Data

STAT-S520

Arturo Valdivia

02-16-23

These slides complement ISI Chapter 7

Population and experiments

The following statements are based on ISI section 1.3.

- When we perform an experiment n times, we observe n outcomes.
  - We call this a sample of size n.
- A population is the set of all outcomes that might have been observed.
- Probability helps us describe the population and understand the data generating process that produced the sample.
- This linkage, of sample to population through probability, is the foundation on which statistical inference is based.

Population and random variables

The following statements are based on ISI Chapter 6

- The distribution of a random variable is a mathematical abstraction of the possible outcomes of an experiment.
- Having identified a random variable of interest, we will often refer to its distribution as the **population**.
- If the goal is to represent an entire population, the probability mass or density function would do.
- Often, we are interested in specific attributes of a population such as mean, median, standard deviation, etc.
  - It is easier to draw inferences about specific population attributes than about the entire population.

Random samples

- We plan to perform an experiment n times.
- Let $X_1, X_2, \ldots, X_n$ the random variables that represent the (possible) values and associated chances for those values before the experiment has been performed.
- We focus on experiments that are replicated under the same conditions each time, so the probability distribution for each random variable is the same and the outcome of any experiment does not influence the outcome of another experiment.
- We then say that the random variables are independent identically distribution (i.i.d) and write:

$$X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$$

Samples

- After the experiment is replicated n times, each random variable has assigned a value to the outcome of each replication:

$$X_1 = x_1, \ldots, X_n = x_n$$

- The set of observed values

$$\vec{x} = \{x_1, \ldots, x_n\}$$

  is called a sample.
- We want to use this sample to make inferences about the population.

The plug-in principle

- Using
$$\vec{x} = \{x_1, \ldots, x_n\}$$
we assume that each x_i has probability $1/n$.
- We construct an "empirical" probability distribution, $\hat{\mathbb{P}}_n$.
- We use $\hat{\mathbb{P}}_n$ to estimate population attributes of $\mathbb{P}$.

Plug-in estimates

- The plug-in estimate of μ, denoted $\hat{\mu}_n$, is the mean of the empirical distribution.

$$\hat{\mu}_n = \sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$ sample mean

- The plug-in estimate of σ^2, denoted $\hat{\sigma}_n^2$, is the variance of the empirical distribution.

$$\hat{\sigma}_n^2 = \sum_{i=1}^{n} (x_i - \hat{\mu}_n)^2 \cdot \frac{1}{n} = \frac{1}{n} \sum (x - \bar{x})^2 \Rightarrow$$

Note: This is not the sample variance

- The plug-in estimate of a population quantile is the corresponding quantile of the empirical distribution.
  - the sample median is the median of the empirical distribution.
  - the sample IQR is the IQR of the empirical distribution.

Plots

- Boxplot
- Normal probability plots (QQ plots)
- Kernel density estimates

Let's work in R