Statistics Homework

September 4, 2020

[1]:
```
import pandas as pd
import math
import matplotlib.pyplot as plt
```

1 14

1.1 Construct Stem and Leaf Display

[2]:
```
data = pd.read_csv("14.csv")
#print(data)

min_index = int(math.floor(min(data["'Rate'"])))
max_index = int(math.floor(max(data["'Rate'"])))
stem_data = {}
for i in range(min_index,max_index+1):
    stem_data[str(i)]=[]
for r in data["'Rate'"]:
    stem = str(int(math.floor(r)))
    stem_data[stem].append(r-math.floor(r))
#print(stem_data)
for e in stem_data:
    out_str = e
    if len(e)<2:
        out_str+=" "
    out_str+=": "
    for v in stem_data[e]:
        out_str+=" " + str(round(v*10,0))
    print(out_str)
```

```
2 :  3.0 2.0
3 :  4.0 7.0 3.0 9.0 5.0 4.0 6.0 7.0 2.0 8.0
4 :  6.0 0.0 8.0 3.0 8.0 1.0 5.0 9.0
5 :  1.0 1.0 6.0 8.0 0.0 4.0 0.0 4.0 5.0 0.0 6.0 1.0 6.0 5.0 9.0 7.0 0.0
6 :  7.0 9.0 4.0 2.0 6.0 4.0 5.0 3.0 2.0 0.0 9.0 6.0 1.0 0.0 7.0 2.0 4.0 6.0 9.0
8.0 9.0 2.0 0.0 3.0 0.0
7 :  1.0 0.0 5.0 5.0 6.0 3.0 5.0 5.0 6.0 2.0 2.0 4.0 3.0 0.0 5.0 8.0 0.0
```

```
8 :  0.0 8.0 3.0 2.0 4.0 4.0 3.0 2.0
9 :  2.0 6.0 8.0 3.0 2.0 0.0 5.0 3.0 7.0 6.0 3.0 6.0 3.0 8.0 1.0
10:  5.0 4.0 8.0 3.0 4.0 2.0 5.0 8.0 4.0 6.0
11:  5.0 2.0 9.0 3.0 9.0 9.0 3.0
12:  3.0 7.0
13:  8.0
14:  3.0 6.0
15:  0.0 3.0 5.0 0.0
16:
17:
18:  9.0
```

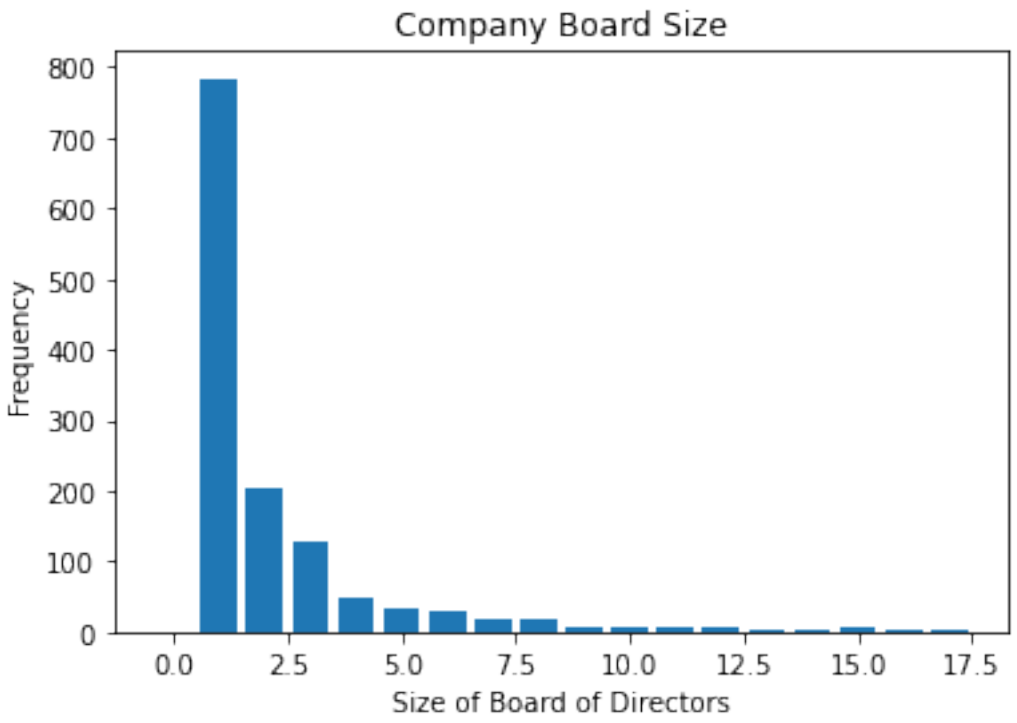
1.2 b. What is a typical, or representative, flow rate?

A typical flow rate is 6 L/min ## c. Does the display appear to be highly concentrated or spread out? The flow rate is concentrated around 6 L/min ## d. Does the distribution of values appear to be reasonably symmetric? If not, how would you describe the departure from symmetry? The data has a skew to towards the higher flow rates. ## e. Would you describe any observation as being far from the rest of the data (an outlier)? the 18.9 L/min entry is an outlier.

2 18. Every corporation has a governing board of directors. The number of individuals on a board varies from one corporation to another. One of the authors of the article "Does Optimal Corporate Board Size Exist? An Empirical Analysis" (J. of Applied Finance, 2010: 57–69) provided the accompanying data on the number of directors on each board in a random sample of 204 corporations.

2.1 a. Construct a histogram of the data based on relative frequencies and comment on any interesting features.

[3]:
```
data = pd.read_csv("18.csv")
plt.bar(data["'Number o'"],data["'Frequenc'"])
plt.title("Company Board Size")
plt.xlabel("Size of Board of Directors")
plt.ylabel("Frequency")
plt.show()
```



The board size centers around 1 and is skewed to the right. The distribution is unimodial and the board size appears to follow Zipf's law. ## c. What proportion of these corporations have at most 10 directors?

[4]:
```
total = sum(data["'Frequenc'"])
less_then_10 = 0
for i in range(0,10+1):
    less_then_10+=data["'Frequenc'"][i]
prop = float(less_then_10)/float(total)
print(str((prop*100.0))+"% of coroprations have less then 10 directors")
```

```
97.55725190839695% of coroprations have less then 10 directors
```

3 22. How does the speed of a runner vary over the course of a marathon (a distance of 42.195 km)? Consider determining both the time to run the first 5 km and the time to run between the 35-km and 40-km 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 times of runners who participated in several different Japanese marathons ("Factors Affecting Runners' Marathon Performance," Chance, Fall, 1993: 24–30). 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 han the early distance

The speed of a runner slows down as the marathon continues. The histogram is unimodal with a skew to the right. A typical difference value is 150 units (The graph in the book failed to include units or a title). Roughly 5% of the runners ran more quickly in the beginning.

4 38. Blood pressure values are often reported to the nearest 5 mmHg (100, 105, 110, etc.). Suppose the actual blood pressure values for nine randomly selected individuals are [118.6, 127.4,138.4,130.0,113.7,122.0,108.3,131.5,133.2]

4.1 a. What is the median of the reported blood pressure values?

[5]:
```
def median(l):

    l.sort()
    list_len = len(l)
    if list_len %2 ==0:
        return (l[int((list_len)/2)-1]+l[int((list_len)/2)])/2
    else:
        return l[int((list_len)/2)]
def mean(l):
    return float(sum(l))/float(len(l))
print("Median Blood pressure is "+str(median([118.6, 127.4,138.4,130.0,113.
 ↪7,122.0,108.3,131.5,133.2]))+" mmHg.")
```

```
Median Blood pressure is 127.4 mmHg.
```

4.2 b. Suppose the blood pressure of the second individual is 127.6 rather than 127.4 (a small change in a single value). How does this affect the median of the reported values? What does this say about the sensitivity of the median to rounding or grouping in the data?

[6]:
```
med = median([118.6, 127.6,138.4,130.0,113.7,122.0,108.3,131.5,133.2])
old = median([118.6, 127.4,138.4,130.0,113.7,122.0,108.3,131.5,133.2])
print("The median increased by "+str(round(med-old,1)))
```

```
The median increased by 0.2
```

That tells me that the median is not very sensitive to changes in outliers.

5 42.

5.1 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 yis relate to the mean and median of the xis? Verify your conjectures.

The sample mean and median will rise by c. For example given the list [1,2,3,4,5,6] the mean is

[7]:
```
print(mean([1,2,3,4,5,6]))
```

```
3.5
```

and the median is

[8]:
```
print(median([1,2,3,4,5,6]))
```

```
3.5
```

If c=1 the mean is:

[9]:
```
print(mean([2,3,4,5,6,7]))
```

```
4.5
```

and the median is:

[10]:
```
print(median([2,3,4,5,6,7]))
```

```
4.5
```

6 44 (a,b,c) Poly(3-hydroxybutyrate) (PHB), a semicrystalline polymer that is fully biodegradable and biocompatible, is obtained from renewable resources. From a sustainability perspective, PHB offers many attractive properties though it is more expensive to produce than standard plastics. The accompanying data on melting point (°C) for each of 12 specimens of the polymer using a differential scanning calorimeter appeared in the article "The Melting Behaviour of Poly(3-Hydroxybutyrate) by DSC. Reproducibility Study" (Polymer Testing, 2013: 215–220).

180.5 181.7 180.9 181.6 182.6 181.6 181.3 182.1 182.1 180.3 181.7 180.5 ## a. Compute the sample range.

[11]:
```
L = [180.5,181.7,180.9,181.6,182.6,181.6,181.3,182.1,182.1,180.3,181.7,180.5]
print("The range is "+str(round(max(L)-min(L),1)))
```

```
The range is 2.3
```

6.1 b. The sample variance s 2 from the definition [Hint: First subtract 180 from each observation.]

[12]:
```
def variance(L):
    m = mean(L)
    return sum(map(lambda x: (x-m)**2,L))/(len(L)-1)
v = variance(L)
print("The median is "+str(round(v,2)))
```

```
The median is 0.52
```

6.2 c. The sample standard deviation

[13]:
```
print(" The sample standard deviation is "+str(round(math.sqrt(v),2)))
```

```
 The sample standard deviation is 0.72
```

7 50 In 1997 a woman sued a computer keyboard manufacturer, charging that her repetitive stress injuries were caused by the keyboard (Genessy v. Digital Equipment Corp.). The injury awarded about $$ 3.5 million for pain and suffering, but the court then set aside that award as being unreasonable compensation. In making this 45 determination, the court identified a "normative" group of 27 similar cases and specified a reasonable award as one within two standard deviations of the mean of the awards in the 27 cases. The 27 awards were (in 1000s) 37, 60, 75, 115, 135, 140, 149, 150, 238, 290, 340, 410, 600, 750, 750, 750, 1050, 1100, 1139, 1150, 1200, 1200, 1250, 1576, 1700, 1825, and 2000. What is the amount that could be awarded under the two- standard deviation rule?

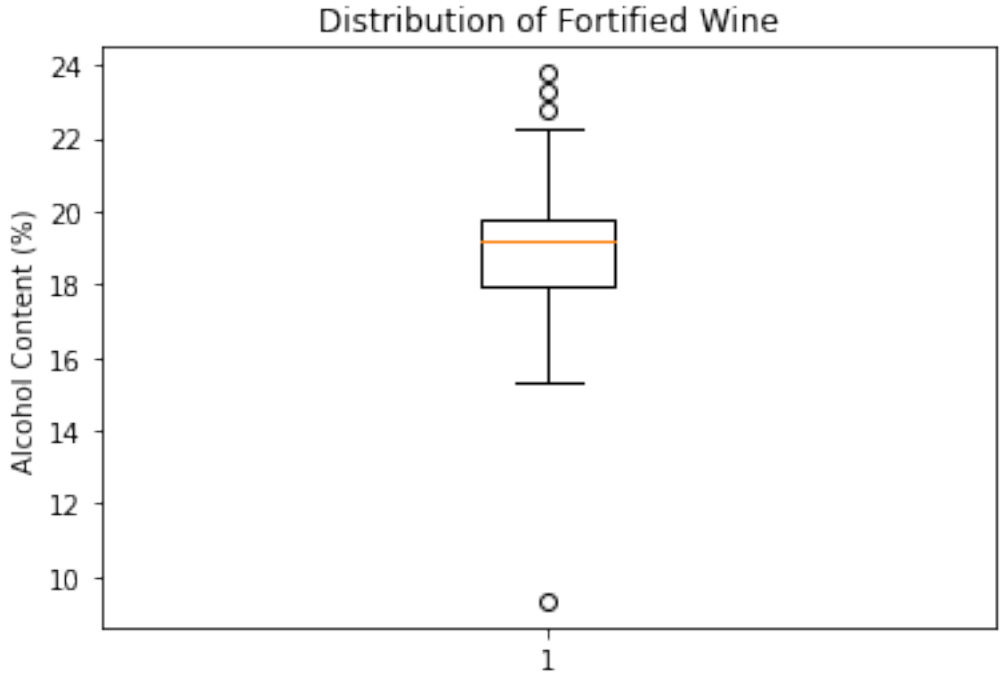
[14]:
```
std = math.sqrt(variance([37, 60, 75, 115, 135, 140, 149, 150,
                          238, 290, 340, 410, 600, 750, 750, 750, 1050, 1100]))
print("The maximum that can be awarded is $"+str(round(1000*std,-2)))
```

```
The maximum that can be awarded is $348100.0
```

8 56 The following data on distilled alcohol content (%) for a sample of 35 port wines was extracted from the article "A Method for the Estimation of Alcohol in Fortified Wines Using Hydrometer Baumé and Refractometer Brix" (Amer. J. Enol. Vitic., 2006: 486–490). Each value is an average of two duplicate measurements. Use methods from this chapter, including a boxplot that shows outliers, to describe and summarize the data

[15]:
```
alcohol = [16.35,18.85,16.2
,17.75,19.58,17.73,22.75,23.78,23.25,19.08,19.62,19.2,20.05,17.85,19.17,19.
↪48,20,19.97,17.48,17.15
,19.07
,19.9
,18.68
,18.82
,19.03
,19.45
,19.37
,19.2
,18
,19.6
```

```
,9.33,21.22
,19.5,15.3,22.25]
plt.boxplot(alcohol)
plt.title("Distribution of Fortified Wine")
plt.ylabel("Alcohol Content (%)")
plt.show()
```



9 Chapter 2.

10 4 Each of a sample of four home mortgages is classified as fixed rate (F) or variable rate (V).

10.1 a. What are the 16 outcomes in S ?

F F F F

F F F V

F F V F

F F V V

F V F F

F V F V

F V V F

F V V V

V F F F

V F F V

V F V F

V F V V

V V F F

V V F V

V V V F

V V V V ## b. Which outcomes are in the event that exactly three of the selected mortgages are fixed rate? V F F F

F V F F

F F V F

F F F V ## c. Which outcomes are in the event that all four mortgages are of the same type? F F F F

V V V V ## d. Which outcomes are in the event that at most one of the four is a variable-rate mortgage? F F F F F F F V

10.2 e. What is the union of the events in parts (c) and (d), and what is the intersection of these two events?

$$c \cup d = \begin{pmatrix} FFFF \\ VVVV \\ FFFV \end{pmatrix}$$

$$c \cap d = \begin{pmatrix} FFFF \end{pmatrix}$$

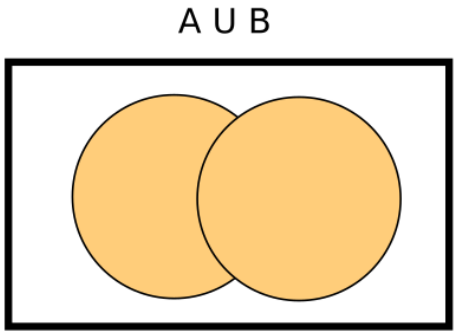
10.3 f. What are the union and intersection of the two events in parts (b) and (c)?

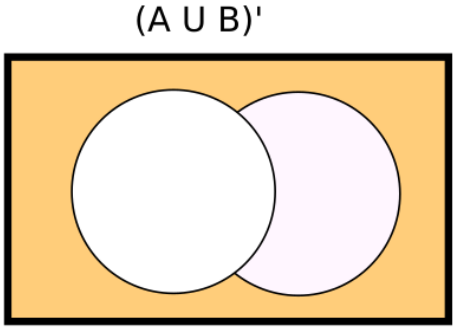
$$b \cup c = \begin{pmatrix} VFFF \\ FVFF \\ FFVF \\ FFFV \\ FFFF \\ VVVV \end{pmatrix}$$

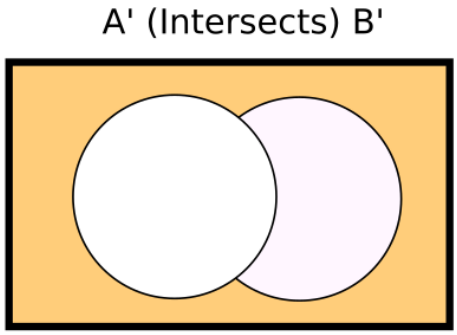
$$b \cap c = \Rightarrow ($$

9 Use Venn diagrams to verify the following two relationships for any events A and B (these are called De Morgan's laws): ## a.

$$(A \cup B)' = A' \cap B'$$



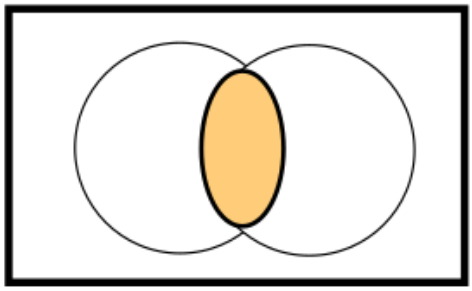




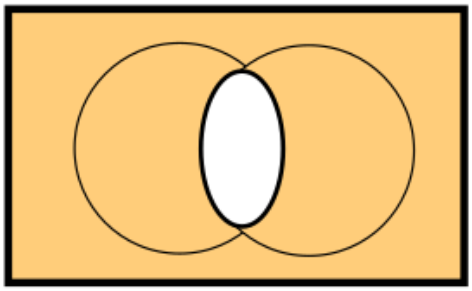
10.4 b.

$$(A \cap B)' = A' \cup B'$$

A (Intersects) B



(A (Intersects) B)'



A' U B'