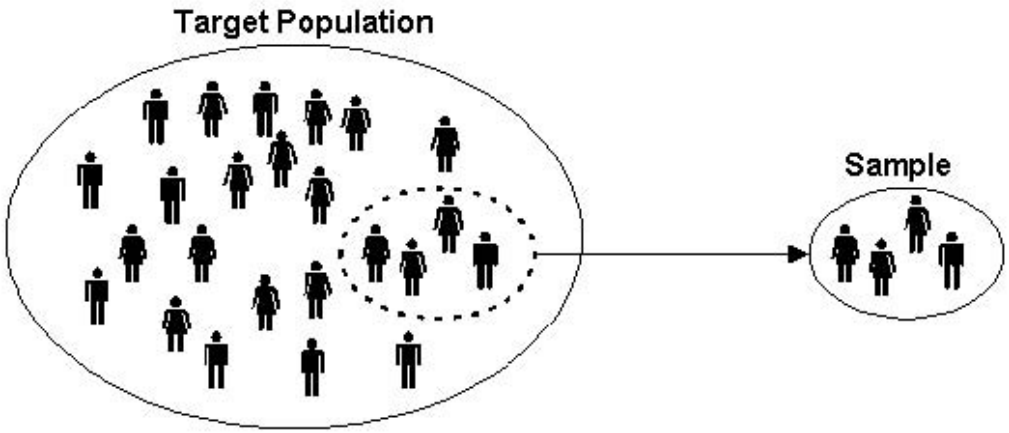
Basic concepts of Statistical Inference

MDT - Master Digital Transformation

May 29, 2020

Population and sample I

A **population** is the collection of all units of interest. A **sample** is a subset of the population.



In a **simple random sample** all members of the population are equally likely to be selected for inclusion in the sample. A simple random sample is meant to be representative of the population.

Population and sample II

The population of interest maybe an actual *physical population*, but it could also be a *hypothetical* population.

Examples of sample and population are:

- Political polls: The population will be all voters, whereas the sample will be the subset of voters we poll.
- Laboratory experiment: The population will be all the data we could have collected if we were to repeat the experiment a large number of times (infinite number of times) under the same conditions, whereas the sample will be the data actually collected by the one experiment.
- Quality control: The population will be the entire batch of items produced, by a machine, whereas the sample will be the subset of items we tested.
- Clinical studies: The population will be all the patients with the same disease, whereas the sample will be the subset of patients used in the study.

Statistical Inference I

Statistical inference aims to obtain information about an underlying population based on a sample collected from it.

In classical **parametric inference** the objective is the value of one (or more) parameter(s), and many classical inferential problems can be identified as being one of three types.

- **Point estimation** refers to providing a single "best guess" of the parameter of interest.
- **Interval estimation** refers to providing a "plausible" interval for the parameter of interest.
- In **hypothesis testing** we start with a hypothesis on the parameter of interest, and we ask if sample data provide sufficient evidence to reject such hypothesis.

Point estimation I

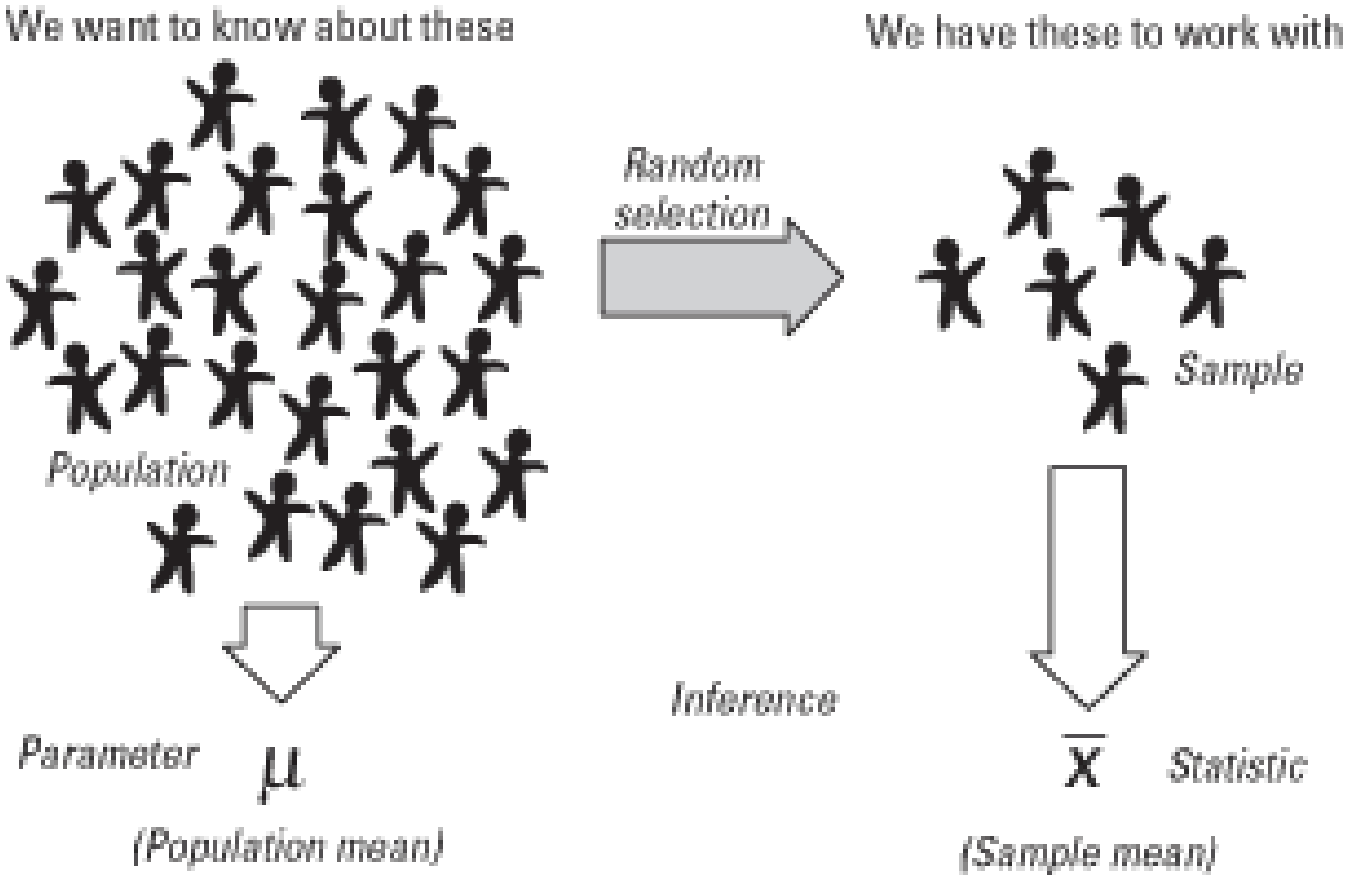
When we consider a variable X having in the population a given distribution, a simple random sample of size n from that population is the set of n independent copies of X, that is a set $X_1, \ldots, X_n$ of iid random variable.

In classical parametric inference on distribution, we assume that the probability distribution of X is known except for the value of a parameter, say θ (or more) which is unknown.

An **estimator** of θ is a function of $X_1, \ldots, X_n$. Once, we observe a realization $x_1, \ldots, x_n$ of the random sample, we have a corresponding value of the estimator which is the **estimate**.

Example. Assume that the heights of people in a population is described by a Normal distribution with known σ^2 and unknown μ. We take a random sample from the population, and using the heights of the people in the sample, we draw some type of conclusion about μ.

Point estimation II



Point estimation III

Let $X_1, \ldots, X_n$ be a random sample from a population with unknown mean μ, a possible estimator of μ is the sample mean

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}.$$

Let $X_1, \ldots, X_n$ be a random sample from a population with unknown variance σ^2, a possible estimator of σ^2 is the sample variance

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}.$$

Estimators are function of random variables, that is they are **statistics** and so they are also random variable having a probability distribution which is called the sampling distribution.

Evaluating estimators

For a parameter there are infinitely many possible estimators. So how can we make sure that we have chosen a good estimator? How do we compare different possible estimators? There are some desirable properties that we would like our estimators to have.

Let $\hat{\theta} = g(X_1, \ldots, X_n)$ be an estimator of θ. The bias of $\hat{\theta}$ is defined by

$$bias(\hat{\theta}) = E[\hat{\theta}] - \theta.$$

We say that $\hat{\theta}$ is unbiased for θ, if $E[\hat{\theta}] = \theta$, for each θ.

Example. $\bar{X}$ and S^2 are examples of unbiased estimators for the mean and the variance of a population, respectively.

The quality of a point estimate is sometimes assessed by the mean squared error, or mse which is defined as

$$MSE[\hat{\theta}] = E[(\hat{\theta} - \theta)^2]$$

The mse can be written as

$$MSE[\hat{\theta}] = V[\hat{\theta}] + bias^2[\hat{\theta}].$$

Confidence intervals I

Interval estimation aims to obtain a interval which contain the parameter with a given *level of confidence*.

For $\alpha \in (0, 1)$, a $1 - \alpha$ confidence interval for a parameter θ is an interval $C_n = (a, b)$ where $a = \mathrm{a}(X_1, \ldots, X_n)$ and $b = \mathrm{b}(X_1, \ldots, X_n)$ are functions of the data such that

$$P(\theta \in C_n) = 1 - \alpha,$$

In words, C_n traps θ with probability $1 - \alpha$. $1 - \alpha$ is said to be the **confidence level** of the interval.
Commonly, 95 percent confidence intervals are used. This corresponds to choosing $\alpha = 0.05$.

Confidence intervals II

Warning! There is much confusion about how to interpret a confidence interval. A confidence interval is not a probability statement about θ, since θ is a fixed while C_n is random! Some texts interpret confidence intervals as follows: if I repeat the experiment over and over, the interval will contain the parameter 95 percent of the time.

Example. Every day, newspapers report opinion polls. For example, they might say that "83 percent of the population favor arming pilots with guns." Usually, you will see a statement like "this poll is accurate to within 4 points 95 percent of the time." They are saying that 83 ± 4 is a 95 percent confidence interval for the true but unknown proportion p of people who favor arming pilots with guns.

Hypothesis test I

In hypothesis testing, we start with some default theory, called a **null hypothesis**, and we ask if the data provide sufficient evidence to reject the theory. If not we retain the null hypothesis.

More formally, suppose that we partition the *parameter space* Θ, i.e. the set of all possible values of θ, into two disjoint sets Θ_0 and Θ_1 and that we wish to test

$$H_0 : \theta \in \Theta_0 \quad \textit{versus} \quad H_1 : \theta \in \Theta_1.$$

We call H_0 the null hypothesis and H_1 the alternative hypothesis. Let $X_1, \ldots, X_n$ be a random sample and let $\mathcal{X}$ be the *sample space*. We test a hypothesis by finding an appropriate subset of outcomes $R \subset \mathcal{X}$ called the **rejection region**, such that

$$\textit{if} \;\; (X_1, \ldots, X_n) \in R \rightarrow \textit{reject} \;\; H_0$$

$$\textit{if} \;\; (X_1, \ldots, X_n) \notin R \rightarrow \textit{retain} \;\; H_0.$$

Hypothesis test II

Usually, the rejection region R is of the form

$$R = \{x : T(x) > c\}$$

where T is a *test statistic* and c is a *critical value*. The problem in hypothesis testing is to find an appropriate test statistic T, i.e. a function of the sample, and an appropriate critical value c.

Example. Let $X_1, \ldots, X_n \sim Bernoulli(p)$ be n independent coin flips. Suppose we want to test if the coin is fair. Let H_0 denote the hypothesis that the coin is fair and let H_1 denote the hypothesis that the coin is not fair. H_0 is called the null hypothesis and H_1 is called the alternative hypothesis. We can write the hypotheses as

$$H_0 : p = 1/2 \quad versus \quad H_1 : p \neq 1/2.$$

Letting $\hat{p}_n = (X_1 + \ldots + X_n)/n$, it seems reasonable to reject H_0 if $T = |\hat{p}_n - 1/2|$ is large.