Statistics and Data Analysis
Descriptive Statistics (2) – Summarization

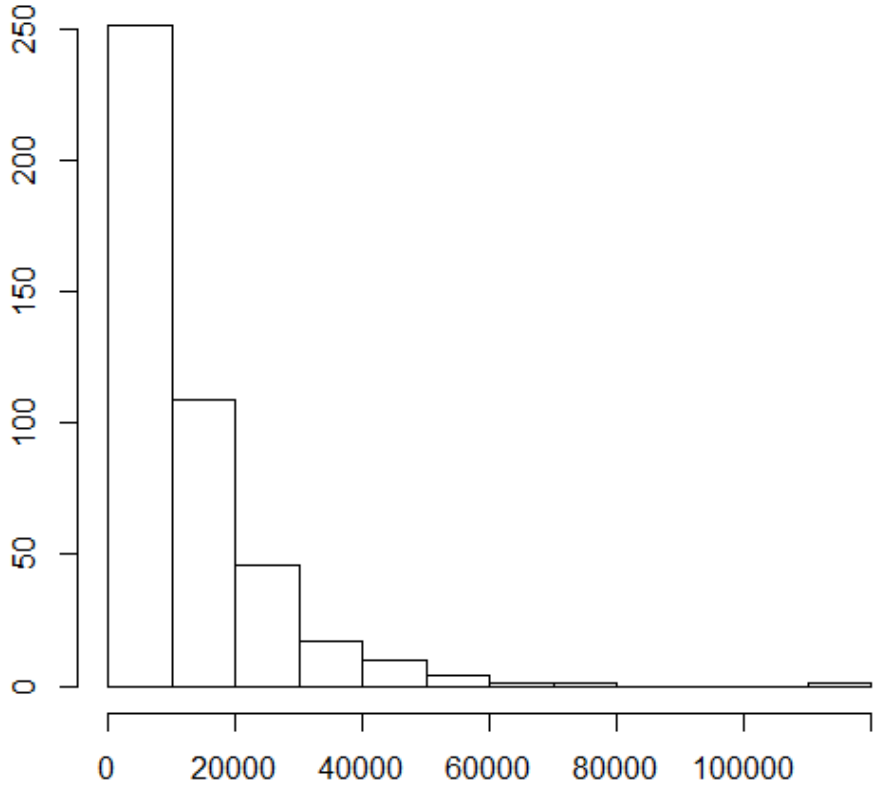
Instructor: Ling-Chieh Kung
Department of Information Management
National Taiwan University

1. Consider the wholesale data in the sheet "Wholesale".
   (a) For the grocery sales in region 1, use the median and mean to describe the central tendency.
   (In MS Excel: Data/Filter, `AVERAGE()` and `MEDIAN()`)
   (b) For the grocery sales in region 1, find the first quartile and 40th percentile.
   (In MS Excel: `QUARTILE.INC()` and `PERCENTILE.INC()`)

2. Consider the wholesale data in the sheet "Wholesale" again. Consider the fresh food sales and its histogram.

(a) For a quantitative data set, the definition of modes is typically changed to be the midpoint(s) of the modal class(es). Find the mode of the fresh food sales.

(b) Without doing any calculation, find the order of mode, median, and mean of the fresh food sales based on the histogram.



3. You are given the mode, median, and mean of a quantitative data set, but you are not given that set. You would like to guess whether the data are skewed to the left, skewed to the right, or symmetric.

   (a) If mode $<$ median $<$ mean, what is your conclusion?

   (b) If mode $>$ median $>$ mean, what is your conclusion?

   (c) If mode $\approx$ median $\approx$ mean, what is your conclusion?

4. Consider the daily bike rental data in the sheet "Bike_Day" again.

   (a) Draw two histograms, one for "cnt" in 2011 and one for "cnt" in 2012.

   (b) For "cnt" in 2011 and "cnt" in 2012, find their modes, medians, and means, respectively.

   (c) Verify that the following is (typically) true:

   - For a skewed-to-the-right distribution, mean $>$ median $>$ mode.
   - For a skewed-to-the-left distribution, mean $<$ median $<$ mode.
   - For a symmetric distribution, mean $=$ median $=$ mode.

5. Consider the daily bike rental data in the sheet "Bike_Day" again.

(a) For "cnt" in 2011 and "cnt" in 2012, compare their modes, medians, and means, respectively. Based on the comparisons, which set of values is considered "larger?"

(b) Draw the two frequency polygons for "cnt" in 2011 and "cnt" in 2012, respectively. Plot them on the same figure. What do you observe?

(c) Draw the two curves for the cumulative frequencies for "cnt" in 2011 and "cnt" in 2012, respectively. Plot them on the same figure. What do you observe?

6. Consider the data sets in "House" and "TeamHeight":

   (a) Find the sample variances and standard deviations for house sizes and prices.
   (In MS Excel: `VAR.S()` and `STDEV.S()`)

   (b) Find the sample coefficients of variation for house sizes and prices. Which variable has higher variability?

   (c) Find the population variance and standard deviation for team heights. Note that this is a set of population data!
   (In MS Excel: `VAR.P()` and `STDEV.P()`)

7. Consider the wholesale data set in the "Wholesales" sheet.

(a) Describe the correlation between grocery and detergents & paper by their correlation coefficient. Are they positively, negatively, or not correlated? How strong the correlation is?
(In MS Excel: `CORREL()`)

(b) Describe the correlation between fresh food and grocery by their covariance and correlation coefficient.

(c) For each of Parts (a) and (b), draw a scatter plot. Does your conclusions in Parts (a) and (b) fit your intuitions from the graphs?

(d) Add linear trend lines to the two scatter plots. For each pair of data, are the correlation coefficient and the slope of the trend line the same or different? Explain why.

8. We have learned how to use frequency distributions to make ungrouped data into grouped data. However, sometimes we may get grouped data only. To calculate the mean, variance, and standard deviation for grouped data, the rule is to assume that all values in one class is of the midpoint value. Now, consider the ungrouped and grouped data in the "Grouped" sheet.

   (a) Find the mean for the grouped data.

   (b) Find the variance and standard deviation for the grouped data.

   (c) Show that the values you obtained in Parts (a) to (b) are different from those calculated from the ungrouped data.

   Note. The median of a grouped data set is NOT calculated based on class midpoints. This is beyond the scope of this course. To know how to do it, look it up by yourself or ask the instructor.