UCLA Stat10 Statistical Reasoning - Midterm Review
Observational Studies, Designed Experiments & Surveys

In early 1997 Whitcoulls bookstores conducted a nation-wide survey. Whitcoulls' shoppers were invited to fill in a short survey. They were asked to list, in order, their three favourite books. Using the results, Whitcoulls published the list "New Zealand's 100 Favourite Books". The top twenty books from this list are given below.

No.	Book	No.	Book
1	The Lord of the Rings	11	April Fool's Day
2	The Power of One	12	Complete Winnie the Pooh
3	Pride and Prejudice	13	The Runaway
4	The Bible	14	Clan of the Cave Bear
5	Wild Swans	15	Long Walk to Freedom
6	The Horse Whisperer	16	Sleepers
7	Cross Stitch	17	Jane Eyre
8	Goosebumps	18	Gone With the Wind
9	The Bone People	19	Wuthering Heights
10	The Hobbit	20	The English Patient

1. Based on the information stated above, the two most obvious types of non-sampling errors that may be present in this survey are:

(1) self-selection bias and interviewer effects.

(2) random errors and non-response bias.

(3) self-selection bias and question effects.

(4) selection bias and self-selection bias.

(5) selection bias and chance errors.

2. A student wants to select and read any 4 books from this top twenty books list. Choose a simple random sample of 4 books for this student. To select the sample you must use the thirty random digits given below. Start at the beginning of the line of random digits given below.

87105 75663 05103 47781 00910 21112

The four books in the random sample are:

(1) Wild Swans, April Fool's Day, The Hobbit, Goosebumps.

(2) The Hobbit, Wild Swans, The Hobbit, The Power of One.

(3) The Hobbit, Wild Swans, The Power of One, April Fool's Day.

(4) Goosebumps, Cross Stitch, The Hobbit, Wild Swans.

(5) Goosebumps, Cross Stitch, The Lord of the Rings, Wild Swans.

3. Consider the following three studies:

Study 1: An animal researcher was interested in cats' abilities to survive surprisingly high falls if they had time to twist round and prepare for the impact. Vets in New York City recorded incidents of cats falling out of apartment windows. The data was divided into three groups: cats that fell from one or two storeys above the ground; cats that fell from three to five storeys above the ground and cats that fell from six or more storeys above the ground. The proportion of cats that survived in each group was then compared.

Study 2: A random sample of 100 students is asked to keep a diary in which they record their clothing expenditures for the next three months. The expenditures of males and females are then compared.

Study 3: A sample of 50 shoppers at an appliance store is split into two groups. One group is shown a television commercial for a new range of appliances that has been filmed in the same style as previous television commercials for the store. The second group is shown a television commercial for the same new range of appliances that has been filmed in a totally new style. An hour after viewing the commercial, each of the shoppers was asked what they could recall about the new range of appliances and a score based on their recollection was recorded. The recall scores were then compared for the two groups.

(i) For each study, describe what "treatment" is being compared.

Study 1:

Study 2:

Study 3:

(ii) Which of the three studies would be described as experiments and which would be described as observational studies?

Study 1:

Study 2:

Study 3:

(iii) For the studies that are observational, briefly explain why an experiment could not be carried out instead.

4. In 1950 two hundred employees from the Christchurch Firestone Tire and Rubber Company became part of a cancer study. These employees were observed until 1996 and any occurrences of cancer within this group were recorded. This study is **best** called:

(1) a double-blind experiment.

(2) a randomised experiment.

(3) a sample survey.

(4) a retrospective observational study.

(5) a prospective observational study.

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

(1) Non-sampling errors are often bigger than the random sampling errors in surveys.

(2) People will sometimes answer a question differently for different interviewers.

(3) Sophisticated sampling projections can always correct the results if the population you are sampling from is different to the one of interest.

(4) Slight changes in the wording of questions can often make a big change to survey results.

(5) Non-response can cause bias in surveys because non-respondents can behave differently from people who respond.

6. A TIME daily poll on the Internet invited readers to make a choice from a given list of options, in response to the following question:

> *"Three times in the last five months, children went on killing sprees. What is fuelling this bizarre and tragic trend?"*

As of 2 June 1998, the largest proportion of respondents (29%) chose the option:

> *"Nurture: The American family is crumbling; permissive parents are raising wild children."*

We wish to use this percentage as an estimate of the proportion of all Americans who believe that *Nurture* is the cause.

Which **one** of the following is **not** a potential source of non-sampling error in this survey?

(1) Question effects.

(2) Self-selection bias.

(3) Selection bias.

(4) Non-response bias.

(5) Transferring findings.

7. Television polls have become commonplace in New Zealand over the last few years. A television sports programme often runs polls on questions such as: *"Do you approve or disapprove of Wayne Smith as the All Black coach?"* Viewers are then invited to phone in their vote at a cost of approximately 99 cents per minute. Identify two sources of bias in this form of survey.

8. TIME magazine, 20 December 1993, reported that 70% of Americans answered "Yes" to the question *"Do you favour stricter gun-control laws?"* The figure was obtained from a telephone poll of 500 adult Americans. Are the following statements true or false? Explain briefly.

(i) The sample was too small to provide any useful results.

(ii) The survey does not take into account the views of homeless people.

(iii) The survey may be inaccurate due to non-response bias.

(iv) The survey should be repeated so that it includes a control group.

9. Two drugs are to be compared. A group of 20 people are each randomly allocated one of the two drugs. Neither the people who were treated nor the doctor who administered the drugs knew who got the drug. Which best describes this situation?

(1) An observational study.

(2) A double blind experiment.

(3) A sample survey.

(4) A case-control study.

(5) A block design.

Graphs, Numerical Summaries, Histograms

1. The weights of 9 stage II engineering students were recorded as part of a class experiment. The weights, in kilograms, of these 9 students were: 70, 75, 60, 102, 67, 85, 97, 60, 70.

(a) Draw a dot plot of the weights of the students.

(b) Comment on the main features in this sample.

2. At one stage in the process of producing silicon chips, a very thin layer of silicon oxide is deposited on a "wafer". The wafer is then broken up into chips. Using the following data from *Technometrics* (1994), draw a stem-and-leaf plot of the thickness of silicon oxide in 30 such chips. The thickness has been measured in a special unit for very small distances called Angstrom units, Å.

840	900	930	940	950	960	970	980	990	990	1000	1000	1000	1010	1010
1030	1030	1030	1040	1040	1050	1050	1050	1050	1050	1070	1070	1100	1100	1120

(a) Complete the stem-and-leaf plot for these 30 thicknesses.

Units: 9 | 5 = 950Å

8	
8	
9	
9	
10	
10	
11	

(b) For this data set, the median is:

(1) 950Å **(2)** 1030Å **(3)** 1010Å **(4)** 1012Å **(5)** 1020Å

(c) The lower quartile for the above data is:

(1) 840Å **(2)** 940Å **(3)** 985Å **(4)** 975Å **(5)** 980Å

(d) Which of the following statements is **not** a feature of the data?

(1) The interquartile range is 70Å.

(2) The range is 270Å.

(3) The mode is 1050Å.

(4) The median is 1020Å.

(5) Those observations with values 1100Å or more represent about 10% of the distribution of thicknesses.

3. A second batch of 6 chips yielded the following values, in Å:

940, 960, 1010, 980, 1040, 970.

The sample mean, $\bar{x}$, and the sample standard deviation, *s*, for this data set are, respectively:

(1) 983 and 50 **(2)** 983 and 33 **(3)** 983 and 36 **(4)** 975 and 50 **(5)** 975 and 36

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

(1) The mean is less affected by outliers than the median.

(2) Outliers affect the standard deviation more than they affect the interquartile range.

(3) The numbers of cars owned by a family is a continuous variable.

(4) Box plots are good at distinguishing between unimodal and bimodal distributions.

(5) When coding qualitative variables (i.e. using numbers to describe the outcomes) it is a good idea to work out the means and medians.

5. Do you agree with the following statements? Discuss.

(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 samples with a small number of observations.

(3) Box plots are not good for comparing centres of location and spreads of data.

(4) Bar graphs cannot be used to display discrete data.

6. Draw a box plot for the following set of data:

18	19	21	21	23	23	23	27	29
29	30	31	35	41	49	55	78	

Five-number summary: (18, 22, 29, 38, 78)

7. Do you agree with the following statements? Discuss.

(1) The distribution from which this sample is drawn is highly skewed.

(2) The interquartile range is 21.

(3) There are no observations greater than 78.

(4) The observation 78 is an outside value for the box plot representing the above data.

(5) The observation 18 is an outside value for the box plot representing the above data.

8. The five-number summary for a set of data is:

(10, 22, 37, 50, 60)

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

(1) Each of the whiskers on the box plot of the data must be greater than 42 units in length.

(2) It is not possible determine the mean of the data from this five-number summary.

(3) At least half of the observations are between 22 and 50 inclusive.

(4) The interquartile range is 28.

(5) None of the observations in the data set is an outside value on the box plot of the data.

Questions 9 to 11 refer to the following information.

The stem-and-leaf plot below shows the annual salaries for the 21 employees in the engineering department of the Technitron company.

Stem-and-leaf plot of SALARY n = 21 Units: 4 | 7 = $47,000

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

9. The median for the SALARY data set is:

(1) $39

(2) $39,000

(3) $11

(4) $11,000

(5) $35,000

10. The upper quartile for the SALARY data set is:

(1) $49,500

(2) $53,000

(3) $51,000

(4) $49,000

(5) $46,000

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

(1) The stem-and-leaf plot is drawn incorrectly because the second to last line should have been omitted, as there are no data values on it.

(2) The stem-and-leaf plot is drawn incorrectly as there is a 0 missing on the second to last line.

(3) The stem-and-leaf plot is drawn correctly despite the fact that there is only one row for stem 2.

(4) The stem-and-leaf plot has been drawn correctly because the length of the plot is such that there is one stem-digit with more leaf-digits than any other stem-digit.

(5) The stem-and-leaf plot is drawn incorrectly because 4 | 7 in the units statement is not a data value.

Exploratory Tools for Relationships

Section A: Types of Variables

1. **(a)** **Quantitative** variables are ____ and counts.

(b) **Qualitative** variables describe _________________ ________________________________.

2. **Quantitative variables** can be either *discrete* or *continuous*.

(a) Variables with **few** *repeated values* are treated as _____________________________________.

(b) Variables with **many** *repeated values* are treated as ___________________________________.

3. **Qualitative variables** can be either *categorical* or *ordinal*.

(a) Variables **with order** are called ___________________________________.

(b) Variables **without order** are called **_____________________________________.**

4. **(a)** To explore the relationship between two **quantitative** variables we use a __________________

_______________________.

(b) To explore relationships between a **qualitative** variable and a **quantitative** variable we use

___________ plots, ________________________________ plots and _______________ plots.

(c) To explore the relationship between two **qualitative** variables we use a ___________________

__________________ of _________________________.

Section B: Two Variables

Questions 1 and 2 refer to the following information.

TVNZ News, 5 August 1997, reported that smoking is on the increase in the high socio-economic group in the USA. It was claimed that the advertising and fashion industries are responsible for this increase. The data shown in the table below is a subset of the data from a study on a large number of people. Each person has measurements made on variables that describe some aspect of their image.

ID	Gender	Weight (kg)	Socio-Ec Status	Smoking Status	Age	…
1	Female	50	High	Smoker	21-30	…
2	Male	75	Low	Smoker	31-40	…
3	Male	68	Middle	Non-smoker	51-60	…
4	Female	55	Middle	Non-smoker	11-20	…

Table 1: Data on People's Images

1. The most appropriate way to begin to explore the relationship between Socio-Economic Status and Smoking Status is to construct a:

(1) two-way table of counts with Socio-Economic Status for the row values and Smoking Status for the column values.

(2) dot plot of Socio-Economic Status for each level of Smoking Status, using the same scale for each plot.

(3) box plot of Socio-Economic Status for each level of Smoking Status, using the same scale for each plot.

(4) frequency table for each of these two variables.

(5) scatter plot of Socio-Economic Status against Smoking Status.

2. The most appropriate way to begin to explore the relationship between Weight and Smoking Status is to construct a:

(1) two-way table of counts with Weight for the row values and Smoking Status for the column values.

(2) dot plot of Weight for each level of Smoking Status, using the same scale for each plot.

(3) box plot of Weight for each level of Smoking Status, using the same scale for each plot.

(4) frequency table for each of these two variables.

(5) scatter plot of Weight against Smoking Status.

Questions 3 and 4 refer to the following information.

A record of quarterly sales revenues and the corresponding advertising costs from a large retail outlet is given below.

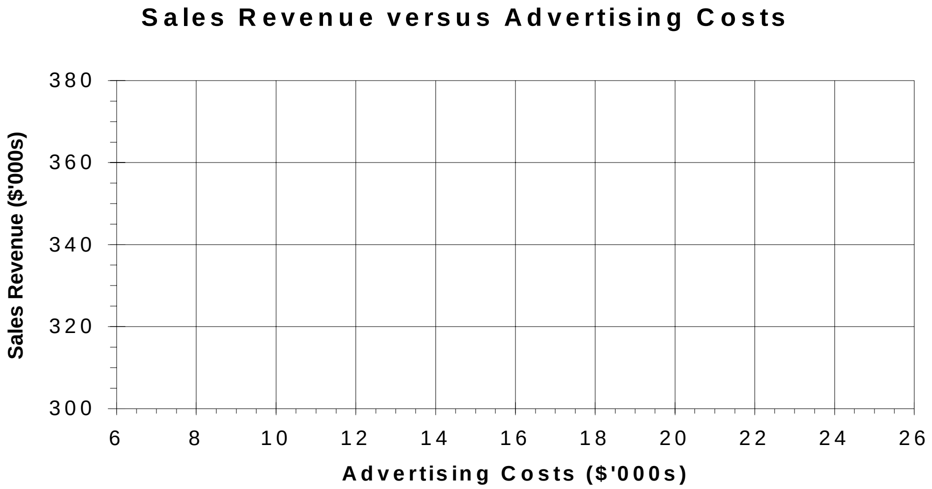
Quarter	1	2	3	4	5	6	7	8
Advertising Costs ($'000s)	10	12	8	20	11	15	10	25
Sales Revenue ($'000s)	342	347	318	350	351	346	345	367

Table 2: Quarterly Advertising Costs and Sales Revenues

3. If we want to investigate the relationship between the quarterly advertising costs and the quarterly sales revenues, then the most appropriate plot to look at is a:

(1) dot plot of the combined sales revenue data and advertising costs data.
(2) back-to-back stem-and-leaf plot of sales revenue and advertising costs.
(3) histogram of the combined sales revenue data and advertising costs data.
(4) dot plot of sales revenue and a dot plot of advertising costs (plotted on the same axes).
(5) scatter plot of sales versus advertising costs.

4. Draw a scatter plot of the above data, fit a trend curve by eye and describe anything interesting you see in the plot.



Interpretation:

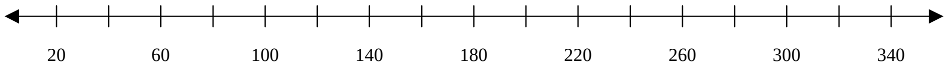
5. The following table gives the lengths (in kilometres) of the major rivers in the South Island.

Flowing into Pacific Ocean				**Flowing into Tasman Sea**			
Clutha	322	Selwyn	95	Buller	177	Hokitika	64
Taieri	288	Ashburton	90	Grey	121	Arahura	56
Clarence	209	Opihi	80	Motueka	108	Mokihinui	56
Waitaki	209	Shag	72	Karamea	80	Wanganui	56
Waiau	169	Kakanui	64	Taramakau	80	Whataroa	51
Waimakariri	161	Waihao	64	Hollyford	76	Waimea	48
Rakaia	145	Waipara	64	Aorere	72	Waitaha	40
Hurunui	138	Pareora	56	Takaka	72	Karangarua	37
Rangitata	121	Conway	48	Arawata	68	Heaphy	35
Ashley	97			Cascade	64	Cook	32
				Haast	64	Waiho	32

Table 3: Lengths of major rivers in the South Island (in kilometres)

(a) For each of the two groups of rivers, find the median, lower quartile and upper quartile.

(b) Draw a side-by-side box plot of the two sets of river lengths.



(c) Describe what you see in the plots.

Chapter 4 – Probabilities and Proportions

1. In 1995 there were 2011 students enrolled in either 528.181 or 528.188 (Stage I Statistics) at the City campus. The numbers of female and male students are given in the following table.

	Females	Males	**Total**
528.181	604	593	1197
528.188	387	427	814
Total	991	1020	2011

(a) Convert the above table of counts into a probability table (to 4 dp).

	Females	Males	Total
528.181			
528.188			
Total			

(b) One of the 2011 students is chosen at random. What is the probability that the student chosen is:

(i) a male taking 528.181?

(ii) a female?

(iii) a female taking 528.188?

(c) Given that a student is taking 528.188, what is the probability that they are male?

(d) What is the probability that a randomly chosen male student is taking 528.188?

2. Consider drivers stopped at random for breath testing. Below is a partially completed probability table providing information about such drivers, with regards to their age (40 or under, over 40) and whether they were (or were not) wearing seat belts.

	40 or under	Over 40	**Total**
Wearing a seat belt	0.484		0.853
Not wearing seat belt		0.081	
Total			1

(a) Complete the table.

(b) What is the probability that a driver stopped at random is not wearing a seat belt?

(c) If a driver stopped at random is not wearing a seat belt, then what is the probability the driver is over 40?

(d) What is the probability that a driver stopped at random is 40 or under?

Chapter 5 – Discrete Random Variables

Section A: Discrete Random Variables

1. Consider the experiment of tossing two fair coins.
The sample space is {HH, HT, TH, TT}.
Let random variable X be the number of tails.

(a) The probability function for this experiment is:

x	
pr($X = x$)	

(b) Find the probability that:

(i) X is more than 1

(ii) X is at least 1

(iii) X is at most 2

2. Random variable Y has the following probability function:

y	5	6	10	12	13	18	25
pr($Y = y$)	0.10	0.07	0.25	0.15	0.03	0.28	0.12

Find the probability that:

(a) Y is more than 12

(b) Y is no more than 10

(c) Y is at least 6

(d) Y is at least 6 and at most 12

(e) Y is at least 10 and at most 12

(f) Y is more than 13 but less than 25.

Section B: Binomial Distribution

1. The owner of a small bookshop estimates that 30% of the customers who enter the shop purchase at least one book. At 10.30am on a particular day there are 7 potential customers in the shop. Assuming that these customers can be regarded as a random sample from the population of all potential customers, calculate the probability that at least two of these people purchase at least one book.

2. The manufacturer of disk drives for a well-known brand of computers expects 5% of the drives to malfunction during the computer's warranty period. Let X be the number of disk drives, in a batch of 10 randomly selected disk drives, which malfunction during this period. X has a Binomial distribution.

(a) Identify n and p, the parameters of the Binomial random variable.

(b) In the context of this exercise, state the assumptions required for X to have a Binomial distribution.

(c) Are the assumptions satisfied here?

(d) Calculate the probability that:

(i) no disk drive will malfunction during the warranty period.

(ii) exactly one disk drive will malfunction during the warranty period.

(iii) at least two disk drives will malfunction during the warranty period.

(iv) between 2 and 5 (inclusive) disk drives will malfunction during the warranty period.

Section C: Poisson Distribution

1. What three conditions are necessary for X to have a Poisson distribution?

(a)

(b)

(c)

2. Use the Poisson tables to calculate the following probabilities.

(a) If $X \sim$ Poisson (λ = 4.8) then pr($X \leq 2$) =

(b) If $X \sim$ Poisson (λ = 0.8) then pr($X = 3$) =

(c) If $X \sim$ Poisson (λ = 2) then pr($X \geq 2$) =

(d) If $X \sim$ Poisson (λ = 9.5) then pr($X < 10$) =

(e) If $X \sim$ Poisson (λ = 6.5) then pr($3 \leq X \leq 8$) =

Section D: Differences Between Poisson and Binomial Distributions

In questions 1 to 3, choose the option (from those given) that is the **best** model of the random variable described.

1. A University department computing laboratory contains 40 Macintosh computers (Macs). 90% of Macs run for more than 1000 hours between breakdowns. At the start of a year the department buys 40 new Macs. The actual distribution of the number of Macs that need to be repaired by the time the laboratory has been open for 1000 hours is:

(a) Poisson (λ = 40 x 0.1) **(b)** Binomial (n = 40, p = 0.1)
(c) Binomial (n = 40, p = 0.9)

2. A journalist researching a story about the Tamaki campus visits a Stage I Statistics lecture of 500 students. She assumes that the composition of the Stage I statistics class is representative of the students at Tamaki. As the students take notes she chooses a random sample of 20 and counts the number who are writing with their left hand. If 15% of all students are left-handed the distribution of W, the number of left-handers in her sample, is:

(a) Poisson (λ = 20) **(b)** Poisson (λ = 20 x 0.15)
(c) Binomial (n = 500, p = 0.15) **(d)** Binomial (n = 20, p = 0.15)

3. There were 385 aggravated robberies and 31 murders in Auckland for the year to June 1991 compared to 407 aggravated robberies and 34 murders in the year to June 1992.

If aggravated robberies occur randomly at a constant rate of 400 per year, then the distribution of X, the number of aggravated robberies in any particular year is:

(a) Poisson ($\lambda = \frac{385+407}{2}$) **(b)** Poisson ($\lambda$ = 400)
(c) Binomial (n = 400, $p = \frac{385}{407}$) **(d)** Binomial (n = 400, p = unknown)

Section E: Choosing an Appropriate Probability Model

In questions 1 to 8, state an appropriate model for the distribution of random variable *X*. Choose from: Binomial, Poisson, or neither Binomial nor Poisson. State the parameters of your model and briefly discuss the assumptions underlying the model.

1. A fire occurred at a warehouse in Wellington. 200 boxes out of 1200 were water damaged when the fire was extinguished. These boxes were mixed up with all of the other boxes while the warehouse was being repaired. The company decided to sell all boxes at a reduced price (informing prospective buyers that some of the boxes were damaged). The purchasing company randomly selects 25 boxes and delivers them to their retail outlet. Let *X* be the number of these 25 boxes that have been damaged.

2. A shuttle bus service between the Tamaki Campus and City Campus has a capacity of 30 passengers. Information collected over a long period of time shows that, on average, 40% of passengers are female. For a Tamaki Campus to City Campus journey, let *X* be the number of male passengers who get on the shuttle before the first female boards the shuttle.

3. Cataracts are a very rare birth defect. In New Zealand they affect, on average, 3 babies out of every 100,000 live births. Let *X* be the random variable for the number of babies born with cataracts in a given year in which 50,000 babies were born.

4. In each round of a certain game, a player tosses three fair coins into the air. The player receives $2 for each head that is tossed. It costs $3.50 to play a round of this game. Let *X* be the number of heads tossed by a player in a single round of this game.

5. The number of deaths due to strokes in the Auckland region each year varies randomly with 550 deaths per year, on average. Let *X* be the number of deaths due to strokes in a given 6-month period.

6. An Auckland car sales company is offering customers an incentive when they purchase a new car. At the end of the month each customer purchasing a car will be offered an opportunity to win either a major prize or a minor prize determined by the spinning of a wheel. There is a one-in-ten chance of winning a major prize (a $2000 refund). If they don’t win a major prize, they win a minor prize (a $500 refund). At the end of a month 45 cars had been sold. Let *X* be the number of customers who receive a major prize.

7. A Stage I Statistics student sits the multi-choice term test of 26 questions. Each question has 5 possible answers, of which only one is correct. Furthermore the student, who has done no work for the test, randomly guesses an answer for each question. Let *X* be the number of questions the student completes before obtaining his or her first correct answer.

8. In a particular semester 1400 students are enrolled in the Stage I Statistics papers. Of these students, 280 had not done any mathematics in the previous two years. One of the streams had 200 students randomly allocated to it. Let *X* be the number of students in that lecture stream who had not done any mathematics in the previous two years.

3. The medical records of a group of diabetic patients presenting at a clinic showed that 50 presented as serious cases, while 36 presented as mild cases. Of the 31 patients aged under 40, 16 presented as mild cases.

(a) Present this information in the table below.

(b) A patient is chosen at random. Find the probabilities that:

(i) the patient is under 40 and has a mild case.

(ii) the patient is at least 40 years old or has a serious case.

(iii) the patient has a serious case and is at least 40 years old.

(c) Of those presenting with serious cases, what proportion are aged under 40?

(d) Of those aged at least 40, what proportion present with mild cases?

4. A bank classifies borrowers as high-risk or low-risk. Of all its loans, 5% are in default. Forty percent (40%) of those loans in default are to high-risk borrowers, while 77% of loans not in default are to low-risk borrowers.

(a) Complete the table.

(b) What percentage of loans is made to borrowers in the high-risk category?

(c) What is the probability that a high-risk borrower will default on his or her loan?

5. According to recent figures from the National Centre of Educational Statistics (US), 17.5% of all bachelor’s degrees are in business. 27% of bachelor’s degrees in business are obtained by women and 48.75% of other degrees are obtained by men.

(a) Complete the table.

(b) What is the probability that a randomly selected recent bachelor’s degree graduate will be a man?

(c) What is the probability that a randomly selected recent bachelor’s degree graduate will be a man with a degree in business?

(d) What is the probability that a randomly selected female recent bachelor’s degree graduate will have a degree in business?

6. (1998 Semester 1 Term Test)

A drinking pattern found by a survey is that 19% of male drinkers and 10% of female drinkers drink alcohol daily. Also, 51% of all drinkers are male (a ‘drinker’ was defined as someone who had consumed alcohol in the previous 12 months).

The probability that a randomly selected drinker from this survey who drinks alcohol daily is female is:

(1) 0.3448
(2) 0.3358
(3) 0.0490
(4) 0.1459
(5) 0.2041

Chapter 6 – Continuous Random Variables

Section A: Probability Density Function Quiz

The probability distribution function of a continuous random variable is represented by a *density curve*. The following quiz is about the density curve.

1. How are probabilities represented?

2. What is the total area under the density curve?

3. What parameter is at the point where the density curve balances?

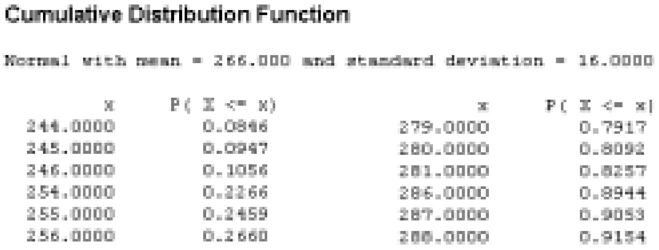
4. When we calculate probabilities for a continuous random variable, does it matter whether interval endpoints are included or excluded?

5. Write down some features of the Normal distribution p.d.f. curve.

6. What are the parameters of the Normal distribution?

Section B: Normal Distribution

1. The natural gestation period for human births, *X*, has a mean of about 266 days and a standard deviation of about 16 days. Assume that *X* is Normally distributed with a mean of 266 days and a standard deviation of 16 days.



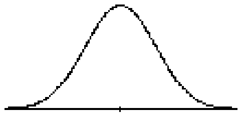
Cumulative Distribution Function

```
Normal with mean = 266.000 and standard deviation = 16.0000

        x       P( X <= x)              x       P( X <= x)
 244.0000         0.0846         279.0000         0.7917
 245.0000         0.0947         280.0000         0.8092
 246.0000         0.1056         281.0000         0.8257
 254.0000         0.2266         286.0000         0.8944
 255.0000         0.2459         287.0000         0.9053
 256.0000         0.2660         288.0000         0.9154
```

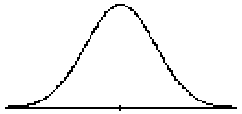
Use the STATA output above to answer the following questions.

Calculate the proportion of women who carry their babies for:

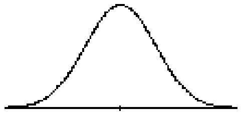
(a) less than 245 days (ie, deliver at least 3 weeks early).



(b) between 255 and 280 days.

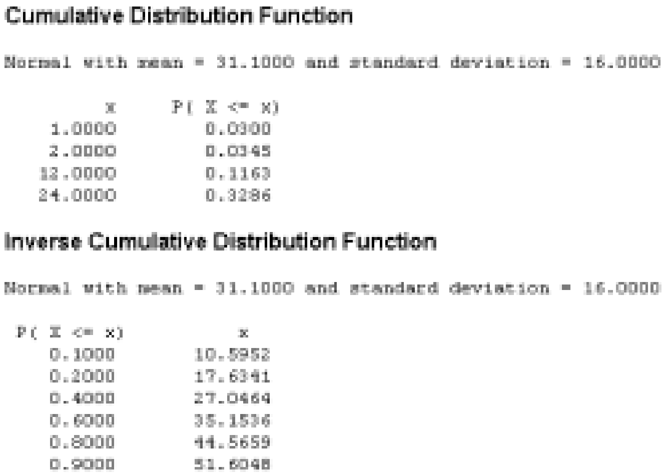


(c) longer than 287 days (ie, the baby is more than 3 weeks overdue).



2. A medical trial was conducted to investigate whether a new drug extended the life of a patient who had lung cancer. Assume that the survival time (in months) for patients on this drug is Normally distributed with a mean of 31.1 months and a standard deviation of 16.0 months.

(a) Use the following STATA output to answer the questions below.



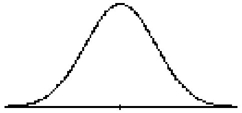
Cumulative Distribution Function

```
Normal with mean = 31.1000 and standard deviation = 16.0000

        x      P( X <= x)
   1.0000        0.0300
   2.0000        0.0345
  12.0000        0.1163
  24.0000        0.3286
```

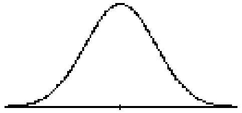
Inverse Cumulative Distribution Function

```
Normal with mean = 31.1000 and standard deviation = 16.0000

 P( X <= x)          x
    0.1000       10.5952
    0.2000       17.6341
    0.4000       27.0464
    0.6000       35.1536
    0.8000       44.5659
    0.9000       51.6048
```

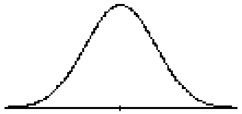
(i) Calculate the probability that a patient survives for no more than one year.



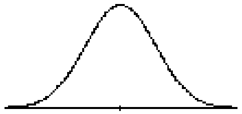
(ii) Calculate the proportion of patients who survive for between one year and two years.



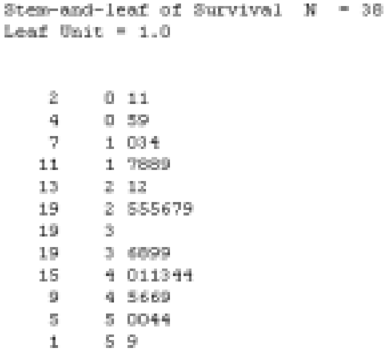
(iii) Calculate the number of months beyond which 80% of the patients survive.



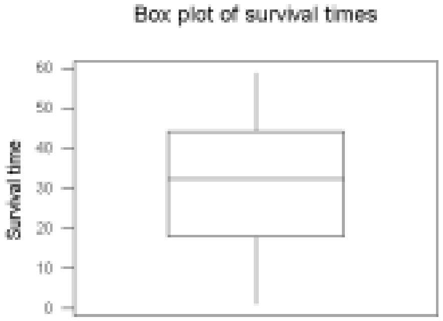
(iv) Calculate the range of the central 80% of survival times.

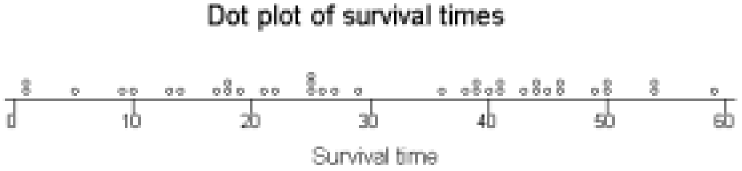


(b) A sample of survival times is taken for 38 patients on this drug. Plots of these 38 survival times are shown below. Use these plots to comment on the validity of the assumption that the survival time is Normally distributed.



```
Stem-and-leaf of Survival  N  = 38
Leaf Unit = 1.0


  2    0 11
  4    0 59
  7    1 034
 11    1 7889
 13    2 12
 19    2 555679
 19    3
 19    3 6899
 15    4 011344
  9    4 5669
  5    5 0044
  1    5 9
```





Comment:

3. The designer of a new aircraft’s cockpit wants to position a switch so that most pilots can reach it without having to change positions. Suppose that among airline pilots the distribution of the maximum distance (measured from the back of the seat) that can be reached without moving the seat is approximately Normally distributed with mean μ = 125cm and standard deviation σ = 10cm.

Cumulative Distribution Function

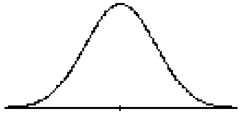
```
Normal with mean = 125.000 and standard deviation = 10.0000

       x          P(X<= x)
   95.0000        0.0013
  115.0000        0.1587
  120.0000        0.3085
  125.0000        0.5000
  135.0000        0.8413
```

Inverse Cumulative Distribution Function

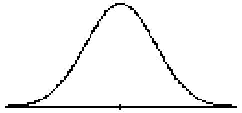
```
Normal with mean = 125.000 and standard deviation = 10.0000

  P(X <= x)         x
    0.0250      105.4004
    0.0500      108.5515
    0.9500      141.4485
    0.9750      144.5996
```

Use the STATA output above to answer the following questions.

(a) If the switch is placed 120cm from the back of the seat, what proportion of pilots will be able to reach it without moving the seat?



(b) What is the maximum distance from the back of the seat that the switch could be placed if it is required that 95% of pilots be able to reach it without moving the seat?



(c) **(i)** If the pilot has a *z*-score of 1.5, what does this mean in this context?

(ii) To what maximum reach does a *z*-score of 1.5 correspond?

Section C: Combining Random Variables

Formulae for Combining Random variables (An extract from the formulae appendix)

For any constants a and b:

$$E(aX + b) = aE(X) + b \qquad sd(aX + b) = |a|sd(X)$$

If X_1 and X_2 are independent random variables:

$$E(a_1X_1 + a_2X_2) = a_1E(X_1) + a_2E(X_2)$$

$$sd(a_1X_1 + a_2X_2) = \sqrt{a_1^2 sd(X_1)^2 + a_2^2 sd(X_2)^2}$$

If $X_1, X_2, \ldots\ldots, X_n$ is a random sample from a distribution with mean μ and standard deviation σ:

$$E(X_1 + X_2 + \ldots\ldots + X_n) = n\mu$$

$$sd(X_1 + X_2 + \ldots\ldots + X_n) = \sqrt{n}\sigma$$

1. If $Y = a_1X_1 + a_2X_2$ is written as $\boxed{Y} = \boxed{a_1} \times \boxed{X_1} + \boxed{a_2} \times \boxed{X_2}$

complete the following by filling in the gaps:

(a) $W = 3X + 2Y$ $\quad \boxed{} = \boxed{} \times \boxed{} + \boxed{} \times \boxed{}$

(b) $T = 3X - 2Y$ $\quad \boxed{} = \boxed{} \times \boxed{} + \boxed{} \times \boxed{}$

(c) $V = Y - X$ $\quad \boxed{} = \boxed{} \times \boxed{} + \boxed{} \times \boxed{}$

2. X and Y are independent random variables. X has a mean of 1 and a standard deviation of 2, and Y has a mean of 3 and a standard deviation of 3. Suppose $W = 2Y - X$. The standard deviation of W, σ_W, is:

(1) 8 **(2)** 40 **(3)** 5 **(4)** $\sqrt{40}$ **(5)** $\sqrt{8}$

3. X is a random variable with a mean of 2 and a standard deviation of 2 and Y is a random variable with a mean of 3 and a standard deviation of 4. If X and Y are independent random variables and $W = 3X - 2Y$ then the standard deviation of W, σ_W, is:

(1) 10 **(2)** $\sqrt{2}$ **(3)** $\sqrt{28}$ **(4)** 100 **(5)** $\sqrt{34}$

4. The true weight of a 40-gram packet of salt and vinegar potato chips is Normally distributed with a mean of 40.25 grams and a standard deviation of 0.099 grams. Let W be the combined weight of 50 packets of potato chips. Assuming that the packets are a random sample from the population of all such packets, then W has a mean, μ_W, and a standard deviation, σ_W, given by:

(1) $\mu_W = 40$g, $\sigma_W = 0.014$g

(2) $\mu_W = 2000$g, $\sigma_W = 0.7$g

(3) $\mu_W = 40$g, $\sigma_W = 4.95$g

(4) $\mu_W = 2012.5$g, $\sigma_W = 4.95$g

(5) $\mu_W = 2012.5$g, $\sigma_W = 0.7$g

5. The true weight of a 200-gram packet of coffee is Normally distributed with a mean of 205 grams and a standard deviation of 5 grams. Let W be the combined weight of 25 packets of coffee. Assuming that the packets are a random sample from the population of all such packets, then W has a mean, μ_W, and a standard deviation, σ_W, given by:

(1) $\mu_W = 5125$g, $\sigma_W = 25$g

(2) $\mu_W = 200$g, $\sigma_W = 1$g

(3) $\mu_W = 5125$g, $\sigma_W = 125$g

(4) $\mu_W = 5125$g, $\sigma_W = 1$g

(5) $\mu_W = 200$g, $\sigma_W = 25$g

6. A gardening business provides two services for customers – garden work and lawn mowing. From experience, the charge to the customer will vary according to the size and state of the garden or lawn. The manager of the business estimates that the charge for a gardening job is Normally distributed with a mean of \$25 and a standard deviation of \$3 while the charge for a lawn mowing job is Normally distributed with a mean of \$15 and a standard deviation of \$2. The charge for each job is independently assessed.

(a) On one particular day the business is contracted to do gardening jobs for six different customers. Let X be the total charge for six gardening jobs. X has a Normal distribution. State the value of each parameter for this distribution.

(b) On the same day the business is also contacted to do mowing jobs for eleven different customers. Let Y be the total charge for eleven mowing jobs. Y will be Normally distributed with a mean of \$165 and a standard deviation of \$6.63.

Let T be the total charge for six gardening jobs and eleven mowing jobs.

(i) Verify the mean and standard deviation of Y.

(ii) Express T in terms of the random variables X and Y.

(iii) Explain why T has a Normal distribution.

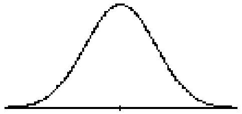
(iv) What is the value of each parameter of the distribution of T? State any assumptions required.

(c) Many customers have their lawn mown once a week. The charge is the same each time the lawn is mown. Let W be the total charge for mowing a randomly chosen lawn once a week for one year. Describe the distribution of W.

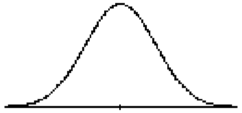
7. A university professor keeps records of his travel time while he is driving between his home and the university. Over a long period of time he has found that his morning travel times are approximately Normally distributed with a mean of 31 minutes and a standard deviation of 3 minutes. His return journey in the evening is also Normally distributed but with a mean of 35.5 minutes and a standard deviation of 3.5 minutes.

Use the STATA output on the next page to answer the following questions.

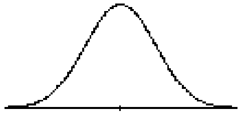
(a) Find the probability that on a typical day he spends more than one hour travelling to and from work.



(b) Find the probability that on a given day his morning journey is longer than his evening journey.



(c) On what proportion of days is the evening journey more than five minutes longer than the morning journey?



(d) Over a five-day working week, what is the distribution of the total time for:

(i) morning journeys?

(ii) evening journeys?

(iii) all journeys?

Cumulative Distribution Function

```
Normal with mean = 66.5000 and standard deviation = 4.60977

         x     P( X <= x)
    1.0000        0.0000
   30.0000        0.0000
   60.0000        0.0793
```

Cumulative Distribution Function

```
Normal with mean = 66.5000 and standard deviation = 6.50000

         x     P( X <= x)
    1.0000        0.0000
   30.0000        0.0000
   60.0000        0.1587
```

Cumulative Distribution Function

```
Normal with mean = -4.50000 and standard deviation = 4.60977

         x     P( X <= x)
  -10.0000        0.1164
   -5.0000        0.4568
    0.0000        0.8355
    5.0000        0.9803
   10.0000        0.9992
```

Cumulative Distribution Function

```
Normal with mean = -4.50000 and standard deviation = 6.50000

         x     P( X <= x)
  -10.0000        0.1987
   -5.0000        0.4693
    0.0000        0.7556
    5.0000        0.9281
   10.0000        0.9872
```

Cumulative Distribution Function

```
Normal with mean = 4.50000 and standard deviation = 4.60977

         x     P( X <= x)
  -10.0000        0.0008
   -5.0000        0.0197
    0.0000        0.1645
    5.0000        0.5432
   10.0000        0.8836
```

Cumulative Distribution Function

```
Normal with mean = 4.50000 and standard deviation = 6.50000

         x     P( X <= x)
  -10.0000        0.0128
   -5.0000        0.0719
    0.0000        0.2444
    5.0000        0.5307
   10.0000        0.8013
```