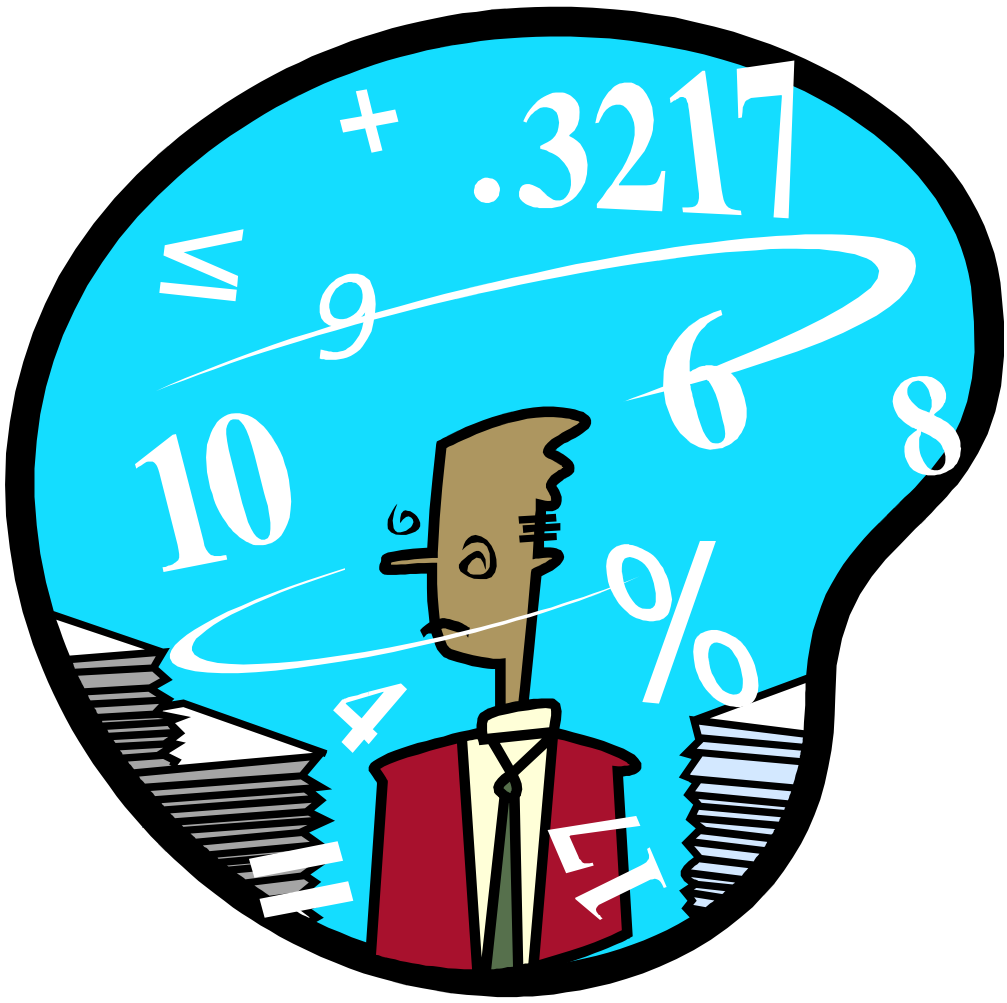
STATS 10X WORKSHOP

EXAM PREP 1: CHAPTERS 1, 2 & 3

SAT 2 & WED 27 OCTOBER 2010



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

Statistical help available at the SLC

The Student Learning Centre (SLC) offers help for STATS 10x by offering:

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

One-on-one help over S2 2010 including exam period

One-on-one assistance for STATS 10x is available at the SLC. Check appointment availability and book at SLC reception in person (third floor, Information Commons building) or by calling 373-7599 ext. 88850.

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

SLC STATS 10x Exam Prep Workshops

Any questions regarding STATS 10x workshops should be forwarded to:

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

These twelve workshops (six different sessions, each repeated twice) are held prior to the exam, from Saturday 2 October until Monday 1 November 2010 (inclusive).

These workshops concentrate on questions reviewing the **basic concepts**, rather than questions on finer details. They are designed to assist students to achieve a pass and **don't cover all material**.

The timetable for these workshops is available at this workshop, at SLC Reception and on Leila's website. Please enrol in each of your preferred workshops by EITHER:

- ***Dropping by*** **the SLC Reception to enrol in person (Room 320, Level 3, Information Commons Building, 11 Symonds Street) OR**
- ***Emailing*** **slc@auckland.ac.nz with your name, ID number, and the name, date and time of the workshop/s you wish to attend OR**
- ***Calling*** **the SLC Reception on 373-7599 ext. 88850 and book over the phone.**

Useful Websites

- SLC webpage: **www.slc.auckland.ac.nz**
- Cecil: **https://cecil.auckland.ac.nz**
- **Leila's website for STATS 10x 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 <u>without</u> 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.

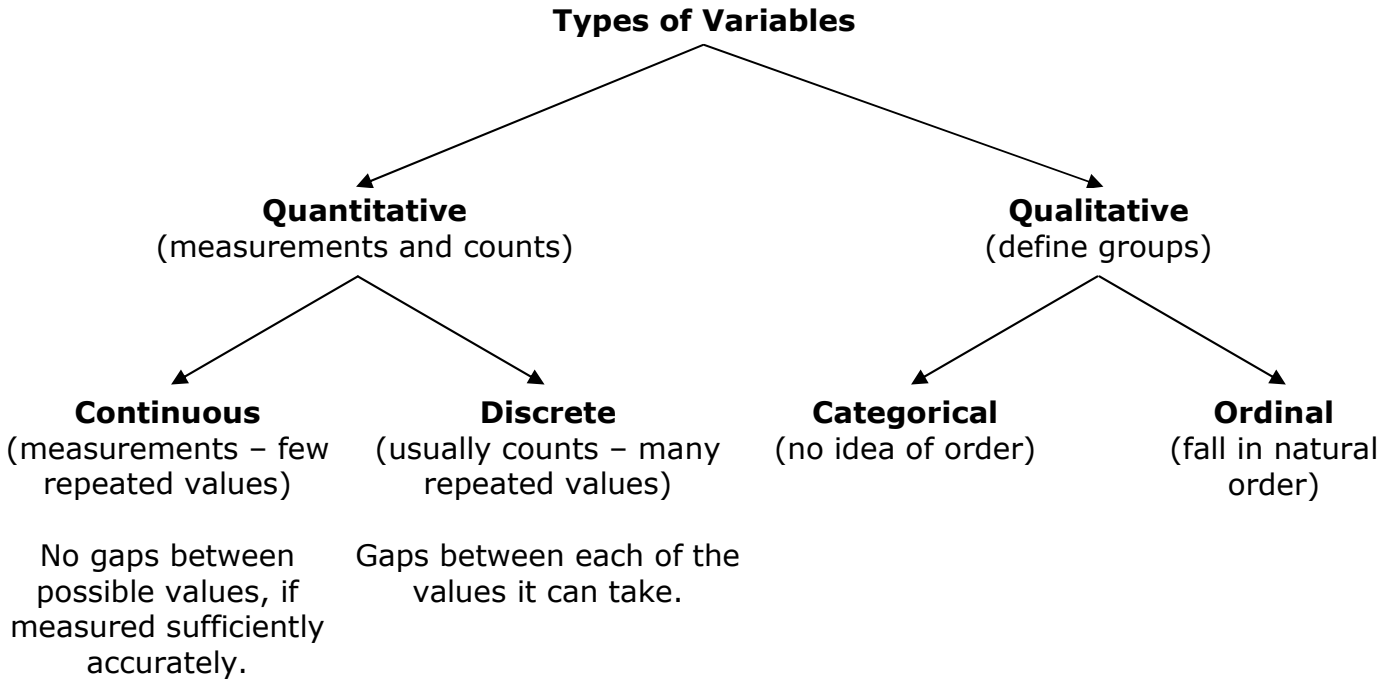
- **Experiments**
  - Experimenter decides who or what receives which treatment (ideally using some form of random allocation)
  - Randomisation used for treatment allocation.
  - Can prove cause and effect
  - Types of experiments include:
    - ***Completely Randomised Design***
      Allocate treatments 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 given no treatment, a placebo or an existing treatment.
  - ***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/or 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 (Q_1) – 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, σ_{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** (σ_{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?

  (tick)
  - A list of numbers (usually separated by commas, tabs or spaces) ☐
  - 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

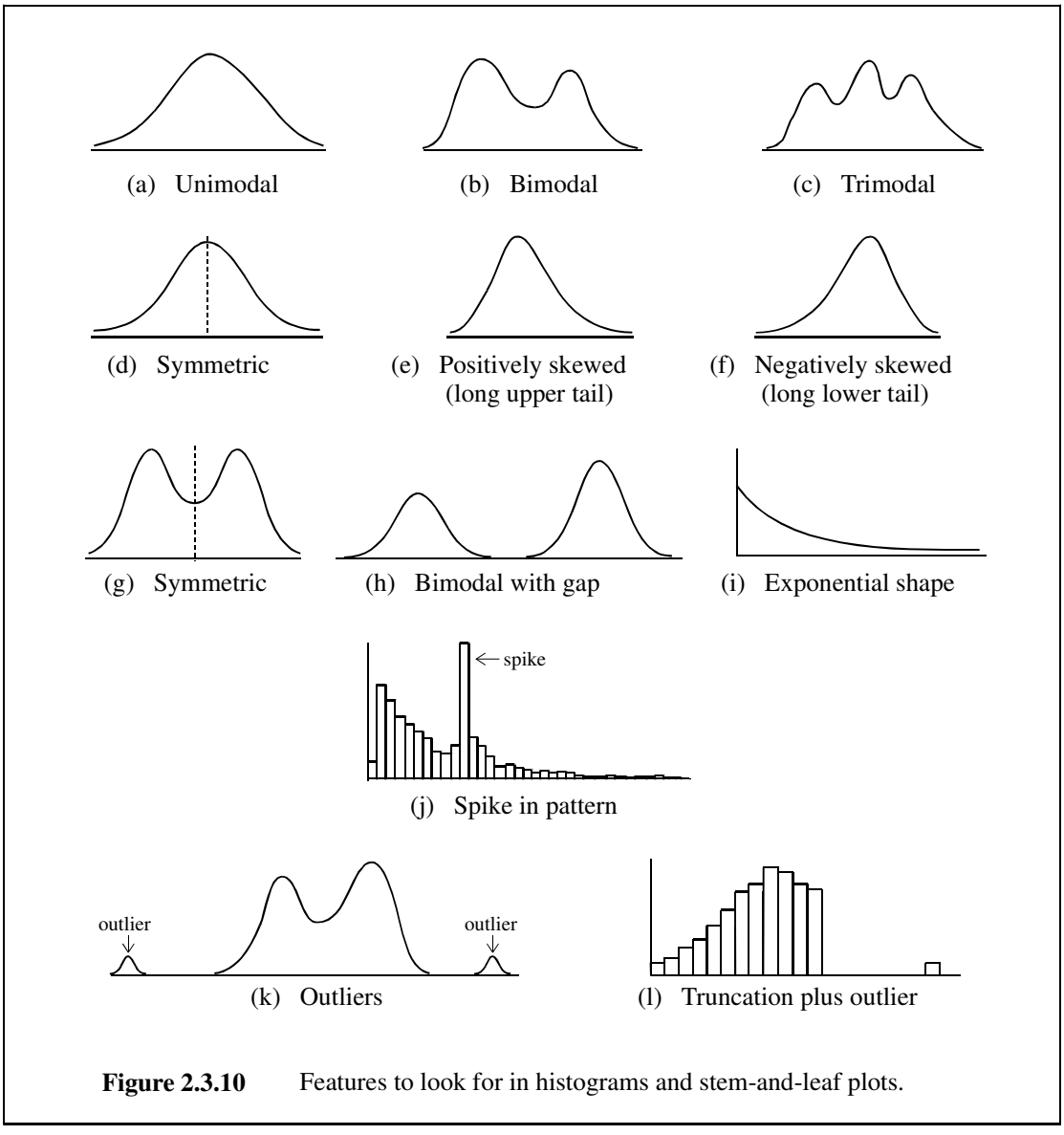


Figure 2.3.10 Features to look for in histograms and stem-and-leaf plots.

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, $n \geq 50$

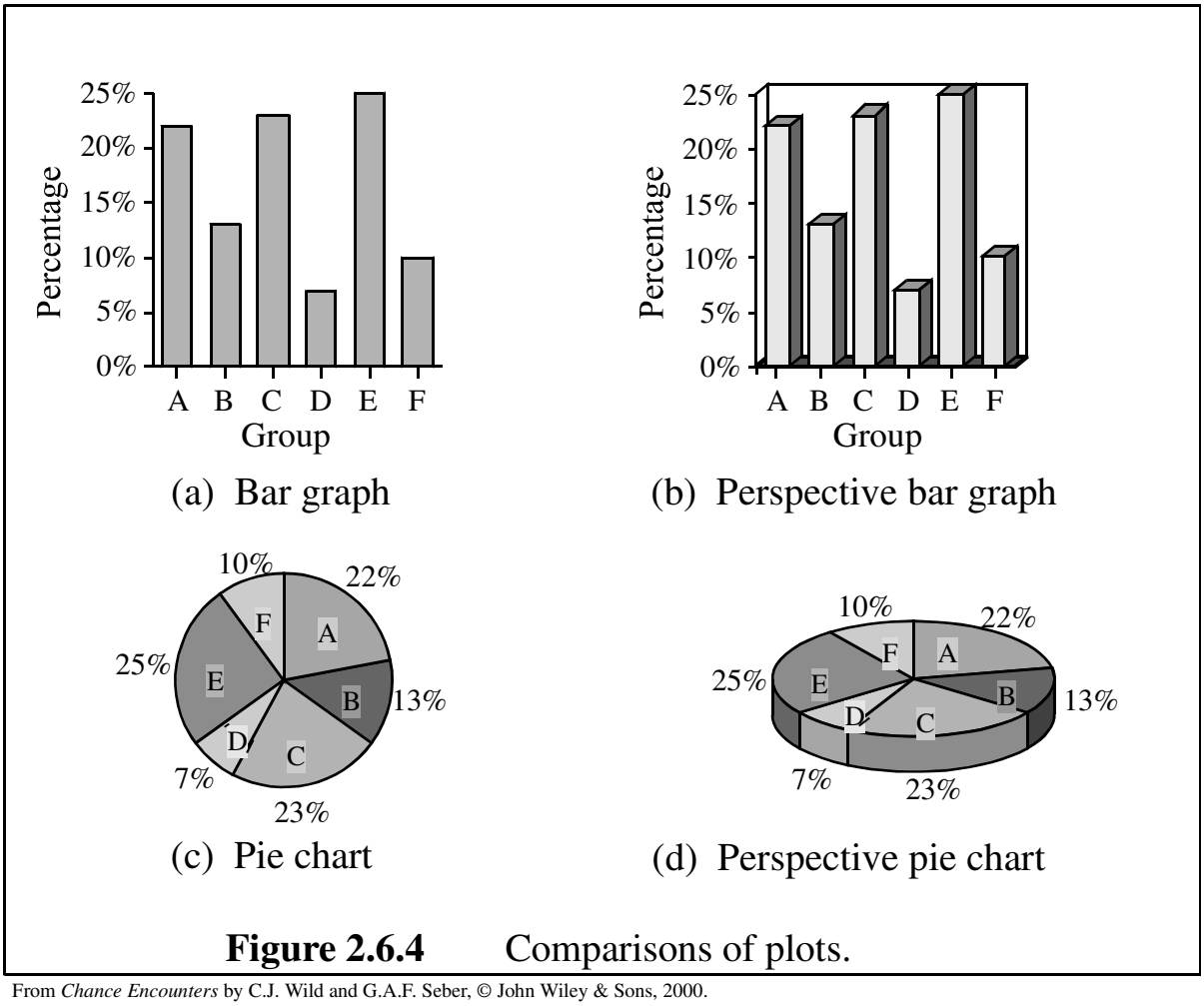


Figure 2.6.4 Comparisons of plots.

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

Exploring Relationships between Two Variables

two quantitative	quantitative vs qualitative	two qualitative

Exploratory tools to use to explore relationship/s

Scatter plot	Side-by-side plots on the same scale	Two-way table of counts or Bar graphs of proportions
Features to look for:		
▪ Trend ▪ Scatter ▪ Strength of relationship ▪ Association ▪ Outliers ▪ Groupings	▪ Any group differences: ○ averages/centres ○ variability/spread ○ shapes • symmetric/skewed • modes ▪ Details of individual groups: ○ outliers, gaps, clusters, groupings Think about reasons *why* these differences, similarities and features are seen	▪ Most common and least common combinations ▪ Differences in distributions (e.g. row or column bar graphs)

Chapters 1, 2 & 3 – Questions

1. A Statistics Department recently carried out a survey to see if the heating was set at a comfortable level. Of the 60 staff, 6 were surveyed. To take a random sample, the 60 staff names were numbered from 1 to 60. The following numbers were extracted from the table of random digits:

   46400 89413 74507 49374 85137 99211 49557 98263

   By starting at the beginning of the line of random digits and taking consecutive pairs of digits, the first six staff selected would be numbered:

   (1) 46 40 41 50 49 37
   (2) 46 40 08 41 37 45
   (3) 46 13 07 49 17 57
   (4) 46 40 08 13 50 11
   (5) 46 13 07 49 17 13

2. All 10 to 14-year-olds at a Hamilton Intermediate School took part in a study of adolescent attitudes towards the television programme *Shortland Street* (reported in the *Listener*). Suppose that the number of students involved was 200. Four of the students interviewed did not watch the programme. The study found that 69% of the children felt that the programme dealt with 'real-life' issues, and 23% reported instances where they had learned or been guided in their own lives by stories from the *Street*.

   These results should not be taken as reflecting the opinions of the general population of New Zealand school children **mainly** because (select **one** only):

   (1) of the "placebo" effect.
   (2) of interviewer effects.
   (3) of non-response bias.
   (4) four of the children surveyed did not watch the programme.
   (5) of selection bias.

3. A study examined the effects of stress on the subsequent onset of diabetes. People living in stressful environments were selected and followed through time to see if they developed diabetes. All instances of diabetes were recorded as they were diagnosed. This type of study is **best** called:

   (1) a longitudinal study.
   (2) a cross-sectional study.
   (3) a sample survey.
   (4) a controlled experiment.
   (5) a double blind experiment.

4. 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:

(1) a well 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.

(4) A badly presented table because there is no 'average column'.

(5) A badly presented table because the GDP is measured in $US for all countries rather than the local currency.

Questions 5 and **6** refer to the following information.

A study was performed to investigate liver damage in 21 patients with Hepatitis C. Two quantitative and two qualitative variables were measured for each patient, as listed below.

Age: The age of each patient, in years.

AST: The mean level of the enzyme AST for each patient.

Gender: The gender of each patient.

Histology: A category indicating the amount of damage to the liver of each patient (mild, medium, severe).

5. One purpose of the study was to investigate the relationship between **AST** and **Age**. Which **one** of the following exploratory tools would be the **most** appropriate to investigate this relationship?

   (1) A scatter plot.

   (2) A side-by-side box plot.

   (3) A side-by-side dot plot.

   (4) A two-way table of counts.

   (5) A bar chart.

6. Consider the dot plots, in the figure below, of **AST** for each level of **Histology**.

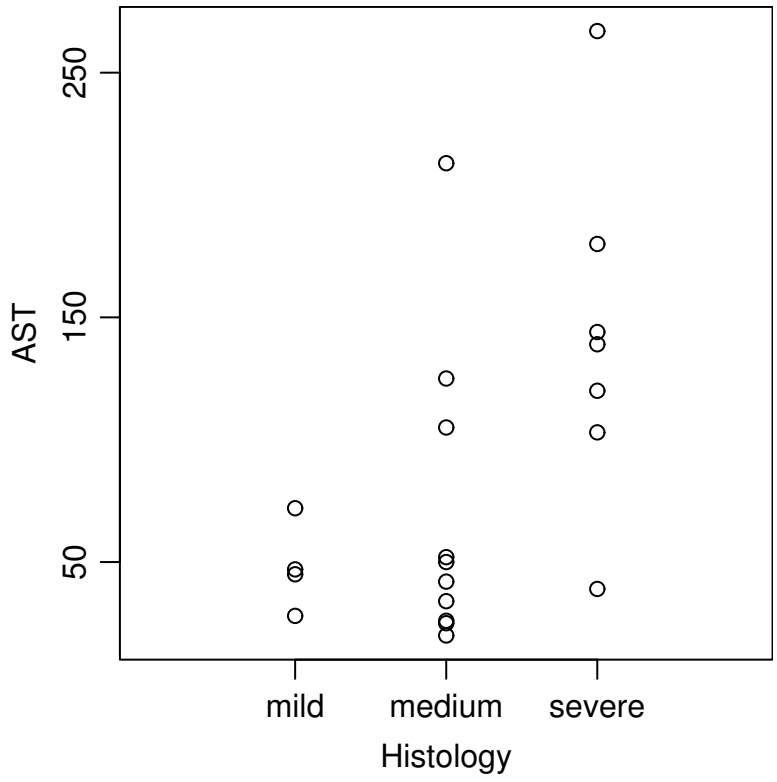


Figure: Dot plots of **AST** for each level of **Histology**

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

(1) The distribution of **AST** is left-skewed (negatively skewed) for patients with medium liver damage.

(2) The distribution of **AST** is reasonably symmetric for patients with severe liver damage.

(3) The value of **AST** is, on average, higher for patients with severe liver damage than it is for patients with only mild damage.

(4) The range of **AST** is similar for patients with severe liver damage and patients with medium damage.

(5) The variability in **AST** is less for patients with mild damage than it is for patients with severe liver damage.

7. A researcher wanted to test whether knowledge of letter sounds helps children to learn to read. To do this, a sample of children was divided into two groups. One group consisted of children with no knowledge of letter sounds; they were given training in the sounds of letters. The second group of children had some knowledge of letter sounds; they were not given any special training. The two groups of children were then tested on their ability to learn to read and their scores compared to see if there was a difference. The **main** problem with this design is (select **one** only):

(1) Insufficient attention is paid to the placebo effect item.

(2) The children were not randomly assigned to "treatments".

(3) The experiment is not an observational study.

(4) The teachers were not "blind" to the treatments.

(5) Children were not "blind" to the treatment they received.

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

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

9. Which one of the following statements about the five number summary below is **false**?

(5.2, 18.1, 19.4, 24.7, 31.1)

(1) The distribution is left skewed.

(2) The interquartile range is 6.6.

(3) There are no observations greater than 31.1.

(4) The mean is 19.4.

(5) The range is 25.9.

10. A recent study looked at the type of school a student attends and that student's performance in bursary examinations. It was found that students at single-sex schools perform significantly better in examinations than those students at coeducational schools. The results from this study alone should not be used to argue the case that single-sex schooling, generally, gives rise to better student performance in examinations **mainly** because:

    (1) there may be a difference between male and female performance in examinations.
    (2) there are many more co-educational schools than single-sex schools in New Zealand.
    (3) the designers of this study would not have been able to use any form of blinding.
    (4) the designers of this study have not used a control group.
    (5) the designers of this study did not allocate each student to the school attended.

Questions 11 and 12 refer to the following concepts:

mean
mode
median
standard deviation
range
interquartile range

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

    (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.

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

    (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.
    (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 the same.

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

(1) It is a good idea to round off numbers when using them in a table for display purposes.

(2) Dot plots should be used for small numbers of observations.

(3) Box plots are good at comparing centres and spreads of quantitative data for two or more groups.

(4) Bar graphs can be used to display discrete data.

(5) The total area under a standardised histogram is equal to the number of observations.

14. A large supermarket chain undertook a study reviewing general operating practices. In one part of the study, checkout operators were classified into four groups, Gp1 through Gp4, with members of Gp1 having the least experience as checkout operators and members of Gp4 having the most experience. A box plot of the number of shoppers processed between 4:30 pm and 6:30 pm on a weekday by checkout operators in each group is shown below.

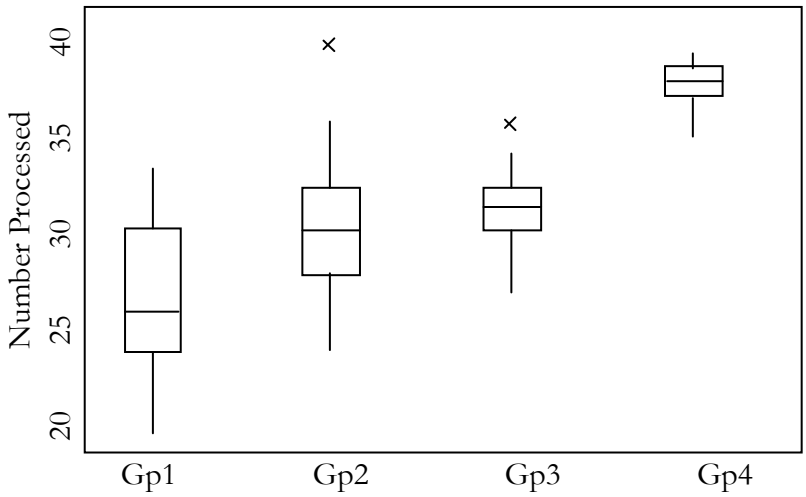


Figure 1: Box Plot of Number of Shoppers Processed by Group

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

(1) The operator who processed the greatest number of shoppers over this time period was a member of Gp2.

(2) The variability of the data as measured by the group interquartile range tends to decrease as the group median increases.

(3) As the experience of a group of operators increases, the average number of shoppers processed increases but the variability of the number of shoppers processed decreases.

(4) Nearly all of the operators in Gp3 processed fewer shoppers than any operator in Gp4.

(5) A scatter plot of group interquartile range versus group median would show no trend.

Question 15 refers to the following information.

The table below shows the total amount of debt owed by each of a university department's fourth year students. The amount of debt owed by each of the 20 students is given to the nearest hundred dollars. The students are ordered by the size of the debt owed. Each student has been assigned an identification number between 1 and 20.

Student	1	2	3	4	5	6	7	8	9	10
Debts ($00's)	0	0	23	37	49	114	124	125	125	127

Student	11	12	13	14	15	16	17	18	19	20
Debts ($00's)	131	151	193	234	260	267	273	281	342	491

Table: Student debt in hundreds of dollars

15. Forty random digits are given below. Select a simple random sample of five (5) students from the population of 20 students by using these random digits. You must start at the beginning of the line of random digits and take consecutive pairs of digits.

    37182 91306 13472 81283 04972 59607 16557 98144

    The sample mean, $\bar{x}$, and the sample standard deviation, s, of the amount of debt owed by the students in your sample are (to the nearest dollar):

    (1) $\bar{x}$ = \$15,520 s = \$9,075
    (2) $\bar{x}$ = \$15,520 s = \$8,117
    (3) $\bar{x}$ = \$20,240 s = \$9,848
    (4) $\bar{x}$ = \$18,640 s = \$6,231
    (5) $\bar{x}$ = \$18,640 s = \$5,573

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

    (1) 'Cross-sectional' and 'longitudinal' studies are observational studies.
    (2) An observational study is often useful for establishing the cause of an effect.
    (3) In an observational study, the researcher compares units that happen to have received each of the treatments.
    (4) An experiment can be conducted only when it is possible (ethically, financially, etc.) for the researcher to allocate treatments to the experimental units.
    (5) A well designed, and well executed experiment can reliably establish causation.

17. On the twelfth anniversary of the (alleged) death of Elvis Presley, a Dallas record company sponsored a national telephone call-in survey. Listeners of over 1000 American radio stations were asked to phone-in (at a charge of $2.50 per call) to voice an opinion on whether or not Elvis was really dead. It turned out that 56% of the callers felt that Elvis was still alive.

    This is an unreliable estimate of the percentage of all Americans who believed that Elvis was still alive at that time **mainly** because:

    (1) of biases from question effects and survey format.
    (2) of bias from transferring findings.
    (3) a pilot study should have been conducted to find out which radio stations should have been included.
    (4) of sampling errors.
    (5) of selection and self-selection biases.

18. At around midnight on the 5th March 1997, the Deputy Prime Minister of New Zealand, the Right Honourable Mr. Winston Peters, allegedly assaulted the MP for Whangarei, the Right Honourable Mr. John Banks. The following evening TV3 ran an opinion poll on the question "Is Mr. Peters fit to be Deputy Prime Minister of New Zealand?". Viewers were invited to dial an 0900 number (at a cost of 99c/min) and record either "yes" or "no". 10% of calls said Mr. Peters was fit to be Deputy Prime Minister while 90% said Mr. Peters was not fit to be Deputy Prime Minister.

    The type of non-sampling errors likely to bias the results the **most** are:

    (1) Behavioural effects and nonresponse bias.
    (2) Self-selection bias and selection bias.
    (3) Nonresponse bias and question effects.
    (4) Transferring findings and behavioural effects.
    (5) Selection bias and transferring findings.

19. We need to determine whether a new treatment (treatment X) is effective in treating a certain disease. Which **one** of the following experiments will give the **clearest** information?

    (1) All the patients are given treatment X.
    (2) Half the patients randomly selected to receive the treatment. The other half receives a placebo.
    (3) Patients that ask for treatment X are randomly allocated to either treatment X or a placebo. Those that do not ask for treatment X are given a placebo.
    (4) Patients arriving in the morning receive treatment X and those arriving on the afternoon receive a placebo. It is randomly decided whether to give treatment X in the morning or in the afternoon.
    (5) Those patients who have had a recent attack of the disease are given treatment X. The others are given a placebo.

Questions 20 to **23** 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

Table: Sports Foundation Grants

20. 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.
(2) round all grants to the nearest thousand dollars.
(3) list the sports in order of the amount of the grant received in the year 2000-2001.
(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'.

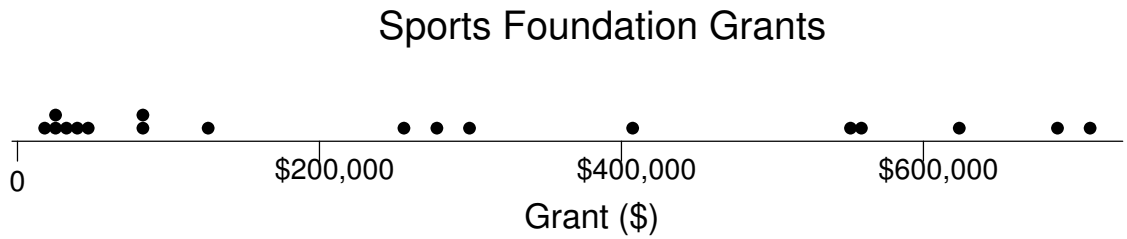


Figure: Sports Foundation Grants, 2000–2001

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

(1) side-by-side box plots with the same scaled *x*-axes.

(2) a labelled bar graph ordered by the size of the grant.

(3) a histogram with equal width class intervals for **Grants** on the *x*-axis.

(4) a scatter plot with **Grants** as the response variable and **Sport** as the explanatory variable.

(5) a pie chart with the sectors labelled and ordered by the size of the grant.

22. The five sports that received the highest grants in the year 2000–2001 are given in the table below.

Sport	Grant
Rowing	$707,265
Cycling	$688,140
Yachting	$622,356
Equestrian	$558,620
Hockey	$554,000

Table: Highest Five Grants Awarded, 2000–2001

The sample mean, $\bar{x}$, and sample standard deviation, s, for these five sports are:

(1) $\bar{x}$ = \$626,076, s = \$71,068

(2) $\bar{x}$ = \$622,356, s = \$71,068

(3) $\bar{x}$ = \$622,356, s = \$63,565

(4) $\bar{x}$ = \$622,000, s = \$71,068

(5) $\bar{x}$ = \$626,076, s = \$63,565

23. Suppose we wish to randomly select five of the sports listed in the table on page 19. The method for randomly selecting the five sports uses the number (in the **No.** column of the table on page 19) associated with each sport and random number digits. Use the row of random digits below to select a simple random sample of five sports. You must start at the beginning of the row and use consecutive pairs of digits.

09874 11018 39090 54804 17130

The five sports selected are:

(1) Judo, Rowing, Yachting, Judo, Cycling

(2) Judo, Shooting, Archery, Wrestling, Swimming

(3) Judo, Shooting, Archery, Cycling, Wrestling

(4) Judo, Rowing, Yachting, Cycling, Boxing

(5) Judo, Hockey, Gymnastics, Boxing, Archery

Question 24 refers to the following stem-and-leaf plot.

Stem-and-Leaf Plot of Engine Size

Units: 2 | 1 = 2100cc

n = 60

1 | 3 3 4
1 | 5 6 6 6 6 6 6 6 7 7 8 8 8 8 9
2 | 0 0 0 0 0 0 0 0 0 0 2 3 3 3 3 4 4 4 4
2 | 5 5 5 5 5 8 8 8 8 9
3 | 0 0 2 2
3 | 5 8
4 | 0 0 0
4 | 9
5 |
5 | 7 7

Figure: Engine sizes of new cars

24. Using the stem-and-leaf plot in given above (of a random sample of 60 models of new cars) which **one** of the following statements is **false**?

(1) 10% of the cars have an engine size greater than 3900cc.

(2) The engine sizes of the cars are negatively (left) skewed.

(3) The distribution is unimodal.

(4) Less than 25% of the cars have an engine size between 2100cc and 2600cc.

(5) The range is 4400cc.

25. An experiment was conducted to study some of the effects of ageing on concentration and balance. Nine elderly and eight young subjects participated in an experiment. Each subject stood barefoot on a "force platform" and was asked to maintain a stable upright position as well as to react as quickly as possible to an unpredictable noise by pressing a hand-held button. The platform automatically measured how many centimetres each subject moved both in the forward/backward and the side-to-side directions.

    A sample of forty young people was taken. The side-to-side sway of this sample is summarised in the table below.

Side-to-Side (cm)	Frequency
1.0	1
1.1	0
1.2	0
1.3	3
1.4	2
1.5	5
1.6	8
1.7	4
1.8	7
1.9	2
2.0	2
2.1	5
2.2	0
2.3	1
Total	**40**

    The sample mean, $\bar{x}$, and the sample standard deviation, s, of the side-to-side sway in this sample are:

    (1) $\bar{x} = 1.70,\ s = 0.2702$
    (2) $\bar{x} = 1.68,\ s = 0.2557$
    (3) $\bar{x} = 1.72,\ s = 0.2757$
    (4) $\bar{x} = 1.70,\ s = 0.2736$
    (5) $\bar{x} = 1.68,\ s = 0.2591$

26. 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 humans 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.

Questions 27 to **29** refer to the following information.

During 1999, 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 107, 108 and 101 refer to the various streams.

	Stream			
Response	107	108	101	**Total**
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: Responses to question regarding Excel experience.

27. The variable Stream is:
    (1) discrete.
    (2) quantitative.
    (3) qualitative.
    (4) dependent.
    (5) continuous.

28. Which one of the following is not a potential source of error in this analysis?
    (1) Behavioural considerations.
    (2) Question effects.
    (3) Nonresponse bias.
    (4) Interviewer effects.
    (5) Sampling error.

29. 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.
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.

30. Which **one** of the following statements is **false**?
(1) Blinding and double blinding are techniques often used by researchers when people are used as experimental units.
(2) Blocking is used in experiments to ensure fair comparisons with respect to factors the experimenter believes are important.
(3) In an experiment, the control group always receives no treatment.
(4) The placebo effect is the response caused in human subjects by the idea that they are being treated.
(5) Randomisation in experiments allows the calculation of the likely size of sampling errors.

ANSWERS

1. **(4)**	2. **(5)**	3. **(1)**	4. **(2)**	5. **(1)**	6. **(1)**
7. **(2)**	8. **(3)**	9. **(4)**	10. **(5)**	11. **(3)**	12. **(3)**
13. **(5)**	14. **(5)**	15. **(1)**	16. **(2)**	17. **(5)**	18. **(2)**
19. **(2)**	20. **(1)**	21. **(2)**	22. **(1)**	23. **(4)**	24. **(2)**
25. **(4)**	26. **(2)**	27. **(3)**	28. **(4)**	29. **(4)**	30. **(3)**