UNIVERSITY OF MANITOBA

DATE: February 9, 2012 MIDTERM EXAMINATION
TIME: 1:00 PM–2:15 PM DURATION: 1.25 hours
EXAMINATION: Statistics for Engineers PAGE: 1 of 10
COURSE: STAT 2220 EXAMINER: A. Forbes

SOLUTIONS

[Do not distribute.]

Multiple-Choice Questions

1. (1 point) Consider the following back-to-back stemplot for male and female scores with 20 and 18 observations respectively.

Male $n = 20$		*Female* $n = 18$
50	7	
8	8	
21	9	
984	10	139
5543	11	5
6	12	669
2	13	77
60	14	08
1	15	244
9	16	55
	17	8
70	18	
	19	
	20	0

Which of the following statements is true?

(A) The median of the males scores is larger than that of the female scores.
(B) One of the male scores is an outlier.
(C) One of the female scores is an outlier.
(D) There is more variability in the female scores than there is the the male scores.
(E) None of the above are true.

Solution:

calculation	Male	Female
Median	114.5	138.5
Q1, Q3	98, 143	126, 154
IQR	45	28
Range	117	99
UF,LF	200.5, 30.5	196, 84

2. (1 point) Which of the following sets of 4 numbers has the largest standard deviation?

(A) 7,8,9,10 (B) 5,5,5,5 (C) 0,1,2,3 **(D) 0,0,10,10** (E) 1,2,3,4

Solution:

- note A, C, E have increments of 1
- B has no variation

---

3. (1 point) Which of the answers below is the five-number summary for the following sample of 17 pulse readings:

$$Data: \ 51, 80, 80, 78, 88, 81, 68, 60, 75, 72, 72, 68, 86, 67, 64, 88, 62$$

$$Sorted\ Data: \ 51, 60, 62, 64, \quad 67, 68, 68, 72, \quad 72, \quad 75, 78, 80, 80, \quad 81, 86, 88, 88$$

(A)	**51**	**65.5**	**72.0**	**80.5**	**88**
(B)	51	64.0	72.0	80.0	88
(C)	51	67.0	73.5	80.0	88
(D)	51	65.5	72.0	80.0	88
(E)	51	64.0	72.0	80.5	88

4. (1 point) Which of the following statements about the least squares regression line is (are) true?

I The slope of the least squares regression line always has the same sign as the coefficient of correlation.

II The least squares regression line is the line that minimizes the sum of residuals.

III The least squares regression line is the line that maximizes the value of the correlation coefficient.

IV The least squares regression line is the line that minimizes the sum of squared residuals.

(A) I only (B) II only (C) I and II (D) III and IV **(E) I and IV**

5. (1 point) When calculating statistics for a dataset, you must be concerned about outliers. **How many** of the following statements are true?

I An outlier in the y-direction will always affect the slope of the regression line.

II An outlier in the x-direction will never affect the slope of the regression line.

III The mean, variance and standard deviation are sensitive to outliers.

IV The median and interquartile range are not sensitive to outliers.

V The coefficient of correlation is sensitive to all outliers.

(A) All (B) None (C) Two **(D) Three** (E) Only one

Solution: Statements III, IV , V are correct

6. (1 point) A simple random sample of size n is the **only type** of sample design that guarantees that

(A) every individual in the population has a known chance of being selected.

(B) every individual in the population has an equal chance of being selected.

(C) every group of n individuals has an equal chance of being selected.

(D) results will not be biased.

(E) all of the above.

7. (1 point) At a local health club, a researcher sampled 75 people whose primary exercise was cardiovascular and 75 people whose primary exercise was strength training. The researcher's objective was to assess the effect of type of exercise on cholesterol levels. Each subject reported to a clinic to have his or her cholesterol measured. The subjects were aware of the purpose of the study, but the technician measuring the cholesterol was not aware of the subjects type of exercise. This is an example of:

(A) a matched pairs design.

(B) an observational study.

(C) a randomized block design.

(D) a completely randomized design.

(E) a double-blind experiment.

Solution: Note that club members were sampled rather than assigned to an exercise

8. (1 point) In conducting an experiment, randomization is used to:

(A) select a random sample from the population to ensure that every unit has the same chance of being in the sample.

(B) create blocks in a block design and hence reduce the variability caused by extraneous factors.

(C) eliminate bias when allocating treatments to experimental units and to control the effects of possible confounding factors.

(D) draw a simple random sample for the treatments in order to eliminate any bias in deciding which treatments to include in the study.

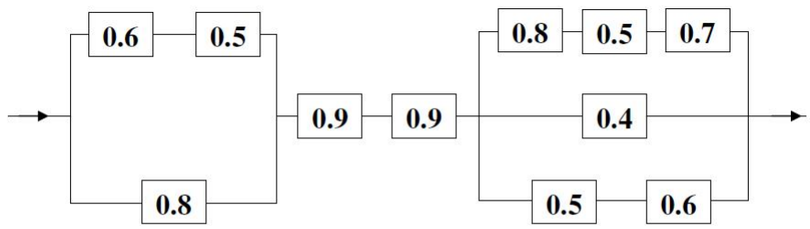
(E) eliminate variability when allocating treatments to experimental units so as to make it easier to see the effect of lurking variables.

Solution: Numerous students selected A, which pertains to a sample rather than an experiment. While you could randomly sample or select the experimental units, quite often only volunteers are available. The key for the experiment is to randomly assign the treatments to the experimental units.

9. (1 point) Consider the following system, with series and parallel components, where the value in each box represents the reliability of that component.

What is the reliability of this system?

(A) 0.4277 (B) 0.3891 (C) 0.4580 **(D) 0.4859** (E) 0.3568



10. (1 point) Consider two independent events A and B. **How many** of the following statements are true?

I $P(A|B) = P(B|A)$.

II Events A and B are mutually exclusive.

III $P(A \cup B) = P(A) + P(B)$.

IV $P(B \cap A) = P(B)P(A)$.

V If event A occurs, then even B also occurs.

(A) All (B) None (C) Two (D) Three **(E) Only one**

Solution:

- Only statement IV is correct
- for independence, $P(A|B) = P(A)$ or $P(B|A) = P(B)$ or $P(A \cap B) = P(A)P(B)$ but not $P(A|B) = P(B|A)$
- II and III are the same, which is different than independence
- V is $B \subset A$

Long-answer Questions

11. This question deals with linear regression and the plots shown below. Calculated values for the dataset include $\bar{x} = 46.67$, $\bar{y} = 384.17$, $s_x = 32.29$ and $s_y = 121.32$. The regression of X on Y indicates that 67.25% of the variation in Y can be explained by X.

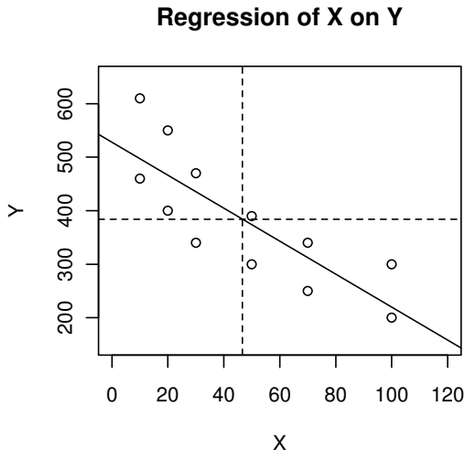
(a) (2 points) Calculate the values for the intercept and slope of the regression line

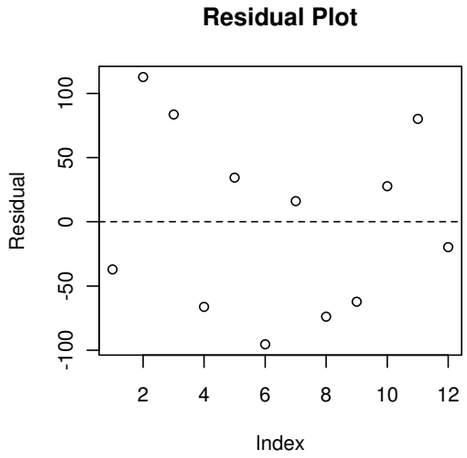
Solution: First of all, $r < 0$ from the plot and $r^2 = 67.25\% = 0.6725$

$$b_1 = r * \frac{s_y}{s_x} = -\sqrt{.6725} * \frac{121.32}{32.29} = -3.081$$

$$b_0 = \bar{y} - b_1 * \bar{x} = 384.17 - (-3.081) * 46.67 = 527.96$$

Note: you can validate the b_0 calculation on the above Regression plot and can also estimate the slope from the plot. Two more reasons to always look a plots before calculating anything.





(b) (2 points) Consider three values of X for which you want to make predictions: $x_1 = 20$, $x_2 = 60$ and $x_3 = 120$. Which of these X values would result in the most reliable prediction and which would result in the least reliable prediction? Explain your reasoning in both cases (please be brief). **No Calculations Required!**

Solution:

- standard error for prediction at x_0 is $\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\Sigma_{i=1}^{n}(x_i-\bar{x})^2}\right)}$
- Least Reliable: x_3 because it is beyond the range of the data; it would require extrapolation (and it is furthest from $\bar{x}$); note that predicting for an x value of 120 does not make it an outlier
- Most Reliable: x_2 because it is closer than x_1 to $\bar{x}$
- the prediction error does not depend on how many points there are around the specific x value nor does it matter if there are actual response values for the specific x value

(c) (2 points) In the dataset, one individual point has an X value of 70 and a Y value of 340. Calculate the residual value for this point and interpret its sign.

Solution:

$$residual = actual - predicted$$

$$e_{70} = y_{70} - \hat{y}_{70} = y_{70} - (b_0 + b_1 x) = 340 - (527.96 - 3.081 * 70) = 27.71$$

Since the residual value is positive, the actual point is above the regression line. Note: you can also validate the calculation on the above Residual plot for index $= 10^{th}$ value (9^{th} and 10^{th} correspond to the two points at $x = 70$).

(d) (1 point) Comment on the strength, form and direction of the relationship between X and Y and indicate if you think outliers are present. Please be brief with your comments - point form is acceptable.

Solution:

- strength: with an r value of 0.82, the association is strong
- direction: negative association since the slope is negative

- form: non-linear relationship as shown, especially on the residual plot (not many students noticed this)
- outliers: none present (although if you didn't notice the non-linearity, you might think there were a few y-outliers)

12. Primers are used on aluminum parts to **improve paint adhesion**. The primer is applied by **spraying** or by **dipping**. The process engineers perform a **factorial experiment** with primers from three different **manufacturers** (denoted **A, B and C**), and the two **application methods**. There were a total of 24 experimental units evenly allocated (**randomly assigned)** over the treatments. After the paint had dried, the **adhesion force** was measured. For this experiment, answer the following questions:

(a) (½ point) What is the purpose of the experiment?

Solution: To determine which primer in combination with which application method results in the best paint adhesion or to improve paint adhesion or to see the effect on paint adhesion due to ...

(b) (½ point) What is the response?

Solution: force of adhesion

(c) (½ point) What are the factors, and the associated levels?

Solution:

- application method has 2 levels: spray and dip
- primer manufacturer has 3 levels: A, B, C

(d) (½ point) How many replicates are there? How many treatments?

Solution:

- 6 treatments since one factor has 2 levels and the other has 3
- 24 experimental units over 6 treatments gives 4 replicates per treatment

(e) (1 point) If one of the treatments resulted in significantly more adhesion, can you conclude the treatment was likely the cause? Why or why not?

Solution: Since this is an experiment (with treatments randomly assigned) rather than an observational study, the variation in the response can be attributed to the treatments.

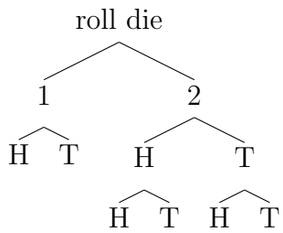
13. (1 point) What are the four principles of experimental design?

Solution:

1. randomization
2. replication (repetition)
3. control (comparison, blocking)
4. factorial experimentation

14. You have a fair coin (H or T) and a fair die with five of the faces having a value of 1 and the sixth face having a value of 2 (i.e. the face values are 1,1,1,1,1,2). You roll the die and then flip the coin the number of times based on the face value of the die - if you roll a 1, you flip the coin once but if you roll a 2 you flip the coin twice. The outcomes of interest are the combinations of the die face value and the H or T from the two coin flips.

(a) (1 point) What is the sample space (i.e. the set of all possible outcomes)?



Solution: Note that the sample space is not just a diagram or a list, but is a **set** of **unique** outcomes

$$S = \{2HH, 2HT, 2TH, 2TT, 1H, 1T\}$$

Also note that all outcomes in the sample space do not have the same probability

$$P(1H) = P(1T) = \frac{5}{6} \cdot \frac{1}{2}$$

$$P(2HH) = P(2TT) = P(2TH) = P(2HT) = \frac{1}{6} \cdot \frac{1}{4}$$

(b) (1 point) What is the probability of getting two tails?

Solution: Needed to show work for full marks

$$P(\textit{two tails}) = P(\{TT\}|\{\textit{rolling a } 2\})P(\{2\}) = \frac{1}{4} \cdot \frac{1}{6} = \frac{1}{24} = 0.04167$$

(c) (2 points) If you have exactly one head from flipping the coin (either once or twice), what is the probability that you rolled a 2?

Solution: Needed to show work for full marks

$$\begin{aligned} P(\{2\}|\{\textit{exactly one head}\}) &= \frac{P(\{2\}\ AND\ \{\textit{exactly one head}\})}{P(\{\textit{exactly one head}\})} \\ &= \frac{P(2HT)+P(2TH)}{P(2HT)+P(2TH)+P(1H)} \\ &= \frac{\frac{1}{6}\frac{1}{4}+\frac{1}{6}\frac{1}{4}}{\frac{1}{6}\frac{1}{4}+\frac{1}{6}\frac{1}{4}+\frac{5}{6}\frac{1}{2}} \\ &= \frac{1}{6} = 0.1667 \end{aligned}$$

15. (1 point (bonus)) Assume for a given dataset of n values you calculate a mean of $\bar{x}$ and variance s_x^2. If you multiply each x_i value by a constant b and add a constant a (i.e. $y_i = a + bx_i$), show that $\bar{y} = a + b\bar{x}$ and $s_y^2 = b^2 s_x^2$.

Solution:

$$\begin{aligned} \bar{y} &= \frac{\Sigma_{i=1}^n y_i}{n} \\ &= \frac{\Sigma_{i=1}^n (a+bx_i)}{n} \\ &= \frac{\Sigma_{i=1}^n a}{n} + b\frac{\Sigma_{i=1}^n x_i}{n} \\ &= a + b\bar{x} \\ s_y^2 &= \frac{1}{n-1}\Sigma_{i=1}^n (y_i - \bar{y})^2 \\ &= \frac{1}{n-1}\Sigma_{i=1}^n (a + bx_i - (a+b\bar{x}))^2 \\ &= \frac{1}{n-1}\Sigma_{i=1}^n (b(x_i - \bar{x}))^2 \\ &= b^2\frac{1}{n-1}\Sigma_{i=1}^n ((x_i - \bar{x}))^2 \\ &= b^2 s_x^2 \end{aligned}$$