MSMS 304 - Biostatistics

Sample Size Determination

Ananda Biswas

Last updated : November 5, 2025

Contents

1 Introduction

Clinical trials should have sufficient statistical power to detect differences between groups considered to be of clinical importance. Therefore calculation of sample size with provision for adequate levels of significance and power is an essential part of planning.

If the sample size is drastically large, then we may put human lives at risk unnecessarily. On the other hand, if the sample size is too small, there is a good chance that the trial will fall short in demonstrating any difference between study groups.

To detect a large difference between two study groups as statistically significant, small sample size is enough. But to detect a small difference between the study groups as significant, we require big enough sample size.

Note : Prevalence of a condition in a region $= \dfrac{\text{total number of cases of the health condition}}{\text{total population in the region}}$.

2 Sample Size for Single Proportion

2.1 Given Absolute Precision

☞ A Case Scenario : A researcher is interested in estimating the prevalence of Type II diabetes amongst the population in Jaipur. It was assumed that the anticipated prevalence of Type II diabetes would not be more than 15% in the population. What is the minimum sample size required to estimate this prevalence at 95% confidence level and 5% absolute precision ? (*Absolute precision is the maximum allowable difference between the estimated prevalence and the true prevalence.*)

Anticipated prevalence $p = 0.15$, confidence coefficient $1 - \alpha = 0.95$, absolute precision $d = 0.05$, $\tau_{\alpha/2} = 1.96$.

$$\begin{aligned} n &\geq \frac{\tau_{\alpha/2}^2 \cdot p \cdot (1-p)}{d^2} \\ &= \frac{(1.96)^2 \cdot 0.15 \cdot 0.85}{(0.05)^2} \\ &= 196. \end{aligned} \tag{1}$$

✎ Write up : It was anticipated that the prevalence of Type II diabetes amongst the population in Jaipur would not be more than 15%. Based on sample size determination for single proportion, considering 95% confidence level and 5% absolute precision, the minimum sample size required to estimate the prevalence was determined to be 196.

From (1), sample size n increases as p increases up to $p = 0.5$ *i.e.* 50%.

Absolute Precision (d)	Anticipated Prevalence (p)		
	0.10	0.15	0.20
0.05	139	196	246
0.07	71	100	126
0.10	35	49	62

As you might notice, for a fixed absolute precision, as anticipated prevalence increases required sample size also increases. This directly follows from the formula (1). The sample size n and p are directly proportional.

2.2 Given Relative Precision

☞ A Case Scenario : A researcher is interested in estimating the prevalence of Type II diabetes amongst the population in Jaipur. It was assumed that the anticipated prevalence of Type II diabetes would not be more than 15% in the population. What is the minimum sample size required to estimate this prevalence at 95% confidence level and 20% relative precision ? (*Relative precision is the maximum allowable error in the estimated prevalence expressed as a proportion of the true value of prevalence. So here allowable error is* 20% *of* 0.15 *i.e.* 0.03.)

Anticipated prevalence $p = 0.15$, confidence coefficient $1 - \alpha = 0.95$, relative precision $d = 0.2$, $\tau_{\alpha/2} = 1.96$.

$$\begin{aligned} n &\geq \frac{\tau^2_{\alpha/2} \cdot p \cdot (1-p)}{(pd)^2} \\ &= \frac{(1.96)^2 \cdot 0.15 \cdot 0.85}{(0.15 \times 0.2)^2} \\ &= 545. \end{aligned} \tag{2}$$

✎ Write up : It was anticipated that the prevalence of Type II diabetes amongst the population in Jaipur would not be more than 15%. Based on sample size determination for single proportion, considering 95% confidence level and 20% absolute precision, the minimum sample size required to estimate the prevalence was determined to be 545.

Relative Precision (d)	Anticipated Prevalence (p)		
	0.10	0.15	0.20
0.15	1537	968	683
0.20	865	545	385
0.25	534	349	246

Here, for a fixed relative precision, as anticipated prevalence increases required sample size decreases. This directly follows from the formula (2) (Notice an additional p^2 in the denominator). So as p increases sample size n decreases.

✎ A few *advantages* of relative precision over absolute precision are noteworthy. For rare diseases, prevalence is too small, not more than 3% to 5%. Then the choice of an absolute precision 10% will result the tolerance boundary of estimated prevalence to go beyond 0 in the negative region. To prevent such scenarios, relative precision is preferred. Another edge case might be for very frequent conditions the prevalence is high, for example 95% or 97%. Then a choice of absolute precision 10% will push the tolerance boundary of estimated prevalence beyond 100%, which is fallacious. This adds to demerits of absolute precision.

3 Sample Size for Single Mean

3.1 Given Absolute Precision

☞ A Case Scenario : A researcher is interested in estimating the mean HDL cholesterol amongst the population of Jaipur. It was assumed that the mean HDL cholesterol would be around 40 mg/dL with a standard deviation of 10 mg/dL. What is the minimum sample size required to estimate the mean HDL cholesterol level at 95% confidence level and 5 mg/dL absolute precision?

Anticipated mean HDL cholesterol $\mu = 40$, standard deviation $\sigma = 10$, confidence coefficient $1 - \alpha = 0.95$, absolute precision $d = 5$, $\tau_{\alpha/2} = 1.96$.

$$\begin{aligned} n &\geq \frac{\tau_{\alpha/2}^2 \cdot \sigma^2}{d^2} \\ &= \frac{(1.96)^2 \cdot (10)^2}{(5)^2} \\ &= 16 \end{aligned} \tag{3}$$

✎ Write up : It was anticipated that the mean HDL cholesterol amongst the population in Jaipur would be around 40 mg/dL with a standard deviation of 10 mg/dL. Based on sample size determination for single mean, considering 95% confidence level and 5 mg/dL absolute precision, the minimum sample size required to estimate the mean HDL cholesterol was determined to be 16.

Absolute Precision (d)	Standard Deviation (σ)		
	5	10	15
3	11	43	97
5	4	16	95
7	2	8	18

As standard deviation increases, for a fixed absolute precision, required sample size increases. This aligns with the formula (3).

3.2 Given Relative Precision

☞ A Case Scenario : A researcher is interested in estimating the mean HDL cholesterol amongst the population of Jaipur. It was assumed that the mean HDL cholesterol would be around 40 mg/dL with a standard deviation of 10 mg/dL. What is the minimum sample size required to estimate the mean HDL cholesterol level at 95% confidence level and 10% relative precision?

Anticipated mean HDL cholesterol $\mu = 40$, standard deviation $\sigma = 10$, confidence level $1 - \alpha = 0.95$, relative precision $d = 0.1$, $\tau_{\alpha/2} = 1.96$.

$$n \geq \frac{\tau_{\alpha/2}^2 \cdot \sigma^2}{(d \cdot \mu)^2} \tag{4}$$

$$= \frac{(1.96)^2 \cdot (10)^2}{(0.1 \cdot 40)^2}$$
$$= 25$$

✎ Write up : It was anticipated that the mean HDL cholesterol amongst the population in Jaipur would be around 40 mg/dL with a standard deviation of 10 mg/dL. Based on sample size determination for single mean, considering 95% confidence level and 10% relative precision, the minimum sample size required to estimate the mean HDL cholesterol was determined to be 25.

Relative Precision (d)	Standard Deviation (σ)		
	5	10	15
0.05	25	97	217
0.10	7	25	55
0.15	3	11	25

As standard deviation increases, for a fixed relative precision and fixed μ, required sample size increases. This aligns with the formula (4).

4 Sample Size Determination for Comparison : Theory

Let X be a random variable representing a characteristic of interest in a population with mean μ and variance σ^2 *i.e.* $X \sim N(\mu, \sigma^2)$. Let us take a sample of size n from the population to estimate the population parameter μ based on defined statistic (say $\bar{X}$) whose sampling distribution follows approximately normal distribution for large n under central limit theorem with parameters μ and variance $\frac{\sigma^2}{n}$ *i.e.* $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

The standard error of $\bar{X}$ will be $SE = \frac{\sigma}{\sqrt{n}}$.

Standardizing the sample mean using the population parameters gives:

$$Z = \frac{\sqrt{n}\left(\bar{X} - \mu\right)}{\sigma} = \frac{\bar{X} - \mu}{SE} \sim N(0, 1).$$

A researcher hypothesizes that the sample is drawn from one of two possible populations.

Null Hypothesis (H_0) : The sample is drawn from a population whose mean is μ_0 and variance σ_0^2 *i.e.*

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma_0^2}{n}\right) \equiv N\left(\mu_0, SE_0^2\right).$$

Alternative Hypothesis (H_1) : The sample is drawn from a population whose mean is μ_1 and variance σ_1^2 *i.e.*

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n}\right) \equiv N\left(\mu_1, SE_1^2\right).$$

Let $\Delta = |\mu_0 - \mu_1|$. We *assume* that σ_0^2 and σ_1^2 are known. For now we also assume $\mu_1 > \mu_0$ *i.e.* $\mu_1 = \mu_0 + \Delta$. In other words, we have an one-sided alternative hypothesis.

Under H_0, $Z = \dfrac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{\sigma_0} = \dfrac{\bar{X} - \mu_0}{SE_0} \sim N(0,1)$.

We reject H_0 at level of significance α if observed $Z > \tau_\alpha$ where τ_α is upper α point for a standard normal distribution *i.e.* $P(Z > \tau_\alpha) = \alpha$ where $Z \sim N(0,1)$.

So H_0 is rejected when

$$\begin{aligned} &\frac{\bar{x} - \mu_0}{SE_0} > \tau_\alpha \\ &\Rightarrow \bar{x} > \mu_0 + \tau_\alpha \cdot SE_0. \end{aligned}$$

Let $1 - \beta$ be the power of the test.

$$\begin{aligned} P[\text{Rejecting } H_0 \text{ when } H_1 \text{ is true}] &= P[\bar{x} > \mu_0 + \tau_\alpha \cdot SE_0 \text{ when } H_1 \text{ is true}] \\ &= P\left[\frac{\bar{x} - \mu_1}{SE_1} > \frac{\tau_\alpha \cdot SE_0 + \mu_0 - \mu_1}{SE_1} \text{ when } H_1 \text{ is true}\right] \\ &= P\left[Z > \frac{\tau_\alpha \cdot SE_0 + \mu_0 - \mu_1}{SE_1}\right] \text{ as } \bar{X} \overset{H_1}{\sim} N\left(\mu_1, SE_1^2\right) \\ &= P\left[Z > \frac{\tau_\alpha \cdot SE_0 - \Delta}{SE_1}\right] \\ \Rightarrow 1 - \beta &= P\left[Z > \frac{\tau_\alpha \cdot SE_0 - \Delta}{SE_1}\right] \\ \Rightarrow \frac{\tau_\alpha \cdot SE_0 - \Delta}{SE_1} &= \tau_{1-\beta} \end{aligned}$$

$$\therefore\ \Delta = \tau_\alpha \cdot SE_0 - \tau_{1-\beta} \cdot SE_1 \tag{5}$$

4.1 Comparison of Means between Two Independent Populations

Let the means of the two populations of the characteristics under study in the populations are μ_1 and μ_2 & the variances are σ_1^2 and σ_2^2 respectively.

To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ or $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : |\mu_1 - \mu_2| = \Delta$.

We know that for a two-tailed test, the difference of anticipated means Δ between the two populations at level of significance α and power $1 - \beta$ can be presented as:

$$\Delta = \tau_{\alpha/2} \cdot SE_0 - \tau_{1-\beta} \cdot SE_1. \tag{6}$$

Let the samples of sizes n_1 and n_2 are taken from the two independent populations. The standard error of the difference of sample means between the two independent populations then can be obtained as:

$$SE(\bar{x_1} - \bar{x_2}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

If the variances of the two populations are not known, these variances are replaced by their unbiased estimates. Let the respective sample variances s_1^2 and s_2^2 be the unbiased estimates of σ_1^2 and σ_2^2.

Under H_0, $SE_0(\bar{x_1} - \bar{x_2}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ and under H_1, $SE_1(\bar{x_1} - \bar{x_2}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

If we further assume that the two variances are unknown but equal *i.e.* $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$SE_0(\bar{x_1} - \bar{x_2}) = SE_1(\bar{x_1} - \bar{x_2}) = SE(\bar{x_1} - \bar{x_2}) = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot \sigma^2}.$$

σ^2 is unknown and it is estimated by the *pooled variance* s^2 which is given by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}.$$

Using the above estimate of σ^2 in SE_0 and SE_1, from (6), we continue as follows :

$$\Delta = \tau_{\alpha/2} \cdot \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot s^2} - \tau_{1-\beta} \cdot \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot s^2}$$

$$\Rightarrow \Delta = \left(\tau_{\alpha/2} - \tau_{1-\beta}\right) \cdot \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot s^2}$$

$$\Rightarrow \left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{\Delta^2}{\left(\tau_{\alpha/2} - \tau_{1-\beta}\right)^2 \cdot s^2}$$

If sample size of second population is k times of the first population, *i.e.*, $n_2 = kn_1$, then we obtain

$$n_1 = \frac{\left(\tau_{\alpha/2} - \tau_{1-\beta}\right)^2 \cdot \left(1 + \frac{1}{k}\right) \cdot s^2}{\Delta^2} \tag{7}$$

Once the sample size for the first population is obtained, the sample size for the second population can be obtained simply by multiplying a predetermined multiplier k, *i.e.*, kn_1.

For $k = 1$, we have $n_1 = \frac{\left(\tau_{\alpha/2} - \tau_{1-\beta}\right)^2 \cdot 2s^2}{\Delta^2} = n_2$.

The above formula can be rewritten as $n_1 = \frac{\left(\tau_{\alpha/2} - \tau_{1-\beta}\right)^2 \cdot 2}{\left(\frac{s}{\Delta}\right)^2}$ where $\frac{s}{\Delta}$ is called the ***Cohen effect size***. Cohen defined

Value	Effect Size
0.2	Small effect size
0.5	Medium effect size
0.8	Large effect size

☞ A Case Scenario : A cross sectional comparative study will be carried out to compare 25−hydroxy vitamin D levels (ng/ml) between tuberculosis patients and normal subjects. The minimum significant difference between the two groups was expected to be 2 ng/ml. The common standard deviation of 25−hydroxy vitamin D levels was 5 ng/ml. What is the minimum sample size required per group to detect this difference with 5% level of significance and 80% power ?

$\Delta = 2$, known $\sigma^2 = 5$, $\alpha = 0.05$, $1 - \beta = 0.8$, equal sized groups. $\tau_{0.025} = 1.96$, $\tau_{0.8} = -0.84$.

$$n_1 \geq \frac{(1.96 - (-0.84))^2 \cdot 2 \cdot 5^2}{2^2} = 98;\ n_2 \geq 98.$$

✎ Write up : The sample size was estimated by using the formula for comparison of two independent mean. Anticipating a minimum significant difference of 2 ng/ml for 25−hydroxy vitamin D levels (ng/ml) between tuberculosis patient and normal subject and a standard deviation of 5 ng/ml, the minimum sample size required is 98 subjects in each group at 5% level of significance and 80% power.

4.2 Comparison of Proportions between Two Independent Populations

Let the proportions of the two populations of the characteristics under study in the populations are P_1 and P_2 & the variances are P_1Q_1 and P_2Q_2 respectively.

To test $H_0 : P_1 = P_2$ against $H_1 : P_1 \neq P_2$ or $H_0 : P_1 - P_2 = 0$ against $H_1 : |P_1 - P_2| = \Delta$.

We know that for a two-tailed test, the difference of anticipated proportions Δ between the two populations at level of significance α and power $1 - \beta$ can be presented as equation (6):

$$\Delta = \tau_{\alpha/2} \cdot SE_0 - \tau_{1-\beta} \cdot SE_1. \tag{8}$$

Let the samples of sizes n_1 and n_2 are taken from the two independent populations. The standard error of the difference of sample proportions between the two independent populations then can be obtained as:

$$SE(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}.$$

If the variances of the two populations are not known, these variances are replaced by their unbiased estimates. Let p_1q_1 and p_2q_2 be the unbiased estimates of P_1Q_1 and P_2Q_2.

Under $H_0 : P_1 = P_2 = P$, $SE_0(p_1 - p_2) = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}$ and

under H_1, $SE_1(p_1 - p_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$.

Replacing the standard errors under H_0 and H_1 in (6), we get

$$\Delta = \tau_{\alpha/2} \cdot \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}} - \tau_{1-\beta} \cdot \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

If sample size of second population is k times of the first population, *i.e.*, $n_2 = kn_1$, then we get

$$\Delta = \tau_{\alpha/2} \cdot \sqrt{\frac{pq}{n_1} + \frac{pq}{kn_1}} - \tau_{1-\beta} \cdot \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{kn_1}}$$

$$\Rightarrow \Delta = \tau_{\alpha/2} \cdot \sqrt{\frac{pq}{n_1}\left(1+\frac{1}{k}\right)} - \tau_{1-\beta} \cdot \sqrt{\frac{1}{n_1}\left(p_1q_1 + \frac{p_2q_2}{k}\right)}$$

$$\Rightarrow \Delta^2 = \frac{1}{n_1}\left[\tau_{\alpha/2} \cdot \sqrt{pq\left(1+\frac{1}{k}\right)} - \tau_{1-\beta} \cdot \sqrt{\left(p_1q_1 + \frac{p_2q_2}{k}\right)}\right]^2$$

$$\Rightarrow n_1 = \frac{\left[\tau_{\alpha/2} \cdot \sqrt{pq\left(1+\frac{1}{k}\right)} - \tau_{1-\beta} \cdot \sqrt{\left(p_1q_1 + \frac{p_2q_2}{k}\right)}\right]^2}{\Delta^2}$$

Once the sample size for the first population is obtained, the sample size for the second population can be obtained simply by multiplying a predetermined multiplier k, *i.e.*, kn_1.

For $k = 1$, we have $n_1 = \dfrac{\left[\tau_{\alpha/2} \cdot \sqrt{2pq} - \tau_{1-\beta} \cdot \sqrt{(p_1q_1 + p_2q_2)}\right]^2}{\Delta^2} = n_2$.

☞ A Case Scenario : A randomized controlled trial will be carried out to compare the effect of two anaesthetic techniques *Spinal Anaesthesia through median approach* & *Spinal Anaesthesia through para-median approach* on post-dural back pain during the early post-operative period. Anticipated incidence of back pain in Median approach is 0.36 and anticipated incidence of back pain in para-median approach is 0.16. We anticipate that the patients receiving Spinal Anaesthesia through 'para-median' approach will have reduction in the incidence of back-pain from those receiving through the 'median approach'. What should be the minimum sample size required per group to detect this difference with 5% level of significance and 80% power ?

$\Delta = 0.36 - 0.16 = 0.2$, known $P_1 = 0.36$, known $P_2 = 0.16$, $\alpha = 0.05$, $1 - \beta = 0.8$, equal sized groups. $\tau_{0.025} = 1.96$, $\tau_{0.8} = -0.84$.

Take $P = \dfrac{P_1 + P_2}{2} = \dfrac{0.36 + 0.16}{2} = 0.26$.

$$n_1 \geq \frac{\left[1.96 \cdot \sqrt{2 \cdot 0.26 \cdot 0.74} - (-0.84) \cdot \sqrt{(0.36 \cdot 0.64 + 0.16 \cdot 0.84)}\right]^2}{0.2^2} = 75;\ n_2 \geq 75.$$

✎ Write up : The sample size is estimated using the sample size formula for comparing two independent proportions. Anticipated incidence of back pain in the median technique as 36% and considering a 20% reduction in the incidence of back pain in the para-median approach as clinically important, at a 5% level of significance and 80% power, the study would require 75 participants in each anaesthetic technique.