I. Descriptive Statistics (1)

Measures of Central Tendency
- Mode, Median, Mean

Measures of Variability
- Range, Interquartile Range, Variance, Standard Deviation

Mode (\(M_o\))
- The most frequently occurring score in a distribution of scores
- It is not influenced by the numerical value of extreme scores in the distribution
- Subject to substantial sampling fluctuation
- Useful for preliminary work
- Easy to obtain (but sometimes not a unique point in the distribution)
- Distributions can be unimodal, bimodal, or multimodal
- Used with nominal scales

Median (\(M_d\))
- A score value in the distribution with equal number of scores above and below it.
- The median is the 50th percentile in a distribution
- Less affected by extremes scores than the mean
- Good choice for strongly skewed distributions
- Used with ordinal scales

Mean (\(\bar{X}\))
- The sum of a set of scores divided by the number of scores summed

\[
\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}
\]

Mean (\(\bar{X}\))
- It is very sensitive to extreme observations
- The measure that best reflects the total of the scores
**Mean (\( \bar{X} \))**

- The most preferred average for *quantitative* data
- It describes the balance point of any distribution
- Key component in complex statistical measures

**Measures of Variability (Dispersion)**

- Numbers that represent how much scores differ from each other and the measure of central tendency in a distribution
- Range
- Interquartile Range (IQR)
- Variance
- Standard Deviation

**Range**

- The numerical difference between the highest and lowest scores in a distribution
- It describes the overall spread between the highest and lowest scores
- It is a relatively unstable measure of variability (depends only on two observations)
- Very easy to compute

**Interquartile Range (IQR)**

- The range for the middle 50 percent of all observations
- The numerical difference between the 3rd quartile and the 1st quartile of a frequency distribution
- It describes the overall spread between the 25th and the 75th percentile
- Relatively easy to compute

**Variance**

- The mean of all squared deviations from the mean
- A quantity expressed in squared units
- Of little use in descriptive statistics
- Important in statistical inference

\[
\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}
\]

**Standard Deviation**

- The square root of the variance
- A *rough* measure of the average amount by which observations deviate on either side of their mean

\[
\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}
\]
Standard Deviation
- A measure of distance
- Value can’t be negative
- Very good in its resistance to sampling variation
- Widely used in advanced statistics

Measures of Central Tendency
- ... and skewness ...

I. Descriptive Statistics (2)
- Introduction to the Normal Distribution
- \( z \) scores

Normal Distribution
- Developed from the work of
  - Bernoulli (1654-1705); de Moivre (1667-1754);
  - Montmort (1678-1719); Gauss (1777-1855)
- Interested in developing mathematical approximations for
  - Probabilities encountered in games of chance
  - Distribution of errors to be expected in observations

Normal Distribution
- Today, the ND has gained importance in many domains such as biology and behavioral sciences
- A large number of natural phenomena show distributions that are good approximations to the normal curve (e.g., distributions of weight, intelligence)
A reminder ...

Population:
- A complete set of actual or potential observations (people, objects, events) that share a common characteristic
- Parameters: A number that describes a characteristic of a population

Sample:
- A subset of observations selected from a population
- Statistic: A (single) number that describes a set of data from a sample

Normal Distribution

- We need to distinguish between
  - Data normally distributed and
  - the normal curve
- The latter is a mathematical abstraction that specifies the relative frequency of a set of scores in a population:
  - Has mode (unimodal), median, and mean with the same value
  - symmetrical
  - continuous
  - asymptotic to $x$-axis

Normal Distribution

- The ND is a theoretical relative frequency distribution
- Relative frequency of a score is
  - the frequency of occurrence of the score divided by the total number of scores in a distribution
- Therefore, RF is expressed as a proportion
  - where the total Cumulative Relative Frequency is 1.0
- Because the ND is a relative frequency distribution
  - the area under the curve is 1.0

Normal Distribution

- The normal curve is not a single curve. There are infinitely many possible normal curves, all described by the same algebraic expression

$$p(\lambda) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\lambda-\mu)^2}{2\sigma^2}}$$

- It has one variable ($\lambda$), three constants ($2, \pi, e$) and two parameters ($\mu, \sigma$)

Normal Distribution

- There are many (infinite!) instances (families) of normal distributions with
  - different means
  - different standard deviations

Standard Normal Distribution

- The standard normal distribution solves the problem of dealing with multiple families of normal distributions
- It is a normal distribution with parameters $\mu$ and $\sigma$ set as:
  - $\mu = 0$
  - $\sigma = 1$
The value of $z$ when a score is transformed into a score on the standard normal distribution

$$z = \frac{x - \mu}{\sigma} \quad (z \text{ score in a population})$$

$$z = \frac{x - \bar{x}}{s} \quad (z \text{ score in a sample})$$

The transformation from a score to a $z$ score (on the SND) locates the original score in terms of how many standard deviations the score is away from the mean.