4} = 3.175$$

Grouped Data

Data available in class intervals that are summarized by a frequency distribution.

Example: Consider the following frequency distribution of the audit time (days) for 20 clients.

Audit time	f_i Frequency	Class Midpoint (M_i)	$f_i M_i$	$M_i - \bar{x}_g$	$(M_i - \bar{x}_g)^2$	$f_i (M_i - \bar{x}_g)^2$
10–14	4	$(10+14)/2 = 12$	$4 \times 12 = 48$	$12-19 = -7$	49	196
15–19	8	$(15+19)/2 = 17$	$8 \times 17 = 136$	$17-19 = -2$	4	32
20–24	5	22	$5 \times 22 = 110$	$22-19 = 3$	9	45
25–29	2	27	$2 \times 27 = 54$	$27-19 = 8$	64	128
30–34	1	32	$1 \times 32 = 32$	$32-19 = 13$	169	169
	20		380			570

(a) Find the sample mean for the grouped data $\bar{x}_g = \frac{\sum f_i M_i}{n} = \frac{380}{20} = 19$

(b) Find the sample variance for the grouped data $s^2 = \frac{\sum f_i (M_i - \bar{x}_g)^2}{n-1} = \frac{570}{19} = 30$

(c) Find the sample standard deviation $s = \sqrt{30} \approx 5.48$

- Sample Mean for grouped data $\bar{x}_g = \frac{\sum_{i=1}^{n} f_i M_i}{n}$ where

  M_i = the middle point of class i

  f_i = the frequency of class i

  n = the sample size

Population Mean for grouped data $\mu_g = \frac{\sum_{i=1}^{N} f_i M_i}{N}$

- Sample Variance for grouped data $s_g^2 = \frac{\sum f_i (M_i - \bar{x}_g)^2}{n-1}$

Population Variance for grouped data $\sigma_g^2 = \frac{\sum_{i=1}^{N} f_i (M_i - \mu_g)^2}{N}$

$\bar{x}_g$ is an estimator for μ_g

s_g^2 is an estimator for σ_g^2