|  |  |
| --- | --- |
| **Course** | ADS-500A: Statistics & Probability for Data Science |
| **Textbook** | *Probability and Statistics for Engineering and the Sciences* (9th), Jay L. Devore |
| **Assignment** | Module 1 – Textbook Assignments |
| **Student Name** | Filipp Krasovsky |

**Exercises Section 1.1 (Page 12) {6 points}**

**5.** Many universities and colleges have instituted supplemental instruction (SI) programs, in which a student facilitator meets regularly with a small group of students enrolled in the course to promote discussion of course material and enhance subject mastery. Suppose that students in a large statistics course (what else?) are randomly divided into a control group that will not participate in SI and a treatment group that will participate. At the end of the term, each student’s total score in the course is determined.

a. Are the scores from the SI group a sample from an existing population? If so, what is it? If not, what is the relevant conceptual population?

|  |
| --- |
| The scores from the SI group are not a sample from an existing population; instead they are part of a conceptual population of all possible score measurements created under similar experimental circumstances. |

b. What do you think is the advantage of randomly dividing the students into the two groups rather than letting each student choose which group to join?

|  |
| --- |
| Allowing students to decide which group to join would create a selection bias problem. |

c. Why didn’t the investigators put all students in the treatment group?

|  |
| --- |
| putting all students in the treatment group would be unproductive in determining whether or not participation in the SI program made a statistically significant impact on the average total score of a student in the class. |

**Exercises Section 1.2 (Page 27) {2 points}**

**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 than the early distance?



*Histogram for Exercise 22*

|  |
| --- |
| The histogram contains a unimodal, positively skewed distribution with a typical difference value in the 50-150 range. Provided the graph is to scale and we don't known the size of *n*, adding up the frequencies of T gives us around 930 participants. provided that T<0 adds up to a frequency of 10, we can say that the relative frequency of runners who ran the late distance faster was about ~.01, or 1 percent. |

**Exercises Section 1.3 (Page 34) {8 points}**

**33.** The **May 1, 2009, issue of *The Montclarian*** reported the following home sale amounts for a sample of homes in Alameda, CA, that were sold the previous month (1000s of USD):

590

815

575

608

350

1285

408

540

555

679

a. Calculate and interpret the sample mean and median.

|  |
| --- |
| The sample mean is 640.5 - if we randomly selected a single data point from the sample, its expected home price value would be approximately 640.5 thousand dollars.  The sample median is 582.5, which similarly means that home prices are distributed around a central value of 582.5 thousand dollars. |

b. Suppose the 6th observation had been 985 rather than 1285. How would the mean and median change?

|  |
| --- |
| The mean would become 610.5, while the median would not change at all. |

c. Calculate a 20% trimmed mean by first trimming the two smallest and two largest observations.

|  |
| --- |
|  |

d. Calculate a 15% trimmed mean.

|  |
| --- |
|  |

**Exercises Section 1.4 (Page 45) {4 points}**

**53.** A mutual fund is a professionally managed investment scheme that pools money from many investors and invests in a variety of securities. Growth funds focus primarily on increasing the value of investments, whereas blended funds seek a balance between current income and growth. Here is data on the expense ratio (expenses as a % of assets, from [www.morningstar.com](http://www.morningstar.com/)) for samples of 20 large-cap balanced funds and 20 large-cap growth funds (“large-cap” refers to the sizes of companies in which the funds invest; the population sizes are 825 and 762, respectively):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Bl | 1.03 1.27 0.94 0.79 | 1.23 1.25 2.86 1.61 | 1.10 0.78 1.05 1.26 | 1.64 1.05 0.75 0.93 | 1.30 0.64 0.09 0.84 |
| Gr | 0.52 0.99 0.91 1.02 | 1.06 1.10 0.79 1.10 | 1.26 1.07 1.39 1.78 | 2.17 1.81 0.62 1.01 | 1.55 2.05 1.52 1.15 |

a. Calculate and compare the values of , , and , for the two types of funds. *You may use Microsoft Excel to calculate the* AVERAGE(), MEDIAN(), and STDEV.S(), rounding (ROUND()) to the nearest three decimal places.

|  |
| --- |
|  |

b. Referencing the comparative boxplot shown to the right for the two types of funds, comment on interesting features*.*

|  |
| --- |
|  |

**Exercises Section 2.2 (Page 65) {4 points}**

**14.** Suppose that 55% of all adults regularly consume coffee, 45% regularly consume carbonated soda, and 70% regularly consume at least one of these two products.

a. What is the probability that a randomly selected adult regularly consumes both coffee and soda?

|  |
| --- |
|  |

b. What is the probability that a randomly selected adult doesn’t regularly consume at least one of these two products.

|  |
| --- |
|  |

**Exercises Section 2.4 (Page 84) {6 points}**

**59.** At a certain gas station, 40% of the customers use regular fuel (), 35% use plus fuel (), and 25% use premium fuel (). Of those customers using regular gas, only 30% fill their tanks (event ). Of those customers using plus, 60% fill their tanks, whereas of those using premium, 50% fill their tanks.

a. What is the probability that the next customer will request plus fuel *and* fill the tank ()?

|  |
| --- |
|  |

b. What is the probability that the next customer fills the tank?

|  |
| --- |
|  |

c. If the next customer fills the tank, what is the probability that regular fuel was requested? Plus? Premium?

|  |
| --- |
|  |

**Exercises Section 2.5 (Page 91) {8 points}**

**87.** Consider randomly selecting a single individual and having that person test drive three different vehicles. Define events , , and , by:

Suppose that , , 0, , , and .

a. What is the probability that the individual likes both vehicle #1 and vehicle #2?

|  |
| --- |
|  |

b. Determine and interpret .

|  |
| --- |
|  |

c. Are and mutually independent events?

|  |
| --- |
|  |

d. If you learn that an individual did not like vehicle #1, what now is the probability that they liked at least one of the other two vehicles?

|  |
| --- |
|  |