Stat Tools : Sample size computation

Context

For a given population, for example the office, one wants to know the proportion of french speeking colleagues. To do so, we might take a sample of the population and ask them the question : Are you french or do you speak french fluently ?
But what is the best sample size to consider ? More precisely, what should be the sample size to get an accurate estimation of the proportion with a confidence of 95% and a margin of error of 5% ?

Mathematics : Cochran's formula

We consider a population of N individuals. Those person are respectively $Y_1, Y_2, \ldots, Y_N$. $\forall i \in [1, N], y_i \in \{0, 1\}$ is the ability of a person to speak in french. Assuming the proportion of french speaking person is p. it is obvious that :

$$\forall i \in [1, N] \qquad y_i \sim \mathcal{B}(p) \quad \mathbf{E}[y_i] = p \qquad Var[y_i] = p(1-p)$$

Let's consider a sample of size $n \leq N$ of the population. The best estimator for p, $\hat{p}_n$ is the empirical mean.

$$\hat{p}_n = \frac{1}{n}\sum_{k=1}^{n} y_k$$

As y_i are independants, by linearity and property of standard deviation :

$$\mathbf{E}[\hat{p}_n] = p \qquad Var[\hat{p}_n] = \frac{p(1-p)}{n}$$

The central limit theorem gives us that :

$$\hat{p}_n \xrightarrow{\mathcal{L}} \mathcal{N}\big(p, Var[\hat{p}_n]\big)$$

We are thus able to build a confidence interval with level of confidence α : (where z_α is the standard z-score for α)

$$\hat{p}_n \in \left[p - z_\alpha\sqrt{\frac{p(1-p)}{n}}, p + z_\alpha\sqrt{\frac{p(1-p)}{n}}\right]$$

Finally, assuming the margin of error Z is equal to half of the size of confidence interval, we get the formula bellow :

$$Z = z_\alpha\sqrt{\frac{p(1-p)}{n}} \qquad \Longrightarrow \qquad n = \frac{z_\alpha^2 p(1-p)}{Z^2}$$

Remarks

- $p(1-p)$ takes value between 0 and 0.25. For $p = 0.5$, the product is maximum and the sample size too. That is the more conservative value ? That is why for an unknown proportion, one may take this value as an estimator.
- This formula applies for a non finite population. In practice if N the population size is unknown or too large to be considered. Again, the formula is more conservative than it's equivalent with a finite set N given bellow.

Other sampling formulas

Cochran for a finite set N

Assuming N is the size of our population.

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \qquad \text{Where} \quad n_0 = \frac{z_\alpha^2 p(1-p)}{Z^2} \qquad \text{(Cochran's formula)}$$

Yamana formula

A much simplier approach that only uses the margin or error Z and the size of the population N.

$$n = \frac{N}{1 + Z^2 N}$$

Python implementation

On ‘code.mazars.global‘ platform in Stat Tools repository, you'll find a simple implementation of all those formulas.

Installation

```
$ pip install stat-tools --index-url https://__token__:xT3QzkWD91p5KDyLUXnn@code.mazars.global/api/v4/projects/2375/packages/pypi/simple
```

CLI (Command line interface)

```
$ stat_tools sample_size --conflvl 0.99 --marginerr 0.01
```

For Windows user, you may probably need to add ‘python -m‘ in front of ‘stat$_{t}ools$‘

Python

```python
from stat_tools import compute_sample_size

margin_error = 0.05
estimated_prop = 0.5
confidence = 0.95
population_size = 600

# Cochran finite
compute_sample_size(
    confidence, margin_error,
    estimated_prop, population_size
)

# Cochran infinite
```

```
compute_sample_size(confidence, margin_error, estimated_prop)

# Yamane
compute_sample_size(
    confidence, margin_error,
    estimated_prop, population_size, how="yamane"
)
```