Sampling Methods in Medical Research

Odisha Social Science
Research & Consultancy

CONTENTS

- Introduction
- Need for and advantages of sampling
- Basic concepts
- Sampling Distribution
- Sampling Theory
- Formulae for computing standard error
- Sampling design or strategy
- Types of sample designs
- Determination of sample size

Learning Objective

- The trainees will be able to adopt suitable Sampling Design to Medical Research.

Sampling : An Introduction

- Selection of some part of an aggregate or totality on the basis of which an inference about the aggregate or totality is made.
- **Sample**: A representative part of the population.
- **Sampling design:** Process of selecting a representative sample.
- **Sample survey:** Survey conducted on the basis of sample.
- **Complete Enumeration Survey or Census inquiry:** A complete enumeration of all the items in the Population.

Need for and Advantages of Sampling

- Sampling has the following advantages over Census.
  - Less Resource (Money, Materials, Manpower & Time)
  - More accuracy: Due to better scope for employing trained manpower.
  - Inspection fatigue is reduced (non-sampling error)– Sampling error can be studied, controlled & probability statement can be made about magnitude.
  - Non-sampling error can not be estimated
  - Only way for destructive enumeration.
  - Only way when population size is infinite.

Disadvantages of sampling

- May not be a proper representative of the population.
- A chance of over estimation and under estimation.
- To estimate population parameter and the statistics should be unbiased. There are some parameter for which we cannot get the unbiased estimation.
- Sampling results may not be equal to the population results.
- Sample survey associated with both sampling and non-sampling errors.

  Census survey: only non-sampling error .

Basic Concept

- **UNIVERSE OR POPULATION:** It is the aggregate of objects from which sample is selected. Total of items about which information is desired aggregate of elementary units (finite or infinite, N) possess at least one common characteristics.
- **POPULATION: TARGET POPULATION AND SAMPLED POPULATION:** Target population is the one for which the inference is drawn. While the sampled population is the one from which sample is selected. This may be restricted to some extent than the target population due to practical difficulties.
- **FRAME:** It is the list of all the sampling units in the population. This should be complete, exhaustive, non-overlapping and up to date.

- **SAMPLING UNITS:** Units possessing the relevant characteristics i.e., attributes that are the object of study (operational definition).
- **SAMPLING DESIGN:** A definite plan for obtaining a sample from the sampling frame

  Refers to technique or procedure adopted by the researcher .
- **PARAMETERS AND STATISTICS:** The statistical constants of the population such as mean, variance etc. are referred to as parameters.

  Statistic: An estimate of the parameter, obtained from a sample, is a function of the sample values.

  A statistic ‘t’ is an unbiased estimate of the population parameter ‘θ’ if expectation of $t = \theta$.

- **SAMPLING ERRORS:** Errors which arise on account of sampling.
- Total Error= Sampling error + Non sampling Error

Reasons for sampling errors :

- Faulty selection of the sample
- Substitution : If difficulty arises in enumerating a particular sampling unit, it is usually substituted by a convenient unit of the population, this leads to some bias
- Faulty demarcation of sampling unit
- Error due to improper choice of the statistics for estimating the population parameters

Non-sampling errors

Non-sampling errors may be due to following reasons.

- **Faulty planning and definitions**
  - data specification being inadequate and inconsistent with respect to the objectives of the survey,
  - error due to the location of the unit and actual measurement of the characteristics, errors in recording the measurement, errors due to the ill designed questionnaire, etc. and
  - lack of trained and qualified investigator & lack of adequate supervisory staff.

- **Response errors:-** This errors are introduced as the result of the responses furnished by respondents and may be due to any of the following reasons.
- Response errors may be accidental due to mis-understanding in a particular question.
- May be due to prestige bias.
- Self interest.
- Bias due to investigation/ investigator.
- Failure of the respondent's memory.

- **Non-response bias**

  Non response biases occur if full information is not obtained on all the sampling unit. A rough classification of the types of non-response is as follows.

  - Non coverage
  - Not-at homes
  - Unable to answer
  - The hard core
- **Compiling errors**
- **Publication errors**

- Non sampling errors are likely to be more serious in a complete enumeration survey as compared to a sample survey.
- In a sample survey, the non sampling errors can be reduced by employing qualified, trained and experienced personnel, better supervision and better equipments for processing and analyzing relatively smaller data as compared to a complete census.
- Sampling error usually decreases with increase of sample size.
- On the other hand, as the sample size increases, the non-sampling error is likely to increase.

Basic Concepts contd.

- **PRECISION:** Range within which the population parameter will lie in accordance with the reliability specified in the confidence level
- **RELIABILITY OR CONFIDENCE LEVEL:** Expected % of times that the actual value will fall within the stated precision limits i.e.. the likelihood that the answer will fall within that range .
- **SIGNIFICANCE LEVEL:** The likelihood that the answer will fall outside the range.
- **SAMPLING DISTRIBUTION:** The aggregate of the various possible values of the statistics under consideration grouped into a frequency distribution is known as the sampling distribution of the statistic.

STANDARD ERROR:

–The standard deviation of a sampling distribution of a statistics its standard error; it is a key to sampling theory.

–Helps in testing whether difference between observed and expected frequency could arise due to chance.

–Gives an idea about the reliability and precision of a sample

–Enables to specify the limits within which the parameters of the population are expected to lie with a specified degree of confidence

Sampling Design or Strategy

- A definite plan for obtaining sample.
- Technique or procedure for selecting items for sample including the size of the sample
- It should be reliable & appropriate to research study and determined before data are collected

IMPORTANT ASPECTS IN SAMPLING DESIGN:

1.Type of population / universe

Structure, Composition & finité or infinité nature.

2. Sampling unit

Individual, group, family, institution, village, district, etc. Natural (e.g., Geographical) or constructed (e.g.. Social entity)

3. Sampling frame / source list

- Representative, comprehensive, correct, reliable& appropriate
- Ready to use or constructed for the purpose

4. Population parameters of specific interest

Important sub-groups in the population

5. Budgetary constraints

Non-probability sample is cheaper.

6. Size of sample

- Adequate to provide an estimate with sufficiently high precision
- Representative to mirror the various patterns and sub-classes of the population

- Neither too large nor too small, but optimum to meet efficiency, (cost) ,reliability (precision) & flexibility
- Higher the precision & larger the variance, the larger the size and more the cost.

7. Types of sample or sampling procedure

- For a given size, cost & precision, choose the one which has a smaller sampling error.

Characteristics of a Good sampling Design

1. Truly representative
2. Should have all the characteristics that are present in the population
3. Having small sampling error
4. Economically viable
5. Systematic bias is controlled (in a better way)
6. Results can be applied to the universe in general with a reasonable level of confidence or reliability
7. Optimum size (adequately large)

Types of Sample Designs

Probability sampling:

- Based on the concept of random selection & probability theory.
  - Simple Random Sampling
  - Complex Random Sampling (mixed sampling) Designs
    - Stratified Sampling
    - Cluster Sampling

- Area Sampling
- Systematic Sampling
- Multistage Sampling
- Sequential Sampling

Non-probability sampling

- Convenience or haphazard sampling
- Purposive / Deliberate sampling
- Judgment Sampling
- Quota Sampling
- Snowball sampling

Non-Probability Sampling

- Not based on probability theory
- Judgment of researcher / organizer plays important role
- Personal elements (bias) has a great chance to enter
- No assurance that every element has some specifiable chance of being included
- Representative-ness is in question
  - –sampling error cannot be measured
  - –saves time and money

1. Convenience or haphazard sampling:

– Selected at the convenience of the researcher

– No way to find representativeness

– Not to be used in descriptive / diagnostic studies & for causal studies

– Useful for formulative / exploratory studies, pilot surveys, testing questionnaires, pre-test phase, formulation of probability/ hypothesis **2.**

Purposive or Deliberate sampling

(I) JUDGEMENT SAMPLING

-Researcher deliberately or purposively draws a sample which he thinks is representative

- Personal biases of investigator have great chance; not possible to estimate sampling error.

(ii) QUOTA SAMPLING

- The selection of the sample is made by the interviewer, who has been given quotas to fill from specified sub-groups of the population.
- For example, an interviewer may be told to sample 50 females between the age of 45 and 60.
- There are similarities with stratified sampling, but in quota sampling the selection of the sample is non-random.

- Anyone who has had the experience of trying to interview people in the street knows how tempting it is to ask those who look most helpful, hence it is not the most representative of samples, but extremely useful.

Advantages

- Quick and cheap to organize.

Disadvantages

- Not as representative of the population as a whole as other sampling methods. Because the sample is non-random it is impossible to assess the possible sampling error.

3.Snowball sampling

In social science research, **snowball sampling** is a technique for developing a research sample where existing study subjects recruit future subjects from among their acquaintances.

Thus the sample group appears to grow like a rolling snowball.

This sampling technique is often used in hidden populations which are difficult for researchers to access; example populations would be drug users or commercial prostitutes.

Probability or Random or Chance Sampling

Sample survey Principles- Based on probability theory

Principle of statistical regularity

This lays down that a moderately large sample chosen at random from a large population almost sure on the average to possess the characteristic of the large population. (King).

Principle of validity

Validity of a sample design we mean that it should enable us to obtain valid tests and estimates about the population parameters.

Principle of optimization

Achieving a given level of efficiency at minimum cost and obtaining maximum possible efficiency with given level of cost.

Probability or Random or Chance Sampling

- **Simple Random Sampling (SRS)**

It is the technique of drawing a sample in a such way that each unit of the population has an equal and independent chance of being included in the sample.

In SRS from a population of N units the probability of drawing any specified unit in any specified draw is 1/N.

The probability that a specified unit is included in the sample is n/N. (n= sample size)

SRS can be defined equivalently as follows:

SRS is the technique of selecting the sample in such a way that each of the ${}^{N}C_{n}$ samples has an equal chance or probability ($p = 1/ {}^{N}C_{n}$) of being selected.

SRS with replacement (SRSWR)

In SRSWR the units selected in the earlier draws are replaced back in the population before the subsequent draws are made. Thus a unit has a chance of being included in the sample for more than once.

SRS without replacement (SRSWOR) – Most common

In SRSWOR the units selected in the earlier draws aren't replaced back in the population before the subsequent draws are made. Thus a unit has only one chance of being included in the sample.

SIMPLE RANDOM SAMPLING

The sample mean is an unbiased estimate of the population mean i.e. $E(\bar{y}_n) = \bar{Y}_N$

$$\bar{y}_n = \frac{\sum y_i}{n} \qquad \bar{Y}_N = \frac{\sum Y_i}{n}$$

The sample mean square is an unbiased estimate of the population mean square i.e. $E(s^2) = S^2$

Where $\mathbf{s^2} = \frac{1}{n-1} \sum [y_i - \bar{y}_n]^2$

S^2 = Mean square for the population

Where $\mathbf{S^2} = \frac{1}{N-1} \sum [Y_i - \bar{Y}_N]^2$

$$S.E(\bar{y}_n) = \sqrt{\frac{N-n}{N}} \frac{S}{\sqrt{n}} \qquad \text{Est } S.E(\bar{y}_n) = \sqrt{\frac{N-n}{N}} \frac{S}{\sqrt{n}}$$

SELECTION OF RANDOM SAMPLES FOR FINITE POPULATION

Lottery method (blind folded or rotating drum)

- All the population units are assigned numbers serially i.e. 1,2,3……N. N= population size
- N numbers of homogeneous chits are prepared
- Then one by one "n" number of chits are selected without replacement.

Merits

Very simple technique.
Based on probability law
Has got no personal bias.

Demerits

If the population size is very large then it is time taking.

Mechanical randomization

Different random number table

- Tipetts (1927) Random Number Table
- Fisher & Yates (1938)
- Kendall & Babington Smith`s (1939)
- Rand Corporation (1955) table of random numbers.
- C.R-Rao, Mitra & Mathai (1966) table of random numbers

Methods of Using random number table for selecting a random sample

- Identify N units in the population with the number 1 to N. Say ‘N’ is an r- digited number.
- Open at random any page of the table.
- Select a, column or row at random.
- Select a r- digited number from the column or row at random.
- Pick up r- digited numbers proceeding forward or backward in a systematic manner along any row or column selected at random.
- Consider only numbers less than equal to N and reject the numbers greater than N.
- Population units corresponding to numbers selected constitute the sample units.
- The procedure is continued till required numbers of units are selected. The procedure is continued till required numbers of units are selected.

Advantages of SRS

- Very simple technique to draw sample.
- It is a probability sampling and has got no personal bias.
- If variability in the population is less the sample provides a representative and the sampling is the best.
- The efficiency of the estimates of the parameter can be ascertained by considering the sampling distribution of the statistic.

Disadvantages

- Sample may over or under represent.
- If the population is heterogeneous SRS is not suitable because it may not provide a proper representative sample.
- Less efficient.
- To draw a SRS a up to date frame is required which may not be available.
- A SRS may result in the selection of the sampling units which are widely spread geographically and in such a case the cost of collecting data may be much in terms of time and money

Stratified Random Sampling (STRS)

The whole heterogeneous population of size (N) is divided in to "K" number of homogeneous subgroups called strata having sizes $N_1, N_2 \ldots\ldots\ldots N_k$.

- Then $n_1, n_2, \ldots\ldots n_k$ number of units are selected from $1^{st}, 2^{nd}, \ldots\ldots k_{th}$ strata by SRS
- $N = \sum Ni$ and $n = \sum n_i$ total sample size
- Stratified factor: Criteria for stratification

Principle Of Stratification

- ✓ Variability within the strata should be as less as possible and variability between strata as more as possible
- ✓ Strata should be mutually exclusive.

Advantages

- ✓ More representative
- ✓ Precision of STRS is more than SRS.
- ✓ Administratively more convenient
- ✓ Problem of the survey within each stratum can be solved independently.

Disadvantages

- ✓ Stratification should be done properly
- ✓ If study relates to multiple characteristics, the division into homogeneous layer is difficult.

Estimate of population Mean and Variance

Let k be the number of strata.

Let Y_{ij}, ($j = 1,2,\ldots N_i$; $i= 1,2,\ldots..k$) be the value of the j^{th} unit in the i^{th} stratum.

$\bar{Y}_{Ni} =$,population mean of i^{th} stratum $= \frac{1}{N_i}\sum Y_{ij}$

$\bar{Y}_N =$ population mean $= \frac{1}{N}\sum\sum Y_{ij} = \frac{1}{N}\sum N_i \bar{Y}_{Ni}$

$= \sum P_i \bar{Y}_{Ni}$

Where $P_i = N_i/N$ is called the weight of the i^{th} stratum.

S_i^2 = population mean square of the i^{th} stratum=

$$\frac{1}{N_i - 1}\sum (Y_{ij} - \bar{Y}_{Ni})^2 , (i=1,2,\ldots..,k)$$

y_{ij} = value of j^{th} sampled unit from i^{th} stratum

$\bar{y}_{ni}$ = mean of sample selected from i^{th} stratum.

s_i^2 = sample mean square of the i^{th} stratum

$$= \frac{1}{n_i - 1}\sum (y_{ij}-\bar{y}_{ni})^2 \; ; (i = 1,2,\ldots\ldots,k)$$

Let $\bar{y}_{st} = \frac{1}{N} \sum N_i \bar{y}_{ni} = \sum p_i \bar{y}_{ni}$ $p_i = n_i/N$

This is an unbiased estimate of the population mean

$$Var(\bar{y}_{st}) = \frac{1}{N^2}\sum N_i(N_i-n_i)\frac{S_i^2}{n_i} = \sum p_i^2(\frac{1}{n_i}-\frac{1}{N_i})S_i^2$$

$$Est\ (Var\ \bar{y}_{st}) = \sum (\frac{1}{ni} - \frac{1}{Ni}) \; p_i^2 s_i^2 = \frac{1}{N^2}\sum N_i(N_i-n_i)\frac{s_i^2}{n_i}$$

Allocation Of Sample Size to various Strata

(a) Proportional allocation

(b) Optimum allocation

(a) Proportional allocation

Allocation of n_i's various strata is called proportional if the sample fraction is constant for each stratum, i.e.,

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} \ldots\ldots = \frac{n_k}{N_k} = \frac{\sum n_i}{\sum N_i} = \frac{n}{N} = C \text{ (constant)}$$

Thus $n_i \alpha N_i$

Thus, in proportional allocation each stratum is represented according to its size.

In proportional allocation , $\mathrm{Var}(\bar{y}_{st})$ is given by

$$\mathrm{Var}(\bar{y}_{st})_{prop} = \left(\frac{1}{n} - \frac{1}{N}\right) \sum P_i S_i^2$$

(b) Optimum Allocation : Another guiding principle in the determination of the n_i's is to choose them so as to:

(1) $\mathrm{Var}(\bar{y}_{st})$ is minimum for fixed sample size 'n'.

(2) $\mathrm{Var}(\bar{y}_{st})$ is minimum for fixed total cost C(say)

(3) total cost C is minimum for fixed value of

$$\mathrm{Var}(\bar{y}_{st}) = V_0 \text{ (say)}$$

Systematic sampling

- In systematic sampling of size ‘n’ the first unit is selected by random number table.
- Then the rest (n-1) units are selected by some pre-determined pattern i.e. every unit at the k^{th} interval.
- Let us suppose that N sampling units are serially numbered from 1 to N in some order and a sample of size n is to be drawn such that N= nk

$$\Rightarrow k = \frac{N}{n}$$

Where k, usually called the sampling interval, is an integer.

- Systematic sampling consists in drawing a random number, say, $i \leq k$ and selecting the unit corresponding to this number and every k^{th} unit subsequently. Thus the systematic sample of size n will consists of the units i, i+k, i+2k, ……i+(n-1)k
- The random number 'i' is called the random start and its value determines as a matter of fact, the whole sample.
- Systematic sample mean is an unbiased estimate of population mean.

- $\text{Var}(\bar{y}_{sys}) = \frac{N-1}{N}.S^2 - \frac{(n-1)k}{N}.S^2_{wsy}$ where

$$S^2_{wsy} = \frac{1}{K(n-1)} \sum\sum (y_{ij} - \bar{y}_{i.})^2$$

S^2 = population mean square.

- A systematic sample is more precise than a simple random sample without replacement if the mean square within the systematic sample is larger than the population mean square. In other words, systematic sampling will yield better results only if the units within the same sample are heterogeneous.
- p_{wst}^2 is the correlation coefficient between deviation from stratum means of pairs of items that are in the same systematic sample.

- The relative efficiency of systematic sampling over stratified random sampling depends upon the values of p_{wst}^2 and nothing can be concluded in general.
- If $p_{wst}^2 > 0$, then $E' < 1$ and thus in the case stratified sampling will provide a better estimate of $\bar{y}_{..}$ However, if $p_{wst}^2 = 0$, then $E' = 1$ and consequently both systematic sampling and stratified sampling provide estimates of $\bar{y}_{..}$ with equal precision.

$$E' = \frac{Var(\bar{y}_{st})}{Var(\bar{y}_{sys})} = \frac{1}{1 + (n-1)\, p_{wst}}$$

Advantages

- Easier to use & less costlier for large population
- Sample is spread more evenly over the entire population
- Elements can be ordered in a manner found in the universe
- Can be used even without list of units in the population

Disadvantages

- Systematic samples are not in general random samples.
- May yield biased estimate if there are periodic features associated with sampling interval.

CLUSTER SAMPLING

1. Divide a large area of interest into a no. of smaller non overlapping areas / clusters
2. Randomly select some of these smaller areas
3. Choose all units in these sample small areas
   - –It is a trade off of economics and precision of sample estimates. i.e. it reduces cost but precision is also reduced
   - –Units in clusters tend to be homogenous & hence increasing sample size improves precision only marginally

Advantages:

- Reduces cost (more reliable per unit cost)
- Better field supervision
- No sampling frame necessary
- Ensures better cooperation of respondents as they are not isolate persons (for intimate data)
- As the cluster size increases the cost decreases

MULTI-STAGE SAMPLING

- Refers to a sampling techniques which is carried out in various stages.
- Population is regarded as made of a number of primary units each of which further composed of a number of secondary units.
- Consists of sampling first stage units by some suitable method of sampling.
- From among the selected first stage units, a sub-sample of secondary stage units is drawn by some suitable method of sampling which may be same as or different from the method used in selecting first stage unit.

Advantages:

II stage units are necessary only for selected I stage units

Flexible & allows different selection procedure

Easier to administer

A large number of units can be sampled for a given cost.

Area sampling: This is basically multistage sampling in which maps, rather than lists or registers, serve as the sampling frame. This is the main method of sampling in developing countries where adequate population lists are rare. The area to be covered is divided into a number of smaller sub-areas from which a sample is selected at random within these areas; either a complete enumeration is taken or a further sub-sample.

SEQUENTIAL SAMPLING

– Some what complex sampling design

– Size of the sample is not fixed in advance

– Size is determined as per mathematical decision rules as the survey progresses on the basis of information yielded

– If decision is taken to accept or reject based on single sample , then it is single sampling, if it is based on two samples it is double sampling.

–One goes on taking samples as long as one desires to do so

Determination of Sample Size

1. Nature of population :Size, Heterogeneous/ homogenous
2. Number of variables to be studied
3. Number of groups & sub-groups proposed
4. Nature of study (qualitative or quantitative)
5. Sampling design or type of sample
6. Intended depth of analysis
7. Precision and reliability
8. Level of non-response (item & unit) expected
9. Available finance and other resources

Sample Size Determination in health studies

ONE SAMPLE SITUATION

- **Estimating a population proportion with specified absolute precision**

  Required information and notation

  a) Anticipated population proportion = P

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

  (c) Absolute precision required on either side of he proportion (in percentage point) = d

$$n = z^{2}_{1-\alpha/2} P(1-P)/d^2$$

- ✓ If it is not possible to estimate P, a figure of 0.5 should be used; since the sample size required is largest when P= 0.5
- ✓ If 'P' is given as a range, the value closest to 0.5 should be used.

- Estimating a population proportion with specified relative precision

  Required information and notation

  a) Anticipated population proportion = P

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

  c) Relative precision = ε

- $$n = z^{2}_{1-\alpha/2}(1-P)/\varepsilon^{2}P$$

- The choice of P for the sample size computation should be as small as possible, since the smaller P is the greater is the minimum sample size.

- **Hypothesis tests for a population proportion**

  Required information and notation

  a) Test value of the population proportion under the null hypothesis = P_0

  b) Anticipated value of the population proportion=P_a

  c) Level of significance = 100 α %

  d) Power of the test = 100(1- β)%

  e) Alternative hypothesis : either $P_a > P_0$ or $P_a < P_0$ (for one sided test) $P_a \neq P_0$ (for two-sided test)

  - For a one- sided test

    $$n = \{z_{1-\alpha}\sqrt{[P_0(1-P_0)]}+z_{1-\beta}\sqrt{[P_a(1-P_a)]}\}^2/(P_0-P_a)^2$$

  - For a two sided test

    $$n = \{z_{1-\alpha/2}\sqrt{[P_0(1-P_0)]}+z_{1-\beta}\sqrt{[P_a(1-P_a)]}\}^2/(P_0-P_a)^2$$

TWO-SAMPLE SITUATIONS

- **Estimating the difference between two population proportions with specified absolute precision**

Required information and notation

a) Anticipated population proportion = P_1 and P_2

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

(c) Absolute precision required on either side of the true proportion (in percentage point) = d

d) Intermediate value = $V= P_1(1- P_1)+ P_2 (1- P_2)$

$$n = z^{2}_{1-\alpha/2} [P_1(1-P_1)+P_2(1-P_2)]/d^2$$

$$n = z^{2}_{1-\alpha/2} V/d^2$$

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

- ✓ If it isn't possible to estimate either population proportion, the safest choice of 0.5 should be used in both cases.
- ➢ The value of V may be obtained directly from table from the corresponding to P_2 (or its complement) and the row corresponding to P_1(or its complement)
- ➢ **Hypothesis test for two population proportion**

  This is designed to test the hypothesis that two population proportions are equal.

  Required information and notation

  a) Test value of the difference between the population proportions under the null hypothesis = $P_1 - P_2 = 0$

  b) Anticipated value of the population proportion = P_1 and P_2

  c) Level of significance = 100 α %

d) Power of the test = $100(1-\beta)\%$

e) Alternative hypothesis : either $P_a > P_0$ or $P_a < P_0$

(for one sided test) $P_a \neq P_0$ (for two-sided test)

$$\mathbf{n = \{z_{1-\alpha}\sqrt{[2\bar{P}(1-\bar{P})]}+z_{1-\beta}\sqrt{[P_1(1-P_1)+P_2(1-P_2)]}\}^2/(P_1-P_2)^2}$$

Where

$$\bar{P} = (P_1+P_2)/2$$

For a two sided test

$$\mathbf{n = \{z_{1-\alpha}\sqrt{[2\bar{P}(1-\bar{P})]}+z_{1-\beta}\sqrt{[P_1(1-P_1)+P_2(1-P_2)]}\}^2/(P_1-P_2)^2}$$

For a one sided test for small proportions

$$n = (z_{1-\alpha}+z_{1-\beta})^2/[0.00061(\arcsin\sqrt{P_2}-\arcsin\sqrt{P_1})^2]$$

For a two sided test for small proportions

$$n = (z_{1-\alpha/2}+z_{1-\beta})^2/[0.00061(\arcsin\sqrt{P_2}-\arcsin\sqrt{P_1})^2]$$

CASE CONTROL STUDIES

Classification of people exposure to the risk and disease

	Exposed	Unexposed
Disease	a	b
No disease	c	d

The odds ratio is then ad/bc.

- **Estimating an odds ratio with specified relative precision**

Required information and notation

(a) Two of the following should be known

- Anticipated probability of " exposure" for people with the disease [a/(a + b)] = P_1^*
- Anticipated probability of " exposure" for people without the disease [c/(c + d)] = P_2^*
- Anticipated odds ratio = OR

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

c) Relative precision = ε

$$n = z^2_{1-\alpha/2} \{1/[P_1^*(1-P_1^*)+ 1/P_2^*(1-P_2^*)]\}/[\log_e(1-\varepsilon)]^2$$

➢ Hypothesis test for an odd ratio

Required information and notation

(a) Test value of the odds ratio under the null hypothesis=OR_0= 1

(b) Two of the following should be known

- ➢ Anticipated probability of " exposure" for people with the disease [a/(a + b)] = P_1^*
- ➢ Anticipated probability of " exposure" for people without the disease [c/(c + d)] = P_2^*
- ➢ Anticipated odds ratio = OR_a

(c) Level of significance = 100 α %

(d) Power of the test = 100(1-β)%

(e) alternative hypothesis = $OR_a \neq OR_0$

$$n = z_{1-\alpha/2}\ [2P_2^*(1-P_2^*)]+z_{1-\beta}\sqrt{P_1^*(1-P_1^*)+P_2^*(1-P_2^*)]}^2/(P_1^*-P_2^*)^2$$

COHORT STUDIES

Estimating a relative risk with specified relative precision

Required information and notation

(a) 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 not exposed to the factor of interest = P_2
- Anticipated relative risk = RR

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

c) Relative precision = ε

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

- **Hypothesis test for a relative risk**

Required information and notation

(a) Test value of the relative risk under the null hypothesis=RR_0= 1

(b) Two of the following should be known

- Anticipated probability of disease in people exposed to the variable = P_1
- Anticipated probability of disease in people not exposed to the variable = P_2
- Anticipated relative risk = RR_a

(c) Level of significance = 100 α %

(d) Power of the test = 100(1-β)%

(e) Alternative hypothesis = $RR_a \neq RR_0$

$$n = \{z_{1-\alpha}\sqrt{[2\bar{P}(1-\bar{P})]}+z_{1-\beta}\sqrt{[P_1(1-P_1)+P_2(1-P_2)]}\}^2/(P_1-P_2)^2$$

$$\bar{P} = (P_1+P_2)/2$$

LOT QUALITY ASSURANCE SAMPLING

- **Accepting a population prevalence as not exceeding a specified value**

Required information and notation

(a) Anticipated population prevalence = P

(b) Population size = N

(c) Maximum number of sampled individuals showing characteristics = d^*

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

The value of n is obtained by solution of the inequality

$$\sum {}^{M}C_{x}\ {}^{(N-M)}C_{(n-x)} / {}^{N}C_{n} < \alpha$$

Where M=NP, for a finite population; or

$$\text{Prob}\{d \leq d^*\} < \alpha \quad \text{i.e.} \quad \sum \text{prob}\,(d) < \alpha$$

or $\sum {}^{n}C_{d}\ P^{d}\ (1-P)^{n-d} < \alpha$ for an infinite population

- **Decision rule for "rejecting a lot"**

Required information and notation

(a) Test value of the population proportion under the null hypothesis $= P_0$

(b) Anticipated value of the population proportion $= P_a$

(c) Level of significance $= 100\ \alpha\%$

(d) Power of the test $= 100(1-\beta)\%$

$$n = [z_{1-\alpha}\sqrt{\{P_0(1-P_0)\}}+z_{1-\beta}\sqrt{\{P_a(1-P_a)\}}]^2/(P_0-P_a)^2$$

$$d^* = [\, nP_0 - z_{1-\alpha}\sqrt{\{nP_0(1-P_0)\}}]$$

INCIDENCE-RATE STUDIES

- **Estimating an incidence rate with specified relative precision**

  Required information and notation

  (a) Relative precision = ε

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

$$n = (z_{1-\alpha/2} / \varepsilon)^2$$

- **Hypothesis tests for an incidence rate**

  Required information and notation

  (a) Test value of the population incidence rate under the null hypothesis= λ_0

  (b) Anticipated value of the population incidence rate $= \lambda_a$

  (c) Level of significance = $100\ \alpha\%$

  (d) Power of the test = $100(1-\beta)\%$

(e) Alternative hypothesis : either $\lambda_a > \lambda_0$ or $\lambda_a < \lambda_0$ (for one sided test) or $\lambda_a \neq \lambda_0$ (for two sided test)

For a one sided test $n = (z_{1-\alpha}\lambda_0 + z_{1-\beta}\lambda_a)^2/(\lambda_0-\lambda_a)^2$

For a two sided test $n = (z_{1-\alpha/2}\lambda_0 + z_{1-\beta}\lambda_a)^2/(\lambda_0-\lambda_a)^2$

- **Hypothesis tests for two incidence rates in follow-up (cohort) studies**

Required information and notation

(a) Test value of the difference between the population incidence rate under the null hypothesis= $\lambda_1 - \lambda_0 = 0$

(b) Anticipated value of the population incidence rate $= \lambda_1$ and λ_2

(c) Level of significance = 100 α%

(d) Power of the test = 100(1-β)%

(e) Alternative hypothesis : either $\lambda_1 - \lambda_0 > 0$ or $\lambda_1 - \lambda_2 < 0$ (for one sided test) or $\lambda_1 - \lambda_2 \neq 0$ (for two sided test)

(f) duration of study (if fixed) = T

For one sided test $n = (z_{1-\alpha}\lambda_0 + z_{1-\beta}\lambda_a)^2/(\lambda_0 - \lambda_a)^2$

For two sided test $n = (z_{1-\alpha}\lambda_0 + z_{1-\beta}\lambda_a)^2/(\lambda_0 - \lambda_a)^2$

For study duration not fixed

For one sided test

$$n = \{z_{1-\alpha}\sqrt{[(1+k)\bar{\lambda}^2]} + z_{1-\beta}\sqrt{(k\lambda_1^2 + \lambda_2^2)}\}^2/k(\lambda_1 - \lambda_2)^2$$

For two sided test

$$n = \{z_{1-\alpha/2}\sqrt{[(1+k)\bar{\lambda}^2]} + z_{1-\beta}\sqrt{(k\lambda_1^2 + \lambda_2^2)}\}^2/k(\lambda_1 - \lambda_2)^2$$

Where $\bar{\lambda} = (\lambda_1+\lambda_2)/2$ and k is the ratio of the sample size for the second group of subjects(n_2) to that for the first group (n_1)

THANK YOU