July 21 - 24, 2021 | Garki Hospital Abuja | Resource Person

Research Methodology Boot Camp

with Epi Info Training

Dr. Adamu Onu
MBBS, FWACP (FM)
MS Epidemiology & Biostatistics
PhD Public Health (Epidemiology)

Target Audience Clinical Researchers, Post-Part 1 Residents, and Others

Important Information

- Limited slots are available on a first come, first served basis
- Laptop running Windows 10 required
- Organized as morning lecture sessions and afternoon hands on coaching sessions

Highlights

- Research Methodology
- Research Design
- Data Management
- Sample Size Calculations
- Test Statistics
- Interpretation of Results
- Report Writing
- Hands-on training sessions
- Statistical consulting sessions

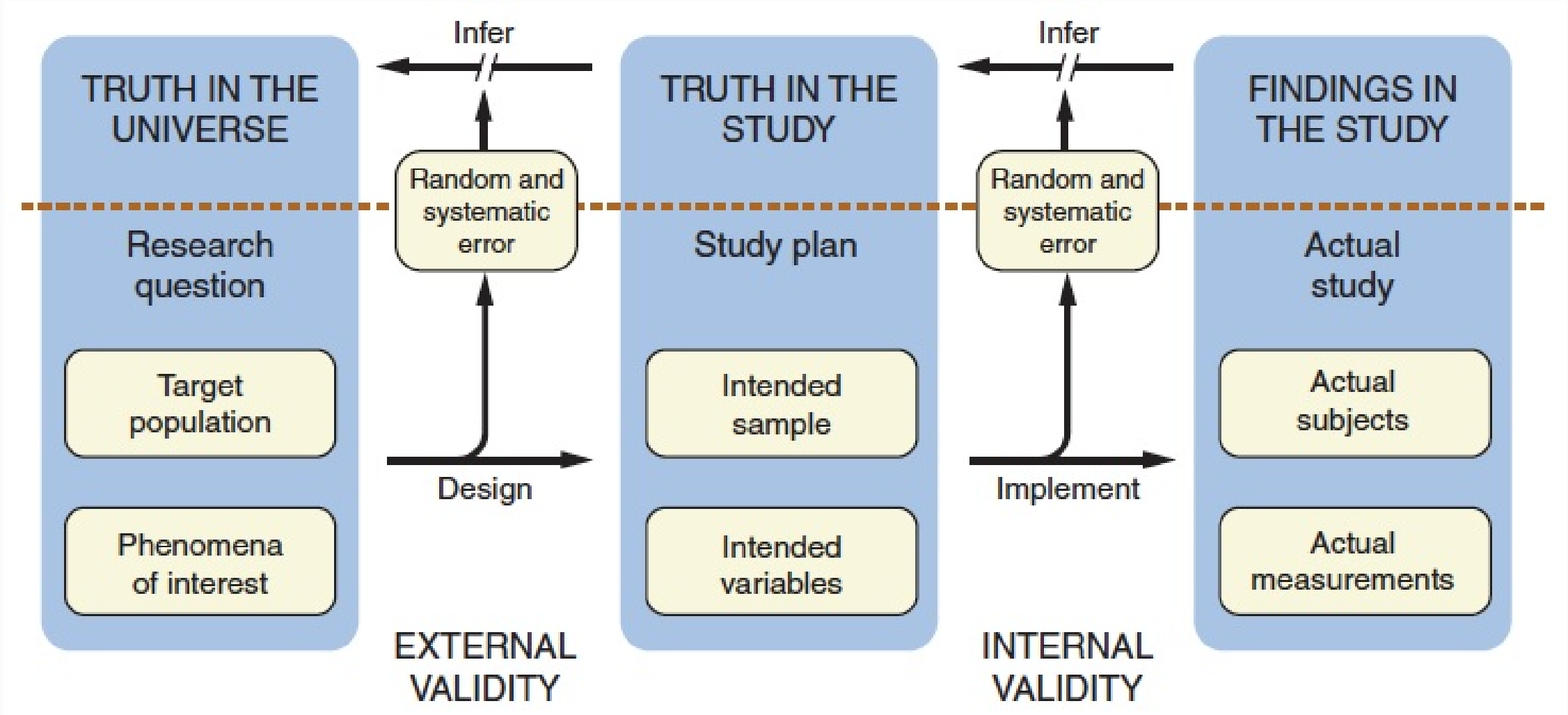
For further details contact
Email: epimetrix@gmail.com
Phone: +234 803 474 9930

epimetrix
health systems
data for health decisions

Sampling Methods and Sample Sizes

Introduction

- Research often involves the observation of a **sample** from some predefined **population** of interest.
- Conclusions drawn from the study are based on generalizing the results from the sample to the entire population from which the sample was drawn.
- The accuracy of the conclusions will depend on:
  - how the samples have been collected, and
  - how representative the sample is of the population.


Infer
Infer
TRUTH IN THE UNIVERSE
Random and systematic error
TRUTH IN THE STUDY
Random and systematic error
FINDINGS IN THE STUDY
Research question
Study plan
Actual study
Target population
Intended sample
Actual subjects
Phenomena of interest
Intended variables
Actual measurements
Design
Implement
EXTERNAL VALIDITY
INTERNAL VALIDITY

Terminology

Population

- All members of a defined group
- *Parameters*:
  - Mean - μ
  - Standard deviation - σ
  - Proportion - π

Sample

- A subset of the entire population
- *Statistics*:
  - Mean - $\bar{x}$
  - Standard deviation - *s*
  - Proportion - *p*

Sampling

> Sampling is a process of choosing a section of the population for observation and study.

- Used when it is impractical to measure an entire population (e.g. prevalence of breast cancer in Nigerian women)
- Reduces cost, time, and personnel required
- Sample should be representative of population as a whole
- Predictive power is based on both sample size and quality of sampling method.

Process of sampling

> What determines a proper sample?

- The sample should be representative of the population.
- Every variable of interest should have the same distribution in the sample as in the population from which the sample is chosen.

Process of sampling

1. The population has to be clearly defined — the sampling frame.
2. How large a sample should be selected?
3. How should the individual units be selected?

Sampling frame

- *A list of all elements (units) in the population*
  - Population surveys – list of people
  - Clinical trials – list of patients
  - Case-control study – list of people with disease and list of people without the disease
- Inclusion and exclusion criteria define the general framework for the population

Sampling Methods

- Once the population has been identified, we need to decide how we are going to choose the sample from the population.
  - Simple random sample
  - Stratified sampling
  - Cluster sampling
  - Systematic sampling
  - Multi-stage sampling

Simple random sample

- This is the most common and the simplest of the sampling methods.
- In this method, the subjects are chosen from the population with **equal probability** of selection.
- One may use a random number table, or balloting or computer programs to draw simple random samples from a given population

Advantages of simple random sample

- It is easy to administer
- It is representative of the population in the long run
- The analysis of data is straightforward

Disadvantages of simple random sample

- The selected sample may not be truly representative of the population, especially if the sample size is small.

Stratified sampling

- When the size of the sample is small and we have some information about the distribution of a particular variable, we can select simple random samples from within each of the subgroups (*strata*) defined by that variable.
- Gender: 50% ♂/50% ♀: Choosing half the sample from males and half from females assures that the sample is representative of the population with respect to gender.
- Stratified sampling can reduce potential confounding by selecting homogeneous subgroups when confounding is an important issue (e.g. in case-control studies).

Cluster sampling

- Using simple random sample in administrative surveys done on large, geographically dispersed populations imposes large costs and is inconvenient.
- In such cases, clusters may be identified (e.g. households) and random samples of clusters will be included in the study:
  - Every member of the cluster will also be part of the study
- This introduces two types of variations in the data:
  - Between clusters – *intercluster* variance
  - within clusters – *intracluster* variance
  - This variation have to be accounted for when analysing data.

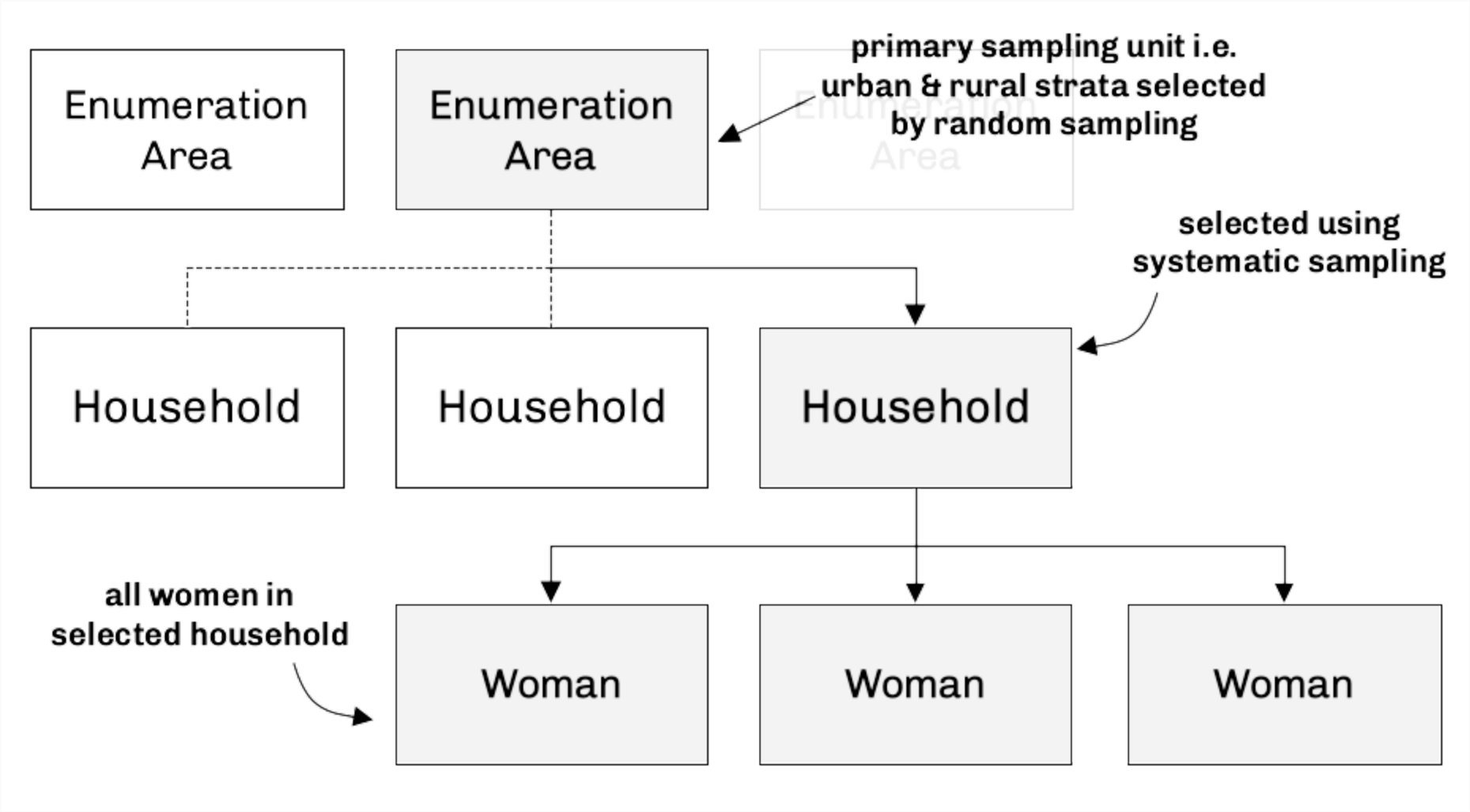
Systematic sampling

- First subject of population chosen at random
- Subsequently others are selected in a systematic way (e.g. every 5th person)
- Other methods preferred

Multi-stage sampling

Many studies, especially large nationwide surveys, will incorporate different sampling methods for different groups, and may be done in several stages.

Multi-stage sampling



Non-probability sampling

- Should be avoided for prevalence studies
- Frequently used in health research (patients who arrive for care become the sample)
- Impossible to avoid selection bias (e.g. poor patients, not too poor, belief in medicine)
- One cannot know the chances that a particular subject will be included in the sample (e.g. probability of a person in Abuja coming to Garki Hospital is not the same for every member of population).

Sample size

- The main determinant of the sample size is how accurate the results need to be.
- This depends on the purpose of the study:
  - **Descriptive** study to determine a *summary measure* of a characteristic
  - **Analytical** study where specific sets of *hypotheses are being tested*.

Sample size for descriptive study

- The main objective is to obtain an estimate of a population parameter e.g. proportion of people who smoke, or average daily caloric intake of the population.
- The sample size required to answer these questions depends on several factors:
  - Measure of interest
  - Underlying probability distribution
  - The sampling distribution of the measure
  - Desired level of accuracy

Sample size for estimating a population proportion (p)

> Suppose we want to conduct a survey to determine the prevalence (π) of a relatively common disease in a community. We want to determine how many people should be observed to obtain a reasonably accurate picture of the prevalence

$$n = p(1-p)\left(\frac{z_{1-\alpha/2}}{\delta}\right)^2$$

> The problem is to calculate the sample size required for estimating the prevalence of the disease within ± 5% of the true value, with 95% confidence.

Specify the parameters

1. Confidence coefficient (1 – α) = 95%
2. Width of the interval (Δ) = 10%
3. Make a guess as to the value of π = 30%

$$n = p(1-p)\left(\frac{z_{1-\alpha/2}}{\delta}\right)^2$$

$$n = 30 \times 70\left(\frac{1.96}{5}\right)^2 = 323$$

> We need a minimum of 323 subjects observed to assure that the 95% confidence interval for the estimated proportion will be within 5% of the true prevalence.

- The above calculation assumes a **simple random sample** from a relatively large population.
- Often the population from which the samples are drawn may be fixed and small, in which case corrections to the above formula is required.

$$n = \text{deff} \times \frac{N\hat{p}\hat{q}}{\frac{\delta^2}{z_\alpha^2}(N-1) + \hat{p}\hat{q}}$$

- Where deff = design effect; N = population size; and $\hat{q} = 1 - \hat{p}$

Sample size for estimating a population average (μ)

> Suppose we want to estimate the average daily caloric intake of people in a community.

$$n = \left[\frac{z_{1-\alpha/2} \cdot \sigma}{\delta} \right]^2$$

- The daily caloric intake is assumed to have a normal distribution around μ, with a standard deviation (σ).

> The problem is to calculate the sample size required for estimating the average daily caloric intake within ± 25 cal. of the true value with 95% confidence.

Specify the parameters

1. Confidence coefficient (1 – α) = 95%
2. Width of the interval (Δ) = 50 cal.
3. Obtain the standard deviation (σ) = 150 cal.

$$n = \left[\frac{z_{1-\alpha/2} \cdot \sigma}{\delta} \right]^2$$

$$n = \left[\frac{1.96 \times 150}{50} \right]^2$$

$$n = 35$$

> We need a minimum of 35 subjects observed to assure that the 95% confidence interval for the estimated average daily caloric will be within ±25 cal. of the true average.

Sample size for analytical studies

- The primary purpose of an analytical study is to test null hypotheses
- The sample size calculations requires the specification of the limits of the type I and type II errors.
- One also has to determine the sample measures used (a proportion, a sample mean, an estimate of RR or OR, etc.) and their sampling distribution.

Testing equality of two proportions: $p_1 = p_2$.

- The sample measures used are the sample proportions
- The sampling distribution used in testing this null hypothesis is either the standard normal distribution (z), or equivalently the chi-square (χ^2).

$$n = \left[\frac{z_{1-\alpha/2}\sqrt{2\bar{p}(1-\bar{p})} - z_{\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}}{\delta} \right]^2$$

Where $\bar{p} = \frac{(p_1 + p_2)}{2}$

N.B.

- The calculation of a type II error, or β depends on a precise definition of “null hypothesis is not true.”
- The simplest way to do this is to define the smallest difference (Δ) in the two proportions that we consider meaningful (*clinically significant difference*) and calculate β under this hypothesis.

> We are interested in determining the sample size required in a clinical trial of a new drug that is expected to improve survival. Suppose the traditional survival rate is 40%, i.e. $p_2 = 0.4$. We are interested in detecting whether the new drug improves survival by at least 10%, i.e. $\delta = 0.10$

Specify the parameters

1. Set the type I error: $\alpha = 0.05$, $\therefore z_{1-\alpha/2} = 1.96$
2. Determine “minimum clinically significant difference”: $\delta = 0.10$
3. Make a guess as to the “proportion” in one group (usually ‘control’): $p_1 = 0.40$, $\therefore p_2 = 0.50$
4. Determine the power required to detect this difference: $1 - \beta = 0.95$, $\therefore z_\beta = -1.645$

Substituting these values in the equation above gives n = 640

> Thus the study would require 640 subjects in each of the two groups to assure a probability of detecting an increase in the survival rate of 10% or more with 95% certainty, if the statistical test used 5% as the level of significance.

Comparison of two population means

- The sampling distribution of the difference of the sample means has an approximately normal distribution.
- The standard error of difference depends on the standard deviations of the measurements in each of the population.
- In the simplest (and most commonly used) scenario, the two standard deviations are considered to be the same.

- We need to determine the minimum difference (Δ) in the means that we are interested in detecting by statistical test: the two types of statistical errors (α and β) and the standard deviation (σ).

$$n = \left[\frac{(z_{1-\alpha/2} - z_{\beta}) \cdot \sigma}{\delta} \right]^2$$

> Suppose we want to test a drug that reduces blood pressure. We want to say the drug is effective if the reduction in blood pressure is 5 mmHg or more, compared with the ‘placebo’. Suppose we know that systolic blood pressure in a population is distributed normally, with a standard deviation of 8 mmHg.

Specify the parameters

1. Set the type I error: $\alpha = 0.05$, $\therefore z_{\alpha/2} = 1.96$
2. Determine “minimum clinically significant difference”: $\delta = 5$ mmHg
3. Power required to detect this difference: $1 - \beta = 0.95$, $\therefore z_{\beta} = -1.645$
4. Standard deviation: $\sigma = 8$ mmHg

$$n = \left[\frac{(1.96 + 1.645) \cdot 8}{5}\right]^2 = 33.3$$

> The sample size in this study will be 34 subjects in each group

Comparison of more than two groups and multivariate methods

- The formulae for these situations are much more complicated.
- Simple formulae for the calculation of sample sizes are not available for multivariate analyses, such as those using multiple linear regression, logistic regression, or comparison of survival curves

> Computer programs are readily available for most of these cases and even situations not discussed here.

> The computations in this presentation are solely for illustrative purposes.

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

- For one-sided test

$$n_1 = \frac{1}{k} \cdot \left[\frac{z_{1-\alpha}\sqrt{(1+k)\bar{\lambda}^2} + z_{1-\beta}\sqrt{(k\lambda_1^2 + \lambda_2^2}}{\lambda_1 - \lambda_2} \right]^2$$

$$\bar{\lambda} = (\lambda_1 + \lambda_2)/2$$

$$k = n_2/n_1$$

Software

- OpenEpi
- StatCalc
- G*Power
- Stata

> When planning studies, one of the crucial steps is in deciding how large the study should be, and appropriate guidance should be sought from experts.

Discussion Session