Descriptive Statistics – Solutions
COR1-GB.1305 – Statistics and Data Analysis

Types of Data

1. The class survey asked each respondent to report the following information: gender; birth date; GMAT score; undergraduate major; time spent studying per week; interest level in the course; industry; job type; expected starting salary; number of dinners out per month; number of pairs of shoes; cups of coffee consumed per week; number of websites visited per day; political party; and presidential vote.

   (a) Which of the variables measured by the survey are categorical/qualitative?

   > **Solution:** gender; major; industry; job type; political party; presidential vote

   (b) Which of the variables measured by the survey are numerical/quantitative?

   > **Solution:** birth date; GMAT score; study time; interest level; dinners out; shoes; cups of coffee; number of websites

2. What type of variable is the answer to the phone prompt "Enter '1' for English, '2' for Spanish."? Why?

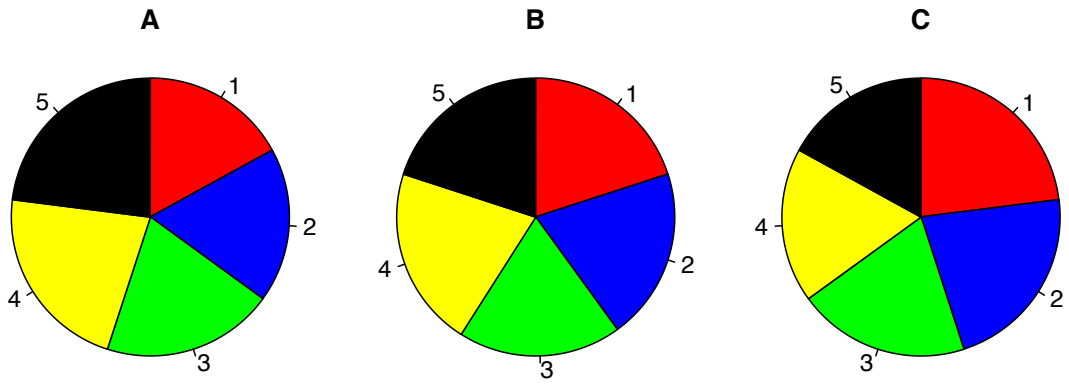
   > **Solution:** Categorical. Even the variable is measured on a numerical scale (1 or 2), the scale is not meaningful.

3. Each Yelp restaurant includes a star rating (1–5). What type of variable is the star rating?

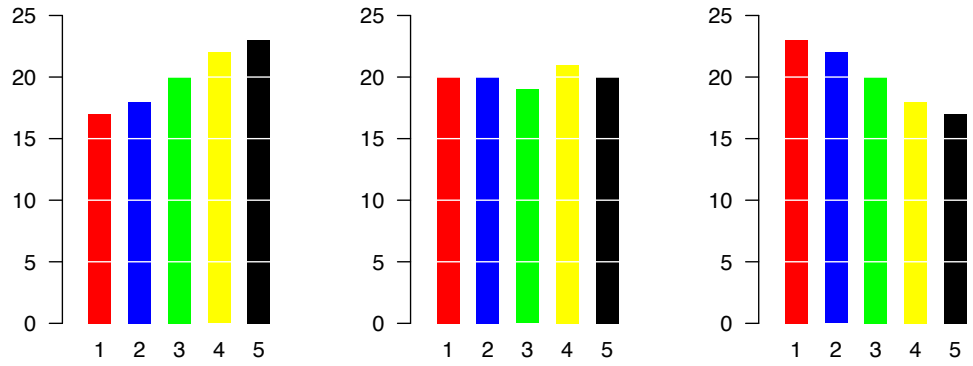
   > **Solution:** An argument can be made for both Categorical and Quantitative. In some ways the numerical scale is meaningful (order matters), but in other ways it is not (the difference between 4 and 5 stars is different than the difference between 1 and 2 stars). Sometimes this type of data is called "Ordinal."

Describing Categorical (Qualitative) Data

4. Use the following pie charts to rank the categories (1–5) by size.



Solution: This is much easier if we have bar charts instead:



The relative ordering of the categories is obvious. The takeaway here is that you should never use a pie chart; a bar chart conveys the same information, and it is much easier to read.

5. List two methods to describe the reported undergraduate majors of the class survey respondents.

Solution: Frequency table; bar chart.

6. Draw what you think the bar chart for the birth months of the survey respondents will look like.

Solution:

Describing Numerical (Quantitative) Data

7. Draw what you think the histogram for "Websites Visited per Day" will look like.

Solution:

8. Draw what you think the histogram for "Dinners per Month" will look like.

Solution:

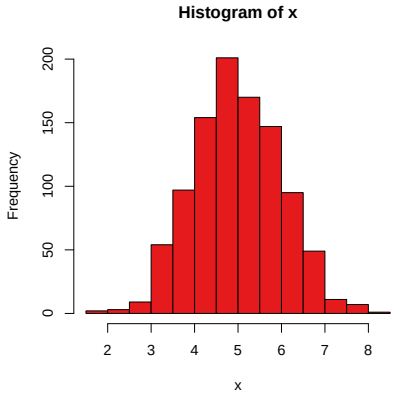
9. Draw what you think the histogram for "Interest in this Class" will look like.

Solution:

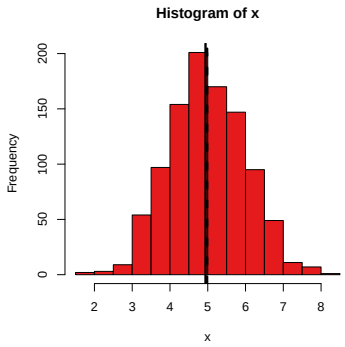
Measures of Central Tendency

10. Here are some histograms. Estimate the mean and median of the data.

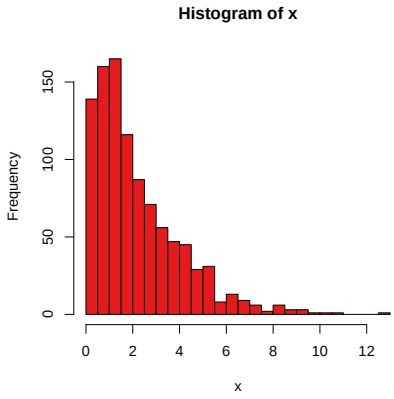
    (a) Symmetric and mound-shaped data.



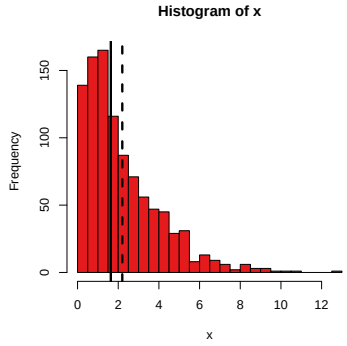
Solution: The median (solid) is roughly in the sample place as the mean (dashed).



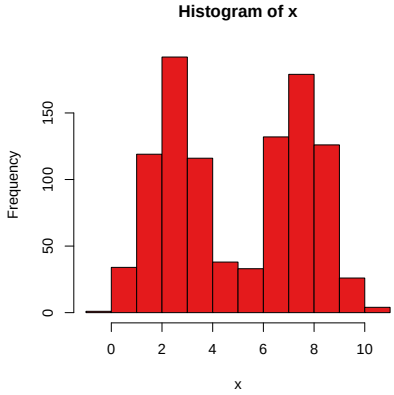
    (b) Skewed data.



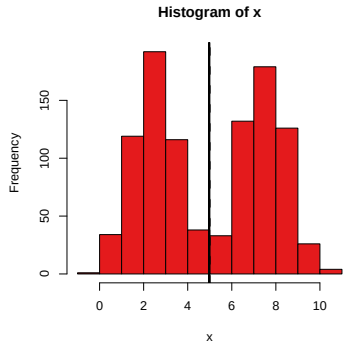
Solution: The mean is pulled to the right by the long tail.



(c) Bimodal data.



Solution: The median and the mean are roughly in the center. Note that neither number conveys much information about the distribution.



11. For the examples (a)–(c) of the previous problem, which is appropriate, the mean or the median?

Solution: This depends on context. If we care about "average" behavior, then mean is typically more appropriate; if we care about "typical" behavior, then median is typically

more appropriate.

(a) Both are appropriate; (b) the median is more appropriate for "typical" behavior; mean is more appropriate for "average" behavior; (c) mean is appropriate for "average"; median is not appropriate.

Standard Deviation and The Empirical Rule

12. Thirty-three respondents to the class survey reported their GMAT scores. The mean score was 720, and the standard deviation was 30. What can you say about the range of scores reported? Assume that the distribution of reported scores is symmetric and mound-shaped.

Solution: We can use the empirical rule to make the following statements:

- For approximately 68% respondents, reported score is between 690 and 750.
- For approximately 95% respondents, reported score is between 660 and 780.
- For approximately 99.7% respondents, reported score is between 630 and 810.

(For the last interval, it is ok to say "between 630 and 800," since it is impossible to score above 800.) In fact the true percentages in those intervals are 67%, 97%, and 100%. When the distribution of the data is symmetric and mound-shaped, the predictions from the empirical rule are usually only accurate for the 68% and 95% intervals.

13. The mean reported expected starting salary was $\$125K$ and the standard deviation was $\$50K$.

(a) Complete the following statement with appropriate values for X and Y: "Approximately 95% of the survey respondents have expected starting salaries between X and Y."

Solution: $X = 125 - 2 \times 50 = 25$; $Y = 125 + 2 \times 50 = 225$.

(b) What assumptions do you need to make for the statement in (a) to be correct? Do you think these assumptions are plausible? How could you check this?

Solution: That the distribution of expected salaries is symmetric and mound-shaped. We could check this with a histogram.

(c) What can we do if the assumptions needed in part (b) are not satisfied?

Solution: Sometimes, we can transform the data (e.g., by taking logarithms) to get a variable that has a symmetric, mound-shaped histogram.

z-scores

14. Your company has an annual profit of \$60MM with a standard deviation of \$5MM. Assume that the distribution of your annual profits is symmetric and mound-shaped.

    (a) Would it be unusual for your company to have an annual profit of \$52MM?

    Solution: No; 95% of the time, profits are between \$50MM and \$70MM.

    (b) Would it be unusual for your company to have an annual profit of \$83MM?

    Solution: Yes; this would happen less than 99.7% of the time.

15. Thirty-five respondents from the class survey reported their expected starting salaries. The histogram of these responses was approximately bell-shaped. The mean and standard deviation (in \$1K/year) was was $\bar{x} = 125$ and $s = 50$. How many standard deviations above or below the mean are the following values?

    (a) A starting salary of \$300K.

    Solution: Let $x_1 = 300$ and let z_1 be the number of standard deviations above of below the mean. Then,

    $$x_1 = \bar{x} + sz_1,$$

    so

    $$z_1 = \frac{x_1 - \bar{x}}{s} = \frac{300 - 125}{50} = 3.5.$$

    Thus, x_1 is 3.5 standard deviations above the mean.

    (b) A starting salary of \$100K.

    Solution: Let $x_2 = 100$. Then,

    $$z_2 = \frac{x_2 - \bar{x}}{s} = \frac{100 - 125}{50} = -0.5.$$

    Thus, x_2 is 0.5 standard deviations below the mean.

    (c) A starting salary of \$200K.

    Solution: Let $x_3 = 200$. Then,

    $$z_3 = \frac{x_3 - \bar{x}}{s} = \frac{200 - 125}{50} = 1.5.$$

    Thus, x_3 is 1.5 standard deviations above the mean.

16. In the previous problem, which of the values are unusual?

Solution: The value $x_1 = 300$ is unusual, since this is 3.5 standard deviations away from the mean. Typical values are within 2 or 3 standard deviations of the mean (here, "typical" means 95% or 99.7% of the time).