Chapter 2 - Summarizing Data

**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

**Answer**

**1. Box Plot Creation**

Below code creates a visual representation called a box plot, which helps summarize and visualize the distribution of a dataset.

**Explanation:**

* **Box Plot Overview:** A box plot (also known as a whisker plot) shows the distribution of a dataset based on five key summary statistics: minimum value, first quartile (Q1), median, third quartile (Q3), and maximum value.
* **Box Plot Features:**
  + Box: The central box represents the interquartile range (IQR) where the middle 50% of the data lies, extending from Q1 to Q3.
  + Median Line: A line inside the box marks the median, which is the middle value of the dataset.
  + Whiskers: Lines extending from the edges of the box to the minimum and maximum values indicate the range of the data outside the IQR.

**Process:**

1. Data: The dataset contains final exam scores of 20 students.
2. Visualization: The box plot visually displays how the scores are distributed and highlights any potential outliers.

A screenshot of a computer

Description automatically generated

**2. Summary Statistics Calculation**

The below code calculates and prints out the key summary statistics of the dataset, which are used to describe the distribution of the data.

**Explanation**:

* **Summary Statistics**: These statistics give a quick overview of the data's central tendency and spread.
  + **Minimum**: The lowest value in the dataset.
  + **First Quartile (Q1)**: The value below which 25% of the data falls.
  + **Median**: The middle value of the dataset, dividing it into two equal halves.
  + **Third Quartile (Q3)**: The value below which 75% of the data falls.
  + **Maximum**: The highest value in the dataset.

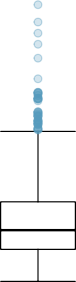
**Process**:

1. **Data**: The same dataset of final exam scores.
2. **Calculation**: Using statistical functions, the code computes these values to summarize the dataset.
3. **Output**: The results provide insights into the distribution and variability of the exam scores.

A screenshot of a computer program

Description automatically generated

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

70  100

6 80

65

60

4

60

40

2

55 20

50 60 70

(a)

0 50 100

(b)

0 2 4 6

(c)

0

(1)

(2)

0

(3)

**Answers**

1. **Histogram (a) and Box Plot (2) Correspondence**: The histogram is bell-shaped, indicating a symmetrical distribution with a peak in the center and tails of equal length on both sides. The center of the distribution is approximately 60.
2. **Histogram (b) and Box Plot (3) Correspondence**: This histogram has a rectangular shape, showing a symmetrical distribution with a central value around 50.
3. **Histogram (c) and Box Plot (1) Correspondence**: The histogram features a long tail on the right side, suggesting a right-skewed distribution with the center around 1.

**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.

1. 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.

1. 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.

1. 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.
2. 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.

**Answers**

**(a) Housing Prices:**

* **Distribution:** Right Skewed
  + Most houses are less expensive, but there are a few very expensive ones that push the average price up.
* **Best Measure for Typical Price:** Median
  + The median gives a better sense of what a typical house costs since it's not affected by those extremely high prices.
* **Best Measure of Variability:** IQR (Interquartile Range)
  + The IQR is better because it focuses on the middle range of house prices and ignores the extreme high prices.

**(b) Housing Prices:**

* **Distribution:** Symmetric
  + Prices are evenly spread around the middle value, without any significant skew.
* **Best Measure for Typical Price:** Median
  + The median is a good measure of a typical price because it represents the middle point in a balanced distribution.
* **Best Measure of Variability:** Standard Deviation
  + The standard deviation works well here because the distribution is balanced and there are no extreme outliers.

**(c) Alcoholic Drinks Consumed by College Students:**

* **Distribution:** Left Skewed
  + Most students drink very little or none, but a few drink a lot, creating a long tail on the right.
* **Best Measure for Typical Consumption:** Median
  + The median gives a better idea of the typical student's drinking habits since it’s not skewed by the few heavy drinkers.
* **Best Measure of Variability:** IQR (Interquartile Range)
  + The IQR is more reliable here because it focuses on the middle 50% of students and isn’t affected by the few who drink excessively.

**(d) Salaries at a Fortune 500 Company:**

* **Distribution:** Right Skewed
  + Most employees earn a moderate salary, but a few high-level executives earn very high salaries, skewing the distribution to the right.
* **Best Measure for Typical Salary:** Median
  + The median gives a better sense of what a typical employee earns without being influenced by the very high salaries of executives.
* **Best Measure of Variability:** IQR (Interquartile Range)
  + The IQR is suitable because it focuses on the range where most salaries fall and is not affected by the very high salaries.

**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.

control treatment

Survival Time (days)

alive 1500

# 1000

dead

500

# 0

control treatment

1. Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.
2. What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.
3. What proportion of patients in the treatment group and what proportion of patients in the control group died?
4. One approach for investigating whether or not the treatment is effective is to use a randomization technique.
   1. What are the claims being tested?
   2. 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.

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

**Answers**

**(a) Mosaic Plot and Survival**

**Question: Is survival independent of whether a patient got a transplant?**

**Answer:**

* Independence: If survival were independent of getting a transplant, then the survival rates would be similar for both groups (transplant and no transplant). In a mosaic plot, if you see big differences in survival between the two groups, it means survival is affected by whether or not the patient received a transplant.

**(b) Box Plots and Treatment Effectiveness**

**Question: What do the box plots suggest about the heart transplant treatment?**

**Answer:**

* Box Plot Insight: Box plots show how the data is spread out. If the box plot for the treatment group (transplant) shows better survival (e.g., higher median survival) compared to the control group (no transplant), it **suggests that the transplant might be effective in increasing lifespan.**

**(c) Proportions of Death**

**Question: What proportion of patients in each group died?**

**Answer:**

* Control Group:
  + 30 out of 34 patients died.
  + Proportion of deaths = 30/34 ≈ 88.2%
* Treatment Group:
  + 45 out of 69 patients died.
  + Proportion of deaths = 45/69 ≈ 65.2%

**(d) Randomization Technique**

(i). Claims Being Tested

**Question: What are we testing?**

**Answer:**

* We’re testing if the heart transplant makes a real difference in survival compared to not getting a transplant.

(ii). Randomization Setup

**Question: Fill in the blanks for the randomization process.**

**Answer:**

**Randomization Setup**

To perform the randomization test, we did below steps:

1. **Cards**:
   * Writing "alive" on 28 cards (patients who survived) and "dead" on 75 cards (patients who did not survive).
2. **Group Sizes**:
   * Size of treatment group = 69
   * Size of control group = 34
3. **Setup**:
   * Shuffle all 69 + 34 = 103 cards.
   * Split them into two groups: one group of size 69 (treatment) and another group of size 34 (control).
4. **Calculate the Difference**:
   * Calculating the difference in the proportion of "dead" cards between the treatment group and the control group.
5. **Repeat**:
   * Repeating this process 100 times to build a distribution of differences under the null hypothesis.
6. **Fraction of Simulations**:
   * Calculating the fraction of simulations where the difference in proportions is as extreme as, or more extreme than, the observed difference from the actual study.

If this fraction is low (usually below 0.05), it suggests that the observed difference is unlikely to be due to chance, leading to rejection of the null hypothesis in favor of the alternative.

**Explanation of Numbers**

* **28 cards labeled "alive"**: This number comes from the total number of patients who survived in the study (since 34 patients in the control group had 30 deaths, implying 4 survived; for the treatment group, 69 patients with 45 deaths implies 24 survived).
* **75 cards labeled "dead"**: This represents the total number of patients who died (30 in the control group and 45 in the treatment group).
* **69 and 34**: These are the sizes of the treatment and control groups respectively.
* **100**: Number of simulations performed to assess the randomization distribution.
* **0**: The distribution is centered at zero, reflecting no effect in the null hypothesis.
* **greater than or equal to the observed difference**: This phrase refers to comparing simulated results with the actual observed difference in proportions.

A screenshot of a computer program

Description automatically generated’’

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 **23.02%**. 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). Simulation Results

**Question: What do the results say about the transplant?**

**Answer:**

Although the observed difference in survival rates between the treatment and control groups is 23.02%, which might seem modest, the null hypothesis should be rejected based on this difference. This is because, despite the seemingly small percentage, if the p-value from the randomization test indicates that such a difference is statistically significant, it suggests that the observed effect is unlikely to be due to random variation. Therefore, rejecting the null hypothesis is appropriate if the p-value is below the chosen significance level.

−0.25 −0.15 −0.05 0.05 0.15 0.25

simulated differences in proportions