MTH208: Worksheet 11

Descriptive Measures of Statistics

The purpose of this module is to determine certain features of a variable in a data set which will describe the nature of the variable in that data set in a general way. The most important features are:

- **Central tendency:** provides the average trend of the variable in appropriate sense.
- **Dispersion:** Scatterness of the variable.
- **Skewness:** Measure of the asymmetry of the frequency distribution.

Throughout, we will denote X as the random variable that represents a realization from a continuous population.

Central Tendency

Any representative and random data from the population is denoted by the random variable X. Since this is a random object, we are interested in its "average behavior".

Sample Mean:

The sample mean is a sample version of the population mean, $E(X)$.

The most familiar notion of average is that of arithmetic mean (AM), denoted by $\bar{x}$, i.e., given data $x_1, x_2, \dots, x_n$,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Sample Median:

Another measure of central tendency of a random variable is the population median, κ_{me} such that

$$\Pr(X \leq \kappa_{me}) = \Pr(X \geq \kappa_{me}) = \frac{1}{2}\,.$$

From data $x_1, x_2, \dots, x_n$, a guess for the population median can be obtained by the sample median $\tilde{x}_{me}$: it is a number such that at least half of the data points are bigger than or equal to it, and at least half of the data points are smaller than or equal to it, i.e.,

$$\sum_{i=1}^{n} \mathbb{I}(x_i \leq \tilde{x}_{me}) = \sum_{i=1}^{n} \mathbb{I}(x_i \geq \tilde{x}_{me})$$

here $\mathbb{I}$ is the indicator function.

Suppose the observations are arranged in ascending or descending order of magnitude, then median is the middle most value in this arrangement. Thus if n is odd, median is the $(n+1)/2$-th observation of the ordered arrangement. If n is even, then any value lying between the $n/2$-th and $(n/2)+1$)-th of the ordered arrangement is a median.

Sample Mode:

The population mode κ_{mo} for a discrete random variable is the value of the random variable associated with the highest mass function function.

The sample mode, $\tilde{x}_{mo}$, of a discrete variable is a value having the highest frequency. If there are more than one values having the highest frequency, then the mode is not unique.

For a continuous variable, the population mode is the value where the probability density function is the largest. The sample modal class is a class with highest frequency. The mode is *ideally the value of the variable with highest frequency density corresponding to the ideal distribution which would be obtained if the total frequency were increased indefinitely and, at the same time the width of the class intervals were decreased indefinitely.*

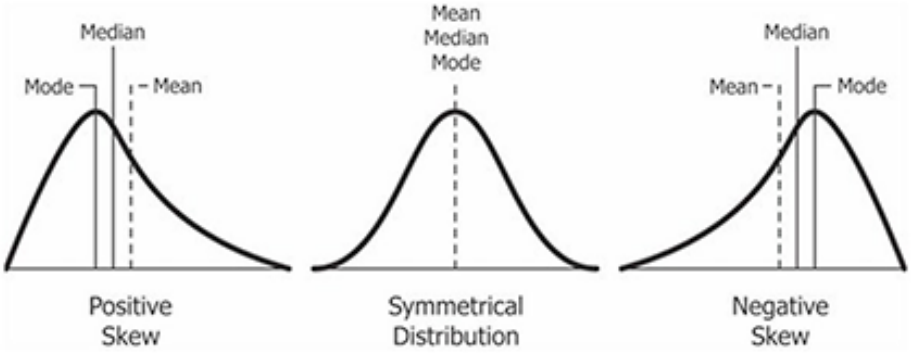
Comparison Between Mean Median and Mode:

(1) Sample mean is unique. Both sample median and sample mode may not be unique.

(2) Although in determining sample mean, sample median and sample mode, all the observations are taken into consideration, in the actual computation only the mean directly uses all the observations. The value of sample mean would change even if a single observation is altered.

(3) Mean is least affected by *sampling fluctuations.*

(4) Under the existence of extreme values, mean is most affected. Median and mode are more robust measures of central tendency than mean.



Univariate data - Visualizations

So far, we have learned how to collect data, clean and process it, and save it. Through courses in the next few years, you will learn how to analyze the data through statistical models.

However, a crucial component of data analysis is data visualization. This can often help ask interesting questions about the data.

We can think of visualizations as:

- single variable visualization - histogram, boxplots
- Multivariable visualization - scatterplot, side-by-side boxplot.

Using the movies data-set scraped in Worksheet 8 we will try to visualize the data. I have provided the dataset of weighted and unweighted ratings in the repository.

Note: In any visualization one must be very clear about what we are trying to visualize. Further axes should be clearly described and legends should be provided.

1. You can find the data in `movie_unweighted.csv` in your GitHub repository. Load the file using an appropriate function. This file contains the data on the ratings (adjusted and unadjusted) of the top 250 movies on IMDb

2. **Histogram**: read the documentation for the `hist()` function that makes a histogram.

   a. Make a histogram of the `ratings` for the top 250 movies. Using the argument `main`, set the title of the histogram to be "Histogram of Ratings" and the label on the x-axis as "Ratings".

   b. Make the histogram again so that the bars are white in color.

   c. Do the following

      i. Do `par(mfrow = c(1,2)` to make two plots next to each other. `c(1,2)` means 1 row, 2 columns.

      ii. Make a histogram of `ratings` and `unweighted` next to each other.

      iii. Use option `xlim = c(7.5, 10)` to define the limits on the x-axis.

      iv. Make sure you change the title using `main` and the x-labels using `xlab` options.

      v. Are both data points positively or negatively skewed?

      vi. Using functions `mean` and `median`, calculate the sample mean and sample median for both variables and add vertical lines using function `abline()` on the plot.

      vii. What is the modal class?

   d. **Measure of dispersion: sample variance**

      The sample variance is a sample version of the population variance

$$\sigma^2 = \mathrm{Var}(X) = E[(X - E(X))^2)\,.$$

The sample variance is:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Often, we look at s instead of s^2, since s is in the same units as the original data x.

e. Using function `var`, find the sample variance of both `ratings` and `unweighted`.

3. **Boxplot**: read the documentation for the `boxplot()` function that makes a boxplot. A boxplot is a summary of a continuous random variable that plot the quantiles of the data. Recall that a population p-quantile κ_p of a random variable X is such that

$$\Pr(X \leq \kappa_p) = p\,.$$

Similar to sample median, the sample quantile $\tilde{\kappa}_p$ is the value such that

$$\sum_{i=1}^{n} \mathbb{I}(x_i \leq \tilde{\kappa}_p) = np$$

In a boxplot, the line in the middle is the sample median of the observations, the bottom box the .25 sample quantiles, the top of the box is the .75 sample quantile.

a. Make a boxplot of the `ratings` of the top 250 movies. Make sure to assign an appropriate title using `main`.

b. Make the boxplot again so that the bars are pink in color.

c. **Measures of dispersion: range**
The range of a data is a the difference between the max and the min of the data. Find the range for both ratings.

d. **Measures of dispersion: interquartile range**
The inter quartile range, or IQR is

$$IQR = \tilde{\kappa}_{.75} - \tilde{\kappa}_{.25}$$

Find the IQR for both ratings using function `quantile()`.

4. **Side-by-Side Boxplots:** Reading the help page for `boxplot()`, make a side-by-side boxplot of `ratings` and `unweighted`. Make sure the axis labels are appropriate.

5. **Overlapped histograms:** Make a plot of histograms of `ratings unweighted` ratings, overlayed on top of each other. You may use the `col = adjustcolor("red", alpha.f = .5)` option to make colors transparent.

   Use `legend()` command to add a legend to the plot.

6. Below is a summary of skewed data and symmetric data. What kind of skew does `ratings` and `unweighted` have?