# 7. INTRODUCTION TO STATISTICAL INFERENCE

**Study time: 50 minutes**

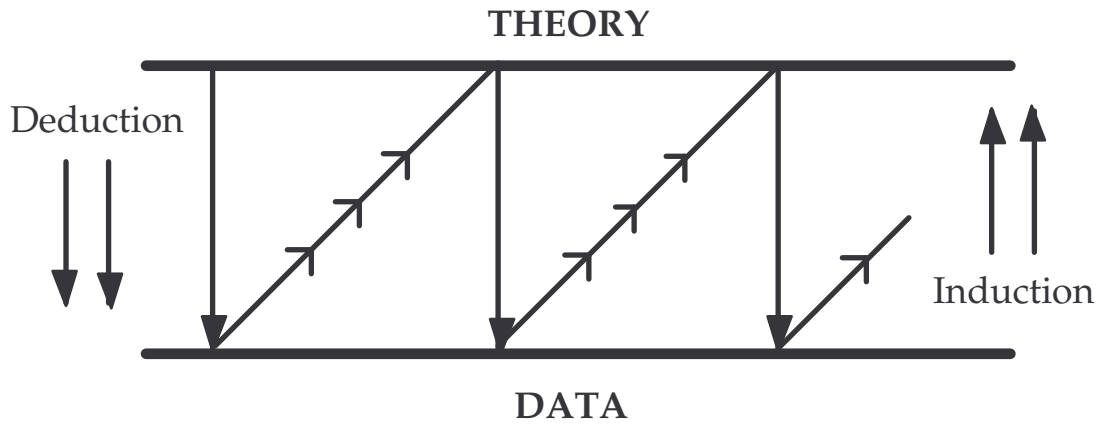
**Aim** - you will be able to

- understand the random sampling term
- use the sampling distribution and their properties

Explication

## 7.1. Introduction – The Scientific Method

Science is a process of systematic learning which proceeds by alternating between inductive and deductive methods of investigation.




The methods of induction and deduction are the connections between data and theory of a science. The deductive method proceeds in a logically consistent fashion to project what data should result from a particular theory. Induction is an informal process which tries to postulate some theory to reasonably explain the observed data.

Statistics to be a complete science must embody both inductive and deductive methods. The first topic of the course, exploratory data analysis, was an attempt to understand observed distributions of data and was therefore an inductive method. Without some theory of randomness however, the ability of EDA methods to induce precise explanations was limited. Therefore, we introduced the theory of probability and discussed its

applications to various hypothetical sample spaces. Probability theory is the basis of deductive methods of statistics.

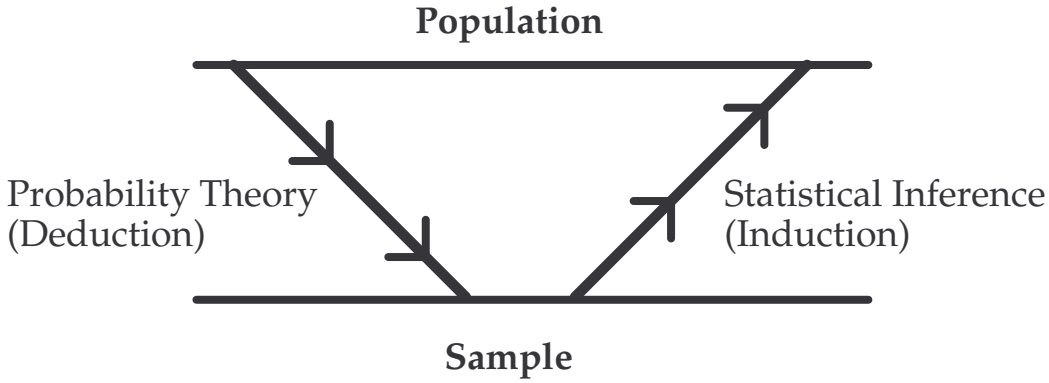
Probability theory proceeds by assuming a hypothetical sample space or population on which a probability measure is defined. Probability distributions of random variables defined on this sample space are then derived mathematically. The probability of any sample observation from the hypothetical population can then be determined.

Deductive Theory of Probability


Known Probability Distribution of Population

Sampling Distributions of Sample Statistics

If probability theory is the deductive method of statistics, then by implication, theory in statistical science must be represented by some well-defined population with a known probability distribution and data by the sample drawn from that population. Statistical inference then becomes the inductive methods for using sample data to make inferences about the probability distribution of the population from which the sample was drawn.

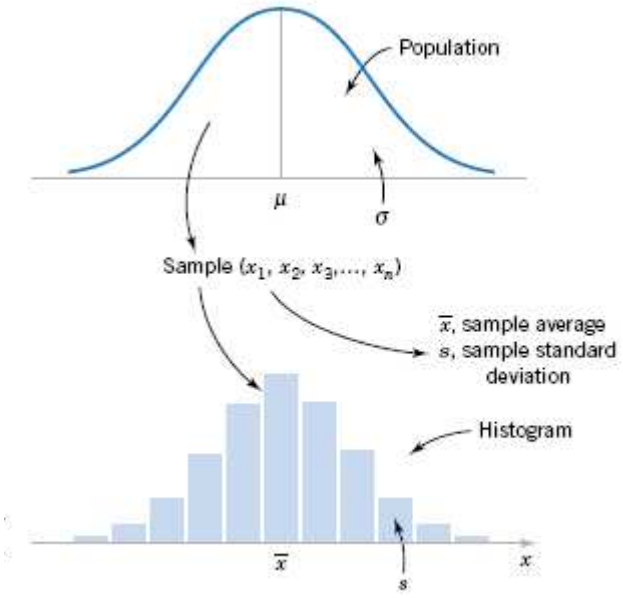



Statistical inference then is the inverse of probability theory. It is the process of making statements about an unknown population on the basis of a known sample from that population.


Viewing statistical science in this light may led us to ask why we have studying probability theory at all. Surely probability theory is answering the wrong question. The hypothetical problem of sampling from a known population never occurs in practice. In statistics, the population is always unknown and we not generate sample data in order to obtain information about the unknown population. In practice, is not statistical inference the only problem of statistical science?

## 7.2. Random Sampling

In most statistics problems, we work with a sample of observations selected from the population that we are interested in studying. Following Figure illustrates the relationship between the population and the sample.




We have informally discussed these concepts before; however, we now give the formal definitions of some of these terms.

*Definition*:
A **population** consists of the totality of the observations with which we are concerned.

In any particular problem, the population may be small, large but finite, or infinite. The number of observations in the population is called the **size** of the population. For example, the number of undefiled bottles produced on one day by a soft-drink company is a population of finite size. The observations obtained by measuring the carbon monoxide level every day is a

population of infinite size. We often use a **probability distribution** as a **model** for a population.

For example, a structural engineer might consider the population of tensile strengths of a chassis structural element to be normally distributed with mean _ and variance. We could refer to this as a **normal population** or a normally distributed population. In most situations, it is impossible or impractical to observe the entire population. For example, we could not test the tensile strength of all the chassis structural elements because it would be too time consuming and expensive. Furthermore, some (perhaps many) of these structural elements do not yet exist at the time a decision is to be made, so to a large extent, we must view the population as **conceptual.** Therefore, we depend on a subset of observations from the population to help make decisions about the population.

*Definition*:
A **sample** is a subset of observations selected from a population.

For statistical methods to be valid, the sample must be representative of the population. It is often tempting to select the observations that are most convenient as the sample or to exercise judgment in sample selection. These procedures can frequently introduce **bias** into the sample, and as a result the parameter of interest will be consistently underestimated (or overestimated) by such a sample. Furthermore, the behavior of a judgment sample cannot be statistically described. To avoid these difficulties, it is desirable to select a **random sample** as the result of some chance mechanism. Consequently, the selection of a sample is a random experiment and each observation in the sample is the observed value of a random variable. The observations in the population determine the probability distribution of the random variable. To define a random sample, let $X$ be a random variable that represents the result of one selection of an observation from the population. Let $f(x)$ denotes the probability density function of $X$. Suppose that each observation in the sample is obtained independently, under unchanging conditions. That is, the observations for the sample are obtained by observing $X$ independently under unchanging conditions, say, $n$ times. Let $X_i$ denote the random variable that represents the $i$th replicate. Then, $X_1, X_2 \ldots X_n$ is a random sample and the numerical values obtained are denoted as $x_1, x_2, \ldots, x_n$. The random variables in a random sample are independent with the same probability distribution $f(x)$ because of the identical conditions under which each observation is obtained. That is, the marginal probability density function of $X_1, X_2 \ldots X_n$ is

$$f(x_1),\ f(x_2),\ldots f(x_n)$$

respectively, and by independence the joint probability density function of the random sample is

$$f_{X1X2,\ldots Xn}(x_1, x_2, \ldots, x_n) = f(x_1)\, f(x_2)\ldots f(x_n).$$

*Definition*:
The random variables $X_1, X_2 \ldots X_n$ are a random sample of size $n$ if (a) the $X_i$'s are independent random variables, and (b) every $X_i$ has the same probability distribution.

To illustrate this definition, suppose that we are investigating the effective service life of an electronic component used in a cardiac pacemaker and that component life is normally distributed. Then we would expect each of the observations on component life in a random

sample of $n$ components to be independent random variables with exactly the same normal distribution. After the data are collected, the numerical values of the observed lifetimes are denoted as $x_1, x_2, \ldots, x_n$.

The primary purpose in taking a random sample is to obtain information about the unknown population parameters. Suppose, for example, that we wish to reach a conclusion about the proportion of people in the United States who prefer a particular brand of soft drink. Let $p$ represent the unknown value of this proportion. It is impractical to question every individual in the population to determine the true value of $p$. In order to make an inference regarding the true proportion $p$, a more reasonable procedure would be to select a random sample (of an appropriate size) and use the observed proportion $\hat{p}$ of people in this sample favoring the brand of soft drink.

The sample proportion, $\hat{p}$ is computed by dividing the number of individuals in the sample who prefer the brand of soft drink by the total sample size $n$. Thus, $\hat{p}$ is a function of the observed values in the random sample. Since many random samples are possible from a population, the value of $\hat{p}$ will vary from sample to sample. That is, $\hat{p}$ is a random variable. Such a random variable is called a **statistic.**

*Definition*:
A **statistic** is any function of the observations in a random sample.

We have encountered statistics before. For example, if $X_1, X_2 \ldots X_n$ is a random sample of size $n$, the **sample mean** $\overline{X}$ the **sample variance** $S^2$, and the **sample standard deviation** $S$ are statistics.

Although numerical summary statistics are very useful, **graphical displays** of sample data are a very powerful and extremely useful way to visually examine the data. In first lecture we presented a few of the techniques that are most relevant to engineering applications of probability and statistics.

## 7.3. Sampling Distribution

Let's assume that given random sample comes from normal distribution:

$\underline{X}=(X_1, \ldots, X_n)'$, $X_i \to N(\mu, \sigma^2)$

1. $\overline{X_n} = \dfrac{\sum_{i=1}^{n} X_i}{n} \to N\left(\mu, \dfrac{\sigma^2}{n}\right)$ … comes from central limit theorem for large number $n$

2. $Z_n = \dfrac{\overline{X_n} - \mu}{\sigma} \cdot \sqrt{n} \to N(0,1)$ … comes from a transformation of previous distribution

3. $\dfrac{S_n^2}{\sigma^2} \cdot (n-1) \to \chi^2(n-1)$ … was explained in $\chi^2$ discussion

where $S_n^2 = \dfrac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2$ ; $\dfrac{S_n^2}{\sigma^2} \cdot (n-1) = \sum_{i=1}^{n} \dfrac{(X_i - \overline{X})^2}{\sigma^2} = \sum_{i=1}^{n} \left( \dfrac{X_i - \overline{X}}{\sigma} \right)^2$

4. $\frac{\overline{X}_n - \mu}{s} \cdot \sqrt{n} \to t_{n-1}$ … was derived in the discussion about using of Student's distribution since:

$$\frac{\frac{\overline{X}_n - \mu}{\sigma} \cdot \sqrt{n}}{\sqrt{\frac{\frac{S^2}{\sigma^2} \cdot (n-1)}{n-1}}} = \frac{\overline{X}_n - \mu}{S} \cdot \sqrt{n}$$

Now assume two samples from the normal distribution
$\underline{X}=(X_1, \ldots, X_n)'$, $X_i \to N(\mu_1, \sigma_1^2)$, $\underline{Y}=(Y_1, \ldots, Y_m)'$, $Y_j \to N(\mu_2, \sigma_2^2)$. Then it holds:

5. $$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \to N(0,1)$$

$$\overline{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n}\right)$$

$$\overline{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{m}\right)$$

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

6. $$\frac{\frac{\frac{S_x^2}{\sigma_1^2} \cdot (n-1)}{n-1}}{\frac{\frac{S_y^2}{\sigma_2^2} \cdot (m-1)}{m-1}} = \frac{\frac{S_x^2}{\sigma_1^2}}{\frac{S_y^2}{\sigma_2^2}} \to F_{n-1,m-1}$$ … explained in F – distribution

Now assume that the variances are the same and unknown: $\sigma_1^2 = \sigma_2^2$. Then can be shown that it holds:

7. $$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_x^2(n-1) + S_y^2(m-1)}} \cdot \sqrt{\frac{n.m.(n+m-2)}{n+m}} \to t_{n+m-2}$$

## Summary of notions

The **random sample** is the special random vector whose elements are independent random variables with the same probability distribution.

If the random sample comes from the normal distribution of probability we can derivate other significant statistics with known distribution from given random sample, e.g. t-statistics $\frac{\overline{X}_n - \mu}{s} \cdot \sqrt{n} \to t_{n-1}$ or two-sample t-statistics:

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_x^2(n-1) + S_y^2(m-1)}} \cdot \sqrt{\frac{n.m.(n+m-2)}{n+m}} \to t_{n+m-2} .$$

These other statistics will be later used for the construction of interval estimation or for hypothesis testing.

## Questions

1. What is the statistical induction?
2. Characterize the term a random sample.