ST517 Sample Final Exam

NOTE to students: This is intended to give you an idea of the type questions the instructor asks and the approximate length of the exam. It does NOT indicate the exact questions or the topics covered. Students should refer to the information posted on Moodle to determine the coverage of the material.

Solutions are presented at the end of this document.

Name (Print Legibly): _____

Instructions:

- Read each question carefully.
- Provide one answer to each question in the space provided.
- For the short answer problems:
  - Show all work or explain your reasoning to receive full credit.
  - Make all work legible.
- You may use a calculator (no cell phones).
- Unless otherwise specified give answers to **3 decimal places**.
- Unless otherwise stated:
  - Use a 5% significance level for hypothesis tests
  - Use a 95% confidence level for confidence intervals

Honor Pledge:

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

Signed: _____

ST517 Sample Final Exam

True/False-Multiple Choice. Circle the best answer (3 points each).

1. **True False** A normal distribution is always centered at zero.
2. **True False** A lower value of AIC generally indicates a better fitting model.
3. **True False** In multiple regression, a t-statistic is used to test the fit of the model overall.
4. **True False** The p-value is the probability that the null hypothesis is correct.
5. **True False** A p-value is always between -1 and 1.
6. **True False** A correlation is always between -1 and 1.
7. **True False** The expected value of the Chi-square distribution is equal to the degrees of freedom for the distribution.
8. **True False** A Chi-square test of homogeneity is used to determine if the distribution of a categorical variable is the same for two or more populations.

Use the following to answer question 9 and 10: Results of a regression analysis exploring the relationship between the cost (in U.S. dollars) of a textbook at the university bookstore and the number of pages in the book are shown below:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	54	16.0	8.52	1.878	0.0657
Slope	54	XXXX	0.0143	6.853	<0.0001

9. Which of the following is a valid interpretation of the p-value in the row labeled 'Slope.'
    a. If there were no relationship between the number of pages and the cost of the book, there would be less than a 0.01% chance of observing a test statistic of 6.853 or larger.
    b. If there were a positive relationship between the number of pages and the cost of the book, there would be less than a 0.01% chance of observing a test statistic of 6.853 or smaller.
    c. If there were no relationship between the number of pages and the cost of the book, there would be less than a 0.01% chance of observing a test statistic as or more extreme than 6.853.
    d. If there were a positive relationship between the number of pages and the cost of the book, there would be less than a 0.01% chance of observing a test statistic as or more extreme than 6.853.

10. What is the value of the estimated slope coefficient?
    a. +0.098
    b. -0.098
    c. +0.0143
    d. It is impossible to tell without knowing the values of *SSxy* and *SSxx.*

11. In multiple regression, the variance inflation factor (VIF)
    a. is always greater than (or equal to) 10.
    b. measures the amount of collinearity in the predictor variables.
    c. indicates which variable is the best predictor of the response.
    d. will be large if the x variable is a good predictor of the response.

12. We would like to create a confidence interval. Which of the following would produce the narrowest interval?
    a. A 80% confidence level.
    b. A 90% confidence level.
    c. A 95% confidence level.
    d. A 99% confidence level.

13. An educational software company wants to assess the usefulness of its software. It runs a "quick vote" poll on a website, asking users to indicate whether they like or dislike the software. Of 900 respondents, 610 said they liked the software. The results of the sample are probably
    a. unbiased, because of the large sample size.
    b. biased, because it is a voluntary response sample.
    c. unbiased, because it is a simple random sample.
    d. biased, because a larger sample should be used.

14. In regression we typically test null hypothesis that the slope is equal to zero. One reason for this is because
    a. If the slope is zero the sum of squares error will be zero.
    b. If the slope is zero, the R^2 will be close to 1.
    c. If the slope is zero, the correlation will be zero.
    d. If the slope is zero, the mean of y-variable will be zero.

15. Which of the following conditions are required for a function to be a valid pmf for a discrete random variable?
    a. The function must be non-decreasing.
    b. Sum of all probabilities must equal 1.
    c. No probabilities can be negative.
    d. Both b. and c. but not a.

16. Consider the following matrix $\mathbf{X} = \begin{bmatrix} 1 & 3 \\ 1 & 5 \\ 1 & 7 \end{bmatrix}$. What is $\mathbf{X'X}$?

    a. $\begin{bmatrix} 10 & 16 & 22 \\ 16 & 26 & 36 \\ 22 & 36 & 50 \end{bmatrix}$

    b. $\begin{bmatrix} 4 & 6 & 8 \\ 4 & 6 & 8 \end{bmatrix}$

    c. $\begin{bmatrix} 3 & 15 \\ 15 & 83 \end{bmatrix}$

    d. $\begin{bmatrix} 4 & 4 \\ 6 & 6 \\ 8 & 8 \end{bmatrix}$

17. The sampling distribution of a statistic is
    a. the probability that we obtain the statistic in repeated random samples of the same size from the same population.
    b. the mechanism that determines whether randomization was effective.
    c. the extent to which the sample results differ systematically from the truth.
    d. the distribution of values taken by a statistic in all possible samples of the same size from the same population.

18. Suppose *X* has a binomial distribution based on 16 trials and a 0.7 probability of success. What is the standard deviation of *X*?
    a. 11. 2
    b. 9.80
    c. 3.36
    d. 1.83

Use the following for questions 19 to 22: Does the type of movie children are watching make a difference in the amount of snacks they will eat? A group of 50 children were randomly assigned to watch either a cartoon or a live action musical (25 to each). Crackers were available in a bowl, and the investigators compared the number of crackers eaten by children while watching the different kinds of movies.

19. In this study the response variable is:
    a. The amount of crackers eaten.
    b. The children.
    c. Does the type of movie make a difference in the amount of snacks eaten?
    d. The type of movie watched.

20. In this study the explanatory variable is:
    a. The amount of crackers eaten.
    b. The children.
    c. Does the type of movie make a difference in the amount of snacks eaten?
    d. The type of movie watched.

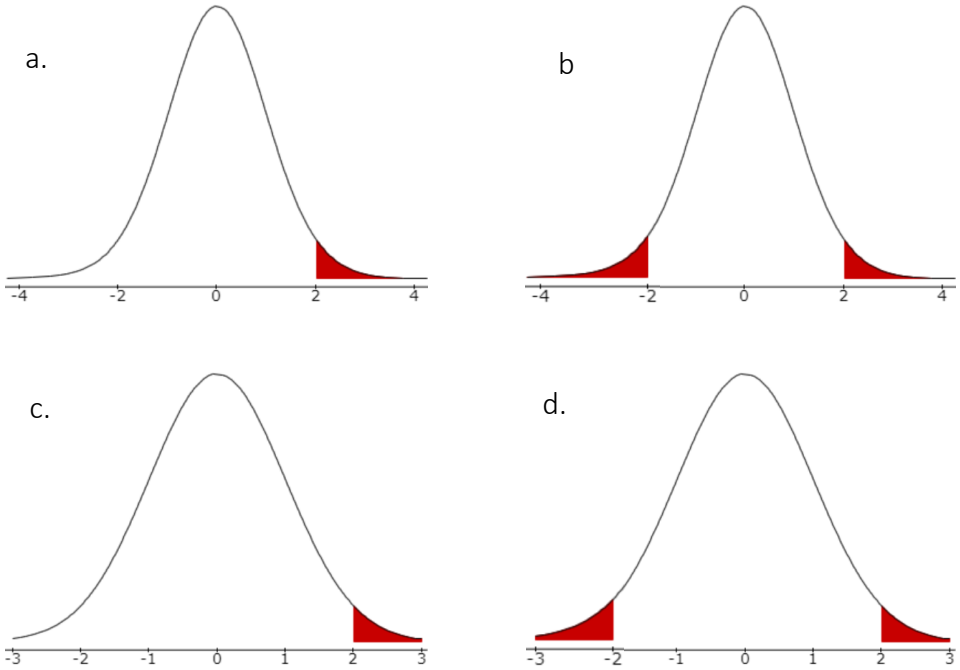
21. This study is best described as:
    a. A placebo controlled experiment.
    b. A matched pairs experiment.
    c. A randomized experiment.
    d. An observational study.

22. In the study described above, one kind of movie was shown at 8 AM (after the children had breakfast) and another at 11 AM (before the children had lunch). It was found that during the movie shown at 11 AM, more crackers were eaten than during the movie shown at 11 AM. The investigators concluded that the different types of movies had an effect on appetite. The results cannot be trusted because
    a. the study was not double blind. Neither the investigators nor the children should have been aware of which movie was being shown.
    b. the investigators were biased. They knew beforehand what they hoped the study would show.
    c. the investigators should have used several bowls, with crackers randomly placed in each.
    d. the time the movie was shown is a lurking variable.

23. An instructor was concerned that the highest score on a recent exam was only 99 (instead of 100). He decided to add one point to everyone's score. The effect of this would be
    a. The standard deviation would increase by 1.
    b. The mean (i.e. expected value) would not change.
    c. The variance would decrease.
    d. The standard deviation would not change but the mean would increase.

24. Which of the following is not a necessary condition for applying the Binomial distribution?
    a. The trials are independent of each other.
    b. The variance is equal to the mean of the distribution.
    c. The probability of a success is the same for all trials.
    d. The number of trials is fixed.

25. We test the hypotheses $H_0: \mu = 20\ vs. H_A: \mu \neq 20$. From a sample of 30 subjects we calculate a test statistic of t=2. Which picture best represents the p-value? [Hint: use the values on the horizontal axis to help you determine the appropriate distribution.]



Use the following for Questions 26 and 27: A high school statistics class wants to estimate the average weight of the chocolate chips per cookie in a generic brand of chocolate chip cookies. They collect a random sample of cookies, rinse away the baked dough, and obtain the weight in grams of the chocolate chips for each cookie. Based on their data, the 95% confidence interval for the average weight of chocolate per cookie goes from 5.65 to 6.35 grams.

26. Which of the following is a valid interpretation?
    a. We believe that the true average weight of chocolate per cookie is between 5.65 to 6.35 grams.
    b. We believe that 95% of all cookies from this generic brand will have between 5.65 to 6.35 grams of chocolate.
    c. There is a 95% probability that the true average weight of chocolate per cookie is between 5.65 and 6.35 grams.
    d. Both a and c but not b.

27. The class would like to use this confidence interval to test the following hypotheses: Ho: μ = 7 vs. Ho: $\mu \neq 7$. What is the appropriate decision?
    a. Reject Ho, since the value of 7 is a reasonable estimate for μ.
    b. Reject Ho, since the value of 7 is not a reasonable estimate for μ.
    c. Fail to reject Ho, since the value of 7 is a reasonable estimate for μ.
    d. Fail to reject Ho, since the value of 7 is not a reasonable estimate for μ.

Use the following for questions 28 and 29: A company that makes candy-coated chocolate pieces annually produces a special holiday mix of candy corn colors. They claim that in this mix, 25% of the candies are yellow, 40% are orange, and 35% are white. Suppose we take a random sample of 50 candies from this mix and find the following counts:

Color	Yellow	Orange	White
Observed count	10	23	17

28. What type of Chi-square test would we use to determine if the company's stated model for the colors is correct?
    a. Goodness of fit
    b. Homogeneity
    c. Independence
    d. Either b. or c. but not a.

29. If the distribution is really as claimed, how many white candies would we expect in the sample?
    a. 0.35
    b. 3
    c. 17
    d. 17.5

30. Researchers surveyed 1,000 randomly selected adults in the United States. A strong, positive, statistically significant correlation was found between income and the number of containers typically recycled in a week. Can the researchers conclude that earning more money causes more recycling among U.S. adults?
    a. No, the study design does not allow causation to be inferred.
    b. No, the sample size is too small to allow causation to be inferred.
    c. Yes, the statistically significant result allows causation to be inferred.
    d. Yes, there is strong evidence that income causes people to care more about the environment.

31. We conduct a regression analysis using shoe size as the explanatory variable and height (in inches) as the response variable. If we had measured height in centimeters instead of inches, which of the following would change?
    a. The value of the slope.
    b. The value of the correlation.
    c. The value of R-square.
    d. Both b. and c. but not a.

32. Recall the BAC data from class. The full dataset had 16 subjects and 4 variables: BAC (which was the response variable) and 3 possible predictors (number of beers, weight, sex). What would be the dimensions of the design matrix for this data?
    a. 16 × 3
    b. 16 × 4
    c. 3 × 16
    d. 4 × 16

33. A political scientist obtains a list of the 3114 undergraduates at her college and mails a questionnaire to 250 students selected at random. Only 100 of the questionnaires are returned. The bias most likely to affect these results is
    a. Selection bias
    b. Undercoverage
    c. Response bias
    d. Non-response bias

34. We use the t-distribution to calculate a confidence interval for the population mean μ. If we increase the sample size from 10 to 20 the interval would become smaller because of
    a. the change in degrees of freedom.
    b. the change in standard error.
    c. both a and b.
    d. none of the above.

Short Answer: Show your work or explain your answers.

35. (6 points) We are able to make inference because statistics have a predictable distribution called a sampling distribution. Explain briefly how the sampling distribution is used in the inferential process of hypothesis testing.

36. There is a basket with 100 marbles inside. The marbles are either green or yellow, but we do not know the number of each color. We want to decide between the hypotheses:
    H_0: Basket has 50 green and 50 yellow vs. H_A: Basket has 20 green and 80 yellow
    We will select a single marble and reject the null hypothesis if that marble is yellow.
    a. (6 points) Define a Type I Error in the context of this problem.

    b. (6 points) Calculate P(Type I error).

37. (6 points) Customers using a self-service soda dispenser take an average of 12 ounces of soda with a standard deviation of 4 ounces. Assume the amounts taken are normally distributed. There is an 13.6% chance that the next 16 customers will take an average between 13 and 14 ounces. Draw a well-labeled picture of this value.

38. (6 points) By definition, $MSE(\widehat{\Theta}) = E\left[(\widehat{\Theta} - \theta)^2\right]$. Show that this is equivalent to $V(\widehat{\Theta}) + \left[Bias(\widehat{\Theta})\right]^2$.

39. (6 points) Researchers are investigating if an herbal supplement (made with extracts from the *kudzu* vine) can curb binge drinking. They recruited 8 people, who regularly consumed three to four drinks per day, to spend two 90-minute sessions consuming beer and watching TV. After the first session, subjects received capsules of kudzu. Findings showed that 75% of the subjects decreased their amount of beer intake (as compared to the first session levels). Researchers will use this data to test the hypotheses that a majority of users of kudzu would decrease their beer intake: H_0: p =0.50 versus H_A: p > 0.50. Write the appropriate formula—with values plugged in—that would be used for the test statistic and p-value.

40. (6 points) A chemical supply company currently has in stock 100lb of a certain chemical, which it sells to customers in 5-lb lots. Let X = the number of lots ordered by a randomly chosen customer, and suppose that X has pmf

x	1	2	3	4
$P(x)$	.2	.3	.3	.2

Compute the expected number of pounds left after the next customer's order is shipped. (*Hint:* The number of pounds left is a linear function of X.)

41. (14 points) We would like to determine if there is a relationship between educational attainment and marital status. Education is measured as: low = high school or less, moderate = completed college, high = completed graduate school (masters, PhD, or higher). Marital status is measured as: never married, married, divorced, or widowed. SAS output of the data is shown at the right. Use this output to conduct the appropriate hypothesis test. [Note: you do not have to verify or recalculate the values in the output.] Be sure to include all necessary steps and all relevant information in each step. Be sure to include a sentence defining the p-value in context.

The SAS System

The FREQ Procedure

Frequency
Expected

Table of status by education

status	education: high	low	moderate	Total
divorced	6 10.927	15 8.543	9 10.53	30
married	57 54.636	48 42.715	45 52.649	150
never	15 32.781	54 25.629	21 31.589	90
widowed	87 66.656	12 52.113	84 64.232	183
Total	165	129	159	453

Statistics for Table of status by education

Statistic	DF	Value	Prob
Chi-Square	6	96.9616	<.0001
Likelihood Ratio Chi-Square	6	104.8332	<.0001
Mantel-Haenszel Chi-Square	1	0.0216	0.8832
Phi Coefficient		0.4626	
Contingency Coefficient		0.4199	
Cramer's V		0.3271	

Use the following for questions 42 to 45: A study was conducted to examine the relationship between the value of a home (in dollars) and the following independent variables: (1) the size of the home (in square feet), (2) the size of the property the home sits on (in acres), and (3) the interaction between the first two variables. Part of the output based on a random sample of 98 homes in Wake Country NC is given below.

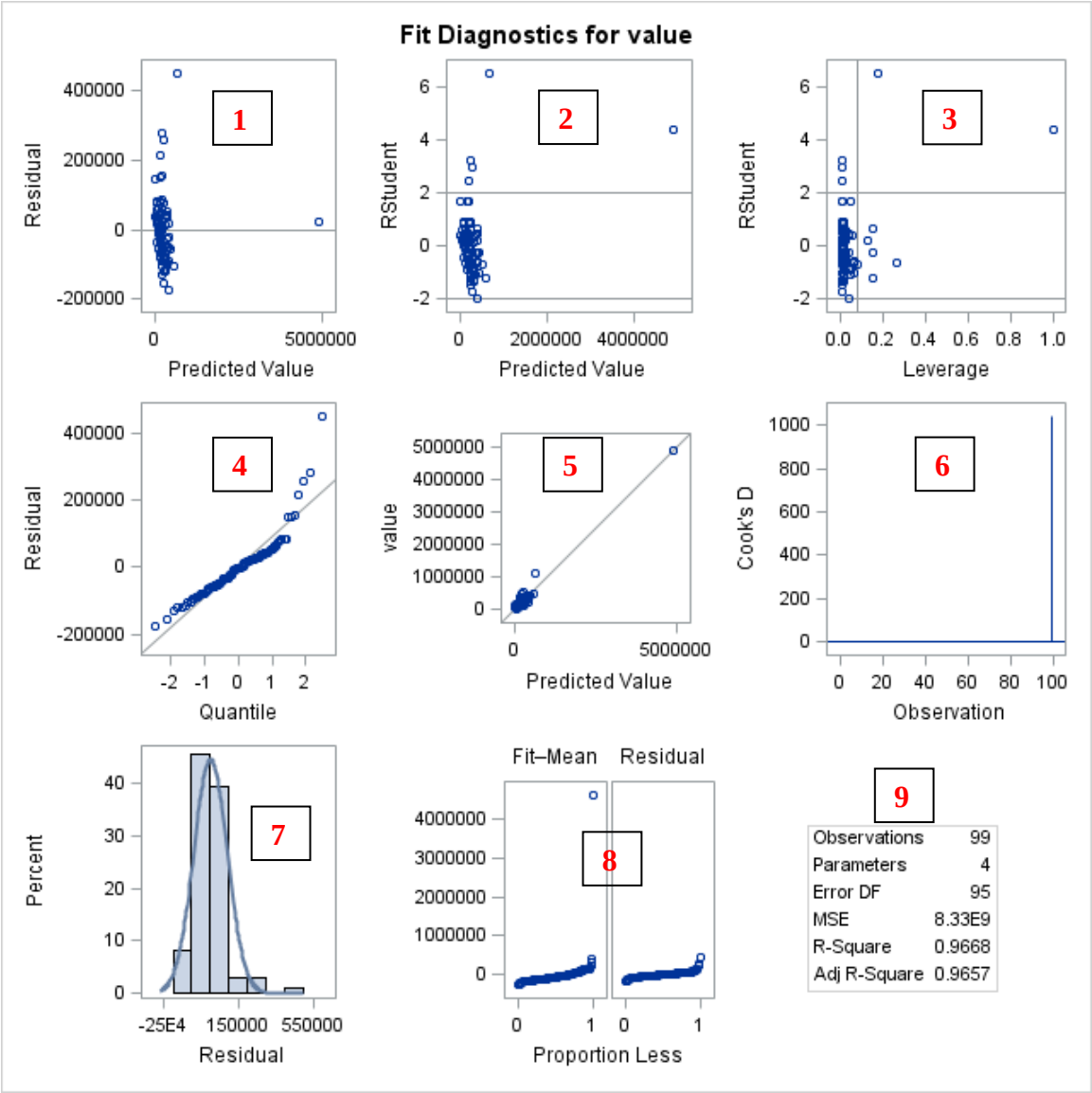
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	-142686	34518	-4.13	<.0001
Sqft	**1**	184.19066	22.45531	8.20	<.0001
Acres	**1**	216992	54551	3.98	0.0001
interaction	**1**	-68.36380	38.76488	-1.76	0.0810

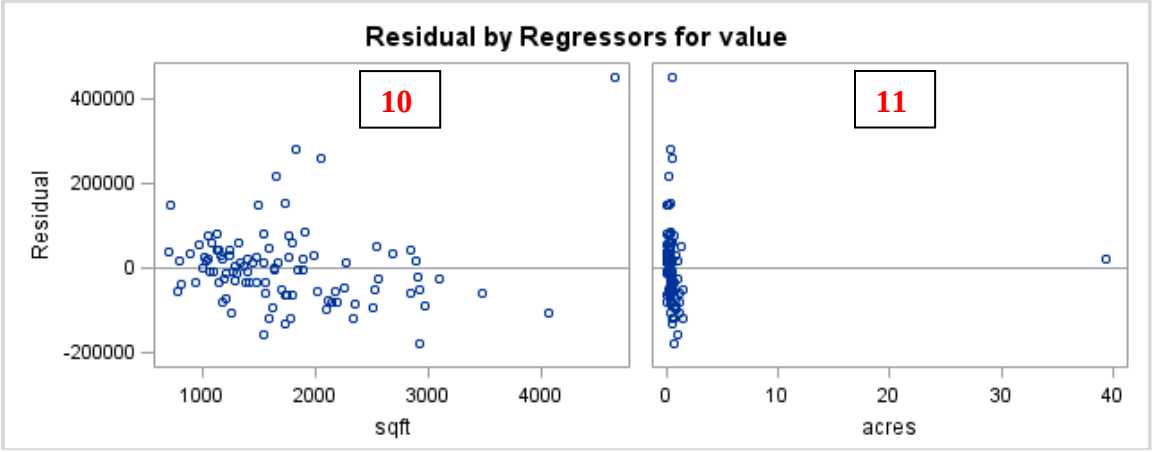
42. (14 points) Conduct the appropriate test to see if there is a significant negative interaction on the value of the home at the 10% level. [Note: you do not have to verify or recalculate the values in the output.] Be sure to include all necessary steps and all relevant information in each step. Be sure to include a sentence defining the p-value in context. Note: you do not need to check the conditions here; this will be addressed in part through question 45.

43. (6 points) What is the predicted value of a 3000 square feet home on 0.5 acres of land?

44. (6 points) A 95% prediction interval for the value of a home with sqft = 2102 and acres = 0.77 is (118086, 483748). Write a sentence interpreting this interval in the context of the problem.

45. (16 points) Residual plots from this model are on the next page. Based on these, is there cause for concern? If so, explain how you know and what you would recommend to fix the problems.


Fit Diagnostics for value
1
Residual
Predicted Value
2
RStudent
Predicted Value
3
RStudent
Leverage
4
Residual
Quantile
5
value
Predicted Value
6
Cook's D
Observation
7
Percent
Residual
8
Fit–Mean
Residual
Proportion Less
9
Observations 99
Parameters 4
Error DF 95
MSE 8.33E9
R-Square 0.9668
Adj R-Square 0.9657


Residual by Regressors for value
10
Residual
sqft
11
acres

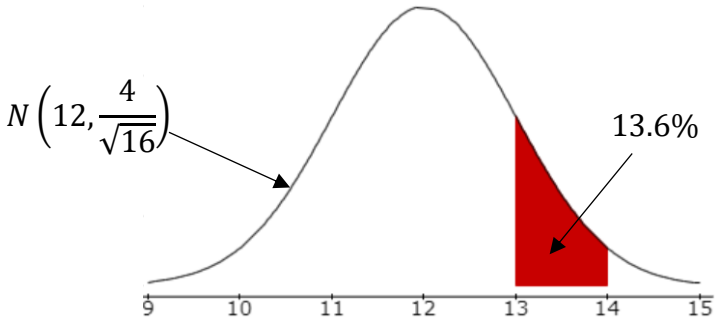
Solutions:

1. False
2. True
3. False
4. False
5. False
6. True
7. True
8. True
9. C
10. A
11. B
12. A
13. B
14. C
15. D
16. C
17. D
18. D
19. A
20. D
21. C
22. D
23. D
24. B
25. B
26. A
27. B
28. A
29. D
30. A
31. A
32. B
33. D
34. C

35. The sampling distribution allows us to quantify the variability in sample statistics. We use the hypothesized mean and estimated standard error of a statistic to calculate the test statistic for the hypothesis test. The distribution of the test statistic (i.e. the null distribution) also derives from the sampling distribution; we then use the null distribution to calculate the p-value, which is the probability of observing a test statistic that is as extreme as or more extreme than our test statistic assuming the null hypothesis is true.

36.
    a. In this context, a Type I Error corresponds to selecting a yellow marble when the basket has 50 green and 50 yellow marbles.
    b. P(Type I Error) = proportion of yellow marbles in the null basket = 50 / 100 = 0.5

37.



38. $MSE(\hat{\theta}) = E\left[\left(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta\right)^2\right]$

$= E\left[\left(\left(\hat{\theta} - E(\hat{\theta})\right) + \left(E(\hat{\theta}) - \theta\right)\right)^2\right]$

$= E\left[\left(\hat{\theta} - E(\hat{\theta})\right)^2 + 2\left(\hat{\theta} - E(\hat{\theta})\right)\left(E(\hat{\theta}) - \theta\right) + \left(E(\hat{\theta}) - \theta\right)^2\right]$

$= E\left(\hat{\theta} - E(\hat{\theta})\right)^2 + 2E\left(\hat{\theta} - E(\hat{\theta})\right)\left(E(\hat{\theta}) - \theta\right) + E\left(E(\hat{\theta}) - \theta\right)^2$

$= E\left(\hat{\theta} - E(\hat{\theta})\right)^2 + \left(E(\hat{\theta}) - \theta\right)^2$

$= V(\hat{\theta}) + \left(Bias(\hat{\theta})\right)^2$

39. Note the small sample size! np_0 = (8)(0.5) = 4 and $n(1\text{-}p_0)$ = (8)(0.5) = 4. Both of these are less than 10, so it would not be appropriate to use the z-test (which assumes a large sample). Instead we use the small sample binomial test from lecture 8.4.

Test statistic = number of kudzu users in the sample who reduced their beer intake
= (8)(0.75) = 6

p-value = $P(X \geq 6) = \binom{8}{6}(0.5)^6(0.5)^2 + \binom{8}{7}(0.5)^7(0.5)^1 + \binom{8}{8}(0.5)^8(0.5)^0$

40. Since orders are sold in 5-lb lots, the weight left = $100 - 5X$
$E(100 - 5X) = 100 - 5E(X)$, where $E(X) = (1)(.2) + (2)(.3) + (3)(.3) + (4)(.2) = 2.5$ lbs
So: $E(100 - 5X) = 100 - (5)(2.5) = 87.5$ lbs

41. H_0: There is a relationship between educational attainment and marital status for the population represented by this sample.

Based on the expected counts, the large sample condition is met. We are not told much else about the sample, so we need to assume it was randomly selected and the counts in each cell are independent in order for the rest of this test to be valid. If this is the case:

Test statistic: $\chi^2 = 96.9616$

Null distribution: χ^2 distribution with $df = (4 - 1)(3 - 1) = 6$

p-value = $P(\chi^2 \geq 96.9616) < 0.0001 \Rightarrow$ Reject the null at the 5% level

There is evidence of a significant relationship between educational attainment and marital status for the population represented by this sample.

42. H_0: $\beta_3 = 0$ vs. H_A: $\beta_3 < 0$, where β_3 represents the true value of the slope coefficient for the interaction between the size of a home and the size of the property.

    Test statistic: $t = -1.76$

    Null distribution: t -distribution with $df = 98 - 3 - 1 = 94$

    p-value $= P(t \leq -1.76) = \frac{0.0810}{2} = 0.0405 \Rightarrow$ Reject the null at the 10% level

    There is evidence of a significant negative interaction between the size of the home and the size of the property on the value of the home.

43. $\hat{y} = -142686 + (184.191)(3000) + (216992)(0.5) - (68.364)(3000)(0.5) = \$415{,}837$

44. We predict* that the value for a house with 2,102 square feet on a 0.77 acre lot will be somewhere between $118,086 and $483,748.

    * Note: here you could also say "We believe" or "We are 95% confident"; the key with the interpretation of a prediction interval is that you talk about an *individual* rather than a *mean*.

45. Note: There are a few potential problems you can comment on here, including
    - There are 2 potential outliers, one of which is influential as indicated by the large Cook's distance (plot 6); these points (or at least one of them) can be seen in plots 1, 2, 3, 4, 5, 8, & 11.
        - Fix: Look into the observations and try to figure out why they are so unusual. If there is a typo in the data, correct it; if not, try fitting the model without the point(s) and report the results of both models.
    - There may be problems with increasing variance in the relationship between sqft and value, as indicated by plot 10.
        - Try a transformation of the variable or the response (though I would deal with the outlier and reevaluate the fit of the model first).
    - The QQ-plot (plot 4) and histogram of residuals (plot 7) show problems with the assumption of normally distributed residuals.
        - Do nothing. The sample size is large enough that we don't need to be concerned about non-normality (thanks to the CLT).