Homework #1 – Due Thursday, Jul. 12, before class

STAT-UB.0001 – Statistics for Business Control

Some of the homework assignments will involve data files. All data files will be available on NYU Classes under the "Datasets" section. You are permitted to work in teams, but each student must independently write up their own solutions (no direct copying).

Electronic submissions are not accepted, but you can write your solutions on a laptop and print it out. Print out and turn in your Minitab plots. See the "Minitab Tips" on NYU Classes for tips on opening files and printing or saving plots.

Problem 1

Much of our understanding of prehistoric peoples comes from caves, such as cave paintings, evidence of fire pits, etc. made nearly 40,000 years ago. Based on these, people tend to believe that most of prehistoric peoples lived in caves for most of their lives. Do you see any problem in this?

Solution: Prehistoric people are associated with caves because that is where the data still exists, not necessarily because most of them lived in caves for most of their lives. If there had been paintings on trees, animal skins or hillsides, they would have been washed away long ago. Similarly, evidence of fire pits, middens, burial sites, etc. are most likely to remain intact to the modern era in caves.

. . . . . .

Problem 2

Consider these values:

$$21 \quad 22 \quad 18 \quad 25 \quad 27 \quad 9 \quad 21 \quad 34 \quad 20 \quad 16$$

Find the mean, median, mode, and standard deviation for these. You need to compute these values by hand. This is for you to get familiar with those formulas.

Solution: The mean is 21.3. Sort the numbers from low to high:

$$\{9, 16, 18, 20, 21, 21, 22, 25, 27, 34\}$$

The median is the average of 5-th and 6-th number, thus 21. The mode is 21 (occurs twice). The sample variance is

$$\begin{aligned} s^2 = \frac{1}{10-1}\Big[&(9-21.3)^2 + (16-21.3)^2 + (18-21.3)^2 \\ &+ (20-21.3)^2 + (21-21.3)^2 + (21-21.3)^2 \\ &+ (22-21.3)^2 + (25-21.3)^2 + (27-21.3)^2 + (34-21.3)^2\Big] \\ &= 44.46 \end{aligned}$$

The sample standard deviation is $s = \sqrt{44.4556} = 6.67$.

. . . . . .

Problem 3

The file `HeightWeight.csv` contains data on 200 records of human heights and weights of 18 years old children. Here, we focus on the the *Weight* (in Pounds) column.

1. Use *Stat* ⇒ *Basic Statistics* ⇒ *Display Descriptive Statistics* to compute the mean, median and interquartile range, for the *Weight*. You will need to obtain the inter-quartile range by hand as the difference between the third quartile (Q_3, the 75th percentile) and the first quartile (Q_1,the 25th percentile).

   Solution: The inter-quartile range is $136.17 - 119.88 = 16.29$.

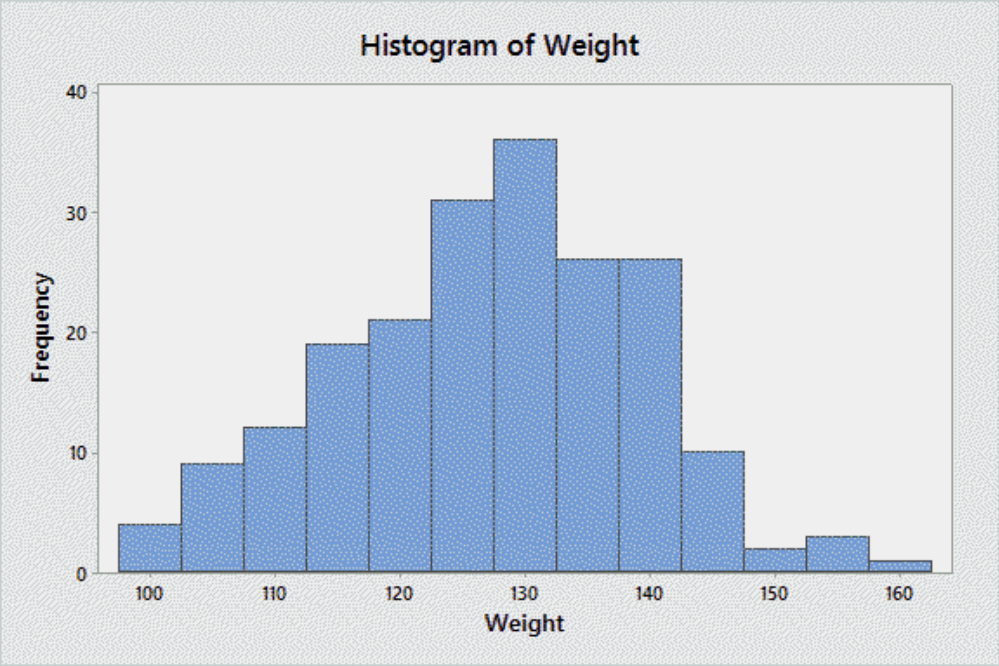
   Descriptive Statistics: Weight

   Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Weight	200	0	127.22	0.846	11.96	97.90	119.88	127.88	136.17	158.96

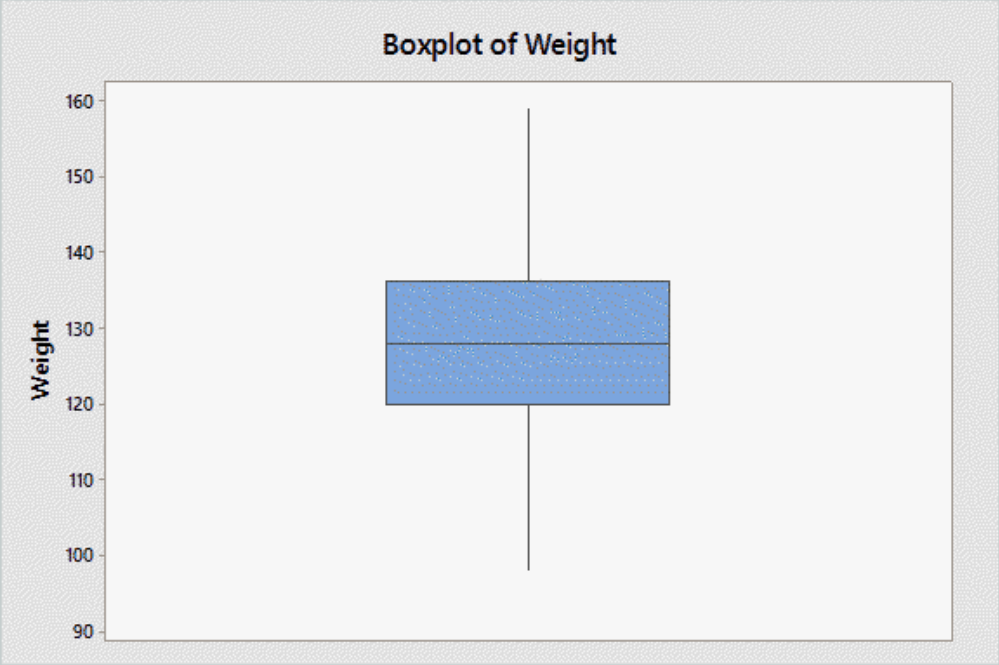
2. Make a histogram of *Weight*. Does the data seem to have a reasonably bell-shaped distribution?

   Solution: It looks reasonably bell-shaped.



3. Make a boxplot of *Weight*. Do you see any outliers?

   Solution: There is no outliers according to the boxplot.



4. Use empirical rules to complete the statements:

   - For approximately 68% individuals, the weight is between [?, ?]
   - For approximately 95% individuals, the weight is between [?, ?]
   - For approximately 99.7% individuals, the weight is between [?, ?]

   Solution:

   - For approximately 68% individuals, the weight is between $[115.26, 139.18]$
   - For approximately 95% individuals, the weight is between $[103.3, 151.14]$
   - For approximately 99.7% individuals, the weight is between $[91.34, 163.1]$

5. (Optional) Look at the data. What are the true percentages in those intervals you just computed? Do they agree with the empirical rules (roughly)?

   Solution: The true percentages are very close to those given by the empirical rules:

   - For 66.5% individuals, the weight is between $[115.26, 139.18]$
   - For 94.5% individuals, the weight is between $[103.3, 151.14]$
   - For 100% individuals, the weight is between $[91.34, 163.1]$

. . . . . .

Problem 4

A World Cup 2018 quarter-final between Sweden and England will happen on Jul 7. Suppose Sweden will score 0-3 goals with equal probability, and same for England.

1. What is the sample space for this quarter-final?

   Solution: The sample space for this quarter-final is all the possible scores for (Sweden : England),

   $$\{(0:0),\ (0:1),\ (0:2),\ (0:3),\\ (1:0),\ (1:1),\ (1:2),\ (1:3),\\ (2:0),\ (2:1),\ (2:2),\ (2:3),\\ (3:0),\ (3:1),\ (3:2),\ (3:3)\}$$

2. What is the probability that Sweden scores more goals than England?

   Solution: There are a total of 16 sample points in sample space, each has probability $\frac{1}{16}$. The event "Sweden scores more goals than England" contains the following 6 sample points:

   $$\{(1:0),\ (2:0),\ (3:0),\ (2:1),\ (3:1),\ (3:2)\}.$$

   Thus the probability is $\frac{6}{16} = \frac{3}{8}$.

. . . . . .

Problem 5

Suppose among all soccer fans, 12% support Sweden, 23% support England, and 4% support both. A soccer fan is to be selected at random.

1. What is the probability that the soccer fan supports either Sweden or England (or both)?

   Solution: Let A="Supports Sweden", B="Supports England". Then by additive rule,

   $$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = 0.12 + 0.23 - 0.04 = 0.31.$$

2. What is the probability that the soccer fan supports neither team?

   Solution: The event "Supports neither team" is the complement of "Supports either Sweden or England (or both)". Thus by comlement rule,

   $$\mathbb{P}((A \cup B)^c) = 1 - \mathbb{P}(A \cup B) = 1 - 0.31 = 0.69.$$

3. What is the probability that the soccer fan supports Sweden but not England?

   Solution: Let C="Supports Sweden but not England", then $A = C \cup (A \cap B)$. Note that C and $A \cap B$ are mutually exclusive. Therefore,

   $$\mathbb{P}(C) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = 0.12 - 0.04 = 0.08.$$

   If any of these is not clear, drawing a diagram would help you.

. . . . . .

Problem 6

Suppose the probability that a World Cup game will have more than 3 goals scored is 30%, and the probability that the game will have more than 4 goals scored is 20%. For a World Cup game that has more than 3 goals scored, what is the probability of it having more than 4 goals scored?

Solution: The question as asking about the conditional probability $\mathbb{P}(\text{scores } 4_+ \mid \text{scores } 3_+)$. By the definition of conditional probability,

$$\begin{aligned}\mathbb{P}(\text{scores } 4_+ \mid \text{scores } 3_+) &= \frac{\mathbb{P}(\text{scores } 4_+ \cap \text{scores } 3_+)}{\mathbb{P}(\text{scores } 3_+)} \\ &= \frac{\mathbb{P}(\text{scores } 4_+)}{\mathbb{P}(\text{scores } 3_+)} \\ &= \frac{0.2}{0.3} = \frac{2}{3}\end{aligned}$$

. . . . . .