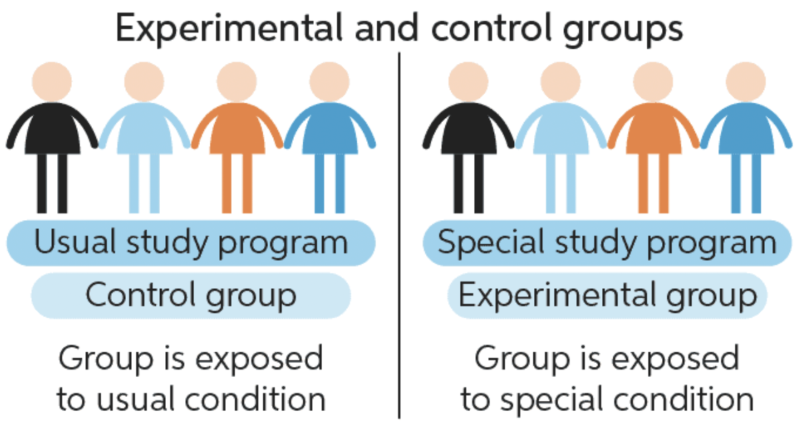
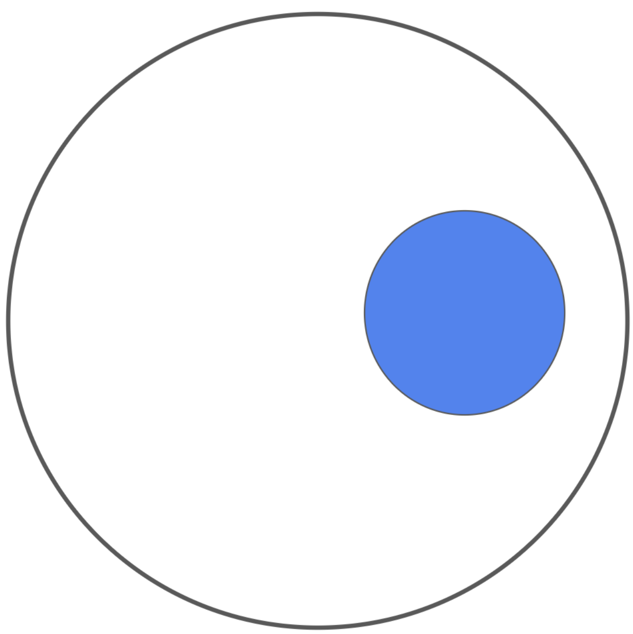
Alberto Andrés Valdés González.
Degree: Mathematical Engineer.
Work position: Data Scientist.
Mail: anvaldes@uc.cl/alberto.valdes.gonzalez.96@gmail.com
Location: Santiago, Chile.

Control Groups

When we can measure the effect of the applicatiion of a logic over people we uses **control groups** and **experimental groups**.



Once we created control and experimental groups we do the question: How big have to be both groups to make inferences with statistical significance?



Thus the real question is: How big have to be an group from a sample to make inferences with statistical significance?

Consider we have n observations on the control/experimental group, then we define:

$$Y = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$$

With $X_i \overset{\text{iid}}{\sim} D(\mu, \sigma)$.

You can see:

$$Z = \frac{Y - \mathbb{E}(Y)}{\sqrt{\mathbb{V}(Y)}} \sim D_Z(0,1)$$

Now we going to compute $\mathbb{E}(Y)$ and $\mathbb{V}(Y)$:

$$\mathbb{E}(Y) = \mathbb{E}\left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i\right) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mu = \frac{1}{n} \cdot n \cdot \mu = \mu$$

$\Rightarrow$

$$\boxed{\mathbb{E}(Y) = \mu}$$

$$\mathbb{V}(Y) = \mathbb{V}\left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i\right) \overset{\text{iid}}{=} \frac{1}{n^2} \cdot \sum_{i=1}^{n} \mathbb{V}(X_i) = \frac{1}{n^2} \cdot \sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

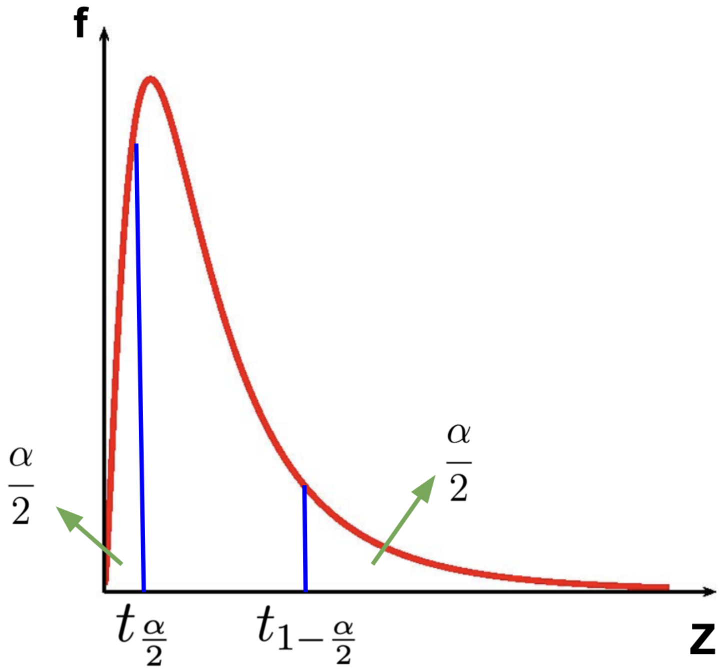
$\Rightarrow$

$$\boxed{\mathbb{V}(Y) = \frac{\sigma}{\sqrt{n}}}$$

$\Rightarrow$

$$Z = \frac{\sqrt{n} \cdot (Y - \mu)}{\sigma} \sim D_Z(0,1)$$

Now its important determine which is the distribution of Z.



Thus:

$$\mathbb{P}\left(t_{\frac{\alpha}{2}} \leq \frac{\sqrt{n} \cdot (Y-\mu)}{\sigma} \leq t_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

The confidence interval is:

$$t_{\frac{\alpha}{2}} \leq \frac{\sqrt{n} \cdot (Y-\mu)}{\sigma} \leq t_{1-\frac{\alpha}{2}}$$

$\Rightarrow$

$$\boxed{\mu + t_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq Y \leq \mu + t_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}}$$

Using the estimations for μ and σ we have:

$$\boxed{\bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \leq Y \leq \bar{x} + t_{1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}}$$

If we consider an ϵ margin of error then:

$$|t_{\frac{\alpha}{2}}| \cdot \frac{S}{\sqrt{n}} \leq \epsilon$$

$$|t_{1-\frac{\alpha}{2}}| \cdot \frac{S}{\sqrt{n}} \leq \epsilon$$

$\Rightarrow$

$$max\{|t_{\frac{\alpha}{2}}|, |t_{1-\frac{\alpha}{2}}|\} \cdot \frac{S}{\sqrt{n}} \leq \epsilon$$

$\Rightarrow$

$$(max\{|t_{\frac{\alpha}{2}}|, |t_{1-\frac{\alpha}{2}}|\})^2 \cdot \frac{S^2}{\epsilon^2} \leq n$$

In this way, the n minimum for a $(1-\alpha)$ level of significance is:

$$\boxed{N_{min} = (max\{|t_{\frac{\alpha}{2}}|, |t_{1-\frac{\alpha}{2}}|\})^2 \cdot \frac{S^2}{\epsilon^2}}$$

Example: If we consider $\alpha = 5\,\%$, $D_Z = Normal$, $X_i \sim Bernoulli(p), \epsilon = 5\,\%$ and $\bar{p} = 10\,\%$.

$t_{\frac{\alpha}{2}} = 1{,}96, t_{1-\frac{\alpha}{2}} = -1{,}96 \;\Rightarrow\; max\{|t_{\frac{\alpha}{2}}|, |t_{1-\frac{\alpha}{2}}|\} = 1{,}96$

$S^2 = \bar{p} \cdot (1-\bar{p}) = 0{,}1 \cdot (1-0{,}1) = 0{,}1 \cdot 0{,}9 = 0{,}09$

$\Rightarrow$

$$N_{min} = \frac{(1{,}96)^2 \cdot 9\,\%}{0,25\,\%} = 138{,}3$$

$\Rightarrow$

$$\boxed{N_{min} = 139}$$
