Dinesh Kumar Walia
Department of Community Medicine
Government Medical College
Chandigarh

- Type of Study Designs
  - Case Control and Cohort Studies
- Level of Significance/Confidence Coefficient
- Concept of Errors in Testing of Hypotheses
  - Type-I error (α)
  - Type-II error (β)
- Power of the Test

SAMPLE SIZE IN ANALYTICAL STUDIES

There are several variations in methodologies for estimating sample size in analytic study or experiments, as in descriptive studies but certain steps are common:

.

Steps:

- State the null hypothesis H_0 and either a one- or two-tailed alternative hypothesis H_1.
- Select the appropriate statistical test from table based on the type of predictor variable and outcome variable in that hypothesis

- Chose a reasonable effect size (and variability, if necessary).
- Set α and β. (if the alternative hypothesis is one-tailed, use a one-tailed α; other-wise, use two-tailed α.)
- Use the formula to estimate the optimum sample size.

FOR ESTIMATING SINGLE PROPORTION

95% CI for p is given by

$p \pm 1.96 [p (1-p)/ \sqrt{n}]$

so that

$n = (1.96)^2 \; p (1-p)/d$

$\approx 4pq/ d^2$

Problem. A local health department wishes to estimate the prevalence of vitamin A deficiency among children under five years of age. How many children should be selected so that prevalence may be estimated within 5% point of true value with 95% confidence, if it is known that the true rate is unlikely to exceed 20%.

P=0.20, conf coeff= 95% d=5%=0.05

n_{opt} = 246.

objective: To find prevalence of malnutrition in under five children

Anticipated population prevalence = 28%

[Past Studies/ pilot surveys]

Permissible Error =5% in absolute terms

p=0.28, q= 0.72

$n_{opt} = 4*0.28*0.72 = 322.56 \sim 323$

ESTIMATION OF OPTIMUM SAMPLE SIZE IN ANALYTICAL STUDIES FOR TESTING EQUALITY OF TWO PROPORTIONS:

$$n_{opt} = z^2_{1-\alpha} V/d^2 ,$$

where,

$$V=P_1 (1- P_1) + P_2 (1-P_2)$$

Hypothesis testing for two population proportions

For one sided test

$n_{opt} = \{z_{1-\alpha}\sqrt{2(P(1-P)} + z_{1-\beta}\sqrt{V}\}^2/d^2$

Where

$P = (P_1+P_2)/2,$

$V=P_1(1-P_1)+P_2(1-P_2)$,

and $d=P_1-P_2$.

For two tailed test

$n_{opt}=\{z_{1-\alpha/2}\sqrt{(2P(1-P)}+z_{1-\beta}\sqrt{V}\}^2/d^2$

Hypothesis testing for two population proportions

For one sided test

$n_{opt} = \{z_{1-\alpha}\sqrt{2(P(1-P)} + z_{1-\beta}\sqrt{V}\}^2/d^2$

Where

$P = (P_1+P_2)/2,$

$V = P_1(1-P_1) + P_2(1-P_2),$

and

$d = P_1 - P_2.$

For two tailed test

$$n_{opt}=\{z_{1-\alpha/2}\sqrt{(2P(1-P)}+z_{1-\beta}\sqrt{V}\}^2/d^2$$

ESTIMATION OF OPTIMUM SAMPLE SIZE IN CASE- CONTROL STUDIES:

(a) Calculating Sample Size When Using The Z- Test In A Case- Control Study

EXAMPLE

If we presume that long- term use of Oral Contraceptive Pill (OCP) increases the risk of Coronary Heart Disease (CHD), we want to know the sample size sufficient to detect an increase in OR of ≥ 30% by means of the case control methodology. The hypothesis can be stated as the proportion of women using OCPs is the same among those with CHD and those without CHD. Assume 20% women without CHD use OCPs (control group).

Relative Odds

= Odds among Exposed /Odds among Unexposed

Odds among Exposed= $P_1 / 1 - P_1$

Odds among Unexposed= $P_2 / 1 - P_2$

This formula gives us the interrelationship between Odds Ratio, P_1 and P_2

- We have decided the need to detect an OR ≥ 1.3, the use of OCPs among women with CHD must be 20+0.3x20=26%

- Choosing α and β to be 5% each, the sample size, using the above formula comes to 2220. This means that we need to study 2220 cases and 2220 controls.

Problem:

The investigator plans a case-control study of whether a history of herpes simplex is associated with lip cancer. A brief pilot study finds that about 30% of persons without lip cancer have had herpes simplex. The investigator is interested in detecting whether the odds ratio for lip cancer associated with herpes simplex infection in 2.5 or more. How many subjects will be required?

The ingredients for the sample size calculation are as follows:

- Null hypothesis: The proportion of cases of lip cancer with a history of herpes simplex is the same as the proportion of controls with a herpes simplex history.
- Alternative hypothesis: The proportion of cases of lip cancer with a history of herpes simplex is greater than the proportion of controls with a herpes simplex history.
- P_2 (proportion of controls expected to have the risk factor) = 0.30;

- P_1 (proportion of cases expected to have the risk factor) = OR * P_2/ (1-P_2 + OR * P_2) = (2.5 * 0.3)/ (1-0.3 + 2.5*0.3) = 0.75/1.45 = 0.52.
- With α (one-tailed) α = 0.025 and power = 90%, β = 1-0.90 = 0.10.

(b) ESTIMATING AN ODDS RATIO WITH SPECIFIED RELATIVE PRECISION

$$n_{opt}=z^2_{1-\alpha/2}\{1/[P_1^*(1-P_1^*]$$

$$+1/[P_2^*(1-P_2^*)]\}/[\log_e(1-\epsilon)]^2$$

(c) HYPOTHESIS TESTS FOR AN ODDS RATIO

$$n_{opt}=\{z_{1-\alpha/2}\sqrt{(2P_2^*(1-P_2^*)}+z_{1-\beta}\sqrt{V^*}\}^2/d^2$$

$$V^*=P_1^*(1-P_1^*)+P_2^*(1-P_2^*),$$

(a) Two of the following should be known:

P_1*= Anticipated probability of “exposure” for people with the disease [a/ (a + b)]

P_2*= Anticipated probability of “exposure “for people without the disease [c/(c+ d)]

Anticipated odds ratio OR

(b) Confidence level 100(1-α) %

(c) Relative Precision ϵ

Example

In a defined area where cholera is posing a serious public health problem, about 30% of the populations are believed to be using water from contaminated sources. A case-control study of the association between cholera and exposure to contaminated water

is to be undertaken in the area to estimate the odds ratio to within 25% of the true value, which is believed to be approximately 2, with 95% confidence. What sample sizes would be needed in the cholera and control groups?

Solution

(a)

Anticipated probability of “exposure “given “disease” ?

Anticipated probability of “exposure “given “no disease”,

(Approximated by overall exposure rate) = 30%

Anticipated OR=2

b) Confidence level 95%

(c) Relative Precision 25%

Example

The efficacy of BCG vaccine in preventing childhood tuberculosis is in doubt and a study is designed to compare the vaccination coverage rates in a group of people with tuberculosis and a group of controls. Available information indicates that roughly 30% of the controls are not vaccinated. The investigator wishes to have an 80% chance of detecting an odds ratio significantly different from 1 at the 5% level. If an odds ratio of 2 would be considered an important difference between the two groups, how large a sample should be included in each study group?

Solution

(a)

Test value of the odds ratio= 1

Anticipated probability of “exposure “for “disease” ?

Anticipated probability of “exposure “for “no disease” = 30%

Anticipated odds ratio = 2

(b) level of significance =5%

(c) Power of the test = 80%

(d) Alternative hypothesis :

odds ratio≠1

(III) CALCULATING SAMPLE SIZE WHEN USING THE t- TEST IN A CASE- CONTROL STUDY:

Example

The research's question is whether serum cholesterol level is associated with stroke. The mean value for cholesterol in controls with-out stroke is

about 200 mg / dl, with a standard deviation of about 20 mg / dl. A few previous studies have detected a difference of about + 10 mg / dl between stroke patients and controls, and other studies have found no difference or even a tendency for serum cholesterol to be lower in stroke patients. How many cases and controls will be needed to detect a difference of 10mg/dl between the two groups? Why was a two-tailed α used?

Solution:

The ingredients for the sample size calculations are as follows:

Null hypothesis: There is no difference in mean serum cholesterol level in stroke cases and controls.

Alternative hypothesis: There is difference in mean serum cholesterol in stroke cases and control.

α (two- tailed) = 0.05

and

β=0.10,

CALCULATING SAMPLE SIZE IN COHORT STUDIES

- USING z- TEST IN A COHORT STUDY

Problem:

The research question is whether elderly smokers have a greater incidence of skin cancer than nonsmokers. A review of pervious literature suggests that the 5 year incidence of skin cancer is about 0.20 in elderly nonsmokers.

How many smokers and nonsmokers will need to be studied to determine whether the 5 year skin cancer incidence is at least 0.30 in smokers? Why was a one-tailed alternative hypothesis chosen?

Solution: The ingredients for the sample size calculation are as follows:

Null hypothesis: The incidence of skin cancer is the same is elderly smokers and nonsmokers.

Alternative hypothesis: The incidence of skin cancer is higher in elderly smokers than nonsmokers.

P_2 (incidence among nonsmokers)=0.20;

P1 (incidence among smokers) = 0.30. The smaller of these values is 0.20, and the difference between them ($P_1 - P_2$) is 0.10.

At α (one-tailed) = 0.05 and power = 80%, $\beta = 1 - 0.80 = 0.20$. Calculate n_{opt}?

ESTIMATING A RELATIVE RISK WITH SPECIFIED RELATIVE PRECISION

$$n_{opt}=z^2_{1-\alpha/2}\{1-P_1)/P_1+(1-P_2)/P_2[\log_e(1-\epsilon)]^2$$

Two of the following should be known:

Anticipated probability of disease in people exposed to the factor of interest P_1

Anticipated probability of disease in people exposed to the factor of interest P_2

Anticipated relative risk

RR

Confidence level $100(1-\alpha)\%$

Relative precision ϵ

Remember,

$RR = P_1/P_2$,

$P_2 = P_1/RR$

Results on minimum sample size for confidence levels of 95% and 90%, and levels of precision of 10%, 20%, 25% and 50% can be calculated.

For determining sample size from RR> 1, the values of both P_2 and RR are needed. Either of these may be calculated, if necessary, provided that P_1 in Known:

$$RR = P_1/P_2$$

And

$$P_2 = P_1/RR$$

If RR<1, the values of P_1 and 1/RR should be used instead.

Example

An epidemiologist is planning a study to investigate the possibility that a certain lung disease is linked with exposure to a recently identified air pollutant. What sample size would be needed in needed in each of two groups, relative risk to within 50% of the true value (which is believed to be approximately 2) with 95% confidence? The disease is present in 20% of people who are not exposed to the air pollutant.

Solution

(a) Anticipated probability of disease given "exposure" ?

Anticipated probability of disease given "no exposure" = 20%

Anticipated relative risk =2

(b) Confidence level = 95%

(c) Relative precision = 50%

- Example
- Two competing therapies for a particular cancer are to be evaluated by a cohort study in a multi-centric clinical trial. Patients will be randomized to either treatment A or treatment B and will be followed for 5 years after treatment for recurrence of the disease. Treatment A is a new therapy that will be widely used if it can be demonstrated that it halves the risk of recurrence in the first 5 years after treatment (i.e. RR=0.5): 35% recurrence is currently observed in patient who have received treatment B.

how many Patients should be studied in each of the two treatment groups if the investigators wish to be 90% confident of correctly rejecting the null hypothesis (RR_0=1), if it is false, and the test is to be performed at the 5% level of significance?

Solution

test value of the relative risk = 1

Anticipated probability of recurrence given treatment A ?

Anticipated probability of recurrence given treatment B 35%

Anticipated relative risk = 0.5

Level of significance = 5%

Power of the test = 90%

Alternative hypothesis

RR ≠1

SAMPLE SIZE CALCULATION FOR COMPARING TWO RESPONSE RATES IN CLINICAL TRIALS

The number of patients needed in an experimental and a control group for comparing two response rates based on α, β, and d have been investigated by Cochran Cox (1957) using sine transformation, which yield the approximate formula:

$$n_{opt} = \frac{(Z_{\alpha} + Z_{\beta})^2}{2(\sin^{-1} P_1 - \sin^{-1} P_2)^2} \, ,$$

where P_1 and P_2 are the response rates to treatments A and B. respectively,

Z_α and Z_β are the upper percentage level α and β. Later Gail and Gart (1973) used the exact test for the determination of sample sizes. Cochran and Cox's tables were then modified according to the exact test (Gehan and Schneiderman 1973).

Example

Suppose 20% patients are predicated to respond to a standard treatment. The comparative trail is to determine whether a new treatment, has a response rate of 50%.

Then P_1 =0.20,

P_2 = 0.50,

and

d=0.30,

Number of patients needed in each treatment group for one-sided test is 36 at a 5% significance level and 80% power and 76 at a 1% significance level and 95% power. The sample size required is roughly proportional to the desired power and selected significance level. Also, two-sided alternative hypothesis require large sample size than one-sided alternatives.

Main Determinants of Study Size

Recommendations when budget is not enough:

- (Estimation) Lower desired precision

(Hypothesis Testing) Lower desired power or increase minimum detectable effect size

- It is not recommended to change confidence levels, significance levels, or variance estimates.
- If after all these changes, budget is still insufficient, one has to decide between;

- Not conducting the study until enough budget has been obtained, or
- Go ahead with the study knowing that the result are likely to be inconclusive (pilot study or exploratory).

- Adjustment to Sample Size

- Non-Response(and attrition)

$$n_2 = n_1/(1\text{-}NR)$$

n_2= final size, n_1= effective size

NR= Non- response (and attrition) rate

Other Considerations

- Data dependencies (e.g., Matching, Repeated Mesures).
- Multivariable methods (e.g.,control for confounding).
- Multiplicity issues (e.g., Multiple testing, endponts, treatments, interim analyses).
- Other variables of interest (e.g., time to event).
- Other hypothesis (e.g., equlvaence).

Software

PASS

(www.ncss.com/pass.html)

nQUERY

(www.statsolusa.com/nquery.html)

EPI-INFO

(www.cdc.gov/epiinfo)

EPIDAT

(www.paho.org/English/SHA/epidat.htm)


THANK YOU