# Sampling techniques

# What is research?

- "Scientific research is systematic, controlled, empirical, and critical investigation of natural phenomena guided by theory and hypotheses about the presumed relations among such phenomena."
  - Kerlinger, 1986

- *Research is an organized and systematic way of finding answers to questions*

# Important Components of Empirical Research

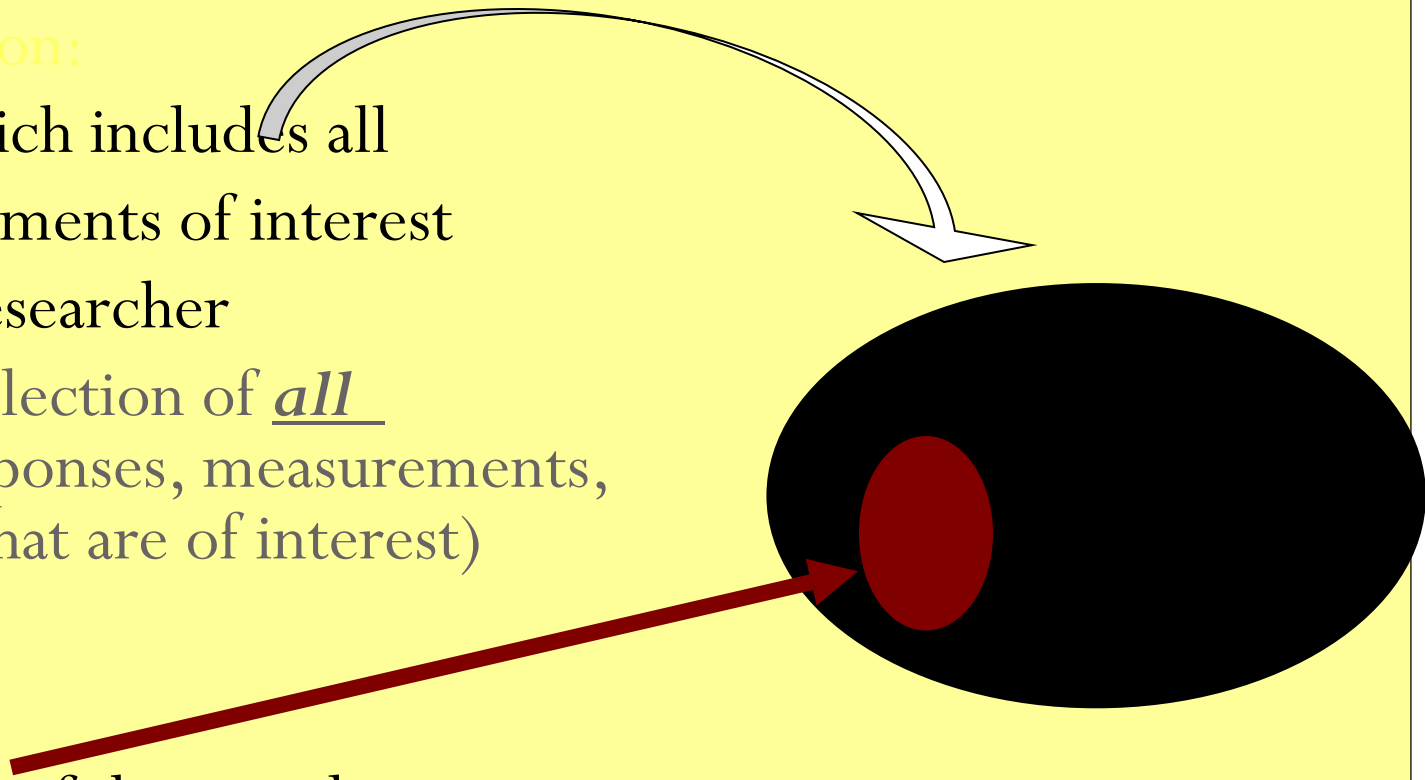
- Problem statement, research questions, purposes, benefits
- Theory, assumptions, background literature
- Variables and hypotheses
- Operational definitions and measurement
- Research design and methodology
- Instrumentation, sampling
- Data analysis
- Conclusions, interpretations, recommendations

# Important statistical terms

Population:

a set which includes all measurements of interest to the researcher

(The collection of ***all*** responses, measurements, counts that are of interest)



Sample:

A subset of the population

# Why sampling?

Get information about large populations

- Less costs
- Less field time
- More accuracy i.e. Can Do A Better Job of Data Collection
- When it's impossible to study the whole population

# SAMPLING

- A sample is “a smaller (but hopefully representative) collection of units from a population used to determine truths about that population” (Field, 2005)
- Why sample?
  - Resources (time, money) and workload
  - Gives results with known accuracy that can be calculated mathematically
- The sampling frame is the list from which the potential respondents are drawn
  - rosters
  - Must assess sampling frame errors

- 3 factors that influence sample representative-ness
  - Sampling procedure
  - Sample size
  - Participation (response)

- When might you sample the entire population?
  - When your population is very small
  - When you have extensive resources
  - When you don’t expect a very high response

**Target Population:**

The population to be studied/ to which the investigator wants to generalize his results

**Sampling Unit:**

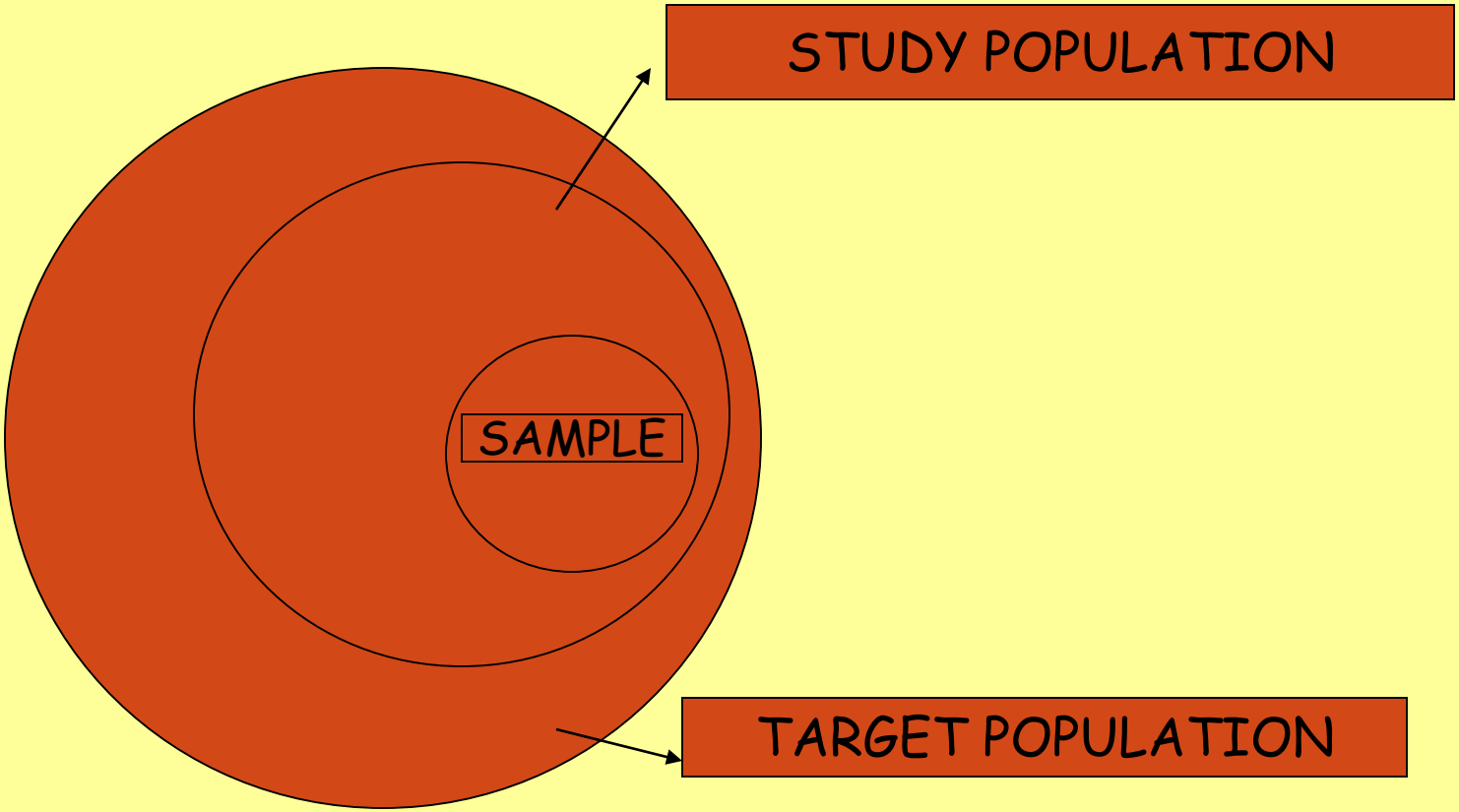
smallest unit from which sample can be selected

**Sampling frame**

List of all the sampling units from which sample is drawn

**Sampling scheme**

Method of selecting sampling units from sampling frame


STUDY POPULATION
SAMPLE
TARGET POPULATION

# Types of sampling

- Non-probability samples
- Probability samples

# Non probability samples

- Convenience samples (ease of access)

sample is selected from elements of a population that are easily accessible

- Snowball sampling (friend of friend….etc.)
- Purposive sampling (judgemental)
  - You chose who you think should be in the study
  - Quota sample

# Non probability samples

Probability of being chosen is unknown

Cheaper- but unable to generalise

potential for bias

# NON PROBABILITY SAMPLING

- Any sampling method where some elements of population have *no* chance of selection (these are sometimes referred to as 'out of coverage'/'undercovered'), or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is nonrandom, nonprobability sampling not allows the estimation of sampling errors..

- *Example: We visit every household in a given street, and interview the first person to answer the door. In any household with more than one occupant, this is a nonprobability sample, because some people are more likely to answer the door (e.g. an unemployed person who spends most of their time at home is more likely to answer than an employed housemate who might be at work when the interviewer calls) and it's not practical to calculate these probabilities*.

# Probability samples

- Random sampling
  - Each subject has a known probability of being selected
- Allows application of statistical sampling theory to results to:
  - Generalise
  - Test hypotheses

# PROBABILITY SAMPLING

- A **probability sampling** scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined.

- . When every element in the population *does* have the same probability of selection, this is known as an 'equal probability of selection' (EPS) design. Such designs are also referred to as 'self-weighting' because all sampled units are given the same weight.

# Types of Samples

- Probability (Random) Samples
- Simple random sample
  - Systematic random sample
  - Stratified random sample
  - Multistage sample
  - Multiphase sample
  - Cluster sample
- Non-Probability Samples
  - Convenience sample
  - Purposive sample
  - Quota

# Process

- The sampling process comprises several stages:
  - Defining the population of concern
  - Specifying a sampling frame, a set of items or events possible to measure
  - Specifying a sampling method for selecting items or events from the frame
  - Determining the sample size
  - Implementing the sampling plan
  - Sampling and data collecting
  - Reviewing the sampling process

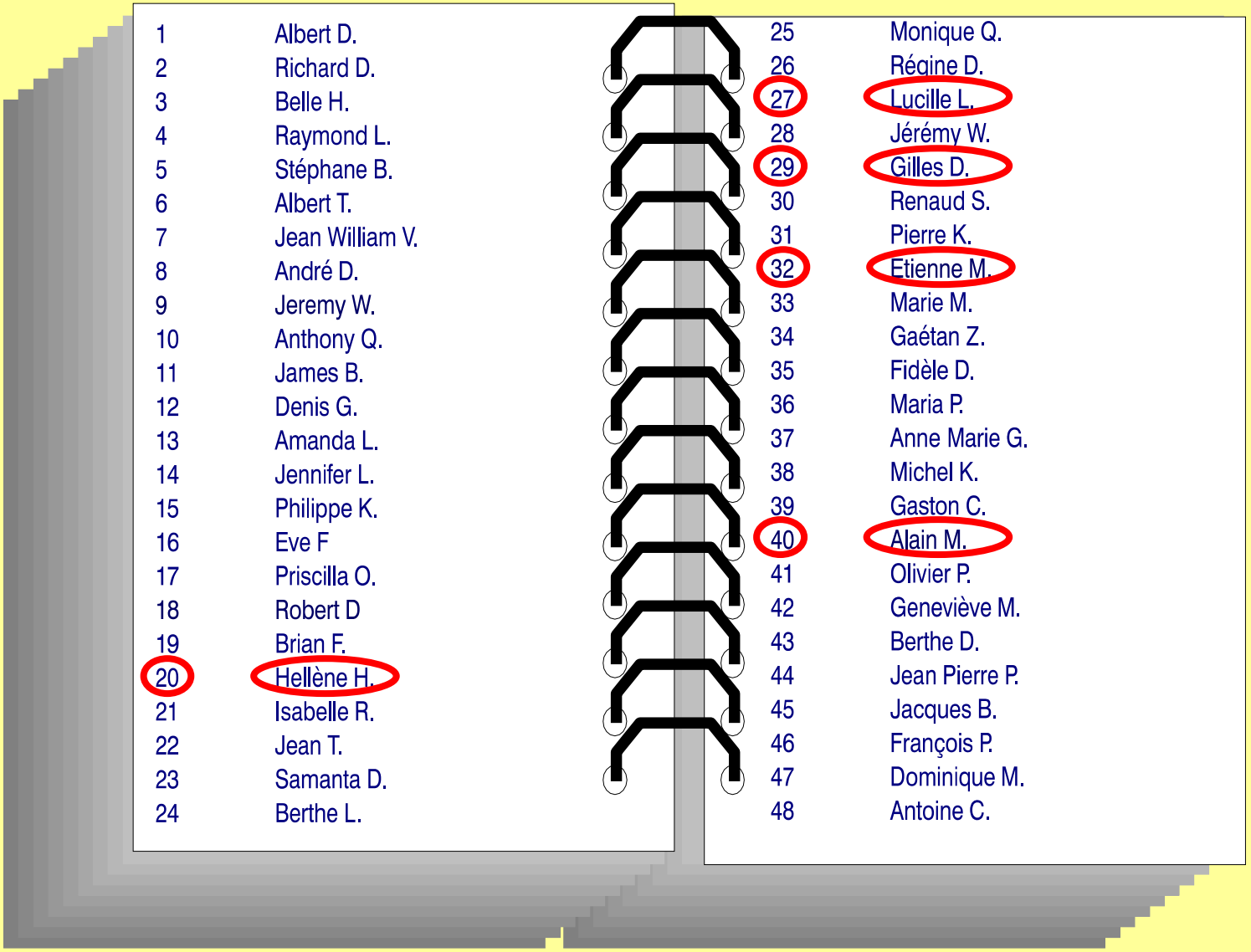
# SIMPLE RANDOM SAMPLING

- Applicable when population is small, homogeneous & readily available
- All subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection.
- It provides for greatest number of possible samples. This is done by assigning a number to each unit in the sampling frame.
- A table of random number or lottery system is used to determine which units are to be selected.

- Estimates are easy to calculate.
- Simple random sampling is always an EPS design, but not all EPS designs are simple random sampling.

- **Disadvantages**
- If sampling frame large, this method impracticable.
- Minority subgroups of interest in population may not be present in sample in sufficient numbers for study.

# Simple random sampling



# Table of random numbers

6 8 4 2 5 7 9 5 4 1 2 5 6 3 2 1 4 0

5 8 2 0 3 2 1 5 4 7 8 5 9 6 2 0 2 4

3 6 2 3 3 3 2 5 4 7 8 9 1 2 0 3 2 5

9 8 5 2 6 3 0 1 7 4 2 4 5 0 3 6 8 6

# Systematic sampling

Sampling fraction

Ratio between sample size and population size

# Systematic sampling

# SYSTEMATIC SAMPLING

- **Systematic sampling** relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list.
- Systematic sampling involves a random start and then proceeds with the selection of every *k*th element from then onwards. In this case, *k*=(population size/sample size).
- It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the *k*th element in the list.
- A simple example would be to select every 10th name from the telephone directory (an 'every 10th' sample, also referred to as 'sampling with a skip of 10').

# STRATIFIED SAMPLING

Where population embraces a number of distinct categories, the frame can be organized into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected.

- Every unit in a stratum has same chance of being selected.
- Using same sampling fraction for all strata ensures proportionate representation in the sample.
- Adequate representation of minority subgroups of interest can be ensured by stratification & varying sampling fraction between strata as required.

# STRATIFIED SAMPLING……

- Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata.

- **Drawbacks** to using stratified sampling.
- First, sampling frame of entire population has to be prepared separately for each stratum
- Second, when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata.
- Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods

# POSTSTRATIFICATION

- Stratification is sometimes introduced after the sampling phase in a process called "poststratification".
- This approach is typically implemented due to a lack of prior knowledge of an appropriate stratifying variable or when the experimenter lacks the necessary information to create a stratifying variable during the sampling phase. Although the method is susceptible to the pitfalls of post hoc approaches, it can provide several benefits in the right situation. Implementation usually follows a simple random sample. In addition to allowing for stratification on an ancillary variable, poststratification can be used to implement weighting, which can improve the precision of a sample's estimates.

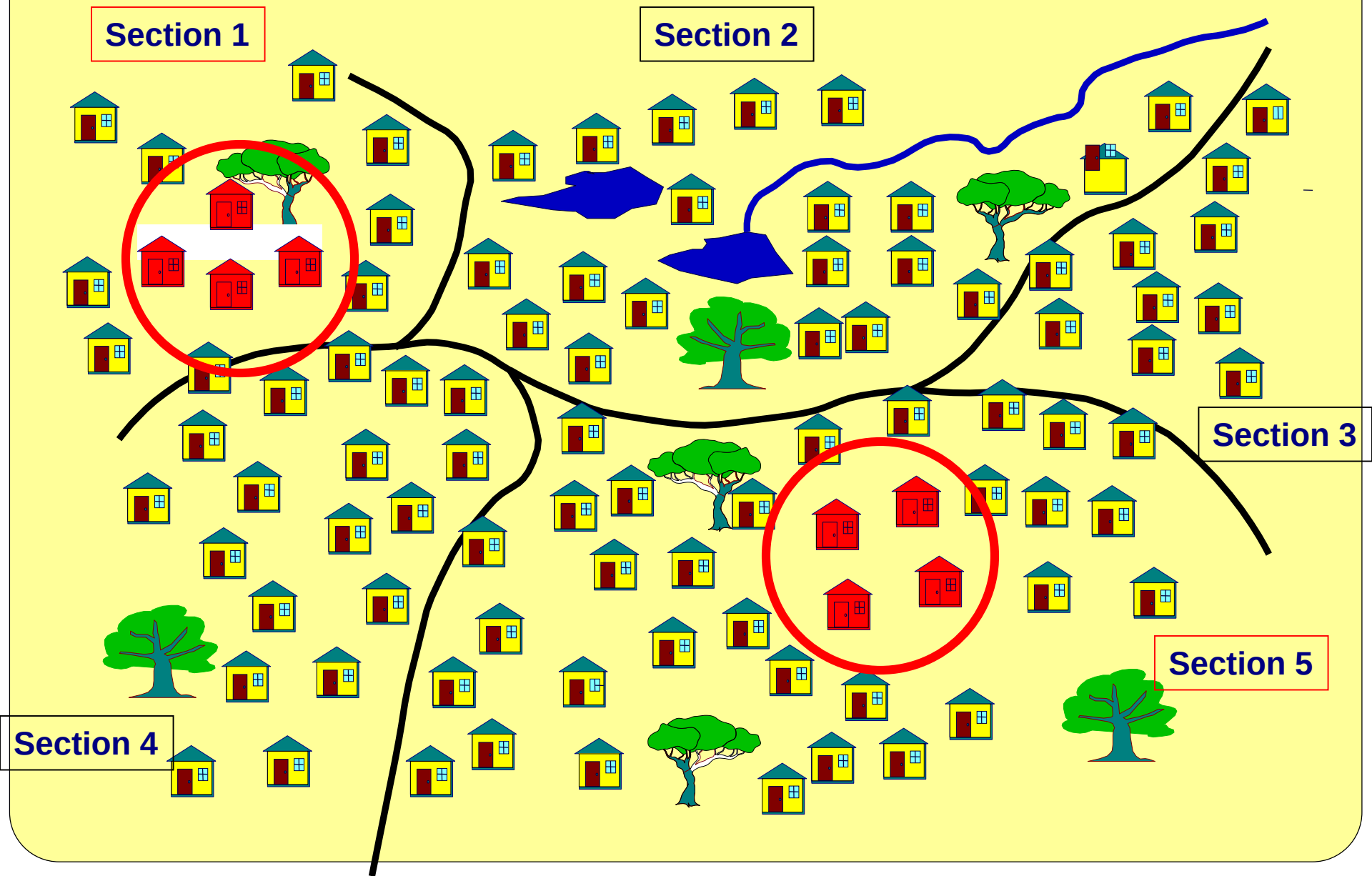
# OVERSAMPLING

- Choice-based sampling is one of the stratified sampling strategies. In this, data are stratified on the target and a sample is taken from each strata so that the rare target class will be more represented in the sample. The model is then built on this biased sample. The effects of the input variables on the target are often estimated with more precision with the choice-based sample even when a smaller overall sample size is taken, compared to a random sample. The results usually must be adjusted to correct for the oversampling.

# Cluster sampling

Cluster: a group of sampling units close to each other i.e. crowding together in the same area or neighborhood

# Cluster sampling



# CLUSTER SAMPLING

- Cluster sampling is an example of 'two-stage sampling' .
- First stage a sample of areas is chosen;
- Second stage a sample of respondents *within* those areas is selected.
- Population divided into clusters of homogeneous units, usually based on geographical contiguity.
- Sampling units are groups rather than individuals.
- A sample of such clusters is then selected.
- All units from the selected clusters are studied.

- Advantages :
- Cuts down on the cost of preparing a sampling frame.
- This can reduce travel and other administrative costs.
- Disadvantages: sampling error is higher for a simple random sample of same size.
- Often used to evaluate vaccination coverage in EPI

- **Identification of clusters**
  - List all cities, towns, villages & wards of cities with their population falling in target area under study.
  - Calculate cumulative population & divide by 30, this gives sampling interval.
  - Select a random no. less than or equal to sampling interval having same no. of digits. This forms 1st cluster.
  - Random no.+ sampling interval = population of 2nd cluster.
  - Second cluster + sampling interval = 4th cluster.
  - Last or 30th cluster = 29th cluster + sampling interval

Two types of cluster sampling methods.

**One-stage sampling**. All of the elements within selected clusters are included in the sample.

**Two-stage sampling**. A subset of elements within selected clusters are randomly selected for inclusion in the sample.

# CLUSTER SAMPLING.....

| | Freq | c f | cluster |
|---|---|---|---|
| I | 2000 | 2000 | 1 |
| II | 3000 | 5000 | 2 |
| III | 1500 | 6500 | |
| IV | 4000 | 10500 | 3 |
| V | 5000 | 15500 | 4, 5 |
| VI | 2500 | 18000 | 6 |
| VII | 2000 | 20000 | 7 |
| VIII | 3000 | 23000 | 8 |
| IX | 3500 | 26500 | 9 |
| X | 4500 | 31000 | 10 |
| XI | 4000 | 35000 | 11, 12 |
| XII | 4000 | 39000 | 13 |
| XIII | 3500 | 44000 | 14,15 |
| XIV | 2000 | 46000 | |
| XV | 3000 | 49000 | 16 |
| XVI | 3500 | 52500 | 17 |
| XVII | 4000 | 56500 | 18,19 |
| XVIII | 4500 | 61000 | 20 |
| XIX | 4000 | 65000 | 21,22 |
| XX | 4000 | 69000 | 23 |
| XXI | 2000 | 71000 | 24 |
| XXII | 2000 | 73000 | |
| XXIII | 3000 | 76000 | 25 |
| XXIV | 3000 | 79000 | 26 |
| XXV | 5000 | 84000 | 27,28 |
| XXVI | 2000 | 86000 | 29 |
| XXVII | 1000 | 87000 | |
| XXVIII | 1000 | 88000 | |
| XXIX | 1000 | 89000 | 30 |
| XXX | 1000 | 90000 | |

- 90000/30 = 3000 sampling interval

# MULTISTAGE SAMPLING

- Complex form of cluster sampling in which two or more levels of units are embedded one in the other.
- First stage, random number of districts chosen in all states.
- Followed by random number of talukas, villages.
- Then third stage units will be houses.
- All ultimate units (houses, for instance) selected at last step are surveyed.

# MULTISTAGE SAMPLING……..

- This technique, is essentially the process of taking random samples of preceding random samples.
- Not as effective as true random sampling, but probably solves more of the problems inherent to random sampling.
- An effective strategy because it banks on multiple randomizations. As such, extremely useful.
- Multistage sampling used frequently when a complete list of all members of the population not exists and is inappropriate.
- Moreover, by avoiding the use of all sample units in all selected clusters, multistage sampling avoids the large, and perhaps unnecessary, costs associated with traditional cluster sampling.

# MULTI PHASE SAMPLING

- Part of the information collected from whole sample & part from subsample.

- In Tb survey MT in all cases – Phase I
- X –Ray chest in MT +ve cases – Phase II
- Sputum examination in X – Ray +ve cases - Phase III

- Survey by such procedure is less costly, less laborious & more purposeful

# MATCHED RANDOM SAMPLING

A method of assigning participants to groups in which pairs of participants are first matched on some characteristic and then individually assigned randomly to groups.

- The Procedure for Matched random sampling can be briefed with the following contexts,
- Two samples in which the members are clearly paired, or are matched explicitly by the researcher. For example, IQ measurements or pairs of identical twins.
- Those samples in which the same attribute, or variable, is measured twice on each subject, under different circumstances. Commonly called repeated measures.
- Examples include the times of a group of athletes for 1500m before and after a week of special training; the milk yields of cows before and after being fed a particular diet.

# QUOTA SAMPLING

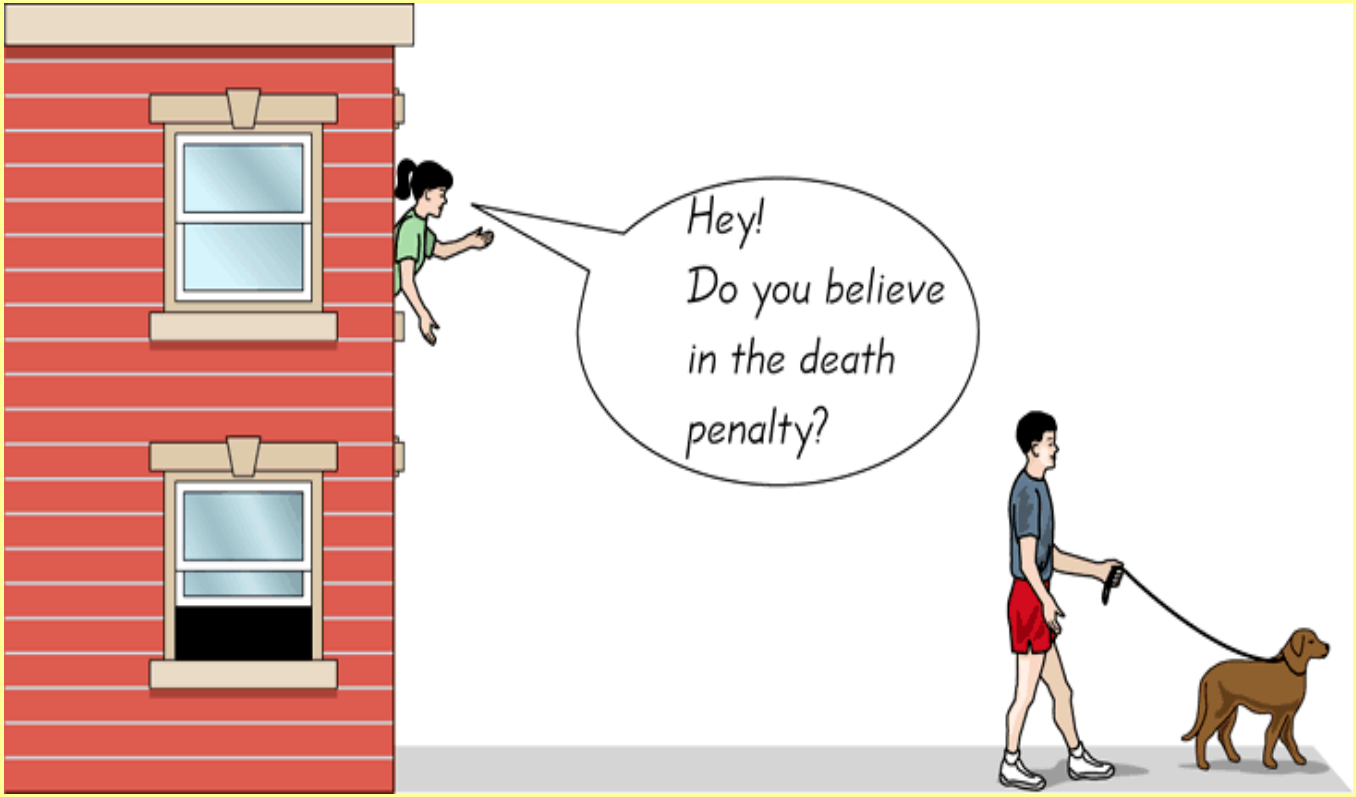
- The population is first segmented into mutually exclusive sub-groups, just as in stratified sampling.
- Then judgment used to select subjects or units from each segment based on a specified proportion.
- For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.
- It is this second step which makes the technique one of non-probability sampling.
- In quota sampling the selection of the sample is non-random.
- For example interviewers might be tempted to interview those who look most helpful. The problem is that these samples may be biased because not everyone gets a chance of selection. This random element is its greatest weakness and quota versus probability has been a matter of controversy for many years

# CONVENIENCE SAMPLING

- Sometimes known as **grab** or **opportunity sampling or accidental or haphazard sampling.**
- A type of nonprobability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, readily available and convenient.
- The researcher using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative enough.
- For example, if the interviewer was to conduct a survey at a shopping center early in the morning on a given day, the people that he/she could interview would be limited to those given there at that given time, which would not represent the views of other members of society in such an area, if the survey was to be conducted at different times of day and several times per week.
- This type of sampling is most useful for pilot testing.
- In social science research, snowball sampling is a similar technique, where existing study subjects are used to recruit more subjects into the sample.

# CONVENIENCE SAMPLING…….

- Use results that are easy to get






42

# Judgmental sampling or Purposive sampling

- - The researcher chooses the sample based on who they think would be appropriate for the study. This is used primarily when there is a limited number of people that have expertise in the area being researched

# PANEL SAMPLING

- Method of first selecting a group of participants through a random sampling method and then asking that group for the same information again several times over a period of time.
- Therefore, each participant is given same survey or interview at two or more time points; each period of data collection called a "wave".
- This sampling methodology often chosen for large scale or nation-wide studies in order to gauge changes in the population with regard to any number of variables from chronic illness to job stress to weekly food expenditures.
- Panel sampling can also be used to inform researchers about within-person health changes due to age or help explain changes in continuous dependent variables such as spousal interaction.
- There have been several proposed methods of analyzing panel sample data, including growth curves.

# Sample size determination in qualitative study

- Probability sampling not appropriate as sample not intended to be statistically representative
- But, sample should have ability to represent salient characteristics in population.
- Sample size taken until point of theoretical saturation

- Sample size is usually small to allow in-depth exploration and understanding of phenomena under investigation
- Ultimately a matter of judgement and expertise in evaluating the quality of information against final use, research methodology , sampling strategy and results is necessary.
- In practice, qualitative sampling usually requires a flexible, pragmatic approach.

- The researcher actively selects the most productive sample to answer the research question.
- This can involve developing a framework of the

  variables that might influence an individual's contribution and will be based on the researcher's practical knowledge of the research area, the available literature and evidence from the study itself.
- This is a more intellectual strategy than the simple demographic stratification of epidemiological studies, though age, gender and social class might be important variables.

.....

- If the subjects are known to the researcher, they may be stratified according to known public attitudes or beliefs
- It may be advantageous to study a broad range of subjects :
  - (maximum variation sample)
  - outliers (deviant sample)
  - subjects who have specific experiences (critical case sample)
  - subjects with special expertise (key informant sample).

- The iterative process of qualitative study design means that samples are usually theory driven ( theoretical sampling) to a greater or lesser extent
- Theoretical sampling necessitates building interpretative theories from the emerging data and selecting a new sample to examine and elaborate on this theory.
- It is the principal strategy for the **grounded theoretical approach .**

# Some suggestions of sample size in qualitative studies

- The smallest number of participants should be 15
- Should lie under 50
- 6-8 participants for FGDs AND at least 2 FGDs per population group

**IMPORTANT**

- Attainment of saturation
- Justification of choice of number

## Sample size determination in quantitative study

Several criteria will need to be specified to determine the appropriate sample size:

- Level of precision,
- Level of confidence or risk,
- Degree of variability in the attributes being measured ( prevalence)
- External validity

- **The Level of Precision-***sometimes called sampling error*
  - range in which the true value of the population is estimated to be.
  - This range is often expressed in percentage points (e.g., ±5 percent).
- **The Confidence Level**
  - based on ideas encompassed under the Central Limit Theorem.
  - E.g a 95% confidence level is selected, 95 out of 100 samples will have the true population value within the range of precision

.....

## Degree of Variability

- refers to the distribution of attributes in the population.
- The more heterogeneous a population, the larger the sample size required to obtain a given level of precision.
- The less variable (more homogeneous) a population, the smaller the sample size.

.....

- A proportion of 50 % indicates a greater level of variability than either 20% or 80%. This is because 20% and 80% indicate that a large majority do not or do, respectively, have the attribute of interest.
- Because a proportion of 0.5 indicates the maximum variability in a population, it is often used in determining a more conservative sample size, that is, the sample size may be larger than if the true variability of the population attribute were used.

.....

- Sample size affects accuracy of representation; Larger sample means less chance of error
- Minimum suggested sample is 30 and upper limit is 1,000

**External validity –** how well sample generalizes to the population, a representative sample is required (not the same thing as variety in a sample)

# Strategies for Determining Sample Size

There are several approaches to determining the sample size.

- Using a census for small populations
- Imitating a sample size of similar studies
- Using published tables
- Applying formulas to calculate a sample size
- Use computer soft ware e.g EPI-info series

# Using a Census for Small Populations

....

- One approach is to use the entire population as the sample.
- Although cost considerations make this impossible for large populations.
- Attractive for small populations (e.g., 200 or less).
- Eliminates sampling error and provides data on all the individuals in the population.
- Some costs such as questionnaire design and developing the sampling frame are "fixed," that is, they will be the same for samples of 50 or 200.
- Finally, virtually the entire population would have to be sampled in small populations to achieve a desirable level of precision

# Using a Sample Size of a Similar Study

- Use the same sample size as those  of studies similar to the one you plan( Cite reference).
- Without reviewing the procedures employed in these studies you may run the risk of repeating errors that were made in determining the sample size for another study.
- However, a review of the literature in your discipline can provide guidance about “typical” sample sizes that are used.

## Using Published Tables

- Published tables provide the sample size for a given set of criteria.
- Necessary for given combinations of precision, confidence levels and variability.
- The sample sizes presume that the attributes being measured are distributed normally or nearly so.
- Although tables can provide a useful guide for determining the sample size, you may need to calculate the necessary sample size for a different combination of levels of precision, confidence, and variability.

Sample Size for ±5%, ±7% and ±10% Precision Levels where Confidence Level Is 95% and P=.5.

| Size of Population | Sample Size (n) for Precision (e) of: | | |
|---|---|---|---|
| | ±5% | ±7% | ±10% |
| 100 | 81 | 67 | 51 |
| 125 | 96 | 78 | 56 |
| 150 | 110 | 86 | 61 |
| 175 | 122 | 94 | 64 |
| 200 | 134 | 101 | 67 |
| 225 | 144 | 107 | 70 |
| 250 | 154 | 112 | 72 |
| 275 | 163 | 117 | 74 |
| 300 | 172 | 121 | 76 |
| 325 | 180 | 125 | 77 |
| 350 | 187 | 129 | 78 |
| 375 | 194 | 132 | 80 |
| 400 | 201 | 135 | 81 |
| 425 | 207 | 138 | 82 |
| 450 | 212 | 140 | 82 |

# Using Formulas to Calculate a Sample Size

- Sample size can be determined by the application of one of several mathematical formulae.
- Formula mostly used for calculating a sample for proportions.

**For example**:

- For populations that are large, the Cochran (1963:75) equation yields a representative sample for proportions.
- Fisher equation, Mugenda etc

# Cochran equation

$$n_0 = \frac{Z^2 pq}{e^2}$$

Where $n_0$ is the sample size,

$Z^2$ is the abscissa of the normal curve that cuts off an area α at the tails;

(1 – α) equals the desired confidence level, e.g., 95%);

e is the desired level of precision,

p is the estimated proportion of an attribute that is present in the population,and q is 1-p.

The value for Z is found in statistical tables which contain the area under the normal curve. e.g Z = 1.96 for 95 % level of confidence

.....

## A Simplified Formula For Proportions

- Yamane (1967:886) provides a simplified formula to calculate sample sizes.
- ASSUMPTION:
  - 95% confidence level
  - P = .5 ;

.....

$$n = \frac{N}{1 + N(e)^2}$$

Where n is the sample size,
N is the population size,
e
is the level of precision.

# Finite population correction for proportions

- With finite populations, correction for proportions is necessary
- If the population is small then the sample size can be reduced slightly.
- This is because a given sample size provides proportionately more information for a small population than for a large population.
- The sample size ($n_0$) can thus be adjusted using the corrected formulae

.....

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}$$

Where n is the sample size
N is the population size.
$n_o$ is calculated sample size for infinite population

# Use of software in sample size determination

Depending on type of study and specific software Some information will be required:

- Population sample size, population standard deviation, population sampling error, confidence level, z –value, power of study etc …
- 80% **power** in a clinical trial means that the **study** has a 80% chance of ending up with a p value of less than 5% in a statistical test (i.e. a statistically significant treatment effect) if there really was an important difference (e.g. 10% versus 5% mortality) between treatments.