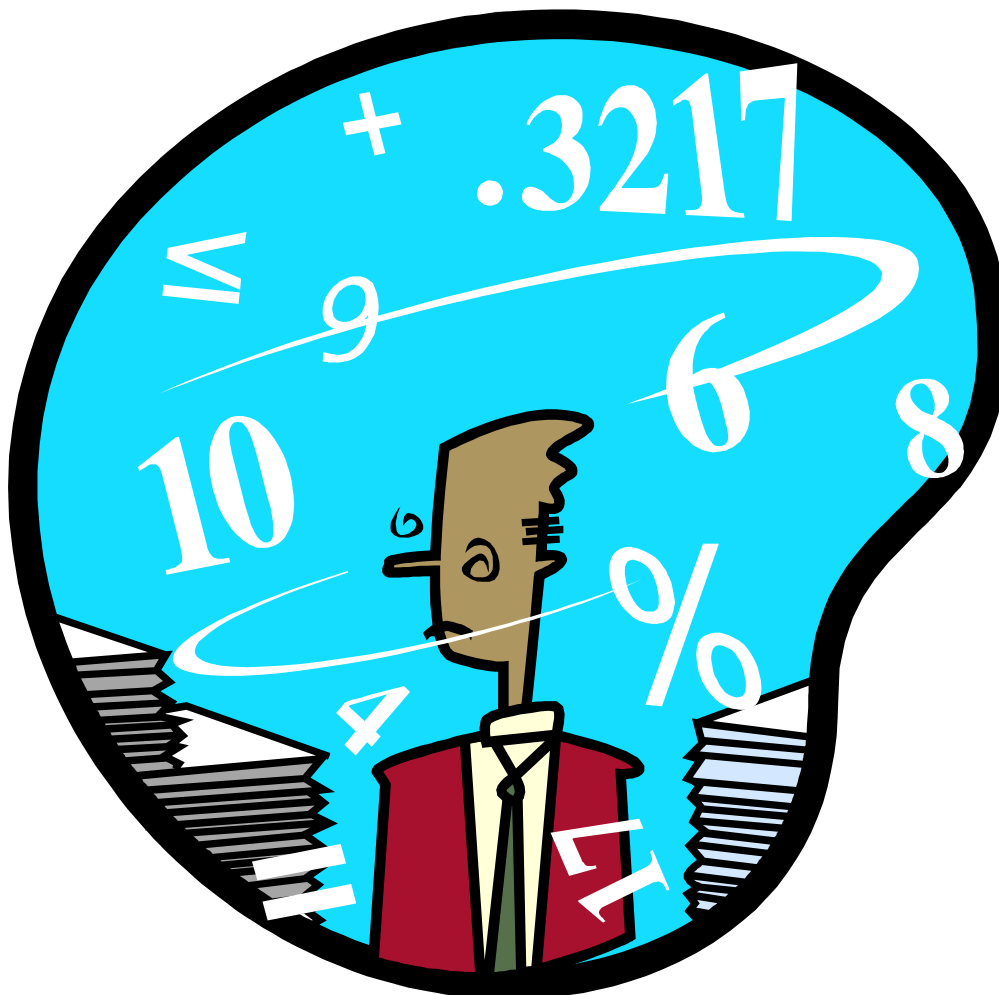


STATS 101/108 WORKSHOP

CHAPTERS 1, 2 AND 3

SATURDAY 31 JULY, 2010



Students **MUST REGISTER** for all workshops with
The Student Learning Centre, 3rd Floor, Information Commons

Student Learning Centre

Topics we teach and can provide advice on include:

- ✓ Essay writing
- ✓ Computer skills
- ✓ Reading and notetaking
- ✓ Memory and concentration
- ✓ Report writing
- ✓ Test and examination skills
- ✓ Thesis and dissertation writing
- ✓ Tutorial skills
- ✓ Research skills
- ✓ Time and stress management
- ✓ Mathematics
- ✓ **Statistics**
- ✓ Oral presentation and seminar skills
- ✓ Language learning
- ✓ Specific learning disabilities
- ✓ Motivation and goal setting
- ✓ Survival skills (in the University system)

Programmes within SLC include:

- Te Puni Wananga
Maori university tutors committed to enhancing Maori students' success
- Fale Pasifika
Pacific Island tutors committed to enhancing success for Pacific Island students
- Learning Disabilities
Learning assessments and academic assistance are available for students with specific learning disabilities and mental health impairments
If you have any special learning requirements, please feel free to discuss this with Leila in person or via email.
- Academic English Conversation Groups
Improve your academic English; develop communication skills including critical/creative thinking and clear expression of ideas and opinions. Weekly class held at the SLC on Thursdays, 3-5pm (during semester)

Statistical help available at the SLC

The Student Learning Centre (SLC) offers help for STATS 101/108 by offering:

- one-on-one tutoring help, and
- a number of workshops

One-on-one help

The SLC employs tutors specifically to help students with one-on-one assistance for STATS 101/108. One-on-one tutoring must be booked at SLC reception on the third floor of the Information Commons in person or by calling 373-7599 X 88850. Enquire at the SLC reception for available times.

Note: SLC tutors are not allowed to help students complete their assignments.

SLC STATS 101/108 Workshops

Any questions regarding STATS 101/108 workshops should be forwarded to:

Leila Boyle
SLC Statistics Co-ordinator
l.boyle@auckland.ac.nz

Workshops are run in a relaxed environment, typically set at a pace for those students that find the Statistics Department's tutorials too fast. All workshops allow plenty of time for questions. In fact, this is encouraged 😊

1) Saturday Workshops

These five 3-hour workshops are held on Saturdays throughout the semester to help students with different sections of the course.

2) Computer Workshops: Excel / SPSS

These three computer-based workshops introduce students to the skills needed for Excel and SPSS (PASW) use in STATS 101/108 assignments.

3) Pre-test Workshops

These three workshops will cover the basics that you need for the test.

4) Pre-exam Workshops

These six workshops will cover the basics that you need for the exam.

Note: All workshops concentrate on questions reviewing the basic concepts, rather than questions on finer details. They are designed to assist students to achieve a pass; they are not designed to cover all material.

The timetable for these workshops is available with this handout. Please enrol in each of your preferred classes at the Student Learning Centre by:

- **Going to the SLC in person**
- **Enrolling online at www.slc.auckland.ac.nz (click on "[SLC Workshops](#)" then click on "[Undergraduate Workshops](#)" then view workshops by "[Statistics](#)").**

Useful Websites

- SLC webpage: www.slc.auckland.ac.nz
- Online enrolment through the SLC site: www.slc.auckland.ac.nz, click on "[SLC Workshops](#)" and then click on "[Undergraduate Workshops](#)"
 - For STATS 101/108 workshops, view by "[Statistics](#)"
 - For SLC workshops aimed at helping students new to computers to learn basic skills as well as specific workshops on using Microsoft Excel, Word, PowerPoint, view by "[Computer skills](#)"
 - For other SLC workshops on skills appropriate for undergraduate students, scroll down the page or view by the appropriate category
- Cecil: <https://cecil.auckland.ac.nz/>
- Leila's website for STATS 101/108 SLC workshop handouts & information: www.stat.auckland.ac.nz/~leila



Revision Notes

Chapter 1 – What is Statistics?

Look at blue pages for good notes and test/exam questions for practice

- **Polls and Surveys**

- **Target population/Population of interest**
Complete set of individuals, objects, or units that we want information about.
- **Study population**
Complete set of units that might possibly be included in the study. Ideally the same as the target population but often different.
- **Sampling frame**
List of units in the study population from which the sample will be drawn.
- **Sampling design**
The way the sample is to be chosen from the sampling frame.
- **Sample**
Subset of units in the study population which information is collected on.
- **Census**
Attempt to sample the whole population.
- **Variable**
A characteristic of each unit that we measure.
- **Parameter**
Numerical characteristic of the population or distribution.
- **Statistic**
A number calculated from the data, usually used to estimate an unknown parameter.

- **Randomisation - Obtain representative samples**

- A representative sample reflects the characteristics in the population.
- Random sampling
Technique where each unit is selected entirely by chance.
- Simple Random Sample (SRS)
Sampling without replacement



- **Errors in Surveys:**

- 1. Sampling errors**

- Arise from taking a sample rather than a census, unavoidable.
 - Also known as chance or random errors.
 - Are bigger in smaller samples than larger ones.
 - Size may be estimated by statistical methods.

- 2. Non-sampling errors**

- Errors that occur during the data collection process → try to minimise in design of survey by using a pilot survey.
 - Can be much larger than sampling errors – always present
 - Can be virtually impossible to correct for afterwards
 - Virtually impossible to determine how large they can be
 - **Selection bias**
Population sampled is not exactly the population of interest.
 - **Non-response bias**
Not everyone in the sample who had been specifically chosen responded. Non-respondents often behave or think differently from respondents.
 - **Self selection**
People themselves decide whether or not to participate.
 - **Question effects**
Wording and sentence structure of questions. Even slight differences in question wording can produce measurable differences in how people respond.
 - **Survey format effects**
Factors such as type of survey (mail/phone/face-to-face interview), question order, layout of written survey, self-administered questionnaire or interviewer, ... etc, can affect the results.
 - **Interviewer effects**
Gender, ethnicity, age of the interviewer, facial expression...etc. Different interviewers asking the same questions may obtain different results.
 - **Behavioural considerations**
Social desirability of answers.
 - **Transferring findings**
Using data gathered from one population and using the results to comment on another.



- **Experiments**

- Experimenter decides who or what receives which treatment (ideally using some form of random allocation)
- Randomisation used for treatment allocation.
- Can prove cause and effect
- Types of experiments include:
 - **Completely Randomised Design**
Allocate treatments to units entirely by chance to try to make the treatment groups as similar as possible.
 - **Randomised Block Design**
Group (block) units by some known factor, then randomly allocate treatments to units within each block to try to balance out the unknown factors.
- **Control group**
Group of experimental units is given no treatment. Treatment effect is estimated by comparing each treatment group with control group
- **Placebo** – inert (inactive) “dummy” treatment
- **Placebo effect** – people show signs of “improvement” when they believe they have taken the real treatment.
- **Blinding**
Prevent people involved in experiment from knowing which experimental subjects have received which treatment.
 - **Single Blind** – subjects themselves
 - **Double Blind** – subjects and people administering the treatments

- **Observational Studies**

- CANNOT prove cause and effect – often useful for identifying possible causes of effects, but cannot reliably establish causation.
- Should use some form of random sampling → representative samples.
- Unit/person/thing “decides” what treatment they want/get.
 - **Cross-sectional:**
A study which observes a group of individuals or units at a point in time. It is a descriptive study, providing a “snapshot” at a particular point in time.
 - **Longitudinal:**
A study which observes the same group of individuals or units over a long period of time. Comprised of a series of cross-sectional studies

Chapter 2 – Tools for Exploring Univariate Data

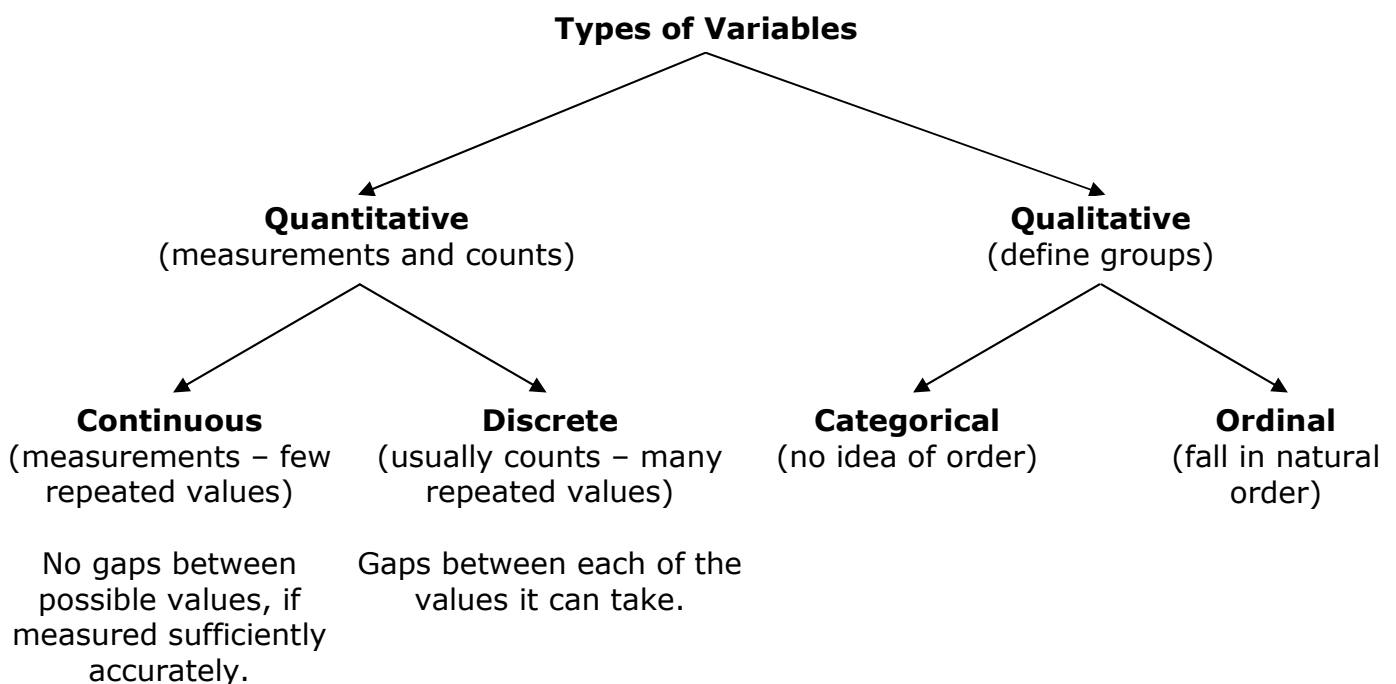
Look at blue pages for good notes and test/exam questions for practice

- **Presentation of Data in Tables**

Guidelines for conveying information quickly and easily:

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

- **Types of Variables**





- **Numerical summaries**

- **Centre**

- Sample mean, \bar{x} (also known as the average or expected value) $\sum \frac{X_i}{n}$ – affected by outliers
- Median (= Med – also known as the 50th percentile) = middle number of the ordered data – not affected by outliers
- Mode, most frequently occurring number/most common value – not affected by outliers, useful for qualitative data

- **Spread**

- Inter-quartile range: IQR = Upper quartile – Lower quartile (middle 50% of data) – not affected by outliers
 - Lower quartile (Q₁) – upper boundary of the lower quarter of the data – not affected by outliers
 - Upper quartile (Q₃) – lower boundary of the upper quarter of the data – not affected by outliers
- Range (maximum – minimum) – affected by outliers
- Sample standard deviation, $\sigma_{n-1} / s / s_x$ – affected by outliers

- **Five number summary** (Min, Q₁, Med, Q₃, Max)

- **Using your scientific/graphics calculator**

Do you know how to use your calculator to find the **mean** (\bar{x}) and **standard deviation** ($\sigma_{n-1} / s / s_x$) of a **sample**? There are three possible types of **sample mean/ standard deviation** problem, can you do all of them?

- A list of numbers (usually separated by commas, tabs or spaces) ^(tick)
- A frequency table with single numbers in the left hand column
- A frequency table with ranges in the left hand column



- **Outliers / Outside values**
- **Shape / Distributions of data**
 - How many modes/peaks does the data have?
 - Unimodal **OR**
 - Bimodal
 - Is the data symmetric or skewed?
 - Positively/right skewed **OR**
 - Negatively/left skewed **OR**
 - Roughly/approximately symmetric

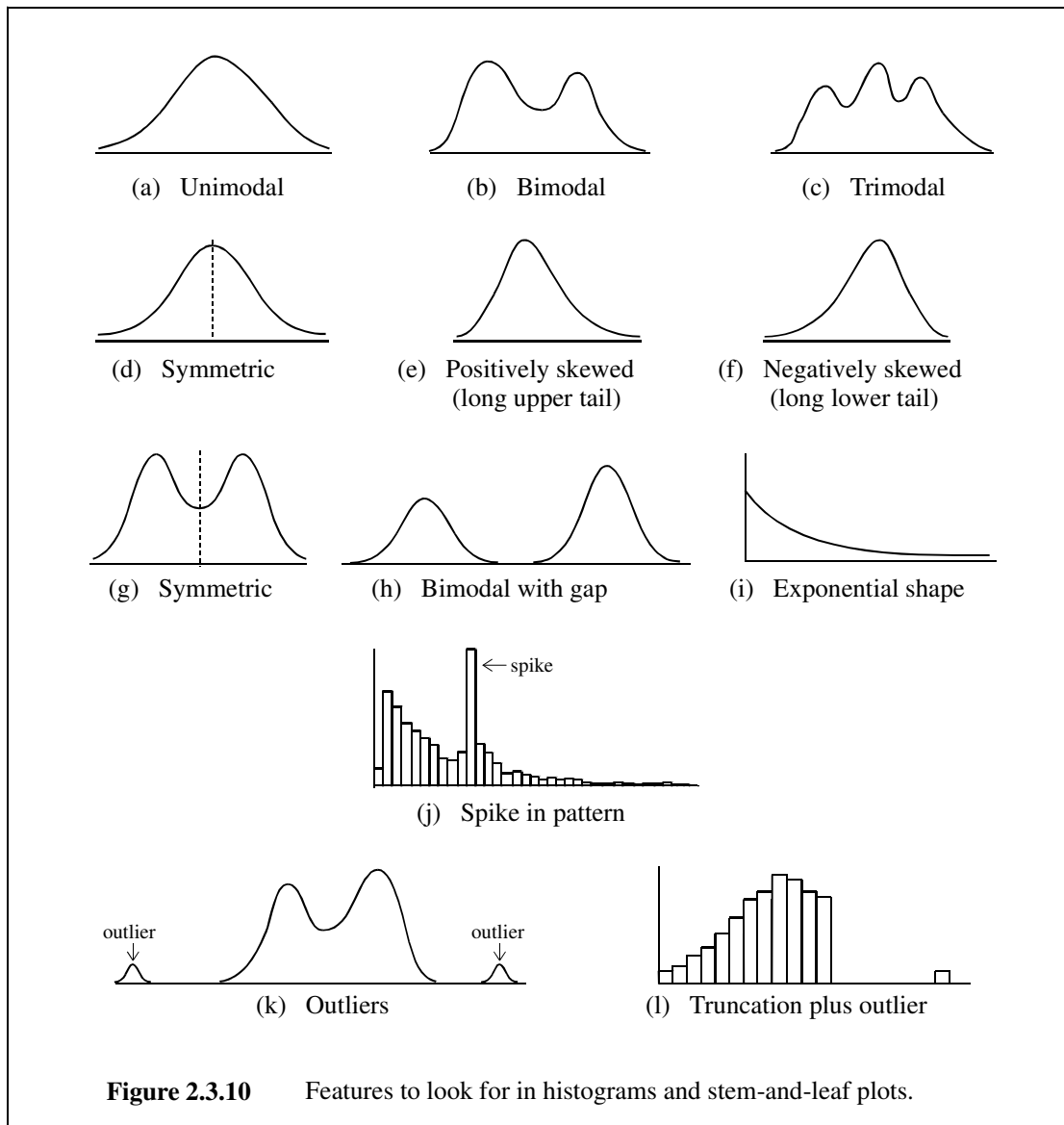


Figure 2.3.10 Features to look for in histograms and stem-and-leaf plots.

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



Types of graphs

- 2D vs 3D – Always use 2D!
- Avoid pie graph if you can!
- Order items sensibly.
- Discrete variables:
 - Bar – order categories by size
- Continuous variables:
 - Dot plot – small data sets, $n \leq 20$
 - Stem-and-leaf plot – moderate data sets, $15 \leq n \leq 150$
 - Box plot – moderate to large data sets, $n \geq 30$
 - Histogram – large data sets, $n \geq 50$

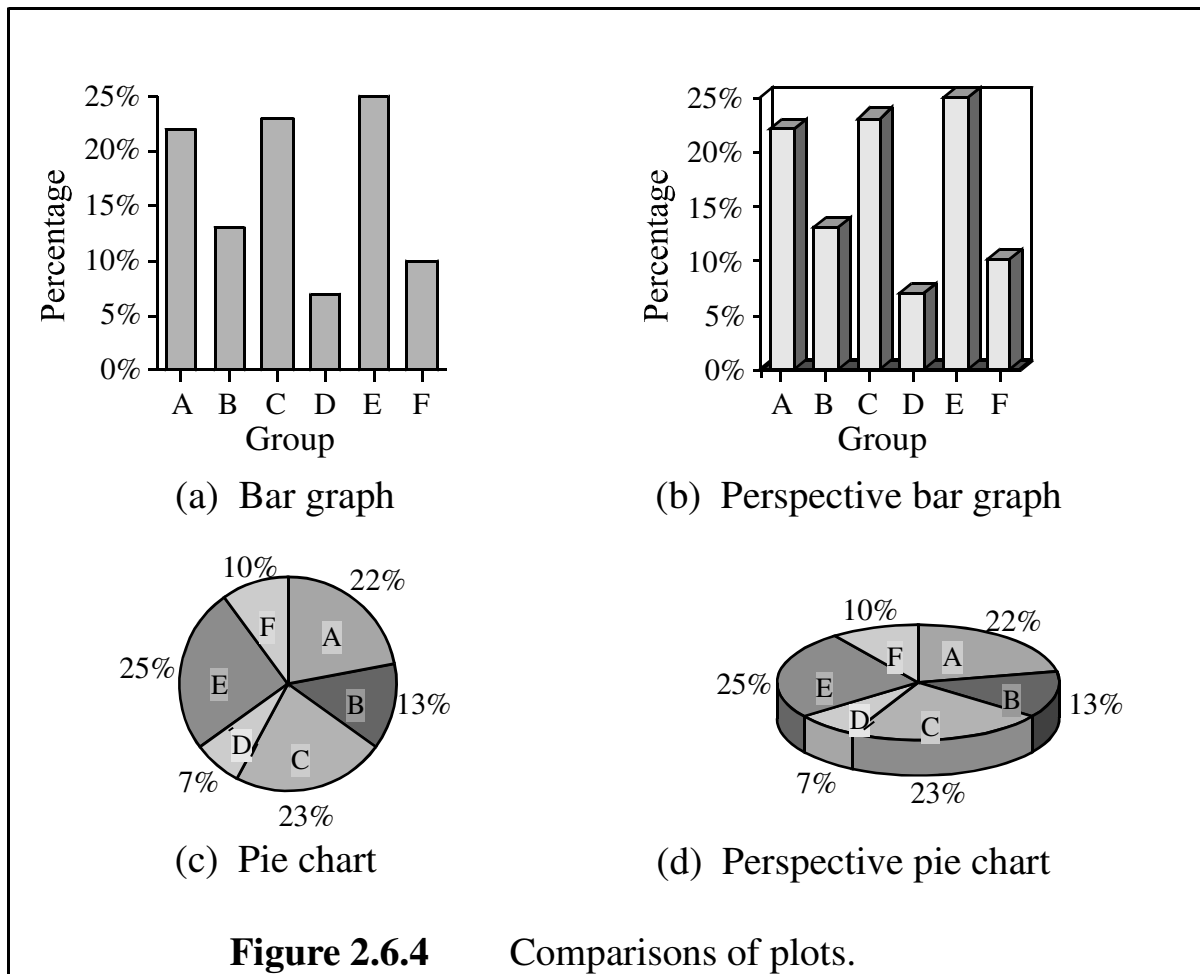
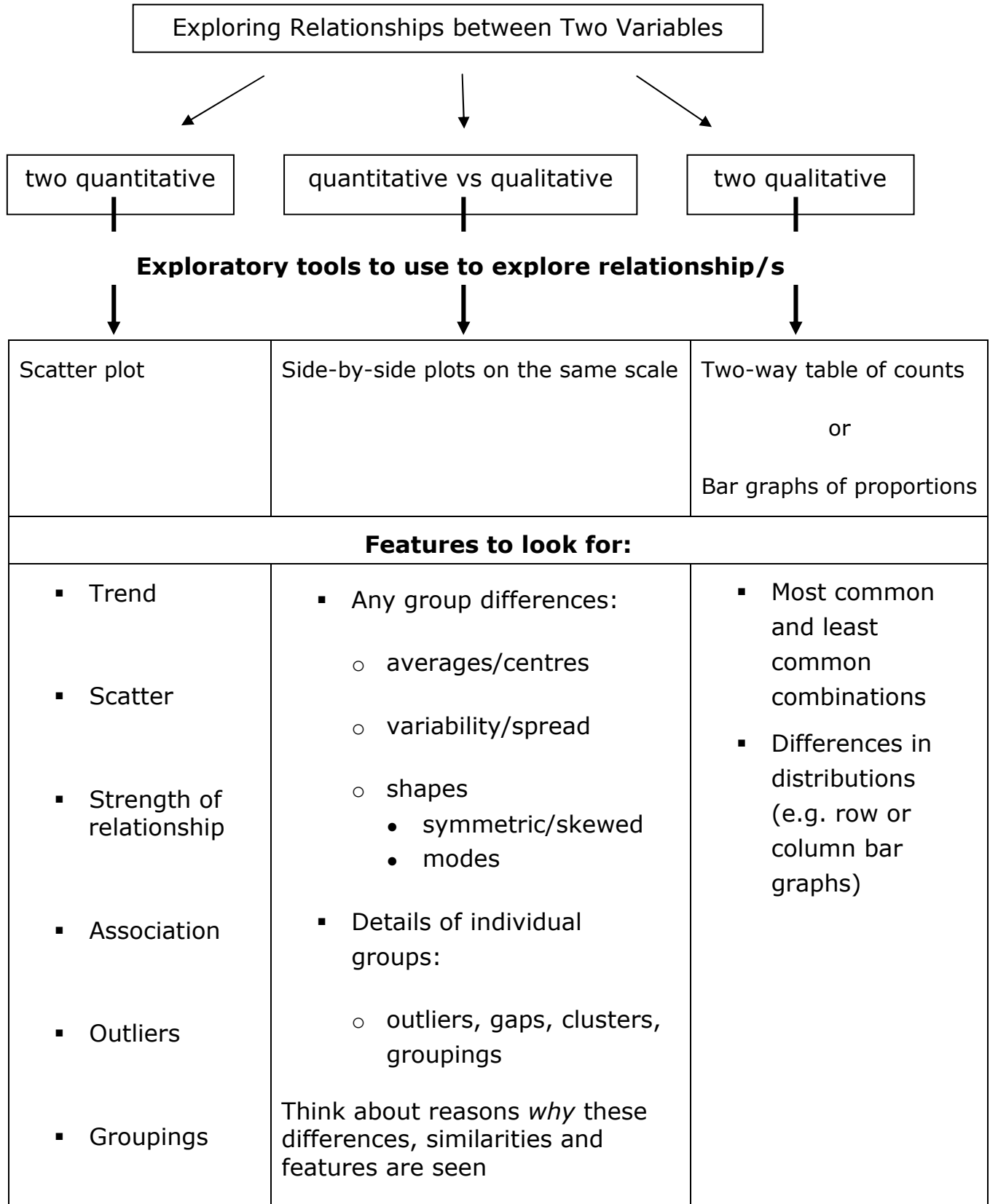


Figure 2.6.4 Comparisons of plots.

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

Chapter 3 – Exploratory Tools for Relationships

Look at blue pages for good notes and test/exam questions for practice



Chapters 1, 2 & 3 – Questions

1. A statistician recently surveyed people in a small town on the mathematics they used in the workplace. Because of financial restraints he could only survey 200 people. To take a random sample he used a list of 661 workplaces in the town (numbered 1...661). The following numbers were extracted from the table of random digits:

03521 10933 34889 74209 31391 62728 99671 77720 13726

If he started sampling at the beginning of the line of random digits and took consecutive digits, the first five workplaces selected would be numbered:

- (1) 35 211 93 334 139
- (2) 35 93 139 162 177
- (3) 35 109 348 313 137
- (4) 35 93 93 139 162
- (5) 35 211 93 334 93

2. *Consumer News* (March 1994) reports on an Australian scientist's claim that weekdays really are warmer than weekends. Adrian Gordon of the Flinders Institute of Atmospheric and Marine Scientists in Adelaide says, "This is not a greenhouse effect. It is the direct warming effect of fuel burning in factories and vehicles during the week". Gordon's findings follow a study of worldwide satellite data for 5000 days between 1979 and 1992. Sunday, he says, is typically the coldest day, with the temperature rising to a peak on Wednesday and then steadily dropping again.

This study could **best** be described as:

- (1) An observational study.
- (2) An experimental study.
- (3) A cluster sampling study.
- (4) A stratified sampling study.
- (5) A randomised block design study.

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

- (1) Nonresponse can cause bias in surveys because non-respondents often tend to behave differently from people who do respond.
- (2) Nonsampling errors are often bigger than the random sampling errors in surveys.
- (3) Well designed experiments use randomisation to avoid subjective and other biases.
- (4) People will sometimes answer a question differently for different interviewers.
- (5) A well-planned observational study is a reliable method for establishing causation.

4. The following is a quotation from the *Sunday Star-Times* (23 March 1994): "New Zealanders have given the thumbs down to Prime Minister Jim Bolger's call for a republic". A nationwide *Sunday Star-Times* opinion poll shows only 28% in favour with 46% against and 26% who don't care or don't know. The telephone poll of 1014 people was taken the day after Mr. Bolger's speech proposing the year 2000 as "an appropriate symbolic moment" to dump the Queen in favour of an elected president.

Which **one** of the following statements about telephone polls is **false**?

- (1) The true error in the poll is larger than the sampling error.
 - (2) Retired people may be over represented.
 - (3) Busy people may be less likely to respond.
 - (4) The errors quoted for polls such as this in the media include non-response errors.
 - (5) There may be different response from different geographical areas.
5. Which one of the following statements is **false**?
- (1) Slight changes in the wording of questions can make a measurable difference to the results of a survey.
 - (2) There are always statistical procedures available to correct results (at the completion of a survey) when the population from which a sample is taken is different from the population of interest.
 - (3) Bias can occur when too many respondents in a survey give an answer which does not reflect their actual behaviour.
 - (4) The outcome of a survey which uses personal interviews may be different from the outcome of the same survey if telephone interviews had been used.
 - (5) The outcome of a survey may be affected by the race and/or gender of the interviewer.
6. 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**?
- (1) Month of birth coded Jan = 1, Feb = 2, ... , Dec = 12 is a qualitative variable.
 - (2) Staff status coded as 1 = full-time, 2 = part-time is a qualitative variable.
 - (3) The distance travelled to the university, recorded to the nearest 5 kilometres, is a quantitative variable.
 - (4) Continuous variables have few repeated values.
 - (5) The length of time it generally takes to find a park, including the time waiting in a queue, is a continuous quantitative variable.



7. *TIME*, 20 September 1993, reports on a long-term study of child development conducted by researchers at the University of Otago. In the study, every child born at Queen Mary Hospital between 1 April 1972 and 31 March 1973 has been monitored at two year intervals for the past 21 years. This study could best be described as:

- (1) a sample survey.
- (2) a randomised experiment.
- (3) an observational study.
- (4) a double blind experiment.
- (5) a cross-sectional study.

8. The ages of 19 people who took part in a tetanus study are given below.

Patient Number	1	2	3	4	5	6	7	8	9	10
Age	41	16	28	28	35	27	40	20	30	17

Patient Number	11	12	13	14	15	16	17	18	19
Age	27	12	50	12	30	16	9	20	40

The mean age of the people in your sample is:

- (1) 23
- (2) 12.6
- (3) 17.0
- (4) 26.2
- (5) 8.4

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

- (1) The scatter plot is a useful tool for investigating relationships between two continuous variables.
- (2) The box plot is a useful tool for comparing several sets of data.
- (3) Blocking is used in experimental design whereas stratification is used in observational studies to refer to the idea of making comparisons only within similar relevant groups.
- (4) The interquartile range can be seriously affected by an outlier.
- (5) The sample mean can be seriously affected by an outlier.

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

- (1) Blocking is used for experiments to ensure fair comparisons with respect to factors the experimenter knows are important.
- (2) Random sampling errors always have an identifiable cause.
- (3) Experiments on human beings should be double blind, if possible.
- (4) Non-response in surveys can cause bias because non-respondents often tend to behave differently from people who do respond.
- (5) Observational studies are not reliable for proving causation.



11. The September 1993 issue of *Consumer* contains figures on the reliability of cars driven in New Zealand. The figures were obtained from a sample of 15,372 *Consumer* readers who responded to a mailed questionnaire.

The **main** problem with using the results to draw conclusions about the reliability of all cars driven in New Zealand is:

- (1) there was no control group
- (2) there was insufficient attention to the placebo effect
- (3) selection bias
- (4) question effects
- (5) interviewer bias

Questions 12 and 13 make use of the data on 1993 model cars in U.S.A, taken from the *Journal of Statistics Education*, v.1, n.1 (1993). The "midrange" price for each of the 21 cars in the small car category is given in the following stem-and-leaf plot:

Units: 6|1 = \$6100

7	4
8	0 3 4 4 6
9	0 1 2 8
10	0 1 3 9
11	1 3 6 8
12	1 2
13	
14	
15	9

12. Which one of the following statements is **not** a feature of the data?
- (1) The distribution of the midrange prices is negatively skewed.
 - (2) The minimum midrange price is \$7,400.
 - (3) The distribution of the midrange prices appears to be unimodal.
 - (4) The range is \$8500.
 - (5) Most small cars have a midrange price of between \$7,400 to \$12,200.

13. The same source gives the engine sizes, in litres, for 8 models of large cars namely:

3.3, 4.6, 3.8, 4.6, 5.7, 4.9, 3.5, 3.8

The mean, \bar{x} , and standard deviation, s , for this data set is:

- (1) $\bar{x} = 4.28, s = 0.81$
- (2) $\bar{x} = 4.28, s = 0.76$
- (3) $\bar{x} = 4.21, s = 0.76$
- (4) $\bar{x} = 4.21, s = 0.81$
- (5) $\bar{x} = 4.21, s = 0.97$

14. The data in the following table is a subset of the data collected from a study on a series of male patients admitted to Greenlane Hospital in Auckland after a heart attack. The table shows the frequency of the ejection fraction for a sample of 45 patients. The ejection fraction is the percentage of blood in the left ventricle of the heart ejected in one heartbeat.

Percentage of blood ejected from the left ventricle in one heartbeat

Interval	Frequency
15-24	1
25-34	3
35-44	9
45-54	13
55-64	13
65-74	6
Total	45

The sample mean, \bar{x} , and the sample standard deviation, s , of this data set are:

- (1) $\bar{x} = 50.37$, $s = 12.10$
 - (2) $\bar{x} = 49.33$, $s = 12.60$
 - (3) $\bar{x} = 51.06$, $s = 12.10$
 - (4) $\bar{x} = 50.37$, $s = 12.24$
 - (5) $\bar{x} = 51.06$, $s = 12.24$
15. *Time*, 17 October 1994, 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.

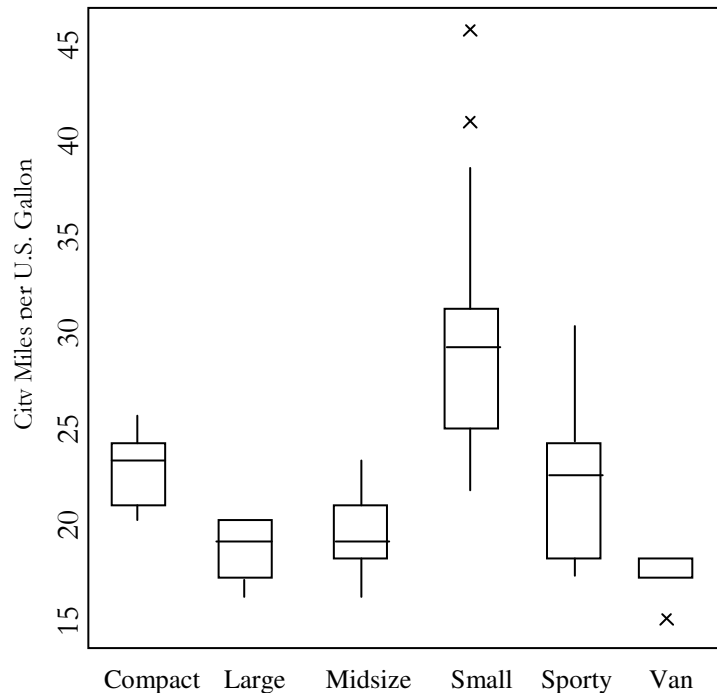
	Number of Sexual Partners						Totals
	None	1	2 - 4	5 - 10	11 - 20	21+	
Women	51	549	616	342	103	51	1712
Men	52	348	365	401	278	296	1740
Totals	103	897	981	743	381	347	3452

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

- (1) Gender is a qualitative variable.
- (2) The number of sexual partners is a continuous quantitative 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.



16. Box plots of the city miles per gallon (*mpg*) ratings for six different types of vehicle are shown below:



Which one of the following is **not** a feature of the data?

- (1) The variability as measured by the interquartile range is very similar for the sporty and small cars.
- (2) The median value of *mpg* for compact cars is slightly higher than sporty cars.
- (3) All types of car contain some cars with *mpg* values less than 23.
- (4) The highest value of *mpg* for large cars is smaller than the lowest value for small cars.
- (5) The only outliers occur in the small car sample.

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

- (1) The means and standard deviations of quantitative variables are meaningless.
- (2) A stem-and-leaf plot of the data for a continuous variable is useful for showing the density of data points along a scale.
- (3) Stem-and-leaf plots show more detail about the distribution of a single variable than boxplots do.
- (4) The sample median is usually unaffected by outliers.
- (5) Bar graphs are a good way of plotting the frequencies of values for a discrete quantitative variable.

Questions 18 to 20 refer to the following information.

The following stem-and-leaf plot of flight duration times for 20 Air New Zealand international flights out of Auckland.

Units: 7|3 = 7.3 hours (i.e. 7 hours 18 minutes)

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

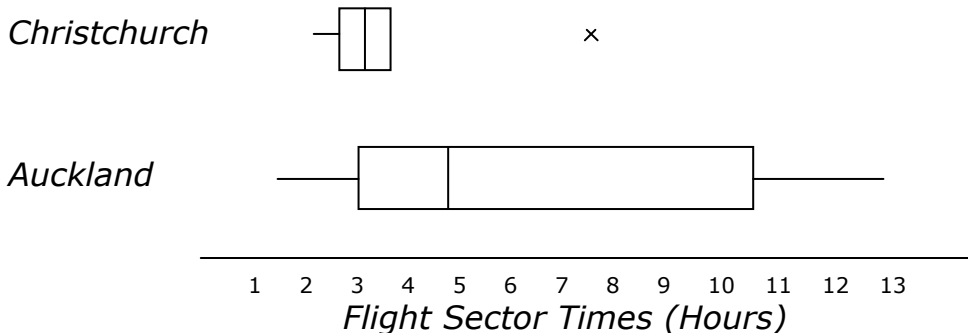
18. Which one of the following statements is **not** a feature of the data?

- (1) The range is 10 hours.
- (2) There are no obvious outliers in this sample.
- (3) The distribution of times appears to be bimodal.
- (4) The distribution of times is highly skewed.
- (5) No flight lasts longer than 12 hours.

19. The sample mean, \bar{x} , and standard deviation, s_x , for this data set are given by:

- (1) $\bar{x} = 4.80, s_x = 3.38$
- (2) $\bar{x} = 6.26, s_x = 3.38$
- (3) $\bar{x} = 6.26, s_x = 12.02$
- (4) $\bar{x} = 4.80, s_x = 3.47$
- (5) $\bar{x} = 6.26, s_x = 3.47$

20. Boxplots of the duration of the 20 Auckland flights and of 6 Air New Zealand international flights from Christchurch are shown below.



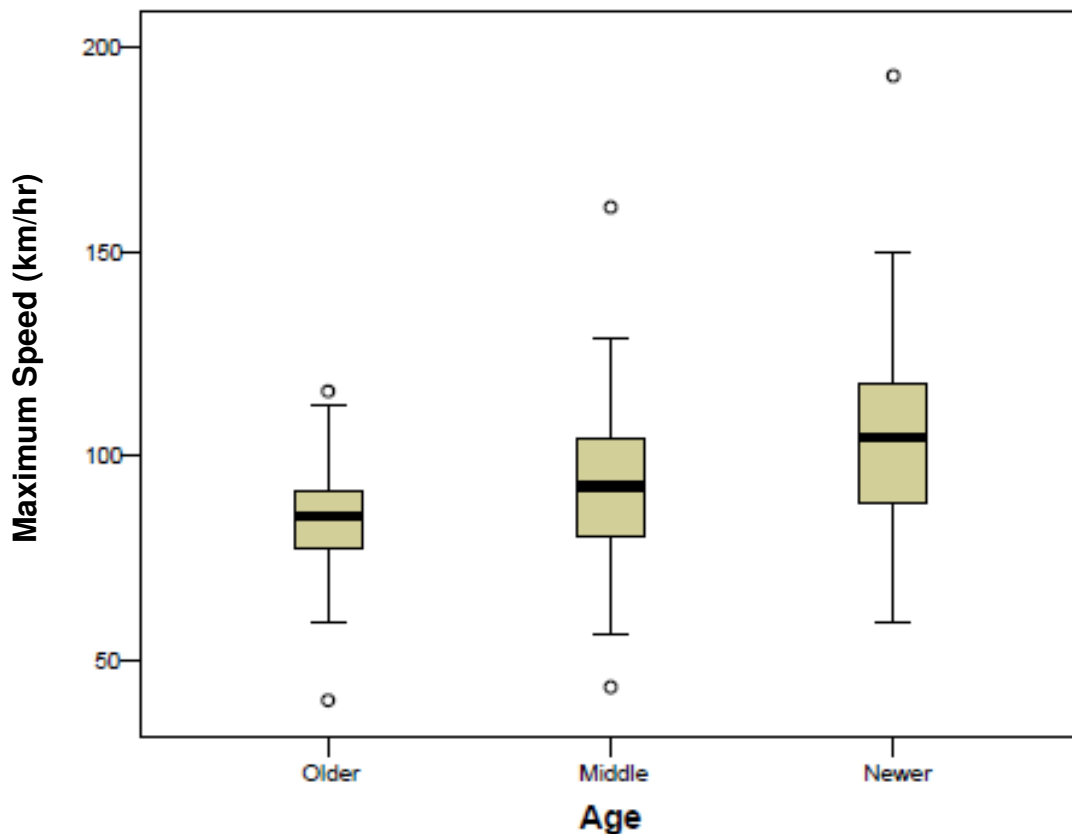
Which of the following statements is **true**?

- (1) The median length of the Auckland flights is shorter than that of the Christchurch flights.
- (2) The shortest of the Christchurch flights is longer than the shortest of the Auckland flights.



- (3) There are substantial numbers of outliers in both data sets.
- (4) The variability, as measured by the interquartile range, is very similar in both data sets.
- (5) Both data sets contain some flights of more than 10 hours duration.

21. Based on the box plots below, which one of the following statements is **false**?
- (1) None of the older roller coasters reach a maximum speed of more than 150 km/hr.
 - (2) More than half of the newer roller coasters reach maximum speeds of more than 100 km/hr.
 - (3) The range of the maximum speeds of the newer roller coasters is approximately 90 km/hr.
 - (4) At least one-quarter of the newer roller coasters have maximum speeds faster than any older roller coasters.
 - (5) The variation in the maximum speeds of the newer roller coasters is greater than that for the older roller coasters.



ANSWERS

- | | | | | | |
|---------|---------|---------|---------|---------|---------|
| 1. (1) | 2. (1) | 3. (5) | 4. (4) | 5. (2) | 6. (3) |
| 7. (3) | 8. (4) | 9. (4) | 10. (2) | 11. (3) | 12. (1) |
| 13. (1) | 14. (5) | 15. (2) | 16. (5) | 17. (1) | 18. (4) |
| 19. (5) | 20. (2) | 21. (3) | | | |