

please get a handout from the back!

Stats 101/101G/108 Workshop

Exploratory Data Analysis

2019

by Leila Boyle



Stats 101/101G/108 Workshops

The Statistics Department offers workshops and one-to-one/small group assistance for Stats 101/101G/108 students wanting to improve their statistics skills and understanding of core concepts and topics.

Leila's website for Stats 101/101G/108 workshop hand-outs and information is here: www.tinyURL.com/stats-10x

Resources for this workshop, including pdfs of this hand-out and Leila's scanned slides showing her working for each problem are available here: www.tinyURL.com/stats-EDA

Leila Boyle

Undergraduate Statistics Assistance, Department of Statistics
Room 303.320 (third floor of the Science Centre, Building 303)
l.boyle@auckland.ac.nz; (09) 923-9045; 021 447-018

Want help with Stats?

Stats 101/101G/108 appointments

Book your preferred time with Leila here: www.tinyURL.com/appt-stats, or contact her directly (see above for her contact details).



Stats 101/101G/108 Workshops

Workshops are run in a relaxed environment, and allow plenty of time for questions. In fact, this is encouraged ☺

Please make sure you bring your calculator with you to all of these workshops!

- **Preparation at the beginning of the semester:**

Multiple identical sessions of a preparation workshop are run at the beginning of the semester to get students off to a good start – come along to whichever one suits your schedule!

- Basic Maths and Calculator skills for Statistics

www.tinyURL.com/stats-BM

- **First half of the semester**

Five theory workshops are held during the first half of the semester:

- ✓ ◦ Exploratory Data Analysis

www.tinyURL.com/stats-EDA

- Proportions and Proportional Reasoning www.tinyURL.com/stats-PPR

- Observational Studies, Experiments, Polls and Surveys

www.tinyURL.com/stats-OSE

- Confidence Intervals: *Means*

www.tinyURL.com/stats-CIM

- Confidence Intervals: *Proportions*

www.tinyURL.com/stats-CIP

- **Second half of the semester**

Five theory workshops and one computing workshop are held during the second half of the semester:

- **Statistics Theory Workshops**

- Hypothesis Tests: *Proportions*

www.tinyURL.com/stats-HTP

- Hypothesis Tests: *Means (part 1)*

www.tinyURL.com/stats-HTM

- Hypothesis Tests: *Means (part 2)*

www.tinyURL.com/stats-HTM

- Chi-Square Tests

www.tinyURL.com/stats-CST

- Regression and Correlation

www.tinyURL.com/stats-RC

- **Computer Workshop:** Hypothesis Tests in SPSS

www.tinyURL.com/stats-HTS

- **Useful Computer Resource:**

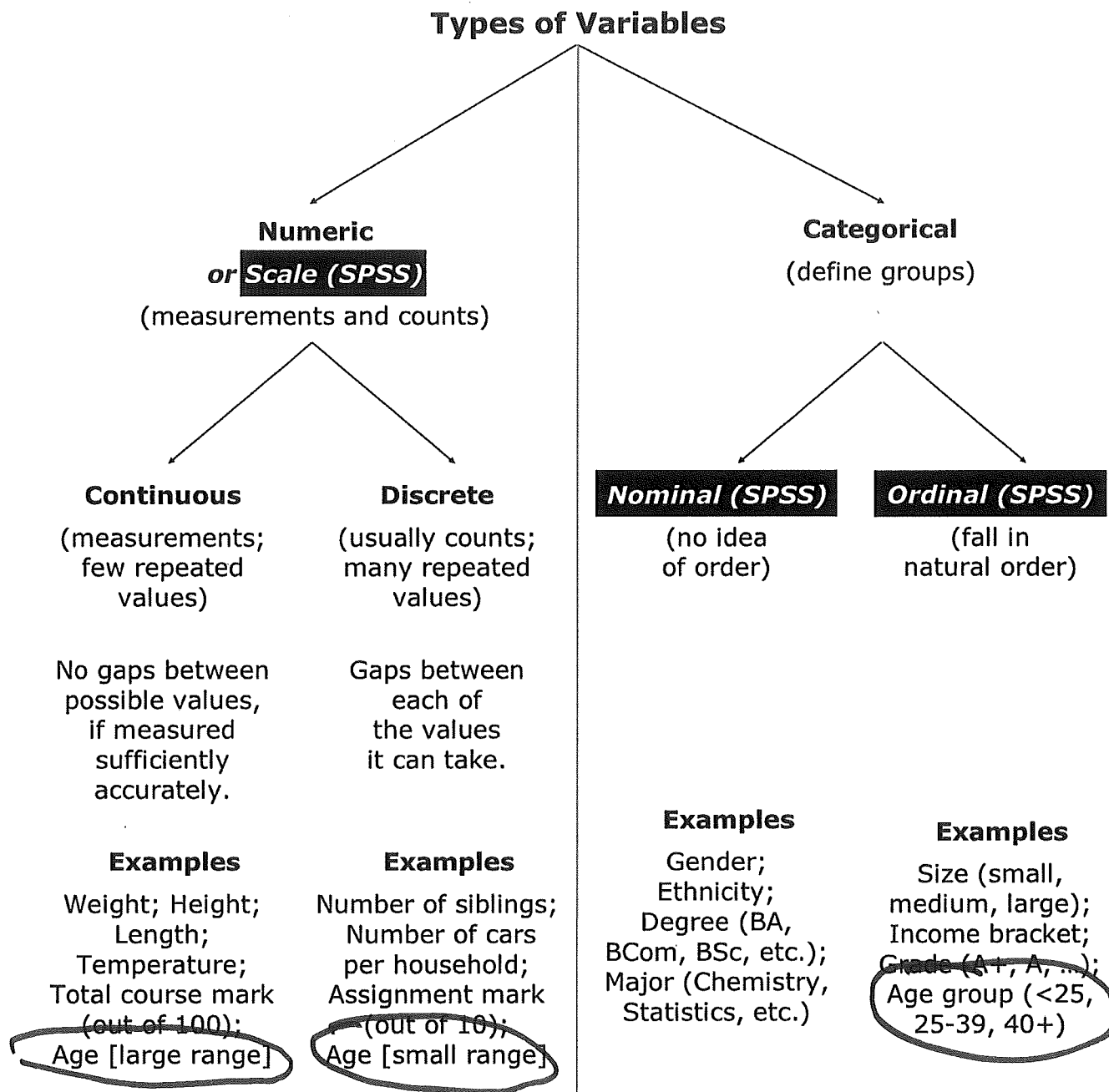
If you haven't used SPSS before, you may find it useful to work your way through this self-paced workshop:

www.tinyURL.com/stats-IS

Exploratory Data Analysis

Tools for Exploring Data

Exploratory data analysis is all about **exploring our data**. We need to use the appropriate tools, though, and to make the correct choices, we need to be clear about what **types of variable/s** our data consists of:



📖 **Useful reference:** Chance Encounters, pages 40 – 42

Useful learning resource: www.learning.statistics-is-awesome.org/dots


Quite often, we start by **entering our data into a table** of some kind:

- **Presentation of Data in Tables**

There are two roles for tables:

1. To convey information quickly and easily.

Guidelines:

- Round drastically
- Arrange the numbers you want compared in columns, not rows
- Sort by appropriately chosen column(s) 
- Use row and/or column averages if appropriate

2. To make data available for detailed checking and/or analysis.

Summarising our data is the next step:

When given a set of raw (numerical) data one of the most useful calculations we can make is finding the **centre** and **spread** of that set of data.

- **Numerical summaries**

• **Centre:** describes the tendency of the observations to bunch around a particular value:

- **Sample mean**, \bar{x} (also known as the average or expected value). The total of all values divided by the total number of values [affected by outliers]
- **Median:** the "middle value". It splits the data in half with half the observations at or above and half at or below [not affected by outliers]
- **Mode**, most frequently occurring number/most common value – not affected by outliers, useful for categorical data

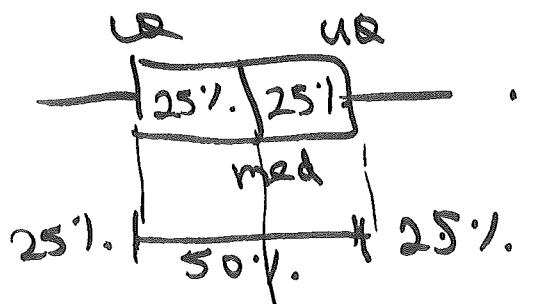
• **Spread:** describes the dispersion of the observed values

- **Sample standard deviation**, σ_{n-1} / s / s_x approximately measures the average of the differences (distances) between the observations and the mean – [affected by outliers]
- **Inter-quartile range (IQR)** gives "the length of the middle half (50%) of the data" [not affected by outliers]

IQR = upper (3rd) quartile – lower (1st) quartile

Note that quartiles come from separating numeric data into 4 groups, each containing equal numbers of values. The lower (1st) quartile is the middle of the lower half of the data and the upper (3rd) is the middle of the upper half of the data.

- **Range** is calculated by largest value – smallest value [affected by outliers]



pg 5-8: (a) - (i)

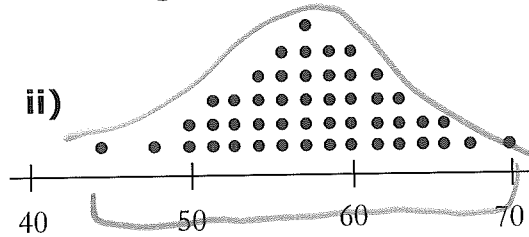
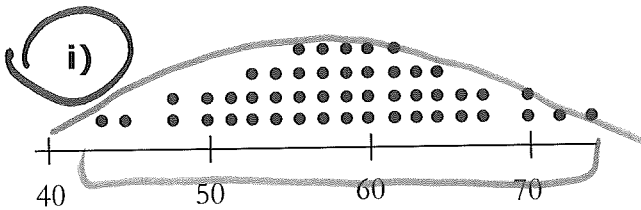
9 Qs \approx 27 mins

restart @ 2.15 pm

Check you understand! Have a go at these problems.

{Note: You don't need to calculate the standard deviations for (a) and (b) - just look at the plots in (a) and the numbers in (b) and think about the spread of the observations - you may find it useful to draw dotplots for (b)!}: **Make sure you have a break!**

(a) Which of the following distributions has the greater standard deviation?

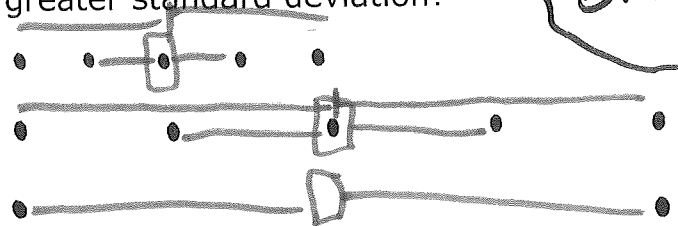


(b) Which of the following lists has the greater standard deviation?

a) 98 99 100 101 102

b) 2 4 6 8 10

c) 2 10



(c) Let Y be the amount the lecturer pays for electricity in a randomly selected thirty-day period and X be the amount of electricity the lecturer uses during that thirty-day period. The variable Y could be treated as:

(1) a categorical variable.

(2) a discrete variable.

(3) a continuous variable. **numeric**

(4) independent of X .

(5) an ordinal variable.

0 1 2 3 4 5 6 7 8 9 10 11 12 13

(d) Recently a travel and parking survey was carried out for University of Auckland staff. Several variables were recorded for each person in the sample. Which **one** of the following statements is **false**?

T (1) Month of birth coded Jan = 1, Feb = 2, ... , Dec = 12 is a categorical variable.

T (2) Staff status coded as 1 = full-time, 2 = part-time is a categorical variable.

(3) The distance travelled to the university, recorded to the nearest 5 kilometres, is a ~~numeric~~ variable. **ord, cate!**

T (4) Continuous variables have few repeated values.

T (5) The length of time it generally takes to find a park, including the time waiting in a queue, is a continuous numeric variable.

(e) In presenting a table to communicate some general features of a set of data, which one of the following statements is **true**?

- F (1) Always order the categories in a table ~~alphabetically~~ *on the column/s of most interest.*
- F (2) Use ~~as little white space as possible~~ *wisely*; ~~compact tables convey information more easily.~~
- T (3) Averages can be helpful for indicating overall patterns in a table.
- F (4) A summary, which highlights the important features of the table, will ~~help~~ *help* ~~confuse~~ the reader.
- F (5) Don't round the original numbers when they are presented in the table, as the rounded numbers will be misleading. ~~Round drastically!~~

Question (f) is about the following information.

Data on 56 hospital births from a single week at the Wellington hospital were collected. Researchers wanted to examine the relationship between various characteristics of the mother and the Apgar Score for the baby. This is a score given to the baby in the first minute after birth and measures the overall physical appearance of the baby.

The characteristics that were measured include:

Age Age of mother in years.

Mass Mass of baby in grams.

Gravida Number of pregnancies including this.

Para Number of births including this.

Term Time to delivery in weeks.

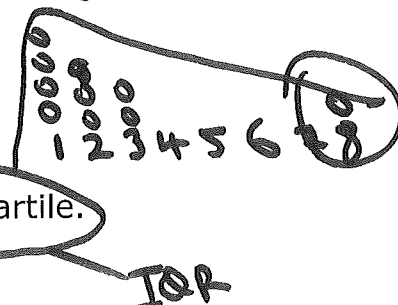
Apgar Apgar score out of 10.

Age	Mass	Gravida	Para	Term	Sex	Apgar
18	3850	1	1	40	M	9
22	2590	1	1	34	M	7
23	3500	1	1	41	M	8
29	2850	2	2	36	F	9
26	3480	3	3	41	M	9
28	3210	2	1	40	F	6
30	3310	8	4	39	F	9
30	4220	3	3	42	M	9
27	4400	2	2	41	M	9
27	2900	1	1	40	F	9

Table: Ten observations from the Apgar Study

(f) Which one of the following numerical summaries is the **best** summary for the ten observations of Gravida reported in the table above?

- (1) The ~~mean~~ and the ~~range~~.
- (2) The ~~median~~ and the ~~standard deviation~~.
- (3) The ~~mode~~ and the ~~median~~.
- (4) The ~~median~~, the ~~lower quartile~~ and the ~~upper quartile~~.
- (5) The ~~mean~~ and the ~~standard deviation~~.



In September 2014 Southern Cross Health Insurance surveyed 1633 adult New Zealanders about their overseas travelling experiences.

One question in the survey asked:

'What is the most annoying thing a nearby passenger can do on a flight?'

The following table shows the responses from the 1633 travellers cross-classified by their age.

<i>cate, nom</i> Annoyance		<i>cate, ord</i> Age (years)				Total
		Under 30	30 to 39	40 to 49	50 and over	
Smell		107	133	117	140	497
Let children misbehave		85	57	78	129	349
Talk loudly		43	47	64	117	271
Recline seat		35	43	39	77	194
Take over armrest		19	22	12	40	93
Other		53	44	57	75	229
Total		342	346	367	578	1633

Table: Most annoying thing a nearby passenger can do on a flight

(g) As used in the table above, which **one** of the following statements about the variables **Annoyance** and **Age** is **true**?

- (1) Both the variables **Annoyance** and **Age** are discrete variables.
- (2) Both the variables **Annoyance** and **Age** are numeric variables.
- (3) The variable **Annoyance** is a discrete variable and the variable **Age** is a numeric variable.
- (4) The variable **Annoyance** is a numeric variable and the variable **Age** is a categorical variable.

(5) Both the variables **Annoyance** and **Age** are categorical variables.

(h) Which **one** of the following statements about numerical summaries for numeric variables is **false**?

N/A ~~(1)~~ This option is no longer examined.

T (2) The sample median is the 50th percentile.

T (3) The interquartile range is not at all sensitive to outliers.

T (4) The sample standard deviation approximately measures the average of the differences between the observations in the sample and the sample mean.

F **(5)** Of the measures of centre, the ~~median~~ *mean* is more sensitive to outliers.

nice definition!

In order to provide a national picture of what is happening in schools and classes, NZCER commissioned a survey of those involved with primary and intermediate schools (Wyllie and Bonne, 2014). The survey was undertaken in July and August 2013 and given to a sample of principals, teachers, Board of Trustees' members and parents. Assume that those chosen were a simple random sample from each of the four groups.

In one part of the survey, all the participants were given a core set of items concerning the challenges facing their school. They were asked to choose which of the items, if any, they believed were the major challenges for their school. The percentages of participants from each of the four groups who felt a particular item was a major challenge are given in Table 4. (Note: Only part of the table is shown.)

Challenge	Principals (n = 180) %	Teachers (n = 713) %	Trustees (n = 277) %	Parents (n = 684) %
Funding	66	60	55	39
Keeping good teachers	25	21	16	28
Large class sizes	18	38	20	24
Improving student behaviour	12	17	11	17
Decreasing bullying	6	8	7	15
Motivating & engaging	21	17	11	13
:	:	:	:	:

Table: Major challenges facing primary and intermediate schools

i.e. sort on this! ↓

(i) Which **one** of the following would **best** improve the table above, if the primary focus was on the major challenges facing principals?

- ☒ (1) Show a table of counts (instead of percentages) and re-order the rows of the table so that the challenges are in alphabetical order.
- ☒ (2) Re-order the rows of the table so that the percentages given for the principals are listed from largest to smallest. *best!*
- ☒ (3) Re-order the columns of the table so that the column for the principals is on the far right. *would be better, but not best*
- ☒ (4) Show a table of counts (instead of percentages) and re-order the columns of the table so that the column for the principals is on the far right.
- ☒ (5) Re-order the rows of the table so that the challenges are in alphabetical order.

Displaying/graphing your data

Common displays of data include tables and graphs such as dotplots, stem-and-leaf plots, boxplots, histograms and bar charts. Choosing which of these to use will depend on the type of variable/s you have collected and the relationships you are attempting to explore.

Displaying continuous numeric variables

Display tools:

- Appropriate plots for continuous numeric variables are:

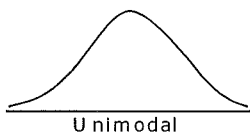
- Dot plot – data sets of any size, n can be anything
- Stem-and-leaf plot – moderate data sets, $15 \leq n \leq 150$
- Box plot – moderate to large data sets, $n \geq 20$
- Histogram – large data sets, $n \geq 50$

- Features to look for and comment on in the above plots are:

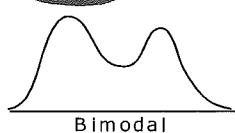
- Centre and spread

- Modality – How many modes/peaks does the data have?

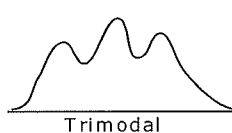
not box!



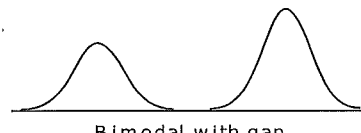
Unimodal



Bimodal

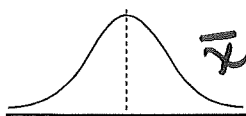


Trimodal



Bimodal with gap

- Symmetry or skewness – Is the data symmetric or skewed?

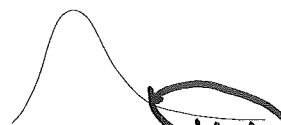


Symmetric



Symmetric

$\bar{x} = med$



Positive/Right skew
(longer upper tail)

$\bar{x} > med$

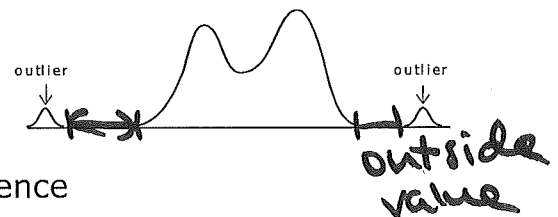


Negative/Left skew
(longer lower tail)

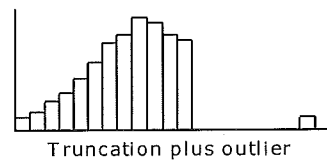
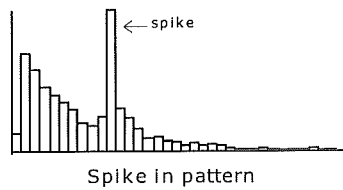
$\bar{x} < med$

- Outliers:

- Are observations which are far from the bulk of the data
- Search for a reason for their existence
- Only delete outliers if they are found to be mistakes



- Abrupt changes:



📖 **Useful reference:** Chance Encounters, pages 58 – 60

Displaying discrete numeric variables

Display tools:

- **Frequency table**

Number of cars per household

	<i>Frequency</i>	<i>Percent</i>	<i>Cumulative Percent</i>
1	64	62.1%	62.1%
2	16	15.5%	77.7%
3	10	9.7%	87.4%
4	5	4.9%	100.0%
Total	103	100.0%	

- **Bar graph**

- 2D vs ~~3D~~ – Always use 2D!
- Similar to histogram (for continuous data) except bars/rectangles are not joined up.
- On the vertical axis EITHER use:
 - frequency to show the actual counts from the sample,

OR

 - percentage to show an estimate of the distribution of the population.
- Bar graphs are very good for presenting relative sizes.

Displaying categorical variables

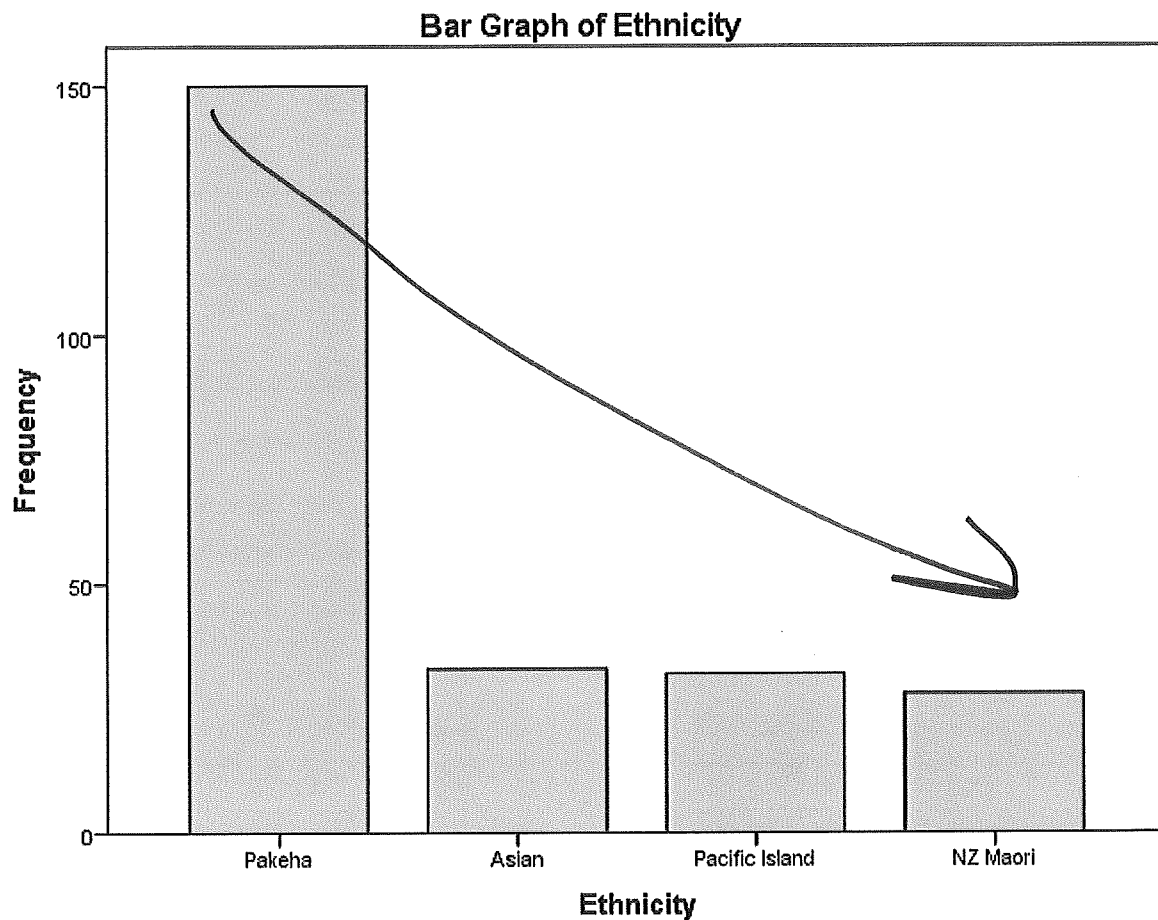
Display tools:

- **Frequency Table**

Ethnicity				
	Frequency	Percent	Valid Percent	Cumulative Percent
Pakeha	150	61.7	61.7	61.7
Asian	33	13.6	13.6	75.3
Valid Pacific Island	32	13.2	13.2	88.5
NZ Maori	28	11.5	11.5	100.0
Total	243	100.0	100.0	

- Used in exactly the same way as for discrete variables.
- Frequency tables often just include the value and frequency columns.

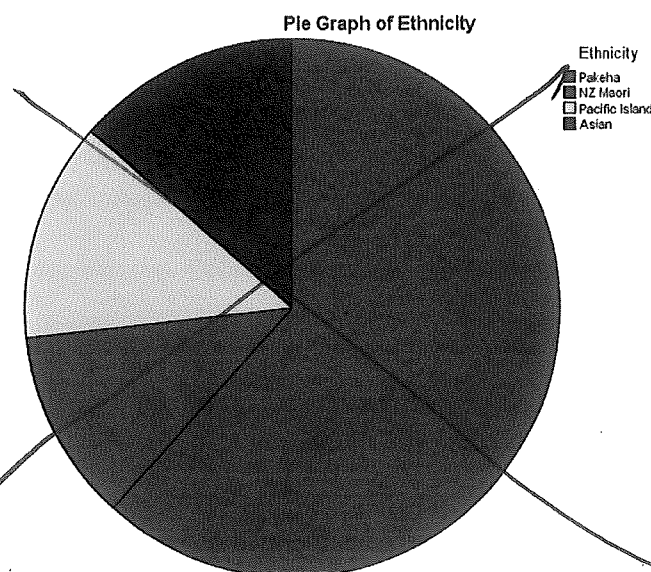
- **Bar Graph**



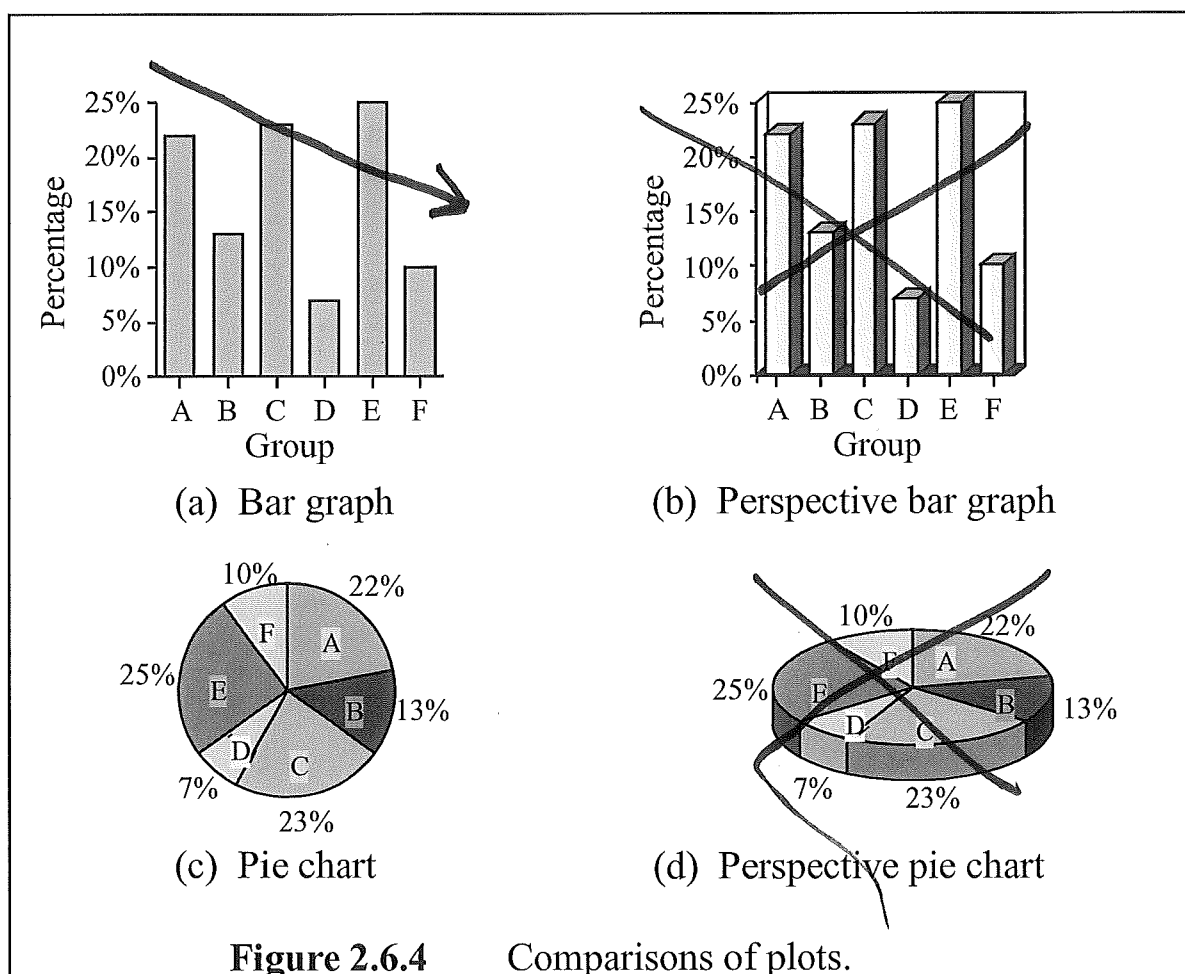
- Order categories sensibly, usually by size (i.e., by frequency unless there is some very compelling reason for some other ordering).
- Avoid using perspective (3-D) bar graphs.

Other forms of plots for categorical variables:

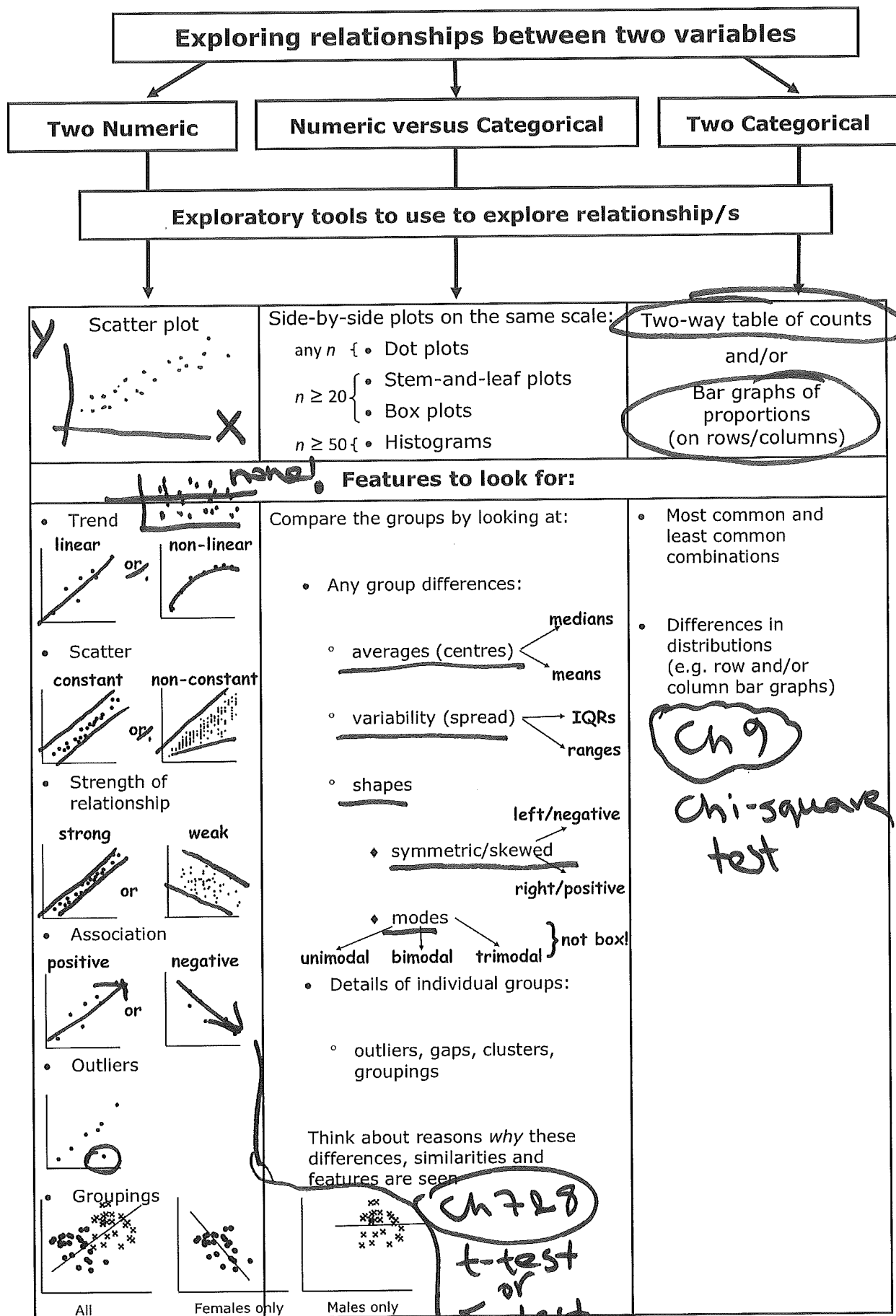
- Segmented bar charts are a better option than pie charts.
- Pie charts do not communicate information as well as bar graphs. Avoid using them!
- Perspective (3-D) pie charts are disastrous – never use them!



Useful reference: Chance Encounters, pages 75 – 79 and pages 81 – 86



From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.



Ch 10

regression & correlation

Ch 7 & 8

t-test or F-test

Exploratory Data Analysis: Questions

1. The table below is taken from an overview article on business in Asia in the March 9th 1996 issue of the *Economist*.

	Population 1994, m	GDP per head at PPP 1994, US\$	% of population under 25 1995 est.
Pakistan	126.3	2,210	62.7
India	913.6	1,290	54.1
Sri Lanka	18.1	3,150	49.2
Bangladesh	117.8	1,350	60.9
Myanmar	45.6	751*	57.1
Thailand	58.7	6,870	48.9
Cambodia	10.0	1,250*	61.4
Vietnam	72.5	1,010*	57.3
Laos	4.7	1,760*	63.0
China	1,190.0	2,510	44.5
Hong Kong	5.8	23,080	33.8
Taiwan	21.1	13,022**	42.2
North Korea	23.5	3,026*	48.1
South Korea	44.6	10,540	42.2
Japan	124.8	21,350	31.2
Malaysia	19.5	8,610	56.1
Singapore	2.8	21,430	37.4
Philippines	66.2	2,800	58.3
Indonesia	189.9	3690	53.8

* 1992

** Estimate

Which one of the following statements is most appropriate?

The table is an example of:

- X (1) a ~~well~~ ^{badly} presented table because the order of the countries is determined according to geography: roughly east to west.
- (2) A badly presented table because its entries are not ordered by magnitude according to population, GDP or % of young people.
- ? (3) A well presented table because it uses the white spaces well.
- X (4) A badly presented table because there is no 'average column'.
- X (5) A ~~badly~~ ^{well} presented table because the GDP is measured in \$US for all countries rather than the local currency.

2. Which **one** of the following statements is **false**?

- T (1) The interquartile range is much less sensitive to the presence of outliers than the range.
- T (2) The range, interquartile range, and sample standard deviation measure the spread of the data.
- T (3) To help characterize a distribution of data, both a measure of 'centre', and a measure of 'spread' are useful.
- F (4) The interquartile range ~~can~~ be seriously affected by an outlier. *won't*
- T (5) The sample mean can be seriously affected by an outlier.



Questions 3 to 5 refer to the following information.

The California Department of Development Services (DDS) is responsible for providing services and support to people with developmental disabilities. A data set (Taylor and Mickel, 2014) was designed to represent a random sample of clients to whom the DDS has provided services and support.

The variables were:

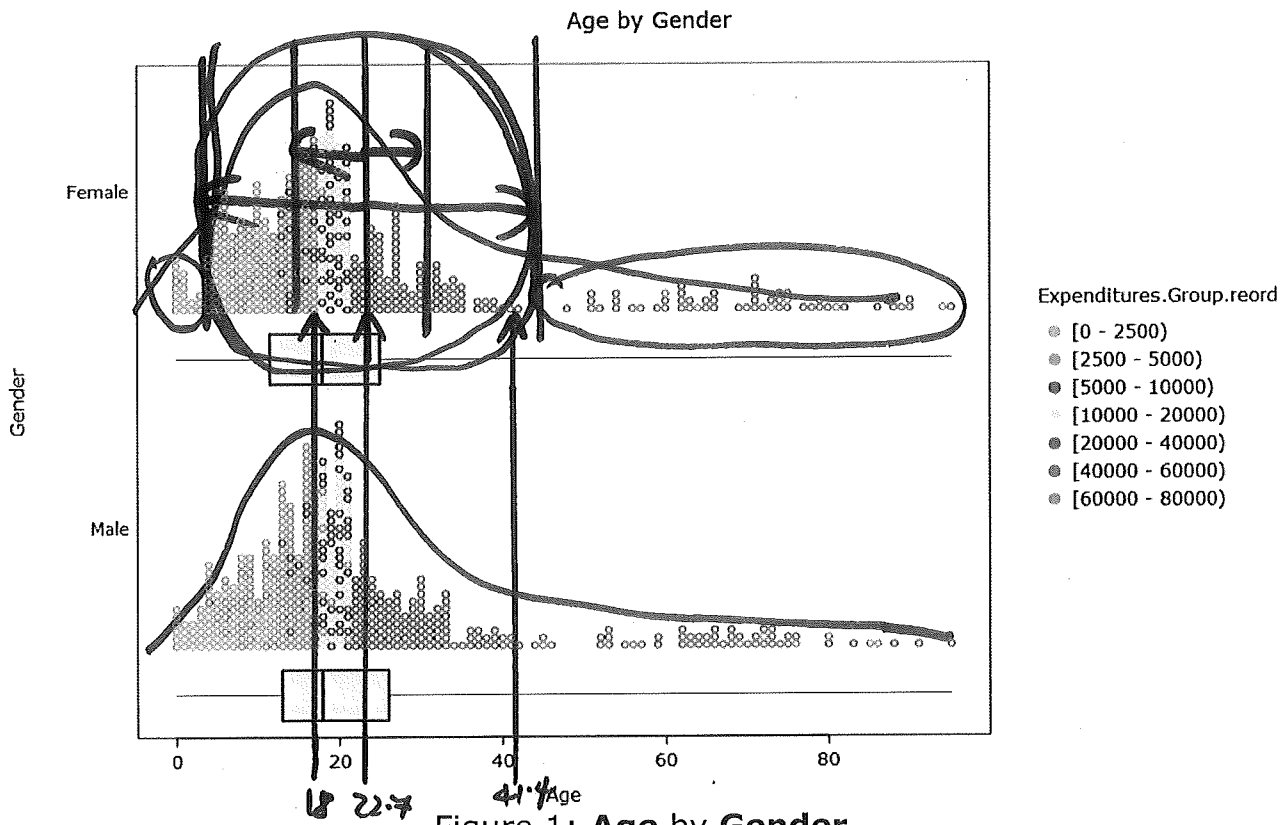
Age	The age of the client, in years	<i>num, cont</i>
Gender	The gender of the client	
	– Female	<i>cate, nom</i>
	– Male	
Ethnicity	The ethnicity of the client	
	– White	
	– Hispanic	
	– Black	<i>cate, nom</i>
	– Multi Race	
	– Asian	
	– American Indian	
	– Native Hawaiian	
	– Other	
Expenditures	The amount of money the client received per year from the DDS, in dollars	<i>num, cont</i>

3. Which **one** of these displays would be the **most suitable** to explore the relationship between the variables **Age** and **Expenditures**?

- (1) Side-by-side dot plots and box plots *num*
- (2) Side-by-side histograms
- (3) Two-way table of counts
- (4) Scatter plot *num*
- (5) Stacked bar charts

Questions 4 and 5 refer to the following additional information.

Side-by-side dot plots and box plots of ^{num}Age and ^{cate}Gender are shown in Figure 1.



4. Which **one** of the following statements is **false**?

- T
T
T
T
F
- (1) The interquartile range of the ages is similar for the males and the females.
 - (2) The average age of the female clients is similar to the average age of the male clients.
 - (3) The clients under 20 years old received less than \$40,000 per year from the DDS.
 - (4) Most of these clients are under 40 years old.
 - (5) For both genders, the distribution of the ages is ^{posi}negatively (left) ^{right}skewed.

5. Which one of the following is the **only** possible **correct** pair of values for the mean and standard deviation of the ages of the **female** clients?

-
- (1) ~~mean = 18.0~~ ~~standard deviation = 13.5~~
 - (2) ~~mean = 18.0~~ ~~standard deviation = 95.0~~
 - (3) mean = 22.7 ~~standard deviation = 5.4~~
 - (4) ~~mean = 41.4~~ ~~standard deviation = 19.0~~
 - (5) mean = 22.7 ~~standard deviation = 19.0~~

Q6-19, also have a break! final discussion @ 3:45pm...

Questions 6 and 7 refer to the following information.

The Human Development Index (HDI) of each country is calculated each year by the United Nations.

The 2014 Human Development Report is available at <http://hdr.undp.org/en>.

Some of the variables reported for each country were:

MaleEmployment	The percentage of males aged 15 years and above who are in paid employment	num, cont
FemaleEmployment	The percentage of females aged 15 years and above who are in paid employment	num, cont
SecondaryEducation	The percentage of those aged 25 years and above who have at least some secondary education categorised into the following groups	
	– 20% or less	Cate, ord
	– 20.1%–40.0%	
	– 40.1%–60.0%	
	– 60.1%–80.0%	
	– more than 80%	
HealthExpenditure	The amount the country spends on health as a percentage of gross domestic product	num, cont
HDI	HDI classification	
	– Low	Cate, ord
	– Medium	
	– High	
	– Very High	

6. Which **one** of the following displays would be the **most suitable** to explore the relationship between the variables **HealthExpenditure** and **HDI**? *cate*

- (1) Stacked bar charts
- (2) Side-by-side dot plots and box plots
- (3) Tile density plot
- (4) Scatter plot
- (5) Two-way table of counts

7. Figure 2 was created to explore possible relationships between **MaleEmployment**, **FemaleEmployment** and **SecondaryEducation**.

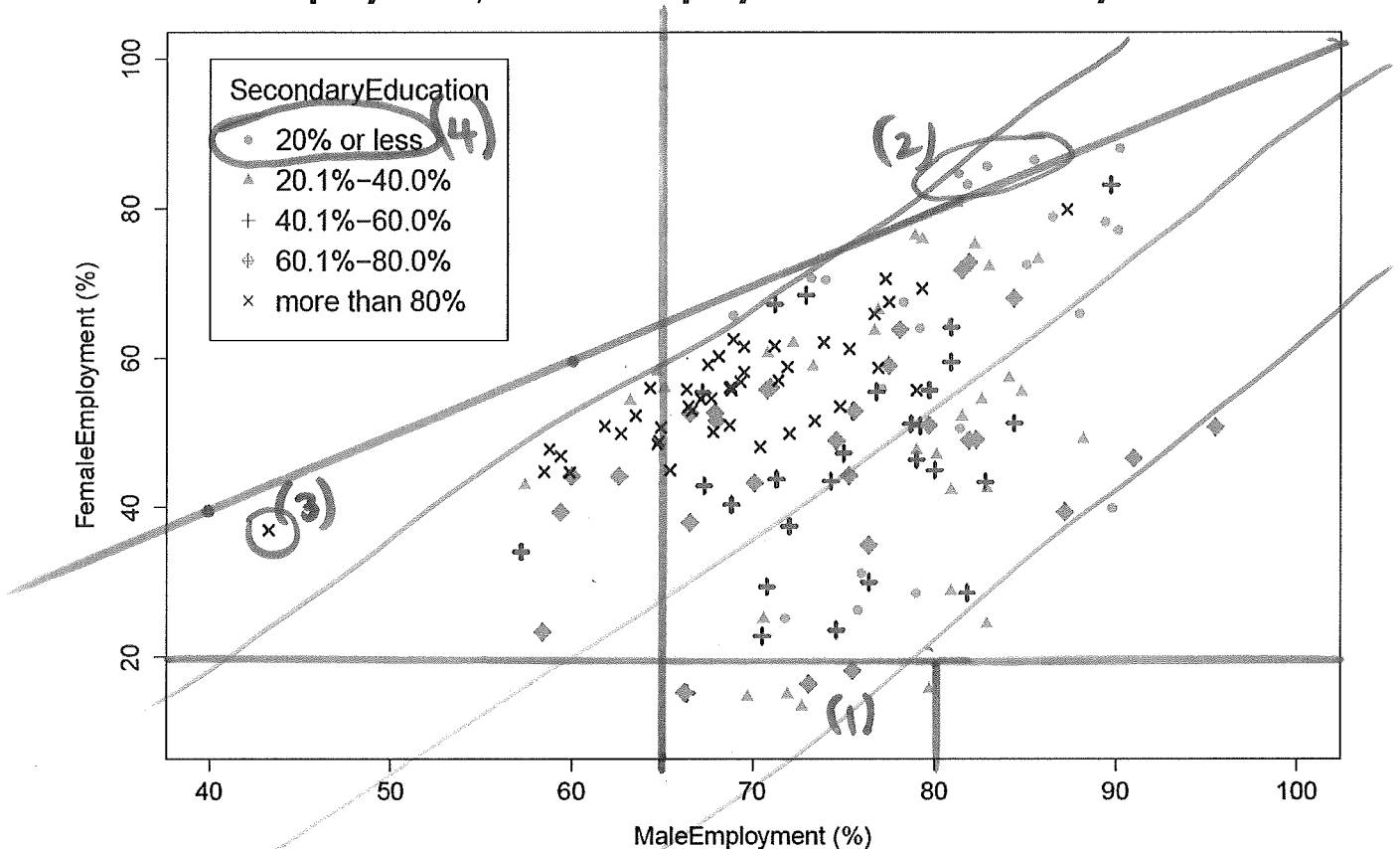


Figure 2: **FemaleEmployment** vs **MaleEmployment**

Which **one** of the following statements is **false**?

- T (1) The countries with less than 20% of females aged 15 years and above in paid employment have between 65% and 80% of males aged 15 years and above in paid employment.
- T (2) For most countries the percentage of males aged 15 years and above who are in paid employment is greater than the corresponding percentage of females aged 15 years and above.
- T (3) The country with the lowest percentage of males aged 15 years and above in paid employment has more than 80% of people aged 25 years and above with at least some secondary education.
- T (4) The countries with 20% or less of people aged 25 years and above with at least some secondary education have more than 65% of males aged 15 years and above in paid employment.
- F (5) There is a strong ~~break~~ linear relationship between the percentage of males aged 15 years and above who are in paid employment and the corresponding percentage of females aged 15 years and above.

Question 8 refers to the following information.

A study (Chijiwa *et al.*, 2015) was conducted to investigate whether or not domestic dogs evaluate humans interacting with one another. Fifty-four domestic dogs and their owners participated in the study. The owners were not told the purpose of the study.

Each dog, along with its owner, was randomly allocated to one of three groups of 18: a control group, a helper group and a nonhelper group. Each dog and its owner participated in four trials under identical conditions.

In each trial the owner and their dog sat between an actor and a neutral person. The owner then tried to take the lid off a container. For those in the helper and nonhelper groups, the owner had been instructed to ask the actor for help to take the lid off. Those in the helper group received help from the actor, while for those in the nonhelper group the actor refused to help. Those in the control group tried to take the lid off but did not ask for help. For all three groups the container with the lid on was then put down and, while the owner watched, both the actor and the neutral person offered the dog a treat at the same time.

The number of times, out of the four trials, that each dog chose to take the treat from the actor (chose the actor) is shown in Figure 3 below. Also shown is the mean for each group (solid line) and the overall mean (dashed line).

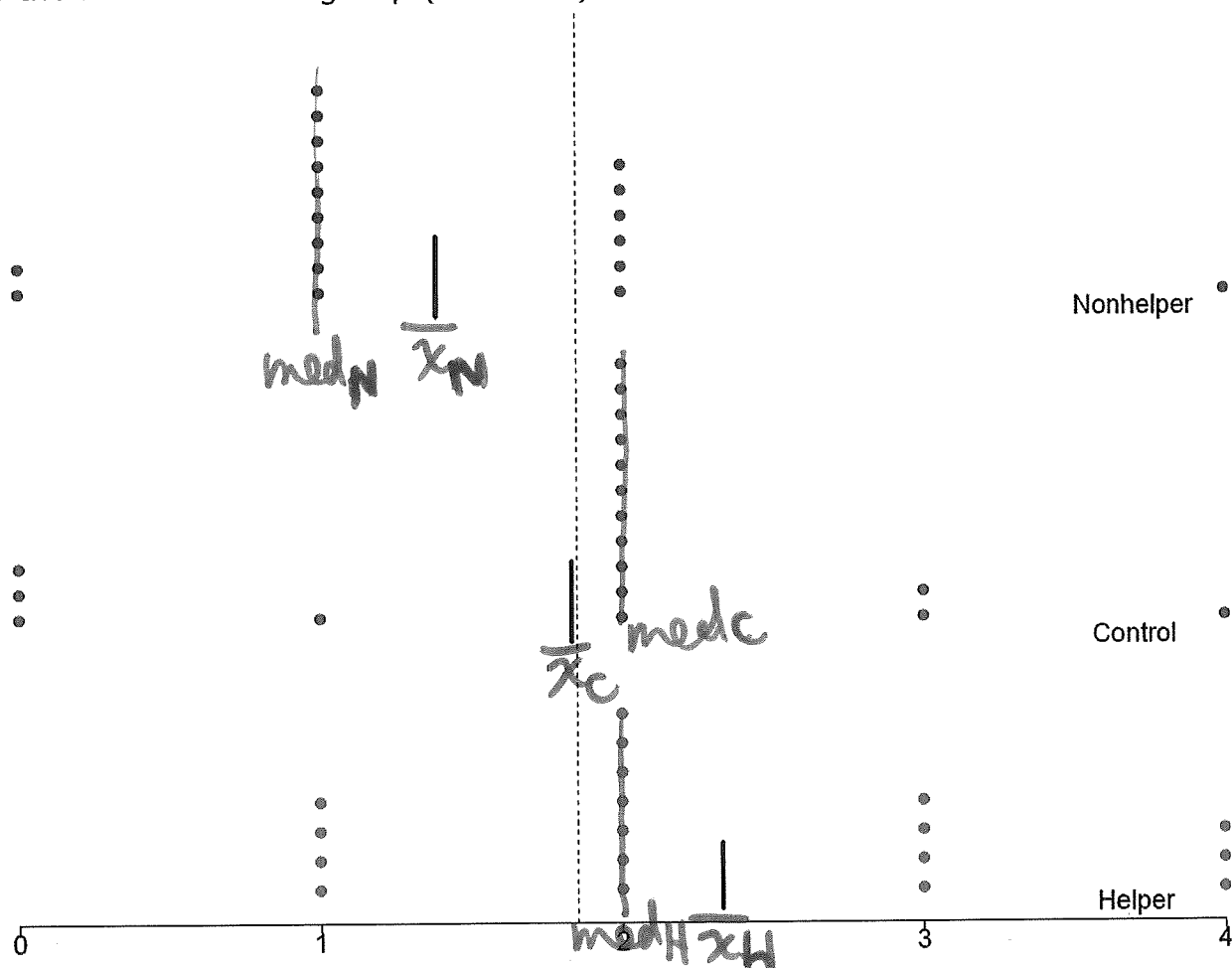


Figure 3: Number of times actor chosen

8. Which **one** of the following statements about the data displayed in Figure 3 is **false**?

T
T
T
F
T

- (1) The helper group has the highest observed mean.
- (2) If the three observed medians were calculated, the nonhelper group would have the lowest observed median.
- (3) The helper group has the lowest observed range.
- (4) At least one dog in each group never chose the actor.
- (5) At least one dog in each group always chose the actor.

Question 9 refers to the following information.

In New Zealand about 3500 people work in community pharmacies and a register of all pharmacists working in New Zealand is maintained. A random sample of pharmacists were asked about the quality of their work life and several measures were recorded. One such measure was called a **Compassion satisfaction** score. This was a measure (from 0 to 50) of the pleasure derived from being able to do their job well. Higher scores represent greater satisfaction. Two other measures recorded were the gender of the pharmacist and the location where the pharmacist worked: City, Suburban, Rural.

cate, non
num, cont
Cate, non

9. Which **one** of the following statements about the measures (variables) described above is **true**?

X
X
X
X
X

- (1) **Gender, Location and Compassion satisfaction** are categorical.
- (2) **Gender and Location** are categorical and **Compassion satisfaction** is numeric.
- (3) **Gender** is ordinal and **Location** and **Compassion satisfaction** are numeric.
- (4) **Gender** is nominal and **Location** and **Compassion satisfaction** are numeric.
- (5) **Gender and Location** are ordinal and **Compassion satisfaction** is nominal.

10. Which **one** of the following statements is **false**?

T
T
T
T
F

- (1) The scatter plot is a useful tool for investigating relationships between two continuous variables.
- (2) Dot plots should be used for small numbers of observations.
- (3) Box plots are good at comparing centres and spreads of numeric data for two or more groups.
- (4) Bar graphs can be used to display discrete data.
- (5) Histograms should be used for small numbers of observations.

large!

Questions 11 to 13 refer to the following information.

Taste.com.au is a food website that provides free recipes. In January 2016, data were scraped from this website to create a data set of 440 *main meal* recipes so that the relationships between different variables could be explored.

Some of the variables selected were:

Total time	The time to prepare and cook the main meal, in minutes <i>num, cont</i>
Cooking time	The time to cook the main meal, in minutes <i>num, cont</i>
Difficulty	The difficulty rating of the recipe (Easy, Hard) <i>cate, ord</i>
Chicken	Whether the main meal contains chicken or not (Chicken, No chicken) <i>cate, nom</i>

iNZight was used to explore the variable **Total time** and its relationship with other variables in the data set. The four different plots created are shown in Figure 4.

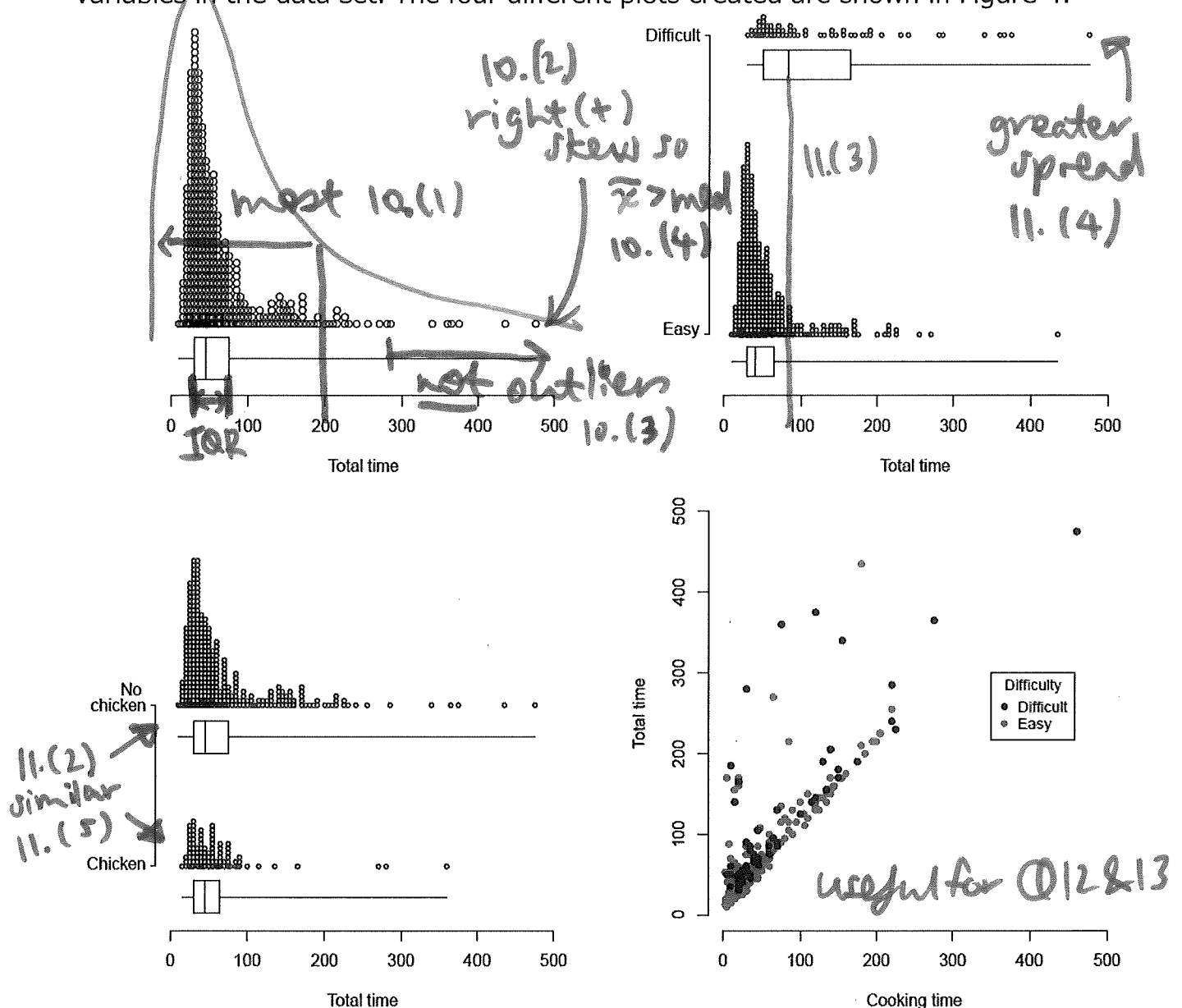


Figure 4: Exploratory data analysis for **Total time**

11. For these recipes which **one** of the following statements is **false** when considering **Total time** by itself?

T
F
T
T
T
T

- (1) Most of recipes have a total time of less than 200 minutes.
- (2) The total times are negatively (left) skewed. *negatively*
- (3) The total times over 300 minutes would not be considered outliers.
- (4) The mean total time is higher than the median total time.
- (5) The interquartile range of the total times is around 45 minutes.

12. Which **one** of the following statements about these recipes is **false**?

F
T
T
T
T

- (1) For both difficult and easy recipes, there is a positive association between the total time and the cooking time with a small amount of scatter. *large*
- (2) Knowing whether the recipe contained chicken or not does not appear to help you predict the total time.
- (3) The recipes that are rated difficult have longer total times on average than the recipes that are rated easy.
- (4) The standard deviation of total time is higher for the recipes that are rated difficult than the recipes that are rated easy.
- (5) The distribution of total times are similar for the recipes that contain chicken and the recipes that do not contain chicken.

13. Which **one** of the following statements is **false**?

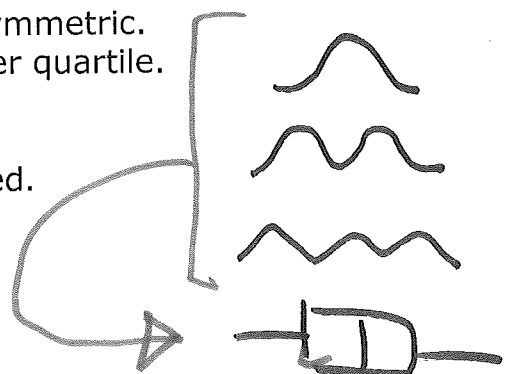
T
T
F
T
T

- (1) The relationship between **Total time**, **Cooking time** and **Difficulty** could also be explored using subsetting.
- (2) **Difficulty** has been added to the scatter plot of **Total time** vs **Cooking time** using colour.
- (3) Side-by-side dot plots could be used to explore the relationship between **Chicken** and **Difficulty**.
- (4) The variable **Cooking time** could be used to create a new categorical variable by defining groups based on times *day 1/2 hour increments!*
- (5) A histogram would be suitable to display **Total time** by itself. *n=440!*

14. Which **one** of the following characteristics **cannot** be detected by looking at a box plot of the data?

✓
✓
X
✓
✓

- (1) That the sample is approximately symmetric.
- (2) That the median is close to the upper quartile.
- (3) That the sample has a single mode.
- (4) That there are outliers.
- (5) That the sample is negatively skewed.



Questions 15 to 19 refer to the following information.

An experiment presented participants with two images of the same shape at different orientations. To obtain the second image, the original image was first rotated at a randomly selected angle between 0 and 180 degrees (at 20° steps). The second image was either this rotated image (Same) or a mirror image of it (Mirror). The observer's task was to press the letter S or the letter M depending on whether the two images were the same or mirror images of each other (Howell, 2011).

Below are examples of the two stimuli. Box 1 is an example where the stimulus is "Same" (the second image in box 1 has been rotated only) and Box 2 is an example where the stimulus is "Mirror" (as well as being rotated, the second image in box 2 is a mirror image.) Table 1 shows some of the results from 600 trials.

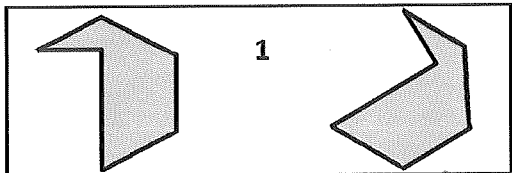
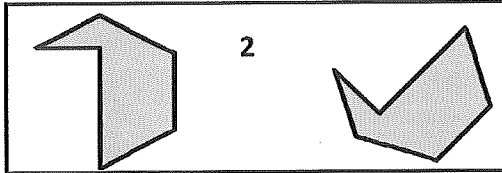
				
<i>label</i>	<i>num, dis</i>	<i>cate, non</i>	<i>cate, non</i>	<i>num, cont</i>
Trial	Angle	Stimulus	Accuracy (1 = Correct 0 = Wrong)	Reaction Time (Seconds)
1	140	Same	1	4.42
2	60	Same	1	1.75
3	180	Mirror	1	1.44
4	100	Same	0	1.74
5	160	Mirror	1	1.94
...
600	40	Mirror	1	1.12

Table 1: Results from 600 trials

15. Which **one** of the following **best** describes the types of variable for **Stimulus**, **Accuracy** and **Reaction Time**?

- ☒ (1) **Stimulus** is categorical, **Accuracy** is numeric and **Reaction Time** is continuous.
- ☐ (2) **Stimulus** is categorical, **Accuracy** is categorical and **Reaction Time** is continuous.
- ☒ (3) **Stimulus** is categorical, **Accuracy** is categorical and **Reaction Time** is discrete.
- ☒ (4) **Stimulus** is discrete, **Accuracy** is numeric and **Reaction Time** is continuous.
- ☒ (5) **Stimulus** is discrete, **Accuracy** is discrete and **Reaction Time** is continuous.

16. Suppose we are interested in seeing if there is a relationship between **Stimulus** and **Accuracy**. The **most** appropriate plots to use would be:

- ☒ (1) Side by side box plots of **Accuracy** for each level of **Stimulus**.
- ☒ (2) Side by side dot plots of **Stimulus** for each level of **Accuracy**.
- ☒ (3) A scatter plot of **Accuracy** against **Stimulus**.
- ☒ (4) Side by side dot plots of **Accuracy** for each level of **Stimulus**.
- ☐ (5) Bar charts of the proportions of **Accuracy** for each level of **Stimulus**.

17. Figure 5 shows the reaction times for both types of stimulus; Same and Mirror.

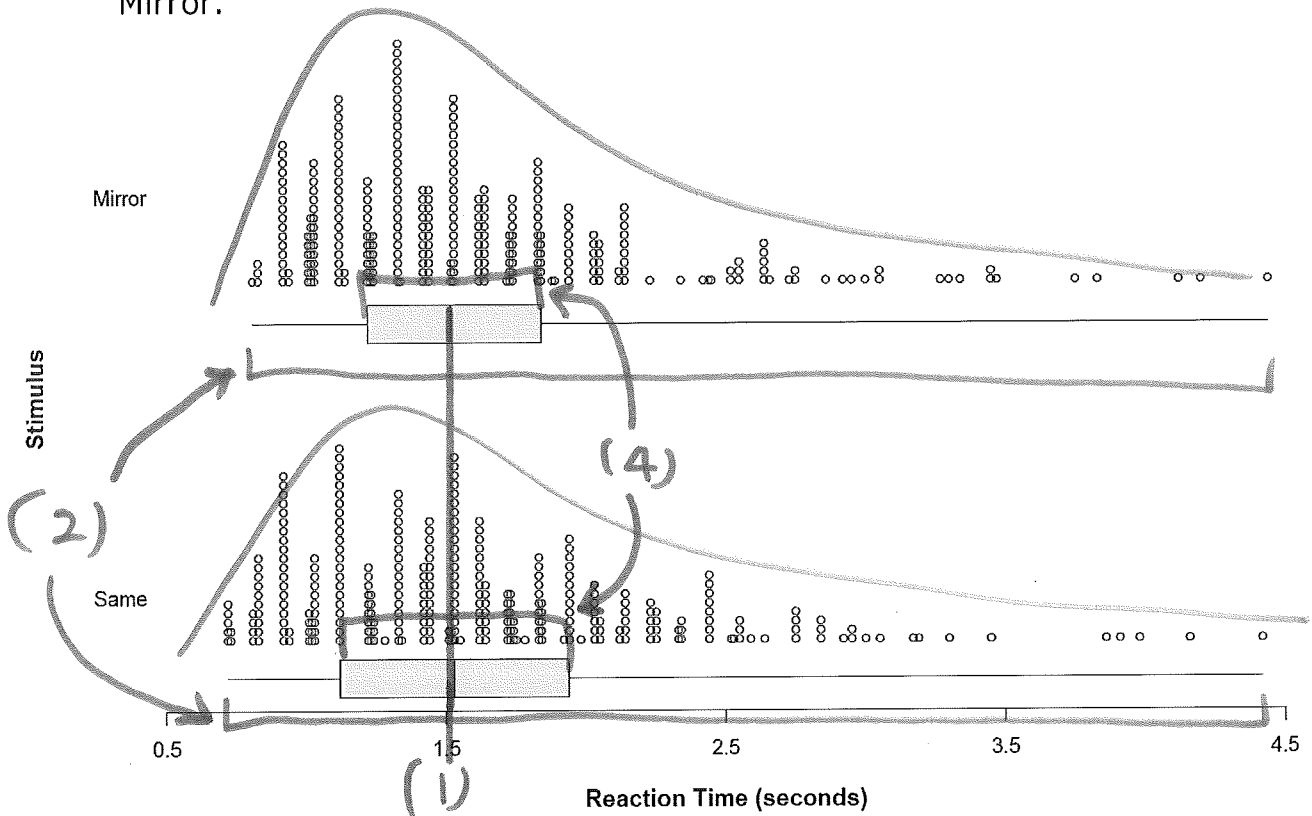


Figure 5: Distribution of Reaction Time by Stimulus

Which **one** of the following statements concerning the plots is **false**?

- T (1) The median reaction times for each stimulus type are similar.
- T (2) The range of the reaction times for each stimulus type are similar.
- T (3) For each stimulus type the distribution of reaction time appears to be unimodal.
- T (4) The interquartile range of the reaction time is smaller when the stimulus is **Mirror** than it is when the stimulus is **Same**.
- F (5) For each stimulus type the mean reaction time is lower than its median.

→ due to positive/right skewness! ^{higher}

Questions 18 and 19 refer to the following additional information.

The reaction times by stimulus are further subsetting by the accuracy of the results.

18. Using Table 2 below, which **one** of the following statements is **false**?

T
F
T
T
T

- (1) When the response was correct, the mean reaction time if the stimulus was **Mirror** was similar to that if the stimulus was **Same**.
- (2) Approximately 28% of the responses that were wrong were when the stimulus was **Mirror**. $\frac{286}{286+12}$
- (3) The experiment consisted of 298 times when the stimulus was **Mirror** and 302 times when the stimulus was **Same**. $\frac{286+12}{286+12+43}$
- (4) On average, when the stimulus was **Mirror**, the reaction time was longer if the response was wrong than if it was correct.
- (5) Approximately 91% of the responses were correct. $\frac{286+259}{600} = \frac{545}{600} \times 100\% = 90.8\% (1dp)$

iNZight Summary									
Primary variable of interest: Reaction Time									
Secondary variable: Stimulus									
Subset by: Accuracy									
Total number of observations: 600									
Summary of Reaction Time by Stimulus, for Accuracy = Correct:									
	Min	25%	Median	75%	Max	Mean	SD	Sample Size	
Mirror	0.81	1.14	1.44 (4)	1.83	4.44	1.604 (4)	0.6192	286	
Same	0.72	1.12	1.53	1.94	4.42	1.619	0.6520	259	
Summary of Reaction Time by Stimulus, for Accuracy = Wrong:									
	Min	25%	Median	75%	Max	Mean	SD	Sample Size	
Mirror	1.33	1.635	1.735 (4)	2.367	3.45	2.085 (4)	0.7529	12	
Same	0.72	1.330	1.530	1.890	3.17	1.664	0.6112	43	

Table 2: Distribution of **Reaction Time** by **Stimulus** subsetting by **Accuracy**

$$\frac{12}{12+43} = \frac{12}{55} \times 100\% = 21.8\% (2)$$

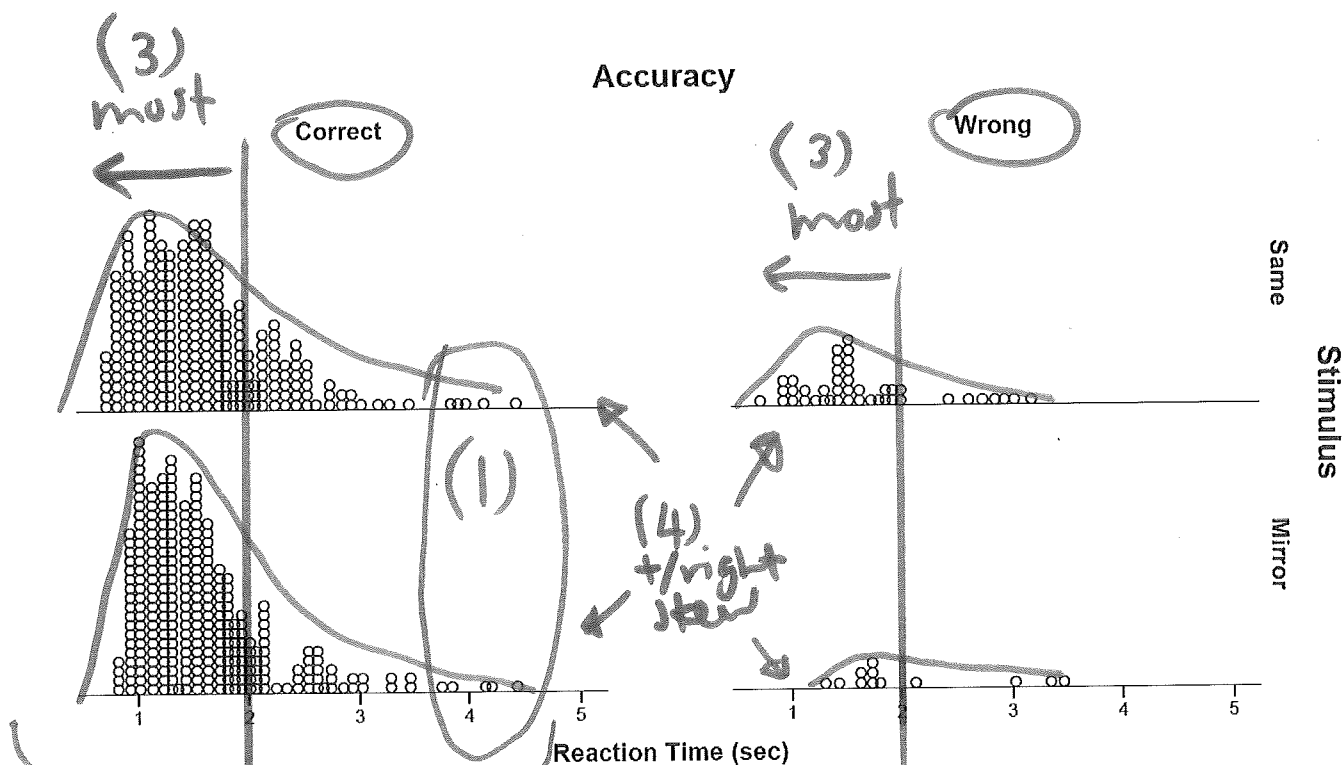


Figure 6: Distribution of Reaction Time by Stimulus subsetting by Accuracy

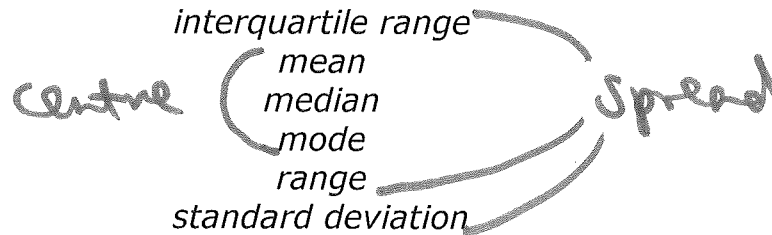
19. Using Figure 6 above, which **one** of the following statements is **false**?

- T (1) The longest reaction times occurred when **Accuracy** was correct.
- F (2) When the response is correct, **Reaction time** is clearly ^{not} dependent on **Stimulus**. → similar shapes, similar summary stats!
- T (3) For all combinations of **Accuracy** and **Stimulus**, the majority of the reaction times were less than 2 seconds.
- T (4) For all combinations of **Accuracy** and **Stimulus** the distributions of reaction time appear to be positively skewed.
- T (5) For both **Stimulus** types, the majority of the responses were correct.

20. A study was carried out to investigate the association between bank interest rates and mortgage interest rates. To explore this relationship the most appropriate display would be a:

- (1) Box plot of bank interest rates and a box plot of mortgage interest rates.
- (2) Histogram of bank interest rates and a histogram of mortgage interest rates, on the same scale.
- (3) Dot plot of bank interest rates and a dot plot of mortgage interest rates.
- (4) Scatter plot of bank interest rates and mortgage interest rates.
- (5) Dot plot of the differences between bank interest rates and mortgage interest rates. • maybe?

Questions 21 and 22 refer to the following concepts:



21. Which one of the following statements is **true**?

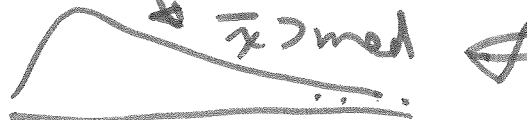
F
F
T
F
F

- (1) Only the interquartile range and the standard deviation are measures of spread.
- (2) All six give us some information about the spread.
- (3) Only the range, the interquartile range and the standard deviation are measures of spread.
- (4) Only the standard deviation tells us about the spread.
- (5) The interquartile range and the range give us the same information.

22. Which **one** of the following statements is **false**?

T
T
F
T
T

- (1) The median is the point such that half of the observations are no larger than it and half are no smaller.
- (2) A distribution can have several modes but only one mean.
- (3) The range is ~~not~~ affected by outliers. *max - min*
- (4) If a distribution is positively skewed then the median will be smaller than the mean.
- (5) If a distribution is symmetric then the mean and the median are about the same.



23. In a recent report on employees' attitudes towards employment law it was stated that the mean hourly rate for salary and wage earners is \$26.92 per hour. The median hourly rate for salary and wage earners is believed to be less than \$20 per hour. Assuming this belief to be true, the **best** explanation of why the difference between these two rates is so large is:

- (1) A mistake must have been made in calculating the mean hourly rate.
- (2) A relatively large number of wage and salary earners have an extremely high hourly rate.
- (3) The sample of wage and salary earners used to determine the mean hourly rate must have been non-random.
- (4) A relatively small number of wage and salary earners have an extremely high hourly rate.
- (5) The distribution of the hourly rates of wage and salary earners is symmetric.

Questions 24 and 25 refer to the following information.

Sports Foundation grants for sports which won the right to represent New Zealand at the Sydney Olympics are shown in the table below.

No.	Sport	1998-1999	1999-2000	2000-2001
01	Archery	\$16,500	\$15,000	\$25,000
02	Athletics	\$515,340	\$485,400	\$299,000
03	Basketball	\$133,100	\$90,000	\$40,000
04	Boxing	\$166,950	\$55,000	\$44,350
05	Cycling	\$678,500	\$747,182	\$688,140
06	Equestrian	\$691,000	\$717,000	\$558,620
07	Gymnastics	\$94,500	\$34,500	\$22,400
08	Hockey	\$498,500	\$478,460	\$554,000
09	Judo	\$153,650	\$93,179	\$124,500
10	Rowing	\$533,100	\$466,700	\$707,265
11	Shooting	\$327,000	\$106,000	\$405,616
12	Softball	\$251,913	\$425,259	\$254,542
13	Swimming	\$431,470	\$205,000	\$280,594
14	Table Tennis	\$26,250	\$3,000	\$29,000
15	Triathlon	\$343,110	\$548,255	\$86,300
16	Weightlifting	\$98,900	\$48,125	\$79,500
17	Wrestling	\$13,520	\$8,000	\$15,000
18	Yachting	\$947,000	\$1,131,000	\$622,356

av. col.
✓
(all same units)

av. row ✓

Table: Sports Foundation Grants

24. Suppose the purpose of this table was to convey the information so that the reader could make visual comparisons between different sports with respect to the size of the grant awarded. **One** change in the presentation of the data which would **not** be an improvement would be to:

- ✓ (1) interchange the rows and columns in the table. *columns = good!*
- ✓ (2) round all grants to the nearest thousand dollars. *round drastically!*
- ✓ (3) list the sports in order of the amount of the grant received in the year 2000-2001. *sort on most recent info!*
- ✓ (4) add a column on the right of the table for the 'Average Amount Awarded per Year (1998-2001)'.
- ✓ (5) add a row at the bottom of the table for the 'Average Amount Awarded per Sport'.

25. The figure below is a dot plot of the grants for each of the 18 sports in the year 2000–2001.

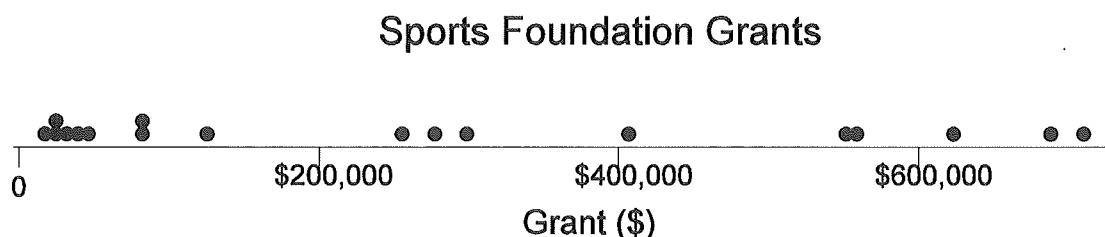


Figure: Sports Foundation Grants, 2000–2001

A **better** graph to highlight the difference between the grants obtained by different sports would be:

- X (1) side-by-side box plots with the same scaled x-axes. *n=18!*
- ? (2) a labelled bar graph ordered by the size of the grant. ✓
- X (3) a histogram with equal width class intervals for **Grants** on the x-axis. *n=18!*
- X (4) a scatter plot with **Grants** as the response variable and **Sport** as the explanatory variable. *cate!*
- ? (5) a pie chart with the sectors labelled and ordered by the size of the grant. X

26. *Time*, an American magazine, reported on a sex survey in America conducted by a Chicago National Opinion Research Centre team. A team of highly trained interviewers interviewed and questioned 3452 subjects. The results of the question "How many sexual partners have you had since you were 18?" are shown in the table below.

	Number of Sexual Partners						Totals
	None	1	2 - 4	5 - 10	11 - 20	21+	
Women	51	549	616	342	103	51	1712
Men	52	348	365	401	278	296	1740
Totals	103	897	981	743	381	347	3452

Which **one** of the following statements is **false** for the above table?

- T
R
I
E
T
- (1) Gender is a categorical variable.
 - (2) The number of sexual partners is a ~~continuous~~ *cate, ord* numeric variable.
 - (3) Two-way tables of counts are useful for investigating the relationship between two categorical variables.
 - (4) Percentages would enable better comparison of the number of sexual partners between men and women.
 - (5) Splitting women/men into several age groups would make the table more informative.

Questions 27 and 28 refer to the following information.

Students enrolled in stage one statistics at the University of Auckland were surveyed regarding their access to, and experience with, computers. The survey was included as a question in an assignment, and students were given marks for completing it (irrespective of the answers they gave). Staff administering the courses wished to use the results of this survey to draw conclusions about future stage one statistics students.

One question asked: 'At the start of the course, how would you describe your Excel experience?'. A total of 918 students answered this question. Each of the 918 answers were classified according to the response given by the student, and the stream the student attended. The results are given in Table 4 below, where 101G, 101 and 108 refer to the various streams.

cat, ord

Response	Stream <i>cat, num</i>			Total
	101G	101	108	
None	15	36	102	153
Very Little	44	89	119	252
Some	74	150	200	424
Lots	9	29	51	89
Total	142	304	472	918

Table 4: Responses to question regarding Excel experience.

27. The variable **Stream** is:

- (1) discrete.
- (2) numeric.
- ☒ (3) categorical.
- (4) dependent.
- (5) continuous.

28. Which of the following plots would together give the **best** display of the data in the table?

- I. ☒ a bar graph of Response.
- II. ☒ a bar graph of Stream.
- III. ☒ a dot plot of Response.
- IV. ☒ a dot plot of Stream. *} not num!*
- V. ☒ a bar graph of Response for each level of Stream.
- VI. ☒ a bar graph of Stream for each level of Response.

- (1) II, V only.
- (2) III, IV only.
- (3) III, IV, V, VI only.
- (4) I, II, V, VI only.
- (5) I, II only.

EXERCISE ANSWERS

- | | | | |
|---------------------------------------------------------|---------------------------------------------------------|-----------|-----------|
| (a) ✓ i) has
the
greater
standard
deviation | (b) ✓ c) has
the
biggest
standard
deviation | (c) ✓ (3) | (g) ✓ (5) |
| | | (d) ✓ (3) | (h) ✓ (5) |
| | | (e) ✓ (3) | (i) ✓ (2) |
| | | (f) ✓ (4) | |

ANSWERS

- | | | | | | |
|-----------|-----------|-----------|-----------|-----------|-----------|
| 1. ✓ (2) | 2. ✓ (4) | 3. ✓ (4) | 4. ✓ (5) | 5. ✓ (5) | 6. ✓ (2) |
| 7. ✓ (5) | 8. ✓ (4) | 9. ✓ (2) | 10. ✓ (5) | 11. ✓ (2) | 12. ✓ (1) |
| 13. ✓ (3) | 14. ✓ (3) | 15. ✓ (2) | 16. ✓ (5) | 17. ✓ (5) | 18. ✓ (2) |
| 19. ✓ (2) | 20. ✓ (4) | 21. ✓ (3) | 22. ✓ (3) | 23. ✓ (4) | 24. ✓ (1) |
| 25. ✓ (2) | 26. ✓ (2) | 27. ✓ (3) | 28. ✓ (4) | | |

WHAT SHOULD I DO NEXT?

Once you've had a go at all of the problems in the handout (check out Leila's scanned slides at www.tinyURL.com/stats-EDA for her additional handwritten notes and workings for every problem), you could:

- Go through the Chapter 1 blue pages. The blue pages relevant to the material in this workshop are:
 - the notes on pages 19 to 28 (except for the left-hand side and the top half of the right-hand side of page 28)
 - the glossary on pages 29 and 30
 - the true/false statements on page 31 (except for z. and cc.)
 - the questions on pages 32 to 39 (except for 10-12 and 20-24)
 - the tutorial material (except for Section B on pages 42-44)
- Try Chapter 1 questions from three of the past five tests that are relevant to this workshop.