Averages

"How difficult is QM1?" "What is the average mark?"
Week 1b, Lecture 2

Topics:
1. Mean
2. Mode
3. Median
4. Order Statistics
5. Minimum, Maximum, Range
6. Percentiles, Quartiles, Interquartile Range
7. Histograms
8. Variance, Standard Deviation, Coefficient of Variation

Definitions:
- **Mode:** the most common value
- **Median:** the middle value
- **Percentiles:** p% of the observations are less than or equal to the p'th percentile
- **Unimodal:** there is a single peak/mode
- **Bimodal:** there are two peaks/modes

Excel Commands:
- mean: AVERAGE
- mode: MODE
- median: MEDIAN
- minimum: MIN
- maximum: MAX
- percentiles: PERCENTILE.EXC
- frequency: FREQUENCY
- variance: VAR.S
- standard deviation: STDEV.S

Note. Bene:
- ⌊i⌋ :round down
- histograms are grouped into equally sized 'bins'
- negatively skewed = long left tail → mean < median
- positively skewed = long right tail → mean > median
- bell shape = single peak → mean ≈ median
- uniform = no peak → mean ≈ median
- measures of dispersion: variance, standard deviation, range and interquartile range

mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

mode: $most\ common\ value$

median

$$if\ n\ is\ odd: x_{\left(\frac{n+1}{2}\right)}$$

$$if\ n\ is\ even: \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

range: $maximum - minimum$

percentiles

$$let\ i = \frac{p}{100}(n+1)$$

$$p'th\ percentile = x_i + (i - \lfloor i \rfloor)(x_{i+1} - x_i)$$

interquartile range

$$IQR = third\ quartile - first\ quartile$$

variance

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

standard deviation

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

coefficient of variation

$$cv = \frac{s}{\bar{x}}$$

Mean

mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

excel: AVERAGE

Definition:
→ let $x_1, x_2, \ldots, x_n$ denote n observations
→ the sample mean is

$$\bar{x} = \frac{x_1 + x_2 + \ldots x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

→ excel command: AVERAGE

Mode

Mode: the most common value of a set

excel: MODE

→ the mode is the **most common value**
→ there may be multiple modes
- *e.g. mode of (1, 3, 4, 4, 5, 6) = 4*
- *mode of (1, 3, 4, 4, 5, 5) = 4 and 5*

→ excel command: MODE
- if a data set contains multiple modes, Excel cannot be relied upon for an accurate answer

Median

median
$\boldsymbol{if\ n\ is\ odd}$: $\boldsymbol{x_{\left(\frac{x+1}{2}\right)}}$
$\boldsymbol{if\ n\ is\ even}$: $\frac{x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}}{2}$

excel: MEDIAN

→ the median is the **middle value** if n is **odd**
- *e.g. median of (42, 10005, 2017) = 2017*

→ the median is the **mean of the two middle values** if n is **even**
- *e.g. median of (42, 10005, 2017, 2015) =* $\frac{2017+2015}{2}$ = 2016

Definition:
→ let $x_1, x_2, \ldots, x_n$ denote n observations
→ sort into ascending order: **(order statistics)**
- $x_{(1)} \leq x_{(2)} \leq x_{(3)} \ldots \leq x_{(n)}$

→ if n is odd: median = $x_{\left(\frac{x+1}{2}\right)}$
→ if n is even: median = $\frac{x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}}{2}$

Maximum, Minimum, Range

range
$= maximum - minimum$

excel: MIN, MAX,

→ the maximum is the highest value
→ the minimum is the lowest value
→ the range is the difference between those two values

Percentiles

Percentile: the percent of the observations that are less than or equal to the p'th percentile

percentile
$\boldsymbol{let\ i} = \frac{\boldsymbol{p}}{\boldsymbol{100}}(\boldsymbol{n}+\boldsymbol{1})$
$\boldsymbol{p'th\ percentile}$
$= \boldsymbol{x_i} + (\boldsymbol{i} - \lfloor \boldsymbol{i} \rfloor)(\boldsymbol{x_{i+1}} - \boldsymbol{x})$

excel: PERCENTILE. EXC

Definition:
→ p% of the observations are less than or equal to the *p*'th percentile
- *e.g. the median is in the 50th percentile*

→ let i = $\frac{p}{100}$ (n + 1)
→ p'th percentile = x_i + (i-⌊i⌋)(x_{i+1}-x_i)
- ⌊i⌋ :round down
  - e.g. ⌊12.1⌋ = 12
    ⌊12.9⌋ = 12

First Quartile (Q_1): the middle number between the smallest number and the median of the data set

Third Quartile (Q_2): the middle value between the median and the highest value of the data set

interquartile range
$\boldsymbol{IQR = third\ quartile - first\ quartile}$

e.g. "Percentiles of QM1 Marks, semester 1 2016"

p	percentile	
10%	50	
25%	55	← first quartile
30%	56.1	
40%	59	
50%	62	← median (second quartile)
60%	65	
75%	69	← third quartile
80%	72	
90%	79	

→ information gathered from this diagram…
- 25% of the students received grades less than or equal to 55
- to be in the top 10%, a student had to achieve a grade over 79
- half of the students received grades less than or equal to 62

→ interquartile range (IQR) = third quartile – first quartile = 69 – 55 = 14

Histograms

excel: FREQUENCY

"Grades at the University of Melbourne"

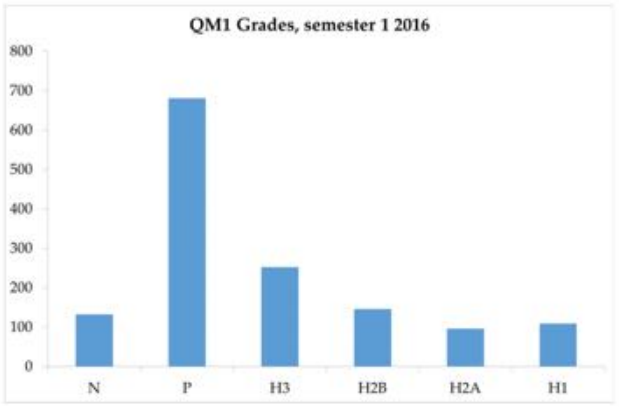
Frequency Distribution of QM1 Grades, Semester 1 2016

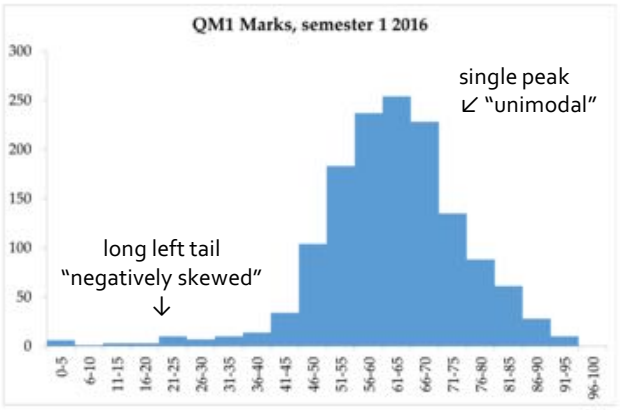
Grade	Mark	Description
H1	80% - 100%	First Class Honours
H2A	75% - 79%	Second Class Honours Division A
H2B	70% - 74%	Second Class Honours Division B
H3	65% - 69%	Third Class Honours
P	50% - 64%	Pass
N	0- 49%	Fail

Grade	Number
H1	109
H2A	96
H2B	146
H3	252
P	681
N	132

→ group all observations into (usually) equally sized "bins" and count:

G	H
Bin	Count
0-5	6
6-10	1
11-15	3
16-20	3
21-25	10
26-30	7
31-35	10
36-40	14
41-45	34
46-50	104
51-55	183
56-60	237
61-65	254
66-70	228
71-75	135
76-80	88
81-85	61
86-90	28
91-95	10
96-100	0

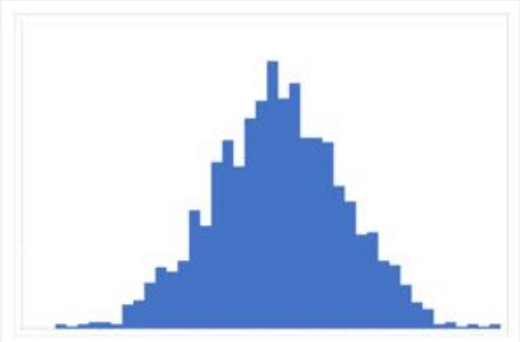




Histogram Shapes

Symmetric and Normal/Bell Shaped

→ unimodal : a single peak
→ mean ≈ median



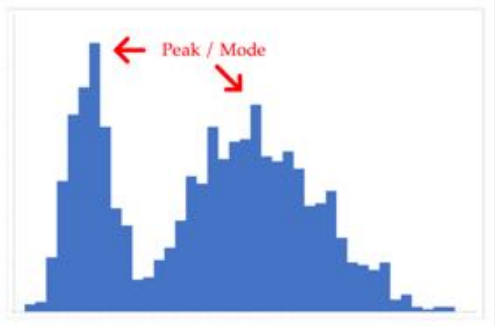
Negatively Skewed

→ unimodal
→ mean < median

Positively Skewed
→ unimodal
→ mean > median

Uniform
→ no distinct peak or mode
→ mean ≈ median

Bimodal
→ has two modes/peaks



Dispersion

variance

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

excel: VAR.S

Variance
Definition:
→ the (sample) variance of observations $x_1, x_2, \ldots, x_n$ is

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

→ units of measurement: the **square of the units** of observations
- *e.g. if x_i are marks in QM1 then s^2 are the squared marks*

→ therefore s^2 is generally uninterpretable

standard deviation

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

excel: STDEV.S

Standard Deviation
Definition:
→ the (sample) standard deviation is the square root of the variance:

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

→ the units of measurement are now the units of the observations
- *e.g. if x_i are marks in QM1 then s is also measured in marks*

Describing Relationships in Data

Week 2a, Lecture 3

Topics:
1. Nominal Associations and Contingency Tables
2. Scatter Plots
3. Covariance
4. Correlation

Definitions:
- **Nominal Data:** data that is not naturally numerical
- **Covariance:** measures whether the relationship between two values is positive or negative
- **Correlation:** measures direction and strength of linear association

Excel Commands:
- covariance: COVARIANCE.S
- correlation: CORREL

Note. Bene:
- scatter plots can be used to display numerical data

covariance

$$cov(x_i, y_i) = s_{xy} - \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- if covariance is **positive**, as y **increases**, x **increases**
- if covariance is **negative**, x and y are **inversely related**

correlation

$$r = \frac{cov(x_1, y_1)}{sd(x_i)sd(y_i)} = \frac{s_{xy}}{s_x s_y}$$

- r close to **+ 1** ⇒ **strong positive** association
- r close to **– 1** ⇒ **strong negative** association
- r close to **0** ⇒ weak or **no linear association**

Nominal Data (Categorical/Qualitative)

Nominal Data: data that is not naturally numerical

→ Nominal Data: data that is not naturally numerical
- *e.g. gender (male, female..); marital status (single, married, divorced…); mode of transport to university (train, tram, car..); degree enrolled in (commerce, arts, science…)*

→ A contingency table can reveal relationships
- *e.g.*

		Degree	
		Commerce	**Arts**
Mode of Transport	**Train**	42%	21%
	Tram	31%	24%
	Bicycle	15%	32%
	Walk	11%	23%
	Helicopter	1%	0%

<u>Relationship</u>

→ data on one characteristic is somehow informative about another
- *e.g. how the mode of transport reflects degrees*
  - *commerce students tend to train or tram, whereas arts students use all modes of transport relatively equally apart from a helicopter*

→ example:

Example. **Leadership in Australian society** (ABS data for 2014)

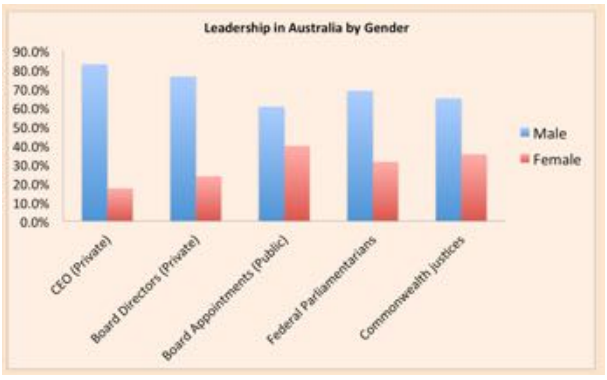
Role	Male	Female
CEOs (private sector)	4,374	914
Board Directors (private sector)	23,873	7,404
Board Appointments (C'wlth govt.)	1,934	1,272
Federal Parliamentarians	154	70
Commonwealth justices/judges	101	55

- *majority of CEOs and board directors are male*
- *in general, there are more men in leadership positions in Australian society*

An Unhelpful Chart

- the scale is too large, only the difference in board directors is properly visible

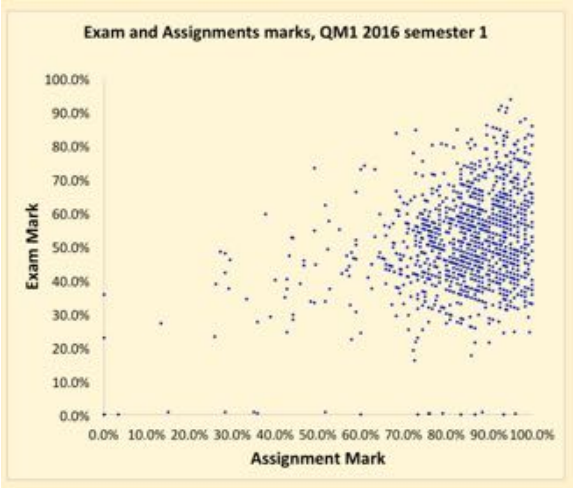
A Helpful Chart



- shows the visible relationship between males and females
- shows the percentage of males and females within each industry

→ these relationships can be displayed with a **scatter plot**

- e.g.



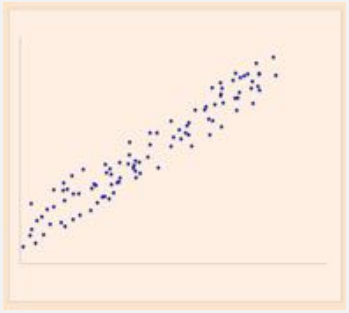
- shows the relationship between assignment marks and exam marks for QM1
- getting a higher score on the exam was more difficult than on the assignments
- there's a positive correlation between the two measurements
- the values have been converted into percentages to make the relationship between the two measurements scaled and more visible

Patterns of Association

Positive and Linear

→ correlation is in a straight line

→ as the x variable increases, the y variable also increases

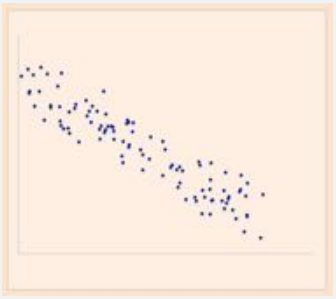


Negative and Linear

→ correlation is in a straight line

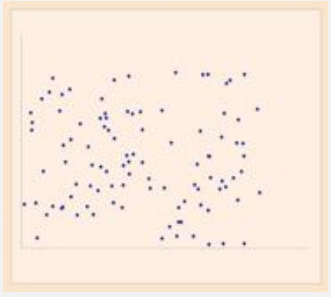
→ as the x variable increases, the y variable decreases

→ inversely related



No Pattern

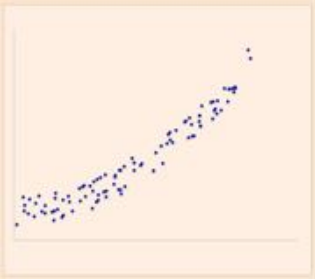
→ no correlation



Positive and Nonlinear

→ positive correlation

→ not in a straight line



Measuring Direction of Association: Covariance

Covariance: this measures whether the relationship between two values is positive or negative

covariance

$$cov(x_i, y_i) = s_{xy} - \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

excel: COVARIANCE.S

definition

→ Covariance between $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$ is

$$\text{cov}(x_i, y_i) = s_{xy} - \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

→ $(x_i - \bar{x})(y_i - \bar{y})$ is positive if:

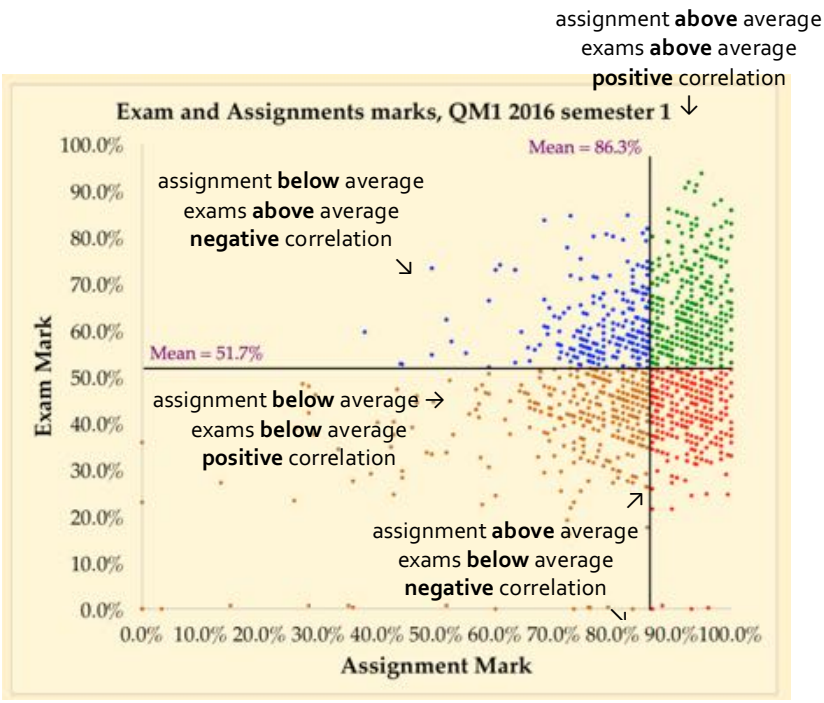
- $x_i > \bar{x}$ and $y_i > \bar{y}$ or…
- $x_i < \bar{x}$ and $y_i < \bar{y}$

→ $(x_i - \bar{x})(y_i - \bar{y})$ is negative if:

- $x_i > \bar{x}$ and $y_i < \bar{y}$ or…
- $x_i < \bar{x}$ and $y_i > \bar{y}$

→ units of measurement is marks^2

e.g. "Exam and Assignment Marks, QM1 2016 Semester 1"



Correlation

Correlation: measures direction and strength of linear association

correlation

$$r = \frac{cov(x_1, y_1)}{sd(x_i)sd(y_i)} = \frac{s_{xy}}{s_x s_y}$$

excel: CORREL

→ Correlation measures **direction** and **strength** of linear association

definition

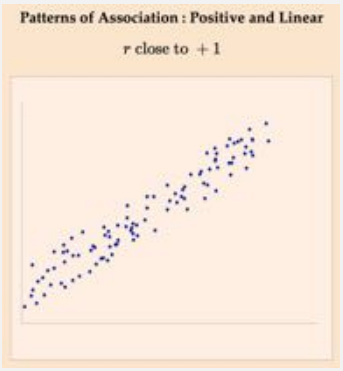
→ Correlation between $x_1,\dots,x_n$ and $y_1,\dots,y_n$ is

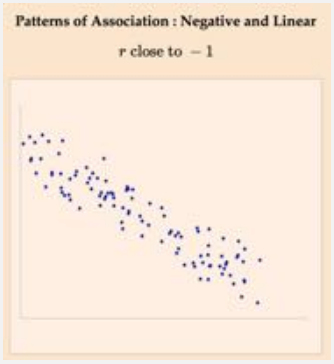
$$r = \frac{\text{cov}(x_1, y_1)}{\text{sd}(x_i)\text{sd}(y_i)} = \frac{s_{xy}}{s_x s_y}$$

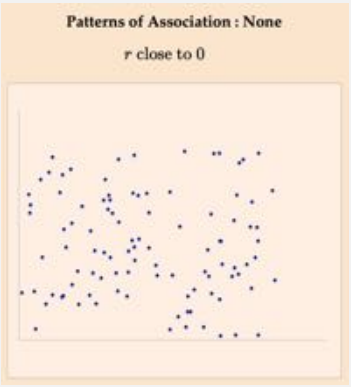
→ $-1 \le r \le 1$:

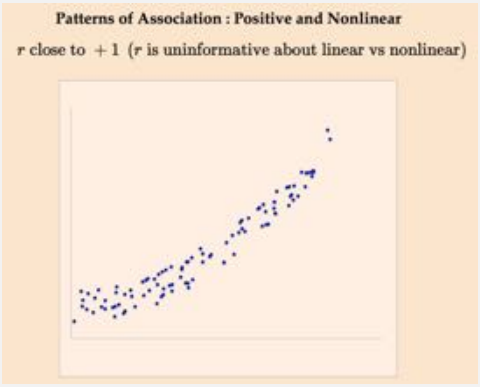
- r close to **+ 1** ⇒ **strong positive** association
- r close to **– 1** ⇒ **strong negative** association
- r close to **0** ⇒ **weak** or no linear association

Patterns of Association: Correlation





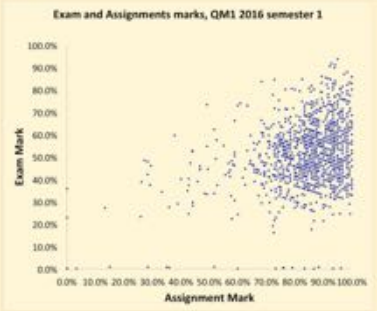




e.g. "QM1 Exam and Assignment Marks"

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{75.96}{14.44 \cdot 15.01} = 0.35$$

s.d. of assignment marks s.d. of exam marks



Your Weight. Average is **82.500**. This was calculated on 01-May-2018.

Year	Study Period	Subject	Short Title	Ver	Mark	Grade Code	Grade Description	Credit Points
2017	Semester 2			1	77	H2A	Second Class Hons A	12.500
2017	Semester 2			1	80	H1	First Class Honours	12.500
2017	Semester 2			1	75	H2A	Second Class Hons A	12.500
2017	Semester 2			1	91	H1	First Class Honours	12.500
2017	Semester 1			1	82	H1	First Class Honours	12.500
2017	Semester 1			1	80	H1	First Class Honours	12.500
2017	Semester 1	ECON10005	Quantitative Methods 1	1	90	H1	First Class Honours	12.500
2017	Semester 1			1	85	H1	First Class Honours	12.500

Email my Statement of Results