STA101 Problem Set 2 - KEY

Summer I, 2021, Duke University

Exercises from the OpenIntro book and 2 additional Problems - 100pts total

Problems include: Chapter 2 exercises 2.10, 2.21, 2.26, 2.34 and 2 additional problems

2.10 (18pts total)

(a) **(3pts for description, 3pts for correctly matching)** The distribution is unimodal and symmetric, and about 95% of the data falls within about 7 units of the center, so the standard deviation will be about 3 or 4. This matches box plot (2).

(b) **(3pts for description, 3pts for correctly matching)** The distribution is uniform and values range from 0 to 100. This matches box plot (3) which shows a symmetric distribution in this range. Also, each 25% chunk of the box plot has about the same width and there are no suspected outliers.

(c) **(3pts for description, 3pts for correctly matching)** The distribution is unimodal and right skewed with a median between 1 and 2. 25th and 75th percentile are near 1 and 2, so the IQR is roughly 1. This matches box plot (1).

2.21 (6pts total)

(a) **(2pts)** We see the order of the categories and the relative frequencies in the bar plot.

(b) **(2pts)** There are no features that are apparent in the pie chart but not in the bar plot.

(c) **(2pts)** We usually prefer to use a bar plot as we can also see the relative frequencies of the categories in this graph.

2.26 (22pts total)

(a) **(2pts for "not independent", 4pts for reasoning)** Proportion of patients who are alive at the end of the study is higher in the treatment group than in the control group. These data suggest that survival is not independent of whether or not the patient got a transplant.

(b) **(2pts for indicating treatment is more effective, 4pts for some justification)** The shape of the distribution of survival times in both groups is right skewed with one very clear outlier for the control group and other possible outliers in both groups on the high end. The median survival time for the control group is much lower than the median survival time for the treatment group; patients who got a transplant typically lived longer. Tying this together with the much lower

variability in the control group, evident by a much smaller IQR than the treatment group (about 50 days versus 500 days), and we can see that patients who did not get a heart transplant tended to consistently die quite early relative to those who did have a transplant. Overall, very few patients without transplants made it beyond a year while nearly half of the transplant patients survived at least one year. It should also be noted that while the first and third quartiles of the treatment group is higher than those for the control group, the IQR for the treatment group is much bigger, indicating that there is more variability in survival times in the treatment group.

(c) **(2pts)** Proportion of patients who in the treatment group that died: $\frac{45}{69} = 0.652$. Proportion of patients who in the control group that died: $\frac{30}{34} = 0.882$

(d) **(2pts for (i), 4pts for (ii), 2pts for reject claim of independence in (iii)**

(i) H_0: The variables group and outcome are independent. They have no relationship, and the difference in survival rates between the control and treatment groups was due to chance. In other words, heart transplant is not effective.
H_A : The variables group and outcome are not independent. The difference in survival rates between the control and treatment groups was not due to chance and the heart transplant is effective.

(ii) 28, 75, 69, 34, 0, -0.23 or lower.

(iii) Under the independence model, only 2 out of 100 times (2-0.23 or lower between the proportions of patients that died in the treatment and control groups. Since this is a low probability, we can reject the claim of independence in favor of the alternate model. There is convincing evidence to suggest that the transplant program is effective.

2.34 (18pts total)

(a) **(2pts for bimodality, 2 pts for outliers)** From the histogram we can see the the distribution is bimodal. In the box plot the more extreme observations, many of which could be considered outliers, are easier to identify.

(b) **(4pts)** Gender may be the reason, it is likely that men and women have different average marathon times.

(c) **(6pts for reasonable comparisons)** The median marathon time for men is about 2.2 hrs while it is about 2.5 hrs for women; therefore, men are faster on average. The minimum marathon time for men is about 2.1 hrs whereas it is about 2.4 hrs for women. The maximum marathon time for men is about 2.5 hrs for men whereas it is about 3.1 hrs for women. Both distributions have apparent outliers on the high end.

(d) **(4pts for identifying a new feature in this plot)** It appears that marathon times decreased greatly between 1970-1975 and remained somewhat steady thereafter. Males consistently had shorter marathon times than females throughout the years. From the box plots of males and females, we could tell that males ran faster "on average", however, we could not tell that the winning male time for each year was better than the winning female time. We also could not tell from the histogram or the box plot that marathon times have been decreasing for males and females throughout the years.

Problem 5 (14pts total)

(a) **(4pts for "decrease")** The late submission has a score below the average of the on-time submissions, so the average will decrease.

(b) **(2pts for calculation, 2pts for right answer)** The total of the first 29 scores is $90 \cdot 29 = 2,610$. Then the total of all 30 scores is $90 \cdot 29 + 76 = 2,686$, and so the average of all 30 scores is $(90 \cdot 29 + 76)/30 \approx 89.5$.

(c) **(2pts for increase, 4pts for intuitive explanation)** Since the 30^{th} observation is more than a standard deviation of average of the first 29 scores, it would increase the standard deviation.

Problem 6 (22 pts total)

(a) **(4pts)** There are 175 liberal voters, and 910 total voters, so $175/910 \approx 19.2\%$ of voters are liberals.

(b) **(4pts)** 262 voters prefer the guest worker option, which is approximately $262/910 \approx 28.8\%$ of voters.

(c) **(4pts)** 28 voters identify as liberal and prefer the guest worker option, which is approximately $28/910 \approx 3.08\%$ of voters.

(d) **(2pts for each calculation)** There are 372 conservative voters, and 179 of them believe illegal workers should leave the country, which is about $179/372 \approx 48.1\%$. Similarly, about $126/363 \approx 34.7\%$ of moderates and $45/175 \approx 25.7\%$ of liberals believe illegal workers should leave the country.

(e) **(2pts for "dependent", 2pts for reasoning)** The percentages of Tampa, FL conservatives, moderates, and liberals who are in favor of illegal immigrants leaving the country are quite different from one another. Therefore, the two variables appear to be dependent.