Statistical Data Analysis

EKA: T510028102

Henrik Schwarz

Learning objectives

- Knowlegde
  - explain relevant data types and their representation for statistical analysis
  - explain probabilities and random variables
  - explain distributions of random variables
  - explain inference and hypothesis testing
  - explain how data may be collected from experiments involving randomness
- Skills
  - choose an appropriate experimental design in respect to a given task
  - perform statistical analyzes on data collected
  - use a statistical tool for analysis and visualization of data
- Competence
  - use statistical methods and tools to interpret experimental data

1 Lecture 1

Table 1: Terms in statistics

Term	Description
Variable	Characteristic or value that can change
Data	The values variables assume
Population	The subjects (human or otherwise) we study
Sample	Subset of the population

- Descriptive statistics vs Inferential Statistics
  - Descriptive statistics: Used to describe data
  - Inferential statistics: Used to make conclusions about

Measures of central tedency (london)

- Mean
  - Division and sum of all values.
  - Calculated $\overline{X} = \frac{X_1+X_2+...+X_n}{n} = \frac{\sum X}{n}$
  - Properties of Mean
    - Uses all data values
    - Unique, usually not part of the values
    - Affected by extremely low or high values (outliers)
- Median
  - Midpoint of the dataset.
  - Calculated by sorting all values in ascending order and then selecting the middle one.
  - If the number of values is odd it will be one value, if the number of values is even it will be the average of two.
  - Properties of Median:
    - Affected less than the mean by extremely low or high values.
- Mode
  - The Mode is the value that appears most often in the dataset.
  - Said to be the most typical case.
  - There may be no mode (all unique)m, one mode (unimodal), two modes (bimodal), or many modes (multimodal).

  - Calculated by sorting all the values, count instances and then select the one (or multiple) that has the most.
  - Properties of the Mode:
    * Easy to compute
    * Can be used with nominal data
    * May not exist

- Midrange
  - The midrange is the average of the lowest and highest value in the dataset.
  - Calculated by $MR = \frac{Lowest+Highest}{2}$.
  - Properties of the Midrange:
    * Easy to compute
    * Affected by **extremely** by low and high values in a dataset.

2 Lecture 2

Measures of variablity(dispersion)

- Range
  - Difference between highest and lowest values in the dataset.
  - $Highest - Lowest$

- Variance
  - Together with standard diviation, it is the measure of how spread out your data is.
  - Variance is the avarage of the squares of distance of each value is from the mean.
  - Population variance: $\sigma^2 = \frac{\sum(X-\mu)^2}{N}$ where X is the value, μ is the mean and N is the number of values.
  - Samlpe variance: $s^2 = \frac{\sum(X-\bar{X})^2}{n-1}$

- Standard Diviation
  - Together with standard diviation, it is the measure of how spread out your data is.
  - Population Standard deviation is $\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$
  - Sample Standard Deviation: $s = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}$ where X is the data, $\bar{X}$ is the mean and $n-1$ is the dataset size minus 1.

- Coefficient of variation
  - the coefficient of variation is the standard deviation divided by the mean expressed as percentage
  - $CV = \frac{s}{\bar{X}} \cdot 100\%$

Measure of position Measures of position indicate the position of a value relative to other values in a set of observations

- Z-score
  - Z score determines how many standard deviations a value is from the mean
  - $z = \frac{x_i - \bar{x}}{s}$ where x_i is the value, $\bar{x}$ is the mean and s is the standard deviation.
- Percentile
  - Percentiles separate the data set into 100 equal groups
  - A percentile rank for a datum represents the percentage of data values below the datum
- Decile and Quartile
  - Deciles - seperate the data set into 10 equal groups
  - Quartiles - seperate the data into 4 equal groups
    - $Q_1 = p_{25}$, $Q_2 = MD$, $Q_3 = P_{75}$
    - $Q_2 = median(Low, High)$, $Q_1 = median(Low, Q_2)$, $Q_3 = median(Q_2, High)$
    - The Interquartile Range $IQR = Q_3 - Q_1$
- Outlier
  - Outlier is an extremely low and high data values when compared to other values
  - Following data values can be considered outliers:
    - less than $Q_1 - 1.5(IQR)$
    - greater than $Q_3 + 1.5(IQR)$