CUNY SPS DATA606 HW2

Chinedu Onyeka

9/11/2021

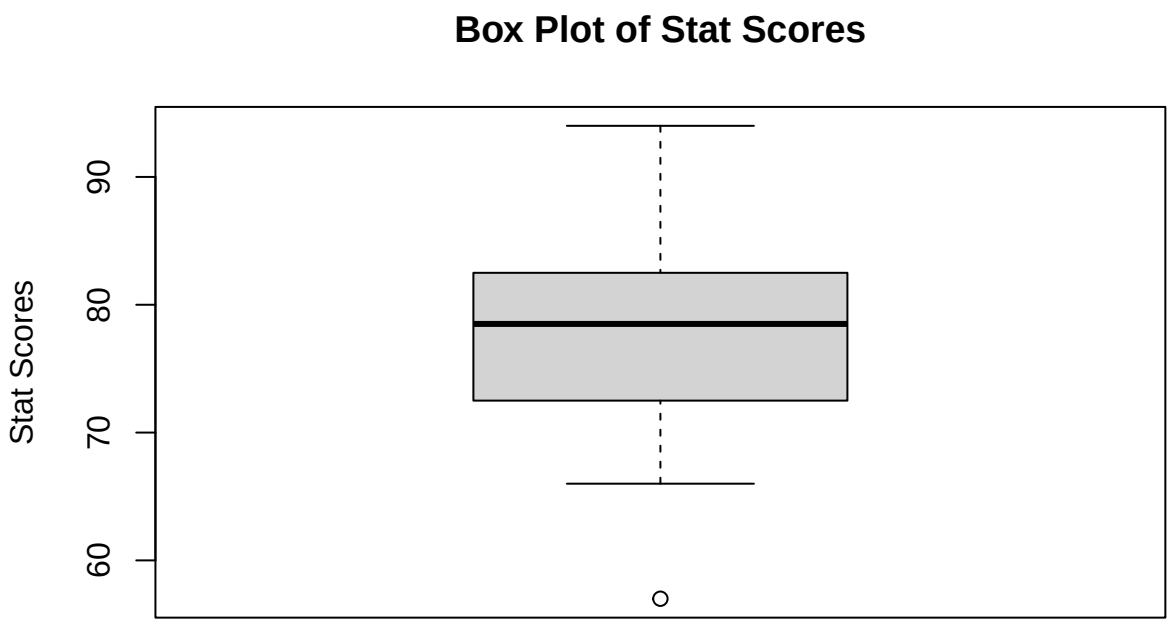
Stats scores. (2.33, p. 78) Below are the final exam scores of twenty introductory statistics students.

57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

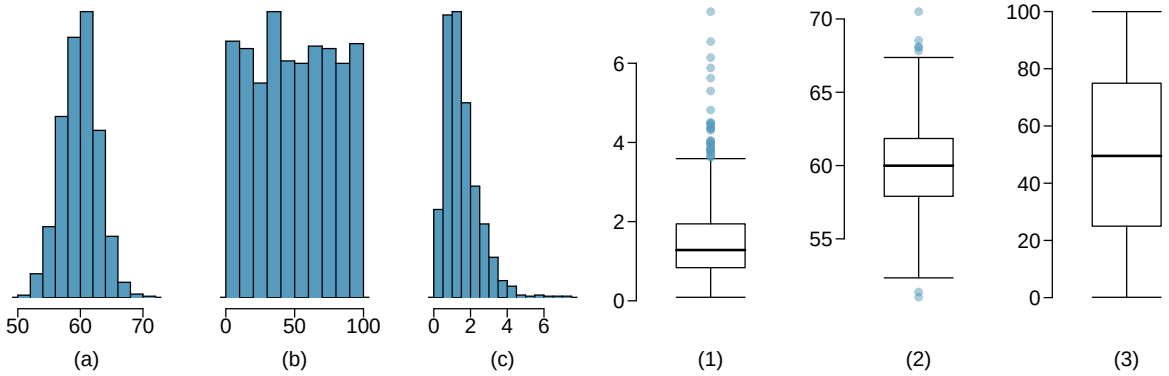
Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

Solution 1 (2.33, p. 78):



Mix-and-match. (2.10, p. 57) Describe the distribution in the histograms below and match them to the box plots.



Solution 2 (2.10, p. 57)

a) Histogram (a) matches boxplot (2). The histogram shows a symetric distribution that resembles that of a normal distribution.

b) Histogram (b) matches boxplot (3). The histogram shows a distribution that resembles a uniform distribution.

c) Histogram (c) matches boxplot (1). The histogram shows a distribution that is skewed to the right.

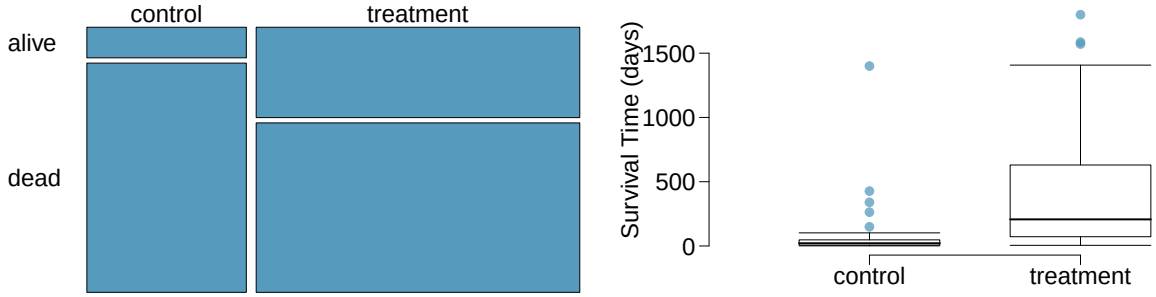
Distributions and appropriate statistics, Part II. (2.16, p. 59) For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(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.
(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.
(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.
(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.

Solution 3 (2.16, p. 59)

(a) - *Right Skewed, use Median and IQR:* The values of the Q1, Q2, and Q3 show that the data is skewed to the right and also a meaningful number of houses cost more than $6million. This information shows that there are a lot of outliers to the right and to get a good representation of the observation, the mean would not be appropriate as it will drag the prices higher because of outliers. Hence, I will use the median to represent this observation and use the IQR to explain the variability since the standard deviation will also be affected by outliers.
(b) - *Symmetric, use Mean and Standard Deviation :* This data is symmetricas can be seen from the Q1, Q2, and Q3 values and only very few houses are above $1.2 million. For this data, I will use the mean (mean = median = mode for symmetric distributions) to represent the observation and use the standard deviation for the variability since there are only few outliers
(c) - *Right Skewed, use Median and IQR:* Since most of the students don't drink, a bigger number of observations will be to the left of the distribution with a few data points on the right for those who drink excessively which will eventually skew the distribution to the right. Also for this data, I will use the median to represent the data so that the few who drink excessively would not end up describing the data for the majority who do not drink. Furthermore, the IQR would be used to explain the variability of the data for the same reason.
(d) - *Right Skewed, use Median and IQR:* A few high level executives who earn much higher salaries will skew the distribution to the right. As a result of this few outliers, the mean would be inappropriate to describe the data. Hence, I will use the median to describe the data and use the IQR to explain the variability in order to minimize the impact of the outliers who earn a much higher salary.

Heart transplants. (2.26, p. 76) The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable *transplant* indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Of the 34 patients in the control group, 30 died. Of the 69 people in the treatment group, 45 died. Another variable called *survived* was used to indicate whether or not the patient was alive at the end of the study.

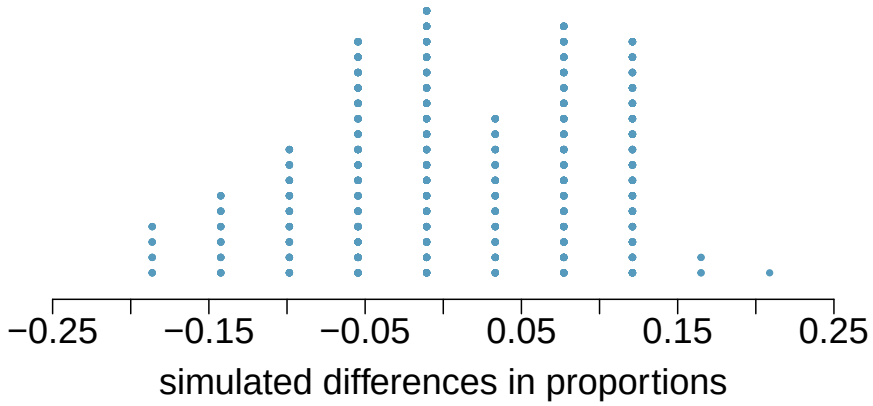


(a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.
(b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.
(c) What proportion of patients in the treatment group and what proportion of patients in the control group died?
(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?
ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.

We write *alive* on _____ cards representing patients who were alive at the end of the study, and *dead* on _____ cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size _____ representing treatment, and another group of size _____ 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 _____. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are _____. 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?



Solution 4 (2.26, p. 76):

(a) Based on the plot, it appears that survival is dependent on whether the patients got a heart transplant since the treatment group has more proportion of patients that survived when compared to the control group.

(b) From the boxplot, it can be seen that the median survival time for the treatment group is much higher than the median survival time of the control group.Hence, the boxplot suggests that the heart transplat treatment appear to be effective since those in the treatment group has a significantly higher survival time when compared to those in the control group.

(c) The proportion of patients in the treatment group that died is $45/69 = 0.65$, while the proportion of patients in the control group that died is $30/34 = 0.88$

(d) (i) The claims being tested is that patients are more likely to survive or have a higher survival time if they take the heart transplant treatment. H_0 : Heart transplant does not increase(has no effect on) lenght of survival H_1 : Heart transplant increases (has effect on) lenght of survival

(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 ***0***. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are ***0***. 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) The results of the simulated differences in proportions is centered at zero (0) and this provides more evidence that the treatment affected survival time and indeed lengthened the survival time. Hence, we can reject the null hypothesis H_0 and accept the alternative hypothesis H_1.