STAT 200 – Exam 2 – Practice Exam Solutions

1. C – 75% of the data falls between the lower quartile and the maximum because 25% of the data falls between each number in a five-number summary (minimum, lower quartile, median, upper quartile, maximum).

2. B – The mean and median are measures of location (center) of a dataset, while the IQR, standard deviation, and range are all measures of spread (variability).

3. A – The mean would be most affected by an outlier because it is a sensitive statistic. The median, upper quartile, and lower quartile are all resistant.

4. D – We need to use $\hat{p}$ rather than p because we are dealing with a sample of 50 adults rather than an entire population.

   $\hat{p}$ = Number in category / Total number
   $\hat{p}$ = 30/50
   $\hat{p}$ = 0.60

5. D – μ is the correct notation because we are dealing with the mean of an entire population.

6. A – Bar graph: Favorite season: (spring, summer, fall, winter)
   B – Histogram: (Number of classes skipped in a semester)
   C – Boxplot: (Hours per day spent studying) and (class standing)

7. B – The variation in the middle 50% of the data spans 4 hours per day. The IQR is from 2 hours per day to 6 hours per day, so the middle 50% of the data spans 4 hours per day.

8. D – 75%. Approximately 75% of the values will fall between Q_1 and the maximum value.

9. D – 75%. Approximately 75% of the values will fall between the minimum and Q_3.

10. B – 2 hours per day. Approximately 75% of the values lie above Q_1.

11. B – The median for method 2 is smaller than it is for method 1. You can see from the side-by-side boxplots that method 2 has a larger median than method 1.

12. A – Categorical. The explanatory variable in this example is gender, which is a categorical variable.

13. B – Quantitative and continuous. The response variable is height in inches, which is a quantitative and continuous variable.

14. B – The median is larger for males. The line in the middle of the "box" portion of the boxplot represents the median. You can see that the median for males is larger than the median for females.

15. C – Both datasets are skewed; however, they are skewed in opposite directions. The median for males is closer to the top of the box, so we know it is skewed to the left. The median for females is closer to the bottom of the box, so we know it is skewed to the right.

16. C – Mean < Median. This dataset will be skewed to the left because the salaries of the three administrative assistants are much lower than the salaries of the lawyers. The smaller salaries of the administrative assistants will have a greater effect on the mean than they will on the median because of the difference in the way the two values are calculated. The mean will always be less than the median when a dataset is skewed to the left.

17. C – 40

    $IQR = Q_3 - Q_1 = 195 - 155 = 40$

18. D – Range is the only sensitive statistic in the list. All of the other answer choices are resistant statistics.

19. C – 7 hours because it is the median value. 50% of values fall above and 50% of values fall below the median.

20. A – 10 is Q1 in the five-number summary. 25% of values fall below Q1 and 75% of values fall above Q1. Since the problem said **at least** you are looking for the values between Q1 and the maximum.

21. B – 10 is once again Q1 in the five-number summary. Since the problem said **at most** you are looking for the values between the minimum and Q1.

22. B – If a dataset is skewed to the left, it is likely that the mean is less than the median. The mean is sensitive, so it is pulled toward the side of the skewedness. The median is resistant, so it is always the middle point of a dataset.

23. A – The association is a positive linear association because it has a positive slope and a straight line fits the best through the points.

24. B – The best straight line through the data is horizontal because there is no relationship, positive or negative, between the two variables.

25. A – The answer is –0.9 because you are looking for the value with an absolute value closest to 1. Make sure to note this problem gives you correlations (r) which can range from –1 to 1. Any value not within this range cannot be the value of the correlation.

26. #1 = –0.75
    #2 = –0.02
    #3 = –0.73
    #4 = –0.99

    Scatterplot #2 has essentially no linear relationship. That is why the correlation is almost 0.

    Scatterplot #4 has almost a perfect negative linear relationship. That is why the correlation is almost –1.

    Scatterplots #1 and #3 are a little trickier. You know than the correlation for #3 will be closer to 0 than the correlation for #1 because they have the same pattern; however, #3 has more observations. When the pattern stays constant and the number of observations increases, correlation will move closer to 0.

27. B – We can use the Empirical (95%) Rule to find the range of ages that would include about 95% of the residents in the building.

    95% Rule – If a distribution is bell-shaped, about 95% of the data falls within two standard deviations from the mean, in either direction.

    95% of data: Mean – 2 * St Dev *to* Mean + 2 * St Dev
    46.5 – 2 * 11.2 *to* 46.5 + 2 * 11.2 **= 24.1 *to* 68.9**

28. A – 56/1626 – The risk that a person who took malaria pills **does** get malaria = Number of people who took malaria pills and did get malaria/Total number of people who took malaria pills

29. B – (70/650)/(56/1626) – The problem tells us we want "those who took the placebo **compared to** those who took the malaria pills." So we need to take the risk of the placebo group divided by the risk of the malaria pills group.

30. A – People taking the placebo are 3.13 times <u>more</u> likely to get malaria than those who received the malaria pills.

31. C – 70/580

    Odds = # in category / # <u>not</u> in category
    Odds = # who took placebo and got malaria / # who took placebo and did <u>not</u> get malaria
    Odds = 70/580

32. Mean = (5+10+3+14+12+1+22+17)/8 = **10.5**

    Median: Remember to list the values in order first: 1, 3, 5, 10, 12, 14, 17, 22

    Since there are an even number of observations, we end up with 10 and 14 as the middle number. We need to find the average of the two middle numbers.

    Median **=** (10+12)/2 = **11.0**

33. C – Standard deviation accurately measures spread regardless of the shape of the data. This statement is **false** because standard deviation accurately measures spread only when dealing with a standard normal distribution.

34. B – 230 to 270
    95% of data is within 2 standard deviations from the mean
    250 - 2(10) = 230
    250 + 2(10) = 270

35. C – If your score is the 30th percentile, then 30% of the class had scores at or <u>below</u> yours. A percentile always represents the percentage at or below a certain value.

36. B – 40%

The vertical axis shows us the frequency of each "bin" of the histogram, and the horizontal access tells us the pop-quiz scores.

Number of observations at or below 60 = 1 + 1 + 2 = 4
Total number of observations = 1 + 1 + 2 + 3 + 2 + 1 = 10

Percentage at or below 60 = 4/10 = .4 = 40%

37. A – 40/200

Because this question asks for the proportion **of males** who had been arrested 1 or 2 times, we take the number of males who have been arrested 1 or 2 times over the **total number of males**.

38. B – 40/400

Because this question asks for the proportion **of students in the study** who were males who had been arrested 1 or 2 times, we take the number of males who have been arrested 1 or 2 times over the **total number of students in the study**.

39. D – 85/420

Risk = (Number in category of interest) / (Total number in group)
Risk **of women** having a heart attack = 85 / 420

40. B – (150/380) / (85/420)

Relative Risk = (Risk for group 1) / (Risk for group 2)

Risk for males = 150/380
Risk for females = 85/420

Because this question asks for the relative risk **of males** having a heart attack **compared to females**, we will call males group 1 and females group 2.

Relative Risk = (150/380) / (85/420)

41. B – 150/380

    Odds = (Number in category of interest) / (Number <u>not</u> in category of interest)
    Odds **<u>for males</u>** = Males who've had a heart attack / Males who haven't had a heart attack

    Odds = 150/380

42. D – It is <u>false</u> that any observation below 3 would be an outlier. In this five-number summary, 3 is the first quartile. That means that 25% of the dataset will fall below 3.

    We can also verify that A, B, and C are true:

    A – IQR = Q3 – Q1 = 14 – 3 = 11

    B – Range = Maximum – Minimum = 74 – (-3) = 77

    C – While the exact requirements for a number to be considered an outlier are not typically covered in this course, we can see that 74 is way higher than any of our other values. This extreme difference makes it clear that 74 is an outlier.

43. C – The vertical axis shows us the frequency of each "bin" of the histogram, and the horizontal access tells us the pop-quiz scores.

    Number of observations above 60 = 3 + 2 + 1 = **6**
    Total number of observations = 1 + 1 + 2 + 3 + 2 + 1 = 10

    Percentage above 60 = 6/10 = .6 = 60%

44. B – 59.1 is the minimum height

45. C – 62.9 to 79.2; 75% of the data lies between Q1 and the maximum. We know this because 25% of the dataset is between each of the numbers in a five-number summary. Since Q1 and the maximum are three numbers apart, about 75% of the dataset would fall between them.

46. A – 25% percent of data falls between Q1 and the median.

47. D – The slowest time of 47 minutes is an outlier because it appears as an asterisk (*) on the boxplot.

48. A – This dataset appears to be skewed to the left because the median is closer to the top of the box.

    Note that we are not considering the outlier because this question tells us to base our answer only on the box portion of the dataset.

49. A – This histogram is positively (right) skewed because it trails off to the right.

50. A – When a dataset is skewed to the right, we expect the mean to be greater than the median. This is because the mean is sensitive and pulled toward the skewedness, while the median is resistant.

51. B – Earlier years tend to have a higher number of injuries among football players. However, we cannot draw a cause-and-effect conclusion from a scatterplot.

52. #1 = –0.61
    #2 = –0.02
    #3 = –0.96
    #4 = 0.76

    Scatterplot #4 has the only positive linear relationship. That is why the correlation is positive.

    Scatterplot #3 has almost a perfect negative linear relationship. That is why the correlation is almost –1.

    Scatterplot #2 has essentially no linear relationship. That is why the correlation is almost 0.

    Scatterplot #1 has a negative linear relationship, but it is a weaker relationship than we saw in Scatterplot #3. That is why the correlation is the weaker negative correlation.

53. C – Discrete – How many times you have skipped class this semester?
    D – Continuous – How long does it take you to walk to your first class?
    B – Nominal – What is your major?
    A – Ordinal – What is your class year?

54. B – This is an observational study because the researcher is not applying a treatment. He is simply sampling people who do drink alcohol and people who do not drink alcohol about their fast food consumption. When dealing with observational studies, we cannot conclude causation.

55. C – A simple random sample could be created by numbering the students in random order and choosing the first 50 students. The key here is that the students are numbered in <u>random order</u>, so the first 50 names on the list are put there randomly.