Stat 2470, 1/13 Discussion Questions Name _____

Instructions: Attempt to answer these questions by reading the textbook or with online resources before coming to class on the date above.

1. What are the steps involved in constructing a stem-and-leaf display? (also called a stemplot)

① Sort the data ② select one or more leading digits to be stems ③ list possible stem values in a vertical column (do not omit values) ④ record the leaf (trailing digits) for each observation beside corresponding stem ⑤ indicate units (in a key) for stems & leaves in display.

2. What are some advantages of stem-and-leaf displays over other types of displays of data?

easy to construct
preserves original data

3. What is the difference between a discrete and a continuous variable?

discrete variable can only take on certain values (like integers)
continuous variable can be any real # value in an interval

4. What are the steps to constructing a histogram?

① group the data into classes of a fixed width (between 5 & 20 depending on the quantity of data ② count the # of data values in each class ③ use the frequency as the y-value to build a bar for each class.

5. What are the two y-axis types that can be used in a histogram?

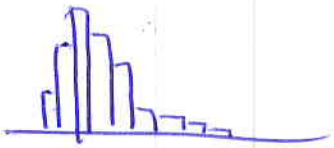
frequency (count)
relative frequency (%)

6. Why are equal class widths so important in constructing histograms? What are the downsides of using unequal class widths?

So that the only difference between classes is the height. Under classes change the area and so usually distort the data.

7. Describe the general shapes of histograms. Draw a sketch of each type.

roughly symmetric



right skewed

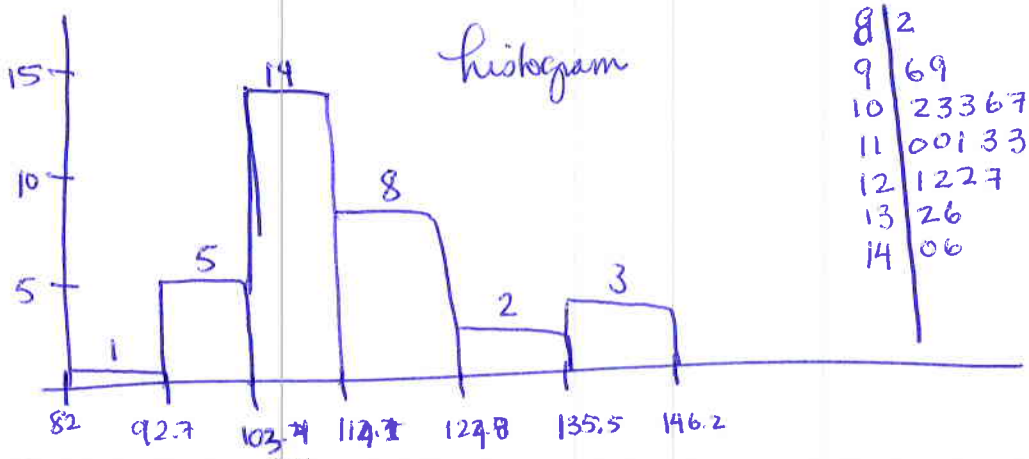
left skewed

8. How do you tell the difference between graphs that are right-skewed vs. left-skewed?

tail determines Skewness. long left tails pull mean below median; long right tails pull mean above median

9. Practice drawing a histogram and a stemplot on the data in problem #28 in section 1.2.

146 − 82 = 64 6 classes → Class width 10.7 [82, 92.7), [92.7, 103.4), etc



Stemplot

8	2
9	69
10	2336788889
11	0013333555889
12	1227
13	26
14	06

Key 8|2 = 82

10. What is the formula for calculating the mean? Describe in words the steps to calculate the formula by hand, and how to do it in the calculator.

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Add up all the data values & divide the sum by the total # in the sample

11. What is the procedure for finding the median by hand? How does it differ for a list with an even number of terms, vs. a list with an odd number of terms? How can you find the mean with the calculator?

① Sorta data ② choose the middle value (if # in list is odd, then median is the $\frac{n+1}{2}$ datapoint. If # in list is even, then average the two values closest to middle (on either side of $\frac{n+1}{2}$)

1Var Stats – enter data into list; then run 1Var Stats. (in TI-84)
mean is $\bar{X}$ & Med is further down.

12. What are the notations for the mean, median, mode and trimmed mean? Label each and say how to read them. How do you tell them apart?

mean $\bar{x}$, median $\tilde{x}$, trimmed mean $\bar{x}_{tr\%}$ ← % of trimming

13. How do you calculate a trimmed mean?

multiply % of trimming by n (sample size). This is # to trim off both ends. if # is an integer, trim & recalculate the mean. If # is not an integer, then trim the data by the integer less & the integer more. Then weight the values according to the decimal value.

14. How do you calculate a trimmed mean when the number of terms you need to trim off is not a whole number of terms?

See above.

15. Why is categorical data treated by using sample proportions instead of means?

you can calculate a numerical value from a list of non-numerical values. cat, dog, bird don't have an average.

16. What is the range of a dataset? How do we find it?

the range is the largest value in the dataset minus the smallest value.

17. What is the formula for the variance of a sample? Describe at least two different formulas for finding it.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = E(X^2) - [E(X)]^2$$

18. How is the variance related to the standard deviation? Which value in the calculator is the sample standard deviation?

s^2 is variance

$s = \sqrt{s^2}$ is the standard deviation

in calculator (II-84) S_x is the sample standard deviation σ_x is the population standard deviation (use that only when doing a census, otherwise, use S_x) found w/ 1Var Stats

19. How is the population variance different? Why two different formulas for the sample and the population?

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

Sample is an estimate so dividing by n-1 allows for more variability than a sample usually provides

20. How do we find the first and third quartiles of a dataset?

- find the median. Divide the dataset in two ① below the median ② above the median. Find the median of each half.

21. The five-number summary is often used to draw a quick boxplot (box-and-whisker plot). What numbers are in the five-number summary?

Min, Q_1, Median, Q_3, Max.

22. Our book refers to f_s as the "fourths spread", but other books call this the IQR or the "inter-quartile range". How is this calculated?

$Q_3 - Q_1$

23. How is the IQR used to find outliers for a boxplot? How are mild and extreme outliers different?

1.5 IQR (or $1.5 f_s$) from the nearest quartile is an outlier (further from that). 3 IQR ($3 f_s$) from nearest quartile (further out) is an extreme outlier.

24. Comparative plots (boxplots, stemplots and histograms) are used throughout this course to compare datasets. Why is this useful? What features can be compared easily this way without extensive computation?

Spread and distribution can be easily compared.

do they take the same values? are their modes (max height) in the same place, similar symmetry? range?