TOPIC 4: ESTIMATION WITH CONFIDENCE INTERVALS

4.1. MAIN TOPICS

Group of observations among we expect to find the real value of the parameter θ being estimated, with a γ level of confidence

Two key issues:

- The precision: it refers to the length of the interval. The shorter the interval the higher the precision and viceversa, the longer the interval the lower the precision
- The level of confidence γ established for the interval containing the parameter:

$$\gamma = 1 - \alpha$$

Where α is the level of significance

Both elements counteract: if we increase the level of confidence then the precision diminishes. Meanwhile if we wish a shorter interval then the level of confidence lowers.

There is an exception: If we rise the sample size (n) we can either increase the level of confidence, keeping the precision, or gain in precision, keeping the level of confidence.

4.2. PIVOTAL QUANTITY METHOD

Let ξ be a random variable with density function *f(x;θ)* with θ unknown.

Then a function *g(X;θ)* with a known probability distribution is defined (*X* is a s.r.s.). That function is called pivot.

Then two values *a* and *b* are obtained verifying that:

$$P(a \leq g(X;\theta) \leq b) = \gamma$$

Being chosen in such a way that each one of them leave a probability of α/2 in both tails of the pivot's distribution.

Assuming that g is a monotone continuous function in θ, we can find the parameter through this expresion:

$$\theta \epsilon [g(X,a); g(X,b)]_{\gamma}$$

Where *g(X,a)* and *g(X,b)* are statistics.

Before the sample has been selected, γ represents the probability for the interval including θ. After, γ is the level of confidence we have for θ being included inside the interval, meaning the percentage of intervals including θ, having sampled the population a very large number of times.

4.3. CONFIDENCE INTERVALS FOR PARAMETERS OF NORMAL DISTRIBUTIONS

ONE POPULATION

❖ C.I. for the mean μ of a normal distribution with a known σ^2 :

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0;1)$$

$$P(-z \leq N(0;1) \leq z) = \gamma$$

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

The precision of the estimation depending on *z*, *n* and *σ*

The sampling error *ε* (also called margin error) is:

$$\varepsilon = z \cdot \frac{\sigma}{\sqrt{n}}$$

It determines the length of the interval, hence the precision of the estimation:

$$lenght\ of\ a\ C.I. = 2\varepsilon$$

Then we can find *n* based on *z* and *ε*:

$$n = \frac{z^2 \cdot \sigma^2}{\varepsilon^2}$$

❖C.I. for the mean μ of a normal distribution with σ^2 unknown

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

$$P(-t \leq t_{n-1} \leq t) = \gamma$$

$$\mu \, \epsilon \left[\bar{x} \mp t \cdot \frac{s_1}{\sqrt{n}} \right]_\gamma$$

The precision of the estimation depending on *t*, *n* and s_1

The sampling error being:

$$\varepsilon = t \cdot \frac{s_1}{\sqrt{n}}$$

And the corresponding calculation of *n*:

$$n = \frac{t^2 \cdot s_1^2}{\varepsilon^2}$$

❖ C.I. for the variance σ^2 of a normal population

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

$$P(k_1 \leq \chi^2_{n-1} \leq k_2) = \gamma$$

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

TWO POPULATIONS: INDEPENDENT SAMPLES

❖ C.I. for the difference between two means

Two independent simple random samples are obtained with sizes n and m, means $\bar{x}$ and $\bar{y}$ and bias-corrected standard deviations s_{1x} and s_{1y} respectively.

❖ Case 1) Population variances are known:

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

$$P(-z \leq N(0;1) \leq z) = \gamma$$

$$(\mu_x-\mu_y) \in \left[(\bar{x}-\bar{y}) \mp z \cdot \sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}\right]_\gamma$$

❖ Case 2) Population variances are unknown but equal:

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

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

$$P(-t \leq t_{n-1} \leq t) = \gamma$$

$$\left(\mu_x - \mu_y\right) \in \left[(\bar{x} - \bar{y}) \mp t \cdot s^* \sqrt{\frac{1}{n} + \frac{1}{m}}\right]_\gamma$$

❖ Case 3) Population variances are unknown and different:

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

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

$$P\left(-t \leq t_g \leq t\right) = \gamma$$

$$\left(\mu_x - \mu_y\right) \in \left[(\bar{x} - \bar{y}) \mp t \cdot \sqrt{\frac{s_{1x}^2}{n} + \frac{s_{1y}^2}{m}}\right]_\gamma$$

❖ C.I. for the quotient between two variances:

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

$$P\left(k_1 \leq F_{(n-1),(m-1)} \leq k_2\right) = \gamma$$

$$\frac{\sigma_x^2}{\sigma_y^2} \in \left[\frac{s_{1x}^2}{k_2 \cdot s_{1y}^2}; \frac{s_{1x}^2}{k_1 \cdot s_{1y}^2}\right]_\gamma$$

TWO POPULATIONS: DEPENDENT SAMPLES

❖ C.I. for the difference between two means

Two dependent simple random samples of size *n* are selected from each population. Then the difference between any pair of observations is calculated:

$$d_i = x_i - y_i \qquad \forall i = 1..n$$

After that the mean and the bias-corrected standard deviation of those differences is constructed:

$$\bar{d} = \bar{x} - \bar{y} \qquad s_{1d} = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n-1}}$$

Finally the C.I. is elaborated as follows:

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

$$P(-t \leq t_{n-1} \leq t) = \gamma$$

$$(\mu_x - \mu_y) \in \left[\bar{d} \mp t \cdot \frac{s_{1d}}{\sqrt{n}}\right]_\gamma$$

4.4. CONFIDENCE INTERVALS FOR PARAMETERS OF NON-NORMAL DISTRIBUTIONS BUT TAKING LARGE SAMPLES

❖ C.I. for the mean of a any population selecting large samples (*n ≥ 30*)

$$\frac{\bar{x} - \mu}{s_1/\sqrt{n}} \sim N(0;1)$$

$$P(-z \leq N(0;1) \leq +z) = \gamma$$

$$\mu \in \left[\bar{x} \mp z \cdot \frac{s_1}{\sqrt{n}}\right]_\gamma$$

The sampling error being:

$$\varepsilon = z \cdot \frac{s_1}{\sqrt{n}}$$

The sample size needed for a given level of confidence γ and a certain error ε arises from:

$$n = \frac{z^2 \cdot s_1^2}{\varepsilon^2}$$

❖ C.I. for the proportion *p* of a dichotomic population (*n ≥ 30*)

As it was explained before, the sampling distribution of the sample proportion when *n* is large is:

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

Estimating the variance and standardizing we obtain the pivot and then the c.i.:

$$\frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} \sim N(0;1)$$

$$P(-z \leq N(0;1) \leq +z) = \gamma$$

$$p \in \left[\hat{p} \mp z \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}\right]_{\gamma}$$

The sampling error is:

$$\epsilon = z \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$$

And the calculation of the required sample size *n* for a given level of confidence γ and a certain precision is:

$$n = \frac{z^2 \cdot \hat{p} \cdot \hat{q}}{\epsilon^2}$$

An important situation occurs when the situation of máximum uncertainty is considered ($p = q = 0{,}5$). Then the estimated standard deviation of *p*, in the c.i., is:

$$\sqrt{\frac{0{,}25}{n}}$$

❖ C.I. for the difference of means in two non-normal populations (independent large samples)

Let ξ_x and ξ_y be two non-normal variables from which two independent s.r.s. *X* and *Y* are obtained, with sizes being *n* and *m*, respectively. Then, the corresponding sample means $\bar{x}$ and $\bar{y}$, and bias-corrected sample standard deviations s_{1x} and s_{1y} are calculated:

The pivot and the c.i. is:

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

$$P(-z\leq N(0;1)\leq +z)=\gamma$$

$$(\mu_x-\mu_y)\in\left[(\bar{x}-\bar{y})\mp z\cdot\sqrt{\frac{s_{1x}^2}{n}+\frac{s_{1y}^2}{m}}\right]_\gamma$$

The consideration of the difference between two population proportions being a particular case of the former:

$$\frac{(\widehat{p_x}-\widehat{p_y})-(p_x-p_y)}{\sqrt{\frac{\widehat{p_x}\widehat{q_x}}{n}+\frac{\widehat{p_y}\widehat{q_y}}{m}}}\sim N(0;1)$$

$$P(-z\leq N(0;1)\leq +z)=\gamma$$

$$(p_x-p_y)\in\left[(\widehat{p_x}-\widehat{p_y})\mp z\cdot\sqrt{\frac{\widehat{p_x}\widehat{q_x}}{n}+\frac{\widehat{p_y}\widehat{q_y}}{m}}\right]_\gamma$$