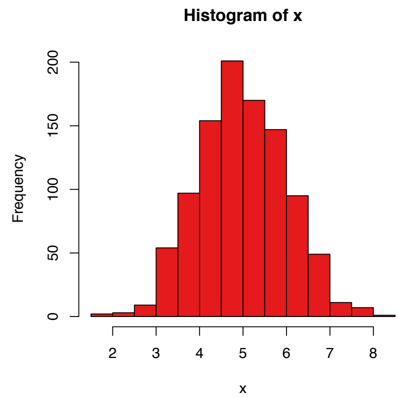
Descriptive Statistics 2 – Solutions
COR1-GB.1305 – Statistics and Data Analysis

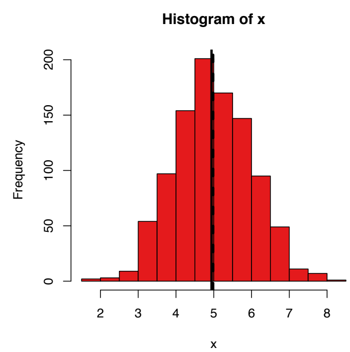
Measures of Central Tendency

1. Here are some histograms. Estimate the mean and median of the data.
   (a) Symmetric and mound-shaped data.

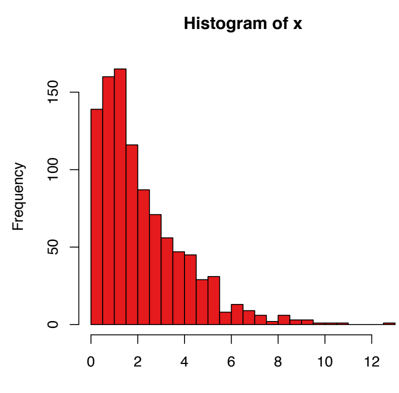


Text

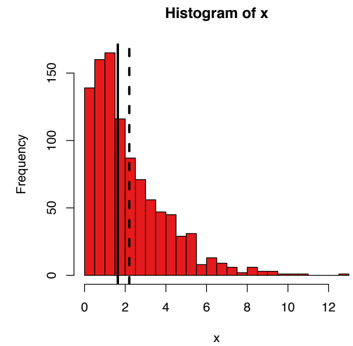
Solution: The median (solid) is roughly in the sample place as the mean (dashed).



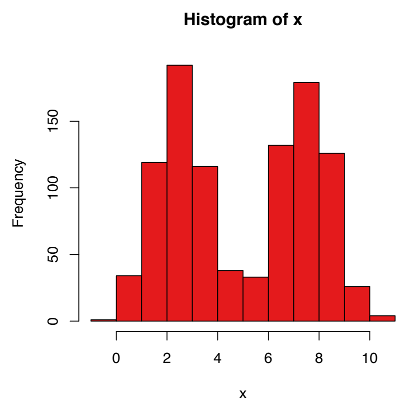
   (b) Skewed data.



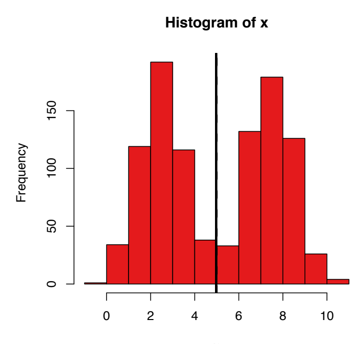
Solution: The mean is pulled to the right by the long tail.



(c) Bimodal data.



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



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

Solution: (a) Both are appropriate; (b) the median is more appropriate; (c) neither is appropriate.

Standard Deviation and The Empirical Rule

3. Sixteen respondents to the class survey reported their GMAT scores. The mean score was 680, and the standard deviation was 28.5. 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 651.5 and 708.5.
- For approximately 95% respondents, reported score is between 623 and 737.
- For approximately 99.7% respondents, reported score is between 594.5 and 765.5.

In fact the true percentages in those intervals are 56%, 94%, 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.

4. The mean reported commute time was 39 minutes, and the standard deviation was 25 minutes.
   (a) Complete the following statement with appropriate values for X and Y: "Approximately 95% of the survey respondents have commute times between X and Y."

   Solution: $X = 39 - 2 \times 25 = -19$; $Y = 39 + 2 \times 25 = 89$. Of course, it's impossible to have a negative commute time, so we could also say $X = 0$.

   (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 commute times is symmetric and mound-shaped. We could check this with a histogram. In fact, there is a slight skew to the right for the commute times, but even with this skewness, there is reasonable agreement with the empirical rule: 82.8% of commute times were within 1 standard deviation of the mean; 96.6% of commute times were within 2 standard deviations of the mean; 96.6% of commute times were within 3 standard deviations of the mean.

   (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. (For the commute times, taking logarithms doesn't fix the symmetry/mound-shaped assumptions, but it does lead to more sensible intervals.

z-scores

5. 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.

6. Thirty respondents from the class survey reported the number of websites they visit on a daily basis. The histogram of these responses was approximately bell-shaped. The mean and standard deviation was was $\bar{x} = 20$ and $s = 24$. How many standard deviations above or below the mean are the following values?

   (a) Visiting 100 websites per day.

   > **Solution:** Let $x_1 = 100$ and let z_1 be the number of standard deviations above of below the mean. Then,
   > $$x_1 = \bar{x} + s z_1,$$
   > so
   > $$z_1 = \frac{x_1 - \bar{x}}{s} = \frac{100 - 20}{24} = 3.33.$$
   > Thus, x_1 is 3.33 standard deviations above the mean.

   (b) Visiting 2 websites per day.

   > **Solution:** Let $x_2 = 2$. Then,
   > $$z_2 = \frac{x_2 - \bar{x}}{s} = \frac{2 - 20}{24} = -0.75.$$
   > Thus, x_2 is 0.75 standard deviations below the mean.

   (c) Visiting 30 websites per day.

   > **Solution:** Let $x_3 = 30$. Then,
   > $$z_3 = \frac{x_3 - \bar{x}}{s} = \frac{30 - 20}{24} = 0.41.$$
   > Thus, x_3 is 0.41 standard deviations above the mean.

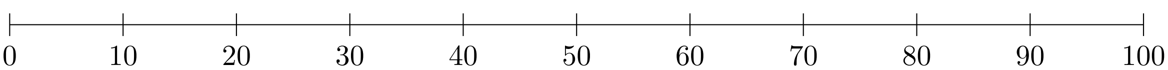
7. In the previous problem, which of the values are unusual?

Solution: The value $x_1 = 100$ is unusual, since this is 3.33 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).

Boxplots

8. Here are the 10 reported answers to the question "How many times do you go out to dinner in a typical month" for the female respondents. Make a boxplot of the data.

$$1, 4, 5, 6, 6.5, 8, 10, 10, 12, 20$$



9. Here are the answers for the 19 male survey respondents. Make a boxplot of the data.

$$4, 4, 4, 4, 4, 5, 5, 6, 8, 8, 8, 10, 10, 12, 13.5, 15, 20, 25, 25$$