LECTURE 2: DESCRIBING DISTRIBUTIONS

AP STATS

Given a data set X we want to see trends and be able to describe the data in a coherent manner. For a quantitative data set, we will be describing the distribution of the values of the data set using SOCS.

Contents

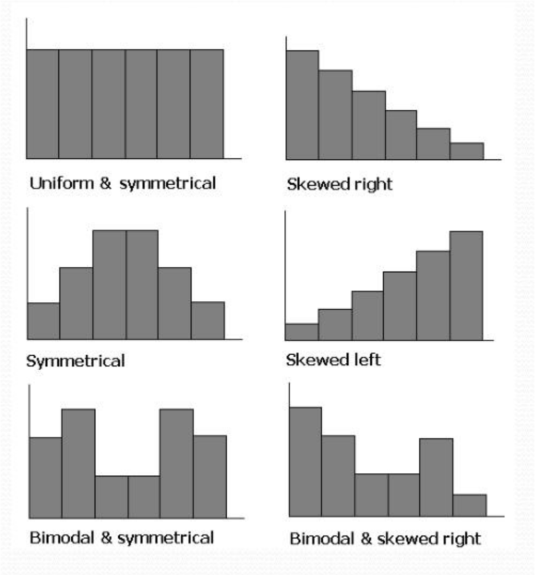
1. Shape

You need to describe the shape of the distribution.

(1) Modes: how many modes are there? If so, **where** are the modes located?
(2) Are there any unusual features?
- Clusters?
- Gaps in data? (A gap is a hole where there is no data)
- Symmetry in data?

The types of shapes:

(1) Uniform
(2) Unimodal/Bimodal/Multimodal
(3) Skew left
(4) Skew right



2. Outliers

These are extreme values that make describing your data *feel weird*. They can appear on either side of the data (or both).

A numerical rule for designating outliers is to calculate the interquartile range (IQR) and call any point outside of $1.5 \times IQR$ from the median "an outlier".

That is, suppose we have the data set

$$X = \{1, 2, 4, 6, 18, 37, 31, 16, 28, 24, 9, 4\}.$$

Then the median is going to be between 9 and 16. So the "median of X" is going to be $(9 + 16)/2 = 12.5$ but we will be using 9 to mark the top of the bottom 50% and the 16 as the bottom of the top 50%. We compute that $Q_3 = (24 + 28)/2 = 26$ and $Q_1 = (4 + 4)/2 = 4$ and so $IQR = 26 - 4 = 22$.

Computing that $1.5 \times IQR = 1.5 \times 22 = 33$, we conclude that if a number is more than 33 below $Q_1 = 4$ or is 33 above $Q_3 = 26$, then it will be an outlier.

3. Center

There are two measures of describing the center:

(1) Median = "the middle" which marks the 50th percentile
(2) Mean = arithmetic mean commonly called "the average"

The median is "harder" to move- you need to adjust more data points to adjust the median$\Longrightarrow$ use MEDIAN when the data is skewed.

4. Spread

The idea of spread is the essence of statistics- to describe spread you need to give me two pieces of data:

(1) Range
(2) Standard Deviation (or Variance)

5. Calculations and Notation

Let $X = \{3, 7, 1, 2\}$. Writing $x \in X$ means that $x = 3$ *or* $x = 7$ *or* $x = 1$ *or* $x = 2$- it just is forced to be one of the objects in the set.

Writing the sum $\sum_{x \in X} 2x$ means we cycle through the entire set X as we write out the sum:

$$\begin{aligned}\sum_{x \in X} 2x &= 2(x = 3) + 2(x = 7) + 2(x = 1) + 2(x = 2) \\ &= 2(3) + 2(7) + 2(1) + 2(2) \\ &= 6 + 14 + 2 + 4 \\ &= 26\end{aligned}$$

With this notation in mind, the formulas for the mean (expected value), variance, and standard deviation are below:

If $X = \{x_1, x_2, x_3, \ldots, x_n\}$, then

Mean $$E(X) = \mu = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Variance $$\mathrm{Var}(X) = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}$$

Standard Deviation $$\sigma_X = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$

So to compute standard deviation of a data set in a primitive manner, you need to first compute the mean and the variance!