AID-521 Mathematics for Data Science

Module: Statistics | Lecture: 1

BASIC CONCEPTS OF STATISTICS

introductory concepts, descriptive statistics (graphical, numerical)

Basic Concepts -- What is Statistics

- The objective of statistics is to make an inference about a population based on information contained in a sample taken from that population.
- The theory of statistics is a theory of information concerned with quantifying information, designing experiments or procedures for data collection, and analyzing data.
- A two-step procedure:
  - enlist a suitable inferential procedure for the given situation,
  - obtain a measure of the goodness of the resulting inference.

Basic Concepts -- What is Statistics

Statistics is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing (model building, estimating model), and interpreting data (via model).

- → *Descriptive Statistics*: organizing, summarizing, and presenting data in the form of tables, graphs, and charts
- → *Inferential Statistics*: methods of drawing inferences and making decisions about the population using the sample; uses probability theory

Basic Concepts -- What is Statistics

A statistical inference is an estimate, a prediction, a decision, or a generalization about the population based on information contained in a sample.

Basic Concepts -- Population & Sample

A population is the collection or set of all objects or measurements that are of interest to the collector.

A sample is a subset of data selected from a population. The size of a sample is the number of elements in it.

Data -- Collection, Types

- → Define the objectives of the problem and proceed to develop the experiment or survey
- → Define the variables or parameters of interest
- → Define the procedures of data-collection and measuring techniques. This includes sampling procedures, sample size, and data-measuring devices (questionnaires, telephone interviews, etc.)

Data -- Collection, Types

Experimental (planned) vs.
Observational (already collected) Data

- → Quantitative (numerical), vs. Non-numerical / qualitative
- → Categorical data (nominal, ordinal)
- → Cross-sectional data
- → Time series data
- → Network data [Spatial data]

Sampling Schemes -- What's a Sample

- Census study: data collected for all units of a population
- A sample, obtained by collecting information from only some members of the population, can be used to represent the population.
- A good sample must reflect all the characteristics (of importance) of the population.
  - representative
  - unbiased

Sampling Schemes -- How to Sample

The choice between different sampling methods that exist depends on

- → the nature of the problem or investigation
- → availability of good sampling frames (a list of all of the population members)
- → desired level of accuracy
- → method of data collection (questionnaires, interviews, experiment)
- → available financial resources

Sampling Schemes -- Random Sampling

A sample selected in such a way that every element of the population has an equal chance of being chosen is called a simple random sample. Equivalently each possible sample of size n has an equal chance of being selected.

- → mimimizes bias from investigator/researcher
- → analytic computations are simple; probabilistic bounds for errors can be computed
- → possible to estimate the sample size for a prescribed error level

Sampling Schemes -- Systematic Sampling

A systematic sample is a sample in which every Kth element in the sampling frame is selected after a suitable random start for the first element. We list the population elements in some order (say alphabetical) and choose the desired sampling fraction.

- → Number the elements of the population from 1 to N.
- → Decide on the sample size, say n, that we need.
- → Choose K = N/n.
- → Randomly select an integer between 1 to K .
- → Then take every K th element.

If there is a correlation or association between successive elements, or if there is some periodic structure, then this sampling method may introduce biases.

Sampling Schemes -- Stratified Sampling

- → Decide on the relevant stratification factors (sex, age, income, etc.).
- → Divide the entire population into strata (subpopulations) based on the stratification criteria. Sizes of strata may vary.
- → Select the requisite number of units using simple random sampling or systematic sampling from each subpopulation. The requisite number may depend on the subpopulation sizes.

Stratified sampling is often used when one or more of the strata in the population have a low incidence relative to the other strata. Compared to random sampling, stratified sampling reduces sampling error.

Sampling Schemes -- Cluster Sampling

In cluster sampling, the sampling unit contains groups of elements called clusters instead of individual elements of the population. A cluster is an intact group naturally available (unlike in stratification) in the field.

→ To obtain a cluster sample, first take a simple random sample of groups and then sample all elements within the selected clusters (groups).

Because it is likely that units in a cluster will be relatively homogeneous, this method may be less precise than simple random sampling.

Sampling Size

Determination of the sample size for a study depends on

- → population size,
- → variation in population,
- → required reliability of the results.

Numerical Description of Data

Characteristics associated with a data set:

→ Central Tendency: sample mean, median, mode

→ Variability: sample variance (s.d), interquartile range

Numerical Description of Data

Let $x_1, x_2, \ldots, x_n$ be a set of sample values.

- Sample mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Mode: most frequently occurring value in the data set
- Sample variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- Sample standard deviation: $s = \sqrt{s^2}$

Numerical Description of Data

Q_1 = lower quartile = the middle number of the half of the data below the median

$Q_2 = M$ = middle quartile (median) = middle number of the ordered data set

Q_3 = upper quartile = middle number of the half of the data above the median

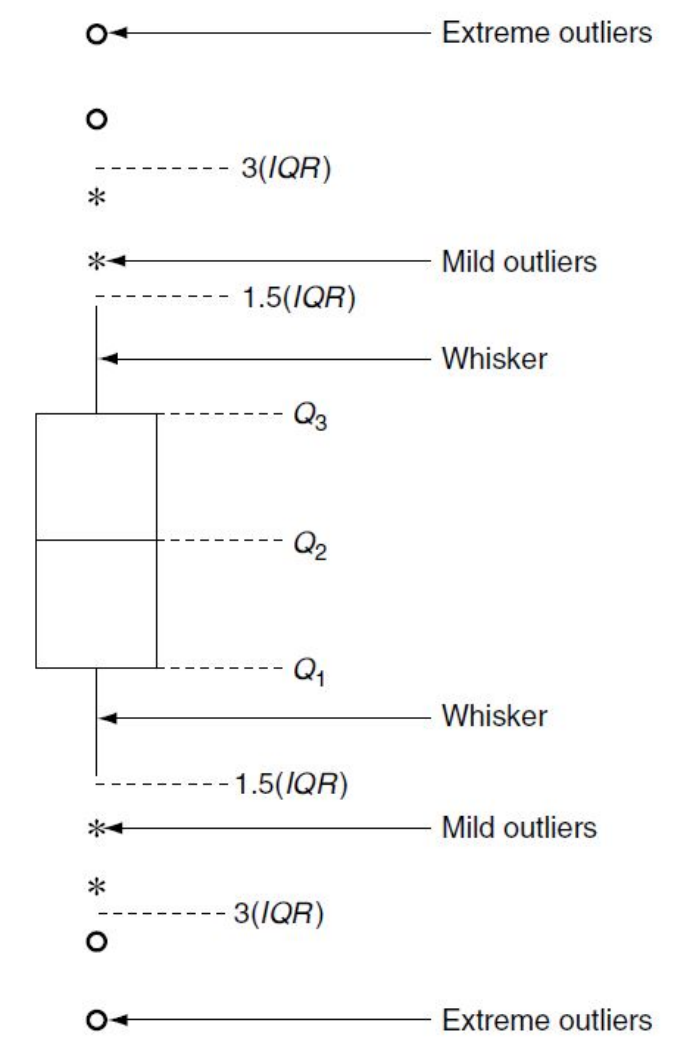
IQR = interquartile range = $Q_3 - Q_1$

Data points outside of $[Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR]$ are potential outliers

Graphical Description of Data

- → frequency table,
- → pie chart,
- → bar graph, Pareto chart,
- → histogram
- → boxplot

Graphical Description of Data -- Boxplot



How Inferences Are Made

The mechanism for making inferences is provided by the theory of probability.

The probabilist reasons from a known population to the outcome of a single experiment, the sample.

In contrast, the statistician utilizes the theory of probability to calculate the probability of an observed sample and to infer from this the characteristics of an unknown population.

Thus, probability is the foundation of the theory of statistics.

Theory & Reality

We are concerned with the theory of statistics and hence with models of reality.

We will postulate theoretical frequency distributions for populations and will develop a theory of probability and inference in a precise mathematical manner.

- → Sampling / Data collection
- → Specification of Population/Data
- → Estimation
- → Inference

You may also read chapter 8 from the book *Mathematics for Data Science.*