Prof. Dinov's UCLA STAT XL 10, Spring 2002

Midterm Study Guide

1. 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.

2. 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.

3. 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.

4. 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.

5. 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.

6. 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.

7. Two drugs are to be compared. A group of 20 people are each randomly allocated to one of the two drugs. Neither the people who were treated nor the doctor who administered the drugs knew who got which 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 senior 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 ($1A=10^{-10}$m).

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.

On August 05, 1997, it was 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.

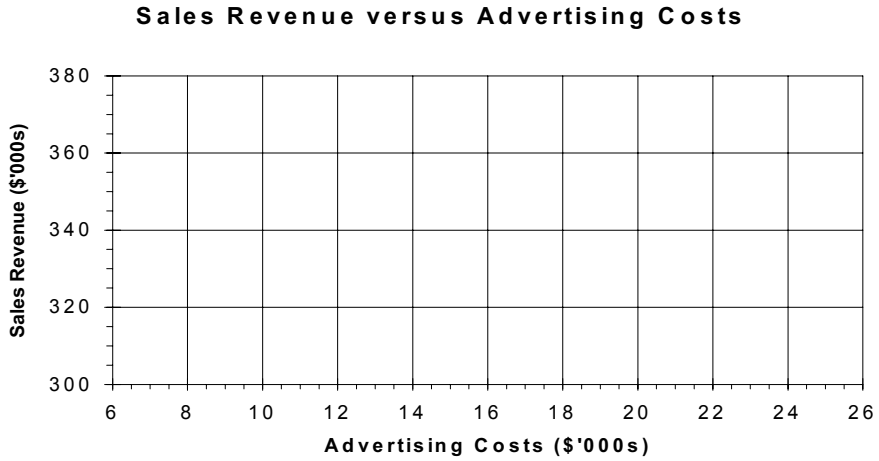
Quarter	1	2	3	4	5	6	7	8
Advertising Costs ($'000s)	10	12	8	20	11	15	10	25
Sales Revenue ($'000s)	342	347	318	350	351	346	345	367

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:

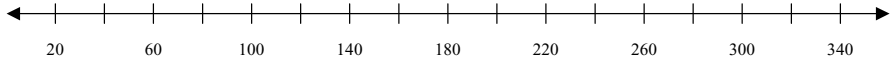
5. The following table gives the lengths (in kilometres) of the major rivers in the South Island.

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.

.

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.

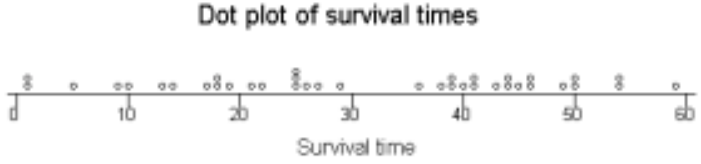
(iii) Calculate the number of months beyond which 80% of the patients survive.

(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 $\mu = 125$cm and standard deviation $\sigma = 10$cm.

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

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

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

1. If $Y = a_1X1 + a_2X_2$ 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) $W = 3X + 2Y$ $\quad \square = \square \times \square + \square \times \square$

(b) $T = 3X - 2Y$ $\quad \square = \square \times \square + \square \times \square$

(c) $V = Y - X$ $\quad \square = \square \times \square + \square \times \square$

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) $\mu_W = 40\text{g}$, $\sigma_W = 0.014\text{g}$

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

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

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

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

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) $\mu_W = 5125\text{g}$, $\sigma_W = 25\text{g}$

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

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

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

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

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.

UCLA, STAT XL 10, Prof. Dinov, Midterm review

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

UCLA Stat 10 Midterm Exam Review

Relationships between Quantitative Variables: Regression and Correlation

Section A: The Straight Line Graph

1. The equation of a line is of the form $y = \beta_0 + \beta_1 x$, where β_0 is the y-intercept and β_1 is the slope of the line. Give the values of β_0 and β_1 for the following lines.

(a) $y = 5 + 3x$

$\beta_0 =$

$\beta_1 =$

(b) $y = 10 - 14x$

$\beta_0 =$

$\beta_1 =$

2. **(a)** What is the equation of a line that has a slope of 2 and a y-intercept of –3?

(b) By how much does the y-value of this line change when

(i) x is increased by 1?

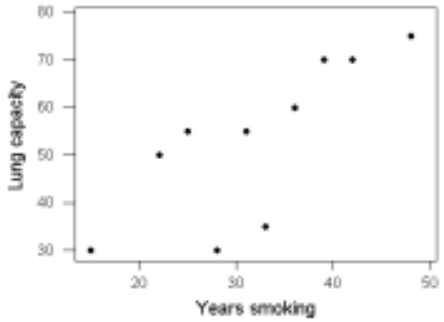
(ii) x is increased by 6?

Section B: Regression

1. Observations on lung capacity, measured on a scale of 0 – 100, and the number of years smoking were obtained from a sample of emphysema patients. One of the uses of the data is to use the number of years smoking to predict lung capacity. The data is shown in the table below. A scatter plot, residual plot, Normal probability plot and *Excel* output are also shown.

Patient	1	2	3	4	5	6	7	8	9	10
Number of years smoking	25	36	22	15	48	39	42	31	28	33
Lung capacity	55	60	50	30	75	70	70	55	30	35

Scatter plot



***Excel* regression output**

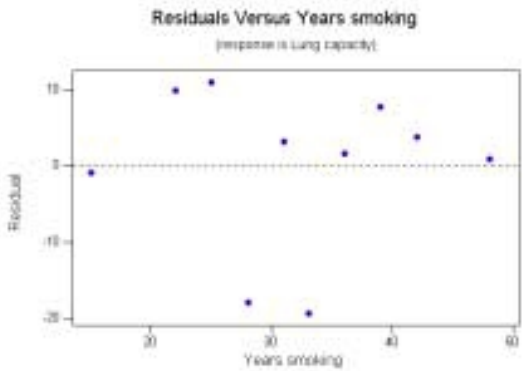
SUMMARY OUTPUT

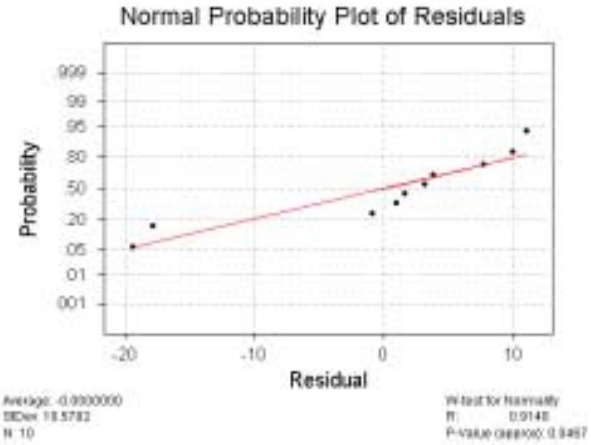
Regression Statistics	
Multiple R	0.773802257
R Square	0.598769933
Adjusted R Square	0.548616175
Standard Error	11.21989008
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1502.912533	1502.912533	11.93868522	0.008627995
Residual	8	1007.087467	125.8859334		
Total	9	2510			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	11.23788345	12.59660909	0.892135603	0.398359228	-17.80996799	40.28573489
Years smoking	1.309157259	0.37889037	3.455240255	0.008627995	0.435433934	2.182880583





(a) Write the equation of the least-squares regression line.

(b) Use the least-squares regression line to predict the lung capacity of an emphysema patient who has been smoking for 30 years.

(c) Patient 1 had smoked for 25 years and had a lung capacity of 55. Calculate the residual (prediction error) for this observation.

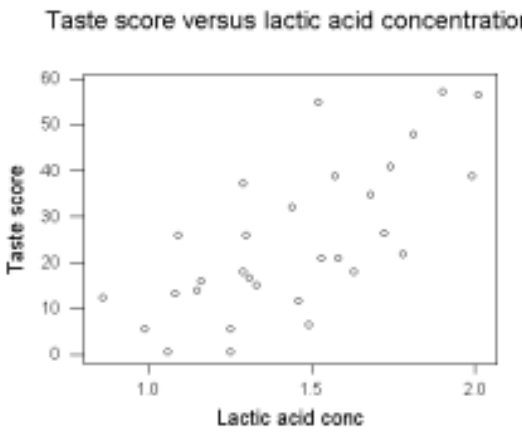
(d) Comment on the appropriateness of using a linear regression model for this data.

(e) Assume that it is appropriate to use a linear regression model for this data. (**Note:** This may not be true.) Carry out a statistical test to see if there is any evidence of an effect of years of smoking on lung capacity. State the hypotheses and interpret the test. If there is evidence of an effect then describe the size of the effect.

(f) **(i)** Find the sample correlation coefficient from the *Excel* output.

(ii) What does *Excel* call it?

2. A study of cheddar cheese from Latrobe Valley investigated the effect on the taste of cheese of various chemical processes that occur during the aging process. One of the aims of the study was to see if the lactic acid concentration could be used to predict the taste score (a subjective measure of taste). Observations were made on 30 randomly selected samples of mature cheddar cheese. A linear regression model is fitted to the data. A scatter plot, residual plot and a Normal probability plot are given below, along with a Normality test and some MINITAB output.



Regression Analysis

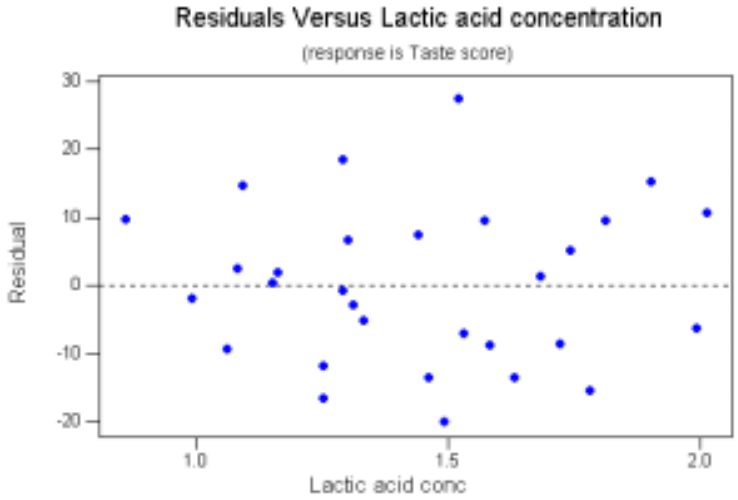
The regression equation is
Taste score = - 29.9 + 37.7 Lactic acid conc

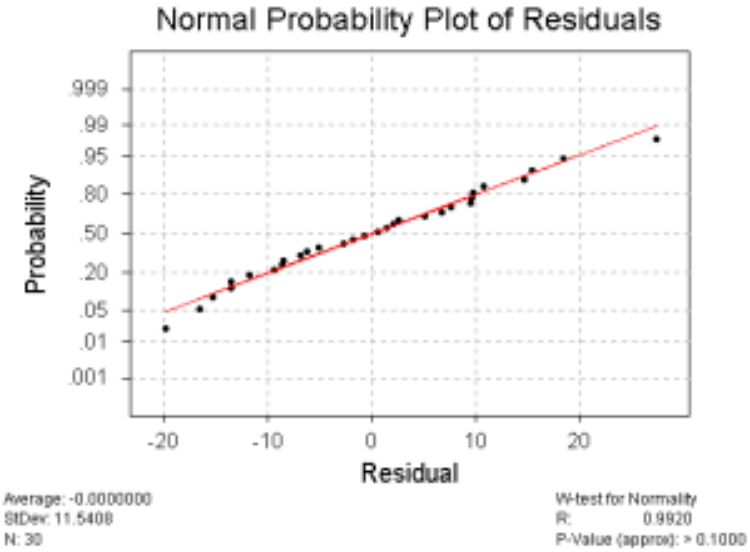
Predictor	Coef	StDev	T	P
Constant	-29.86	10.58	-2.82	0.009
Lactic a	37.720	7.186	5.25	0.000

S = 11.75 R-Sq = 49.6% R-Sq(adj) = 47.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3800.4	3800.4	27.55	0.000
Residual Error	28	3862.5	137.9		
Total	29	7662.9			





(a) One of the observations had a lactic acid concentration of 1.46 and a taste score of 11.6. Calculate the residual for this observation.

(b) Comment on the appropriateness of using a linear regression model for this data.

(c) Assume that it is appropriate to use a linear regression model for this data. (**Note:** This may not be true.) Carry out a statistical test to see if there is any evidence of an effect of lactic acid concentration on taste score. State the hypotheses and interpret the test. If there is evidence of an effect then describe the size of the effect. (Note: For a 95% confidence interval with $df = 28$, the t-multiplier is 2.048.)

(d) The researcher wanted to predict the taste score of a cheddar cheese with a lactic acid concentration of 1.8 and used MINITAB to produce the following output.

Predicted Values

Fit	StDev Fit	95.0% CI	95.0% PI
38.04	3.35	(31.18, 44.90)	(13.02, 63.05)

Use the STATA output to interpret the following:

(i) The "Fit" value of 38.04.

(ii) ???

(iii) ???

(e) The fitted least-squares regression line indicates that for each increase of 0.05 in lactic acid concentration we expect that, on average, the taste score will:

(1) increase by approximately 1.9 units.

(2) decrease by approximately 28.0 units.

(3) increase by approximately 37.7 units.

(4) increase by approximately 18.9 units.

(5) decrease by approximately 29.9 units.

(f) The fitted least-squares regression line can be used to predict taste scores for samples of mature cheddar from the Latrobe Valley. Cheese that has a lactic acid concentration of 1.30 has a predicted taste score of:

(1) 24.5

(2) 19.2

(3) 49.0

(4) 78.9

(5) 25.9

Section C.

1. Which **one** of the following statements regarding the sample correlation coefficient, r, is **false**?

(1) The value of r is an indication of the strength of linear association between the two variables.

(2) In the calculation of the value of r, it does not matter which one of the variables is designated as X and which one is designated as Y.

(3) If the sample correlation coefficient equals 1, then there is a perfect linear association between the two variables for these observations.

(4) The value of r must be between 0 and 1 inclusive.

(5) The value of r may be near 0 when there is a non-linear relationship between the two variables.

2. In the theory of inference, which **one** of the following is **not** an assumption for the linear regression model?

(1) The mean of the errors is 0 for all X-values.

(2) The errors are not independent.

(3) The standard deviation of the errors is the same for all X-values.

(4) The relationship between X and Y variables can be summarised by a straight line.

(5) The distribution of the errors is Normal for all X-values.

3. Consider using a scatter plot to investigate the relationship between a response variable Y and an explanatory variable X. The scatter plot indicates that there is a strong, negative, linear relationship between X and Y and that there are no outliers in the data. Which **one** of the following statements is **false**?

(1) The trend line explains most of the differences we see between the values of Y in the scatter plot.

(2) There are no points that are unusually far from the trend curve.

(3) Y changes, on average, by a fixed amount for each unit change in X.

(4) The value of Y tends to decrease as the value of X increases.

(5) If a new scatter plot was produced that only used a limited range of the X-values, then the relationship would look stronger.

4. Which **one** of the following statements regarding the sample correlation coefficient, r, is **false**?

(1) A value of r near 1 does not necessarily mean there is a causal relationship between the two variables.

(2) The value of r cannot be less than -1.

(3) In calculating r, it is not necessary to define one of the random variables as the response and the other as the explanatory variable.

(4) A negative value of r indicates a negative association between the two variables.

(5) A value of r equal to 0 indicates that there is no relationship between the two variables.

5. Which **one** of the following statements regarding linear regression and correlation analysis is **false**?

(1) In an analysis of the correlation between two variables, we do not single out either variable to have a special role.

(2) Using regression techniques, we can never determine whether a causal relationship exists between two variables.

(3) An outlier on a scatter plot should be removed if it is found to be an error.

(4) A strong relationship plotted for a limited range of x-values may appear weaker than it actually is.

(5) The least-squares regression technique minimises the sum of the squared prediction errors.

6. Which **one** of the following statements is **not** an assumption of the linear regression model?

(1) The relationship between the X variable and the Y variable is linear.

(2) All random errors are independent.

(3) The X-values are Normally distributed.

(4) The standard deviation of the random errors does not depend on the X-values.

(5) For any X-value, the random errors are Normally distributed (with a mean of 0).

7. Which **one** of the following statements about the sample correlation coefficient, r, between two variables X and Y is **false**?

(1) A value of r close to 1 implies a causal relationship exists between X and Y.

(2) A value of $r = 0$ does not necessarily mean that X and Y are unrelated.

(3) A value of $r = 0$ indicates that no linear relationship exists between X and Y.

(4) A value of $r = 1$ indicates that a perfect positive linear relationship exists between X and Y.

(5) A value of $r = -1$ indicates that a perfect negative linear relationship exists between X and Y.

8. Which **one** of the following statements about linear regression and correlation is **false**?

(1) A regression relationship is of the form:

observation = trend + residual scatter.

(2) In analyses of the correlation type, no variables are singled out to have a special role; all variables are treated symmetrically.

(3) Correlation coefficients provide a better means of detecting a relationship between two continuous variables than a scatter plot.

(4) The fitted trend line is often useful for prediction purposes.

(5) Lines fitted to data using the least-squares method do not allow us to reliably predict the behaviour of Y outside the range of x-values for which we have collected data.

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

(1) The two main components of a regression relationship are 'trend' and 'scatter'.

(2) The larger the amount of scatter, the smaller the size of the absolute value of the correlation coefficient, r.

(3) A correlation coefficient of $r = 0$ means that there is no linear relationship between the two variables, whereas a negative correlation coefficient indicates an association, the strength of which depends on its absolute value.

(4) A small value of the absolute value of the correlation coefficient, r, indicates a weak linear relationship.

(5) In the interpretation of a correlation coefficient, r, one variable is always treated as the response variable and the other as the explanatory variable.

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

(1) A linear regression model should never be used without first examining the appropriate scatter plot.

(2) Outliers in the values of the explanatory variable can have a big influence on the fitted regression line.

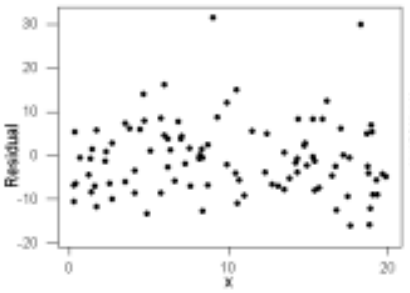
(3) The residuals are computed to be $x_i - \hat{y}_i$.

(4) If the assumption of constant error standard deviation is valid, we would expect to see a patternless horizontal band in a plot of the residuals versus the explanatory variable.

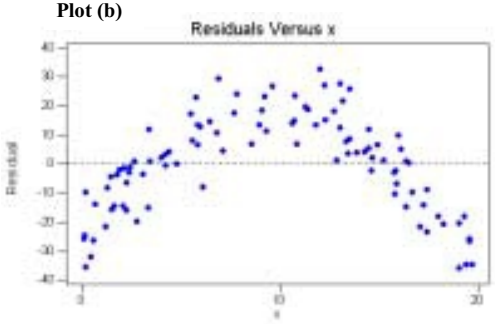
(5) We can investigate the distribution of the errors by looking at a stem-and-leaf plot of a hjistogram of the residuals.

Questions 11 and 12 refer to the following set of residual plots.

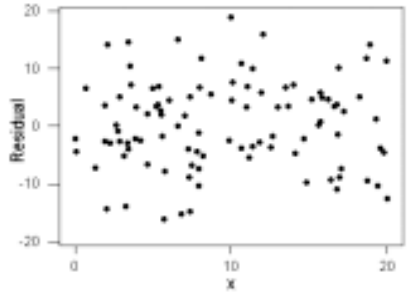
Plot (a)



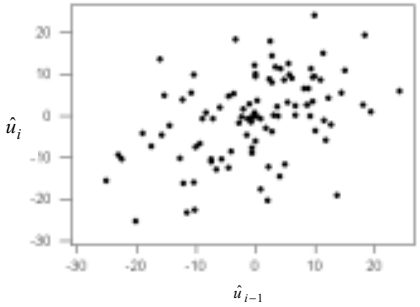
Plot (b)



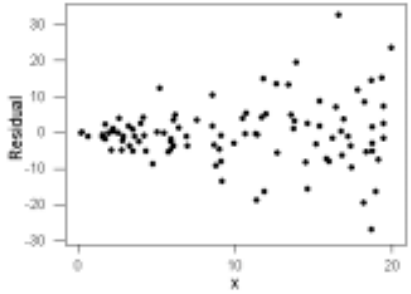
Plot (c)



Plot (d)



Plot (e)



11. Which **one** of the plots does **not** indicate problems with the assumptions underlying the linear regression model?

(1) (a)
(2) (b)
(3) (c)
(4) (d)
(5) (e)

12. Which **one** of the plots indicates that the variability of the error term is **not** independent of x?

(1) (a)
(2) (b)
(3) (c)
(4) (d)
(5) (e)