STAT22000 Winter/Summer 2020 Homework 1 Solutions

Problems to Turn In: due **midnight on Friday, June 26, on Gradescope**.

1. Refer to the description of the Aircraft-Wildlife Collisions data at

   `https://www.openintro.org/data/index.php?data=birds`

   (a) What is a case in this data set?

   (b) Determine whether each of the five variables: `ac_mass`, `effect`, `num_engs`, `height`, and `bird_struck`, is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is nominal or ordinal.

Answer:

(a) *[1pt]* A case is an aircraft-wildlife collision

(b) *[5pts, 1pt each]*

- `ac_mass` is **categorical** and **ordinal**. Note that the numerical value of the mass of an aircraft is not available. We just know whether the mass falls in one of the five intervals: "2250 kg or less", "2251-5700 kg", "5701-27000 kg", "27001-272000 kg", and "above 272000 kg", which are ordered.
- `effect` is **categorical** and **nominal** as the 5 categories: "None", "Aborted Take-off", "Precautionary Landing", "Engine Shut Down", "Other", are not ordered. Though some of the categories are ordered in severity, like "None" is the least severe one among the 5 categories, it's hard to say whether "Aborted Take-off" or "Precautionary Landing" is more severe, and we are not able compare the category "Other" with the other 3 since the actual effect of the collision is not recorded.
- `num_engs` is **numerical** and **discrete** as it has only 4 possible values: 1, 2, 3, and 4.
- `height` is **numerical** and **continuous** as the values can have any number of decimal places
- `bird_struck` is **categorical** and **ordinal** since it falls in one of the 5 ordered categories: 0, 1, 2-10, 11-100, Over 100. Note the actual number of birds/wildlife struck was not recorded unless it's 0 or 1.

---

2. Suppose that the cases in a study are the purchases that Yibi made on amazon.com in the past 12 months. Identify each of the following is a variable or not a variable. If it is a variable, determine whether it is numerical or categorical.

   (a) How much did Yibi spend on those purchases in total?

   (b) Was the purchase shipped to Yibi or to someone else?

   (c) Did Yibi spend more on purchases sent to others than on purchases sent to herself?

Answer: *[3pts, 1pt each]* Note that variables are things that can be recorded for each case (one purchase on Amazon), not an overall question or measure that pertains to the entire dataset.

(a) "The amount Yibi spent on those purchases in total" is NOT a variable since it pertains all the purchases, not a single purchase. (Note that "the amount Yibi spent on the purchase" is a variable since it just describe one purchase.)

(b) "Whether the purchase was shipped to Yibi or to someone else" is a variable since it describe a single purchase.

(c) "Whether Yibi spend more on purchases sent to others or on purchases sent to herself" is NOT a variable since it involves more than one purchase.

---

3. An investigator has a data file showing family incomes for 1,000 subjects in a certain study. The minimum \$5,800 a year to \$98,600 a year. By accident, the highest income in the file gets changed to \$986,000.

   (a) Does this affect the mean? If so, by how much?

   (b) Does this affect the median? If so, by how much?

Answer:

(a) [*2pts*] Yes, it will affect the mean. The mean will go up by $(\$986,000 - \$98,600)/1000 = \$887.4$.

(b) [*2pts*] No, the median will not be affected. This is one advantage of the median – it is not affected by outliers.

---

4. Below are the final exam scores of 25 introductory statistics students.

   42, 53, 63, 76, 76, 78, 80, 85, 86, 86, 87, 87, 88, 88, 89, 89, 90, 91, 92, 94, 95, 95, 96, 96, 97

   Below are the five number summary.

Min	Q1	Median	Q3	Max
42	80	88	92	97

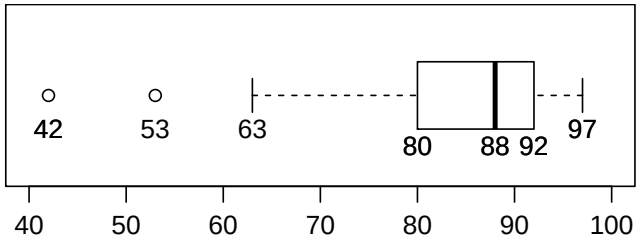
   (a) Use the 1.5 IQR rule to identify outlier(s) if any.

   (b) Please make a boxplot for the data manually (not by R) and indicate on the plot the values the boundaries and the middle line of the box represent respectively, and show how you determine the values the two whiskers extend to.

Answer:

(a) [*2pts*] The IQR of the data set is $Q3 - Q1 = 92 - 80 = 12$. By the 1.5 IQR rule, the upper fence is $Q3 + 1.5\ \text{IQR} = 92 + 1.5 \times 12 = 110$ and the lower fence is $Q1 - 1.5\ \text{IQR} = 80 - 1.5 \times 12 = 62$. There are two potential outliers: 42, 53 as they fall outside of the two fences.

(b) [*4pts =1pt for the boxplot + 1pt for Q1,Q2,Q3 + 1pts for the values of the whiskers + 1pt for showing how the whiskers are determined*] The boxplot can be made using the R commands above.



[*Students may mark the values on the plot or describe them in words. It's okay if the values of the outliers are not marked*]

[*1pt*] The midpoint of the box is at the median, 88. The boundaries of the box are at the first quartile $Q1 = 80$ and the third quartile $Q3 = 92$.

[*2pts = 1pts for the values of the whiskers + 1pt for showing how the whiskers are determined*] The two whiskers extends to the min and max values that are not outliers, which are 63 and 97.

---

5. For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. 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) Number of alcoholic drinks consumed by college students in a given week. Assume that more than half of these students don't drink since they are under 21 years old, and only a few drink excessively.

   (c) The average on a history exam (scored out of 100 points) was 85, with a standard deviation of 15.

Answer: [*2pts each = 1pt for the answer + 1pt for the reasons*]

(a) [*2pts*] **Right skewed** with potential outliers on the right tail. This is because Q1 = 350K is closer to the median = 450K than Q3 = 1000K is. Moreover, the houses that cost more than 6000K are more than 1.5 IQR above Q3. Both indicate the right tail is longer than the left tail.

(b) [*2pts*] **Right skewed**. There would be some students who did not consume any alcohol, but this is the minimum since students cannot consume fewer than 0 drinks. There would be a few students who consume many more drinks than their peers, giving the distribution a long right tail.

(c) [*2pts*] **Left skewed**. Since the mean is at 85, and it is not possible to score above 100 on the exam, the length of the right tail of the histogram won't be over $100 - 85 = 15$. If the length of the left tail is 15 or less, the standard deviation of the distribution can not be as big as 15. So the length of the left tail must be longer than 15, and hence longer than the right tail. So we would expect this distribution to be left skewed.
