Biostatistics

Ass. Proff. Dr. Ban Nadhum

Lecture 6

Tests of statistical significance

Confidence Intervals (CI):

- The mean of a sample is only a **"point estimate"** of the mean for the entire population. Although this sample mean may truly reflect the population mean, there is uncertainty in this value.
- **Confidence intervals are constructs used to describe the range of values possible for this estimate.**
- **The 95%confidence interval represents 95% confidence that the lower and upper limits of this interval include the true mean of the population.**
- **The general expression for the confidence interval is:**
- **Confidence interval = point estimate(mean) ± (critical factor x SE)**

- **For 90%, 95%, 99% confidence intervals, the critical factor is 1.64, 1.96, and 2.58 respectively.**
- **Confidence level:** Is the probability value attached to a given confidence interval, it can be expressed as a percentage (90%,95%, 99%)

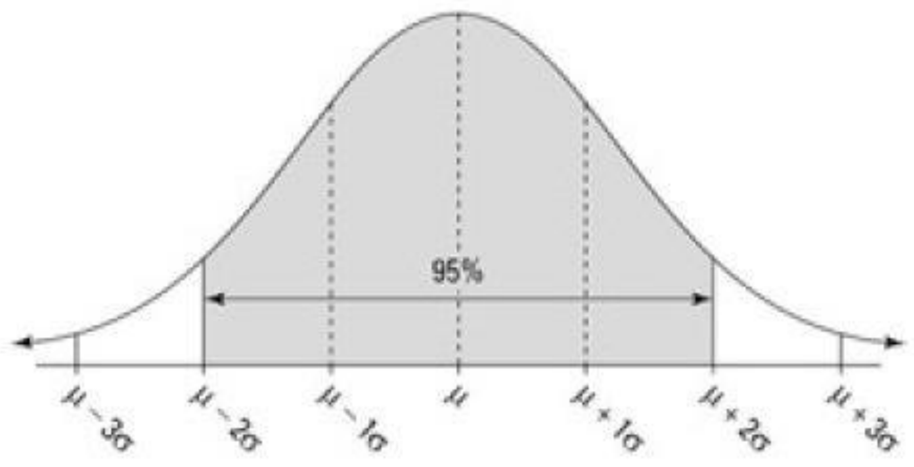
- **Procedure (steps) for statistical tests:**

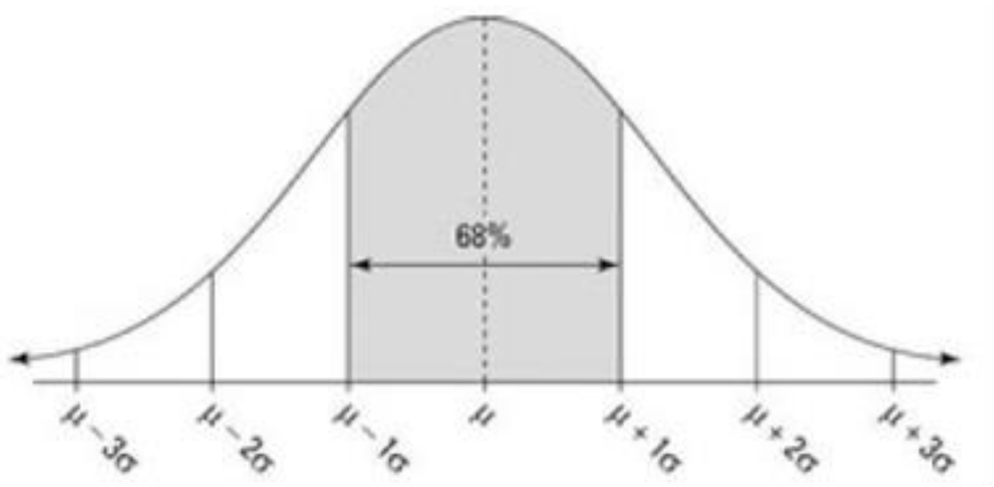
1. **State the null hypothesis:**

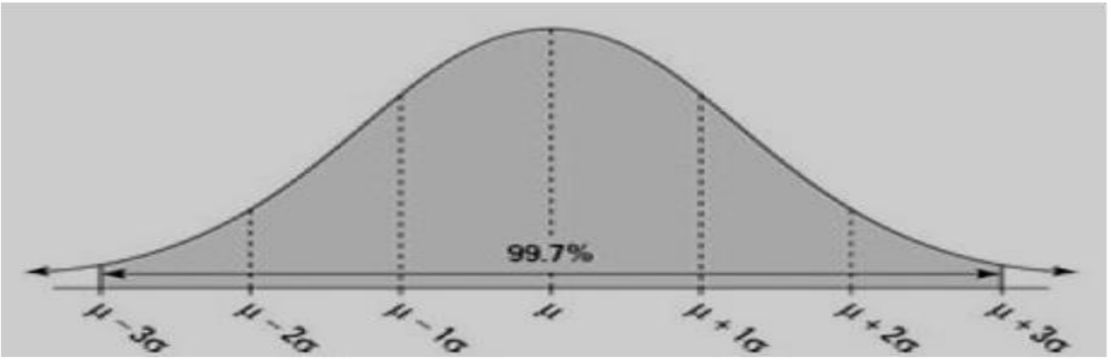
- Statistical hypothesis is that which is stated in such a way that can be evaluated by appropriate statistical techniques or tests.
- **Null hypothesis (H_0)** is the hypothesis to be tested (null = no difference), it states that there is no statistical or real difference between the sample mean and the population mean, and if there is any difference it is due to chance

  Alternate hypothesis (H_A): the opposite of the null hypothesis, it states that there is a statistical or real difference between the sample mean and the population mean, and it is not due to chance.

Normal


95%
μ − 3σ
μ − 2σ
μ − 1σ
μ
μ + 1σ
μ + 2σ
μ + 3σ


68%
μ − 3σ
μ − 2σ
μ − 1σ
μ
μ + 1σ
μ + 2σ
μ + 3σ


99.7%
μ − 3σ
μ − 2σ
μ − 1σ
μ
μ + 1σ
μ + 2σ
μ + 3σ

- **α error: reject the null hypothesis even if it is true.**
  β error: accept null hypothesis even if it is false
- **2. State the level of significance** (($\propto$))= 0.05, or 0.01, 0.1 level.
- **P< 0.05** means if the probability of finding a difference by chance (due to the sampling process) is less than 5%, means that the difference must be significant. **The** level of significance, that at the area of rejection will reject true null hypothesis $H_{0,}$ then we select a small value of (($\propto$)) in order to make the probability of rejection of true H0 small.
- The most counted values of⟨$\propto$⟩

- 3. **Choose the test statistic.**
- **Steps of hypothesis :**

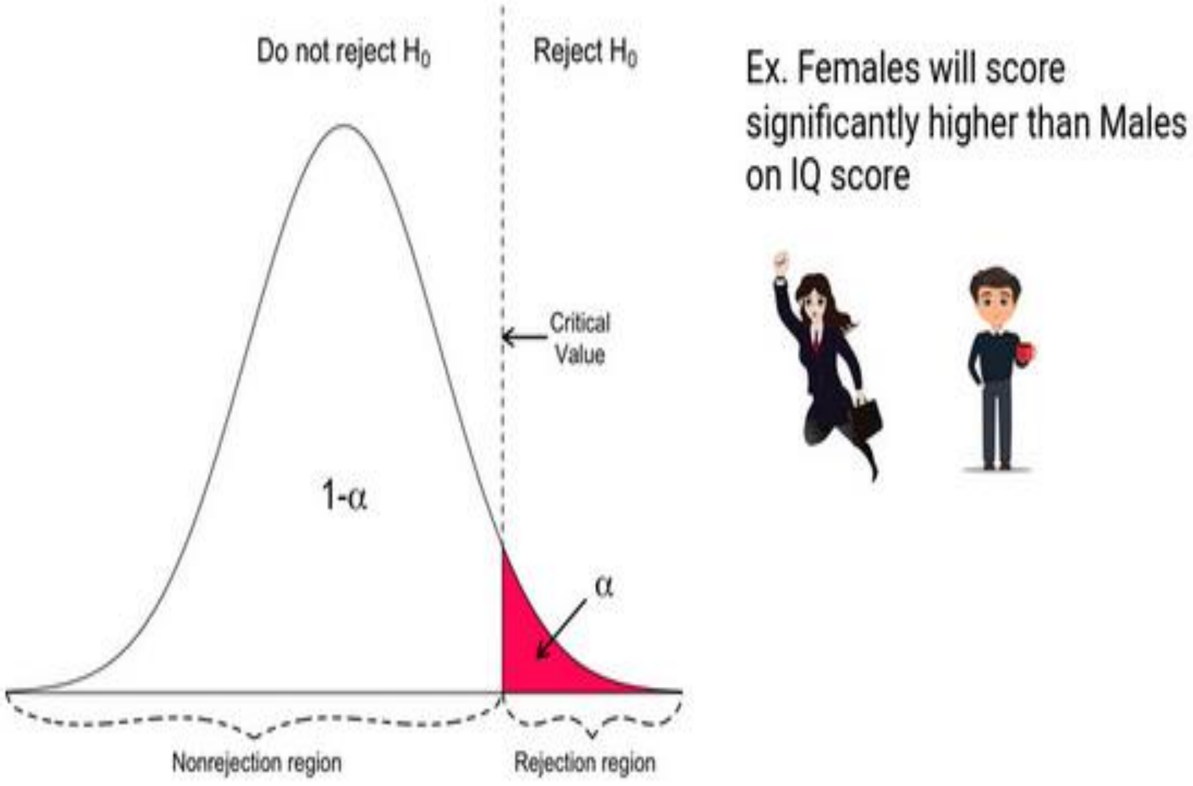
1. **Data**
2. **Assumption of normal distribution**
3. **Set the Hypothesis**
4. **Level of Significance**
5. **Formula**
6. **Conclusion**

- **The test statistic (tests of significance):**
- The test statistic is some statistic that may be computed from the date of the sample. It tests the conflict between what has been assumed (by the null hypothesis) and what is found.
- **There are many types of tests for different types of data:**

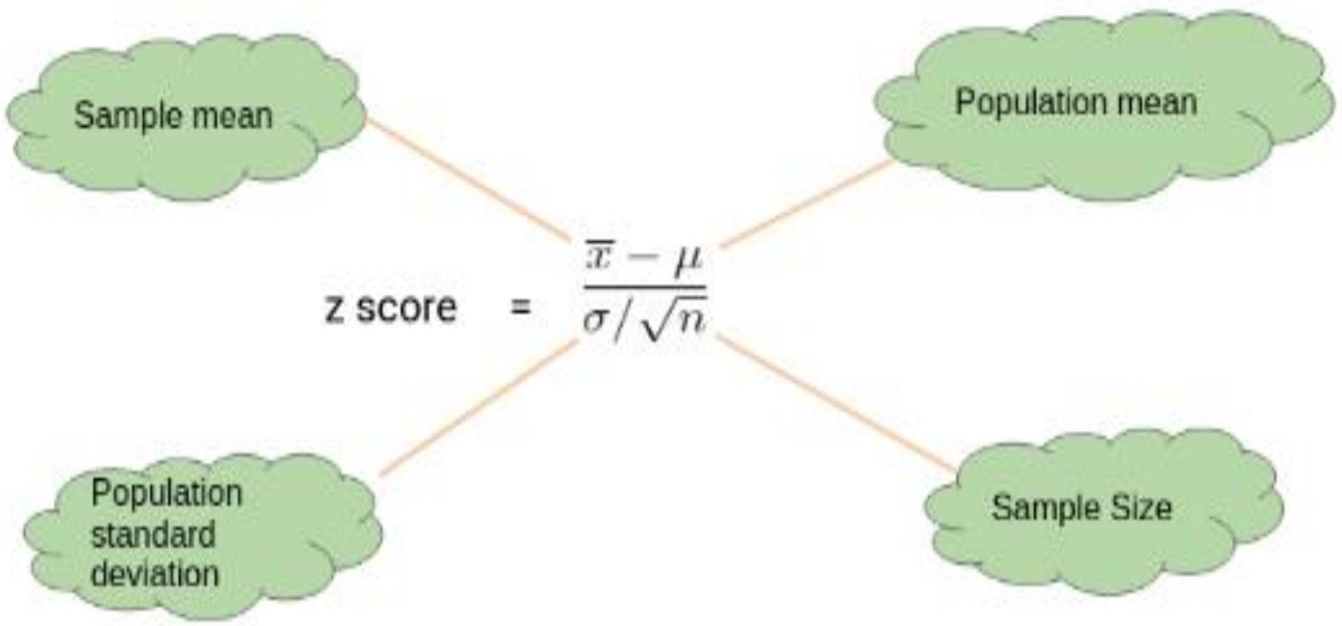
Z test **T-test** **Chi-squared test (X^2 test)**

- The most counted values of⟨∝⟩ are 1%, 5 %, 10%, 20%, mostly use 5%---0.05

⟨∝⟩ - level of significance	confidence level 100 - ∝	critical factor
5 %	95%	1.96
1%	99%	2.58
10%	90%	1.64



- **Z test:** is applied for both quantitative and qualitative normally distributed data.
- 1-For quantitative data:
- **a. to test whether a sample is drawn from/ or belong to a population when there is a big sample size (n > 30)**
  - **The test statistic will be:**
- **Z = X”- µ / SE (x”)**
- **Z = X”- µ / σ**



- **Examples**
- **1: One sample mean compares the normal population:**
- **The mean systolic blood pressure of 130 men of age 50-65 years who are put on a special diet for two years is 145.5 mm Hg. If the mean systolic blood pressure of the normal population is 125.0 mm Hg and the standard deviation is 20 mm Hg, how reasonable is it to conclude that the systolic blood pressure of that sample is not different from that of the population?**

1- Data: X” = 145.5 μ = 125 σ = 20 n = 130

2- Null hypothesis (H_0): there is no significant difference between the mean systolic blood pressure of 130 men and the mean systolic blood pressure of the population OR: X” = μ
- Alternate hypothesis (H_A): there is a significant difference between the mean systolic blood pressure of 130 men and the mean systolic blood pressure of the population OR: X” ≠ μ

3- Assumption of normal distribution

4- Level of significance: 0.05

5-Test statistic: Z test Z= X”- μ / σ/√N Z= (145.5 – 125) / (20/√130) Z= 11.7

C. I at level 95% = X”± cf x SE= 145.5 ±1.96 X 20/√130 = 148.93 -142.2

- 6- Conclusion:
- a. The calculated Z(11.7) is>1.96 at 0.05 level,
- b. P<0.05
- c. There is a significant difference between the mean of systolic blood pressure of the sample and the mean of the population at 0.05 level. .
- We reject the Null hypothesis **(H_0)** and accept the alternate

2. Tow means for quantitative data (unknown population means): To test the difference between two means of two samples when ($n_1 + n_2 > 30$) the test statistic will be

- **Example:** A study for calcium levels in men and women ages 60 years is summarized in the following table. P=0.01

- $$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{V_1}{n1} + \frac{V2}{n2}}}$$

Gender	N	$\bar{x}$	Standard deviation
Men	16	8	2
Women	25	6	3
Total	41		

Z t

- Steps of hypothesis:

- 1. Data: $\bar{X1}$ = 8, σ1 =2, n1 =16 ------- $\bar{X2}$ = 6, σ2 =3, n2 = 25
- 2. Assumption of normal distribution
- 3. Set the Hypothesis:
- 4. Ho: there is no significant difference between the mean of calcium levels in men and women.
- HA: there is a significant difference between the mean calcium levels in men and women.
- **Or $\bar{X1} = \bar{X2}$ $\bar{X1} \neq \bar{X2}$**

- 5. Level of Significance: ∝ = 0.01

6. Formula: $\dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{V_1}{n1} + \frac{V2}{n2}}}$

-
- Z = 8 – 6 / √ 4 / 16 + √9 /25 = 2 / 0.5 + 0.6 = 1.82
- 6. Conclusion: a. Calculated Z < Tabulated (2.58). b. p >0.01,
- c. Accept Ho & reject HA. d. There is no significant deference between 2 means.

Z Test

- **2-** For qualitative data:

to test whether a sample is drawn from/ or belong to a population (for proportion):

$$Z = (\bar{P} - P) / \frac{\sqrt{pq}}{\sqrt{n}}$$

- **$\bar{P}$ = proportion of sample   q = 1-P**
- **P = proportion of population**

a. **to test whether 2 samples has different proportion**

- **$p = (r_1+r_2) / (n_1+n_2)$**
- **$SE\ (p_1-p_2) = \sqrt{\{p\ (1-p) \times\ (1/n_1 + 1/n_2)\}} = \sqrt{\{pq \times\ (1/n1 + 1/n2)\}}$**
- **$Z = (p_1 - p_2)\ / SE\ (p_1-p_2)$**
- **OR :**
- **$Z = \dfrac{(\bar{P1} - \bar{P2}) - (p1 - p2)}{\sqrt{P1 \times q1} / \sqrt{n} + \sqrt{P2 \times q2} / (\sqrt{n})}$**

Z Test

- **Example 1: The mean systolic blood pressure of 130 men of age 50-65 years whom are put on special diet for two years is 145.5 mm hg. If the mean systolic blood pressure of the normal population is 125.0mm hg and a standard deviation is 20 mm hg, how reasonable is it to conclude that the systolic blood pressure of that sample is not different from that of the population?**
- **Solution:**
- 1- Data: X” = 145.5 μ = 125 σ = 20 n = 130
- 2. Null hypothesis: there is no significant difference between the mean systolic blood pressure of 130 men and the mean systolic blood pressure of the population
- - Alternate hypothesis: there is a significant difference between the mean systolic blood pressure of 130 men and the mean systolic blood pressure of the population
- 3. Assumption of normal distribution
- 4- level of significance: 0.05
- 5-test statistic: Z test = **Z= X”- μ / σ /√n**
- Z= (145.5 – 125) / (20/√130) Z= 11.7
-

- 6- Conclusion: The calculated Z is > 1.96 at 0.05 level So P< 0.05
- Thus there is a significant difference between pressure
- of 130 men and the mean systolic blood pressure of the population at 0.05 level.
- We reject Null hypothesis, and accept the alternate hypothesis at 0.05 level.
- \- At 0.01 level, the calculated Z (11.7) is > **2.58**
- **Example 2:**
- **A survey, a total of 88 households used a river for water supply, 49 of them had episodes of diarrhea against 10 from 36 households using the well water. Is there a statistically significant difference in the proportions with episodes of diarrhea between the households using river and well water supply?**
- **Solution:**
- 1. Null hypothesis: there is no significant difference in the proportions with
- episodes of diarrhea between the households using river and well water supply.

– Alternate hypothesis: there is a significant difference in the proportions with episodes of diarrhea between the households using river and well water supply.

2. Assumption of normal distribution

- 3. level of significance: 0.05

$Z = (0.556 - 0.227) / 0.0988$ $Z = 3.32$

6- Conclusion:

1- The calculated Z is > 1.96 at 0.05 level

2- So $P < 0.05$

3- There is a significant difference in the proportions with episodes of diarrhea between the households using river and well water supply, and the difference is not due chance at 0.05 level.

4- We reject Null hypothesis, and accept the alternate hypothesis at 0.05 level.

- **$p = (r_1 + r_2) / (n_1 + n_2)$**
- **$SE(p_1 - p_2) = \sqrt{p(1-p) \times (1/n_1 + 1/n_2)}$**
- **$Z = (p_1 - p_2) / SE(p_1 - p_2)$**
- 5. Data:
- $r_1 = 49$, $n_1 = 88$ $r_2 = 10$, $n_2 = 36$
- . test statistic: Z test for proportion is chosen
- $P_1 = 49/88 = 0.556$
- $P_2 = 10/36 = 0.227$
- $P = (49 + 10) / (88 + 36) = 0.476$
- **$SE(p_1 - p_2) = \sqrt{0.476(1 - 0.476)(1/88 + 1/36)}$**

  $= 0.0988$

Z test

- P” = proportion of sample
- p =proportion of population
- ((Should n x p > 5 & n (1-p) > 5 ------ valid sample size))
- * set the hypothesis;
- H0 **p > 0.04** HA **P ≤0.04**
- *** Z = (P”-** P) / √ (Pq) / √n
- Where √ Pq / √n = √ (0.31 x 0.69)/ √150 = 0.001426 (variance)
- $\frac{0.40-0.31}{0.001426}$ =2.38
- **1-** 2.38 > 1.64 occur in rejection area
- 2- hat **P < 0.04**

3- There is a significant difference in the proportions not due chance at 0.05 level.

4- We reject Null hypothesis, and accept the alternate hypothesis

- **CONFIDENCE INTERVAL FOR(Z) TEST ;TO ESTIMATE THE POPULATION PARAMETER:**
- **1. FOR QUANTITATIVE DATA:**
- **A. FOR ONE MEAM :**
- **X”± P (C.F) (SD/(√(N)**
- THE CI OF THE MEAN AT 95% C . LEVEL = X” ± 1.96 X (SD/(√(N)
- THE CI OF THE MEAN AT 99% C .LEVEL = X” ± 2.58 X (SD/(√(N)
- **B. FOR 2 MEANS :**
- **C.I OF 2MEAN AT 95% C. LEVEL =** $\bar{X}_1 - \bar{X}_2 \pm 1.96 \times \sqrt{V_1/N_1 + V_2/N_2}$
- **C.I OF 2 MEANS AT 99% C .LEVEL =** $\bar{X}_1 - \bar{X}_2 \pm 2.58 \times \sqrt{V_1/N_1 + V_2/N_2}$
- **3. FOR QUALITATIVE DATA**:
- THE PARAMETERS HERE ARE THE SAMPLE PROPORTION (P”), THE POPULATION PROPORTION (P), AND THE STANDARD ERROR OF THE PROPORTION SE(P) .
- **SE(P) = √P”Q”/√N** Q”= 1-P