SAMPLING AND ESTIMATION (FR: ECHANTILLONAGE ET ESTIMATION)

In this chapter, we will be interested in studying a given character in a population, whose proportion is p. In some cases, p is known (sampling), in some others, just supposed known (decision making) and sometimes unknown (estimation). This character is studied on samples of size n of the population.
In the whole chapter, n and p satisfy the conditions :

$n \geq 30$ $\quad\quad$ $np \geq 5$ $\quad\quad$ $n(1-p) \geq 5$

Examples :

1. We want to decide if a given coin is fair or not. We will suppose it is and test this assumption. It is a situation of **sampling** (and decision making).
2. Out of 100 TV's tested before delivery, 5 have a problem. We want to induce the proportion of defective TV's in this production. It is an **estimation** situation.

1. RELATIVE FREQUENCY RANDOM VARIABLE

Property :
The random variable X which measures the number of individuals having the studied character in one sample follows a binomial distribution with parameters (n,p).
$X \sim \mathscr{B}(n; p)$

Definition :
The random variable F which measures the relative frequency of individuals having the studied character in one sample is $F = \dfrac{X}{n}$.

Notes :

1. F takes the values $0, \dfrac{1}{n}, \dfrac{2}{n}, ... \dfrac{n-1}{n}, 1$.
2. F doesn't follow a binomial distribution since its values are not integers.

2. FLUCTUATION INTERVAL AND DECISION MAKING

2.1 FLUCTUATION INTERVAL

In this paragraph, p is known.

Definition :
An asymptotic fluctuation interval of the random variable F at threshold 95% is an interval (defined from n and p) which contains F with a probability which gets closer to 0.95 as the value of n increases.

Note :
It is the interval you've calculated in 1° with the binomial distribution using your calculator.

Definition :
The asymptotic fluctuation interval at threshold 0.95 of the relative frequency random variable F is defined by : $\left[p-1.96\times\frac{\sqrt{p(1-p)}}{\sqrt{n}};p+1.96\times\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$

Note : It is smaller (so more accurate) than the one seen in 2° : $\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$

Example:

An urn contains a large amount of white and black balls. The proportion of white balls is 0.4 .
We take out of the urn, randomly, 50 balls and we want to find the fluctuation interval at threshold 0.9 (so with $\alpha=0.1$). Thanks to the calculator, we have $u_{0.1}\approx1.645$ with 3dps and thus $I_{50}=\left[0.4-1.645\times\frac{\sqrt{0.4\times0.6}}{\sqrt{50}};0.4+1.645\times\frac{\sqrt{0.4\times0.6}}{\sqrt{50}}\right]=[0.286;0.514]$.

"Picking 50 balls, the relative frequency of the white balls is in the interval $[0.286;0.514]$ white a probability roughly 0.9"

Note : With 500 balls, we would get $I_{500}=[0.364;0.436]$. For the same threshold 0.9, the interval is more than 3 times smaller.

2.2 DECISION MAKING

In this paragraph, the proportion of the studied character is <u>supposed to be equal</u> to p.

We measure the relative frequency f of the studied character in a sample with size n and we calculate the fluctuation interval at threshold 0.95 as defined previously. Then we apply the following rule :

Decision rule :
If f belongs to the asymptotic fluctuation interval at threshold 0.95, then we accept the hypothesis we made on p.
If f doesn't belong to the asymptotic fluctuation interval at threshold 0.95, then we reject the hypothesis we made on p (with a 5 % risk to do a mistake).

Note :

1. In the first case, the error risk is unknown.
2. 5% mistake risk means that the probability one has to reject wrongly the hypothesis made on p (**knowing that** it is true) is roughly 5%. (It's a conditional probability)

3. ESTIMATION AND CONFIDENCE INTERVAL

In this paragraph, the proportion p of the studied character is unknown.

Definition :

The confidence interval at threshold 0.95 is defined by : $\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$

f being the observed relative frequency on a sample with size n.

Example :

A wholesaler has just received 2.5 t of potatoes (in 25kg bags) which are meant to have a size 35-55. He takes out every bag one potato and measure it : 17 out of the 100 potatoes have not the expected size. Do you think he has to accept this batch ?

Bullet points of the chapter

- ✓ Knowing the asymptotic fluctuation interval at 95%
- ✓ Estimating with an interval an unknown proportion from a sample
- ✓ Calculating the required size of a sample to get, with a given accuracy, an estimation of a proportion with a confidence level 95%