Math 661-102, Spring 2013

Final Exam

Name:____________

Student ID:_____________

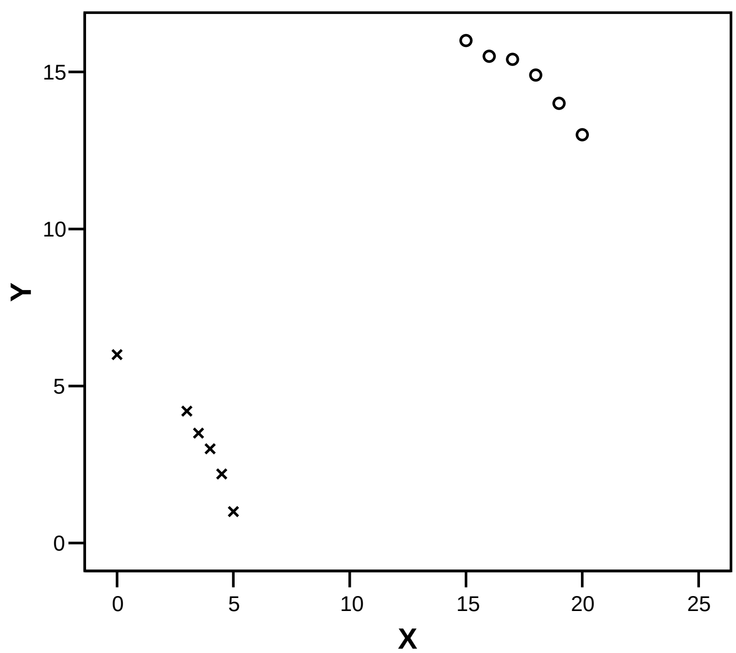
May, 14 **Must show all work to get full credit!**

I pledge I have not violated the NJIT Honor Code______________________

1. In order to select a sample of undergraduate students in the United States, a simple random sample of four states is selected. From each of these states, a simple random sample of two colleges or universities is then selected. Finally, from each of these eight colleges or universities, a simple random sample of 20 undergraduates is selected. The final sample consists of 160 undergraduates. What sampling technique is being used? (Quiz 2) (4 point)
   A) Simple random sampling
   B) Stratified random sampling
   C) Multistage sampling
   D) Convenience sampling

Answer C). It is not a simple random sample nor just stratification and D) is false.

2. The scatterplot below represents a small data set. The data were classified as either of type 1 or type 2. Those of type 1 are indicated by x's and those of type 2 by o's.



What do we know about the overall correlation of the data in this scatterplot? (4 points)

   A) It is positive.
   B) It is negative, because the o's display a negative trend and the x's display a negative trend.
   C) It is near 0, because the o's display a negative trend and the x's display a negative trend, but the trend from the o's to the x's is positive. The different trends will cancel each other out.
   D) It is impossible to compute for such a data set.

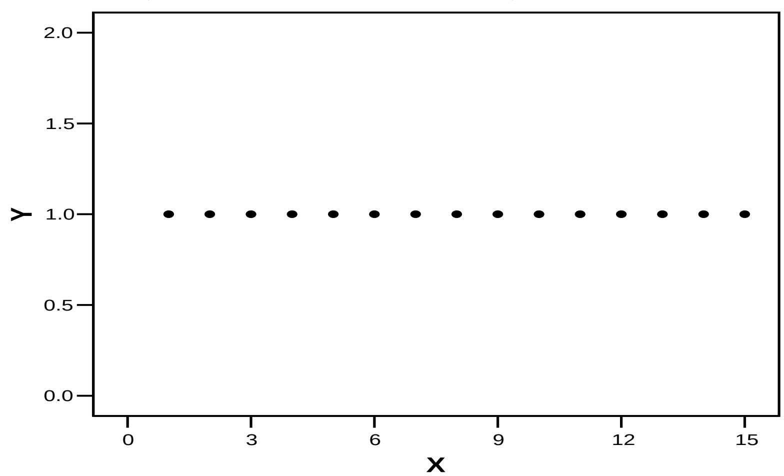
Answer A). B) is false and C) is false. D) is also false.

3 Suppose several of the units selected into a random sample cannot be found or contacted during the conducting of a survey. What can we say about this situation? (Quiz 2) (4 point)

A) This sample contains a lurking variable.
B) This will likely produce nonresponse bias in the sample results.
C) The sample results will suffer from response bias.
D) This situation came about because of interviewer inefficiency.
E) The sample selected is suffering from undercoverage.

Answer B . There are no lurking variables. C) is false. D) is not logical E) is false.

4. A scatterplot of a variable *y* versus a variable *x* produced the scatterplot shown below. The value of *y* for all values of *x* is exactly 1.0. (4 point)



What do we know about the correlation between *x* and *y*?

A) It is +1 because the points lie perfectly on a line.
B) It is either +1 or –1, because the points lie perfectly on a line.
C) It is 0 because *y* does not change as *x* increases.
D) None of the above.

Answer is C). A) B) and D) are false.

5. Many high school students take either the SAT or the ACT. However, some students take both. Data were collected from 60 students who took both college entrance exams. The average SAT score was 912 with a standard deviation of 180. The average ACT score was 21 with a standard deviation of 5. The correlation between the two variables equals 0.817. What fraction of the variation in the values of the SAT scores is accounted for by the linear relationship between SAT and ACT scores? (4 point)

A) 66.7%
B) 81.7%
C) 90.4%
D) Cannot be determined from the information given.

Answer is A). The rest are false or not correct answers.

6. Typographical and spelling errors can be either “non-word errors” or “word errors.” A non-word error is not a real word, as when “the” is typed as “teh." A word error is a real word, but not the right word, as when “lose” is typed as “loose.” Non-word errors make up 20% of all typographical and spelling errors. A human proofreader will catch 90% of

non-word errors and 70% of word errors. What percentage of typographical and spelling errors will the proofreader catch? Given the proofreader caught a typographical and spelling error what is the probability that it is a word error? (reference problem 4.128, page 293) (15 points).

The P("non-word error" and "caught") = P(N and C) = P(N) P(C | N) = (0.2)(0.9)=0.18.

The P("word error" and "caught") = $P(N^c$ and C) = $P(N^c)$ $P(C \mid N^c)$ = (0.8)(0.7) = 0.56.

A proofreader should catch about P(C) = 0.18 + 0.56= 0.74 = 74% of all typographical and spelling errors.

$P(N^c \mid C) = P(N^c)\, P(C \mid N^c)/ P(C)$ = 0.56/0.74 = 0.7568.

7. The financial records of businesses may be audited by state tax authorities to test compliance with tax laws. Suppose the auditors examine a simple random sample of 160 sales records out or 10,000 available. One issue is whether each sale was correctly classified as subject to state sales tax or not. Suppose that 800 of the 10,000 are incorrectly classified. Use normal distribution to find the approximate probability that the number of misclassified sale in the sample is more than 13. Give justification for all approximations used. (Midterm exam, Reference Examples 5.14 and 5.22, page 315 and page 326, respectively) (16 points)

p = probability of a sale being misclassified = $\frac{8}{100}$ = 8%.

X: # out of n = 160 that are misclassified.

Since $\frac{n}{N} = \frac{160}{10,000} = .016 < .05$. Hence

X can be treated as Binomial (n = 160, p = .08)

$P(X > 13) \overset{\text{continuity adjust.}}{=} P(X > 13.5) = P\left(Z > \frac{13.5 - 12.8}{\sqrt{12.8 \times (.92)}} = 0.2\right)$

$np = (.08)(160) = 12.8 > 10$ & $n(1-p) = 147.5 > 10$. Hence, normal approximation to Binomial is good. Ans = 1 − 0.5793 = **0.4207**

8. (a) – (c) A study of iron deficiency among infants compared samples of infants following different feeding regimens. One group contained breast-fed infants; while the children in another group were fed a standard baby formula without any iron supplements. Here are summary results on blood hemoglobin levels at 12 months of age.

Group	**n**	$\bar{x}$	**s**
Breast-fed	23	13.3	1.7
Formula	19	12.4	1.8

a. Is there significant evidence that the mean hemoglobin level is higher among breast-fed babies? State H_0 and H_a and carry out an appropriate test. Find the P-value. What is your conclusion? (10 points)

(a) We test H_0: $\mu_b = \mu_f$; H_a: $\mu_b > \mu_f$.

$SE_D \doteq 0.5442$ and $t = 1.654$, for which $P = 0.0532$ (df $= 37.6$) or 0.0577 (df $= 18$); there is not quite enough evidence to reject H_0 at $\alpha = 0.05$.

b. Give a 95% confidence interval for the mean difference in hemoglobin level between the two populations of infants. (10 points)

The confidence interval depends on the degrees of freedom used; see the table.

df	t^*	Confidence interval
37.6	2.0251	–0.2021 to 2.0021
18	2.101	–0.2434 to 2.0434

c. State the assumptions that your procedures in parts (a) and (b) require in order to be valid. (problem reference 7.85, page 457-8) (5 points)

We need two independent Simple Random Samples from Normal populations.

9. According to literature on brand loyalty, consumers who are loyal to a brand are likely to consistently select the same product. This type of consistency could come from a positive childhood association. To examine brand loyalty among fans of Chicago Cubs, 371 Cubs fans among patrons of a restaurant located in Wrigleyville were surveyed prior to a game at Wrigley Field, the Cubs' home field. The respondents were classified as "die-hard fans" or "less loyal fans." Of the 134 die-hard Cubs fans, 90.3% reported that they had watched or listened to Cubs games when they were children. Among the 237 less loyal fans, 67.9% said that they had watched or listened to Cubs games as children.

a. Find the number of die-hard Cubs fans who watched or listened to games when they were children. Do the same for the less loyal fans. (4 points)

(a) $X_1 =$ 121 =(0.903)(134) die-hard fans and $X_2 =$ 161 =(0.679)(237) less loyal fans who watched or listened to Cubs games as children.

b. Use a significance test to compare the die-hard fans with the less loyal fans with respect to their childhood experiences relative to the team. Use alpha = 0.05. Clearly state the hypotheses and draw conclusion. (10 points)

H_0: $p_1 = p_2$ versus H_1: $p_1 \neq p_2$.

$\hat{p} = \frac{121+161}{134+237} =$ 0.7601 and

$SE_{Dp} =$ 0.04615, so we

find z = 4.85

($P < 0.0001$)—strong evidence of a difference in proportions of childhood experience in the two groups.

c. Express the results with a 96 % confidence interval for the difference in proportions. (Page 507 problem # 8.88) (10 points)

For a 96% confidence interval, $SE_D =$ 0.03966 and the interval is

(0.903 - 0.679) ± (2.05)(0.03966), i.e., 0.224 ± (2.05)(0.03966), 0.224 ± 0.081303

(0.142697, 0.30530)