245: Fundamentals of Statistics

Jianqing Fan — Frederick L. Moore'18 Professor of Finance

Problem Set #1 Fall 2021

Due Wednesday, September 22, 2021 at 11:30pm on Canvas

Please include your written answers, R code, and figures (when applicable). Please do not write your name in your homework submission. Upload your assignment as a single PDF file on Canvas with the format "PS1_XX.pdf", where XX is your assigned anonymous ID. Find your anonymous ID number in the 'Grades' tab on Canvas. Look at the 'Score' column of the 'Anonymous ID for Homeworks' row.

1. Consider the population consisting of all computers of a certain brand and model, and focus on whether a computer needs service while under warranty.

   (a) Pose two probability questions based on selecting a sample of 100 such computers.

   (b) What inferential statistics question might be answered by determining the number of such computers in a sample of size 100 that need warranty service?

2. A study of the relationship between age and various visual functions (such as acuity and depth perception) reported the following observations on the area of scleral lamina (mm^2) from human optic nerve heads:

   ```
   2.75 2.62 2.74 3.85 2.34 2.74 3.93 4.21 3.88
   4.33 3.46 4.52 2.43 3.65 2.78 3.56 3.01 3.00
   ```

   (a) Compute the median and IQR of the data by hand.

   (b) Calculate $\sum x_i$ and $\sum x_i^2$.

   (c) Use this to calculate mean and standard deviation.

   (d) Suppose that there are 39 such observations. By accident, the value 3.93 is recorded as 39.3. How much does this affect the average? How much does this affect the median?

3. In order to investigate whether "Poor gets poorer and rich gets richer", download the income data from tax records in year 2000 from
   `http://fan.princeton.edu/fan/classes/245.html`.
   Draw the histogram of the distribution using the break points (in thousand dollars):
   ```
   > breaks = c(0,5,10,15,20,25,30,40,50,75,100, 200, 500,1000)
   ```

   (a) What is the Shape of the distribution?

   (b) Which interval is more crowded? (10, 15) or (50, 75)?

   (c) Which interval has more families? (10, 15) or (50, 75)?

   (d) What is the density of the block (50, 75)?

   (e) How many percents of families have income between \$50K and \$110K?

4. Consider adjusted closing prices of SP500 index, IBM stock, Apple Inc., and Johnson & Johnson from Jan. 1, 2000 to September 8, 2016 from `http://fan.princeton.edu/fan/classes/245.html`.

   (a) Give a time series plot of the prices of Apple Inc. and Johnson & Johnson.

   (b) Compare the distributions of their returns of these 4 assets by a boxplot.

   (c) Compute their summary statistics using function `summary()`

   (d) Compute their interquartile ranges (IQRs) and their standard deviations. What are the rank of the risk profiles of those stocks? (These are used to validate the Capital Asset Pricing Model).

   Note: All the problems in parts (b)–(d) refer to the log-returns of the stock instead of stock prices.

5. Simulate a pool of size 900 from a large population with 45% supporting candidate A by using the R function `rbinom(900,1,0.45)`.

   (a) Execute the above program 10 times manually and report those 10 poll errors, which is sample proportion -0.45. What are the ranges of such errors?

   (b) Do this 1000 times using R-program and records the poll errors automatically by running

```
x=NULL                              #creating data structure

for (i in 1:1000) x=c(x,mean(rbinom(900,1,0.45)))
             #compute sample proportion and padding the results

x = x - 0.45 # compute poll errors
```

   Summarize those poll errors using R-function `summary()`. What are the average and the standard deviation of those errors?

   (c) Plot the histogram of the poll errors in (b). Try both `hist(x,freq=F,col="blue")` and `hist(x,freq=T,col="blue", main="Histogram of poll errors", xlab="poll errors")` What is the shape of the distribution?