SAMPLE SIZE

Prof Dr Najlaa Fawzi

Sample Size?

Questions:
How large should my sample be?

Answer:
It depends...

...large enough to be an accurate representation of the population

...large enough to achieve statistically significant results

The sample size is an important feature of any study in which the goal is to make inferences about a population from a sample.
In practice, the sample size used in a study is determined based on the expense of data collection, and the need to have sufficient statistical power.

Finding a sample size can be one of the most challenging tasks in statistics and depends upon many factors including the size of your original population.

Sample size should be estimated early in the design phase of the study, when major changes are still possible.

The choosing of sample size depends on non-statistical considerations and statistical considerations.
The non-statistical considerations may include availability of resources, manpower, budget, ethics and sampling frame.

The statistical considerations will include the desired precision of the estimate of prevalence and the expected prevalence problem.

- **“ How large should my sample be in order for it to be representative” ?**

- **Larger samples are not necessarily better – how representative a sample it depends on the sampling technique used *and* the size of the population.**

- **Determining sample size is dependent of how much error you are prepared to accept in your sample.**

Factors Affecting Sample Size

1-Variability of the population characteristic under investigation.

2-Level of confidence desired in the estimate, most often 95% (within 2 SD).

3- Degree of precision desired in estimating the population characteristic, e.g. sampling error = +/- 2.

4- Other . . .

Sampling Error and Confidence

- **The larger the sample size the more likely error in the sample will decrease.**
- **But, beyond a certain point increasing sample size does not provide large reductions in sampling error.**
- **Accuracy is a reflection of the sampling error and confidence level of the data.**

- **If a sample has been selected according to probability we can assess the level of confidence.**
- **Confidence levels will allow you to state, with a certain level of confidence, that the sample findings would also be found in the population.**

Determining Sample Size

1- What data do you need to consider

2- Variance or heterogeneity of population

3-The degree of acceptable error (confidence interval)

4-Confidence level

Generally, we need to make judgments on all these variables

Calculating sample size

Specific method used depends on

- **The specific aim(s)/objective(s).**
- **The study design, including the planned number of measurements per ‘subject’.**
- **The outcome(s) and predictor(s).**
- **The proposed statistical analysis plan.**

Will also need to consider:

- Accrual/Enrollment (response rate for questionnaires).
- Drop-outs (i.e., lost to follow-up) and missing data.
- Budgetary constraints.

Determination of sample size for estimated means

The objective in interval estimation are to obtain narrow interval with high reliability .

Width of the interval is determined by the magnitude of the quantity

(reliability coefficient) x(SE of the estimator)

The total width of interval is twice this amount , this usually called the precision of the estimate or the margin of error

For a given SE , increasing reliability means a larger reliability coefficient .

But a larger reliability coefficient for a fixed SE makes for a wider interval.

The only way to obtain a small SE is to take a large sample

How large a sample?

That depends on the size of SD of the population , the desired degree of reliability ,and the desired interval width.

d = reliability of coefficient x SE of the estimator

$$d = Z \times \delta / \sqrt{n}$$

$$n = Z_2 X\ \delta_2 / d_2$$

Estimation of δ_2 : **The population variance is , as a rule , unknown .**

As a result, δ_2 has to be estimated .

The frequently used sources of estimates for δ_2 are the following :

1- A pilot or preliminary sample may be drawn from the population , and the variance computed from this sample may be used as an estimator of δ2 .

2- Estimates of δ2 may be available from previous or similar studies.

3- If it thought that the population from which the sample is to be drawn is approximately normally distributed , one may use the fact that the range is approximately equal to six SD and compute

$$\delta = R/6$$

Rule of thumb: the value of standard deviation is expected to be 1/6 of the range.

This method requires some knowledge of the smallest and largest value of the variable in the population.

A health department nutritionist ,whishing to conduct a survey among a population of teenage girls to determine their average daily protein intake [measured in grams].

C.I for protein intake about 10 grams ,
Margin of error 5g of the population mean.

95% C.I. population SD 20 g from previous study .

$$n = (1.96)^2 \times (20)^2 / (5)^2 = 61.47 = 62$$

(we round up to the next- largest whole number)

Determination Of Sample Size For Estimating Proportions

Assuming a random sample will selected from an infinite population , or when the sampled population is large enough .

The following formula used:

$$n = Z_2 P q / d_2$$

$$q = 1-p$$

Estimating P: P , the proportion in the population possessing the characteristic of interest , since this is the parameter we are trying to estimate , will be unknown . One solution is to take a pilot sample and compute an estimate to be used in place of P in the formula for n.

Sometimes an investigator will have some notion of an upper bound for P that can be used in the formula .

If it is impossible to come up with a better estimate ,one may set P equal to 0.5 and solve for n , since P= 0.5 in the formula yields the maximum value of n, this procedure will give a large enough sample for the desired reliability and interval width.

It may, however , be larger than needed and result in more expensive sample than if a better estimate of P had been available

This procedure should be used only if one is unable to arrive at a better estimate of P.

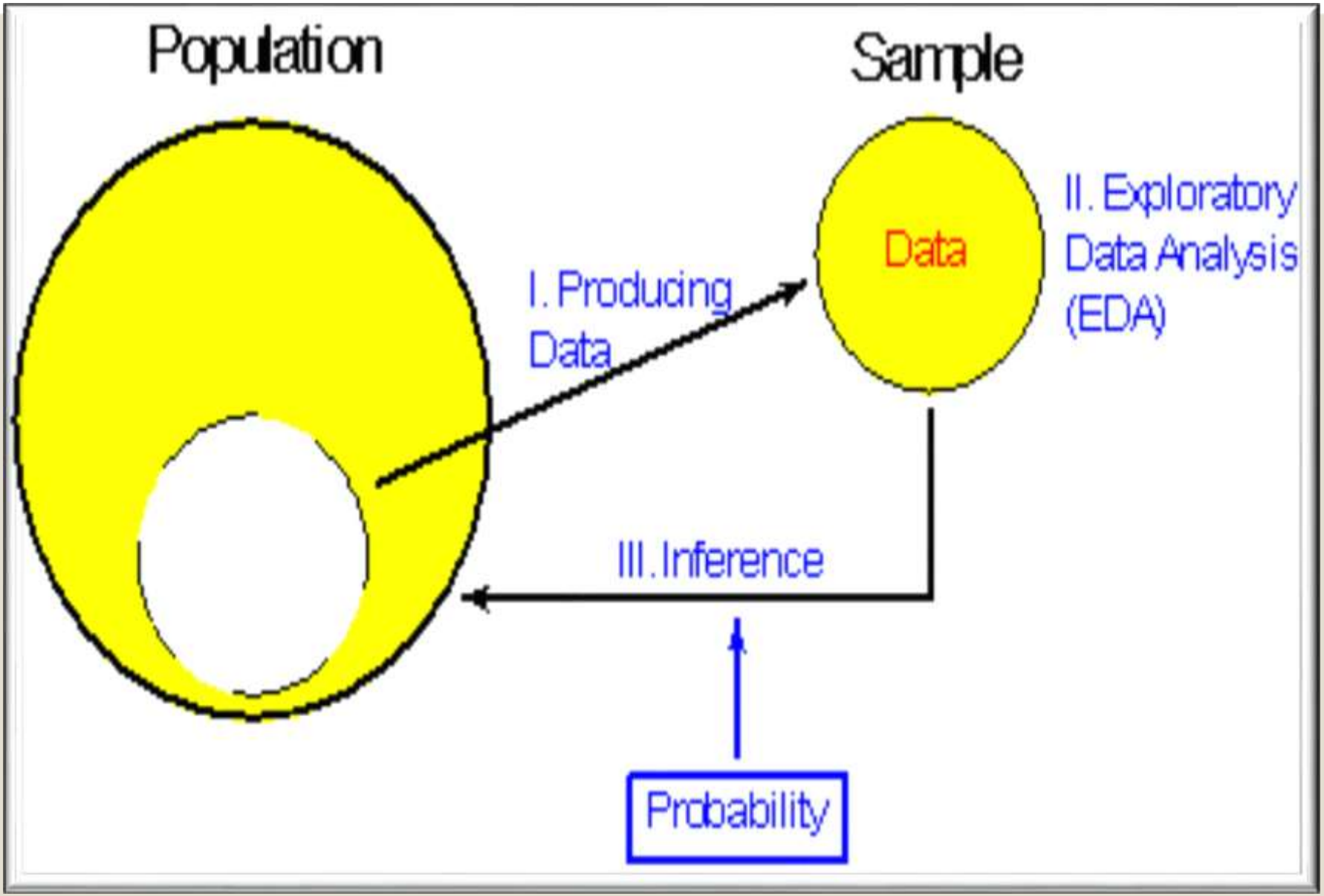
A health planning agency wishes to know, for certain area, what proportion of patients admitted to hospitals for the treatment of trauma die in the hospital .

A 95 percent confidence interval is desired , the width of the interval must be 0.06 , and the population proportion , from other evidence , is estimated to be 0.20 . How large a sample is needed?

$$n = Z_2 P q / d_2$$

$$n = (1.96)_2 * (0.20) * (0.80) / (0.03)_2$$

$$n = 682.88 = 683$$


Population
Sample
Data
II. Exploratory Data Analysis (EDA)
I. Producing Data
III. Inference
Probability

Step 1: Base Sample-size Calculation

The appropriate sample size for a population-based survey is determined largely by three factors:

(I) the estimated prevalence of the variable of interest – chronic malnutrition in this instance,
(ii) the desired level of confidence and
(iii) the acceptable margin of error.

For a survey design based on a simple random sample, the sample size required can be calculated according to the following formula.

$$N = \frac{t^2 \times p(1-p)}{m^2}$$

Description:

n = required sample size
t = confidence level at 95% (standard value of 1.96)
p = estimated prevalence of malnutrition in the project area
m = margin of error at 5% (standard value of 0.05)

Example

In project in Morocco, it has been estimated that roughly 30% (0.3) of the children in the project area suffer from chronic malnutrition. This figure has been taken from national statistics on malnutrition in rural areas. Use of the standard values listed above provides the following calculation.

Calculation:

$$n = \frac{1.96^2 \times .3(1-.3)}{.05^2}$$

$$n = \frac{3.8416 \times .21}{.0025}$$

$$n = \frac{.8068}{.0025}$$

$$n = 322.72 \sim 323$$

Step 2: Design Effect

The anthropometric survey is designed as a cluster sample (a representative selection of villages), not a simple random sample. To correct for the difference in design, the sample size is multiplied by the design effect (D).
The design effect is generally assumed to be 2 for nutrition surveys using cluster-sampling methodology.

Example
n x D = 323 x 2 = 646

Step 3: Contingency

The sample is further increased by 5% to account for contingencies such as non-response or recording error.

Example

n + 5% = 646 x 1.05 = 678.3 ˜ 678

What is the population size?
If you don't know

How many people are there to choose your random sample from? The sample size doesn't change much for populations larger than 20,000.

Your recommended sample size is **348**

This is the minimum recommended size of your survey. If you create a sample of this many people and get responses from everyone, you're more likely to get a correct answer than you would from a large sample where only a small percentage of the sample