Class: Statistics

Name: William Horn

Date: Sept 3, 2017

Assignment: Homework 1

All calculation done using python.

Problems:

2, 9, 12abc, 15, 22, 33ab, 34ab, 42, 46, 58

Question 2:

Problem: For each of the following hypothetical populations, give a

plausible sample of size 4

a. all distances that might result when you throw a football.

answer:

20 yards, 30 yards, 15 yards, 5 yards

b. all companies listed on the New York stock exchange.

answer:

Apple, Google, GMC, Microsoft

c. all students at your college or university

answer:

William Horn, Monica Mikes, Brady Ingersoll, Trenton Friend

d. all grade point averages of students at your college or

university

answer:

2.3, 4.0, 3.8, 3.5

Question 9:

problem: in a famous experiment carried out in 1882, Michelson and

Newcomb obtained 66 observations on the time it too for light to travel

between two locations in a Washington, D.C. a few of the measurements

(coded in a certain manner) were 31, 23, 32, 36, -2, 26, 27 and 31.

a. why are these measurements not identical?

answer: whatever instrument they were using to measure the

elapsed time could have had some inaccuracy.

b. is this an enumerative study? why or why not?

answer: no, the population is the time it took for the light to

travel between the two locations. they only got to measure 66

times, out of an infinite sample, so the study is analytical.

Question 12:

problem: the accompanying summary data on ceo2 particle size (nm) under

certain experimental conditions were read from a graph in the article

"nanoceria energetics of surfaces, interfaces and water absorption"

3.0-<3.5 3.5-<4.0 4.0-<4.5 4.5-<5.0 5.0-<5.5

5 15 27 34 22

5.5-<6.0 6.0-<6.5 6.5-<7.0 7.0-<7.5 7.5-<8.0

14 7 2 4 1

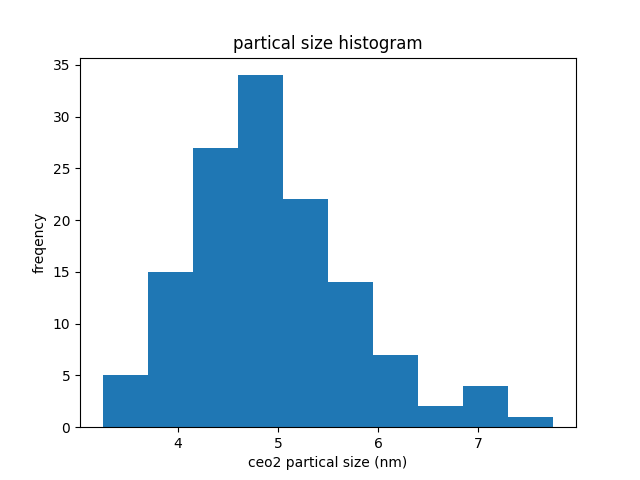
a. What proportion of the observations are less than 5?

answer: 34 + 27 + 15 + 5 = 81

b. What proportion of the observations are less than 6?

answer: 7 + 2 + 4 + 1 = 14

c. Make a histogram with relative frequency on y-axis and comment on its features.



answer: Most of the data is towards the lower forth, making

this histogram into have a skewed right shaped.

Question 15:

problem: Do running times of American movies differ somehow from running

times of French movies? The author investigated this question by randomly

selecting 25 recent movies of each type, resulting in the following

running times

Am: 94 90 95 93 128 95 125 91 104 116 162 102 90 110 92 113 116 90

97 103 95 120 109 91 138

Fr: 123 116 90 158 122 119 125 90 96 94 137 102 105 106 95 125 122

103 96 111 81 113 128 93 92

Construct a stem an leaf plot with the stems in the middle and

having the Am. leaves go out to the left and the Fr. leaves

out to the right. Comment on some of the interesting features.

answer:

- The American movies seem to be skewed towards 90 giving the

graph a reversed J shape.

- The French movies peak in two places (90, 120 minutes) so the

plot has a bimodal shape.

- Also, I think that a larger sample size would help increase

the accuracy of these plots as well possibly even changing

there shapes.

Question 22:

problem: How does the speed of a runner vary over the course of a

marathon? Consider determining both the time to run the first 5km and the

time to run between the 35km and 40km points and then subtracting the

former time from the latter time. A positive value of this difference

corresponds to a runner slowing down toward the end of the race. The

accompanying histogram is based on time of runners who participated in

several different Japanese marathons

What are some interesting features of this histogram? What is a typical

difference value? Roughly what proportion of the runners ran the late

distance more quickly than the early distance?

answer: Almost all of the runners in the sample slowed down during the

later stages of the marathon, with the most common proportion being

around 150. Only around 5-10 runners have a negative proportion, which

from the histogram looks to be less then 1%.

Question 33:

problem: The following home sale amounts were reported for a sample of

homes in Alameda, CA that were sold the previous month (in 1000s of $)

590 815 575 608 350 1285 408 540 555 679

- a. Calculate and interpret the sample mean and median.

answer: mean = 640.5, median = 582.5.

Because of the outlier 1285, the mean get pulled up higher then the

median, which is not really affected.

- b. Suppose the 6th observation had been 985 rather than 1285. How

would the mean and median change?

answer: The mean would be decreased because 985 < 1285 but because 985

is above the upper half, the median would not be affected.

Question 34:

problem: The following data on concentration (EU/mg) in settled dust

for one sample of urban homes and another of farm homes was kindly

supplied

Urban: 6.0 5.0 11.0 33.0 4.0 5.0 80.0 18.0 35.0 17.0 23.0

Farm: 4.0 14.0 11.0 9.0 9.0 8.0 4.0 20.0 5.0 8.9 21.0 9.2 3.0 2.0 0.3

- a. Determine the mean for each sample, how do they compare?

answer:

Farm mean = 8.56

Urban mean = 21.54

Per the sample given, urban homes have a much higher

concentration then farm homes.

- b. Determine the sample median for each sample? How do they compare?

Why is the urban median so different from the mean of that sample?

answer:

Farm median = 8.9

Urban median = 17.0

The farm median is quite like its mean, but urban mean and median

are quite different. Because the urban data has the outlier 80 in its

sample, its mean get pulled up. The farm doesn't have such an extreme

outlier so it doesn't.

Question 42:

problem:

- a. If a constant c is added to each xi in a sample, yielding yi =

xi + c, how do the sample mean and median of the y\_is relate to the

mean and median of the x\_is? Verify your conjectures.

answer:

All the values in the sample are growing proportionally, so the

median will only go up or down by c. However, the mean takes into

account all the values, each one of which got c added to it. So the

sum(y\_is) = sum(x\_is) + (n \* c) where n is the number of values in

the sample. So when c > 0 and n != 0, sum(y\_is) > sum(x\_is) making

the mean mean(y\_is) > mean(x\_is). The opposite is true when c is

negative.

- b. If each x\_i is multiplied by a constant c, yielding y\_i =

c\*x\_i, answer the question of part (a). Again, verify your

conjecture.

answer:

sum(y\_is) = sum(x\_is) \* c

diff = mean(y\_si) - mean(x\_si)

= (1 / n) \* (sum(x\_is) \* c - sum(x\_is))

= (sum(x\_is) / n) \* (c - 1)

The median will be changed by a factor of c, while the mean will

change by a factor of c - 1.

Question 46:

problem:

The data is observations about the temps of pvc frames, there are 3

groups cooler, control, warmer

cooler: 1.59 1.43 1.88 1.26 1.91 1.86 1.90 1.57 1.79 1.72 2.41 2.34

0.83 1.34 1.76

control: 1.92 2.00 2.19 1.12 1.78 1.84 2.45 2.03 1.52 0.53 1.90

warmer: 2.57 2.60 1.93 1.58 2.30 0.84 2.65 0.12 2.74 2.53 2.13 2.86

2.31 1.91

answer:

- a. Compare measures of center for the three different samples.

cooler: mean = 1.706, median = 1.76

control: mean = 1.75272727273, median = 1.9

warmer: mean = 2.07642857143, median = 2.305

the temperature seems to increase from cool to control to warmer,

along with the variability. Through each of the 3 tests the

difference between the mean and the median increases.

- b. Calculate interpret and compare the standard deviations for the

three different samples.

cooler: standard dev = 0.40

control: standard dev = 0.53

warmer: standard dev = 0.77

Throughout the cooler, control, warmer, the standard deviation is

increasing, which means that the values in each of the samples

are becoming increasingly spread apart.

- c. Do the fourth spreads for the three samples convey the same

message as do the standard deviations about relative

variability?

cooler: fs = 0.39

control: fs = 0.365

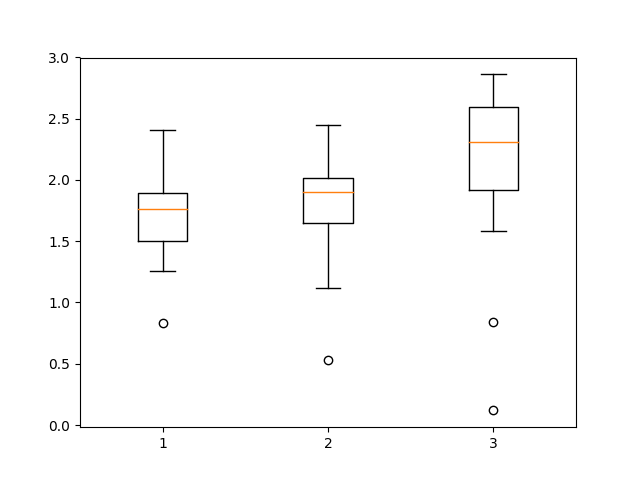
warmer: fs = 0.6775

Overall yes, but the control does not show as high a variability

as it did in the standard deviation.

- d. Construct a comparative boxplot and comment on any interesting

features.



cooler control warmer

Warmer has 2 outliers, with cooler and control having only one.

Also, the spread between Q1 and Q3 on the warmer sample is larger

with cooler and control being roughly the same.

Question 58:

problem:

A company utilizes two different machines to make parts. During a

single shift, a sample of n=20 parts was produced by each, and a

critical dimension was measured for all parts made. Compare and

contrast the boxplots (in book).

answer:

While machine one has an outlier, that is only because the spread

is so small. Comparing machine one to machine two, even though

machine one has the outlier, that parts made are more consistent

overall. Which would most likely be more desirable.