

Data Analysis and Decision Making (DADM)

Lecture 2

What I do while waiting for my toast to pop up



Topic: Presenting and Reporting Data

- Data visualisation
- Summarising data
- Data in Action: Evaluating business reports

Copyright © 2020 McGraw-Hill and Catherine Leighton.

Review Lecture 1



Learning Objectives (LOs)

- Summarize qualitative data with frequency distributions.
- Understand and interpret pie charts, bar charts and Pareto diagrams.
- Summarize quantitative data with frequency distributions.
- Understand and interpret histograms, polygons, and ogives.
- Understand and interpret scatterplots.
- Calculate and interpret the arithmetic mean, median, and mode.
- Calculate and interpret the range, variance, standard deviation, and coefficient of variation.
- Apply the empirical rule.
- Interpret the correlation coefficient.

BUSINESS STATISTICS | Jaggia, Kelly

Summarizing Qualitative Data

- A **frequency distribution** (for qual. data) groups data into categories and records how many observations fall into each category.
- To calculate the **relative frequency** divide each category's frequency by the sample size.
- To express relative frequencies in terms of percentages, multiply each proportion by 100%.

	Frequency	Relative Frequency	Percentage
MCom	30	0.31	30.93
MProfAcc	19	0.20	19.59
MAppFin	14	0.14	14.43
MBILM	13	0.13	13.40
MDS	10	0.10	10.31
MMark	7	0.07	7.22
MCom&MIntComLaw	1	0.01	1.03
MIndEng&Mgt	1	0.01	1.03
Grad Cert Com	1	0.01	1.03
HDR PhD Prelim	1	0.01	1.03
Total	97	1	100

BUSINESS STATISTICS | Jaggia, Kelly

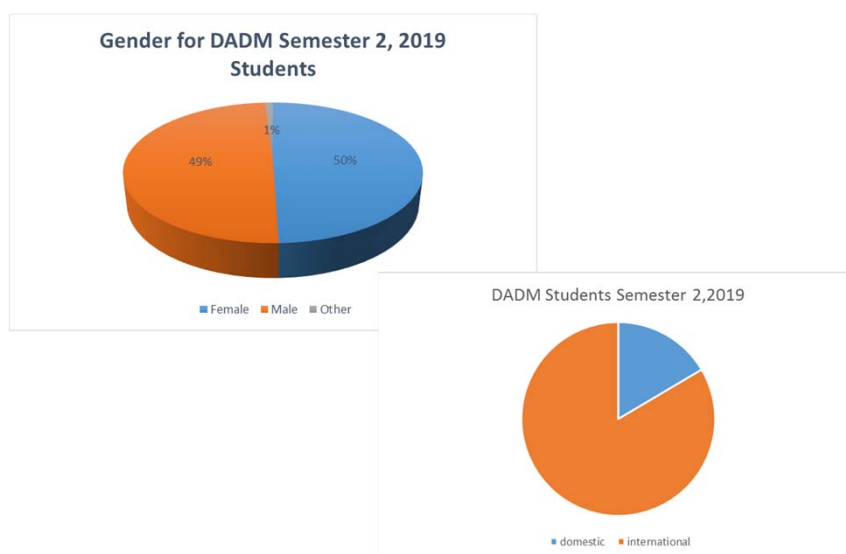
Summarizing Qualitative Data

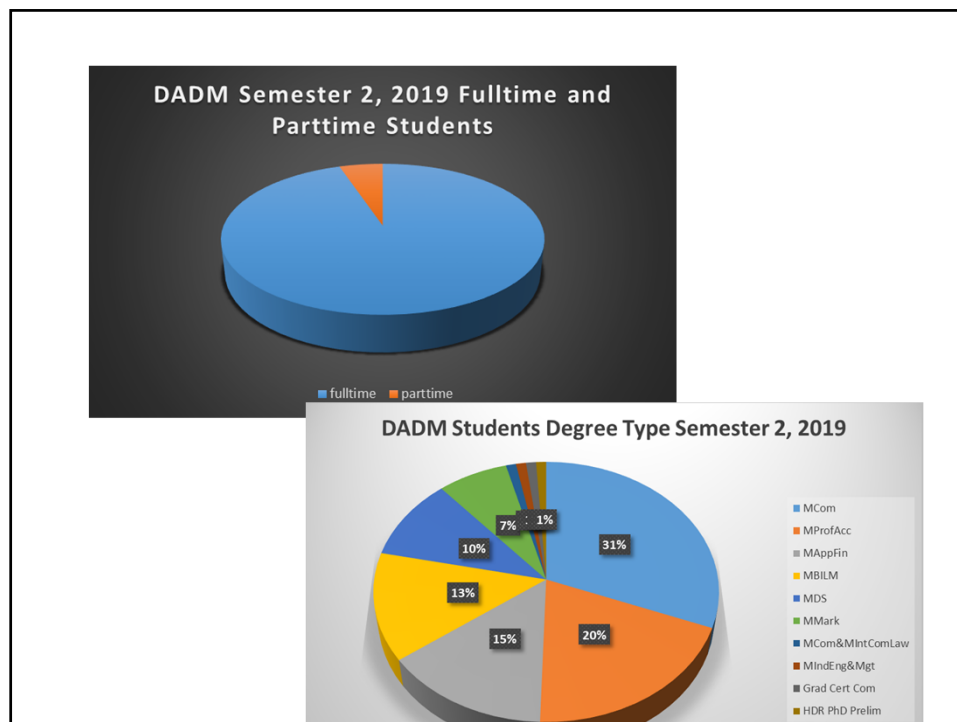
- A **pie chart** is a segmented circle whose segments portray the relative frequencies of the categories of a qualitative variable.
- A **bar chart** depicts the frequency or the relative frequency for each category of the qualitative data as a bar rising vertically from the horizontal axis.

BUSINESS STATISTICS | Jaggia, Kelly

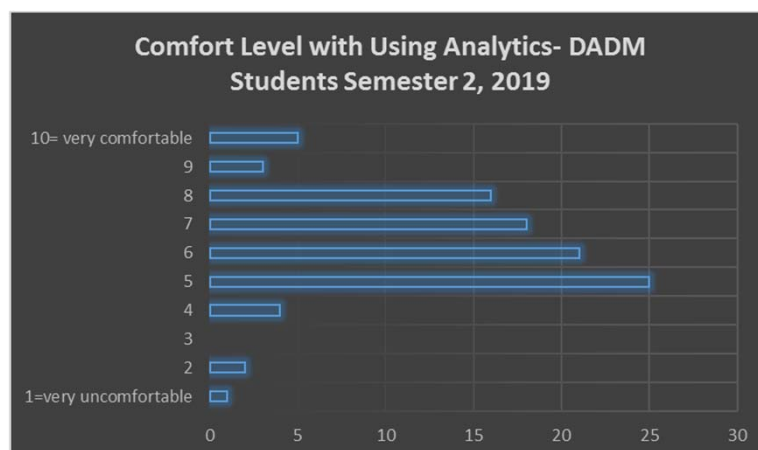
Pie Charts:

Can also be used to determine composition by constituent parts





Bar Charts: Bars represent Categories (Note: the example below treats the numbers as Ordinal data)



We use bar charts to compare things and provide us with a way to display data to make it easy to read

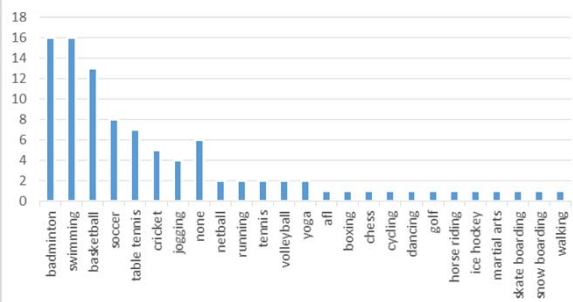
Pie vs Bar

Favourite Sport



■ badminton ■ swimming ■ basketball ■ soccer ■ table tennis
 ■ cricket ■ jogging ■ none ■ netball ■ running
 ■ tennis ■ volleyball ■ yoga ■ afl ■ boxing
 ■ chess ■ cycling ■ dancing ■ golf ■ horse riding
 ■ ice hockey ■ martial arts ■ skate boarding ■ snow boarding ■ walking

Favourite Sport

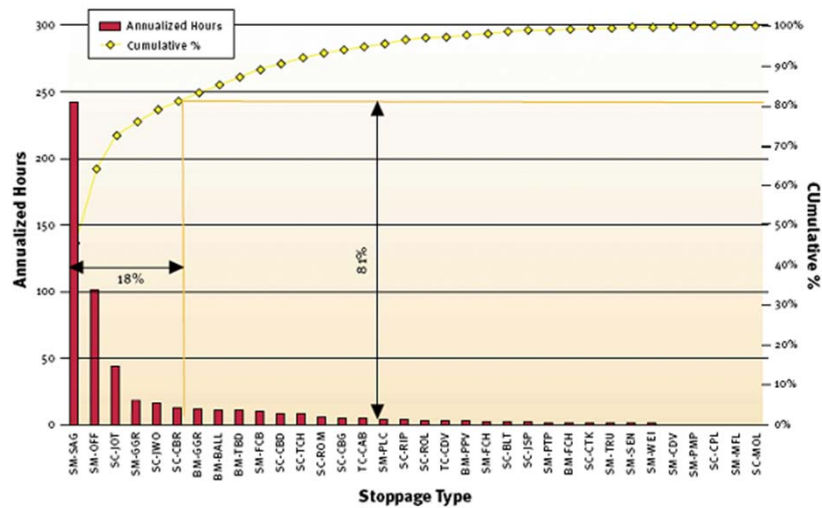


Pareto Analysis

- 80/20 rule
- *“80% of the land in Italy was owned by 20% of the population”*
- Critical few vs. Trivial many



Pareto Diagram



Misleading Graphical Displays

To avoid misleading:

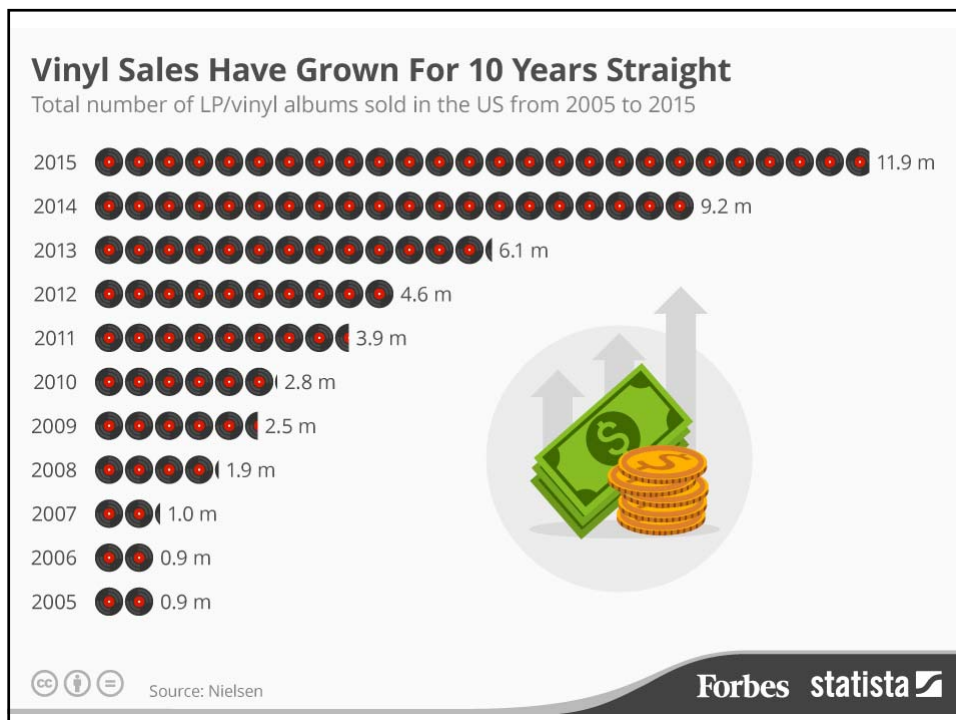
- Present data in a clear, simple way.
- Strive for clarity.
- Avoid chart junk.
- Do not compress the vertical axis.
- Do not stretch the vertical axis.
- Axes should be clearly marked with numbers and clearly labelled.
- Give your chart a title.
- Interpret your chart in the body of your report. Don't assume the reader understands it.
- The vertical axis should (in most cases) include zero.
- For bar charts, bars should be the same width.
- Include a key for different categories.

PLEASE NOTE: Excel does not know these rules! You will always need to fix up an Excel graph.

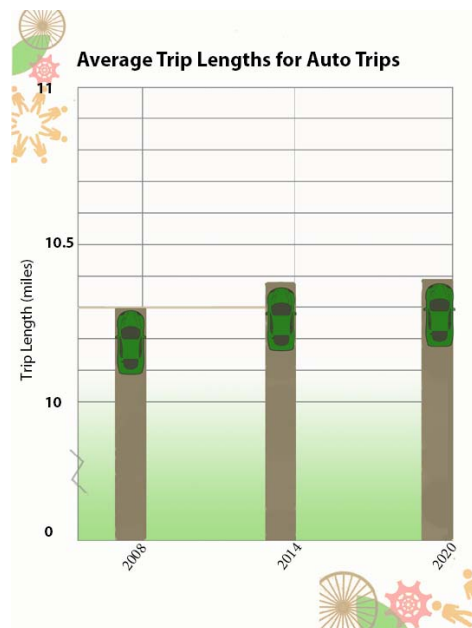
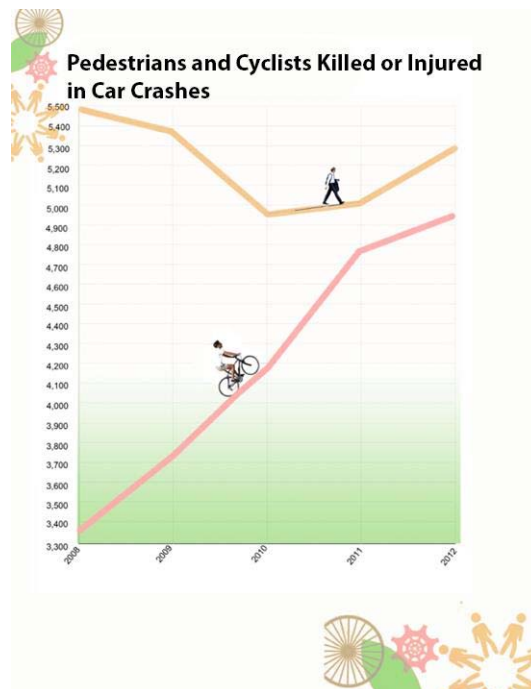
With HSBC Premier,
we put you at the
centre of all we do.

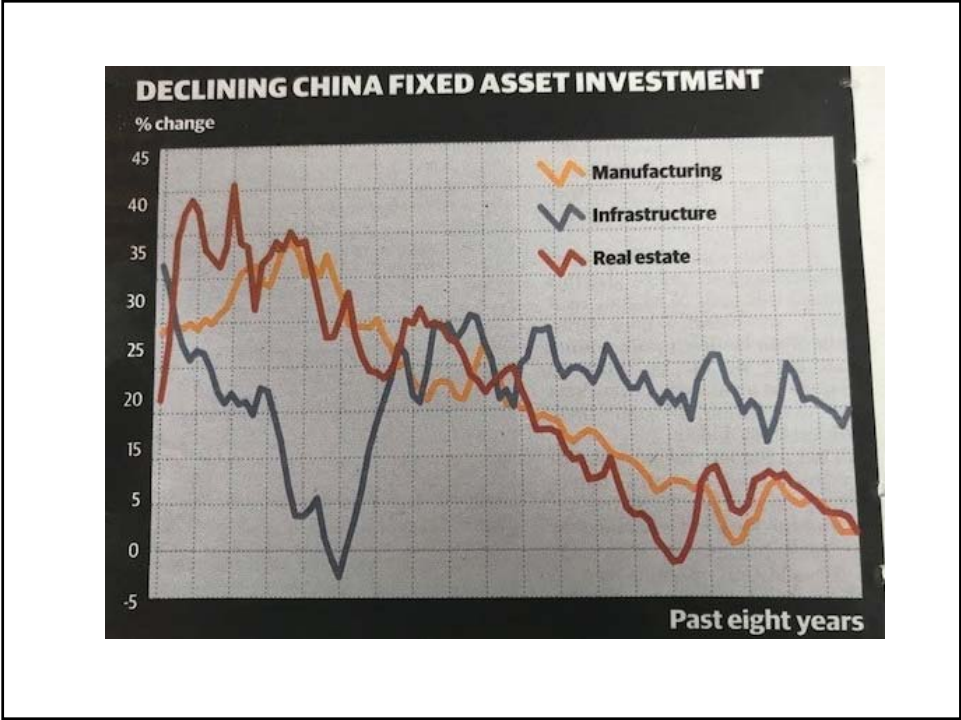
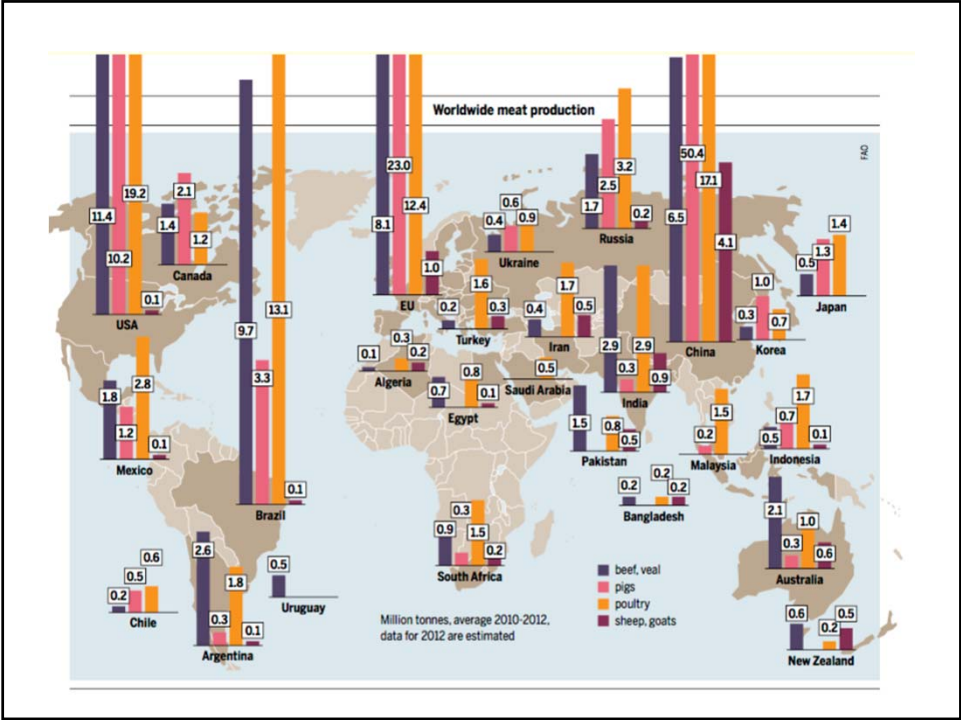


HSBC Premier

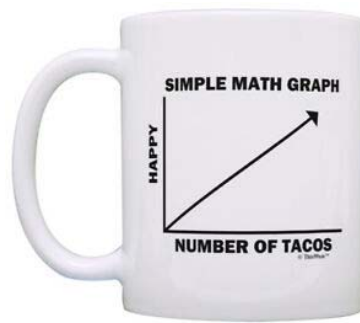


LA MTA Report April 2015





Coffee Break



Summarizing Quantitative Data

- A **frequency distribution** for quantitative data groups data into intervals called classes, and records the number of observations that fall into each class.
- Guidelines when constructing frequency distribution:
 - ❑ Classes are *mutually exclusive*.
 - ❑ Classes are *exhaustive*.
 - ❑ The number of classes usually ranges from 5 to 20. This is a guideline, not an absolute rule.

Summarizing Quantitative Data

- The raw **height data** from our class survey have been converted into a frequency distribution in the following table.

Class	Frequency	Cumulative Frequency
140.1 up to 150	1	1
150.1 up to 160	13	$13+1=14$
160.1 up to 170	34	$14+34=48$
170.1 up to 180	36	$48+36=84$
180.1 up to 190	12	$84+12=96$
190.1 up to 200	1	$96+1=97$
Total	97	

- A **cumulative frequency distribution** specifies how many observations fall below the upper limit of a particular class.

BUSINESS STATISTICS | Jaggia, Kelly

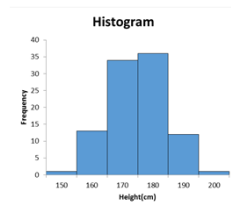
Summarizing Quantitative Data

Here are the relative frequency and the cumulative relative frequency distributions for your height data.

Class	Frequency	Relative Frequency	Cumulative Relative Frequency
140.1 up to 150	1	$1/97=0.010$	$=0.010$
150.1 up to 160	13	$13/97=0.134$	$=0.010+0.134$ $=0.144$
160.1 up to 170	34	$34/97=0.351$	$=0.144+0.351$ $=0.495$
170.1 up to 180	36	$36/97=0.371$	$=0.495+0.371$ $=0.866$
180.1 up to 190	12	$12/97=0.124$	$=0.866+0.124$ $=0.990$
190.1 up to 200	1	$1/97=0.010$	$=0.990+0.010$ $=1$
Total	97	1	

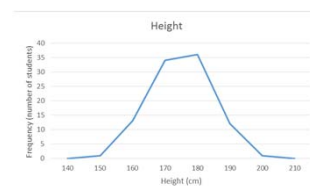
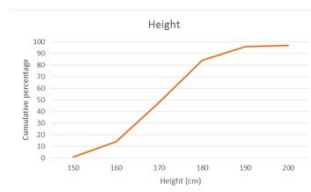
Summarizing Quantitative Data

✓ Histograms



✓ Polygons

✓ Ogives

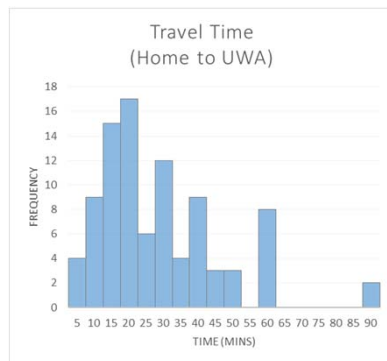


BUSINESS STATISTICS | Jaggia, Kelly

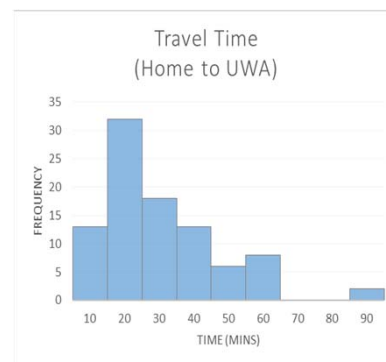
Summarizing Quantitative Data

- A **histogram** is a visual representation of a frequency or a relative frequency distribution.
 - ✓ Bar height represents the respective class frequency (or relative frequency).
 - ✓ Bar width represents the class width.

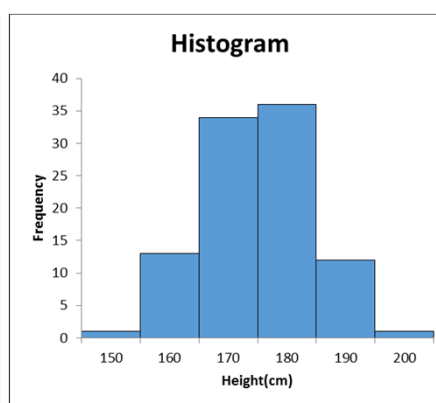
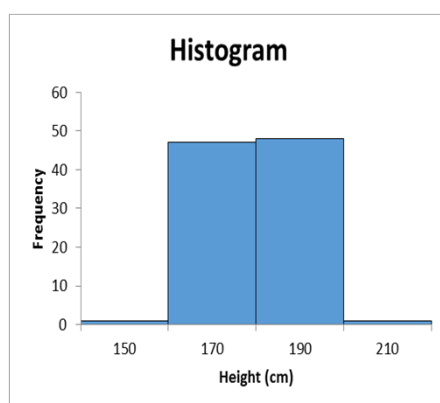
BUSINESS STATISTICS | Jaggia, Kelly



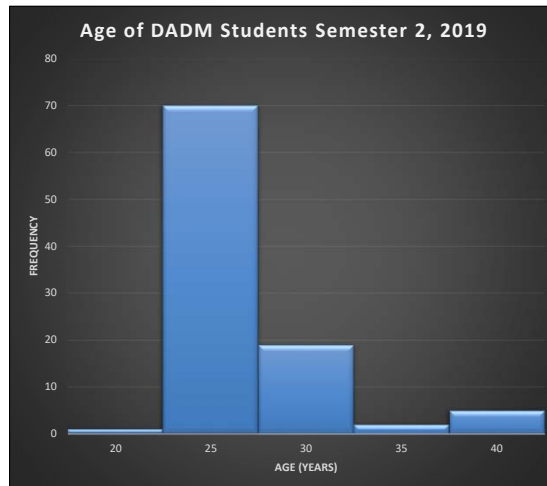
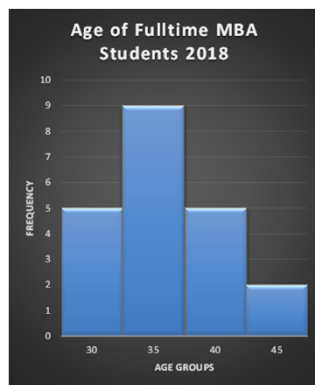
Histogram: It usually deals with continuous data and the bars represent the frequency of the range of that data. How many groupings/bins/classes?



Histogram: It usually deals with continuous data and the bars represent the frequency of the range of that data

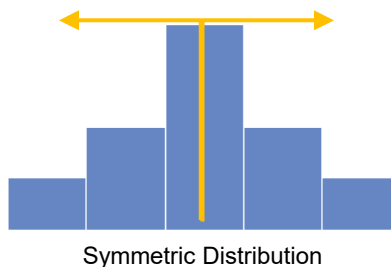


Histogram: It usually deals with continuous data and the bars represent the frequency of the range of that data



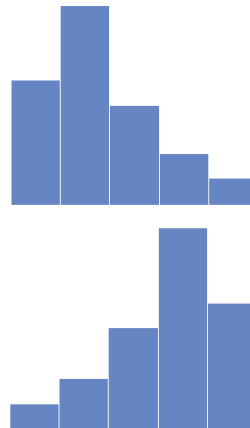
Summarizing Quantitative Data

- Shape of Distribution: typically symmetric or skewed
 - ✓ Symmetric—mirror image on both sides of its center.



Summarizing Quantitative Data

- Skewed distribution
 - ✓ Positively skewed - data form a long, narrow tail to the right.
 - ✓ Negatively skewed - data form a long, narrow tail to the left.



BUSINESS STATISTICS | Jaggia, Kelly

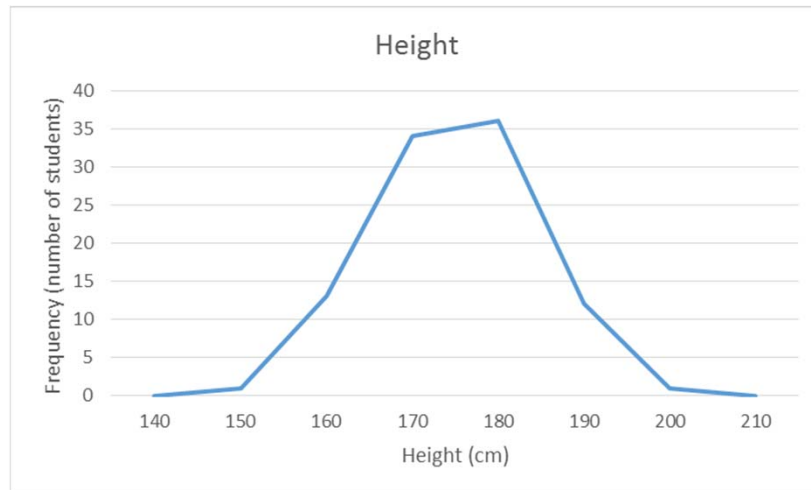
Summarizing Quantitative Data

- A **polygon** is a visual representation of a frequency or a relative frequency distribution.
 - ✓ Plot the class midpoints on x-axis and associated frequency (or relative frequency) on y-axis.
 - ✓ Neighboring points are connected with a straight line.

BUSINESS STATISTICS | Jaggia, Kelly

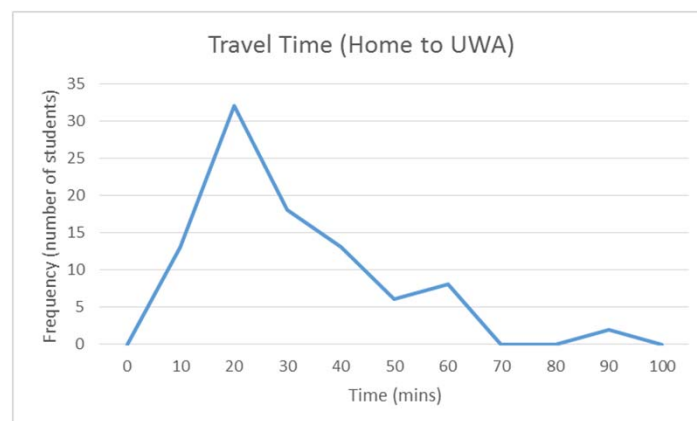
Summarizing Quantitative Data

- Here is a polygon for height.



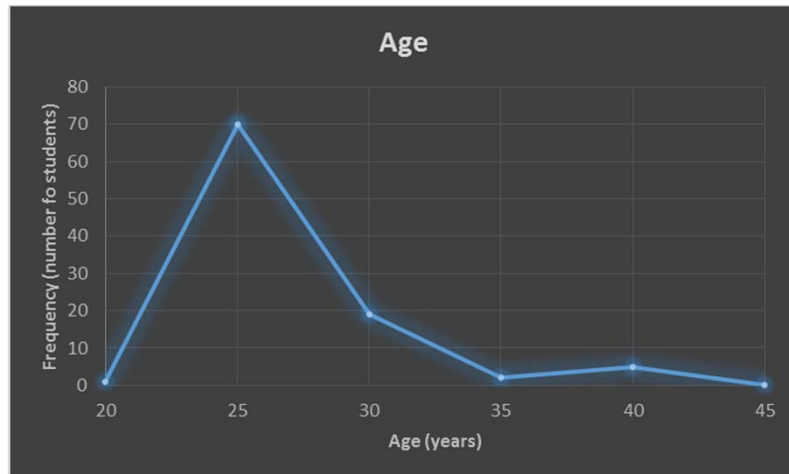
Summarizing Quantitative Data

- Here is a polygon for distance to UWA.

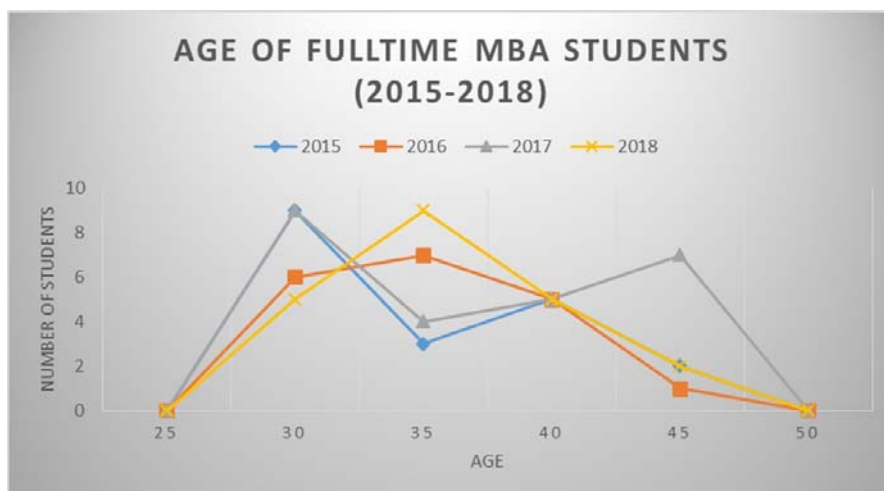


Summarizing Quantitative Data

- Here is a polygon for age.



Q- When would you use a polygon in place of a histogram?



Summarizing Quantitative Data

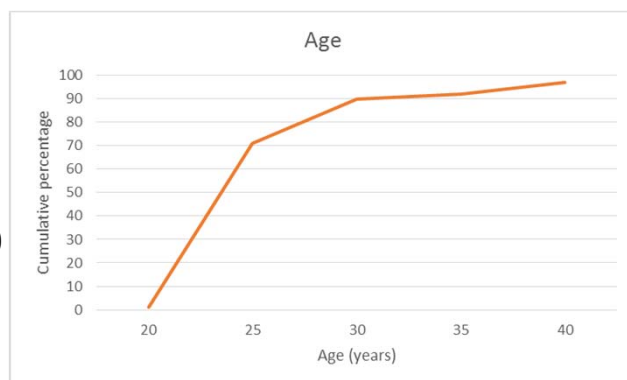
- An ogive is a visual representation of a cumulative frequency or a cumulative relative frequency distribution.
 - ✓ Plot the cumulative frequency (or cumulative relative frequency) of each class above the upper limit of the corresponding class.
 - ✓ The neighboring points are then connected.

BUSINESS STATISTICS | Jaggia, Kelly

Summarizing Quantitative Data

Here is an ogive for the age data.

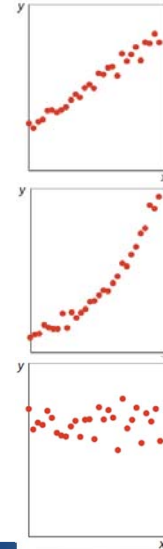
Approximate the percentage of students aged 30 years and under.



Scatterplots

Graphical tool to help determine if two variables are related.

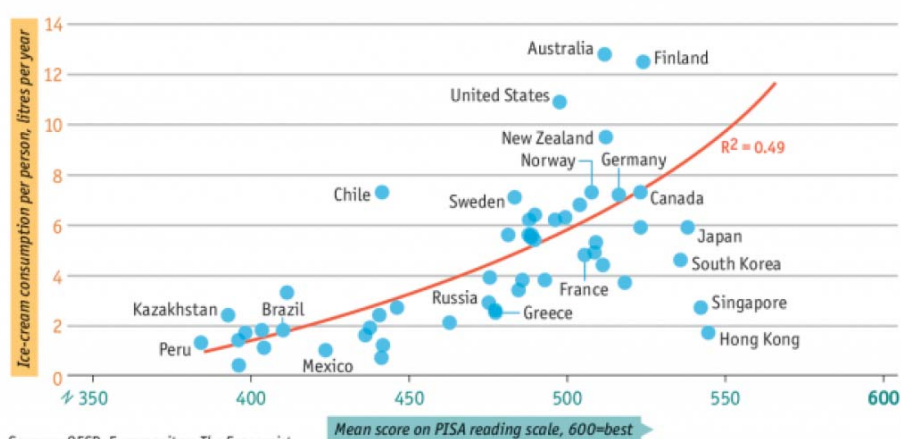
- **Linear** relationship: upward or downward-sloping trend of the data.
- **Curvilinear** relationship: for example as x increases, y increases at an increasing (or decreasing) rate.
- **No relationship**: data are randomly scattered with no discernible pattern.



BUSINESS STATISTICS | Jaggia, Kelly

Ice-cream consumption and PISA educational performance scores

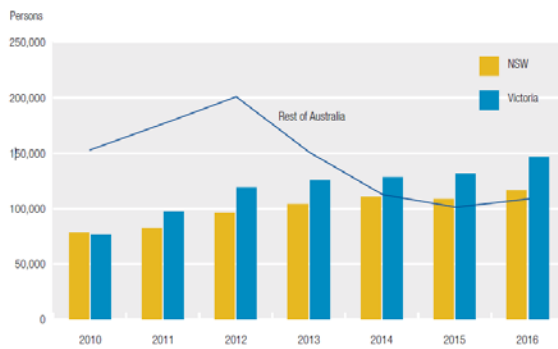
2012



Sources: OECD; Euromonitor; The Economist

BUSINESS STATISTICS | Jaggia, Kelly

FIGURE 2
ANNUAL POPULATION CHANGE, NSW, VICTORIA, AND REST OF AUSTRALIA

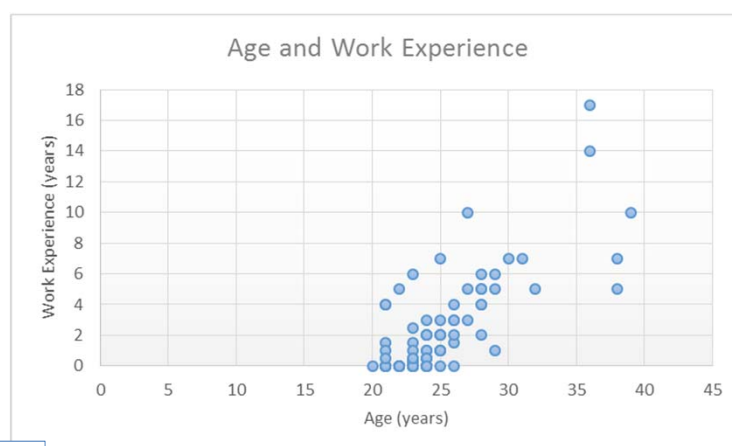


Data source: Australian Bureau of Statistics, Australian Demographic Statistics, Cat no. 3101.

BUSINESS STATISTICS | Jaggia, Kelly

Scatterplots

Q: How would you describe this relationship?

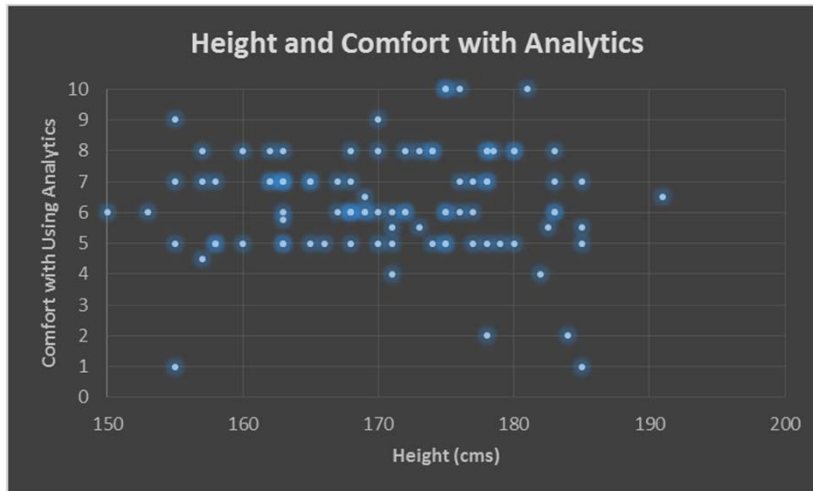


Dependent
Variable or y

Independent
Variable or x

Scatterplots

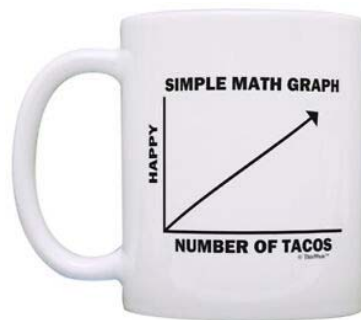
Q: How would you describe this relationship? Does it makes sense to consider this relationship?



TED Talk- Hans Rosling The Best Stats You've Ever Seen

http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

Coffee Break



Stats Joke

A statistician can have their head in an oven and their feet in ice, and they will say that on the average they feel fine.

Measures of Central Location

- The **arithmetic mean** is a primary measure of central location. It is often referred to it as the **average**.

Sample Mean

$$\bar{x} = \frac{\sum x_i}{n}$$

Population Mean

$$\mu = \frac{\sum x_i}{N}$$

- The mean is sensitive to outliers.

BUSINESS STATISTICS | Jaggia, Kelly

Measures of Central Location

- The **median** is another measure of central location that is not affected by outliers.
- The **mode** is another measure of central location.
 - ✓ The most frequently occurring value in a data set
 - ✓ Used to summarize qualitative data
 - ✓ A data set can have no mode, one mode (unimodal), or many modes (multimodal).

BUSINESS STATISTICS | Jaggia, Kelly

Average Aussie News Article

The West Australian | thewest.com.au | Thursday, April 11, 2013 | NEWS 9

CENSUS RESULTS

Retail mum, 37, typical Australian

By Shane Wright
Economics Editor

Forget the Akubra-wearing bloke with a sheep dog and a farm. The average Australian is a 37-year-old woman living in the suburbs with her two children. She works as a shop assistant to help pay the \$1800-a-month mortgage.

A breakdown of the 2011 Census by the Australian Bureau of Statistics has revealed a country long departed from the images penned by Henry Lawson and Banjo Paterson. In Australia's first census in 1911, the average person was a 24-year-old man who eked out life as a farmer and was Anglican. By 1961, the average Australian was a 29-year-old man, locally born, who worked as a clerk.

But in the 2011 Census, with women now outnumbering men across the country, the average Australian was a woman.

And instead of working the land, the average Australian works in a shop, is likely to drive a private car to work and be a Catholic.

Catholicism became the biggest religious group in the early 1960s. The percentage of Australians with a mortgage has grown from 27 per cent in 2001 to 35 per cent in 2011. The monthly mortgage pay-

THE AVERAGE AUSTRALIAN

Sex	Female
Age	37
Born	Australia
Job	Sales assistant
Religion	Catholic
Marital status	Married
Children	Two
House	Three bedroom
Mortgage	\$1800 a month
Living	Capital city suburb
Height	161.8cm
Weight	71.1kg
Source	ABS

capital city and nearly 90 per cent live in an urban area.

The bureau admits no one fits the exact average — least of all West Australians. The average West Australian is a man, 36, who drives a truck for a living. The stark difference from the average Australian is solely because of the State's mining sector, which has led to an increase in the number of men across WA.

The only other part of Australia where the average person is a male is the Northern Territory, where the average person is a 31-year-old member of the defence forces.

Strong growth in transport, from mining trucks to vans making



Face of WA: Truck driver Corey McCluskey fits the profile of the average West Australian. Picture: Michael O'Brien

Average is just fine for some of us

By Jayne Rickard and Ryan Miller

Maria Miceli did not realise just how average her life was. Until yesterday the bubbly shop assistant, 37, had no idea her age, job, marital status and number of children made her the average face of a modern-day Australia.

Australian Bureau of Statistics data has revealed the average Australian to be 37 years old, female, married with two children and working as a shop assistant.

Mrs Miceli laughed when told how closely she fitted the profile of the "average" Australian and said it felt strange to know so many people were living a similar life.

The mother of two young girls said she loved her part-time job at



Nice fit: Shop assistant Maria Miceli.

a change of routine from being at home.

"I look forward to it," Mrs Miceli said.

"The customers are nice and you get a bit of a rapport with the locals. And it's close to home, so it's

resources sector, the average West Australian is a 36-year-old male truck driver.

Success truckie Corey McCluskey moved to WA from Victoria three months ago to take advantage of the mining boom and better pay.

"I'm the face of WA, hey? I'll probably scare everybody off now," Mr McCluskey said.

Mr McCluskey, a father of three children who turns 37 in August, has been a truck driver since he was 19 and said he wouldn't want to do anything else.

"I don't claim to be anything," he said. "I just do my job and I love the job and it's all I ever wanted to do."

"I like the freedom, being my own boss, I suppose. It's a lot of fun but you have to enjoy your own company a bit so job can be inspiring some-

Measures of Dispersion

- Measures of dispersion gauge the variability of a data set.
- Measures of dispersion include:
 - ✓ Range
 - ✓ Variance and Standard Deviation
 - ✓ Coefficient of Variation (CV)

Measures of Dispersion

- **Range**

Range = Maximum Value – Minimum Value

- ✓ It is the simplest measure.
- ✓ It focusses on extreme values.

BUSINESS STATISTICS | Jaggia, Kelly

Measures of Dispersion

- **Variance and standard deviation**

- For a given sample,

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \text{and} \quad s = \sqrt{s^2}$$

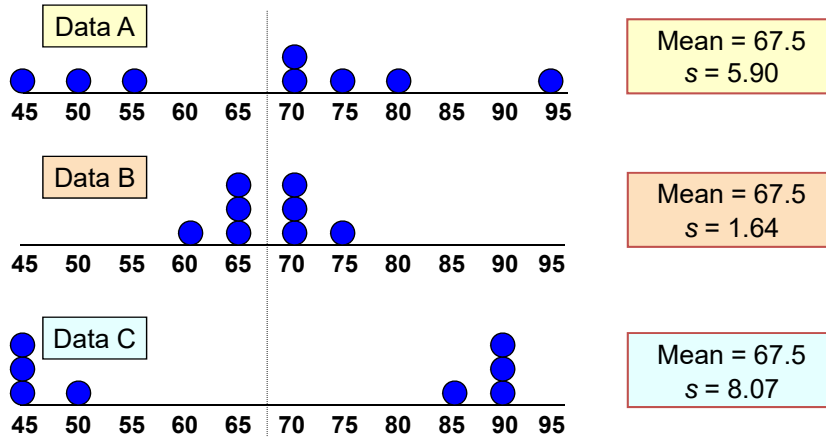
- For a given population,

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \quad \text{and} \quad \sigma = \sqrt{\sigma^2}$$

BUSINESS STATISTICS | Jaggia, Kelly

Comparing Standard Deviations

Same mean, but different standard deviations:



Measures of Dispersion

- **Coefficient of variation (CV)**
 - ✓ CV adjusts for differences in the magnitudes of the means.
 - ✓ CV is unitless, allowing easy comparisons of mean-adjusted dispersion across different data sets.

$$\text{Sample CV} = \frac{s}{\bar{x}}$$

$$\text{Population CV} = \frac{\sigma}{\mu}$$

DADM Student Results- Summary Stats

	Age	Height	Work Experience	Travel Time to UWA	Comfort with Analytics
Mean	24.891	170.685	2.191	28.413	6.307
Standard Error	0.407	0.938	0.328	1.862	0.179
Median	24	171	1	25	6
Mode	23	163	0	15	6
Standard Deviation	3.904	8.995	3.148	17.856	1.717
Sample Variance	15.241	80.916	9.911	318.839	2.947
Kurtosis	3.942	-0.685	6.357	1.650	1.474
Skewness	1.896	-0.166	2.235	1.195	-0.467
Range	19	41	17	89	9
Minimum	20	150	0	1	1
Maximum	39	191	17	90	10
Sum	2290	15703	201.6	2614	580.25
Count	92	92	92	92	92

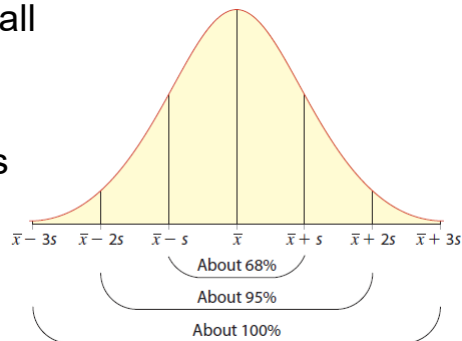
$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \text{and} \quad s = \sqrt{s^2}$$

Analysis of Relative Location- We will discuss this in our next lecture

- The Empirical Rule:**

- ✓ Approximately 68% of all observations fall in the interval $\bar{x} \pm s$.
- ✓ Approximately 95% of all observations fall in the interval $\bar{x} \pm 2s$.
- ✓ Almost all observations fall in the interval $\bar{x} \pm 3s$.



Correlation

- The **correlation coefficient** (r_{xy} or ρ_{xy}) describes both the direction and strength of the relationship between x and y .
- The scatterplot visually displays this relationship.

BUSINESS STATISTICS | Jaggia, Kelly

Interpreting Correlation Coefficients

- Sign (+/-) **direction** of relationship --> direct or inverse relationship
- Values--> **strength** of relationship

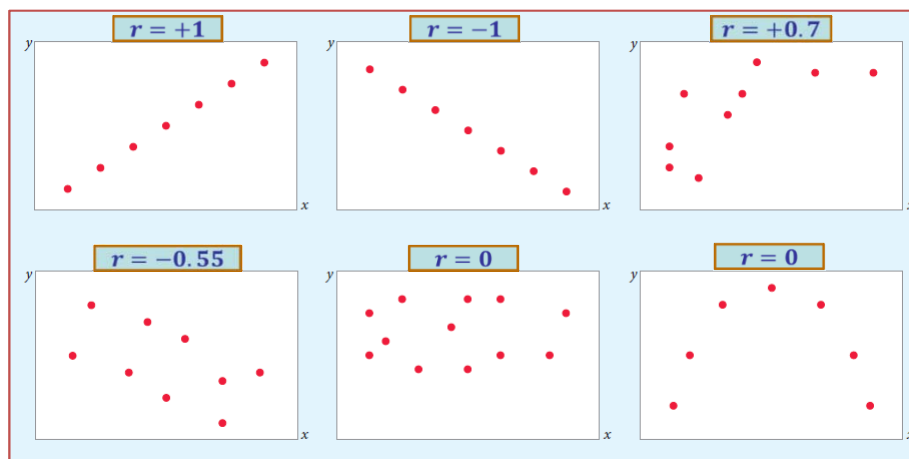
Rule of Thumb

- $|.71| - |1.00|$ Strong
- $|.41| - |.70|$ Moderate
- $|.01| - |.40|$ Weak
- 0 No relationship

Correlation Matrix

	<i>Age</i>	<i>height</i>	<i>work experience</i>	<i>time to UWA</i>	<i>comfort</i>
Age	1				
height	-0.002	1			
work experience	0.760	0.009	1		
time to UWA	0.112	-0.106	0.155	1	
comfort	-0.103	-0.013	0.012	0.302	1

Correlation between Two Variables



Sample Statistics versus Population Parameters- It's all Greek to me!

Measure	Population Parameter	Sample Statistic
Mean	μ	\bar{x}
Variance	σ^2	s^2
Standard Deviation	σ	s
Correlation	ρ	r
Size	N	n

Summary

- Summarize qualitative data by forming frequency distributions.
- Understand and interpret pie charts and bar charts.
- Summarize quantitative data by forming frequency distributions.
- Understand and interpret histograms, polygons, and ogives.
- Understand and interpret a scatterplot.
- Calculate and interpret the arithmetic mean, median, and mode.
- Calculate and interpret the range, variance, standard deviation, and coefficient of variation.
- Apply the empirical rule.
- Interpret the correlation coefficient.

Next Lecture....

- Brief revision of Lecture 2
- Topic:
 - Dealing with Risk and Uncertainty

SEE YOU THEN



BUSINESS SCHOOL