STATISTICS

Module 2 *of*
MATHEMATICS AND STATISTICS

Marina Santacroce

marina.santacroce@unibocconi.it
Department of Decision Sciences, Bocconi University

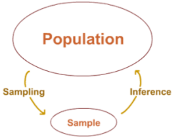
L14: Distributions of sampling statistics

References: S.Ross *Introduction to* Probability and Statistics for Engineers and Scientists: 6.1, 6.2, 6.4, 6.5

Outline

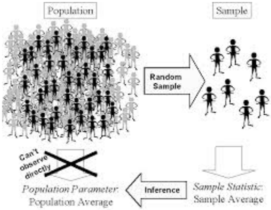
Inference and sampling statistics

Inferential statistics deals with drawing conclusions for a population by observing a sample of individuals suitably chosen.



- A *sample (random sample) from the distribution* F is a set of independent random variables $X_1, \ldots X_n$, with a common distribution F.

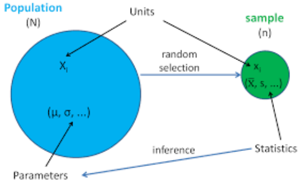
The sample is used to make *inference* on F.



Inference and sampling statistics

The population distribution F is not completely specified.

- If F is known up to some unknown parameters $\Rightarrow$ *parametric inference*



- If F is unknown $\Rightarrow$ *nonparametric inference*

Statistics distribution

- A *statistic* is a random variable whose value is determined by the values of the sample $X_1, \dots, X_n$.
- We will concentrate on the distribution of two *statistics*:

  the sample mean and the sample variance.

  In the study of the *annual income, weight, social media usage, ...* in a certain population,

$\Rightarrow$ the annual income (weight, social media usage,...) related to each individual is thought as a value assumed by a random variable with

(population) mean μ and *(population) variance* σ^2.

Given a random sample $X_1, \dots, X_n$ from the population we define:

$$\overline{X} = \frac{1}{n}\sum_{i:1}^{n} X_i \qquad \text{(sample mean)}$$

and

$$S^2 = \sum_{i:1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n-1} \qquad \text{(sample variance).}$$

Sample mean: properties and distribution

- the *expectation of the sample mean* is equal to the *population mean*:

$$\mathbb{E}(\overline{X}) = \mu$$

- the *variance of the sample mean* is equal to the *population variance divided by n*:

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

- If the random sample is from a *normal* population, i.e. $X_i \sim \mathrm{N}(\mu, \sigma^2)$, the *sample mean is also normal*:

$$\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$$

CLT (*without normal assumption*) the distribution of the sample mean suitably standardized *tends* to the one of *the standard normal* as n goes to infinity:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z \sim \mathrm{N}(0,1), \text{ as } n \to \infty$$

$\Rightarrow$ we can *approximate* the distribution of the sample mean with a normal with parameters μ and $\frac{\sigma^2}{n}$ for n *sufficiently large* (regardless the original distribution of the population).

Sample variance and (joint) distribution

Recall the definition of the sample variance $S^2 = \sum_{i:1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$ then

$$\mathbb{E}(S^2) = \mathbb{E}\left(\sum_{i:1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n-1}\right) = \sigma^2.$$

Theorem: If $X_1, \ldots, X_n$ is a random sample from a *normal* population with mean μ and variance σ^2, then

$\Rightarrow$ $\overline{X}$ and S^2 are independent random variables:

- the sample mean $\overline{X}$ is normally distributed with mean μ and variance $\frac{\sigma^2}{n}$;
- the sample variance S^2 multiplied by $\frac{n-1}{\sigma^2}$ is a chi square with $n-1$ degrees of freedom;

$$\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n}) \qquad \text{and} \qquad \frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}.$$

Corollary: Under the same hypotheses of the theorem, $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ has a t-distribution with n-1 degrees of freedom,

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Sampling from a finite population

- If the population has N individuals a random sample of n elements is one of $\binom{N}{n}$ possible choices of n individuals from the population.
- The random sample can be also selected sequentially, selecting the first element at random among N, the second at random among $N-1$, and so on.

Suppose *pN individuals in the population have a certain characteristic*, let

$X_i = 1$ if the element i in the sample has the characteristic and 0 otherwise.

$X = \sum_{i:1}^{n} X_i \Rightarrow$ *number of individuals in the sample with the characteristic*;

$\overline{X} \Rightarrow$ *proportion of individuals in the sample with the characteristic*.

- If $X_1, \ldots, X_n$ is a random sample from a *finite* population of size N and with Np having a certain characteristic, then

$\Rightarrow$ X_i are Bernoulli random variables with parameter p;

$\Rightarrow$ the X_i, $i : 1, \ldots n$ are **not** independent;

$\Rightarrow$ if N is adequately large with respect to n, the X_i are *approximately independent* and X is *approximately binomial* with parameters n and p, with $\mathbb{E}(X) = np$ and $\mathrm{Var}(X) = np(1-p)$.

$\Rightarrow$ The sample mean $\overline{X}$ has $\mathbb{E}(\overline{X}) = p$ and *(approximately)* $\mathrm{Var}(\overline{X}) = \frac{p(1-p)}{n}$.

Sampling from a finite population: exercise

Suppose the 45% of the population of a certain town favors a certain candidate in an upcoming election. If a random sample of size 200 is chosen out of the 100000 inhabitants of the town, find:

a) the expected value and standard deviation of the number of members of the sample which favors the candidate;

b) the probability that more than half the members of the sample favor the candidate.