UCLA STAT 10 Statistical Reasoning - Midterm Review Solutions

Observational Studies, Designed Experiments & Surveys

.

.

1. **(i)** The treatment being compared is:

Study 1: the number of storeys from which the cat fell.

Study 2: the gender of the student.

Study 3: the style of the commercial.

(ii) Study 1: An observational study. There is no allocation (by the researcher) of cats to the number of storeys of the fall. Results are simply observed for cases that happen.

Study 2: An observational study. There is no allocation (by the researcher) of subjects (the students) to the groups (male or female).

Study 3: An experiment. The researcher allocates which commercial is to be watched by each subject (shopper).

(iii) It is not possible to do an experiment for study 1 due to ethical and moral considerations. To do an experiment a sample of cats would have to be allocated a height and then thrown out of a window at that height.

It is not possible to do an experiment for study 2 as the researcher cannot allocate a gender to a student.

2. (5)

3. (3)

4. (4)

5. Self-selection, selection bias.

6. **(i)** False. A sample of 500 is large enough to give a useful indicative result.

(ii) True. Homeless people cannot be contacted by telephone.

(iii) True. Telephone polls will usually have some non-response bias. There is no indication of how vigorously non-respondents were followed up.

(iv) False. Control groups are not necessary in polls or surveys. Control groups are often used when comparisons want to be made (ie, in observational studies and experiments).

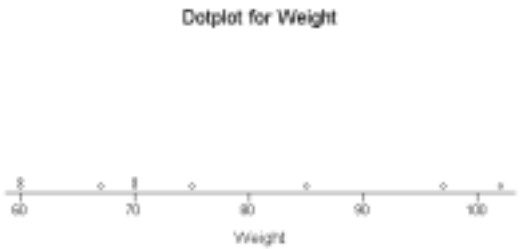
7. (2)

Dinov's STAT 10 Intro to Statistical Reasoning - Midterm Exam Review

.

Graphs, Numerical Summaries, Histograms

1. **(a)**



(b) The centre of these data is about 75kg.
The data is skewed to the right (positively skewed).
There are gaps at weights at around 80kg and 90kg.

2. **(a)** Units: 9 | 5 = 950Å

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

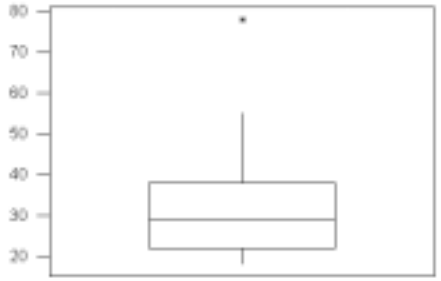
(b) (5) **(c)** (5) **(d)** (2)

3. (3)

4. (2)

5. **(1)** Agree
(2) Agree
(3) Disagree
(4) Disagree

6.



7. **(1)** Disagree, the skew is only moderate.
(2) Disagree, the interquartile range is 16.
(3) Agree.
(4) Agree.
(5) Disagree.

8. (1)

9. (2)

10. (3)

11. (3)

.

Exploratory Tools for Relationships

Section A: Types of Variables

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

(b) **Qualitative** variables describe **group membership**.

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

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

3. (a) Variables **with order** are called **ordinal**.

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

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

(b) To explore relationships between a **qualitative** variable and a **quantitative** variable we use **dot** plots, **stem-and-leaf** plots and **box** plots.

(c) To explore the relationship between two **qualitative** variables we use a **two-way table** of **counts**.

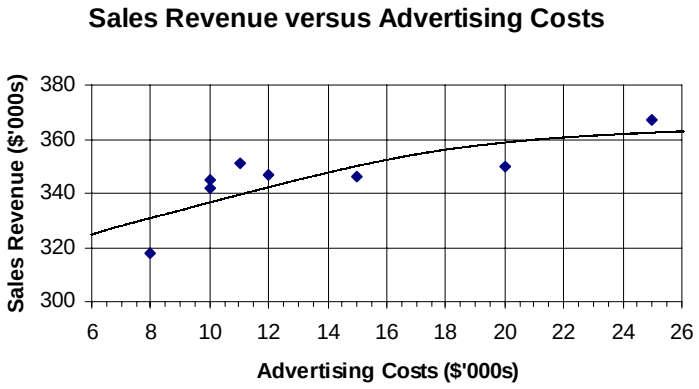
Section B: Two Variables

1. (1)

2. (3)

3. (5)

4.

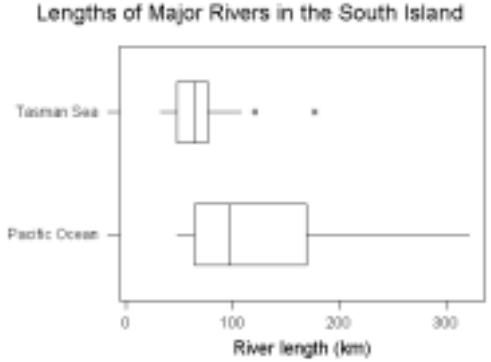


Interpretation:
As advertising costs increase, sales revenue also increases.
The relationship is not linear – the increase in sales revenue decreases as advertising costs increase.
There are no outliers in this data.
The amount of scatter about the trend curve is small.

5. (a) **Pacific Ocean rivers:**
Med = 97
$Q_1 = 64$
$Q_3 = 169$

Tasman Sea rivers:
Med = 64
$Q_1 = 48$
$Q_3 = 76$

(b)



Note: For the rivers flowing into the Pacific Ocean:
IQR = 105, 1.5 x IQR = 157.5, Q_3 + 1.5 x IQR = 326.5, Q_1 – 1.5 x IQR = -93.5
There are no outside values, the whiskers end at 48 (lower) and 322 (upper).
For the rivers flowing into the Tasman Sea:
IQR = 28, 1.5 x IQR = 42, Q_3 + 1.5 x IQR = 118, Q_1 – 1.5 x IQR = 6
177 and 121 are outside values, the whiskers end at 32 (lower) and 108 (upper).

(c) On average, the rivers flowing into the Pacific Ocean are longer.

The lengths of the rivers flowing into the Pacific Ocean have a larger spread than the lengths of the rivers flowing into the Tasman Sea.

The lengths of the rivers flowing into the Pacific Ocean are skewed to the right (positively skewed).

The Grey River and Buller River are outliers amongst the rivers flowing into the Tasman Sea.

.

Probabilities and Proportions

1. (a)

	Females	Males	**Total**
Stat10	0.3003	0.2949	0.5952
Stat11	0.1924	0.2123	0.4048
Total	0.4928	0.5072	1

(b) **(i)** $\frac{593}{2011} = 0.2949$ **(ii)** $\frac{991}{2011} = 0.4928$ **(iii)** $\frac{387}{2011} = 0.1924$

(c) $\frac{427}{814} = 0.5246$

(d) $\frac{427}{1020} = 0.4186$

2. (a)

	40 or under	Over 40	**Total**
Wearing a seat belt	0.484	**0.369**	0.853
Not wearing seat belt	**0.066**	0.081	**0.147**
Total	**0.55**	**0.45**	1

(b) 0.147

(c) $\frac{0.081}{0.147} = 0.551$

(d) 0.55

3. (a)

	Under 40	40 or over	Total
Mild cases	16	20	36
Serious cases	15	35	50
Total	31	55	86

(b) **(i)** $\frac{16}{86} = 0.1860$ **(ii)** $\frac{20+35+15}{86} = 0.8140$ **(iii)** $\frac{35}{86} = 0.4070$

(c) $\frac{15}{50} = 0.3$

(d) $\frac{20}{55} = 0.3636$

4. (a)

	High-risk	Low-risk	Total
In default	40% of 0.05 = 0.02	0.03	0.05
Not in default	0.2185	77% of 0.95 = 0.7315	0.95
Total	0.2385	0.7615	1

(b) 23.85%

(c) $\frac{0.02}{0.2385} = 0.0839$

5. (a)

	Female	Male	Total
Business degree	27% of 0.175 = 0.04725	0.12775	0.175
Other degree	0.42281	48.75% of 0.825 = 0.40219	0.825
Total	0.47006	0.52994	1

(b) 0.52994

(c) 0.12775

(d) $\frac{0.04725}{0.47006} = 0.1005$

6. (2)

Discrete Random Variables

Section A: Discrete Random Variables

1. (a)

x	**0**	**1**	**2**
$\text{pr}(X = x)$	**0.25**	**0.5**	**0.25**

(b) **(i)** $\text{pr}(X > 1) = 0.25$ **(ii)** $\text{pr}(X \geq 1) = 0.75$ **(iii)** $\text{pr}(X \leq 2) = 1$

2. (a) $\text{pr}(Y > 12) = 0.03 + 0.28 + 0.12 = 0.43$

(b) $\text{pr}(Y \leq 10) = 0.10 + 0.07 + 0.25 = 0.42$

(c) $\text{pr}(Y \geq 6) = 1 - \text{pr}(Y < 6) = 1 - 0.10 = 0.9$

(d) $\text{pr}(6 \leq Y \leq 12) = 0.07 + 0.25 + 0.15 = 0.47$

(e) $\text{pr}(10 \leq Y \leq 12) = 0.25 + 0.15 = 0.4$

(f) $\text{pr}(13 < Y < 25) = 0.28$

Section B: Binomial Distribution

1. Let X be the number of customers who purchase at least one book.
$X \sim \text{Binomial}\ (n = 7, p = 0.3)$
$\text{pr}(X \geq 2) = 1 - \text{pr}(X \leq 1) = 1 - 0.329 = 0.671$

2. (a) $n = 10, p = 0.05$

(b) There is a fixed number of trials, 10. Each disk drive is a trial.
Each trial has 2 outcomes: Disk drive malfunctions or disk drive does not malfunction.
The disk drives are independent.
The probability that a disk drive malfunctions is constant.

(c) The first two assumptions will be satisfied.
The disk drives may not be independent. Disk drives could be made from the same batch of materials or may have the same systematic fault.
The probability of a disk drive malfunctioning will not be constant because it will depend on how a disk drive is used.

(d) **(i)** $\text{pr}(X = 0) = 0.599$
(ii) $\text{pr}(X = 1) = 0.315$
(iii) $\text{pr}(X \geq 2) = 1 - \text{pr}(X \leq 1) = 1 - 0.914 = 0.086$
(iv) $\text{pr}(2 \leq X \leq 5) = \text{pr}(X \leq 5) - \text{pr}(X \leq 1) = 1.00 - 0.914 = 0.086$

Continuous Random Variables

Section A: Probability Density Function Quiz

1. Areas under the density curve represent probabilities. The probability that a random observation falls between a and b is equal to the area between the density curve and the x-axis from $x = a$ and $x = b$.

2. The total area under the curve equals 1.

3. The population mean μ is the point where the density curve balances.

4. No, because for a continuous random variable:

$\text{pr}(a \le X \le b) = \text{pr}(a < X \le b) = \text{pr}(a \le X < b) = \text{pr}(a < X < b)$ = area under the curve between a and b.

5. The curve is bell-shaped, symmetrical and centred at μ. The standard deviation σ governs the spread.

6. The parameters are μ and σ.

Section B: Normal Distribution

1. **(a)** $\text{pr}(X < 245) = 0.0947$

(b) $\text{pr}(255 < X < 280) = \text{pr}(X < 280) - \text{pr}(X < 255) = 0.8092 - 0.2459 = 0.5633$

(c) $\text{pr}(X > 287) = 1 - \text{pr}(X < 287) = 1 - 0.9053 = 0.0947$

2. Let X be the survival time in months of a cancer patient on this drug.

(a) **(i)** $\text{pr}(X \le 12) = 0.1163$

(ii) $\text{pr}(12 < X < 24) = \text{pr}(X < 24) - \text{pr}(X < 12) = 0.3286 - 0.1163 = 0.2123$

(iii) $\text{pr}(X > x) = 0.8$ therefore $\text{pr}(X < x) = 0.2$ and so $x = 17.6341$.

80% of the patients live beyond 17.6 months.

(iv) $\text{pr}(a < X < b) = 0.8$

$\text{pr}(X < a) = 0.1$ and so $a = 10.5932$

$\text{pr}(X < b) = 0.9$ and so $b = 51.6048$

The range of the central 80% of survival times is from 10.6 to 51.6 months.

(b) There are some doubts about the validity of the assumption that survival times are Normally distributed. Although the data is roughly symmetrical, there is a gap in the centre which could indicate bimodality of survival times. The tails seem too short for the underlying distribution to have a Normal distribution.

3. Let X be the maximum distance reached by a pilot without moving the seat.

(a) $\text{pr}(X \ge 120) = 1 - \text{pr}(X \le 120) = 1 - 0.3085 = = 0.6915$

(b) $\text{pr}(X \ge x) = 0.95$ therefore $\text{pr}(X < x) = 0.05$ and so $x = 108.5515$.

The maximum distance at which the switch should be placed is 109cm.

(c) **(i)** That this pilot's maximum reach is 1.5 standard deviations above the mean.

(ii) $x = 125 + 1.5 \times 10 = 140$cm. A z-score of 1.5 corresponds to a maximum reach of 140cm.

Section C: Combining Random Variables

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

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

(c) $\boxed{V} = \boxed{1} \times \boxed{Y} + \boxed{-1} \times \boxed{X}$

2. (4)

3. (1)

4. (5)

5. (1)

6. **(a)** Let G_i be the charge for a randomly chosen gardening job. $G_i \sim \text{Normal}(\mu = 25, \sigma = 3)$

$X = G_1 + G_2 + G_3 + G_4 + G_5 + G_6$

$\text{E}(X) = \text{E}(G_1 + G_2 + G_3 + G_4 + G_5 + G_6) = 6 \times \text{E}(G_i) = 6 \times 25 = 150$

$\text{sd}(X) = \text{sd}(G_1 + G_2 + G_3 + G_4 + G_5 + G_6) = \sqrt{6} \times \text{sd}(G_i) = \sqrt{6} \times 3 = 7.35$

(b) **(i)** Let M_i be the charge for a randomly chosen mowing job. $M_i \sim \text{Normal}(\mu = 15, \sigma = 2)$

$Y = M_1 + M_2 + \ldots + M_{11}$

$\text{E}(Y) = \text{E}(M_1 + M_2 + \ldots + M_{11}) = 11 \times \text{E}(M_i) = 11 \times 15 = 165$

$\text{sd}(Y) = \text{sd}(M_1 + M_2 + \ldots + M_{11}) = \sqrt{11} \times \text{sd}(M_i) = \sqrt{11} \times 2 = 6.63$

(ii) $T = X + Y$

(iii) T has a Normal distribution because it is a combination of Normally distributed random variables.

(iv) $\text{E}(T) = \text{E}(X + Y) = \text{E}(X) + \text{E}(Y) = 150 + 165 = 315$

$\text{sd}(T) = \text{sd}(X + Y) = \sqrt{\text{sd}(X)^2 + \text{sd}(Y)^2} = \sqrt{7.35^2 + 6.63^2} = 9.90$

In order to calculate the standard deviation of T we had to assume that X and Y are independent random variables.

(c) Let M be the charge for a randomly chosen mowing job. $M \sim \text{Normal}(\mu = 15, \sigma = 2)$

$W = 52M$

$\text{E}(W) = 52\text{E}(M) = 52 \times 15 = 780$

$\text{sd}(W) = |52|\text{sd}(M) = 52 \times 2 = 104$

$W \sim \text{Normal}(\mu = \$780, \sigma = \$104)$

7. Let M be the morning travel time and N be the evening travel time.

(a) Want pr($M + N > 60$)

E($M + N$) = E(M) + E(N) = 31 + 35.5 = 66.5 minutes

sd($M + N$) = $\sqrt{\text{sd}(M)^2 + \text{sd}(N)^2} = \sqrt{3^2 + 3.5^2}$ = 4.60977, assuming M and N are independent.

$M + N$ ~ approx. Normal (μ = 66.5, σ = 4.60977)

pr($M + N > 60$) = 1 – pr($M + N \leq 60$) = 1 – 0.0793 = 0.9207

(b) Want pr($M > N$) = pr($M – N > 0$)

E($M – N$) = E(M) – E(N) = 31 – 35.5 = –4.5 minutes

sd($M – N$) = $\sqrt{\text{sd}(M)^2 + \text{sd}(N)^2} = \sqrt{3^2 + 3.5^2}$ = 4.60977, assuming M and N are independent.

$M – N$ ~ approx. Normal (μ = –4.5, σ = 4.60977)

pr($M – N > 0$) = 1 – pr($M – N < 0$) = 1 – 0.8355 = 0.1645

(c) Want pr($N – M > 5$)

E($N – M$) = E(N) – E(M) = 35.5 – 31 = 4.5 minutes

sd($N – M$) = $\sqrt{\text{sd}(N)^2 + \text{sd}(M)^2} = \sqrt{3.5^2 + 3^2}$ = 4.60977, assuming N and M are independent.

$N – M$ ~ approx. Normal (μ = 4.5, σ = 4.60977)

pr($N – M > 5$) = 1 – pr($N – M < 5$) = 1 – 0.5432 = 0.4568

(d) **(i)** Let $T_M = M_1 + M_2 + M_3 + M_4 + M_5$, where M_i ~ approx. Normal (μ = 31, σ = 3)

E(T_M) = E($M_1 + M_2 + M_3 + M_4 + M_5$) = 5E($M$) = 5 x 31 = 155 minutes

sd(T_M) = sd($M_1 + M_2 + M_3 + M_4 + M_5$) = $\sqrt{5} \times \text{sd}(M) = \sqrt{5} \times 3$ = 6.7082 minutes, assuming independence of the morning travel times.

T_M ~ approx. Normal (μ = 155min, σ = 6.7082min)

(ii) Let $T_N = N_1 + N_2 + N_3 + N_4 + N_5$, where N_i ~ approx. Normal (μ = 35.5, σ = 3.5)

E(T_N) = E($N_1 + N_2 + N_3 + N_4 + N_5$) = 5E($N$) = 5 x 35.5 = 177.5 minutes

sd(T_N) = sd($N_1 + N_2 + N_3 + N_4 + N_5$) = $\sqrt{5} \times \text{sd}(N) = \sqrt{5} \times 3.5$ = 7.8262 minutes, assuming independence of the morning travel times.

T_N ~ approx. Normal (μ = 177.5min, σ = 7.8262min)

(iii) Let $T = T_M + T_N$

E(T) = E($T_M + T_N$) = E(T_M) + E(T_N) = 155 + 177.5 = 332.5 minutes

sd(T) = sd($T_M + T_N$) = $\sqrt{\text{sd}(T_M)^2 + \text{sd}(T_N)^2} = \sqrt{6.7082^2 + 7.8262^2}$ = 10.3078 minutes, assuming independence of ten travel times in the week.

UCLA STAT 10 Midterm Exam Review Solutions
Relationships between Quantitative Variables: Regression and Correlation

Section A: The Straight Line Graph

1. **(a)** $\beta_0 = 5,\ \beta_1 = 3$

(b) $\beta_0 = 10,\ \beta_1 = -14$

2. **(a)** $y = -3 + 2x$

(b) **(i)** 2

(ii) 12

Section B: Regression

1. **(a)** $\hat{y} = 11.238 + 1.309x$

(b) Predicted lung capacity = 11.238 + 1.309 x 30 = 50.5

(c) Predicted lung capacity = 11.238 + 1.309 x 25 = 44.0
Residual = Observed value – predicted value = 55 – 44.0 = 11

(d) **'**Years smoking' is used to predict lung capacity.
'Years smoking' is a quantitative variable and 'Lung capacity' is continuous and random.
There is a possible linear trend but the observations (28, 30) and (33, 35) are possible outliers which cause concern with the appropriateness of the model.
The residuals versus 'Years smoking' plot along with the *P-value* for the *W*-test for Normality indicates some concern with the assumption that the errors are Normally distributed.

(e) H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$
P-value = 0.0086
There is strong evidence of a linear relationship between years of smoking and lung capacity.
With 95% confidence, we estimate that for every additional year of smoking an emphysema patient's lung capacity increases by between 0.44 and 2.18 units.

(f) **(i)** $r = 0.774$

(ii) *Excel* calls it Multiple R.

2. **(a)** For $x = 1.46$, $\hat{y}$ = -29.86 + 37.72 x 1.46 = 25.2
Residual = Observed value – predicted value = 11.6 – 25.2 = –13.6

(b) The lactic acid concentration is used to predict the taste score.
The lactic acid concentration is quantitative, and the taste score is continuous and random.
The scatter plot shows a linear trend with scatter about that trend.
From the plot of residuals versus lactic acid concentration there is no concern with the assumption that the errors are Normally distributed with mean 0 and with the same standard deviation for each value of *X*.

(c) H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

P-value = 0.000

There is strong evidence of a linear relationship between lactic acid concentration and taste score.

95% confidence interval for β_1 is:

$37.720 \pm 2.048 \times 7.186 = (23.0, 52.4)$

With 95% confidence, we estimate that for every increase of one unit in the lactic acid concentration the taste score increases by between 23.0 and 52.4 units.

(d) **(i)** We predict that, on average, cheddar cheese with a lactic acid concentration of 1.8 will have a taste score of 38.04.

(ii) With 95% confidence, we estimate that the mean taste score for cheddar cheese with a lactic acid concentration of 1.8 will be somewhere between 31.2 and 44.9.

(iii) With 95% confidence, we predict that the next piece of cheddar cheese with a lactic acid concentration of 1.8 will be somewhere between 13.0 and 63.1.

(e) Estimated slope = 37.72

Estimated increase in taste score for a 1 unit change in lactic acid concentration is 37.72.

Estimated increase in taste score for a 0.05 unit change in lactic acid concentration is $0.05 \times 37.72 = 1.886$.

(1) is the correct response.

(f) (2)

Section C.

1. (4)

2. (2)

3. (5)

4. (5)

5. (2)

6. (3)

7. (1)

8. (3)

9. (5)

10. (3)

11. (3)

12. (5)