DAWSON COLLEGE
DEPARTMENT OF MATHEMATICS
201-BZS-05
PROBABILITY AND STATISTICS
FALL 2015
FINAL EXAM

Name:_________________________ Date: December 24th, 2015

Student Number:________________ Time: 9:30 – 12:30

Grade: _______ / 116

Examiner: Matthew MARCHANT

Instructions:

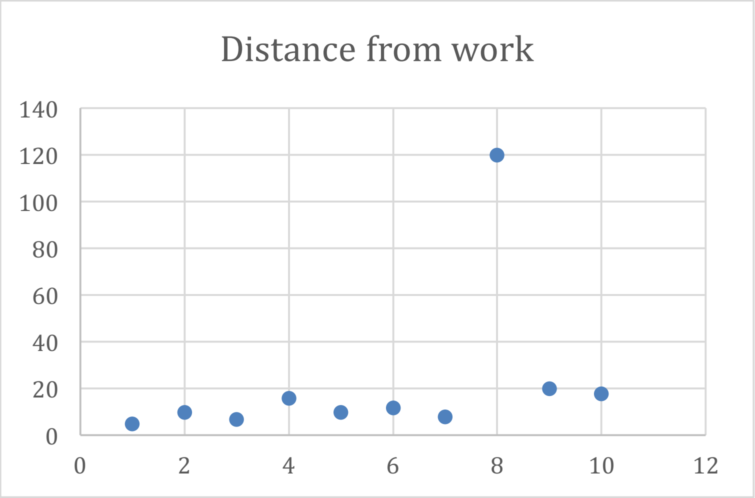
1. No books or notes are permitted.
2. Only calculators without text storage and graphical capability are permitted.
3. Please show all your work clearly.
4. Please justify all your answers.
5. Cheating will result in a minimum penalty of zero in your exam grade.
6. Unless otherwise stated, round your answer to 2 decimal places.

1. [10 marks] The commuting distance was determined for each of 10 employees at Acme Manufacturing. One of the employees lives in another town and has a large commuting distance. The 10 distances were as follows:

5, 10, 7, 15, 10, 12, 8, 120, 20, 18

a. Sketch the dot plot (use employee number as the x-axis)
b. Find the mean. By how much does the outlier affect the mean?
c. What statistic do you expect is more representative of the population variability, the standard deviation or the interquartile mean? Why?

ANS:a.



b. mean=22.6, mean without outlier=11.78 c. Interquartile mean because it is a robust statistic and is less sensitive to outliers and we have an outlier which is the value 120 minutes.

2. [5 marks] A quality-control technician selects 120 assembled parts from an assembly line and records the following information concerning these part:
A: defective or non-defective
B: the employee number of the individual who assembled the part
C: the weight of the part
a. What is the population?
b. What is the sample?
c. Give the types of the three variables (ie. Quantitative/Continuous).

ANS:

a. All of the parts that have ever been produced in the factory.
b. The 120 parts selected from the assembly line
c. A: qualitative ordinal
B: qualitative nominal
C: quantitative continuous

3. [15 marks] A pair of fair dice is rolled once.
Let E = the event of a sum of 8
Let F = the event of a product of 15

Let G = the event of doubles

Let H = the event where a 4 and a 1 are obtained

a. Draw the Venn diagram for the 4 events
b. Find: P(E), P(F), P(G) and P(H)
c. Find: $P(E \cap F), P(E \cap G), P(E \cap H), P(E|G)$
d. Are E and H independent? Why?
e. Are E and G independent? Why?
f. Are E and G mutually exclusive? Why?
g. Are H and E mutually exclusive? Why?

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

ANS:

b. P(E)=5/36
P(F)=2/36=1/18
P(G)=6/36=1/6
P(H)=2/36=1/18

c. P(E and F)=2/36=1/18
P(E and G)=1/36
P(E and H)=0
P(E given G)=1/6

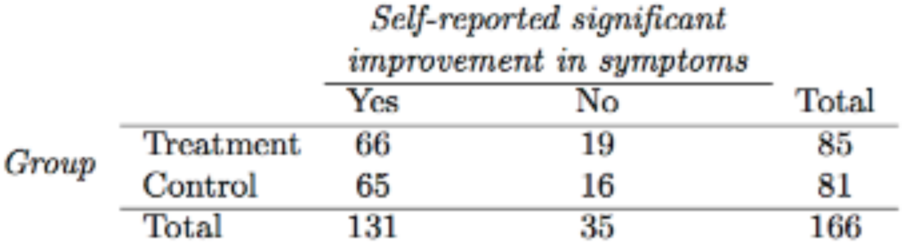
d. No, mutually exclusive events cannot be independent.
e. No, P(E given G) is not equal to P(E)
f. No, P(E and G) is not equal to 0.
g. Yes, P(E and H) is equal to 0.

4. [10 marks] A company has 10 identical machines that produce nails *independently*. The probability that a machine will break down is 0.1. Define a random variable X to be the number of machines that will break down in a day.

a. What is the appropriate probability distribution for X?
b. Give the expression for the probability that r machines will break down.
c. Compute the probability that at least 1 machines will break down.
d. What is the expected number of machines that will break down?
e. What is the variance of the number of machines that will break down?

ANS:
a. Binomial
b. P(X=r)=C(10,r)p^r x q^(10-r)
c. P(X>=1)=1-P(X<1)=1-P(X=0) = 0.65
d. E(X) = 1
e. Var(X)= 0.9

5. [12 marks] Participants of a study with sinusitis received either an antibiotic or a placebo and were asked at the end of a 10-day period if their symptoms had improved. The responses are summarized in the table below:



		Self-reported significant improvement in symptoms		
		Yes	No	Total
Group	Treatment	66	19	85
	Control	65	16	81
	Total	131	35	166

a. Comment on whether or not we can make a causal statement.
b. Set up hypotheses (Give Ho and Ha) to test whether the proportion of patients who reported significant improvement in symptoms is **greater** in the treatment group than in the control group.
c. Assume that the test statistic follows the Normal distribution and obtain i) the SE ii) the test statistic iii) *p*-value and iv) complete the hypothesis test at the 0.05 level of significance.

ANS:
a. Yes, it's an experiment
b. Ho: p_treat-p_control=0
Ha: p_treat-p_control>0
c. p_treat_hat = 66/85 = 0.78
p_control_hat = 65/81 = 0.80
p_treat_hat - p_control_hat = -0.026
p_hat_pooled = (0.78*85 + 0.80*81)/(85+81) = 0.79
SE = 0.063
Z*=-0.41
p_value(one sided) : 0.66, Not less than alpha=0.05. Fail to reject Ho at 0.05 L.O.S.

6. [12 marks] The following sample data pertain to the shipments received by a large firm from three different vendors. Test at the 0.01 level of significance whether the quality level of the items received and the vendor are *independent*.

	Number rejected	Number imperfect but acceptable	Number perfect	
Vendor A	12	23	89	∑=124
Vendor B	8	12	62	∑=82
Vendor C	21	30	119	∑=170
	∑=41	∑=65	∑=270	∑=376

   a. State the hypotheses.
   b. Check the conditions.
   c. Obtain the test statistic.
   d. Obtain the p-value and state the conclusion.

ANS:

a. Ho: Vendor and quality are independent
   Ha: Vendor and quality are not independent

b. Conditions:

1. Hard to say if we have less than 10% of population, we will assume that we do.
2. Expected values all greater than 5
3. df greater than 2.

c. Chi-SquareValue: 1.3

d. p_value: >0.3, greater than 0.01, therefore fail to reject Ho at 0.01 L.O.S.

7. [8 marks] Suppose the probability of having the disease is 0.001. If a person has the disease, the probability of a positive test result is 0.90. If a person does not have the disease, the probability of a negative test result is 0.95.
   For a person selected at random from the population, what is the probability they are infected given they have tested positive?

   Note: Use the following notation:
   dis: a person selected at random has the disease
   $\overline{dis}$: a person selected at random does not have the disease
   pos: a person selected at random tests positive for the disease
   $\overline{pos}$: a person selected at random tests negative for the disease

ANS: 0.018

8. [10 marks] Consider the following data for the time to commute to work for 10 employees:

Employee	Time to commute (min)
1	11.3
2	14.7
3	16.4
4	16.5
5	17
6	19.9
7	22.3
8	23.3
9	26.1
10	26.2

   a. Obtain the class width (use 4 classes or bins).
   b. Add a column for frequency.
   c. Add a column for relative frequency.
   d. Sketch the relative frequency histogram.

ANS:

a. 3.725

b. c.

bin	bin upper boundaries	freq	rel_freq
1	15.3	2	0.2
2	19.025	3	0.3
3	22.75	2	0.2
4	26.475	3	0.3

9. [8 marks] You are interested in learning about opinions of Dawson students about a proposed climate change policy.
   a. Explain how you would use stratified sampling to obtain sample data.
   b. Explain how you would use cluster sampling to obtain sample data.
   c. Which of the two methods, stratified or cluster sampling will take less time?
   d. Which of the two methods, stratified or cluster sampling is likely to give a more representative sample?

ANS:

a. Form stratas. One possibility is to use the programs students are enrolled in. Perform simple random sampling within these strata.

b. Form clusters (groupings that are self similar) and perform simple random sampling to select clusters and then perform simple random sampling within these clusters.

c. Cluster sampling would take less time as we would sample fewer students.

d. Stratified sampling because we are using more samples and not leaving out groups which is risky.

10. [8 marks] A student receives emails according to a Poisson distribution with an average of 53.5 e-mails every week.

a. Calculate the probability that the student receives exactly 115 e-mails in a 15-day period.

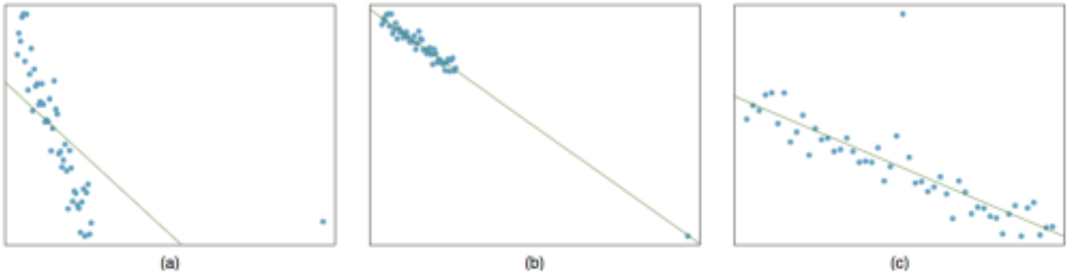
b. Calculate the probability that the time in between two emails is greater than 2 hours.

ANS:

a. 0.0372

b. 0.529

11. [6 marks] Identify the outliers in the scatterplots shown below, and determine if they are influential. Explain your reasoning.



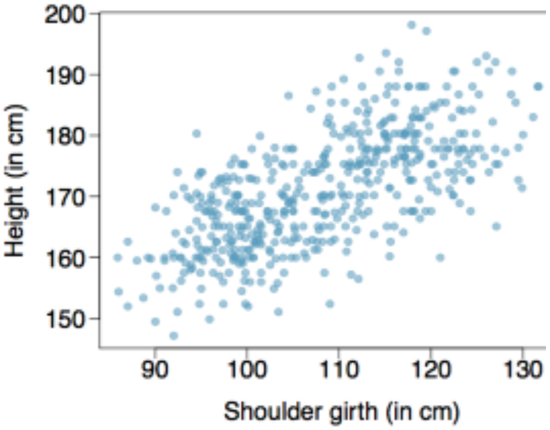
ANS:

a. Certainly influential as the slope has been strongly affected by the outlier. This is because it a high leverage point (big horizontal distance from the center of the data set).

b. Not at all influential because it falls directly on the line of best fit for the cluster furthest to the left.

c. Not influential because small leverage (small horizontal distance from the center of the data set).

12. [12 marks] Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender for 507 physically active individuals. The scatterplot below shows the relationship between height and shoulder girth (over deltoid muscles), both measured in centimeters.



The mean shoulder girth is 108.20 cm with a standard deviation of 10.37 cm. The mean height is 171.14 cm with a standard deviation of 9.41 cm. The coefficient of linear correlation between height and shoulder girth is 0.67.

c. Write the equation of the regression line for predicting height.

d. Interpret the slope and the intercept in this context.

e. Calculate R^2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application.

f. A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model.

g. The student from part (d) is 160 cm tall. Calculate the residual, and explain what this residual means.

h. A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

ANS:

a. y_hat = 105.4 + 0.61x
b. When shoulder girth is 0 the height is approx. 105.4 cm (which doesn't make physical sense)
   For an increase in 1 cm of shoulder girth, the height should increase by 0.61 cm.

c. R^2 = 0.4, is the reduction in variability by the regression line.

d. Y_hat(100)=166.15

e. E=6.15: Vertical distance between observation (160cm) and predicted value (166.16)

f. Y_hat(56)=139.41. Greater than the typical height of a one-year-old. Applying the linear model for x-values beyond the x-values in the data set is not recommended.

Formulas:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

$$P(A) = \frac{n(A)}{n(S)}$$

$$P(\mathrm{A} \cup \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{A} \cap \mathrm{B})$$

$$P(A|B) = \frac{n(A \cap B)}{n(B)} \quad P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A)P(B|A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

$$P(A) = \frac{n(A)}{n(S)}$$

$$P(\mathrm{A} \cup \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{A} \cap \mathrm{B})$$

$$P(A|B) = \frac{n(A \cap B)}{n(B)} \quad P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A)P(B|A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$\mu = E[x] = \sum_{all\ x} xP(x)$$

$$\sigma^2 = E[(x-\mu)^2] = \sum_{all\ x} (x-\mu)^2\, P(x)$$

$$P(X=k)=C(n,k)p^k q^{n-k}$$
$$\mu = np, \qquad \sigma^2 = npq$$

$$P(x)=\frac{C(k,x)\cdot C(N-k,n-x)}{C(N,n)}$$
$$\mu=\frac{nk}{N} \;, \quad \sigma^2=\frac{nk(N-k)(N-n)}{N^2(N-1)}$$

$$P\left[X=k\right]=\frac{\alpha^k}{k!}e^{-\alpha}$$

$$P\left(a\le x\le b\right)=P\left(a<x<b\right)=\int_a^b f\left(x\right)dx$$
$$\mu=E\left[X\right]=\int_{-\infty}^{\infty} xf\left(x\right)dx$$
$$\sigma^2=E\left[\left(x-\mu\right)^2\right]=E\left[X^2\right]-\mu^2$$

$$f(x)=\begin{cases}\lambda e^{-\lambda x} & x\ge 0\\ 0 & x<0\end{cases}$$
$$SE=\frac{\sigma}{\sqrt{n}}\approx\frac{S_x}{\sqrt{n}}$$

C.I. : point estimate $\pm$ $Z^{*}SE$

Test statistic: $Z=\dfrac{\text{point estimate - null value}}{SE}$

$$SE_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \quad , \; SE_{\hat{p}}=\sqrt{\frac{p(1-p)}{n}}$$
$$SE_{\hat{p}_1-\hat{p}_2}=\sqrt{SE_{\hat{p}_1}^2+SE_{\hat{p}_2}^2}=\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}} \;, \quad \hat{p}=\frac{\hat{p}_1 n_1+\hat{p}_2 n_2}{n_1+n_2}$$
$$\chi^2=\sum_{i=1}^{k}\frac{\left(O_i-E_i\right)^2}{E_i},\; df=\left(k-1\right) \quad , \quad \chi^2=\sum_{ij}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \;, \; df=(R-1)\times(C-1)$$

$$\hat{y}_i=b_0+b_1x_i \quad , \quad b_0=\bar{y}-b_1\bar{x} \quad b_1=\frac{s_y}{s_x}R$$