STATISTICAL ANALYSIS OF DATA
POST GRADUATE 2019-2020

PROF DR NAJLAA FAWZI

Statistical Inference :

Is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample drawn from that population.

Inferential Statistics

- **Generalize from samples to pops**
- **Hypothesis testing**
- **Relationships among variables**

Make predictions

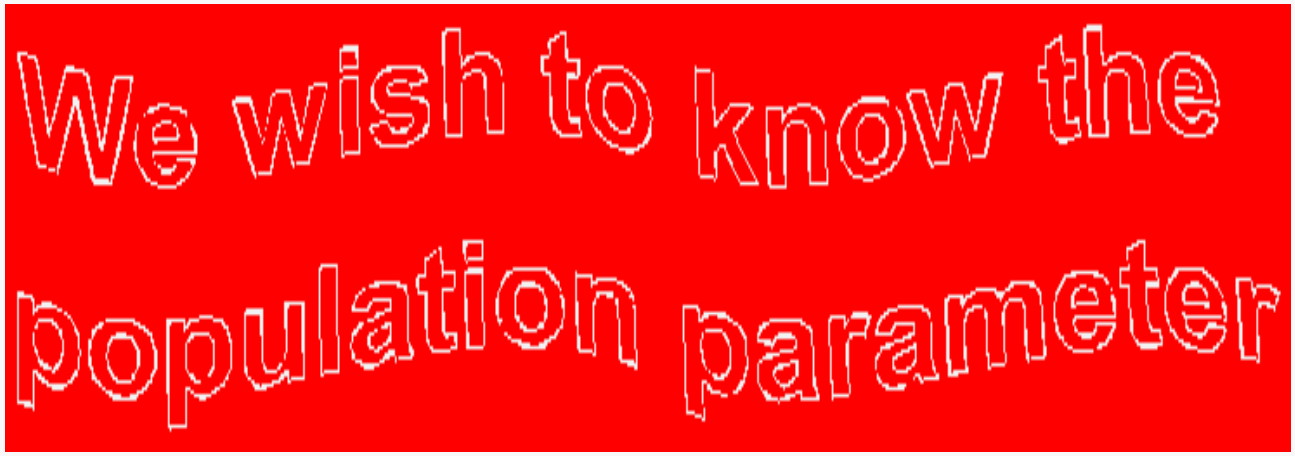
Estimation is the first general area of statistical inference.

The second general area , hypothesis testing.

The process of estimation entails calculating , from the data of a sample , some statistic that is offered as an approximation of the corresponding parameter of the population from which the sample was drawn.

The rational behind estimation in the health sciences field rests on the assumption that the workers in this field have an interest in the parameters , such as means & proportions , of various populations.

If this is the case, there is a good reason why one must rely on estimating procedures to obtain information regarding these parameters .



For each of the parameters , we can compute two types of estimate :

a point estimate and an interval estimate

A POINT ESTIMATE: is a single numerical value used to estimate the corresponding population parameters.

AN INTERVAL ESTIMATE: consists of two numerical values defining a range of values that, with a specified degree of confidence , most likely includes the parameters being estimated.

Confidence interval

A range of values (interval) that act as good estimates of the unknown population parameter.

What are confidence intervals?

confidence intervals provide a range about the observed effect size.

This range is constructed in such a way that we know how likely it is to capture the true – but unknown – effect size.

Confidence interval

It is the interval within which a parameter value is expected to lie with certain confidence levels , as could be revealed by repeated samples.

It is the interval (range) around the mean of population in which the means of multiple samples same population are dispersed.

The specified probability is called the confidence level, and the end points of the confidence interval are called the confidence limits.

If independent samples are taken repeatedly from the same population , and the confidence interval is calculated , then a certain percentage (confidence level) of the intervals will include the unknown population parameter .

Confidence Interval For A population Mean

CONFIDENCE LIMITS:

The properties of normal curve for mean values are

1. **It is symmetrical**
2. **It is a bell shaped curve**
3. **Mean = median= mode**
4. **A- population mean ± 1SE limits include 68.27% of the sample mean**

B- population **MEAN ± 1.96 SE** limits include **95%** of the sample mean values.

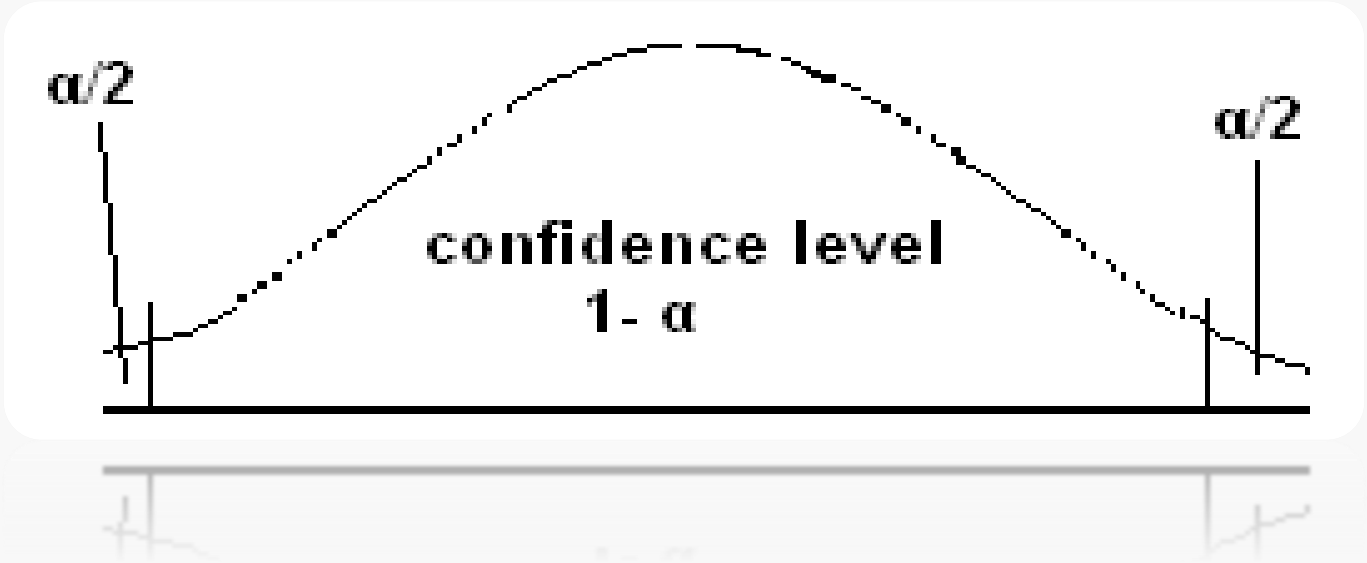
C- population **MEAN ± 2.58 SE** limits include **99%** of the sample meanvalues.

D- population **MEAN ± 3.29 SE** limits include **99.9%** of the sample mean values.

INTERVAL ESTIMATEION (Confidence Interval _C.I_)

It is consistent of 2 numerical values (upper & lower values) defining the interval within which the unknown parameter lies with certain degree of confidence.

These values (upper & lower values) depend upon the confidence level which is equal to 1- α, where α is the probability of error.



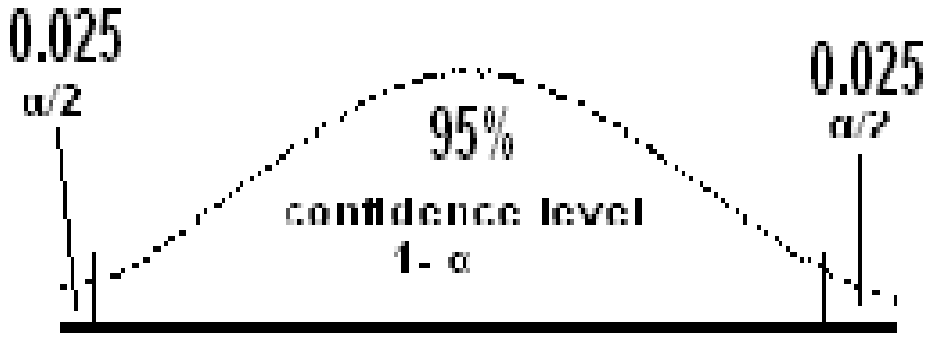
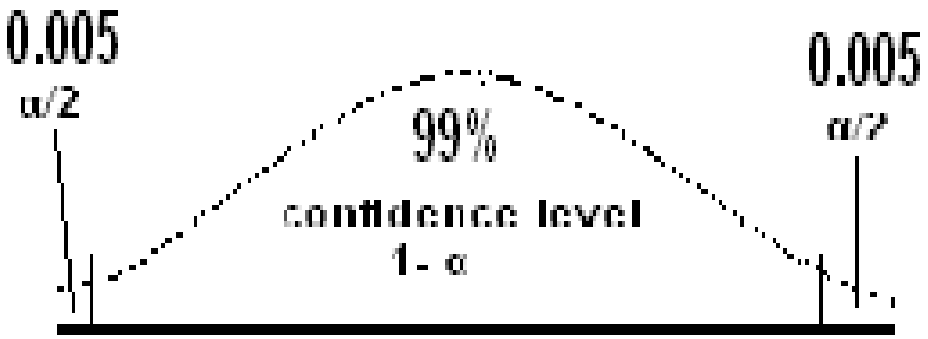
from figure : **Total AUC = 1→**

if α (probability of error) is 10% (or 0.1)

then [α/2=0.05] will be on each side of the confidence interval and the confidence level at 1- α will be 0.9 or 90%

α = 0.05, α/2=0.025, & confidence level is 95%

and for α = 0.01, then α/2 = 0.005 & confidence level is 99%.

probability of error	CURVES	Z value (no.of δ)
α = 0.1		1.645
α = 0.05		1.96
α = 0.01		2.58

Interval Estimate Components

Estimator ± (reliability coefficient) x (standard error)

The value of *Z* is referred to as the reliability Coefficient.

Reliability coefficient for estimates drawn from the population or large samples is the value of Z corresponding to the confidence level:

1.645	for 90%CI
1.96	for 95%CI
2.58	for 99%CI

Confidence interval (CI) is central and symmetrical around (m) so that there is α/2 chance that the parameter is more than the upper limit & α/2 chance to be lower than the lower limit.

Margin of Error

In a confidence interval, the range of values above and below the sample statistic is called the margin of error.

For example, suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the vote. The newspaper states that the survey had a 5% margin of error and a confidence level of 95%. These findings result in the following confidence interval: We are 95% confident that the independent candidate will receive between 25% and 35% of the vote.

Example-1

The mean of indirect serum bilirubin level for (16) four-day old infants was found to be 5.98mg/dl; Assume that bilirubin levels in 4- day old infants are approximately normally distributed with SD of 3.5 mg /dl.

Find 90%, 95% & 99% C.I for population mean .

$$\%C.I \text{ for } \mu = m \pm (Z * \delta / \sqrt{n})$$

$$90\%C.I \text{ for } \mu = 5.98 \pm (1.645 * 3.5/\sqrt{16}) = 4.54 \text{ to } 7.42$$

$$95\%C.I \text{ for } \mu = 5.98 \pm (1.96 * 3.5/\sqrt{16}) = 4.265 \text{ to } 7.695$$

$$99\%C.I \text{ for } \mu = 5.98 \pm (2.58 * 3.5/\sqrt{16}) = 3.72 \text{ to } 8.24$$

Factors Affecting Width of a Confidence Interval

Factors that determine the size of a standard error and therefore the width of a confidence interval:

1. **The amount of variability in the sample: The greater the variability (for sample means, the larger the standard deviation), the wider the confidence interval.**

$$SE = \frac{sd}{\sqrt{n}}$$

As SD increases, SE increases: When the variability among subjects in a sample is large, the variability among the means of repeated samples will also be large

2. The sample size

The greater the sample size, the more narrow the confidence interval.

$$SE = \frac{sd}{\sqrt{n}}$$

As sample size increases, SE decreases: As more individuals in the population are sampled, the estimate of the population parameter will become more precise

How precisely does the sample statistic estimate the population parameter?

To illustrate the calculation and interpretation of confidence intervals we'll use the HR data from a sample of 84 adults:

The sample mean HR was 74.0 bpm
The sample standard deviation was 7.5 bpm

$$\bar{x} = 74.0,\ \ sd = 7.5,\ \ SE = \frac{sd}{\sqrt{n}} = \frac{7.5}{\sqrt{84}} = \frac{7.5}{9.2} = 0.8$$

90% CI: 74.0 ± 1.645(0.8) = 74.0 ± 1.3 = 72.7, 75.3

95% CI: 74.0 ± 1.960(0.8) = 74.0 ± 1.6 = 72.4, 75.6

99% CI: 74.0 ± 2.575(0.8) = 74.0 ± 2.1 = 71.9, 76.1

For the HR example:

- **In our sample of 84 adults, we can be 95% confident that the that the true population heart rate is between 72.4 and 75.6 bpm**
- **If we increase our sample size to 500 adults, we can be 95% confident that the true population heart rate is between 73.34 and 74.657 bpm**

The interpretation of the confidence interval

As we increase the level of confidence, the interval widens because the larger the range between the lower and upper bounds, the more confident we can be that the interval contains the true mean.

1- The narrower the interval , the more precise our estimate of the population value (and the more confidence we have in our study value as an estimate of the population value).

The width of confidence interval is directly related to the level of confidence; smallest with 90% (but not reliable level of confidence, i.e. high probability of error).

The largest with 99% (a highly confident estimation but very wide range)

and that's why 95% level of confidence is the most practical to use.

2- Another factor can affect the width of confidence level, that the width of the interval is inversely related to the square root of the sample size, i.e. inversely related to the sample size, so we can decrease the width of this interval by taking larger samples whenever it is feasible.

Small sample size means high variability (large sample variance and standard deviation) and consequently a large confidence interval, so the explanation of a large confidence interval is either a small sample size &/or a high confidence level (99%).

3- The range of values in the interval is the range of the populationvalues most dependable with the data from our sample or study.

Precision is the degree to which a figure (such as an estimate of a population mean) is immune from random variation.
The width of the confidence interval reflects precision—the wider the confidence interval, the less precise the estimate.

Because the width of the confidence interval decreases in proportion to the square root of sample
size, precision is proportional to the square root of sample size.
So to double the precision of an estimate, sample size must be multiplied by 4; to triple precision, sample size must be multiplied by 9; and to quadruple precision, sample size must be multiplied by 16.

Increasing the precision of research therefore requires disproportionate increases in sample size; thus, very precise research is expensive and time-consuming.

Confidence Intervals in Medical Research

Confidence intervals (usually 95%) around sample means are commonly reported in published medical research.

Other sample statistics that are commonly reported with confidence intervals include:

- **difference between 2 means**
- **proportions**
- **differences between 2 proportions**
- **correlations**
- **relative risks**
- **odds ratios**

An example of the use of confidence intervals

Ramipril is an angiotensin-converting enzyme (ACE) inhibitor which has been tested for use in patients at high risk of cardiovascular events.

In one study published in the *New England Journal of Medicine*, a total of 9,297 patients were recruited into a randomized, double-blind, controlled trial.

The key findings presented on the primary outcome and deaths are shown below.

Incidence of primary outcome and deaths from any cause

Outcome	Ramipril group (n=4,645) number (%)	Placebo group (n=4,652) number (%)	Relative risk (95% CI)
Cardiovascular event (including death)	651 (14.0)	826 (17.8)	0.78 (0.70–0.86)
Death from non-cardiovascular cause	200 (4.3)	192 (4.1)	1.03 (0.85–1.26)
Death from any cause	482 (10.4)	569 (12.2)	0.84 (0.75–0.95)

Example-2: A sample of (10), 12-years old boys their mean height was 59.8 inches & population standard deviation of 2 inches and a sample of (10), 12-year old girls of mean height 58.5 inches & population standard deviation of 3 inches. Assuming normality; find 90%C.I for the difference in means of height between girls & at this age.

$$\%C.I \text{ for } (\mu 1-\mu 2) = (m1-m2) \pm (Z \times \sqrt{[(\delta_1 2/n1) + (\delta_2 2/n2)]})$$

$$90\% \text{ C.I for } (\mu 1-\mu 2) = (59.8-58.5) \pm (1.645 \times \sqrt{[\{(2)_2/10\} + \{(3)_2/10\}]})$$

$$= -0.58 \text{ to } 3.18 \text{ (not significant)}$$

N.B) if the interval includes null value (zero in addition & subtraction or one

in multiplication & division); results will be not significant.

Example -3 **In a survey, 300 adults were interviewed , 123 said that they had yearly medical check up.**

Find the 95%C.I for the proportion of adults having yearly check up.

$$\%C.I \text{ for } P = p \pm [Z \times \sqrt{[\{p(1-p)\}/n]}$$

$$P = 123/300 = 0.41$$

$$1 - p = 1 - 0.41 = 0.59$$

$$95\%C.I \text{ for } P = 0.41 \pm [1.96 \times \sqrt{(0.41 \times 0.59/300)}] = 0.35 \text{ to } 0.47$$

Example-4

200 patients suffering from a certain disease were randomly divided into two equal groups, the 1st group received new treatment, 90 patients recovered within 3 days, out of 2nd group who received the standard treatment 78 patients recovered within 3 days.

Find the 95%C.I for the difference in proportions between the two treatments groups.

%C.I for (P1 -P2) =

$$(p_1 - p_2) \pm (Z \times \sqrt{[p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2]})$$

(p1 -p2) = (90/100) - (78/100) = 0.12

95%C.I for (P1-P2) =

$$0.12 \pm (1.96 \times \sqrt{[(0.9 \times 0.1)/100 + (0.78 \times 0.22)/100]})$$

95%C.I for (P1-P2) = 0.02 to 0.22