606_HW_2

Salma Elshahawy

9/5/2019

Chapter_2 - Summarizing Data

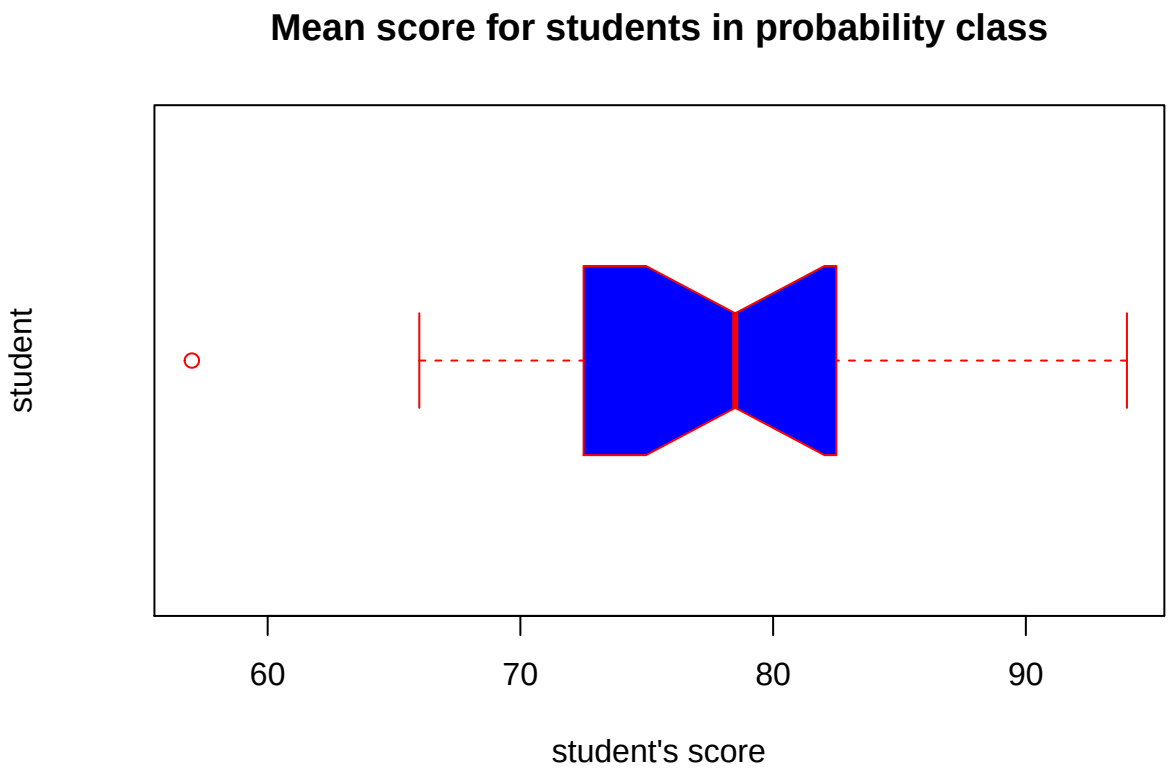
1. Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

```
library(ggplot2)
student_df <- data.frame("SN" = 1:20,"score" = c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81
(student_df)
```

```
##    SN score
## 1   1    57
## 2   2    66
## 3   3    69
## 4   4    71
## 5   5    72
## 6   6    73
## 7   7    74
## 8   8    77
## 9   9    78
## 10 10    78
## 11 11    79
## 12 12    79
## 13 13    81
## 14 14    81
## 15 15    82
## 16 16    83
## 17 17    83
## 18 18    88
## 19 19    89
## 20 20    94
```

Drawing boxplot

```
b <- boxplot(student_df$score, main = "Mean score for students in probability class", xlab = "student's s
```



Pulling out the boxplot data to compare with the given summary

```
b
```

```
## $stats
##      [,1]
## [1,] 66.0
## [2,] 72.5
## [3,] 78.5
## [4,] 82.5
## [5,] 94.0
##
## $n
## [1] 20
##
## $conf
##          [,1]
## [1,] 74.96701
## [2,] 82.03299
##
## $out
## [1] 57
```

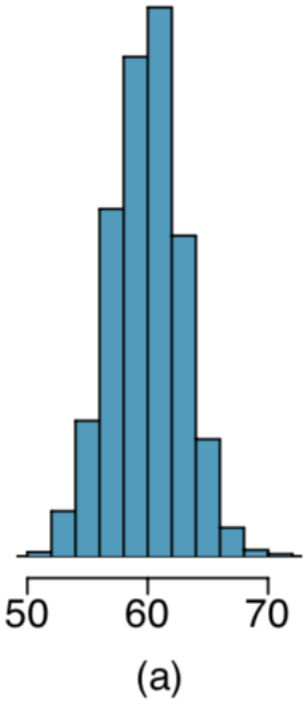
```
##
## $group
## [1] 1
##
## $names
## [1] ""
```

From the observation we can conclude that: 1. Q1 = 72.5 2. Q2(median) = 78.5 3. Q3 = 82.5 4. Max = 94 5. Min = 66

Outlaier point is 57 with index(1) in the dataframe.

---

2. Mix-and-match: Describe the distribution in the histograms below and match them to the box plots.



a.

This graph has a symmetric, **single-peaked(unimodal)** distribution where the histogram forms an approximate mirror imagewith respect to the center of the distribution.

This histogram is match with boxplot number #2

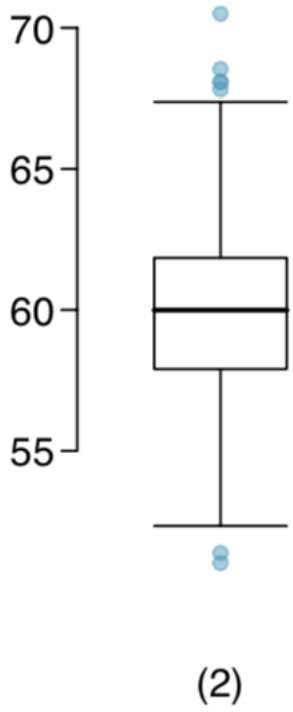
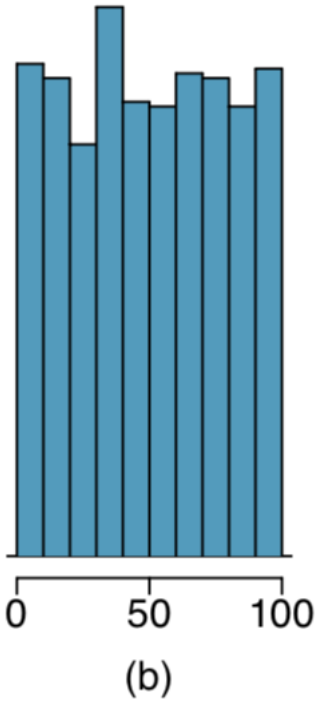


Figure 1: boxplot matching_2



b.

This graph has a symmetric, bimodal **uniform** distribution.

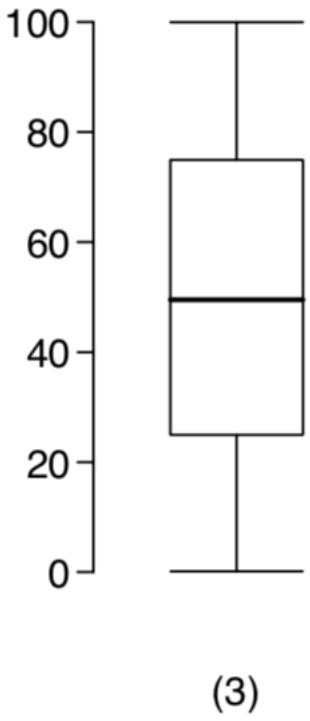


Figure 2: boxplot matching_3

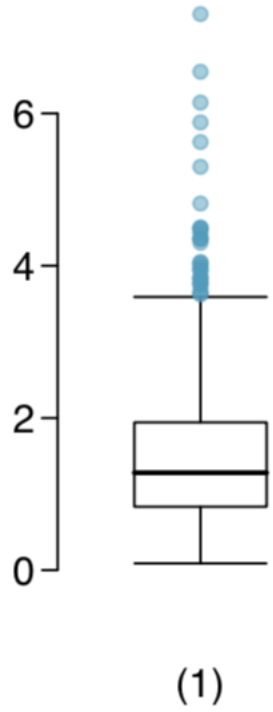
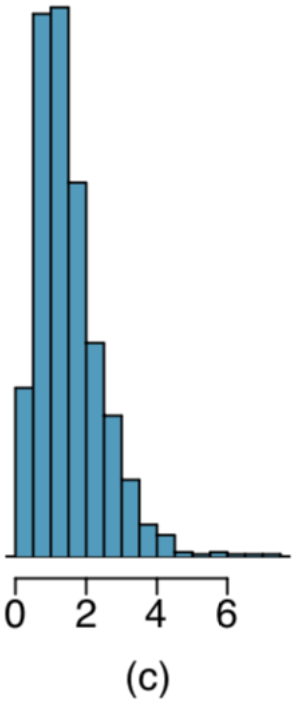


Figure 3: boxplot matching_1



c.

This graph is **skewed-right** distribution.

Distributions and appropriate statistics, Part II

a. Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000.

Skewed-Left Distribution, mean < median. The distribution of house prices are likely left skewed as there is a natural boundary at 0 and meaningful number of houses cost more than $6M. Therefore the center would be best described by the median, and variability would be best described by the IQR.

b. Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000.

Skewed-Right Distribution. The distribution of house prices are likely right skewed as there is a natural boundary at 0 and only a few number of houses cost more than $1.2M. Therefore the center would be best described by the median, and variability would be best described by the IQR.

c. Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

The distribution of number of alcoholic drinks consumed is likely **right skewed** as there is a natural boundary at 0 and only a few drinks are allowed. Therefore the center would be best described by the median, and variability would be best described by the IQR.

d. Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than the all other employees.

The distribution of annual salaries are likely **left skewed** as there is a natural boundary at 0 and only a few people have much higher salaries. Therefore the center would be best described by the median, and variability would be best described by the IQR. The IQR is a much better measure of variability in the amounts earned by nearly all of employees. The standard deviation gets affected greatly by the two high salaries, but the IQR is robust to these extreme observations.

Heart transplants:

a. Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.

The variables survival time and transplant are not independent. The difference in the survival time between who got transplant or not was not due to chance, and heart transplant affected the rate of survival time.

b. What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.

It suggestes that the median survival time for patient who had treatment is much higher than median for the control group. In addition, the max survival time for patients who had treatment is over 1500 days compared to the control group who lived around 100 days without treatment.

c. What proportion of patients in the treatment group and what proportion of patients in the control group died?

```
#propotion of patients died in thetreatment group
prop_treat = 45/69
prop_treat
```

```
## [1] 0.6521739
```

```
prop_cont = 30/34
prop_cont
```

```
## [1] 0.8823529
```

d. One approach for investigating whether or not the treatment is effective is to use a randomization technique.

i. What are the claims being tested?

Heart transplant increase lifespan

---

ii. We write alive on ***28*** cards representing patients who were alive at the end of the study, and dead on ***75*** cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size ***69*** representing treatment, and another group of size ***34*** representing control. We calculate the difference between the proportion of dead cards in the treatment and control groups (treatment - control) and record this value. We repeat this 100 times to build a distribution centered at ***around 0***. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are **due to chance**. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis should be rejected in favor of the alternative.

---

iii. What do the simulation results shown below suggest about the effectiveness of the transplant program?

We can conclude that the data provide strong evidence that the transplant provides a longer lifespan in this clinical setting.