SAMPLE SIZE CALCULATIONS

Determining how large a sample should be drawn is a big part of practical statistics.

More samples amount to more information, which translates into narrower confidence intervals for unknown population parameters.

In these notes, we derive formulas for the sample sizes needed to achieve a desired narrowness of $(1-\alpha)^*100\%$ C.I.s for the population mean μ, the population proportion p, and for a difference in population means $\mu_1-\mu_2$ or proportions p_1-p_2.

ONE-SAMPLE SETTING

For a mean μ:

Let $X_1,\dots,X_n$ be iid rvs with mean μ, variance σ^2.

Then a large-n $(1-\alpha)^*100\%$ C.I. for μ has the form

$$\bar{X}_n \pm \underbrace{z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{\text{M.E.}}.$$

The part added and subtracted is called the margin of error (M.E.).

The M.E. decreases as the sample size n grows, making the C.I. narrower.

We approach sample size calculation as follows:

> Choose an upper bound M^* for the M.E. and find the smallest n such that M.E. $\le M^*$.

The value of M^* comes from the scientific researcher, and it reflects the degree of accuracy with which it is desired to estimate the unknown parameter.

To achieve M.E. $\leq M^*$, we choose the smallest n such that

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq M^* \quad \Leftrightarrow \quad n \geq \left(\frac{z_{\alpha/2}}{M^*}\right)^2 \sigma^2,$$

i.e., we choose $n = \left\lceil \left(\frac{z_{\alpha/2}}{M^*}\right)^2 \sigma^2 \right\rceil$,

round up to next integer

where $\lceil x \rceil$ is the smallest integer greater than or equal to x.

To make this calculation in practice, we replace σ^2 with an estimate $\hat{\sigma}^2$ from a previous or "pilot" study.

Example: Researchers want to build a 95% C.I. for μ with a M.E. no greater than $1/2$. It is believed that $\sigma \approx 2$. What sample size n should they take?

Take

$$n = \left\lceil \left(\frac{1.96}{1/2}\right)^2 (2)^2 \right\rceil = \lceil 61.4656 \rceil = 62.$$

($z_{0.05/2}$, $\hat{\sigma}^2$, M^*)

For a proportion p:

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$.

Then a large-n $(1-\alpha)^*100\%$ C.I. for p has the form

$$\hat{p}_n \pm \underbrace{z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}}_{\text{M.E.}}$$

To achieve M.E. $\leq M^*$, we choose the smallest n such that

$$z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq M^* \quad \Leftrightarrow \quad n \geq \left(\frac{z_{\alpha/2}}{M^*}\right)^2 p(1-p),$$

i.e., we choose $n = \left\lceil \left(\frac{z_{\alpha/2}}{M^*} \right)^2 \hat{p}(1-\hat{p}) \right\rceil$.

In practice, we replace p with either

(i) an estimate $\hat{p}$ from a previous or pilot study.

(ii) the value $p = 1/2$, for which the M.E. is maximized at any fixed sample size n. Using $p = 1/2$ in our sample size calculation gives the most conservative (largest) value of n.

Example: Suppose we wish to build a 99% C.I. for the proportion of voters who will select a certain candidate. We want a M.E. of no more than 2 percentage points. What sample size do we need?

Take

Use $p = 1/2$ to err on the side of larger n.

$$n = \left\lceil \left(\frac{\overbrace{2.576}^{z_{0.01/2}}}{\underbrace{.02}_{M^*}} \right)^2 \tfrac{1}{2}(1-\tfrac{1}{2}) \right\rceil = \lceil 4{,}147.36 \rceil = 4{,}148.$$

TWO-SAMPLE SETTING

For comparing means:

For large n_1, n_2, the interval

$$\bar{x} - \bar{y} \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

is an approximate $(1-\alpha)^*100\%$ C.I. for $\mu_1 - \mu_2$.

Let $n = n_1 + n_2$ be the total sample size and let

$$\gamma = n_1/n$$

$$1-\gamma = n_2/n,$$

so that $\gamma \in (0,1)$ is the proportion of observations coming from population 1.

Then the M.E. for a given n and $\gamma \in (0,1)$ can be written

$$M(n,\gamma) = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{\gamma n} + \frac{\sigma_2^2}{(1-\gamma)n}}\,.$$

We observe

(i) $M(n,\gamma)$ is strictly decreasing in n.

(ii) For each n, $M(n,\cdot)$ is minimized at $\gamma_{opt} = \frac{\sigma_1}{\sigma_1+\sigma_2}$.

(iii) For each n, $M(n,\cdot)$ has the minimum $z_{\alpha/2}\frac{(\sigma_1+\sigma_2)}{\sqrt{n}}$, i.e.

$$M(n,\gamma_{opt}) = z_{\alpha/2}\frac{(\sigma_1+\sigma_2)}{\sqrt{n}}\,.$$

To achieve a M.E. $\le M^*$, we choose n_1 and n_2 as follows:

- Find smallest real number n^* such that

$$z_{\alpha/2}\frac{(\sigma_1+\sigma_2)}{\sqrt{n^*}} \le M^*\,,$$

i.e. get

$$n^* = \left(\frac{z_{\alpha/2}}{M^*}\right)^2(\sigma_1+\sigma_2)^2\,.$$

← May not be integer-valued

- Then set $n_1 = \left\lceil\left(\frac{\sigma_1}{\sigma_1+\sigma_2}\right)n^*\right\rceil$, $n_2 = \left\lceil\left(\frac{\sigma_2}{\sigma_1+\sigma_2}\right)n^*\right\rceil$, $n = n_1+n_2$.

In practice, we replace σ_1 and σ_2 with some estimates $\hat{\sigma}_1$ and $\hat{\sigma}_2$ from a previous or pilot study.

Note that we draw a larger sample from the population with the larger variance.

To find γ_{opt}: Find value of γ which minimizes $M(n,\gamma)$.

Since $M(n,\gamma) > 0$, the same value of γ minimizes $M^2(n,\gamma)$.

So we set the derivative of $M^2(n,\gamma)$ w.r.t. γ equal to zero and solve for γ:

$$\frac{\partial}{\partial\gamma} M^2(n,\gamma) = \frac{\partial}{\partial\gamma} z_{\alpha/2}^2 \left[\frac{\sigma_1^2}{\gamma n} + \frac{\sigma_2^2}{(1-\gamma)n}\right] = \frac{z_{\alpha/2}^2}{n}\left[-\frac{\sigma_1^2}{\gamma^2} + \frac{\sigma_2^2}{(1-\gamma)^2}\right] = 0$$

$$\Leftrightarrow \quad \frac{\sigma_2^2}{(1-\gamma)^2} = \frac{\sigma_1^2}{\gamma^2}$$

$$\Leftrightarrow \quad \frac{\gamma}{1-\gamma} = \frac{\sigma_1}{\sigma_2}$$

$$\Leftrightarrow \quad \gamma = \frac{\sigma_1}{\sigma_1+\sigma_2}, \quad \text{so} \quad \gamma_{opt} = \frac{\sigma_1}{\sigma_1+\sigma_2}.$$

To get $M(n,\gamma_{opt})$: Under γ_{opt}, we get

$$M(n,\gamma_{opt}) = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{\gamma_{opt}\cdot n} + \frac{\sigma_2^2}{(1-\gamma_{opt})\cdot n}}$$

$$= z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{\left(\frac{\sigma_1}{\sigma_1+\sigma_2}\right)n} + \frac{\sigma_2^2}{\underbrace{\left(1-\frac{\sigma_1}{\sigma_1+\sigma_2}\right)}_{\frac{\sigma_2}{\sigma_1+\sigma_2}} n}}$$

$$= z_{\alpha/2}\sqrt{\frac{\sigma_1(\sigma_1+\sigma_2) + \sigma_2(\sigma_1+\sigma_2)}{n}}$$

$$= \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\underbrace{\sigma_1^2 + 2\sigma_1\sigma_2 + \sigma_2^2}_{(\sigma_1+\sigma_2)^2}}$$

$$= z_{\alpha/2}\frac{(\sigma_1+\sigma_2)}{\sqrt{n}}.$$

Example: Suppose we have $\hat{\sigma}_1 = 2$ and $\hat{\sigma}_2 = 3$ from a previous study. We wish to build a 99% C.I. for $\mu_1 - \mu_2$ which has a M.E. ≤ 0.5. Find n_1 and n_2.

Get

$$n^* = \left(\frac{2.576}{0.5}\right)^2 (2+3)^2 = 663.4897,$$

and then set

$$n_1 = \left\lceil \left(\frac{2}{2+3}\right) 663.4897 \right\rceil = 266$$

$$n_2 = \left\lceil \left(\frac{3}{2+3}\right) 663.4897 \right\rceil = 399, \quad n = 266 + 399 = 665.$$

For comparing proportions:

For large n_1, n_2, the interval

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

is an approximate $(1-\alpha)^*100\%$ C.I. for $p_1 - p_2$.

By the same arguments as those in the previous section, we may justify choosing n_1 and n_2 as follows:

- Find smallest real number n^* such that

$$z_{\alpha/2} \left(\frac{\sqrt{\hat{p}_1(1-\hat{p}_1)} + \sqrt{\hat{p}_2(1-\hat{p}_2)}}{n^*} \right) \leq M^*$$

i.e. get

$$n^* = \left(\frac{z_{\alpha/2}}{M^*}\right)^2 \left(\sqrt{\hat{p}_1(1-\hat{p}_1)} + \sqrt{\hat{p}_2(1-\hat{p}_2)} \right)^2 .$$

- Then set

$$n_1 = \left\lceil \left(\frac{\sqrt{\hat{p}_1(1-\hat{p}_1)}}{\sqrt{\hat{p}_1(1-\hat{p}_1)} + \sqrt{\hat{p}_2(1-\hat{p}_2)}} \right) n^* \right\rceil ,$$

$$n_2 = \left\lceil \left(\frac{\sqrt{\hat{p}_2(1-\hat{p}_2)}}{\sqrt{\hat{p}_1(1-\hat{p}_1)} + \sqrt{\hat{p}_2(1-\hat{p}_2)}} \right) n^* \right\rceil \quad \text{and} \quad n = n_1 + n_2 .$$

In practice, we replace $\hat{p}_1$ and $\hat{p}_2$ with some estimates $\hat{p}_1$ and $\hat{p}_2$ from a previous or pilot study.