# Research Summary

اللي ينتظرون الملخص

**The red ( important for sure ) but the highlighted red more important it's from the gifts 🎁**
**For the last 2 lectures for more examples give the lectures a look but according to the doctor ( I'll give you a scenario and you have to choose the appropriate test, no calculations )**
**وصلنا الى نهاية الفلم شكرا لمتابعتكم**

MED439
KING SAUD UNIVERSITY

## Editing file

Black: in male AND female slides

Red : important

Gray: extra information

Yellow = note

# L16 Sampling Techniques to select study subjects

| | |
|---|---|
| **What is Sampling** | Sampling is the process or technique of selecting a study sample of appropriate characteristics and of adequate size. |
| **Why to use Sampling in Research ?** | ˙Unable to study all members of a population<br>˙Reduce selection bias<br>˙Save time and money<br>˙Measurements may be better in sample than in entire population<br>˙Feasibility |
| **Definitions** | - Population – group of things (people) having one or more common characteristics "a set which includes all measurements of interest to the researcher (The collection of all responses, measurements, or counts that are of interest)"<br>- Sample – representative subgroup of the larger population ˙Used to estimate something about a population (generalize) and ˙Must be similar to population on characteristic being investigated " (A subset of the population) |
| **Sampling Frame** | This is the complete list of sampling units in the target population to be subjected to the sampling procedure.<br>Completeness and accuracy of this list is essential for the success of the study. |
| **Sampling Units** | These are the individual units / entities that make up the frame just as elements are entities that make up the population. |
| **Sampling Error** | This arises out of random sampling and is the discrepancies between sample values and the population value. |
| **Sampling Variation** | - Due to infinite variations among individuals and their surrounding conditions.<br>- Produce differences among samples from the population and is due to chance. |
| **Example:** | In a clinical trail of 200 patients we find that the efficacy of a particular drug is 75%<br>If we repeat the study using the same drug in another group of similar 200 patients we will not get the same efficacy of 75%. It could be 78% or 71%.<br>"Different results from different trails though all of them conducted under the same conditions" |
| **Representativeness (validity)** | A sample should accurately reflect distribution of relevant variable in population<br>(Person e.g. age, sex - Place e.g. urban vs. rural - Time)<br>Representativeness essential to generalise<br>Ensure representativeness before starting, Confirm once completed |

**Validity of a Study Two components of validity: Internal validity and External validity**

| | |
|---|---|
| **Internal validity** : A study is said to have internal validity when there have been **proper selection of study group and a lack of error in measurement.**<br>›For example, it is Concerned with the appropriate measurement of exposure, outcome, and association between exposure and disease | External validity: External validity implies the ability to generalize beyond a set of observations to some universal statement. |

| | |
|---|---|
| **How to sample ?** | In general, 2 requirements<br>1. Sampling frame must be available, otherwise develop a sampling frame.<br>2. Choose an appropriate sampling method to draw a sample from the sampling frame. |

**The Sampling Design Process**

Define the Population then Determine the Sampling Frame —> Select Sampling Technique(s) —> Determine the Sample Size —> Execute the Sampling Process

**Types of Sampling Methods**

| Probability Sampling | Non- probability Sampling |
|---|---|
| Simple Random Sampling (SRS) , Stratified Random Sampling , Systematic Random Sampling , Cluster Sampling,Multistage Sampling | Deliberate (Quota) Sampling , Convenience Sampling , Purposive Sampling , Snowball Sampling , Consecutive Sampling |

# L16 Sampling Techniques to select study subjects

## Simple Random Sampling

Numbering all subjects as a list and using random numbers to select required subjects from that list.

Equal probability

**Techniques :**

Lottery method

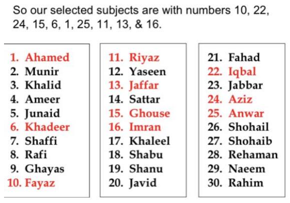
Table of random numbers

**Advantage**

Most representative group

**Disadvantage**

Difficult to identify every member of a population


So our selected subjects are with numbers 10, 22, 24, 15, 6, 1, 25, 11, 13, & 16.

1. Ahamed
2. Munir
3. Khalid
4. Ameer
5. Junaid
6. Khadeer
7. Shaffi
8. Rafi
9. Ghayas
10. Fayaz
11. Riyaz
12. Yaseen
13. Jaffar
14. Sattar
15. Ghouse
16. Imran
17. Khaleel
18. Shabu
19. Shanu
20. Javid
21. Fahad
22. Iqbal
23. Jabbar
24. Aziz
25. Anwar
26. Shohail
27. Shohaib
28. Rehaman
29. Naeem
30. Rahim

How to select a simple random sample

1. Define the population
2. Determine the desired sample size
3. List all members of the population or the potential subjects

**For example:**

4th grade boys who have demonstrated problem behaviors, lets select 10 boys from the list

Another example: Simple random sampling ( Estimate hemoglobin levels in patients with sickle cell anemia )
1. Determine sample size - 2. Obtain a list of all patients with sickle cell anemia in a hospital or clinic - 3. Patient is the sampling unit - 4. Use a table of random numbers to select units from the sampling frame - 5. Measure hemoglobin in all patients - 6. Estimate the levels (normal & abnormal) of hemoglobin

## Systematic random Sampling

**Technique: Select a random starting point and then select every $k^{th}$ subject in the population**

Use "system" to select sample (e.g., every 5th item in alphabetized list, every 10th name in phone book)

**Advantage**

- Quick, efficient, saves time and energy

**Disadvantage**

- Not entirely bias free; each item does not have equal chance to be selected
- System for selecting subjects may introduce systematic error
- Cannot generalize beyond population actually sampled

Example

If a systematic sample of 500 students were to be carried out in a university with an enrolled population of 10,000, the sampling interval would be:

I = N/n = 10,000/500 =20

All students would be assigned sequential numbers. The starting point would be chosen by selecting a random number between 1 and 20. If this number was 9, then the 9th student on the list of students would be selected along with every following 20th student. The sample of students would be those corresponding to student numbers 9, 29, 49, 69, ........ 9929, 9949, 9969 and 9989.

## Stratified Random Sampling

Divide the population into at least two different groups with common characteristic(s), then draw subjects randomly from each group (group is called strata or stratum)

**Technique**

- Divide population into various strata (ex. divide population by age/sex)(Stratification variable)
- Randomly sample within each strata
- Sample from each strata should be proportional

**Advantage**

Better in achieving representativeness on control variable

**Disadvantage**

- Difficult to pick appropriate strata
- Difficult to Identify every member in population

**Good sampling method to represent various segments of non-homogenous target population**

Sampling in Epidemiology

Stratified random sample

¨Assess dietary intake in adolescents

1. Define three age groups: 11-13, 14-16, 17-19
2. Stratify age groups by sex
3. Obtain list of children in this age range from schools
4. Randomly select children from each of the 6 strata until sample size is obtained
5. Measure dietary intake

## Cluster (Area) random sampling

Randomly select groups (cluster) – all members of groups are subjects

**Advantage**

- More practical, less costly

**Conclusions should be stated** in terms of cluster (sample unit – school)

**Sample size** is number of clusters

Example: Randomly selecting <u>multiple schools</u> then sampling <u>all the students</u> in them

Appropriate when

- you can't obtain a list of the members of the population
- have little knowledge of population characteristics
- Population is scattered over large geographic area

## Multistage random sampling

<u>Stage 1</u>

randomly sample clusters (schools)

<u>Stage 2</u>

randomly sample class rooms from the schools selected

<u>Stage 3</u>

**random sample of students from class rooms (if its cluster sampling, <u>all students</u> will be part of the sample)**

## Random Selection/Sample vs. Random Assignment/Allocation

| Random Selection = | Random Assignment = |
|---|---|
| Every member of the population has an equal chance of being selected for the sample.(Choosing which potential subjects will actually participate in the study) | Every member of the sample (however chosen) has an equal chance of being placed in the experimental group or the control group.(Random assignment allows for individual differences among test participants to be averaged out.)"Deciding which group or condition each subject will be part of" |

# L16 Sampling Techniques to select study subjects

## Non-probability Sampling

### Deliberate (Quota) Sampling

Similar to stratified random sampling

**Technique**

- Quotas set using some characteristic of the population thought to be relevant
- Subjects selected non-randomly to meet quotas (usu. convenience sampling)

**Disadvantage**

- selection bias
- Cannot set quotas for all characteristics important to study

### 2-Convenience Sampling

**Technique**

“Take them where you find them” (non-random sampling)

Dr: it is where the sample is taken from a group of people easy to contact or to reach like posting your questionnaire on social media. Its very unfavorable and reduces the power of a research paper.

Intact classes, volunteers, survey respondents (low return), a typical group, a typical person

**Disadvantage:**

- Selection bias

### 3- Purposive Sampling

- Purposive sampling (**criterion-based sampling**) ¨Establish criteria necessary for being included in study and find sample to meet criteria.
- Solution: Screening

Obtain a sample of larger population and then those subjects that are not members of the desired population are screened or filtered out.

EX: want to study smokers but can’t identify all smokers

### 4- Snowball Sampling

In snowball sampling, an initial group of respondents is selected.

- After being interviewed, these respondents are asked to identify others who belong to the target population of interest. **(ask the responders to identify other responders)**
- Subsequent respondents are selected based on the referrals

### 5- Consecutive sampling

- ¨Outcome of 1000 consecutive patients presenting to the emergency room with chest pain
- ¨Natural history of all 125 patients with HIV-associated TB during 5 year period
- Explicit efforts must be made to identify and **recruit ALL persons with the condition of interest**

## Choosing probability vs. non-probability sampling method

## Prof. Shaffi: Very important table

| Probability sampling | Evaluation Criteria | Non-probability sampling |
|---|---|---|
| Conclusive | Nature of research | Exploratory |
| Larger sampling errors | Relative magnitude sampling vs non-sampling error | Larger non-sampling error |
| High<br>[Heterogeneous] | Population variability | Low<br>[Homogeneous] |
| Favorable | Statistical Considerations<br>(all statistical analysis are based on the assumption that the sample is random - use proper random sample technique - ) | Unfavorable |
| High | Sophistication Needed | Low |
| Relatively Longer | Time | Relatively shorter |
| High | Budget Needed | Low |

# L16 Practical Session: How to apply Sampling Techniques?

| | **How to apply sampling techniques?** |
|---|---|
| **Q1** | **What do you mean by 'sample' and population? Explain with a simple example.** |
| **A1** | A small portion/group of subjects selected from a wider group of subjects is called a **sample**. This wider group is called **population**.<br>**Example:** In particular hospital 1000 deliveries occurred in particular year and out of these we take 100 deliveries for our research study. These 1000 deliveries is our population and these 100 deliveries is our sample. |
| **Q2** | **Why do you study only a sample of patients? Write down points only.** |
| **A2** | ● **To save money and time**<br>● To facilitate data collection that we use for research analysis particularly when the population being studied is larger.<br>● Sampling permits you to draw conclusions about complex situations.<br>● To obtain enough data to answer the research questions without having to query the entire population |
| **Q3** | **What do you call that sample where subjects are selected without any bias?** |
| **A3** | Random sample (favorable)<br>**Do not confuse random sample with randomization**<br>Random sample: process of choosing a sample randomly from a population.<br>Randomization: assigning patients randomly to groups that receive different treatments. |
| **Q4** | **What do you call that sample where subjects are selected as you wish?** |
| **A4** | Convenience sample (non-random sample) (prone to a lot of bias, selection bias) |
| **Q5** | **Give names of some of the random sampling techniques you know.** |
| **A5** | 1. Simple random sampling - **Cannot be performed without having a list**<br>2. Stratified random sampling<br>3. Systematic random sampling - sample is selected according to a random starting point but with **a fixed, periodic interval.**<br>4. Cluster random sampling - sample represents a **population not an individual**, ex. you will compare all individuals in a school to all individuals who go to another school.<br>5. Multistage random sampling |
| Study No. 1 | In a big hospital, every year 500 cases of MI (myocardial infarction) are reported. We want to study their physiological profile-their BP, cholesterol level, lipoprotein level, BMI, etc. Resources permit us to do investigations only for 50 cases. **How do you select a simple random sample (SRS)** of 50 cases out of these 500 cases? Explain the crude way as well as easy way to select this sample. |
| | We will write ID numbers of these 500 cases in 500 similar looking slips and roll them and put in a bowl and shuffle well and then take 50 slips one by one. The patients whose ID numbers are picked up is our sample. This method of sampling is called **simple random sampling.** This is the **crude way** and difficult to do.<br>**Easy way** is take 50 random numbers within 1 to 500 from the computer or form the random number tables and the patients whose ID numbers are selected, will be our sample. |
| Study No. 2 | A researcher wants to take a random sample of 100 cases from 1000 deliveries that occurred in maternity hospital in the last year. He has taken one random number out of 1 and 10 say, 5. Then he took a case having ID No.5. Then he took cases having ID numbers 15, 25, 35, 45……995 as his sample. What method of sampling the researcher adopted here? |
| | The research has adopted systematic sampling. Why? There is a fixed and periodic interval |
| Study No. 3 | Consider one more hospital where 1000 MI cases were reported last year. He wants to do a study one these cases. As these number of cases is large, he wants to take a sample of 100 cases. And also, as the physiological parameters of these cases would be different in overweight and less weight cases, the researcher wants to divide these 500 MI cases into two groups one with overweight/obese(that is BMI>=25) cases and the other less weight(BMI<25) cases and that both these groups to represent in his sample of 100 cases. Then he took a sample of 50 patients at random from each of these two groups. |
| | ➔ **What are these two groups called in sampling?** These groups are called **strata in sampling.**<br>➔ **What is the sampling method adopted here to select a sample of 100 cases?** This method of sampling is called **stratified sampling.**<br>➔ **Why did the researcher adopt this sampling method?** He adopted this sampling method because that both strata that is, overweight and of less weight cases to be represented in the sample |
| Study No. 4 | It was decided to estimate prevalence of diabetes in KSA. He had limited resources. So, he divided entire KSA into 5 regions as north, south, east, west and central. Then he made 10 contiguous geographical areas in each of these five regions. Then he selected one area at random from each region. He collected data from all the eligible individuals from each selected area and he found 5000 individuals from these five selected areas. Then he collected relevant data from all these individuals. |
| | ➔ **What type of sampling method he used here? Cluster sampling (key words: all the eligible individuals from each selected area, all individuals)**<br>➔ **Why did he adopt this method?** He used cluster sampling because he had **limited resources and he does not need sampling frame that is list of all the individuals of entire KSA**, which is difficult to get. It's enough if he has list of clusters, and he could collect the data from all individuals of the selected clusters only. In this way, he saves lot of resources by not traveling widely to take a simple randomness sample. |
| Study No. 5 | Health authorities asked an epidemiologist to find out the prevalence of anemia in high school children of standard VI to X in a district of an African country. There are 60 schools in this district. And each school has standard VI to X classes. He wanted to use multistage sampling method to estimate the prevalence of anemia in high school children of standard VI to X in that district. How he would have done multistage sampling method in this situation? |
| | First, he needs only the list of these 60 schools. In the first stage he can select 5 schools among 60 schools at random, and form each of the selected school, out of five standards VI to X, select two standards at random. This is second stage of selection.<br>So our sample consists of 200 students (5x2x20). This is his sample of students from whom he has to collect data to estimate the prevalence of anemia of high school children of that district.<br>1st stage - randomly select from list of 60 schools<br>2nd stage - randomly select from VI - X classes<br>3rd stage - randomly select from students list |

# L17 How many study subjects are required ? (Estimation of Sample size)

How to **calculate sample size?**

Most Important: sample size calculation is **an educated guess**
It is more appropriate for studies **involving hypothesis testing**
There is no magic involved; only statistical and mathematical logic and some algebra
Researchers need to know something about what they are measuring and how it varies in the population of interest

**SAMPLE SIZE:**

How many subjects are needed to assure a given probability of detecting a statistically significant effect of a given magnitude if one truly exists?

**POWER**
احتمالية وجود
significant
في عدد صغير

If a limited pool of subjects is available, what is the likelihood of finding a statistically significant effect of a given magnitude if one truly exists?

Before We Can Determine Sample Size We Need To Answer The Following:

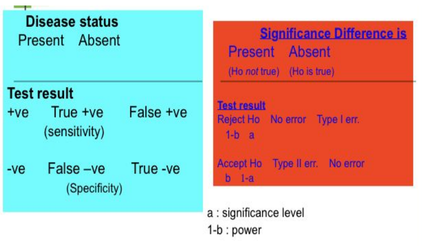
1. What is the primary objective of the study?
2. What is the main outcome measure?Is it a continuous or dichotomous outcome?
3. How will the data be analyzed to detect a group difference?
4. How small a difference is clinically important to detect?
5. How much variability is in our target population?
6. What is the desired a (alpha )and b(beta)?
7. What is the anticipated drop out and non-response % ?

Where do we get this knowledge?

Previous published studies
Pilot studies
If information is lacking, there is no good way to calculate the sample size

Type I error:

Rejecting H0 when H0 is true
■ α: The type I error rate.



| | Disease status | |
|---|---|---|
| | Present | Absent |
| Test result | | |
| +ve | True +ve<br>(sensitivity) | False +ve |
| -ve | False –ve<br>(Specificity) | True -ve |

| | Significance Difference is | |
|---|---|---|
| | Present<br>(Ho *not* true) | Absent<br>(Ho is true) |
| Test result | | |
| Reject Ho | No error<br>1-b | Type I err.<br>a |
| Accept Ho | Type II err.<br>b | No error<br>1-a |

a : significance level
1-b : power

Type II error:

Failing to reject H0 when H0 is false
■ β: The type II error rate
■ Power (1 - β): Probability of detecting group difference given the size of the effect (Δ) and the sample size of the trial (N)
**Accepting the null hypothesis when it is false**

Estimation of Sample Size by Three ways:

By using (1) Formulae (manual calculations) (2) Sample size tables or Nomogram (3) Softwares

## SAMPLE SIZE FOR ADEQUATE PRECISION

In a descriptive study,
- Summary statistics (mean, proportion)
- Reliability (or) precision
- By giving "confidence interval"
- Wider the C.I – sample statistic is not reliable and it may not give an accurate estimate of the true value of the population parameter

Sample size formulae for reporting precision
For single mean : $n = Z2\alpha\ S^2 / d^2$ where S=sd (s )
For a single proportion : $n = Z2\alpha P(1- P)/d2$
Where , $Z\alpha$ =1.96 for 95% confidence level
$Z\alpha$ = 2.58 for 99%

Problem 1 (Single mean)

A study is to be performed to determine a certain parameter(BMI) in a community. From a previous study a sd of 46 was obtained. If a sample error of up to 4 is to be accepted. How many subjects should be included in this study at 99% level of confidence?

**Answer**

$n = (Z\alpha/2)2\ \sigma 2 / d2$
σ: standard deviation = 46
d: the accuracy of estimate (how close to the true mean)= given sample error =4
Zα/2: A Normal deviate reflects the type I error. For 99% the critical value =2.58
$2.58^2 . 46^2 / 4^2 = 880.3$ ~~ 881

Problem 2 (Single proportion)

It was desired to estimate proportion of anemic children in a certain preparatory school. In a similar study at another school a proportion of 30 % was detected. **Compute the minimal sample size required** at a confidence limit of 95% and accepting a difference of up to 4% of the true population.

**Answer**

$n = (Z\alpha/2)2\ p(1-p) / d2$
p: proportion to be estimated = 30% (0.30)
d: the accuracy of estimate (how close to the true proportion) = 4% (0.04)
Zα/2: A Normal deviate reflects the type I error For 95% the critical value =1.96
$n= 1.96^2 . 0.3 (1-0.3 ) / 0.04^2 = 504.21$ ~~ 505

# L17 How many study subjects are required ? (Estimation of Sample size)

## SAMPLE SIZE FOR ADEQUATE power

Three bits of information required to determine the sample size
Type I & II errors
Clinical effect
Variation

Researcher fixes probabilities of type I and II errors
■ Prob (type I error) = Prob (reject H0 when H0 is true) = α
Smaller error ⇒ greater precision ⇒ need more information ⇒ need larger sample size
■ Prob (type II error) = Prob (don't reject H0 when H0 is false) = β
■ Power =1- β
More power ⇒ smaller error ⇒ need larger sample size

Quantities related to the research question (defined by the researche :

- α = Probability of rejecting H0 when H0 is true which's called significance level of the test
- β = Probability of not rejecting H0 when H0 is false,is called statistical power of the test
- Size of the measure of interest to be detected
- Difference between two or more means Difference between two or more proportions Odds ratio, Relative risk, Correlation, Regression coefficients Change in R2, etc
- The magnitude of these values depend on the research question and objective of the study (for example, clinical relevance)

**Clinical Effect Size**

What is a meaningful difference between the groups
It is truly an estimate and often the most challenging aspect of sample size p

- Large difference – small sample size
- Small differences – large sample size
- Cost/benefit

**All statistical tests are based on the following ratio:**

Test Statistic =
Difference between parameters / v / √n
As n ↑ v/√n ↓ Test statistic ↑ ( where's v=Variation)

**Sample size formulae for comparing two means**

n =2 S2 (Zα+ Zβ)2 /d2 where S=sd; d= difference
two proportions :
Zα= 1.96 for 95% confidence level Zα = 2.58 for 99% confidence level ;
Zβ= 0.842 for 80% power Zβ= 1.282 for 90% power

$$n=\frac{(Z_\alpha+Z_\beta)^2((p_1q_1)+(p_2q_2))}{(p_1-p_2)^2}, where\ \ q_1=(1-p_1), q_2=(1-p_2)$$

**Example 1:**

Does the consumption of large doses of vitamin A in tablet form prevent breast cancer?
Suppose we know from our tumor- registry data that incidence rate of breast cancer over a 1-year period for women aged 45 – 49 is 150 cases per 100,000. Women randomized to Vitamin A vs. placebo
Group 1: Control group given placebo pills.
Expected to have same disease rate as registry (150 cases per 100,000)
Group 2: Intervention group given vitamin A tablets.
Expected to have 20% reduction in risk (120 cases per 100,000)
Want to compare incidence of breast cancer over 1 year
Planned statistical analysis: Chi-square test to compare two proportions from independent samples . H0: p1 = p2 vs. HA: p1 p2

**Answer**

Test H0: p1 = p2 vs. HA p1 ≠ p2
• Assume 2-sided test with α=0.05 and 80% power
• p1 = 150 per 100,000 = .0015
• p2 = 120 per 100,000 = .0012 (20% rate reduction)
• Δ = p1 – p2 = .0003
• z1-α/2 = 1.96 z1-β = .84
• n per group = 234,882 ( **Too many to recruit in one year!)**

$$n=\frac{(Z_\alpha+Z_\beta)^2((p_1q_1)+(p_2q_2))}{(p_1-p_2)^2}, where\ \ q_1=(1-p_1), q_2=(1-p_2)$$

**Example 2:**

Does a special diet help to reduce cholesterol levels?
Suppose an investigator wishes to determine sample size to detect a 10 mg/dl difference in cholesterol level in a diet intervention group compared to a control (no diet) group

- Subjects with baseline total cholesterol of at least 300 mg/dl randomized

Group 1: A six week diet intervention - Group 2: No changes in diet
Investigator wants to compare total cholesterol at the end of the six week study
Planned statistical analysis: two sample t-test (for independent samples)(comparison of two means) H0:µ1 =µ2 vs. HA:µ1 ≠µ2

**Answer**

**Sample Size Formula**
To Compare Two Means From Independent Samples: H0: µ1 = µ2
1. α level
2. β level (1 – power)
3. Expected population difference (Δ= |µ1 - µ2|)
4. Expected population standard deviation (σ1 , σ2)
**Continuous Outcome**
(2 Independent Samples)
• Test H0: µ1 = µ2 vs. HA: µ1 ≠ µ2
• Two-sided alternative
• Assume outcome normally distributed with:

$$n_{per/group}=\frac{(2S^2)(z_\alpha+z_\beta)^2}{d^2}$$

S= standard deviation; d=difference between two means ; Zα= 1.96 for 95% confidence level; Zβ= 1.28 for 90% power
**Test H0: µ1=µ2 vs. HA : µ1≠µ2**
• Assume 2-sided test with α=0.05 and 90% power
• d = µ1 - µ2 = 10 mg/dl
• σ1= σ2 = (50 mg/dl)
• zα = 1.96 zβ = 1.28
• n per group = 525
• Suppose 10% loss to follow-up expected, adjust n = 525 / 0.9 = 584 per group

**Problem (comparison of two means)**

A study is to be done to determine effect of 2 drugs (A and B) on blood glucose levels . From previous studies using those
drugs, Sd of BGL of 8 and 12 g/dl were obtained respectively.

- A significant level of 95% and a power of 90% is required to detect a mean difference between the two groups of 3g/dl. How many subjects should be includein each group?

Answer:
Answer

```
      (SD1 + SD2)2
n = -------------------- * f(α,β)
      Δ2
```

$$n=\frac{(8^2+12^2)x10.5}{3^2}=242.6\sim 243$$

*in each group*

Objective: To observe whether feeding milk to 5 year old children enhances growth.
Groups: Extra milk diet - Normal milk diet
Outcome: Height ( in cms.)
Assumptions or specifications: Type-I error (α) =0.05
Type-II error (β) = 0.20 i.e., Power(1-β) = 0.80
Clinically significant difference (Δ) =0.5 cm.,
Measure of variation (SD.,) =2.0 cm., ( from literature or "Guesstimate")

Using the appropriate formula:

```
      2(SD)²
  N=  ---------- f(α,β)
         Δ²
```

2(2)² / (0.5)² *7.9 = 252.8 ( in each group)

# L17 Practical Session: How to calculate Sample Size?

**Study No. 1**

We want to estimate the mean hemoglobin of Saudi females. The standard deviation is around 5 grams/deciliter and we wish to estimate the true mean to within 2 grams/deciliter with 95% confidence. What is the required sample size?

1. Outcome variable = mean hemoglobin (continuous)
2. Type of study = descriptive

According to the outcome variable and study type we will use single mean formula

Findings: Zα = 1.96 for 95% confidence interval, S = 5, d = 2
n = Zα2 S2 / d2
n = 1.962 x 52 / 22 = 24.01 ~ 24
n = 24 + 20% non-response rate = 24 + 4.8 = 28.8 ~ 29

**Study No. 2**

**A researcher wanted to estimate average/mean number of cigarettes smoked per week by undergraduate students studying in a certain city. How many students are to be selected in to the sample such that the estimate of mean number of cigarettes smoked is to be within 2 of the true average with 95% confidence? (Based on a pilot study, it was found that the Sd. of number of cigarettes smoked is 30**

1. Outcome variable = mean number of cigarettes (continuous)
2. Type of study = descriptive

According to the outcome variable and study type we will use single mean formula

Findings: Zα = 1.96 for 95% confidence interval, S = 30, d = 2
n = Zα2 S2 / d2 n = 1.962 x 302 / 22 = 864.36 ~ 864
n= 864 + 20% non-response rate = 864 + 172.8 = 1036.8 ~ 1037

**Study No. 3**

We wish to estimate the proportion of Saudi males who smoke. What sample size do we require to achieve a 95% confidence interval of width ± 5% (that is to be within 5% of the true value)? A study some years ago found approximately 30% were smokers?

1. Outcome variable = proportion of Saudi males who smoke (categorical)
2. Type of study = descriptive

According to the outcome variable and study type we will use single proportion formula

Findings: Zα = 1.96 for 95% confidence interval, P = 0.3, d = 0.05
n = Zα2P(1-P) / d2 n = 1.962 x 0.3 x (1-0.3) / 0.052 = 322.6944 ~ 323
n = 323 + 20% non-response rate = 323 + 64.6 = 387.6 ~ 388

**Study No. 4**

An epidemiologist was asked to estimate the Knowledge level (%) towards Covid-19 in a particular community. How many subjects he should select, if the resulting estimate is to fall within 10% (width of confidence interval) of the true proportion with 95% confidence? What will happen to sample size if width of confidence interval is 5%. (As no literature is available researcher assumes that only 30% of subjects had good knowledge level)

1. Outcome variable = knowledge level (categorical)
2. Type of study = descriptive

According to the outcome variable and study type we will use single proportion formula

Findings: Zα = 1.96 for 95% confidence interval, P = 0.3, d = 0.1
n = Zα2P(1-P) / d2
n = 1.962 x 0.3 x (1-0.3) / 0.12 = 80.6736 ~ 81
n = 81 + 20% non-response rate = 81 + 16.2 = 97.2 ~ 97
What will happen to sample size if width of confidence interval is 5%?
Findings: Zα = 1.96 for 95% confidence interval, P = 0.3, d = 0.05
n = Zα2P(1-P) / d2
n = 1.962 x 0.3 x (1-0.3) / 0.052 = 322.6944 ~ 323
n = 323 + 20% non-response rate = 323 + 64.6 = 387.6 ~ 388

# L17 Practical Session: How to calculate Sample Size?

**Study No. 1**

An epidemiologist wants to test whether an iron supplement for pregnant women will increase their Hb level. One group of women will receive new supplement and the other group the usual supplement. From a pilot study the sd of Hb is 4 g/dl and is assumed to be same for both groups. what is the sample size required to test the hypothesis of no difference in mean Hb level at 99% level of confidence and 90% power of detecting an increase of 2 g/dl.

1. Outcome variable = hemoglobin level (continuous)
2. Type of study = analytical

According to the outcome variable and study type we will use two means formula

Findings: Zα = 2.58 for 99% confidence interval, Zβ = 1.282 for 99% power, S = 4, d = 2
n = 2S2 (Zα+Zβ)2 / d2, per arm
n = 2 x 42 x (2.58+1.282)2 / 22 = 119.320 ~ 119 n = 119 + 20% non-response rate = 119 + 23.8 = 142.8 ~ 143, per group
Total sample size = 143 x 2 = 286

**Study No. 2**

**Suppose it has been estimated that the rate of caries is 800 per 1000 school children in one district and 600 per 1000 in another district. What is the sample size required from each district to determine whether the difference is significant at the 95% level if we wish to have an 90% of chance of detecting the difference if it is real?**

1. Outcome variable = rate of caries (categorical)
2. Type of study = analytical

According to the outcome variable and study type we will use two proportions formula

Findings: Zα = 1.96 for 95% confidence interval, Zβ = 1.282 for 99% power, p1 = 800/1000 = 0.8, p = 600/1000 = 0.6, q1 = 1-0.8 = 0.2, q2 = 1-0.6 = 0.4, difference = p1 -p2= 0.8 - 0.6= 0.2

n = (Zα+Zβ)2((p1q1)+(p2q2)) / (p1-p2)2, per arm, where q1= (1-p1), q2 = (1-q2)
n = (1.96+1.282)2 x ((0.8x0.2) + (0.6x0.4)) / 0.22 = 105.106 ~ 105
n = 105 + 20% non-response rate = 105 + 21 = 126, per group
Total sample size = 126 x 2 = 252

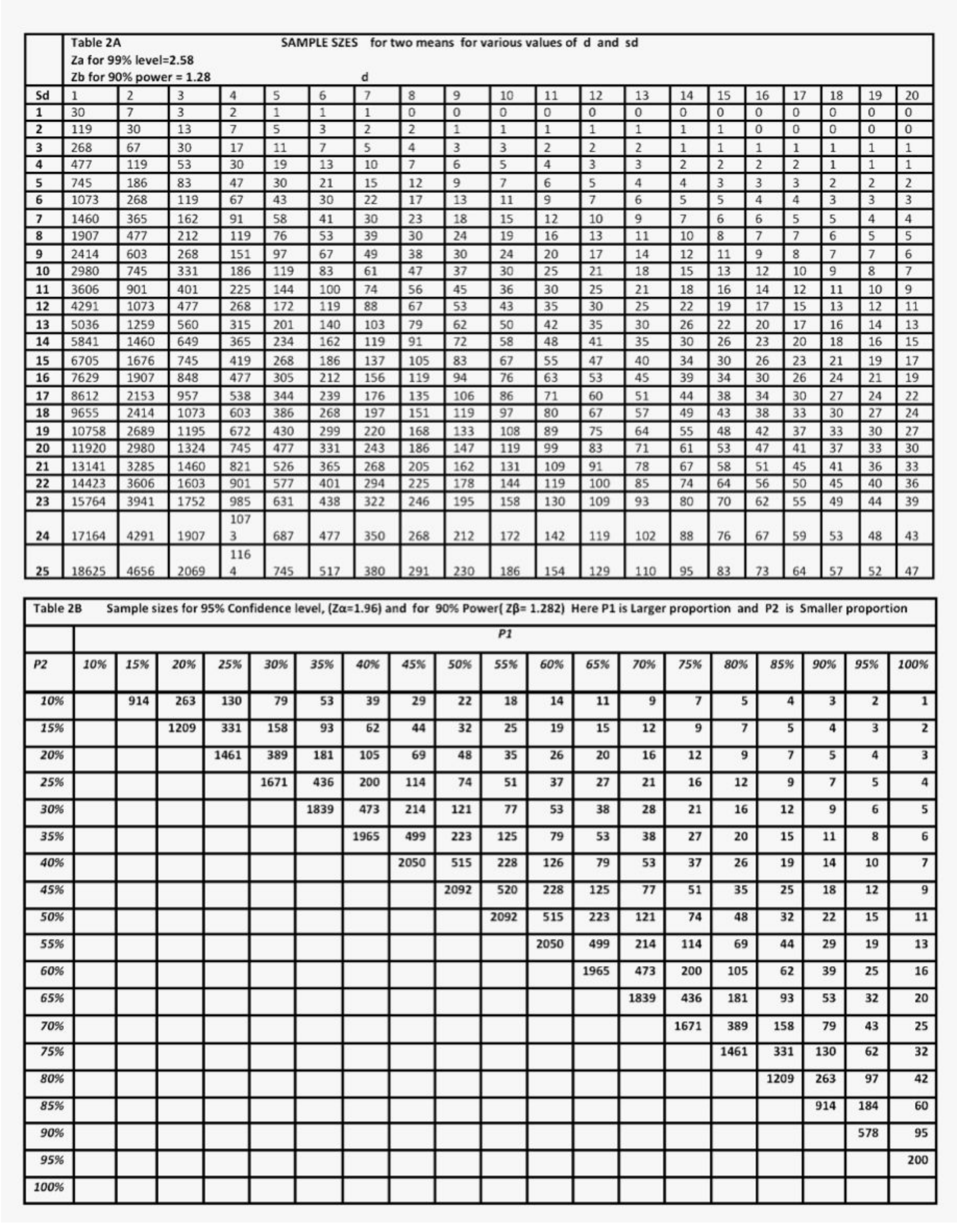


Table 2A SAMPLE SZES for two means for various values of d and sd
Za for 99% level=2.58
Zb for 90% power = 1.28
d

| Sd | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 30 | 7 | 3 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 119 | 30 | 13 | 7 | 5 | 3 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | 268 | 67 | 30 | 17 | 11 | 7 | 5 | 4 | 3 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 4 | 477 | 119 | 53 | 30 | 19 | 13 | 10 | 7 | 6 | 5 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |
| 5 | 745 | 186 | 83 | 47 | 30 | 21 | 15 | 12 | 9 | 7 | 6 | 5 | 4 | 4 | 3 | 3 | 3 | 2 | 2 | 2 |
| 6 | 1073 | 268 | 119 | 67 | 43 | 30 | 22 | 17 | 13 | 11 | 9 | 7 | 6 | 5 | 5 | 4 | 4 | 3 | 3 | 3 |
| 7 | 1460 | 365 | 162 | 91 | 58 | 41 | 30 | 23 | 18 | 15 | 12 | 10 | 9 | 7 | 6 | 6 | 5 | 5 | 4 | 4 |
| 8 | 1907 | 477 | 212 | 119 | 76 | 53 | 39 | 30 | 24 | 19 | 16 | 13 | 11 | 10 | 8 | 7 | 7 | 6 | 5 | 5 |
| 9 | 2414 | 603 | 268 | 151 | 97 | 67 | 49 | 38 | 30 | 24 | 20 | 17 | 14 | 12 | 11 | 9 | 8 | 7 | 7 | 6 |
| 10 | 2980 | 745 | 331 | 186 | 119 | 83 | 61 | 47 | 37 | 30 | 25 | 21 | 18 | 15 | 13 | 12 | 10 | 9 | 8 | 7 |
| 11 | 3606 | 901 | 401 | 225 | 144 | 100 | 74 | 56 | 45 | 36 | 30 | 25 | 21 | 18 | 16 | 14 | 12 | 11 | 10 | 9 |
| 12 | 4291 | 1073 | 477 | 268 | 172 | 119 | 88 | 67 | 53 | 43 | 35 | 30 | 25 | 22 | 19 | 17 | 15 | 13 | 12 | 11 |
| 13 | 5036 | 1259 | 560 | 315 | 201 | 140 | 103 | 79 | 62 | 50 | 42 | 35 | 30 | 26 | 22 | 20 | 17 | 16 | 14 | 13 |
| 14 | 5841 | 1460 | 649 | 365 | 234 | 162 | 119 | 91 | 72 | 58 | 48 | 41 | 35 | 30 | 26 | 23 | 20 | 18 | 16 | 15 |
| 15 | 6705 | 1676 | 745 | 419 | 268 | 186 | 137 | 105 | 83 | 67 | 55 | 47 | 40 | 34 | 30 | 26 | 23 | 21 | 19 | 17 |
| 16 | 7629 | 1907 | 848 | 477 | 305 | 212 | 156 | 119 | 94 | 76 | 63 | 53 | 45 | 39 | 34 | 30 | 26 | 24 | 21 | 19 |
| 17 | 8612 | 2153 | 957 | 538 | 344 | 239 | 176 | 135 | 106 | 86 | 71 | 60 | 51 | 44 | 38 | 34 | 30 | 27 | 24 | 22 |
| 18 | 9655 | 2414 | 1073 | 603 | 386 | 268 | 197 | 151 | 119 | 97 | 80 | 67 | 57 | 49 | 43 | 38 | 33 | 30 | 27 | 24 |
| 19 | 10758 | 2689 | 1195 | 672 | 430 | 299 | 220 | 168 | 133 | 108 | 89 | 75 | 64 | 55 | 48 | 42 | 37 | 33 | 30 | 27 |
| 20 | 11920 | 2980 | 1324 | 745 | 477 | 331 | 243 | 186 | 147 | 119 | 99 | 83 | 71 | 61 | 53 | 47 | 41 | 37 | 33 | 30 |
| 21 | 13141 | 3285 | 1460 | 821 | 526 | 365 | 268 | 205 | 162 | 131 | 109 | 91 | 78 | 67 | 58 | 51 | 45 | 41 | 36 | 33 |
| 22 | 14423 | 3606 | 1603 | 901 | 577 | 401 | 294 | 225 | 178 | 144 | 119 | 100 | 85 | 74 | 64 | 56 | 50 | 45 | 40 | 36 |
| 23 | 15764 | 3941 | 1752 | 985 | 631 | 438 | 322 | 246 | 195 | 158 | 130 | 109 | 93 | 80 | 70 | 62 | 55 | 49 | 44 | 39 |
| 24 | 17164 | 4291 | 1907 | 107<br>3 | 687 | 477 | 350 | 268 | 212 | 172 | 142 | 119 | 102 | 88 | 76 | 67 | 59 | 53 | 48 | 43 |
| 25 | 18625 | 4656 | 2069 | 116<br>4 | 745 | 517 | 380 | 291 | 230 | 186 | 154 | 129 | 110 | 95 | 83 | 73 | 64 | 57 | 52 | 47 |

Table 2B Sample sizes for 95% Confidence level, (Zα=1.96) and for 90% Power( Zβ= 1.282) Here P1 is Larger proportion and P2 is Smaller proportion

| | P1 | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P2 | 10% | 15% | 20% | 25% | 30% | 35% | 40% | 45% | 50% | 55% | 60% | 65% | 70% | 75% | 80% | 85% | 90% | 95% | 100% |
| 10% | | 914 | 263 | 130 | 79 | 53 | 39 | 29 | 22 | 18 | 14 | 11 | 9 | 7 | 5 | 4 | 3 | 2 | 1 |
| 15% | | | 1209 | 331 | 158 | 93 | 62 | 44 | 32 | 25 | 19 | 15 | 12 | 9 | 7 | 5 | 4 | 3 | 2 |
| 20% | | | | 1461 | 389 | 181 | 105 | 69 | 48 | 35 | 26 | 20 | 16 | 12 | 9 | 7 | 5 | 4 | 3 |
| 25% | | | | | 1671 | 436 | 200 | 114 | 74 | 51 | 37 | 27 | 21 | 16 | 12 | 9 | 7 | 5 | 4 |
| 30% | | | | | | 1839 | 473 | 214 | 121 | 77 | 53 | 38 | 28 | 21 | 16 | 12 | 9 | 6 | 5 |
| 35% | | | | | | | 1965 | 499 | 223 | 125 | 79 | 53 | 38 | 27 | 20 | 15 | 11 | 8 | 6 |
| 40% | | | | | | | | 2050 | 515 | 228 | 126 | 79 | 53 | 37 | 26 | 19 | 14 | 10 | 7 |
| 45% | | | | | | | | | 2092 | 520 | 228 | 125 | 77 | 51 | 35 | 25 | 18 | 12 | 9 |
| 50% | | | | | | | | | | 2092 | 515 | 223 | 121 | 74 | 48 | 32 | 22 | 15 | 11 |
| 55% | | | | | | | | | | | 2050 | 499 | 214 | 114 | 69 | 44 | 29 | 19 | 13 |
| 60% | | | | | | | | | | | | 1965 | 473 | 200 | 105 | 62 | 39 | 25 | 16 |
| 65% | | | | | | | | | | | | | 1839 | 436 | 181 | 93 | 53 | 32 | 20 |
| 70% | | | | | | | | | | | | | | 1671 | 389 | 158 | 79 | 43 | 25 |
| 75% | | | | | | | | | | | | | | | 1461 | 331 | 130 | 62 | 32 |
| 80% | | | | | | | | | | | | | | | | 1209 | 263 | 97 | 42 |
| 85% | | | | | | | | | | | | | | | | | 914 | 184 | 60 |
| 90% | | | | | | | | | | | | | | | | | | 578 | 95 |
| 95% | | | | | | | | | | | | | | | | | | | 200 |
| 100% | | | | | | | | | | | | | | | | | | | |

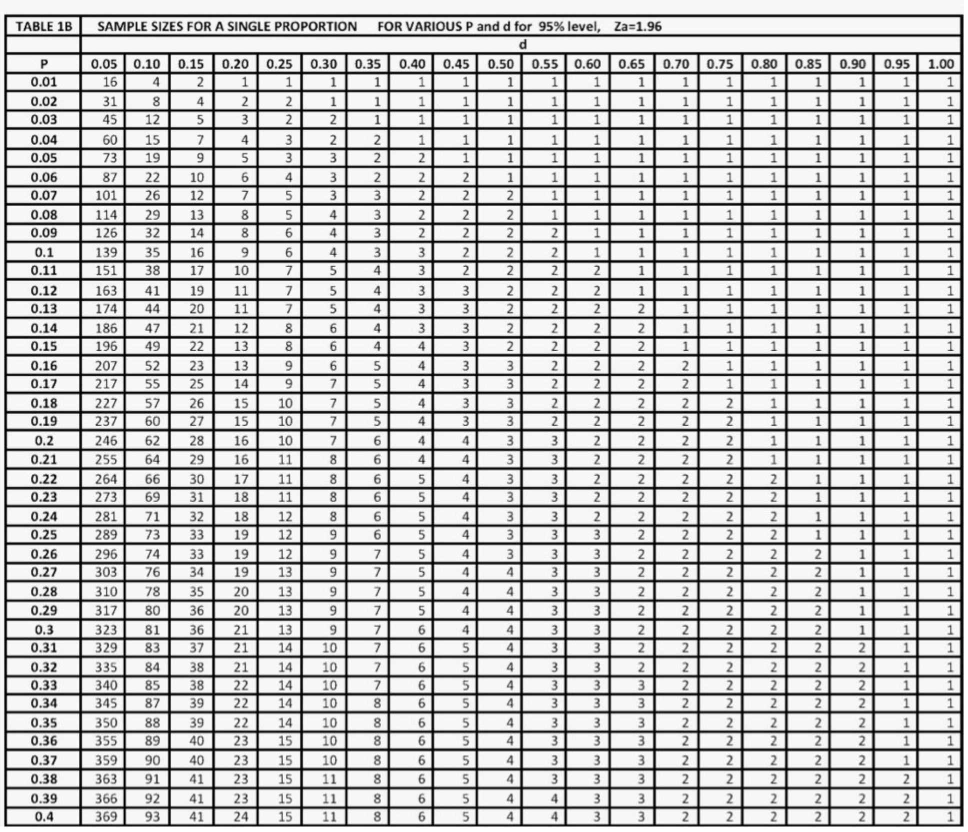


TABLE 1B SAMPLE SIZES FOR A SINGLE PROPORTION FOR VARIOUS P and d for 95% level, Za=1.96

| | d | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 | 0.55 | 0.60 | 0.65 | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 | 0.95 | 1.00 |
| 0.01 | 16 | 4 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.02 | 31 | 8 | 4 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.03 | 45 | 12 | 5 | 3 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.04 | 60 | 15 | 7 | 4 | 3 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.05 | 73 | 19 | 9 | 5 | 3 | 3 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.06 | 87 | 22 | 10 | 6 | 4 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.07 | 101 | 26 | 12 | 7 | 5 | 3 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.08 | 114 | 29 | 13 | 8 | 5 | 4 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.09 | 126 | 32 | 14 | 8 | 6 | 4 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.1 | 139 | 35 | 16 | 9 | 6 | 4 | 3 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.11 | 151 | 38 | 17 | 10 | 7 | 5 | 4 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.12 | 163 | 41 | 19 | 11 | 7 | 5 | 4 | 3 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.13 | 174 | 44 | 20 | 11 | 7 | 5 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.14 | 186 | 47 | 21 | 12 | 8 | 6 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.15 | 196 | 49 | 22 | 13 | 8 | 6 | 4 | 4 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.16 | 207 | 52 | 23 | 13 | 9 | 6 | 5 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.17 | 217 | 55 | 25 | 14 | 9 | 7 | 5 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.18 | 227 | 57 | 26 | 15 | 10 | 7 | 5 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| 0.19 | 237 | 60 | 27 | 15 | 10 | 7 | 5 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| 0.2 | 246 | 62 | 28 | 16 | 10 | 7 | 6 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| 0.21 | 255 | 64 | 29 | 16 | 11 | 8 | 6 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| 0.22 | 264 | 66 | 30 | 17 | 11 | 8 | 6 | 5 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 |
| 0.23 | 273 | 69 | 31 | 18 | 11 | 8 | 6 | 5 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 |
| 0.24 | 281 | 71 | 32 | 18 | 12 | 8 | 6 | 5 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 |
| 0.25 | 289 | 73 | 33 | 19 | 12 | 9 | 6 | 5 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 |
| 0.26 | 296 | 74 | 33 | 19 | 12 | 9 | 7 | 5 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |
| 0.27 | 303 | 76 | 34 | 19 | 13 | 9 | 7 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |
| 0.28 | 310 | 78 | 35 | 20 | 13 | 9 | 7 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |
| 0.29 | 317 | 80 | 36 | 20 | 13 | 9 | 7 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |
| 0.3 | 323 | 81 | 36 | 21 | 13 | 9 | 7 | 6 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |
| 0.31 | 329 | 83 | 37 | 21 | 14 | 10 | 7 | 6 | 5 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| 0.32 | 335 | 84 | 38 | 21 | 14 | 10 | 7 | 6 | 5 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| 0.33 | 340 | 85 | 38 | 22 | 14 | 10 | 7 | 6 | 5 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| 0.34 | 345 | 87 | 39 | 22 | 14 | 10 | 8 | 6 | 5 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| 0.35 | 350 | 88 | 39 | 22 | 14 | 10 | 8 | 6 | 5 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| 0.36 | 355 | 89 | 40 | 23 | 15 | 10 | 8 | 6 | 5 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| 0.37 | 359 | 90 | 40 | 23 | 15 | 10 | 8 | 6 | 5 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| 0.38 | 363 | 91 | 41 | 23 | 15 | 11 | 8 | 6 | 5 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| 0.39 | 366 | 92 | 41 | 23 | 15 | 11 | 8 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| 0.4 | 369 | 93 | 41 | 24 | 15 | 11 | 8 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |

You can also calculate the sample size for a single proportion by using this table. Depending on the variables you are given, choose a row (d) and a column (P) for a given a confidence interval of 95% and a Za of 1.96. e.g.,Question 3

# L21 Basic concepts and terminology in Biostatistics

## Statistics

- **Statistics** is the science of conducting studies to collect, organize, summarize, analyse, present, interpret and draw conclusions from data.
- Date: any value that have been collected. Singular: Datum (Set of values of one or more variables recorded on one or more observational units)
- What is Statistics? It's a sequence of collecting , Characterizing ,Presenting of data and interpreting results for Decision- Making
- statistics are used in many fields such as Public Health & Medicine Epidemiology, Pharmacology, Genetics and Business , Environment and government.
- Dataset: Data for a set of variables collection in group of persons.
- Data Table: A dataset organized into a table, with one column for each variable and one row for each person.

## Biostatistics

Biostatistics: is **the science that helps in managing medical uncertainties and variability of data** (methods used in dealing with statistics in the fields of medicine, biology and public health for planning, conducting and analyzing data which arise in investigations of these branches.)

Examples of biostatistics:

1. **Medical Statistics**: Deals with application of statistical methods to the study of diseases (risk factors, prognostic factors, etc..), efficacy of new treatments or vaccine, etc...
2. **Health Statistics** : Deals with application of statistical methods to varied information of public health importance.
3. **Vital Statistics** :Is the ongoing collection of government agencies of data relating to vital event such as births and deaths which are deemed reportable by local health authorities.

## Variables

**are unit of data collection whose value can vary and are defined into types according to the level of mathematical scaling which can be carried out on the data.**

## Variables Scales

### 1- Nominal scale variables(ex:measuring the "presence or absence" of a symptom?)

- A type of categorical data in which **objects fall into unordered categories.**
- Studies measuring nominal data must ensure that each category is mutually exclusive (**no overlap like Male / Female**) and the system of measurement needs to be exhaustive.
- dichotomies ( Binary data ) A type of categorical data that have only two responses i.e. Yes or no ( categorical data in which there are only two categories.But it can be more than two categories such as "blood groups" or Smoking status- smoker, non-smoker, past smoker)
- Nominal scale is **Least complex, simple measure of whether objects are the same or different.**

### 2- Ordinal scale variables

- Ordinal data is data that comprises(consists) of categories which can be ranked in ordered.
- Similarly with nominal data the distance between each category cannot be calculated but the categories can be ranked above or below each other. (Low stress - Moderate stress – Severe Stress)
- Same as **nominal but adds a measure of order** to what is being observed.
- **Nominal and ordinal data are used for Categorical data (Qualitative data)**
- **Example:cancer stages (I, II, III & IV)**

### 3- Interval scale variables

- Interval scale variables: Fahrenheit temperature scale: **(zero is arbitrary)** -40 degrees is not twice as hot as 20 degrees. You can't compare since no zero reference (there is negative).
- IQ tests: one who has 120 IQ is not twice as much as 60 IQ
- Question:Can we assume that attitudinal data represents real quantifiable measured categories?
- (i.e., That 'very happy' is twice as happy as plain 'happy' or that 'very unhappy' means no happiness at all). "Statisticians not in agreement on this". **NO! You can't quantify it; why ? negative values are included**.

builds on ordinal by adding more information on the range between each observation by allowing us to measure the distance between objects.

### 4- Ratio scale variables

- The distance between any two adjacent units of measurement (intervals) is the same and there is a meaningful zero point. (Includes absolute zero)
- Income: someone earning SAR20,000 earns twice as much as someone who earns SAR10,000.
- **Negative values are not included but we can be compare values (best scale for comparison)**

Ex: Height , Weight , Age , **BMI**

## Categories of data

### Primary data

- **What principal investigator collect** (new data) through observation, questionnaire, record form, interviews, survey

### Secondary data

- **Collected from other source** ,which data already is collected ex: census, medical record, registry and routinely kept records

## Clinimetrics

- Clinometric : A science called clinometries in which qualities are converted to meaningful quantities by using the scoring system. (Categorical data converted into quantitative data)

Examples:

- (1) Apgar score based on appearance, pulse, grimace , activity and respiration is used for neonatal prognosis.
- (2) Smoking index: no. of cigarettes, duration, filter or not, whether pipe, cigar etc.,
- (3) APACHE (Acute Physiology and Chronic Health Evaluation) score: to quantify the severity of condition of a patient

Why do we need to know what type of data? The data type influences the type of statistical analysis techniques

# L21 Basic concepts and terminology in Biostatistics

| | | |
|---|---|---|
| **Categorical (Qualitative) Data** | The objects being studied are grouped into categories based on some qualitative trait.<br>• The resulting data are merely **labels or categories.**<br>• Nominal and Ordinal scales will be used for categorical data or qualitative data. | |
| **Quantitative data** | The objects being studied are ‘measured’ based on some quantitative trait.<br>• The resulting data are set of numbers.<br>• Interval and Ratio scales will be used to measure quantitative data.<br>CONTINUOUS DATA can be converted into QUALITATIVE DATA<br>Ex :Wt. (In kg.) : Under wt, normal & over wt. Ht. (In cm.): Short, medium & tall | |
| **Types of quantitative data:** | 1- Discrete | 2- Continuous: |
| | Only certain values are possible (**There are gaps between the possible values)**. Implies counting. A whole number (Ex: Number of Children) | Continuous: Theoretically, with a fine enough measuring device. Implies measuring. (There is a **decimal**.) (Ex: Haemoglobin levels,2.5..)Age ( in years),Height( in cms.) , Weight (in Kgs.) , Sys.BP, Hb., Etc |

# Practical session: Scales of measurement and type of variables

## Question 1

**Q1) Name type of measurement scale for the following :**

| Measurement scale | Type |
|---|---|
| **Education status** (Literate / Illiterate) | Nominal scale |
| **Outcome of a newborn baby** (Boy / Girl) | Nominal scale |
| **Body mass Index** (weight(kg)/Height$^2$(m)) | Ratio scale |
| **Blood sugar level** (Quantitative variable) | Ratio scale |
| **Cholesterol level** (Quantitative variable) | Ratio scale |
| **Immunization status of the child** (Yes / No) | Nominal scale |
| **Grades of Exam Result** (A+,A,B+,B, etc.,) | Ordinal scale |

## Question 2

**Q2) A sample data of a study is given below. Name the type of variable:**

| Pt ID | Age (years) | sex | Marital status | Education | BMI | CD4 cell count | Viral load | ESR at 1 hour |
|---|---|---|---|---|---|---|---|---|
| 1 | 34 | 1 | 1 | 1 | 20.1 | 351 | 728000 | 35 |
| 2 | 30 | 2 | 1 | 1 | 17.8 | 33 | 11300 | 25 |
| 3 | 48 | 2 | 2 | 1 | 25.1 | 179 | 53900 | 30 |
| 4 | 40 | 1 | 1 | 2 | 17.6 | 235 | 498000 | 34 |
| 5 | 36 | 1 | 1 | 1 | 18.1 | 70 | 7360 | 19 |
| 6 | 25 | 1 | 2 | 2 | 17.3 | 86 | 400 | 10 |
| 7 | 29 | 2 | 1 | 1 | 16.9 | 228 | 750000 | 39 |
| 8 | 25 | 1 | 1 | 2 | 17.3 | 67 | 83400 | 22 |
| 9 | 38 | 2 | 2 | 1 | 22.5 | 27 | 14300 | 23 |
| 10 | 40 | 1 | 2 | 2 | 17.5 | 41 | 290000 | 30 |

Sex (1=Male, 2=Female); Marital status (1=single, 2=married); Education( 1=illiterate, 2=literate)

| Age (years) | Sex | Marital status | Education | BMI | CD4 cell count | Viral load | ESR at 1 hr |
|---|---|---|---|---|---|---|---|
| QNV (Discrete) | QLV (Nominal) | QLV (Nominal) | QLV (Nominal) | QNV (Continuous) | QNV (Discrete) | QNV (Discrete) | QNV (Discrete) |

**QLV** =Qualitative variable. **QNV**= Quantitative variable

## Question 3

**Q3)** Classify the following variables as: **- Quantitative** (discrete or continuous). **- Qualitative** (ordinal or nominal).

| Variable | Type |
|---|---|
| **White blood cells per deciliter of whole blood** | Quantitative variable (continuous) |
| **Leukemia rates in geographic regions** (cases per 100,000 people) | Quantitative variable (continuous[1]) |
| **Presence of type II diabetes mellitus** (yes or no) | Qualitative variable (nominal) |
| **Body weight** (kilograms) | Quantitative variable (continuous) |
| **Low-density lipoprotein level** (mg/dl) | Quantitative variable (continuous) |
| **Grade in course** coded : A, B, C, D or F | Qualitative variable (ordinal) |
| **Course credit** (pass or fail) | Qualitative variable (nominal) |

| Variable | Type |
|---|---|
| **Satisfaction**: 1=very satisfied, 2= satisfied, 3= neutral, 4= unsatisfied, 5= other | Qualitative variable (ordinal) |
| **Treatment group** : 1 = active treatment, 2 = placebo | Qualitative variable (nominal) |
| **The number of road accidents in KSA during Ramadan month.** | Quantitative variable (discrete) |
| **The number of boys in a family** | Quantitative variable (discrete) |
| **The length of time that a cancer patient and survives after diagnosis.** | Quantitative variable (either discrete or continuous[2]) |
| **The number of previous miscarriages an expectant mother had.** | Quantitative variable (discrete) |

1. It is continuous since there is "per 100,000" meaning it will be a fraction not a whole number. E.g: rate of leukemia in Riyadh is 5.5 per 100,000.
2. It depend on how data is presented. E.g: continuous: 1.5 months, discrete: 45 days.

# **L23** Description of Data (Using Summary & Variability measures)

**Data collection**

1- Data Presentation: Tabulation, Diagrams and Graphs
2- Descriptive statistics: measures of location, measures of dispersion, measures of skewness & kurtosis
3- Inferential Statistics: estimation, hypothesis testing, point estimate, interval estimate
4- Inferential statistics: univariate analysis, multivariate analysis

**Describing Data Numerically**

- central tendency: arithmetic mean, median, mode, geometric mean(used in labs like measure AB levels), harmonic mean(like when you measure velocity from Riyadh to Jeddah start by car then train then a walk)
- Quartiles
- Variation: range, interquartile range, variance and standard deviation
- Shape: skewness

**Measures of Central Tendency**

A statistical measure that identifies a single score as representative for an entire distribution. The goal of central tendency is to find the single score that is **most typical or most representative of the entire group** , There are 3 common measures of central tendency:
1- the mean. 2- the median. 3- the mode

**Mean**

It's called Arithmetic mean as well, The most common measure of central tendency , Affected by extreme values (outliers),

**Example**

Calculate the mean of the following data: 1 5 4 3 2
1-Sum the scores ( ΣX): 1 + 5 + 4 + 3 + 2 = 15
2-Divide the sum (ΣX) = 15) by the number of scores (N = 5): 15 / 5 = 3 3- Mean = x = 3

**Median**
alternative term used for "median"Is Q2

- The median is simply another name for the 50th percentile
- It is the score **in the middle or center**; half of the scores are larger than the median and half of the scores are smaller than the median
- Not affected by extreme values (unlike the mean)

In an ordered array, the median is the "**middle**" number
- If n or N is odd, the median is the middle number. - If n or N is even, the median is the average of the two middle numbers

**Example**

What is the median of the following scores: 24 18 19 42 16 12
**1**. Sort the scores from highest to lowest : 42 24 19 18 16 12
**2**. Determine **the middle score: middle** = (N + 1) / 2 = (6 + 1) / 2 = 3.5 **3**. **Median** = average of 3rd and 4th scores: (19 + 18) / 2 = 18.5

**Central tendency (النزعة المركزية)**

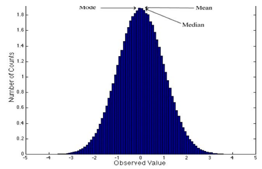
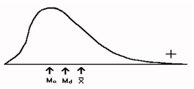
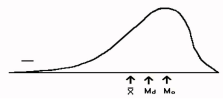
Mean is the most frequently used but is sensitive to extreme scores
● e .g. 1 2 3 4 5 6 7 8 9 10 Mean = 5.5 (median = 5.5)
● 1 2 3 4 5 6 7 8 9 20 Mean = 6.5 ( median = 5.5 )
● e.g. 1 2 3 4 5 6 7 8 9 100 Mean = 14.5 (median = 5.5)

**Mode**

Value that occurs most often , Not affected by extreme values (Ex: 20 student got 90 out of 100 )
- Used for either numerical or categorical(nominal)data
- There may be no mode or there may be several modes

**Shape of Distributions**

Distributions can be either symmetrical or skewed ( becoming narrow on one side ), depending on whether there are more frequencies at one end of the distribution than the other.

| Symmetrical distribution | Skewed distribution **Understand the diagrams well** |
|---|---|
| A distribution is symmetrical if the frequencies at the right & left tails of the distribution are identical, so that if it is divided into two halves, each will be the image of the other.<br>★ **In a symmetrical distribution the mean, median, and mode are identical.**<br><br> <br>Bell-Shaped (also known as symmetric" or "normal") | Few extreme values on one side of the distribution or on the other<br>**Positively** skewed distributions: distributions which have few extremely high values (**Mean>Median>mode** )<br>**- positively (skewed to the right)**,it tails off toward larger values<br><br> <br>- Negatively skewed distributions: distributions which have few extremely low values(**Mean<Median**)<br>**- negatively (skewed to the left)** – it tails off toward smaller values<br><br> |

**Choosing a Measure of Central tendency**
**Prof: it's important**

- IF variable is **Nominal**.. —> **Mode** , Ex: **number of people who smoke**
- IF variable is Ordinal... —> Mode or Median (or both)
- IF variable is Interval-Ratio and distribution is Symmetrical… —> Mode, Median or Mean
- IF variable is Interval-Ratio and distribution is Skewed… —> Mode or Median

● 7,8,9,10,11 n=5, x=45, X=45/5=9 S.D.=1.58, 1.58 —> Less variability
● 3,4,9,12,15 n=5, x=45, X=45/5=9 S.D.= =4.74, 4.74 —> There is variability
● 1,5,9,13,17 n=5, x=45, X=45/5=9 S.D.= =6.32, 6.32 —> High variability

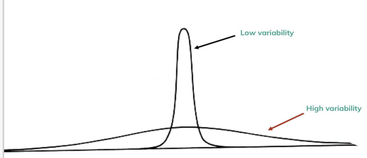
Variability will be low when: 1-Accurate measurements. 2-Good sample size

**Dispersion(التشتت) (Variability)**

Measures of dispersion summarize differences in the data, how the numbers differ from one another.

**Example**

⊙ Series I : 70 70 70 70 70 70 70 70 70 70. ( no variation / dispersion)
⊙ Series II : 66 67 68 69 70 70 71 72 73 74. (low variation / dispersion)
⊙ Series III :1 19 50 60 70 80 90 100 110 120. (high variation/ dispersion)

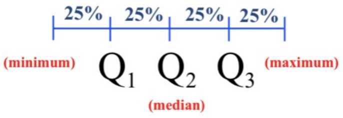


**Important figure**
single summary figure that describes the spread of observations within a distribution.

# L23 Description of Data (Using Summary & Variability measures)

### Quartiles

divides ranked scores 4 four equal parts



| Q1 = **equivalent to 25th**<br>(lower quartile) (first quartile) | Q2<br>(**median**) (middle quartile) (second quartile) | Q3 = **equivalent to 75th**<br>(upper quartile) (third quartile) |
|---|---|---|
| $Q_1 = \frac{n+1}{4}$ th | $Q_2 = \frac{2(n+1)}{4} = \frac{n+1}{2}$ th | $Q_3 = \frac{3(n+1)}{4}$ th |

### Example

Calculate the quartiles from this score (data): 6, 3, 1,7,4, 9, 4

1- rank the score (data): 1, 3, 4, 4, 6, 7, 9

2- n is the number of observation.. x1, x2 ...xn, in this case it equals 7

$$Q1 = \frac{(7)+(1)}{4} = 2 \qquad Q2 = \frac{[2][(7)+(1)]}{4} = 4 \qquad Q3 = \frac{[3][(7)+(1)]}{4} = 6$$

Q1 = (second observation) = 3  Q2 = (fourth observation) = 4  Q3 = (sixth observation) = 7

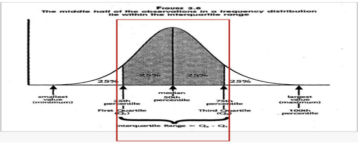
1, **3**, 4, **4**, 6, **7**, 9

### IQR (interquartile range) الانحراف الربيعي

The interquartile range is Q3-Q1, Range of the middle half of scores.

**50% of the observations in the distribution** are in the interquartile range.

The following figure shows the interaction between the quartiles, the median and the interquartile range



### Range

Difference between the smallest and largest observations

Ex: marks of student: 52, 76, **100**, 36, 86, 96, 20, **15**, 57, 64, 64, 80, 82, 83, 30, 31, 31, 31, 32, 37, 38, 38, 40, 40, 41, 42, 47, 48, 63, 63, 72, 79, 70, 71, 89

Range: 100-15 = 85, take the difference It doesn't consider the other data

### Percentile & Quartiles

- **Maximum is 100th percentile:** 100% of values lie at or below the maximum
- Median is 50th percentile: 50% of values lie at or below the median
- Any percentile can be calculated. But the most common are **25th (1st Quartile)** and **75th (3rd Quartile).**

Please Check the way of Locating Percentiles in a Frequency Distribution in our team

### Variance

Deviations of each observation from the mean, then averaging the sum of squares of these deviations.or we can say it's Mean of all squared deviations from the mean.

### Standard deviation

**" ROOT- MEANS-SQUARE-DEVIATIONS" best measurement to calculate variability when data is following normal distribution**

- To "undo" the squaring of difference scores, take the square root of the variance.
- Return to original units rather than squared units.
- **Measures the variation of a variable** in the sample.
- Technically:
- Rough measure of the **average amount by which observations deviate from The mean**

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

### Example

- Data: X =
- N =6
- Mean=
- Variance=
- **Standard Deviation=**

$$\bar{X} = \frac{\sum X}{N} = \frac{42}{6} = 7$$

$$s^2 = \frac{\sum(\bar{X} - X)^2}{N} = \frac{28}{6} = 4.67$$

$$s = \sqrt{s^2} = \sqrt{4.67} = 2.16$$

Interpretation: All 6 values on average are deviating by 2.16. On average each student is different from other by 2.16.

| X | $X - \bar{X}$ | $(X - \bar{X})^2$ |
|---|---|---|
| 6 | -1 | 1 |
| 10 | 3 | 9 |
| 5 | -2 | 4 |
| 4 | -3 | 9 |
| 9 | 2 | 4 |
| 8 | 1 | 1 |
| 42 | 0 | 28 |

### WHICH MEASURE TO USE ? Prof: it's important

- Distribution of data is Symmetric? use: **mean and Standard Deviation , ex: (monthly income of study subjects)**
- Distribution of data is skewed? use: Median and Quartiles

## Exploring data VERY IMPORTANT

### Graphical illustrations

1- Categorical data-**qualitative**-:

- **Bar chart**
- Clustered bar charts (two categorical variables)
- Pie charts

2- Continuous data-**quantitative**-:

- Histogram (can be plotted against a categorical variable)
- Box & Whisker plot (can be plotted against a categorical variable)
- **Stem and Leaf plot**
- **Scatter plot (2 continuous variables )(observe the relationship between two quantitative variables.)**

### Descriptive statistics

1- categorical data :

- **Frequency**
- Percentage (row, column or total)

2- Continuous data (Measure of location) :

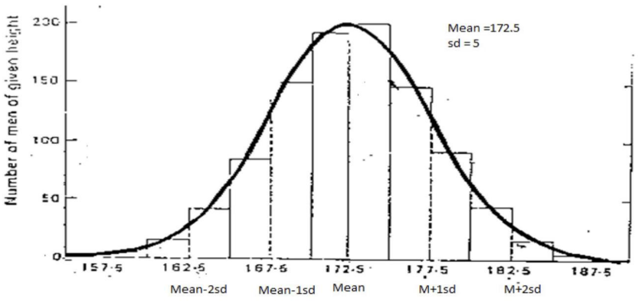
- Mean
- **Median**

3- Continuous data (Measure of variation) :

- Standard deviation
- Range (Min,Max)
- **Interquartile range (LQ, UQ)**

# L24 Practical Session: Normal distribution

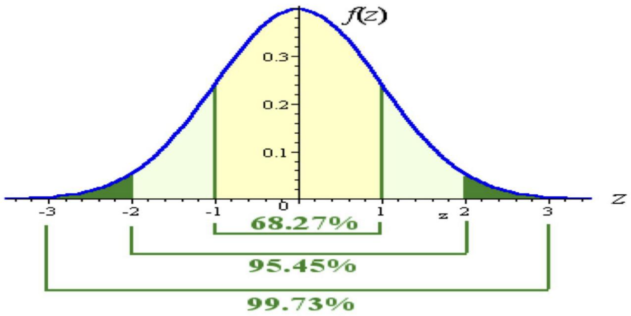
**Question 1**
using the NORMAL curve shown below, answer the following



A. The normal curve is a bell shaped curve.
B. The total area under the curve is equal to 1 (100%) .
C. 68% of the area lies between (mean-sd) and (mean+sd).
D. 95% of the area lies between (mean-2sd) and (mean+2sd).
E. 99% of the area lies between (mean-3sd) and (mean+3sd).
F. Normal distribution can be standardized in terms of a quantity called

Observation - Mean
**Z =. ------------------------------ „ what do you call this Z** : standard normal deviate Standard deviation
SD

**Question 2:**
standardized normal curve (mean 0 and variance 1) is shown



Looking at the graph, fill up the followi
A. what is the area lies between (-1 ≤ Z ≤ 1) ? 68.27%
B. what is the area lies between (-2 ≤ Z ≤ 2) ? 95.45%
C. what is the area lies between (-3 ≤ Z ≤ 3) ? 99.73%

**Question 3:**
To find the shaded area under normal curve from mean to z value 1.45
using z tables.

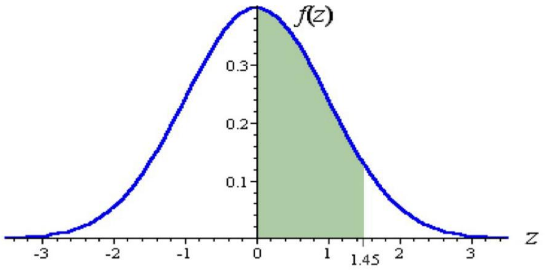
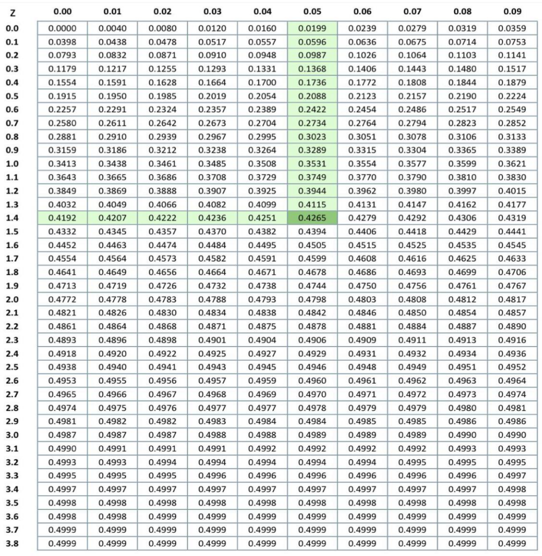


Table : Standard Normal Distribution – Area from 0 to Z value



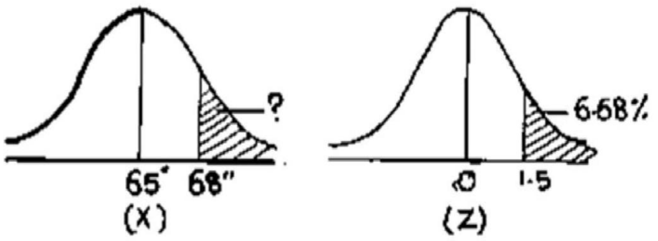
| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0000 | 0.0040 | 0.0080 | 0.0120 | 0.0160 | 0.0199 | 0.0239 | 0.0279 | 0.0319 | 0.0359 |
| 0.1 | 0.0398 | 0.0438 | 0.0478 | 0.0517 | 0.0557 | 0.0596 | 0.0636 | 0.0675 | 0.0714 | 0.0753 |
| 0.2 | 0.0793 | 0.0832 | 0.0871 | 0.0910 | 0.0948 | 0.0987 | 0.1026 | 0.1064 | 0.1103 | 0.1141 |
| 0.3 | 0.1179 | 0.1217 | 0.1255 | 0.1293 | 0.1331 | 0.1368 | 0.1406 | 0.1443 | 0.1480 | 0.1517 |
| 0.4 | 0.1554 | 0.1591 | 0.1628 | 0.1664 | 0.1700 | 0.1736 | 0.1772 | 0.1808 | 0.1844 | 0.1879 |
| 0.5 | 0.1915 | 0.1950 | 0.1985 | 0.2019 | 0.2054 | 0.2088 | 0.2123 | 0.2157 | 0.2190 | 0.2224 |
| 0.6 | 0.2257 | 0.2291 | 0.2324 | 0.2357 | 0.2389 | 0.2422 | 0.2454 | 0.2486 | 0.2517 | 0.2549 |
| 0.7 | 0.2580 | 0.2611 | 0.2642 | 0.2673 | 0.2704 | 0.2734 | 0.2764 | 0.2794 | 0.2823 | 0.2852 |
| 0.8 | 0.2881 | 0.2910 | 0.2939 | 0.2967 | 0.2995 | 0.3023 | 0.3051 | 0.3078 | 0.3106 | 0.3133 |
| 0.9 | 0.3159 | 0.3186 | 0.3212 | 0.3238 | 0.3264 | 0.3289 | 0.3315 | 0.3304 | 0.3365 | 0.3389 |
| 1.0 | 0.3413 | 0.3438 | 0.3461 | 0.3485 | 0.3508 | 0.3531 | 0.3554 | 0.3577 | 0.3599 | 0.3621 |
| 1.1 | 0.3643 | 0.3665 | 0.3686 | 0.3708 | 0.3729 | 0.3749 | 0.3770 | 0.3790 | 0.3810 | 0.3830 |
| 1.2 | 0.3849 | 0.3869 | 0.3888 | 0.3907 | 0.3925 | 0.3944 | 0.3962 | 0.3980 | 0.3997 | 0.4015 |
| 1.3 | 0.4032 | 0.4049 | 0.4066 | 0.4082 | 0.4099 | 0.4115 | 0.4131 | 0.4147 | 0.4162 | 0.4177 |
| 1.4 | 0.4192 | 0.4207 | 0.4222 | 0.4236 | 0.4251 | 0.4265 | 0.4279 | 0.4292 | 0.4306 | 0.4319 |
| 1.5 | 0.4332 | 0.4345 | 0.4357 | 0.4370 | 0.4382 | 0.4394 | 0.4406 | 0.4418 | 0.4429 | 0.4441 |
| 1.6 | 0.4452 | 0.4463 | 0.4474 | 0.4484 | 0.4495 | 0.4505 | 0.4515 | 0.4525 | 0.4535 | 0.4545 |
| 1.7 | 0.4554 | 0.4564 | 0.4573 | 0.4582 | 0.4591 | 0.4599 | 0.4608 | 0.4616 | 0.4625 | 0.4633 |
| 1.8 | 0.4641 | 0.4649 | 0.4656 | 0.4664 | 0.4671 | 0.4678 | 0.4686 | 0.4693 | 0.4699 | 0.4706 |
| 1.9 | 0.4713 | 0.4719 | 0.4726 | 0.4732 | 0.4738 | 0.4744 | 0.4750 | 0.4756 | 0.4761 | 0.4767 |
| 2.0 | 0.4772 | 0.4778 | 0.4783 | 0.4788 | 0.4793 | 0.4798 | 0.4803 | 0.4808 | 0.4812 | 0.4817 |
| 2.1 | 0.4821 | 0.4826 | 0.4830 | 0.4834 | 0.4838 | 0.4842 | 0.4846 | 0.4850 | 0.4854 | 0.4857 |
| 2.2 | 0.4861 | 0.4864 | 0.4868 | 0.4871 | 0.4875 | 0.4878 | 0.4881 | 0.4884 | 0.4887 | 0.4890 |
| 2.3 | 0.4893 | 0.4896 | 0.4898 | 0.4901 | 0.4904 | 0.4906 | 0.4909 | 0.4911 | 0.4913 | 0.4916 |
| 2.4 | 0.4918 | 0.4920 | 0.4922 | 0.4925 | 0.4927 | 0.4929 | 0.4931 | 0.4932 | 0.4934 | 0.4936 |
| 2.5 | 0.4938 | 0.4940 | 0.4941 | 0.4943 | 0.4945 | 0.4946 | 0.4948 | 0.4949 | 0.4951 | 0.4952 |
| 2.6 | 0.4953 | 0.4955 | 0.4956 | 0.4957 | 0.4959 | 0.4960 | 0.4961 | 0.4962 | 0.4963 | 0.4964 |
| 2.7 | 0.4965 | 0.4966 | 0.4967 | 0.4968 | 0.4969 | 0.4970 | 0.4971 | 0.4972 | 0.4973 | 0.4974 |
| 2.8 | 0.4974 | 0.4975 | 0.4976 | 0.4977 | 0.4977 | 0.4978 | 0.4979 | 0.4979 | 0.4980 | 0.4981 |
| 2.9 | 0.4981 | 0.4982 | 0.4982 | 0.4983 | 0.4984 | 0.4984 | 0.4985 | 0.4985 | 0.4986 | 0.4986 |
| 3.0 | 0.4987 | 0.4987 | 0.4987 | 0.4988 | 0.4988 | 0.4989 | 0.4989 | 0.4989 | 0.4990 | 0.4990 |
| 3.1 | 0.4990 | 0.4991 | 0.4991 | 0.4991 | 0.4992 | 0.4992 | 0.4992 | 0.4992 | 0.4993 | 0.4993 |
| 3.2 | 0.4993 | 0.4993 | 0.4994 | 0.4994 | 0.4994 | 0.4994 | 0.4994 | 0.4995 | 0.4995 | 0.4995 |
| 3.3 | 0.4995 | 0.4995 | 0.4995 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4997 |
| 3.4 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4998 |
| 3.5 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 |
| 3.6 | 0.4998 | 0.4998 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 |
| 3.7 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 |
| 3.8 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 |

# L24 Practical Session: Normal distribution

**Question 4**
If the distribution of heights of persons in a city has mean height 65”

a) Find the Proportion of persons whose height exceeds 68“

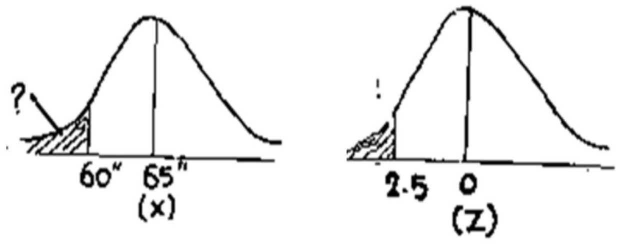
- **Answer:** Normal deviate = Z = (X-mean)/sd = (68-65)/2 = 1.5
  The z table gives areas from 0 to z. But, now we want the area from z to infinity. This gives us proportion of persons whose height exceeds 68’. We know the area from 0 to infinity is 0.5. so, if we subtract area of 0 to 1.5 from 0.5, we get the area from z to infinity.
  Area from 1.5 to infinity = (0 to infinity) – (0 to 1.5)
  = 0.5 – 0.4332
  = 0.0668 = 6.68% That is, there are nearly 7% of persons whose height exceeds 68"



Proportion = percentage , probability = number

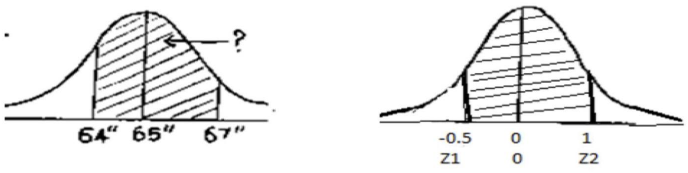
b) Find the proportion of persons whose height is less than 60"

- **Answer:** compute Normal deviate = Z = (X-mean)/sd = (60-65)/2 = -2.5
  We want the area from -∞  to –2.5 because we want the proportion of persons whose height is less than 60”. The z tables give areas from 0 to z .We know the area from 0 to infinity is 0.5. So, if we subtract value of 0 to 2.5 from 0.5, we get the area from z to infinity. Area from 2.5 to infinity = (0 to infinity) – (0 to 2.5) = 0.5 - 0.4938 = 0.0062 = 0.6% There are nearly 0.6% of persons whose height is <60”



c) Proportion of persons whose height is in between 64" & 67”

- **Answer:** First, find Normal deviate for 64” Z1 = (64 – 65)/2 = -0.5 Next, find Normal deviate for 67” Z2 = (67 – 65)/2 = 1 We want the area from -0.5 to 1 Z table gives area from 0 to z. We know area from -0.5 to 0 is same as area from 0 to 0.5 Hence, answer to the problem is to add the areas Area from 0 to 0.5 and from 0 to , Area from 0 to 0.5 = 0.1915 Area from 0 to 1 = 0.3413 Area from -0.5 to 1 = 0.5328  = 53.28%
  There will be 53% of persons whose height is in between 64" & 67”



# L24 Practical Session: Normal distribution

**Question 5**
suppose cholesterol level in a healthy population follows normal

a) What percentage of population is likely to have a level more than 210 mg/dl?

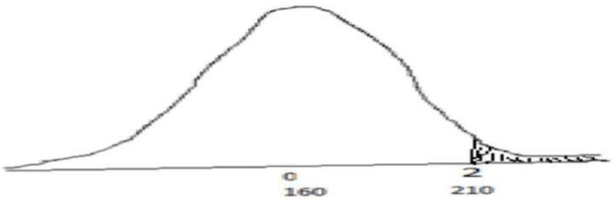
- **Answer:** first we draw rough normal curve showing which area we want to find. Next, we want the area from 210 to the end. So, we find the z value corresponding to 210 mg/dl and find the area from this z to the end.

  Z = (x-mean)/sd = (210-160)/25 = 50/25 = 2

  Area from 2 to end = (0.5) – (area from 0 to 2) “get it from the table”

  = 0.5 – 0.4772 = 0.0228 = 2.3%

  2.3% population is likely to have a level more than 210 mg/dl



b) What percentage of population is likely to have a level between 110 and 210 mg/dl?

- **Answer:** here also draw rough normal curve and shade the area from 110 and 210 mg/dl and then find the z values corresponding to 110 and 210 find the area from the z tables.

  Let z1 = (x-mean)/sd = (110-160)/25 = -50/25 = -2

  Let z2 = (x-mean)/sd = (210-160)/25 = 50/25 = 2

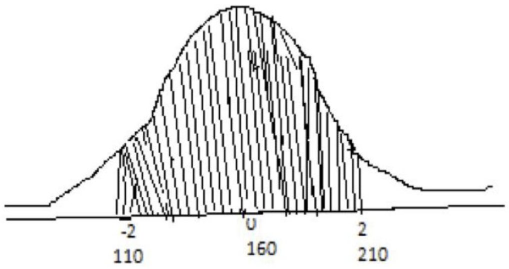
  Area from -2 to 2 = (area from -2 to 0) + (area from 0 to 2)

  Area from -2 to 0 = area 0 to 2 because of symmetry,

  Area from -2 to 2 = 0.4772+0.4772

  = 0.9554 = 95.54%

  95.5% of the population is likely to have a level between 110 and 210 mg/dl



c) What percentage of population is likely to have a level below 160 mg/dl?

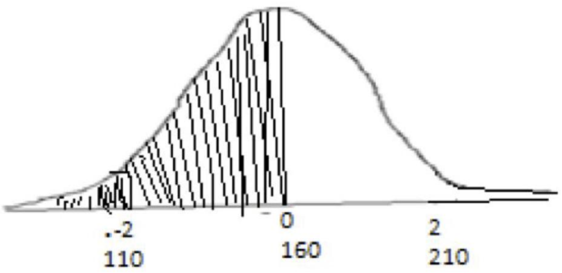
- **Answer:** here also first draw rough normal curve and shade the area up to 160 mg/dl and find z value corresponding to 160.

  Let z = (x-mean)/sd = (160-160)/25 = 0

  Area up to 0 = 0.5 = 50%

  50% of population is likely to have a level below 160 mg/dl.

  (NB: without calculation z itself, we can tell because 50% lie below mean value 160 mg/dl)



# **L25** Statistical significance using p-value

## Importance of inferential statistics

1-Using inferential statistics, we make **inferences about population** (taken to be unobservable) **based on a random sample** taken from the population of interest.
2-We can generate the parametre from the statistic

## Overview

| Parameter | Statistic | Is risk factor X associated with disease Y? |
|---|---|---|
| ● Numbers that summarize data for an entire **population**.<br>● E.g. Average height of all 25-year-old men (population) in KSA.<br>● Not always possible to measure because it needs the actual value in the population. | ● Numbers that summarize data from a **sample**.<br>● E.g. The height of the members of a sample of 100 such men are measured; the average of those 100 numbers is a statistic.<br>● Always possible to measure because it doesn't need the actual value in the population | ⊙ From the sample, we compute an estimate of the effect of X (risk factor) on Y (disease or outcome) (e.g. risk ratio if cohort study):<br>○ Is the effect real? Did chance play a role?<br>○ Why worry about chance?<br>→ Because of sampling variability...you only get to pick one sample!<br>When we take different samples it's going to give us different values because of the variation in each individual , how we can be sure it's real effect or just a chance? By testing the significance (p value and CI) |

| Significance testing | Interpreting the results |
|---|---|
| ⊙ The interest is generally in comparing two groups:<br>○ Significance testing can only be done if we have 2 comparison groups (it can't be applied to purely descriptive research)<br>○ (e.g., risk of outcome in the treatment and placebo group)<br>⊙ The statistical test depends **on the type of data and the study design.**<br>○ (eg. odds ratio in case-control or cross-sectional studies, and relative risk in RCTs and cohort studies) | ⊙ Make inferences from data collected using laws of probability and statistics, You have to use these two concepts:<br>○ Tests of significance (p-value).<br>○ Confidence intervals. |

## Hypothesis Testing

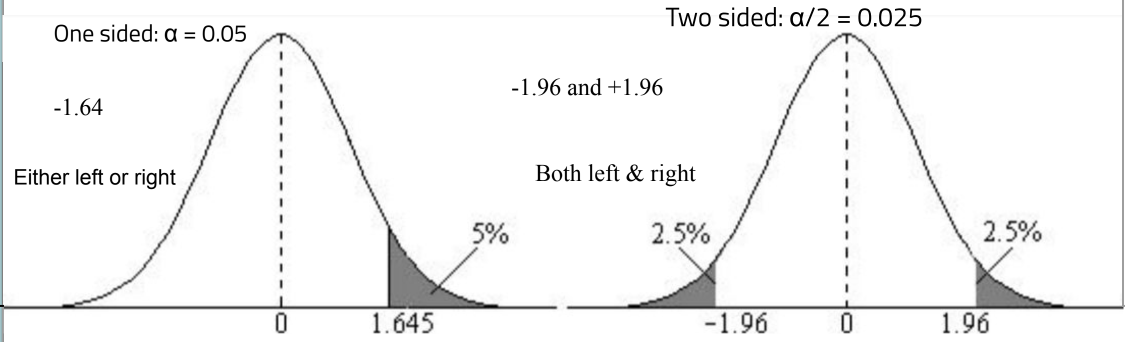
| Null hypothesis (Ho) | Alternative hypothesis (HA) |
|---|---|
| ● There is no association between the predictors (associated factors) and outcome variable in the population. | ● The proposition that **there is an association** between the predictors and outcome variable. |
| ● Assuming there is no association, statistical tests estimate the probability that the association is due to chance. | ● We do not test this directly but accept it by default if the statistical test rejects the null hypothesis. |
| ● States the assumption (numerical) to be tested. | ● The opposite of the null hypothesis, challenges the status quo.(means not = ) |
| ● Begin with the assumption that the null hypothesis is TRUE. | ● Is generally the hypothesis that is believed to be true by the researcher. |
| ● Always contains the '=' sign. | ● Never contains just the '=' sign. |

One and Two Sided Tests
Hypothesis tests can be one or two sided (tailed):

| One tailed tests are directional | Two tailed tests are not directional |
|---|---|
| ● H0: $\mu 1- \mu 2= 0$<br>● HA: $\mu 1- \mu 2 > 0$ or HA: $\mu 1- \mu 2 < 0$<br>⊙ One sided test:<br>○ A statistical hypothesis test in which alternative hypothesis has only one end. So, it will tell you if there is a relationship between variables in single direction.(37)<br>If you truly know the drug has one effect | ● Ho: $\mu 1- \mu 2= 0$<br>● HA: $\mu 1- \mu 2\neq 0$<br>⊙ Two sided test:<br>○ A statistical hypothesis test in which alternative hypothesis has two end. So, it will tell you if there is a relationship between variables in both direction.(37)<br>When you don't know whether the diet will decrease the weight or increase it |

## Hypothesis Testing= Rejection region=

○ Rejection region: set of all test statistic values for which H0 will be rejected.
○ Level of significance, α: Specified before an experiment to define rejection region.



Type-I and Type-II Errors

| | |
|---|---|
| ● Probability of rejecting H0 when H0 is true.<br>● Called significance level of the test.<br>α | ● Probability of not rejecting H0 when H0 is false.<br>● 1-β called statistical power of the test.<br>β |

## Diagnosis and statistical reasoning

| Significance Difference: | | | Disease status: | | |
|---|---|---|---|---|---|
| Test result | Present (Ho not true) | Absent (Ho is true) | Test result | Present (Ho not true) | Absent (Ho is true) |
| Reject Ho | No error (1-β) | Type I error (α) | +ve | **True +ve (Sensitivity)** | False +ve |
| Accept Ho | Type II error (β) | No error (1-α) | -ve | **False -ve** | True -ve (specificity) |

# **L25** Statistical significance using p-value

## Significance testing of example

| | | |
|---|---|---|
| Mortality<br>IV nitrate<br>PN | Subjects with acute MI<br><<br>? | Mortality<br>No nitrate<br>PC |

- ⊙ Suppose we do a clinical trial to answer the above question.
- ⊙ Even if IV nitrate has no effect on mortality, due to sampling variation, it is very unlikely that PN = PC
- ⊙ Any observed difference between groups may be due to treatment or a coincidence (or chance).

## Null Hypothesis (Ho)

- ⊙ There is no association between the independent and dependent/outcome variables.
- ○ Formal basis for hypothesis testing.
- ⊙ In the example, Ho: ”The administration of IV nitrate has no effect on mortality in MI patients” or PN- PC= 0

## Obtaining P values:

| Trial | Number dead (randomized) IV nitrate | (control) | Risk Ratio | 95% C.I. | P value |
|---|---|---|---|---|---|
| Chiche | 3/50 | 8/45 | 0.33 | (0.09, 1.13) | 0.08<br>How did we get this p-value? |
| Bussman | 4/31 | 12/29 | 0.24 | (0.08, 0.74) | 0.01 |
| Flaherty | 11/56 | 11/48 | 0.83 | (0.33, 2.12) | 0.70 |
| Jaffe | 4/57 | 2/57 | 2.04 | (0.39, 10.71) | 0.40 |
| Lis | 5/64 | 10/76 | 0.56 | (0.19, 1.65) | 0.29 |
| Jugdutt | 24/154 | 44/156 | 0.48 | (0.28, 0.82) | 0.007 |

From 437:
⊙ In the table, there are the 6 studies in the first column, sample size of iv nitrate patients and control in the second and third column. So in IV nitrate (in chiche study) 50 patients were randomized, yet 3 have died (people who died\ total) and we are interested to know how we got the p value and its interpretation?

## Example of significance testing

- ⊙ In the Chiche trial:
- ○ pN= 3/50 = 0.06; pC= 8/45 = 0.178
- ⊙ Null hypothesis:
- ○ H0: pN– pC= 0 or pN= pC
- ⊙ Statistical test:
- ○ Two-sample proportion

## Test statistic for Two Population Proportions

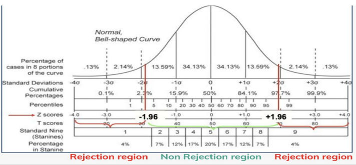
⊙ The test statistic for p1 – p2 is a Z statistic:

$$Z = \frac{(p_N - p_C) - (P_N - P_C)_o}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_N} + \frac{1}{n_C}\right)}}$$

Pc → Observed difference
(PN- Pc)o → Null hypothesis
nN → # of subjects in IV nitrate group
nC → # of subjects in control group

$$\bar{p} = \frac{X_N + X_C}{n_N + n_C}, \quad p_N = \frac{X_N}{n_N}, \quad p_C = \frac{X_C}{n_C}$$

## Testing significance at 0.05 level



Zα/2 = 1.96
Reject H0if Z < -Z α/2 or Z > Z α/2

## Significance testing

### Two Population Proportions

$$Z = \frac{(0.06 - 0.178)}{\sqrt{0.116(1-.116)\left(\frac{1}{50} + \frac{1}{45}\right)}} = -1.79$$

where: $\bar{p} = \frac{3+8}{45+50} = 0.116, \quad p_N = \frac{3}{45} = 0.06, \quad p_C = \frac{8}{50} = 0.178$

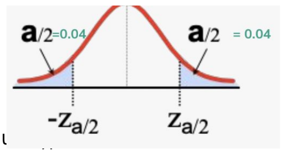
### Statistical test for p1– p2

⊙ Two Population Proportions, Independent Samples:

$$Z = \frac{(0.06 - 0.178)}{\sqrt{0.116(1-.116)\left(\frac{1}{50} + \frac{1}{45}\right)}} = -1.79$$

Two-tail test:
H0: pN– pC= 0
H1: pN– pC≠ 0

⊙ Zα/2= 1.96
→ Reject H0 if Z < -Zα/2 or Z > Zα/2
→ Since -1.79 is > than -1.96, we fail to reject the nu
→ The actual p-value = P (Z<-1.79) + P (Z>1.79)= 0.08



## P-value

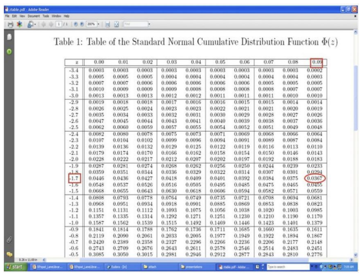

ztable.pdf - Adobe Reader
File Edit View Document Tools Window Help
start

Table 1: Table of the Standard Normal Cumulative Distribution Function Φ(z)

| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| -3.4 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0002 |
| -3.3 | 0.0005 | 0.0005 | 0.0005 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0003 |
| -3.2 | 0.0007 | 0.0007 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0005 | 0.0005 | 0.0005 |
| -3.1 | 0.0010 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0007 | 0.0007 |
| -3.0 | 0.0013 | 0.0013 | 0.0013 | 0.0012 | 0.0012 | 0.0011 | 0.0011 | 0.0011 | 0.0010 | 0.0010 |
| -2.9 | 0.0019 | 0.0018 | 0.0018 | 0.0017 | 0.0016 | 0.0016 | 0.0015 | 0.0015 | 0.0014 | 0.0014 |
| -2.8 | 0.0026 | 0.0025 | 0.0024 | 0.0023 | 0.0023 | 0.0022 | 0.0021 | 0.0021 | 0.0020 | 0.0019 |
| -2.7 | 0.0035 | 0.0034 | 0.0033 | 0.0032 | 0.0031 | 0.0030 | 0.0029 | 0.0028 | 0.0027 | 0.0026 |
| -2.6 | 0.0047 | 0.0045 | 0.0044 | 0.0043 | 0.0041 | 0.0040 | 0.0039 | 0.0038 | 0.0037 | 0.0036 |
| -2.5 | 0.0062 | 0.0060 | 0.0059 | 0.0057 | 0.0055 | 0.0054 | 0.0052 | 0.0051 | 0.0049 | 0.0048 |
| -2.4 | 0.0082 | 0.0080 | 0.0078 | 0.0075 | 0.0073 | 0.0071 | 0.0069 | 0.0068 | 0.0066 | 0.0064 |
| -2.3 | 0.0107 | 0.0104 | 0.0102 | 0.0099 | 0.0096 | 0.0094 | 0.0091 | 0.0089 | 0.0087 | 0.0084 |
| -2.2 | 0.0139 | 0.0136 | 0.0132 | 0.0129 | 0.0125 | 0.0122 | 0.0119 | 0.0116 | 0.0113 | 0.0110 |
| -2.1 | 0.0179 | 0.0174 | 0.0170 | 0.0166 | 0.0162 | 0.0158 | 0.0154 | 0.0150 | 0.0146 | 0.0143 |
| -2.0 | 0.0228 | 0.0222 | 0.0217 | 0.0212 | 0.0207 | 0.0202 | 0.0197 | 0.0192 | 0.0188 | 0.0183 |
| -1.9 | 0.0287 | 0.0281 | 0.0274 | 0.0268 | 0.0262 | 0.0256 | 0.0250 | 0.0244 | 0.0239 | 0.0233 |
| -1.8 | 0.0359 | 0.0351 | 0.0344 | 0.0336 | 0.0329 | 0.0322 | 0.0314 | 0.0307 | 0.0301 | 0.0294 |
| -1.7 | 0.0446 | 0.0436 | 0.0427 | 0.0418 | 0.0409 | 0.0401 | 0.0392 | 0.0384 | 0.0375 | 0.0367 |
| -1.6 | 0.0548 | 0.0537 | 0.0526 | 0.0516 | 0.0505 | 0.0495 | 0.0485 | 0.0475 | 0.0465 | 0.0455 |
| -1.5 | 0.0668 | 0.0655 | 0.0643 | 0.0630 | 0.0618 | 0.0606 | 0.0594 | 0.0582 | 0.0571 | 0.0559 |
| -1.4 | 0.0808 | 0.0793 | 0.0778 | 0.0764 | 0.0749 | 0.0735 | 0.0721 | 0.0708 | 0.0694 | 0.0681 |
| -1.3 | 0.0968 | 0.0951 | 0.0934 | 0.0918 | 0.0901 | 0.0885 | 0.0869 | 0.0853 | 0.0838 | 0.0823 |
| -1.2 | 0.1151 | 0.1131 | 0.1112 | 0.1093 | 0.1075 | 0.1056 | 0.1038 | 0.1020 | 0.1003 | 0.0985 |
| -1.1 | 0.1357 | 0.1335 | 0.1314 | 0.1292 | 0.1271 | 0.1251 | 0.1230 | 0.1210 | 0.1190 | 0.1170 |
| -1.0 | 0.1587 | 0.1562 | 0.1539 | 0.1515 | 0.1492 | 0.1469 | 0.1446 | 0.1423 | 0.1401 | 0.1379 |
| -0.9 | 0.1841 | 0.1814 | 0.1788 | 0.1762 | 0.1736 | 0.1711 | 0.1685 | 0.1660 | 0.1635 | 0.1611 |
| -0.8 | 0.2119 | 0.2090 | 0.2061 | 0.2033 | 0.2005 | 0.1977 | 0.1949 | 0.1922 | 0.1894 | 0.1867 |
| -0.7 | 0.2420 | 0.2389 | 0.2358 | 0.2327 | 0.2296 | 0.2266 | 0.2236 | 0.2206 | 0.2177 | 0.2148 |
| -0.6 | 0.2743 | 0.2709 | 0.2676 | 0.2643 | 0.2611 | 0.2578 | 0.2546 | 0.2514 | 0.2483 | 0.2451 |
| -0.5 | 0.3085 | 0.3050 | 0.3015 | 0.2981 | 0.2946 | 0.2912 | 0.2877 | 0.2843 | 0.2810 | 0.2776 |

- ● After calculating a **test statistic we convert this to a p-value by comparing its value to distribution of test statistic under the null hypothesis.**
- ● Measure of how likely the test statistic value is under the null hypothesis:
- ● **p-value ≤ α** ⇒ Reject H0 at level α (significant)
- ● p-value > α ⇒ Do not reject H0 at level α

### What is a P-value?

- ‘p’ stands for **probability**.
- Tail area probability based on the observed effect.
- Calculated as the probability of an effect as large as or larger than the observed effect (more extreme in the tails of the distribution), assuming null hypothesis is true.

⊙ Size of the P-value is related to:
The sample size.
The effect size or the observed association or difference.

⊙ **Measures the strength of the evidence against the null hypothesis**

○ Smaller p- values indicate stronger evidence against the null hypothesis

P values give no indication about the clinical importance of the observed association.

⊙ A very large study may result in very small p-value based on a small difference of effect that may not be important when translated into clinical practice.

⊙ Therefore, **important to look at the effect size and confidence intervals.**

## Stating the Conclusions of our Results

### p-value is small

⊙ we reject the null hypothesis or, equivalently, we accept the alternative hypothesis.
⊙ “Small” is defined as a p-value ≤ α,
where α = acceptable false (+) rate (usually 0.05).

### p-value is not small

⊙ we conclude that we cannot reject the null hypothesis or, equivalently, there is not enough evidence to reject the null hypothesis.
⊙ “Not small” is defined as a p-value> α, where α = acceptable false (+) rate (usually 0.05).

**Size of the p-value is related to the sample size ( larger = significant p value )**
**Size of the p-value is related to the effect size or the observed association or difference**

**Example:**

(1)
If a new antihypertensive therapy reduced the SBP by 1 mmHg as compared to standard therapy we are not interested in swapping to the new therapy.

(2)
However, if the decrease was as large as 10 mmHg, then you would be interested in the new therapy.

(3)
Thus, it is important to not only consider whether the difference is statistically significant by the possible magnitude of the difference should also be considered.

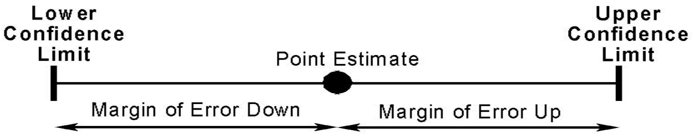
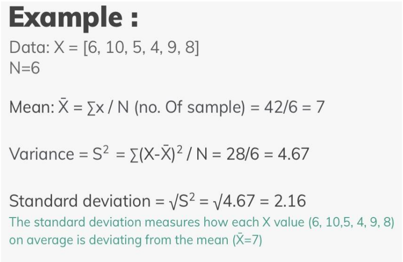
# **L25** Statistical significance using p-value

| Statistically significant Vs not statistically significant | Statistically significant | Not statistically significant |
|---|---|---|
| | ● Reject Ho | ● Do not reject Ho |
| | ● Sample value not compatible with Ho. | ● Sample value compatible with Ho. |
| | ● Sampling variation is an unlikely explanation of discrepancy between Ho and sample value.(يعنى significant not caused by chance or variation) | ● Sampling variation is a likely explanation of discrepancy between Ho and sample value. |

| Clinical importance vs. statistical significance | **Statistically significant & clinically important.** | **Not statistically significant BUT clinically important.** | **Statistically significant BUT NOT clinically important.** |
|---|---|---|---|
| | ○ This is where there is an important, meaningful difference between the groups and the statistics support this. | ○ This is most likely to occur if your study is underpowered and you do not have a large enough sample size to detect a difference between groups. | ○ If you have enough participants, even the smallest differences can become statistically significant.<br>○ just because a treatment is statistically significantly better than an alternative treatment, does not necessarily mean that these differences are clinically important. |

| **Reaction of investigator to results of a statistical significance test** | | | |
|---|---|---|---|
| | | Statistical Significance | |
| | | Not significant | significant |
| Practical importance of observed effect | Not important | 0 | Annoyed |
| | important | very sad | Elated |

# L26 Statistical Significance of Data II (95% CI)

| | |
|---|---|
| **Two forms of estimation** | Point estimation = single value,e.g. (mean, propration,RR,OR,etc)<br>Interval estimation = range of value, e.g Confidence interval |
| **confidence interval= Estimation** | -A range of values so defined that there is a specified probability that the value of a parameter lies within it.<br>-Components of CI:<br> Lower Confidence Limit<br>Point Estimate<br>Upper Confidence Limit<br>Margin of Error Down<br>Margin of Error Up<br>-Relying on information from a sample will always lead to some level of uncertainty, confidence interval is a range of values that tries to **quantify this uncertainty**<br>For example , 95% CI means that under repeated sampling 95% of CIs would contain the true population parameter<br>- Suppose α =0.05, we cannot say: "with probability 0.95 the parameter μ lies in the confidence interval."<br>- We only know that by repetition,  95% of the intervals will contain the true population parameter (μ)<br>- We are 95% sure that the TRUE parameter value is in the 95% confidence interval" |
| **Statistical inference is based on sampling variability** | -Sample statistics: We summarize a sample into one number e.g. a mean.<br>-Sample variability:If we could repeat an experiment many, many times with different samples on the same number of subjects, the resultant sample statistic would not always be the same (because of chance).<br>Standard error: A measure of the sampling variability.Don't get confused with the terms of standard deviation and standard error<br>-What is the difference?<br>**Standard Error means how much is the Variability among different samples**. while **standard deviation is how much the the values in one sample deviating on average from mean** (Variability in one sample).<br>-Smaller Standard Error/Deviation indicates a good precision and vice viscera. |
| **How to calculate CI** | P = point of estimate, a value drawn form sample (a statistic). أي قيمة إحصائية mean,proportion,difference of tow means or proportions,OR,RR etc...<br>**Zα** = standard Normal deviate for α,if α= 0.05, Zα= 1.96 (~95% CI).ثابت<br>σ or S= standard deviation<br>n=number of sample (sample size)<br>SE or $S\bar{X}$ of the mean(sem)=$\sigma/\sqrt{n}$ calculate the standard error from the standard deviation<br>Standard error of the mean(sem) should be always smaller than standard deviation(S)<br><br>**Example :**<br>Data: X = [6, 10, 5, 4, 9, 8]<br>N=6<br>Mean: $\bar{X} = \sum x / N$ (no. Of sample) = 42/6 = 7<br>Variance = $S^2 = \sum(X-\bar{X})^2 / N$ = 28/6 = 4.67<br>Standard deviation = $\sqrt{S^2} = \sqrt{4.67}$ = 2.16<br>The standard deviation measures how each X value (6, 10,5, 4, 9, 8) on average is deviating from the mean ($\bar{X}$=7)<br>**General formula: CI = P +/- $Z_\alpha$ x SE.**<br>(Sample statistic) +/- [(confidence level)  (standard error ) |
| **Most commonly used CI:** | 1-CI 90% corresponds to α 0.10<br>2-CI 95% corresponds to α 0.05<br>3-CI 99% corresponds to α 0.01<br><br>90% confidence interval: NARROWER than 95% ( $\bar{X}$ ± 1.65sem )<br>99% confidence interval:WIDER than 95% ( $\bar{X}$ ± 2.58sem ) |
| **CHARACTERISTICS OF CI'S** | -CI is for both descriptive and analytical studies studies while the P value is only for analytical(Comparison    group ).<br>-**The width of C.I. depends on**:Increasing the sample size will increase precision and narrow C.I<br>○ Sample size By reducing sample size 1-Bell shape converted to skewed 2-increase variability 3- widen C.I and included 0 (larger Sample size narrow CI and smaller variability and smaller P value that all indicate significance)<br>○ Variability smaller variability the narrow CI and the more precision<br>○ Degree of confidence 90% narrower than 95% which gives more precision<br>-The (im) precision of the estimate is indicated by the width of the confidence interval.<br>-The wider the interval the less precision, The narrower interval more precision. |
| Properties of Standard Error (SE) | SE increases with smaller sample size , For any confidence level, large samples reduce the margin of error<br>2. SE increases with larger standard Deviation<br>3. SE increases with larger z values |
| Statistic and Parameter | An observed value drawn from the sample is called a statistic<br>The corresponding value in population is called a parameter<br>We measure, analyze, etc statistics and translate them as parameters |

# **L26** Statistical Significance of Data II (95% CI)

## Interpretation

The result considered statically significant:

1-When P value less than or equal to alpha(α) but because we mostly use 95% C.I ( α=0.05 )

2- In descriptive study **CI for a difference between two means: Does the interval include 0 (zero)?** we take the difference between 2 mean which can =0(ex;p:50-50) which includes no significant

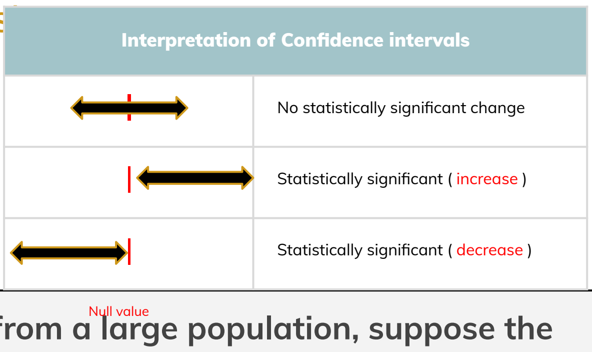
3- In analytical study,when the **Confidence interval in Odds ratio Or Relative risk (risk ratio) do not include one**.here we divide not taking the difference so it's either = 1 where there's no association or ris

The Result considered clinically significant when it has significant effect size.

**Width of the confidence interval (CI)**

○ A narrow CI implies high precision.

○ A wide CI implies poor precision (usually due to inadequate sample size).


**Interpretation of Confidence intervals**

No statistically significant change

Statistically significant ( increase )

Statistically significant ( decrease )

## Standard error of the mean

**In representative sample of 100 observations of heights of men, drawn at random from a large population, suppose the sample mean is found to be 175 cm (standard deviation = 10cm).**
**Can we make any statement about the population mean ?**

- We can not say that population mean is 175 cm because we are uncertain As to how much sampling fluctuation has occurred.
- What to do instead is to determine a range of possible values for population mean, with 95% of confidence[3].
- This range is called the 95% confidence interval and can be important adjuvant to significance test.

**In the example , n= 100, sample mean= 175, S.D,=10,**
**and the S, Error = 10/√100 = 1**

- Using the general format of confidence interval :
  Statistic ± confidence factor x Standard Error of statistic
- **Therefore the 95% confidence is 175 ± 1.96 * 1 = 173 to 177**
- That is, if numerous random sample of size 100 are draw. And 95% confidence interval is computed for each sample, **the population mean will be within the computed intervals in 95% of instances.**

## confidence intervals

| Example | Interpretation | Example | Interpretation |
|---|---|---|---|

### Example 1

- 100 KKUH student 60 do daily exercise (p =0.6).
- What it the proportion of student do daily exercise in the KSU.

$$SE(p) = \sqrt{\frac{pq}{n}}$$

$$=> 95\%\ CI = 0.6 \pm 1.96 \sqrt{\frac{0.6 \times 0.4}{100}}$$

$$= 0.6 \pm 1.96 \times \frac{0.5}{10}$$

$$= 0.6 \pm 0.1 = 0.5\ ;0.7$$

If someone repeat the study,We are 95 % confident that people who do daily exercise between 0.5(50%) to 0.7(70%).
(The closer the interval the better the precision is. And vice versa)

### Example 3: CI of difference between proportions (p1- p2)

- 50 patients with drug A, 30 cured (p1=0.6)
- 50 patients with drug B, 40 cured (p2=0.8)

$$95\%\ CI\ (\mathrm{P1}-\mathrm{P2}) = (\mathrm{P1}-\mathrm{P2}) \pm 1.96 \times SE\ (\mathrm{P1}-\mathrm{P2})$$

$$SE(\mathrm{P1}-\mathrm{P2}) = \sqrt{\frac{p1q1}{n1}+\frac{p2q2}{n2}}$$

$$= \sqrt{\frac{(0.6\times 0.4)}{50}+\frac{(0.8\times 0.2)}{50}} = \sqrt{0.008} = 0.09$$

$$=> 95\%\ CI\ (\mathrm{P1}-\mathrm{P2}) = [0.2-(0.09\times 1.96)]\ ;[0.2+SE\ (0.09\times 1.96)]$$

$$= 0.024\ ;0.3764 = 2.4\%\ to\ 37.6\%$$

-Statically significant because the C.I doesn't include zero value
-The Confidence interval is wide due to low sample size (poor precision).

### Example 2:CI of the mean

- 100 newborn babies, mean BW = 3000 (SD = 400) grams, what is 95% CI?
- 95% CI = X̄ + 1.96 (SEM)

$$SEM = \frac{SD}{\sqrt{n}}$$

$$=> 95\%\ CI = 3000 \pm 1.96\ (\frac{400}{\sqrt{100}})$$

$$= 3000 \pm 80 = (3000-80)\ ;\ (3000+80)$$

$$= 2920\ ;3080$$

We are 95% confident that true population mean lies within 2920 to 3080

### Example 4: CI for difference between 2 means

- **Mean systolic BP:**

**- 50 smokers = 146.4 (SD 18.5) mmHg**

**- 50 non-smokers = 140.4 (SD 16.8) mmHg**

- X̄1-X̄2 = 6.0 mmHg
- 95% CI(X̄1-X̄2) = (X̄1-X̄2) ± 1.96 x SE (X̄1-X̄2)
- SE (X̄1-X̄2) = S x √(1/n1+1/n2)

$$S = \sqrt{\frac{(n1-1)s1^2+(n2-1)s2^2}{(n1+n2-2)}}$$

$$S = \sqrt{\frac{(49\times 18.6)+(49\times 16.2)}{98}} = 17.7$$

$$SE(\bar{X}1-\bar{X}2) = 17.7\ \sqrt{\frac{1}{50}+\frac{1}{50}} = 3.53$$

$$95\%\ CI = 6.0 \pm (1.96\times 3.53) = -1.0\ ;13.0$$

The result is Not statically significant because the confidence interval include Zero value, so we accept $H_0$

# L26 Statistical Significance of Data II (95% CI)

## Interpretation

- If a 95% CI includes the null effect, the P-value is > 0.05 (and we would fail to reject the null hypothesis)
- If the 95% CI excludes the null effect, the P-value is < 0.05 (and we would reject the null hypothesis)

### Interpreting confidence intervals

**Example:** The following finding of non-significance in a clinical trial on 178 patients:

| Treatment | Success | Failure | Total |
|---|---|---|---|
| A | 76 (75%) | 25 | 101 |
| B | 51 (66%) | 26 | 77 |
| Total | 127 | 51 | 178 |

- Chi-square value = 1.74 ( p > 0.1) (non –significant) i.e. there is no difference in efficacy between the two treatments.
- The observed difference is 75% - 66% = 9%
- and the 95% confidence interval for the difference is:-4% to 22%
- This indicates that compared to treatment B, treatment A has at best an appreciable advantage (22%) and at worst a slight disadvantage (-4%).
- This inference is more informative than just saying that the difference is non significant.

### Interpreting confidence intervals

| Trial | Number dead / Randomized | | Risk Ratio | 95% C.I. | P value |
|---|---|---|---|---|---|
| | Intravenous nitrate | Control | | | |
| Chiche | 3/50 | 8/45 | 0.33 | (0.09,1.13)[1] | 0.08 |
| Wide interval: suggests reduction in mortality of 91%(1-0.09) and an increase of 13%(1-0.13) | | | | | |
| Flaherty | 11/56 | 11/48 | 0.83 | (0.33,2.12) | 0.70 |
| Jaffe | 4/57 | 2/57 | 2.04 | (0.33,10.71) | 0.40 |
| Reduction in mortality as little as 18%(1-0.82), but little evidence to suggest that IV nitrate is harmful | | | | | |
| Jugdutt | 24/154 | 44/156 | 0.48 | (0.28,0.82) | 0.007 |

[1] 1-Not significant as confidence interval including 1.

### Interpreting confidence intervals

Which of the following odds ratios for the relationship between various risk factors and heart disease are statistically significant at the 0.05-significance level? Which are likely to be clinically significant?

| Odds ratios | Statistically significant | Clinically significant | Reason |
|---|---|---|---|
| Odds ratio for every 1-year increase in age: 1.10 (95% Cl: 1.01-1.19)<br>1.1 means increase in the risk | ✓ | ✓ | C.I does not include 1<br>Significant effect size |
| Odds ratio for regular exercise (yes vs no): 0.50 (95% Cl: 0.30-0.82) | ✓ | ✓ | C.I does not include 1<br>Significant effect size |
| Odds ratio for high blood pressure (high vs normal): 3.0 (95% Cl: 0.90-5.30)<br>It's clinically significant because you can tell if you can maintain pressure as normal the risk of heart disease increased 3 times | | ✓ | C.I include 1<br>Significant effect size |
| Odds ratio for every 50-pound increase in weight: 1.05 (95% Cl: 1.01-1.20) | ✓ | | C.I does not include 1<br>Insignificant effect size |

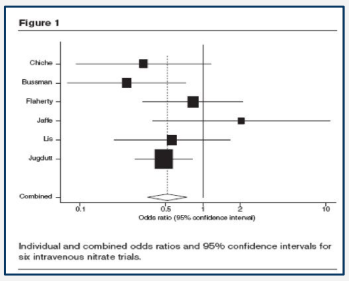


Figure 1

Individual and combined odds ratios and 95% confidence intervals for six intravenous nitrate trials.

**The figer name is Forest plot**
-The size of square indicates effect size.
- Diamond shape indicate sum of confidence intervals.

# **L26** Statistical Significance of Data II (95% CI)

| Comparison between p values and confidence interval | |
|---|---|
| P- value hypothesis testing | CI estimating |
| Gives you the probability that the result is merely caused by chance or not by chance, **it does not give the magnitude and direction of the difference.** | Indicates estimate of value in the population given one result in the sample, **it gives the magnitude and direction of the difference** |
| Provides a measure of strength of evidence against the Ho<br>Does not provide information on magnitude of the effect.<br>Affected by sample size and magnitude of effect:<br>interpret with caution! | How confident are we about the true value in the source population<br>Better precision with large sample size<br>Much more informative than P-value |
| "Is there a statistically significant difference between the two treatments?“ (or two groups) | "What is the size of that treatment difference?", and "How precisely did this trial determine or estimate the treatment difference?" |
| Analytical only | Analytical and descriptive |

# L Practical Session:
# Statistical Significance ( p-value and 95% CI)

## Definition of P-Value:

**Mark correct and false statements as: (Yes/No)**

| | |
|---|---|
| A. A “p” stands for probability and it ranges from 0 to 1. | **Yes** |
| B. A p-value of ≤ 0.05 is considered as not statistically significant. | **No** |
| C. A p-value of > 0.05 is considered as statistically significant | **No** |
| D. Statistically significant is more important than clinical significant. | **No** |
| E. The p-value is the probability of getting an outcome as extreme as or more extreme than the actually observed outcome (sample) under the null hypothesis. | **Yes** |
| F. Usually the null hypothesis is a statement of "no effect", "no difference" or "=0" and we are eager to find evidence against it. | **Yes** |
| G. When large samples are available, even small deviations from the null hypothesis will be significant. | **Yes** |

## Conclusions based on P-Value:

- There are two groups of employees (Teaching staff and Hospital staff)
- H0: Mean Income 1 = Mean Income 2
  - You draw a random sample of size 30 from each population.

**Statistical Test-result: p = 0.016** **Mark correct and false conclusions as: (Yes /No)**

| | |
|---|---|
| A. Statistically, the mean income of the two employee groups is equal. | **No** |
| B. With probability 0.016 teaching staff has the same mean income as hospital staff. | **No** |
| C. The sample data is not compatible (p=0.016) with the null hypothesis: the mean income in the two groups is equal. | **Yes** |
| D. We could not find a significant (at level 0.05) difference in mean income of two groups. | **No** |
| E. Data did not show a significant difference in mean income of two groups. | **No** |
| F. The sample data is compatible (p=0.016) with the null hypothesis that teaching and hospital staff have the same mean income. | **No** |

## Conclusions based on P-Value:

**Statistical Test-result: p = 0.09** **Mark correct and false conclusions as: (Yes /No)**

| | |
|---|---|
| A. Mean income in the two groups did not differ significantly (p=0.09). | **Yes** |
| B. Mean income in the two groups differs significantly (p=0.09). | **No** |
| C. The null hypothesis, that the mean income of teaching and hospital are equal, is rejected at significance level α= 0.05. | **No** |
| D. The null hypothesis, that the mean income of teaching and hospital are equal, is not rejected at significance level α=0.05. | **Yes** |

## Definition of Confidence Interval:

**Mark correct and false conclusions as: (Yes /No)**

| | |
|---|---|
| A. A confidence interval always covers the true parameter. | **No** |
| B. A confidence interval covers the true parameter with a given probability. | **No** |
| C. A confidence interval covers the statistic with a given probability. | **No** |
| D. In 100 repeated samples, 95% its confidence intervals will contain the true parameter. | **Yes** |

# L Practical Session:
# Statistical Significance ( p-value and 95% CI)

**Duality of P-value and 95% confidence intervals:**

- **Which of the 4 statements given below are either consistent or inconsistent by both p-values and 95% confidence intervals? And also comment on the width of the confidence interval where ever it is consistent.**

A) A study comparing BMI (each 50 male and female) reported **mean difference (male-female) = 6.0, p = 0.10, CI 95% = [-1 to 40]**

**Answer:** The mean BMI difference between male and female in the target population is **not statistically significant** (p=0.10, which is >0.05), also the 95% confidence interval for difference of mean value of BMI **included the null value "zero" (of no difference).** Hence both p-value and 95% CI are **consistent. The width of confidence interval is large due to small sample size, which indicates low precision o**f the estimate.

**B)** A study comparing BMI (each 500 male and female) reported **means difference (male-female) = 10.5, p = 0.01 CI 95% = [-2; 15]**

**Answer:** The p-value and 95% CI are **inconsistent.** Because p=0.01 which is <0.05 (statistically significant), where as 95% CI included the null value "zero" (of no difference).

**C)** A study comparing Systolic BP (each 50 male and female) reported **mean difference (male-female)=8.0, p = 0.0001 CI 95% = [-2; 20]**

**Answer:** The p-value and 95% CI are **inconsistent**. Because p=0.0001 which is <0.05 (highly statistically significant), where as 95% CI included the null value "zero" (of no difference).

**D)** A study comparing Systolic BP (each 500 male and female) reported **mean difference (male-female) = 7.5, p = 0.0001 CI 95% = [4.5; 12.0]**

**Answer:** The mean Systolic BP difference between male and female in the target population is highly statistically significant (p=0.0001, which is <0.05) also the 95% confidence interval for difference of mean value of Systolic BP does not included the null value "zero" (of no difference). Hence both p-value and 95% CI are **consistent**. The **width of confidence interval is small due to large sample size, which indicates a good precision of the estimate.**

In a sample of 100 children taken from a rural community, it was found anemia prevalence as 35%. Construct 95 % confidence interval for the prevalence of anemia for that community and give your inference. Also comment on the width of confidence interval.

**[Use: 95% CI for population proportion = p ± confidence factor × S.error of (p), Where confidence factor=1.96 and S.error of (p) = 4.8]**

**Solution:**
95% confidence limits are (0.35±1.96 x 0.048) = 0.2559 to 0.4441 = **26% to 44%**

With 95% confidence, we expect that the anemia prevalence in the population will be as minimum as 26% and as high as 44%. The width of 95% confidence interval is wide, due to sample size of 100 children, which indicates lack of precision of the estimate.

To examine the hypothesis that the low birth weight babies have a higher risk of coronary diseases in later life, a study was conducted in 100 low birth weight babies and in 100 babies born with normal weight. It was found that 15% among the former 10% in the latter had lifetime incidence of chronic diseases. Obtain 95% CI for the difference in proportions in these two groups. Is there a statistically significant difference in the incidence of coronary diseases of low birth weight babies and babies born with normal weight?

**[Use: 95% CI for (P1-P2) = (p1-p2) ± confidence factor × S.error of (p1-p2), Where confidence factor=1.96 and S.error of (p1-p2) = 0.0466]**

**Solution:**
95% C I for (P1-P2) is [(0.15-0.10) ± 1.96 x 0.0466] = - 0.0431 to 0.1432 = **-4.31% to 14.32%**

The CI shows that coronary diseases in low birth weight babies would be as higher as 14.32% when compared to normal group. As the 95% confidence intervals for difference of proportions (incidence of coronary disease) included "zero" (null value of no difference), it can inferred that there is **no statistically significant** difference between low birth weight babies and normal weight babies.

# L Practical Session: Statistical Significance ( p-value and 95% CI)

## ˘nterpretation of p-values and 95% confidence intervals in the following abstract:

**Title: The Outcome of Extubation Failure in a Community Hospital Intensive Care Unit: A Cohort Study** Seymour CW, Martinez A, Christie JD, Fuchs BD. Critical Care 2004, 8:R322-R327 (20 July 2004)

**Introduction:** Extubation failure has been associated with poor intensive care unit (ICU) and hospital outcomes in tertiary care medical centers. Given the large proportion of critical care delivered in the community setting, our purpose was to determine the impact of extubation failure on patient outcomes in a community hospital ICU.

**Methods:** A retrospective cohort study was performed using data gathered in a 16-bed medical/surgical ICU in a community hospital. During 30 months, all patients with acute respiratory failure admitted to the ICU were included in the source population if they were mechanically ventilated by endotracheal tube for more than 12 hours. Extubation failure was defined as reinstitution of mechanical ventilation within 72 hours (n= 60), and the control cohort included patients who were successfully extubated at 72 hours (n =93).

**Results:** The primary outcome was total ICU length of stay after the initial extubation. Secondary outcomes were total hospital length of stay after the initial extubation, ICU mortality, hospital mortality, and total hospital cost. Patient groups were similar in terms of age, sex, and severity of illness, as assessed using admission Acute Physiology and Chronic Health Evaluation II score ($P > 0.05$). Both ICU (1.0 versus 10 days; $P < 0.01$) and hospital length of stay (6.0 versus 17 days; $P < 0.01$) after initial extubation were significantly longer in reintubated patients. ICU mortality was significantly higher in patients who failed extubation (odds ratio = 12.2, 95% confidence interval [CI] = 1.5–101; $P < 0.05$), but there was no significant difference in hospital mortality (odds ratio = 2.1, 95% CI = 0.8–5.4; $P < 0.15$). Total hospital costs (estimated from direct and indirect charges) were significantly increased by a mean of US$33,926 (95% CI =US$22,573–45,280; $P < 0.01$).

**Conclusion:** Extubation failure in a community hospital is univariately associated with prolonged inpatient care and significantly increased cost. Corroborating data from tertiary care centers, these adverse outcomes highlight the importance of accurate predictors of extubation outcome.

**What is the sample size in each of Extubation failure and successfully extubated groups?**

**Answer:** Extubation failure =60 and successfully extubated =93

**On what basis the authors had mentioned that the patients groups were similar?**

**Answer:** By using $P >0.05$.

**Is there a statistically significant difference in ICU and hospital length of stay initial extubation in re-intubated patients? If yes what are the corresponding p-values?**

**Answer:** Yes, Both ICU (1.0 versus 10 days; $P < 0.01$) and hospital length of stay (6.0 versus 17 days; $P < 0.01$) after initial extubation were significantly longer in reintubated patients.

**How to interpret ICU mortality odds ratio =12.2? Is it statistically significant?**

**Answer:** The odd s of ICU mortality is 12.2 times higher in patients who failed extubation, when compared with the patients who successfully extubated. Yes the Odds ratio is statistically significant as p-value is <0.05.

**What is the interpretation of its 95% confidence interval: 1.5 - 101? Why this confidence interval is very wide?**

**Answer:** This study shows an odds ratio of 12.2. If this study is repeated 100 times, 95 times the odds ratio lies within 1.5 and 101. The confidence interval is wide due to small sample size, which indicates that the odds ratio of 12.2 is not a precise estimate.

# L27 Statistical tests to Observe the statistical significance of Quantitative variables

| Choosing the appropriate Statistical test | **Based on the three aspects of the data**<br>**(Types of variables، Number of groups being compared & Sample size.)** |
|---|---|

| Test | Z-test | Student's t-test |
|---|---|---|
| Study variable | Qualitative | Qualitative |
| Outcome variable | Quantitative or Qualitative | Quantitative |
| Comparison | ● Sample mean with population mean<br>Example: The education department at a university has been accused of “grade inflation” in medical students with higher GPAs than students in general.<br>● Two sample means<br>Example: Weight Loss for Diet vs Exercise | ● sample mean with population mean - Whether the sample mean is equal to the predefined population mean?-<br>● Two means (Independent samples)- Whether the CD4 level of patients taking treatment A is equal to CD4 level of patients taking treatment B ? -<br>● paired samples. - Whether the treatment conferred any significant benefit ? - |
| Sample size | larger in each group (>30) & standard deviation is known | each group <30 can be used even for large sample size |
| Z- value & t-Value | “Z and t” are the measures of: How difficult is it to believe the null hypothesis?<br>● **High z & t values:** Difficult to believe the null hypothesis - accept that there is a real difference.significant<br>● **Low z & t values:** Easy to believe the null hypothesis - have not proved any difference.no significant | |

## Karl Pearson Correlation Coefficient

| Study variable & Outcome variable | Quantitative |
|---|---|

- A number called the correlation measures both the direction and strength of the linear relationship between two related sets of quantitative variables.

| Measurement of correlation | **1. Scatter Diagram** **2. Karl Pearson's coefficient of Correlation** |
|---|---|
| Scatter diaphragm | ● **Using the axes**<br>- X-axis horizontally - Y-axis vertically - Both axes meet: origin of graph: 0/0<br>- Both axes can have different units of measurement. - Numbers on graph are (x,y) |

- A correlation coefficient (**r**) provides a quantitative way to express the **degree of linear relationship between two variables.**
- Range: **r** is always between -1 and 1
- Sign of correlation indicates **direction**:
  - high with high and low with low -> positive (Ex. Height and Weight, Age and BP)
  - high with low and low with high -> negative (Ex. Duration of HIV/AIDS and CD4 CD8, Price and Demand, Sales and advertisement expenditure)
  - no consistent pattern -> near zero
- Magnitude (absolute value) indicates **strength**:
  (-.9 is just as strong as .9) - 0.10 to 0.40 weak - 0.40 to 0.80 moderate - 0.80 to 0.99 high - 1.00 perfect

| About “r” | **1- r** is not dependent on the units in the problem **2- r** ignores the distinction between explanatory and response variables<br>**3- r** is not designed to measure the strength of relationships that are not approximately straight line. **4- r** can be strongly influenced by outliers. |
|---|---|
| Limitations | 1. Correlation coefficient is appropriate measure of relation only when relationship is linear.<br>2. Correlation coefficient is appropriate measure of relation when equal ranges of scores in the sample and in the population.<br>3. Correlation doesn't imply causality. |

# **L28** Statistical tests to observe the statistical significance of Categorical variables

| test | Study variable | Outcome variable | Comparison | Sample size | Expected frequency |
|---|---|---|---|---|---|
| **Chi-square** | Qualitative | Qualitative | Two or more proportions<br>E.g :<br>(two proportions) :**Prevalence** of exercise among female and male.<br>(more than two proportions):<br>1- prevalence of exercise among gp A, gp B, female gp.<br>2- prevalence of hypertension among 4 age gps. | **> 20** | >5 |
| **Fisher's exact** | | | Two proportions | < 20 | |
| **Macnemar's test (for paired samples)** | | | Two proportions | Any | |
| **Z-test** | | | - Sample proportion with population proportion<br>- two sample proportions | Larger in each group (>30) | |

| | | |
|---|---|---|
| **Chi-square Test** | **Purpose** | To find out whether the association between two categorical variables are statistically significant. |
| | **Null Hypothesis** | There is no association between two variables. |
| | **Requirements** | ✓ The data must be in the form of **frequencies** counted in each of a set of categories. Percentages cannot be used.<br>✓ The total number observed must exceed 20.<br>✓ The expected frequency under the $H_0$ hypothesis in any one fraction must not normally be less than 5.<br>✓ All the **observations must be independent** of each other. In other words, one observation must not have an influence upon another observation. |
| | **Application** | ● Testing for **independence (or association).**<br>● Testing for **homogeneity**. (Similarity)<br>● Testing of **goodness-of-fit.** |
| **Fisher's Exact Test** | The method of Yates's correction was useful when manual calculations were done. Now different types of statistical packages are available. Therefore, it is better to use Fisher's exact test rather than Yates's correction as it gives exact result. | |
| **McNemar's Test** | **When to use** | When we have a **paired sample**s and both the exposure and outcome variables are **qualitative variables (Binary).** |
| | **Situation** | Two paired binary variables that form a particular type of 2 x 2 table. e.g. matched case-control study or cross-over trial.<br>cross-over trial (is when you give treatment A to the group and after a while you give the same group treatment B so the group become a comparison group, another situation is when you give first group treatment A and second group treatment B and after awhile you cross the treatments between the groups). |

| Test | Equation |
|---|---|
| Ch-square | $X^2 = \sum\left[\frac{(o - e)^2}{e}\right]$ |
| Fisher's exact | $= \frac{R_1!R_2!C_1!C_2!}{n!a!b!c!d!}$ |
| MacNemar's test | $X^2 = \frac{(\lvert f-g \rvert - 1)^2}{f+g}$ |
| Z-test | $z = \frac{p-P}{\sqrt{\frac{pq}{n}}}$ ; $Z = \frac{P_1 - P_2}{\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}}$ |

# L Practical Session: Using appropriate statistical tests

What are the three criteria to use, in selecting the appropriate statistical test?

1- Type of variables.
2- Number of groups being compared. 3- Sample size

One of the best indicators of the health of a baby is his/her weight at birth.
Birth weight of >2500 gms is considered normal. A researcher wants to test whether birth weight of babies born last year in a region are normal. He took a sample of 100 babies and calculated mean and SD (Standard deviation) of the birth weights.
What test he should do to test his hypothesis that the birth weight of babies normal?

Outcome variable: Birth weight
Type of variable: Quantitative
How many groups: 1
Sample size: 100 (large) (more than 30)
Statistical test: The best test for this case is Z-test for single mean and we can also use student's t-test since its used for small and large sample size.

A team of scientists wants to test a new medication to see if it has either a positive or negative effect on intelligence, or no effect at all. In the population, the average IQ is 100 with a standard deviation of 15. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affect intelligence, using alpha = 0.05? Using an appropriate statistical test they concluded that medication has significantly affected intelligence.
What is the statistical test they used here?

Outcome variable: IQ
Type of variable: Quantitative
How many groups: 1
Sample size: 30 (small)
Statistical test: Student's t-test for single mean
Remember whenever you see mean or average its quantitative data and whenever you see proportion, out of and the frequency its categorical (qualitative) data.

research survey claims that 9 out of 10 doctors recommend aspirin for their patients with headaches. To test this claim, a random sample of 100 doctors is obtained. Of these 100 doctors, 82 indicate that they recommend aspirin. Is this claim accurate? Using an alpha of 0.05 with a two-tailed test, it was concluded that the claim that 9 out of 10 doctors recommend aspirin for their patients can't be rejected?
What is the statistical test used here?

Outcome variable: Recommend aspirin or not recommend Type of variable: Qualitative (nominal)
How many groups: 1
Sample size: 100 (large)
Statistical test: Z-test for single proportion.

A statistics teacher wants to compare his two classes to see if they performed any differently on the tests he gave that semester. Class A had 25 students with an average score of 70, standard deviation 15. Class B had 20 students with an average score of 74, standard deviation 25. Using alpha 0.05, did these two classes perform differently on the tests?
Using an appropriate statistical test, he concluded that there was no significant difference between the performances of Class A and Class B.
What is the statistical test the teacher has used ?

Outcome variable: Score
Type of variable: Quantitative
How many groups: 2 (class A and class B)
Sample size: 20 and 25 (small) (less than 30)
Statistical test: Student's t-test for independent samples (two means). Degrees of freedom: n1+n2-2 = 20+25-2= 43

We wish to test the proportion of smokers in a region is 15%. Taking a random sample of 320 persons in that region and found the proportion as 22%.What is an appropriate test here to test the hypothesis that sample proportion is not equal to proportion of smokers in that region?

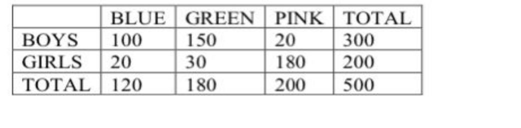
Outcome variable: Proportion of smoking (Smoker or non-smoker) Type of variable: Qualitative (nominal)
How many groups: 1
Sample size: 320 (large) (more than 30)
Statistical test: Z-test for single proportion. We can't apply the chi-square here because its a single proportion and Chi-Square test is for 2 or more proportion.

# L Practical Session: Using appropriate statistical tests

Researchers want to test the effectiveness of a new anti-anxiety medication. In clinical testing, 64 out of 200 people taking the medication report symptoms of anxiety. Of the people receiving a placebo, 92 out of 200 report symptoms of anxiety. Is the medication working any differently than the placebo? Test this claim using alpha = 0.05. what is the appropriate statistical test we can use in this situation?

Outcome variable: Symptoms of anxiety (present or absent) Type of variable: Qualitative (nominal)
How many groups: 2 (medication and placebo)
Sample size: 200 and 200 (large) (more than 30)
Statistical test: Z-test for two proportions.

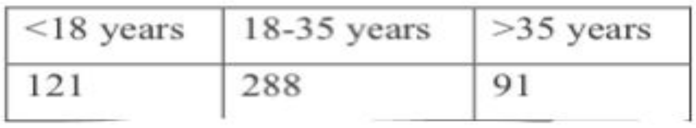
To test the association between gender and favorite color a study has been done on 500 college boys and girls are asked which is their favorite color: blue, green, or pink?
Results are shown below:
What is the appropriate statistical test we can use in this situation?



| | BLUE | GREEN | PINK | TOTAL |
|---|---|---|---|---|
| BOYS | 100 | 150 | 20 | 300 |
| GIRLS | 20 | 30 | 180 | 200 |
| TOTAL | 120 | 180 | 200 | 500 |

Outcome variable: Color
Type of variable: Qualitative (nominal)
How many groups: 2
Sample size: 500 (large) (more than 20)
Statistical test: Chi-square test for independence (or association). Whenever you see association you know by default that it is Chi-Square test.
The degrees of freedom: (r-1)(c-1) = (2-1)(3-1) = 2

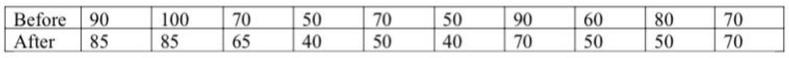
In 2010, ages of a random sample of 500 individuals from the same small town was taken.. below are the results:
Using alpha = 0.05, would you conclude that the population distribution of ages is equally distributed? What is the appropriate statistical test we can use in this situation?



| <18 years | 18-35 years | >35 years |
|---|---|---|
| 121 | 288 | 91 |

mean or proportions ? Proportions
Outcome variable: Age
Type of variable: Qualitative (age in category) (ordinal)
How many groups: 1
Sample size: 500 (large)(more than 20)
Statistical test: Chi-square test for homogeneity (to see whether the values are distributed equally or not). In this example it isn't uniformly distributed (not homogenous because 288 which is 57.6% of 500 are in the 18-35 years category so we need to provide statistical evidence

9-Researchers want to test a new weight loss pill. The following is the weights (kg) of 10 people before and after taking the pill.
How to find the effect of this pill on weight loss? What test will you do in this situation using alpha = 0.05? What is the degrees of freedom?



| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Before | 90 | 100 | 70 | 50 | 70 | 50 | 90 | 60 | 80 | 70 |
| After | 85 | 85 | 65 | 40 | 50 | 40 | 70 | 50 | 50 | 70 |

Outcome variable: Weight
Type of variable: Quantitative
How many groups: 1
Sample size: 10 (small) (less than 30)
Statistical test: Student's t-test for paired samples (dependent samples). (because it's before and after, with small sample size)
The degrees of freedom: (n-1) = (10-1) = 9 (n-1 because its one sample)

When the chi-squared test for 2x2 table is not valid (when the expected numbers are <5) What is an alternative test we use?

Answer : Fisher's exact test.
Remember, the only alternative test for Chi-Square test when the table is 2x2 with a small sample size which results in an expected value less than 5 is Fisher's exact test.
Fisher's exact test is only for 2x2 table, not bigger dimensions tables.

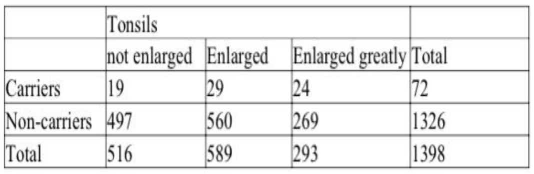
A researcher wants to test the mean systolic blood pressure of Saudi females of Dammam city is 120 mm/hg with 95% confidence. He took a random sample of 525 Saudi females and found the mean systolic blood pressure as 110 mm/hg .
What is an appropriate test here to test his hypothesis?

Outcome variable: Mean systolic blood pressure
Type of variable: Quantitative
How many groups: 1
Sample size: 525 (large) (more than 30)
Statistical test: Z-test for single mean (Z test because sample size is large)

# L Practical Session: Using appropriate statistical tests

The following data describe numbers of children with different sized palatine tonsils and their carrier status for Strep. pyogenes. What is the statistical test used to observe an association between carrier status and size of tonsils?

Outcome variable: Size of tonsils
Type of variable: Qualitative (ordinal)
How many groups: 2 (carriers and non-carriers)
Sample size: 1398 (large) (more than 20)
Statistical test: Chi-square test for association (or independence).
The degrees of freedom: (r-1)(c-1) = (2-1)(3-1) = 2
Can we calculate the odds ratio for this table? No, because it's 3 columns and 2 rows.



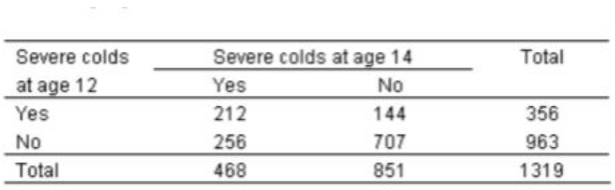
| | Tonsils | | | |
|---|---|---|---|---|
| | not enlarged | Enlarged | Enlarged greatly | Total |
| Carriers | 19 | 29 | 24 | 72 |
| Non-carriers | 497 | 560 | 269 | 1326 |
| Total | 516 | 589 | 293 | 1398 |

14. A researcher wants to test the mean HB of a pregnant women of Malaz area is 12 g/dl. He took a random sample of 20 and found that the mean score is 11g/dl and standard deviation is 34 g/dl. Could this sample originate from a population of mean = 12 g/dl? What is an appropriate test here?

Outcome variable: Hemoglobin
Type of variable: Quantitative
How many groups: 1
Sample size: 20 (small) ( less than 30)
Statistical test: Student's t-test for single mean. ( you can't apply z-test here because sample size is small)
The degrees of freedom: (n-1) = (20-1) = 19

15. A research team claims that their new drug increases the birth weight of babies. In order to test this, he took a random sample of 75 women for treatment group and 75 for Control group and at the end of the study period it was found Average birth Weight 3100 g and SD 420g for treatment group and for control **average** weight was 2750 g and SD 425g. What is an appropriate test to be done here?

Outcome variable: Birth weight
Type of variable: Quantitative
How many groups: 2 (treatment group and control group)
Sample size: 75 and 75 (large)(more than 30)
Statistical test: Z-test for two means. You can also apply student's t-test for independent samples because it can be used for small and large samples. Degrees of freedom: n1+n2-2 = 75 + 75 - 2 = 148



| Severe colds at age 12 | Severe colds at age 14 | | Total |
|---|---|---|---|
| | Yes | No | |
| Yes | 212 | 144 | 356 |
| No | 256 | 707 | 963 |
| Total | 468 | 851 | 1319 |

16. In an epidemiological survey, 1319 schoolchildren were assessed symptoms of severe cold at the age of 12 and again at the age of 14. At age 12, 356 (27%) children were reported to have severe colds in the past 12 months compared to 468 (35.5%) at age 14. what test is to be used to test these proportions?Was there a significant increase of the prevalence of severe cold?

Outcome variable: Symptoms of severe cold
Type of variable: Qualitative (nominal)
How many groups: 1
Sample size: 1319 (large) (more than 20)
Statistical test: McNemar's chi-square test. (Because they are related "paired dependent sample" Follow up of the same group)(same sample at the age of 12, same sample at the age of 14) Degrees of freedom: n-1 = 2-1= 1

17. A researcher wants to quantify the linear relationship between systolic blood pressure and age of his study subjects. What is the appropriate plot so as to observe the relationship and what statistical measure he has to apply to quantify this relationship?
Solution:

Outcome variable: Systolic blood pressure and age
Type of variable: Quantitative
What is the aproperite plot: Scatter plot. By putting age on x-axis and systolic BP on y-axis. Statistical measures: Karl pearson of correlation coefficient.

**18. What is the range of correlation coefficient?**

Between -1 and +1
-1 to 0 → negative correlation
0 to +1 → positive correlation
The + and - gives us the direction and the values gives us the magnitude.

19. What are the statistical tests to use, for test of association and for the measure of association? Solution:

- Test the association: Chi-square test (will see either there is association or not)
- Measure the association: Odds ratio (for cross sectional, prospective study and case control) or relative risk (for retrospective study and RCT).

20. What are the degrees of freedom for 3 x 4 & 2 x 3 contingency tables? Solution:

Answer :(R-1)(C-1)
*** first number is the rows, second number is the columns (Rows x Columns)**
- 3 x 4 table = (3-1)(4-1) = 6
- 2 x 3 table = (2-1)(3-1) = 2

## Leader

**Rania almutiri**

## The work done by

**Raina almutiri**

**Renad alhomaidi**

**Haya Alanazi**

**Hessah Fahad**

**Shatha Aldhohair**

**Samar almohammedi**

**Nourah Alklaib**

**shatha Aldossari**

**Sara alobed**

**Noura almasoud**

**Abdulaziz alomar**