Topic 2: INTRODUCTION TO STATISTICAL INFERENCE

2.1. KEY ISSUES

A) ***Objetive***: to study one or more variables in a population

B) ***Method***:

1. Sampling selection:
   i. Simple random sampling (s.r.s.)
   ii. Stratified sampling
   iii. Cluster sampling and other techniques
2. We asume the variable following certain probability distribution
3. The parameters in the population are estimated based on the sample (parametric statistics). Then we study:
   i. Properties of point estimators
   ii. Methods for obtaining the point estimator (moments and maximum-likelihood)
   iii. Estimation procedures:
      - Point estimation and
      - Confidence interval estimation
4. Hypothesis testing is applied
5. Procedures for model criticism (nonparametric statistics):
   i. Goodness of fit tests
   ii. Tests of Independence (runs test, autocorrelation test)

iii. Nonparametric tests for paired samples (sign test and Wilcoxon signed rank test) and nonparametric tests for independent random samples (Mann-Whitney U test and Wilcoxon rank sum test).

Example:

A) Variable to study: the income earned by an individual in Madrid

B) Method:

1. A s.r.s. of n=1000 people must be selected
2. The variable is assumed to follow a N(μ;σ) distribution.
3. Parameters μ and σ^2 are estimated:

i. Maximum likelihood estimators are obtained:

$$\mu_{ML}^{*} = \bar{x}$$

$$\sigma_{ML}^{2^{*}} = S^2$$

ii. ML estimators are unbiased, efficient and normally distributed random variables in big samples.

iii. Estimation process:

a. Point estimation: sample mean and simple variance realized in a given simple.

b. Interval estimation:

$$\mu \in \left[\bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}\right]_{\gamma} \quad with\ \sigma\ known$$

$$\mu \in \left[\bar{x} \pm t \cdot \frac{s_1}{\sqrt{n}}\right]_{\gamma} \quad with\ \sigma\ unknown$$

$$\sigma^2 \epsilon \left[\frac{n \cdot s^2}{k_2}; \frac{n \cdot s^2}{k_1} \right]_\gamma$$

4. Hipothesis testing:

$$H_0: \mu = 30000 \text{ €}$$

$$H_1: \mu \neq 30000 \text{ €}$$

5. A normality goodness of fit test is carried out.

2.2. SAMPLING TECHNICS. DISTRIBUTIONS OF SAMPLING STATISTICS

THE FIRST STEP: TO TAKE A SAMPLE FROM A POPULATION

- *Population*: a group of elements with one or more common features (people, families, firms, objects, etc.).

X_1 X_2 X_3 … X_i …X_N

i. We are interested in studying some of those features
ii. Ideal situation: to implement a census
iii. But when N is big this may take a long time and be very expensive
iv. In these circumstancies a sample must be selected

- *Sample*: a set of data selected from a population, with the intention of being representative of the latter:

x_1 x_2 $x_3 \ldots x_i \ldots$ x_n

- The effectiveness of inference is based on this process
- Ideal situation: the sample becomes a population on a small scale
- What if there are deviations from such perfect scenario. The process is under control whenever:
    i. Those deviations aren't systematic and
    ii. Due to randomness

SAMPLING PROCEDURES

A) Non probabilistic sampling[1]:

- The selection system is *subjective*
- Convenience sampling, judgement sampling, snowball sampling.
- It may be useful:
    a. In qualitative, pilot or exploratory studies
    b. When the researcher does not wish to make inferences over the population

B)Probabilistic sampling:

- It is based on *randomness*
- Hence, sampling becomes a random phenomenon
- Each element selected from the population is an event with a probability of happening
- And so is a given sample
- It is an *objective* (scientific) method allowing for the error produced to be measured and controlled.

[1] https://explorable.com/non-probability-sampling

PROBABILISTIC SAMPLING TECHNICS

B.1) **Simple Random Sampling (s.r.s.)**[2]: Each element in the population has the same probability of being selected.

a. This allows for the sample resembling the structure of the population, *on average*
b. It is appropiate when there is *homogeneity* in the population regarding the variable under study
c. It is the easiest sampling method
d. Two requirements: all the members of the population must be in a list and a random selection tool has to be implemented
e. Two kind of sampling:
   i. ***WITH REPLACEMENT***:
      1. The same element can be repeated *n* times
      2. It simplifies the calculus, provided that the elements in the sample are *INDEPENDENT*
      3. Appropiate in infinite populations

   ii. ***WITHOUT REPLACEMENT***:
      1. Appropiate in finite populations when the sample size is bigger than 5% of the entire population
      2. In these cases the estimation is more accurate, but calculus are more complicated due to independence is not verified

[2] http://www.statcan.gc.ca/edu/power-pouvoir/ch13/prob/5214899-eng.htm

f. Each observation in the sample is a random variable following the same distribution than the variable under study:

$$Let\ \xi\ be\ a\ random\ variable\ with\ density\ function\ f(x;\theta)$$
$$and\ X = (x_1, x_2, \dots x_n)\ a\ s.r.s.$$

Then:

$$f(x_1;\theta) = f(x_2;\theta) = \cdots = f(x_n;\theta) = f(x;\theta)$$

g. Moreover, in case of replacement, the density function of the whole sample may be written as follows:

$$f(x_1; x_2 \dots x_n; \theta) = f(x_1;\theta) \cdot f(x_2;\theta) \cdot \dots \cdot f(x_n;\theta)$$

Hence, any observation in a s.r.s. with replacement is *independent*, *identically distributed* following the same probability distribution than that of the population from which has been selected:

$$X = (x_1, x_2, \dots x_n)\ s.r.s. \rightarrow x_i \sim i.i.d.\ f(x_i;\theta)$$

B.2) **Stratified sampling**: The population is divided into mutually exclusive strata whose components are expected to behave in a different way regarding the variable or variables under consideration due to certain characteristic (age, gender, region of residence,…). Hence, the population is heterogeneous.

a. Strata are homogeneus inside them but heterogeneus among them
b. After the strata are defined, a sample must be taken from each one of them. Although any sampling technique can be used, it is common to use s.r.s.
c. The number of observations to select from each strata can be proportional either to its size or to its variability
d. We can draw inferences about specific subgroups in the population
e. This technique requires lower sample size than the s.r.s. method

B.3) **Cluster sampling**: The population can be clasified in different clusters, which are groups of heterogeneous elements, hence providing a similar variability to that existing in the population analyzed.

a. Then a number of clusters are selected randomly
b. After that, one may either survey all the units included in the selected clusters or just a s.r.s.
c. It is cheaper than s.r.s. in the whole population

B.4) **Sistematic sampling**: it is a kind of s.r.s. The elements in the population must be listed

a. Let *k* be the integer nearer to:

$$\frac{N}{n}$$

b. Then, an element of the population is chosen at random among the first *k*. That one will be the first observation in the sample: x_1

c. The second observation in the sample will be that occupying the $x_1 + k$ position

d. The third one will fall at the $x_1 + 2k$ position, and so on till completing the sample: $x_1 + (n\text{-}1)k$

Final considerations:

- If there is lack of information about the variable under study in the population we will apply s.r.s.
- In other case, population is divided in groups either homogeneus (stratum) or heterogeneus (clusters) and then apply s.r.s.
- Whenever a sampling procedure is implemented, the technique employed must be clearly explained.
- There are no good or bad samples, just good or bad sampling procedures.
- The bigger the sample size, the better the estimation. However the improvement in precision decreases from certain sample size
- The expenses involved must be taking into account.

SECOND STEP: PROBABILITY DISTRIBUTION

- A given distribution is assumed for the variable under study.
- Later on this hypothesis will be contrasted using goodness of fit tests

THIRD STEP: ESTIMATION

ESTIMATOR (or point estimator)

- An estimator is a statistic aimed to give values to an unknown parameter in the population
- It is a function of the sampling data not having unknown parameters:

$$\sum x_i \qquad \bar{x} \qquad s^2$$

- They are random variables.
- Its probability distribution is called sampling distribution and depends on that one followed by the population.
- The values taken by the estimator come from all posible samples with a given size that can be selected.
- Corresponding probabilities depend on the sampling technique used.

SAMPLING DISTRIBUTIONS OF ESTIMATORS (s.r.s.)

ONE POPULATION

Sample mean:

$$\bar{x} = \frac{\sum x_i}{n}$$

1) General case: ξ is any RV with mean μ and variance σ^2:

$$E(\bar{x}) = \mu \qquad V(\bar{x}) = \frac{\sigma^2}{n}$$

If n is low the sampling distribution depends on that one followed by the population

2) If n is high (n≥30) then CLT applies:

$$\bar{x} \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

3) ξ follows a N(μ;σ); then due to the additive property (no matters n):

$$\bar{x} \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

4) ξ follows a N(μ;σ) with σ^2 unknown. In such a case, we use the *bias-corrected sample standard deviation* s_1 instead of σ:

$$\frac{\bar{x} - \mu}{\frac{s_1}{\sqrt{n}}} \sim t_{n-1}$$

or alternatively, using the *sample standard deviation s*:

$$\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n-1}}} \sim t_{n-1}$$

Sample variance

$$s^2 = \frac{\sum(x_i-\bar{x})^2}{n} = \frac{\sum x_i^2}{n} - \bar{x}^2$$

With mean:

$$E(s^2) = \frac{n-1}{n}\sigma^2$$

And variance (general case):

$$V(S^2) = \frac{m_4-\sigma^4}{n} - \frac{2(m_4-2\sigma^4)}{n^2} + \frac{m_4-3\sigma^4}{n^3}$$

where $m_r = E[x-E(x)]^r$

And variance (only when $\xi \sim N(\mu;\sigma)$):

$$V(s^2) = \frac{2(n-1)\sigma^4}{n^2}$$

Moreover (**key issue**) if $\xi \sim N(\mu;\sigma)$:

$$\frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}$$

hence:

$$s^2 \sim \frac{\sigma^2}{n}\chi^2_{n-1}$$

Bias-corrected sample variance

$$s_1^2 = \frac{ns^2}{n-1} = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

with mean:

$$E(s_1^2) = \sigma^2$$

and only in normal populations:

$$\frac{(n-1)s_1^2}{\sigma^2} \sim \chi^2_{n-1} \quad ; \qquad V(s_1^2) = \frac{2\sigma^4}{n-1}$$

Sample proportion

Let a population following a Bernouilli distribution:

$$\xi \sim B(1; p)$$

And the R.V. X= "number of successes in the sample", which follows a B(n,p).

We define $\hat{p}$="proportion of successes obtained in n elements".

Then:

$$\bar{x} \equiv \hat{p} = \frac{\sum x_i}{n} = \frac{B(n;p)}{n}$$

$$E(\hat{p}) = p \qquad V(\hat{p}) = \frac{pq}{n}$$

Furthermore, if *n* is high (base on CLT):

$$\hat{p} \sim N\left(p; \sqrt{\frac{pq}{n}}\right)$$

TWO POPULATIONS

A) Two normal populations (independent samples).

Two independent samples coming, respectively, from the two variables to be compaired are taken:

$$Let\ (x_1, x_2, \dots x_i, \dots x_n)\ be\ a\ s.r.s\ with\ x_i \sim N(\mu_x; \sigma_x) \forall i$$

$$Let\ (y_1, y_2, \dots y_j, \dots y_m)\ be\ a\ s.r.s\ with\ y_j \sim N(\mu_y; \sigma_y) \forall j$$

Difference between two sample means

Case 1) The population variances are known:

$$\bar{x} - \bar{y} \sim N\left(\mu_x - \mu_y; \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}\right)$$

The difference between two proportions being a particular case:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2; \sqrt{\frac{p_1 q_1}{n} + \frac{p_2 q_2}{m}}\right)$$

Case 2) The population variances are unknown but equal

$$\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{s^* \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

$$s^* = \sqrt{\frac{(n-1)s_{1x}^2 + (m-1)s_{1y}^2}{n+m-2}}$$

or:

$$s^* = \sqrt{\frac{ns_x^2 + ms_y^2}{n+m-2}}$$

Case 3) The population variances are unknown and different:

$$\frac{(\bar{x}-\bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_{1x}^2}{n} + \frac{s_{1y}^2}{m}}} \sim t_{n+m-2-g}$$

$$g = \frac{\left[(m-1)\frac{s_{1x}^2}{n} - (n-1)\frac{s_{1y}^2}{m}\right]^2}{(m-1)\frac{s_{1x}^4}{n^2} - (n-1)\frac{s_{1y}^4}{m^2}}$$

Quotient between sample variances

$$\frac{s_{1x}^2 / \sigma_x^2}{s_{1y}^2 / \sigma_y^2} \sim F_{(n-1),(m-1)}$$

or:

$$\frac{{ns_x^2}/{(n-1)\sigma_x^2}}{{ms_y^2}/{(m-1)\sigma_y^2}} \sim F_{(n-1),(m-1)}$$

B) Two normal populations (dependent samples, matched pairs).

The variables are normal distributed and the observations in one sample depend on the observations in the other sample. Concretely, they are matched pairs of data.

$$Let\ (x_1, x_2, \dots x_i, \dots x_n)\ be\ a\ s.r.s\ with\ x_i \sim N(\mu_x; \sigma_x) \forall i$$

$$Let\ (y_1, y_2, \dots y_j, \dots y_n)\ be\ a\ s.r.s\ with\ y_j \sim N(\mu_y; \sigma_y) \forall j$$

Observe that $n = m$.

Difference between two sample means

First of all, we build n pairs with the respective differences:

$$d_i = x_i - y_i$$

Second, we come up with the mean and the bias-corrected standard deviation

$$\bar{d} = (\bar{x} - \bar{y})$$

$$s_{1d} = \sqrt{\frac{n}{n-1}\left(\frac{\sum d_i^2}{n} - \bar{d}^2\right)}$$

Then we get the corresponding distribution:

$$\frac{\bar{d} - (\mu_x - \mu_y)}{s_{1d}/\sqrt{n}} \sim t_{n-1}$$

or:

$$\frac{\bar{d} - (\mu_x - \mu_y)}{s_d/\sqrt{n-1}} \sim t_{n-1}$$

C) Two non normal populations (independent samples with big size).

Difference between two sample means

<u>General case:</u> two independent samples with n and m high.

The difference between two sample means will be aproximated to a normal distribution with mean μ_x-μ_y and variance being equal to the sum of the estimated variances of the respective sample means:

$$\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_{1x}^2}{n} + \frac{s_{1y}^2}{m}}} \sim N(0;1)$$