

Review of Basic Statistical Concepts

The purpose of this review is to summarize the basic statistical concepts. But first, some vocabulary

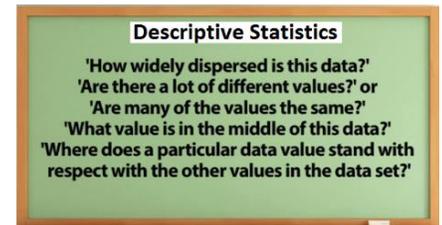
population - a designated group **data** - information collected from the population **sample** – a subset of the population

Basic Statistics can be divided into 3 parts:

- 1) Descriptive Statistics:** In descriptive statistics we collect data, organize it (usually with a table), and then summarize an aspect of a whole set of data (either a population or a sample) with just **one number** (ex: average) or with **a figure** (ex: graph or table). The number or figure is a description of our observations, our data. Although descriptive statistics give us useful information, descriptive stats from a sample cannot be used to make generalizations about the population.

Here are some **terms from descriptive statistics** you should know:

mean, median, mode, range, variance, standard deviation, skew, box-and-whisker, histogram, 5-number summary, frequency distribution, dot plot or statplot, bar graph, time-series, correlation coefficient “r”, coefficient of determination “r²”, causation



- 2) Probability:** Statistics is about data while probability is about chance. They are inverses of each other: in descriptive statistics we observe something that did happen. In probability theory we consider some underlying process which has some randomness or uncertainty modeled by random variables, and we figure out what could happen.

Here are some **terms from probability** you should know:

sample space, event, complement, independent events, mutually exclusive events, addition law

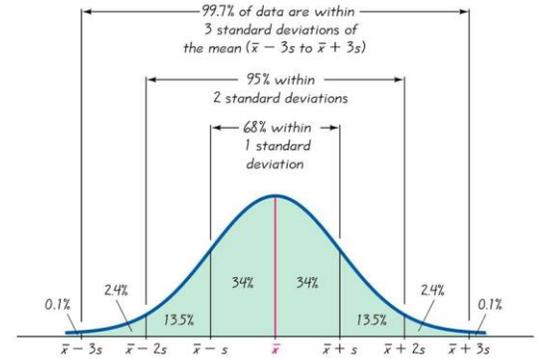


- 3) Inferential Statistics:** Inferential statistics starts with a sample and then generalizes to a population. Unlike descriptive statistics, the information about a population is not stated as a number. Instead we express these parameters as a range of potential numbers, along with a degree of confidence that takes into account chance factors that can occur when these

samples are used to reach conclusions (to make inferences) about the overall larger population from which the sample is drawn.

Here are some **terms from inferential statistics** you should know:

normal (bell) curve, z-score, central limit theorem, type 1 and type 2 error, level of significance, degrees of freedom, sampling distribution, confidence interval, hypothesis testing, null hypothesis, alternative hypothesis, chi-square, probability distribution, 5-number summary, empirical rule



In inferential statistics we are bouncing back and forth between the population and samples. So, to keep things straight we have a different set of variables for each. Take a look at this comparison:

| Meaning | Population version | Sample version |
|--|-----------------------------------|----------------------------------|
| numerical summary of data | parameter | statistic |
| number of elements | N | n |
| mean | μ read as "mew" or "mu" | \bar{x} read as "x bar" |
| median | n/a | M or Med or \bar{M} |
| standard deviation | σ "sigma" | s |
| variance | σ^2 "sigma squared" | s^2 |
| interquartile range | n/a | IQR = $Q_3 - Q_1$ |
| z score = $(x - \text{mean})/\text{sd}$ (number of standard deviations a data value is from the mean) | Z $Z = \frac{x - \mu}{\sigma}$ | z $z = \frac{x - \bar{x}}{s}$ |
| linear correlation coefficient (value between -1 and 1 showing negative or positive correlation) | r or ρ "rho" | R |
| proportion | p | \hat{p} read as "p hat" |

Resources

Understanding descriptive and inferential statistics:

<https://www.youtube.com/watch?v=edEXEYvG4Wk>