Analytics Primer

Final Exam

Honor Pledge:

I certify that I have not received or given unauthorized aid in taking this exam.

*Signed:*______________________________________

Printed Name: KEY

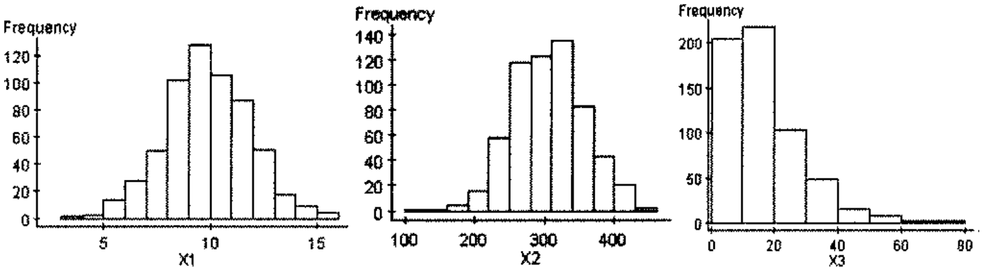
*Date:*____________________

There are 32 questions. The first 20 are worth 2 points each. The last 12 are worth 5 points each. There will be some True/False, Multiple Choice, and Short Answer questions.

1. Briefly describe the difference between stratified random sampling and cluster sampling.

   In stratified sampling, the population is split into groups (called strata) and every group is sampled from.
   In cluster sampling, the population is split into groups (called clusters) and only some groups are sampled from.

Consider the following histograms of population variables labeled X1, X2, and X3. Use these to answer questions 2 and 3.



2. Which population distribution would the mean and median would be furthest apart if all the variables were put on the same scale?
   a. X1.
   b. X2.
   [✓ c.] X3.
   d. It is impossible to tell.

3. If you repeatedly took large samples, calculated their sample means, and graphed the resulting sampling distribution from each of these populations separately, then which of these sampling distributions is normal?
   a. X1 and X2 only.
   [✓ b.] All of these populations (X1, X2, and X3).
   c. None of these because sample means never follow a normal distribution.
   d. Need more information.

4. Which statistic is resistant to outliers?
   a. Variance.
   b. Mean.
   c. Range.
   [✓ d.] None of the above.

5. A standard deviation is any number between -1 and 1 only.
   - a. True.
   - [✓ b.] False.

6. Briefly state the Central Limit Theorem.

   Average of a large sample follows an approximate Normal distribution.

7. The sampling distribution of a statistic describes how possible values of the statistic vary
   - a. Within the population of interest.
   - b. Within the selected sample.
   - c. Between the sample and the population.
   - [✓ d.] Among different possible samples.

8. Confidence intervals are fixed quantities that have the same value for every sample from a population.
   - a. True.
   - [✓ b.] False.

9. Briefly describe the difference between a t-distribution and a standard Normal distribution.

   The t-distribution has thicker tails than the standard Normal, which allows us to use it when estimating sample std. deviations in mean tests.

10. A manager for a large food company was testing whether a machine that bags potato chips was producing bags that were not either too full or too empty. To test this claim the manager conducted a two-sided hypothesis test for mean weight of bags off the line. The manager sampled 140 bags and concluded to not reject the null hypothesis and that the machine was working properly. While the machine went through regular maintenance, it was discovered that the machine was working incorrectly and the manager had made the wrong conclusion. Which of the following is true?
    - a. The manager must have calculated something incorrectly because hypothesis tests are always correct if no calculation error is made.
    - b. The manager committed a type I error.
    - [✓ c.] The manager committed a type II error.
    - d. Both b and c, but not a.

11. P-values are often reported for many statistical tests. In your own words, briefly describe what a p-value is.

The probability we got our sample (or one more rare) given the null hypothesis is true.

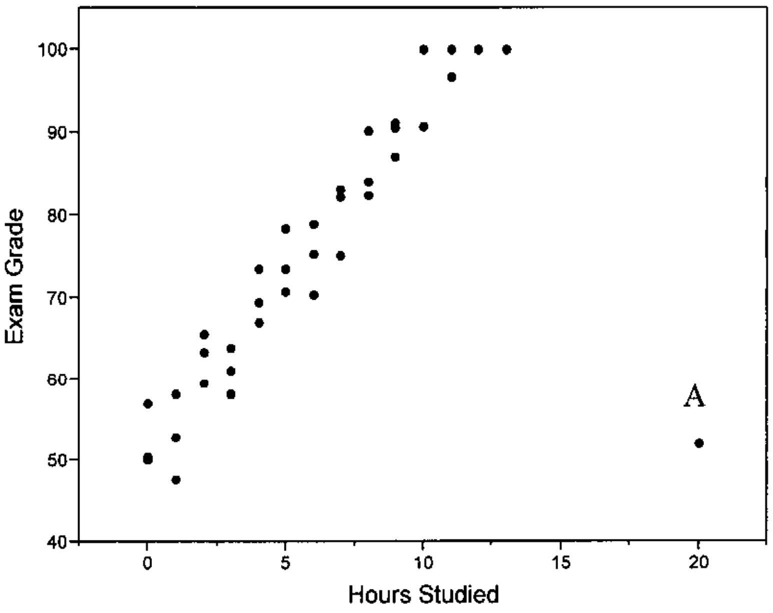
12. Which of the following is a characteristic of the F-distribution?
    a. Symmetric.
    [✓ b.] Right Skewed.
    c. May take positive or negative values.
    d. Only has one set of degrees of freedom.

13. State two assumptions to the multiple linear regression model.

    1. Normality of Errors   Equal variance of errors   Independence of errors

    2. Linearity of means   No perfect collinearity

An instructor of a college literature course records the grades of the students on the final exam for the course along with the students admitted time spent on studying for the exam. Use the following plot to answer question 14.



14. There is an outlier in this data (marked with the letter "A"). If we removed this outlier the following would happen to the correlation coefficient.
    a. Get closer to zero.
    [✓ b.] Get closer to one.
    c. Stay the same.
    d. Need more information.

A data analyst is trying to interpret the following results of a multiple linear regression with an $R^2 = 0.725$. Use these results to answer question 15.

Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value
Intercept	78.21	33.53	68	2.333	0.0226
X_1	-5.81	1.32	68	-4.402	<0.0001
X_2	3.64	0.47	68	7.745	<0.0001

15. What is the value of the correlation coefficient r between the response variable and the first independent variable X_1?
    a. 0.851.
    b. 0.725.
    c. -0.851.
    [✓ d.] Not enough information.

16. A researcher wants to test if there is a difference between the effectiveness of three different types of fertilizers on crop yield. Which type of analysis can answer this question?
    a. Simple linear regression with only one variable for type of fertilizer.
    b. Multiple linear regression with 2 dummy variables for type of fertilizer.
    c. One-way ANOVA.
    [✓ d.] Both b and c, but not a.

17. Drug analysts are trying to determine if the four top-selling blood pressure medicines have the same effect on blood pressure. The one-way ANOVA they perform is an example of which of the following?
    a. A random effects model.
    [✓ b.] A fixed effects model.
    c. Both a random and fixed effects model.
    d. Need more information.

18. Briefly describe why the Tukey-Kramer comparison method is needed in an ANOVA compared to the traditional two-sample t-tests.

Traditional t-tests only control comparisonwise error rates, not experimentwise error rates. Therefore, the more we do, the lower the overall accuracy. T-K corrects this.

19. Performing a Chi-square test for association will only help determine if an association exists, another separate statistic must be used to determine the strength of the association.
    [✓ a.] True.
    b. False.

20. Which Chi-square test is used for ordinal categorical variables?
    a. Pearson Chi-square test.
    [✓ b.] Mantel-Haenszel Chi-square test.
    c. Likelihood Ratio Chi-square test.
    d. There are no Chi-square tests for ordinal categorical variables.

A fast food restaurant is interested in answering some questions about different food options that were ordered by their customers. The restaurant calculated that 40% of all of the customers ordered hamburgers, 67% ordered a side of French fries, and 29% of all customers had both hamburgers and French fries. Use this information to answer questions 21 and 22.

21. (5 points) What is the probability that any one random customer orders either a hamburger **or** a side of French fries?

$$P(H \text{ or } FF) = P(H)+P(FF)-P(H \text{ and } FF)$$

$$= 0.40 + 0.67 - 0.29$$

$$= 0.78$$

22. (5 points) What is the probability that a customer orders French fries given that they already ordered a hamburger?

$$P(FF|H) = \frac{P(H \text{ and } FF)}{P(H)} = \frac{0.29}{0.4} = 0.725$$

A large law firm's board of directors is brain-storming ideas for a commercial. One of members of the board remembered that the success rate for all of the firm's cases was 68% and mentions that it should be used in the commercial. Use this information to answer question 23.

23. (5 points) What is the probability that you randomly select 159 cases from this firm and more than 115 of them were successful?

$$\hat{p} = \frac{115}{159} = 0.723$$

$$z = \frac{\hat{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.723 - 0.68}{\sqrt{\frac{0.68(1-0.68)}{159}}} = 1.16$$

$$P(z > 1.16) = 0.1230$$

A large retailer wants to determine the average sales (in millions of dollars) across all of their stores nationwide. However, it is too costly and time consuming to contact every store. The manager in charge of the task wants to use a sample size of 250 stores and does not know if the sales of the stores nationwide follow a Normal distribution. Use this information to answer questions 24 – 26.

24. (5 points) Does the sample meet all of the needed qualifications/assumptions to conduct a confidence interval? Explain.

Yes, large sample of 250 stores.

25. (5 points) An analyst working on the project thinks the sample of 250 stores is too big for the margin of error of $0.75 million specified by the vice president wanting to know the information. The analyst knows the standard deviation of previous studies of this type was $4.5 million. Calculate the estimated sample size needed for the desired margin of error and a 95% confidence interval.

$$n = \frac{z^2 \hat{\sigma}^2}{e^2} = \frac{1.96^2 \cdot 4.5^2}{0.75^2} = 138.3 \approx 139$$

26. (5 points) The manager doesn't listen to the suggestion by the analyst and collects a sample of 250 stores with sample average sales of $2.37 million with a sample standard deviation of $4.72 million. Calculate a 95% confidence interval for the true average sales of all stores.

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

$$2.37 \pm 1.984 \cdot \frac{4.72}{\sqrt{250}}$$

$$2.37 \pm 0.59 \quad \text{OR} \quad (1.78, 2.96)$$

An analyst working for a hospital is trying to develop a model that predicts the average cost in dollars of a stay at the hospital for incoming patients. The analyst comes up with the following model from a sample of 68 patients. Use the following information to answer questions 27 – 30.

$$x_1 = Age\ of\ patient\ in\ years$$
$$x_2 = \begin{cases} 1 & if\ patient\ is\ male \\ 0 & if\ patient\ is\ female \end{cases}$$
$$x_3 = \begin{cases} 1 & if\ patient\ has\ insurance \\ 0 & if\ patient\ does\ not\ have\ insurance \end{cases}$$
$$x_4 = Number\ of\ visits\ to\ the\ hospital\ over\ the\ past\ year$$

$$\hat{y}_i = 647.12 + 21.57x_{1,i} - 582.36x_{2,i} - 1137.90x_{3,i} + 185.22x_{4,i}$$

$$SSE = 418741 \qquad TSS = 948796$$

27. (5 points) What is the expected cost of an incoming 43 year old female with insurance coverage that has visited the hospital on two previous occasions over the past year?

$$\hat{y} = 647.12 + 21.57(43) - 582.36(0) - 1137.9(1) + 185.22(2)$$
$$= 807.17$$

28. (5 points) Interpret the coefficient for x_4 in terms of the problem.

Holding all else constant, every extra stay at the hospital over the past year lead to an average increase in cost of $ 185.22.

29. (5 points) Calculate the R^2 and R_A^2 for this model.

$$R^2 = 1 - \frac{SSE}{TSS} = 0.559$$

$$R_A^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$
$$= 1 - (1 - 0.559)\left(\frac{67}{63}\right) = 0.531$$

30. (5 points) Use the following information to conduct a hypothesis test to determine if the variable x_3 is significant in the model (the coefficient is different than zero). Be sure to state all the parts of the hypothesis test.

$$SSE = 418741 \qquad R_3^2 = 0.24 \qquad \sum_{i=1}^{n}(x_{3,i} - \bar{x}_3)^2 = 196$$

$H_0: \beta_3 = 0$

$H_a: \beta_3 \neq 0$

$t = \frac{b_3 - 0}{s_{b_3}}$

$= \frac{-1137.9}{7.66}$

$= -148.51$

$s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}}$

$= 81.53$

$s_{b_3} = \frac{s_\varepsilon}{(1-R_3^2)\sqrt{\Sigma(x_{3,i}-\bar{x}_3)^2}}$

$= 7.66$

$P(|t| > +148.51) < 0.001$

REJECT H_0

Insurance coverage is significant in predicting cost of hospital stay.

A researcher is conducting a study at a large university on the east coast. The researcher is trying to determine whether income bracket of parents of incoming freshmen (separated into 4 groups: lower class, middle class, upper-middle class, upper class) helps determine the overall GPA at the end of freshmen year at a college within the university. The researcher wants to block by the 10 majors offered by the college. The researcher samples one person from each income group and major to collect a total sample of 40. Assume all of the needed assumptions are met. Use the following information to answer questions 31 and 32.

Source	DF	Sum of Squares	Mean Square	F Ratio	P-Value
Blocks	9	14.5	1.61	1.14	0.3703
Between	3	54.6	18.2	12.9	<0.0001
Within	27	38.2	1.41		
Total	39	107.3			

31. (5 points) Fill in the missing spaces in the table above.

DONE

32. (5 points) Did the survey need to be blocked by major? Explain.

No, blocking was not needed because it was insignificant in our model.
(P-value = 0.3703 > 0.05)