# BIOSTATISTICS

## FINAL

**Lecture 8 (Types of t-tests for qualitative data)**

**Given by: Dr. Nedal Al-Nawaiseh**

**Done by Doctor:**


**Abdullah Daradkh**

**Mahmoud Al-Otoom**



MUTAH UNIVERSITY

# Chi-square ($X^2$) test

→ **Variables are dichotomous & should be independent & mutually exclusive**

Correlation study ( association ): researcher measures two variables, understands and assess the statistical relationship between them with no influence from any extraneous variable.

**It is used when we want to study whether there is a relation between certain condition and certain characteristic.**

## *SUMMARY OF STATISTICAL TESTS:*

| | | |
|---|---|---|
| **CATEGORICAL DATA** | **Enough data** | **Too little data (<5 in a cell)** |
| **Any r x c table** | **Chi-square** | **Fisher's Exact** |
| **CONTINUOUS DATA** | **Normal (even if transformed to normal) or large n** | **Not normal: (non-parametric tests)** |
| **One (group) sample** | **1-sample t-test** | **Kolmogorov-Smirnov** |
| **Two samples** | **2-sample t-test** | **Mann-Whitney U or Rank Sum** |
| **Paired data** | **1-sample t-test on paired differences (paired t-test)** | **Wilcoxon Signed-Rank** |
| **Three or more samples** | **Analysis of variance (ANOVA)** | **Kruskal-Wallis** |

**Examples:**

1. **Relation between smoking and lung cancer.**
2. **Relation between consanguinity and congenital anomalies.**
3. **Relation between occurrence of breast cancer and presence of family history.**
4. **Relation between diabetes and prolonged healing of wounds.**
5. **Relation between the technique of vaccination (scratch or multiple pressure) and the success of vaccination.**

**So, to do the test, first we should make the table, and in this case. the type of table is a contingency table or an association table or two by two table.**

**N.B.:**

**Chi-square test must be calculated from the absolute observed and expected numbers and NOT from percentages of other proportions.**

---

**Let us take an example which is the technique of vaccination and successful of vaccination and the following table illustrates the example:**

| Result | Technique | | Total |
|---|---|---|---|
| | Scratch | Multiple pressure | |
| Successful | (1) 66 | (3) 85 | 151 |
| Unsuccessful | (2) 34 | (4) 15 | 49 |
| Total | 100 | 100 | 200 |

**→The problem here is to determine whether or not there is any significant difference in the success rate of the two methods of vaccination.**

**The success rate is 66% (66 children out of 100) for the scratch technique and 85% (85 children out of 100) for the multiple pressure technique.**

**Is this difference just due to chance or is the success rate significantly higher when the multiple pressure technique is used? (critical value = 1.96 ^2 = 3.84)**

**1)Null hypothesis ($H_0$): No difference between the success rates of the two techniques.**

**2)Alternative hypothesis ($H_1$): There is a significant difference.**

**3) First we have to calculate the expected frequencies (E) for each cell.**

**These are the frequencies to be expected if there were no differences between the two techniques and E for each cell**

$$E = \frac{\text{Total of column} \times \text{total of row of that cell}}{\text{Grand total}}$$

**The actual or observed (O) frequencies and the expected (E) frequencies in each cell are then compared by a measure called $X^2$ (chi square) and this is called the calculated $X^2$.**

**If O and E of each cell are equal; the value of $X^2$ will be zero.**

**The greater the difference between O and E, the greater the value of $X^2$.**

**The value on this calculated $X^2$ is compared with the critical value of $X^2$ using a statistical table of the $X^2$ distribution and our comparison must depend on:**

**4)Degree of freedom = (c - 1) × (r - 1) and here in this example = (2 - 1) × (2 - 1) = 1 × 1 = 1.**

**Level of significance and usually = 5% (critical value of $X^2$ by d.f = 1 at 5% level of significance = 3.84.**

**5) If the calculated $X^2$ > critical value of $X^2$, so we accept $H_1$ and reject $H_0$, if calculated $X^2$ < critical value of $X^2$ we accept $H_0$ and reject $H_1$.**

$$E_1 = \frac{100 \times 151}{200} = 75.5$$

$$E_2 = \frac{100 \times 49}{200} = 24.5$$

$$E_3 = \frac{100 \times 151}{200} = 75.5$$

$$E_4 = \frac{100 \times 49}{200} = 24.5$$

$$X^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2} + \frac{(O_3-E_3)^2}{E_3} + \frac{(O_4-E_4)^2}{E_4}$$

$$= \frac{(66-75.5)^2}{75.5} + \frac{(34-24.5)^2}{24.5} + \frac{(85-75.2)^2}{75.5} + \frac{(15-24.5)^2}{24.5}$$

$$= \frac{(9.5)^2}{75.5} + \frac{(9.5)^2}{24.5} + \frac{(9.5)^2}{75.5} + \frac{(9.5)^2}{24.5} = 9.7581$$

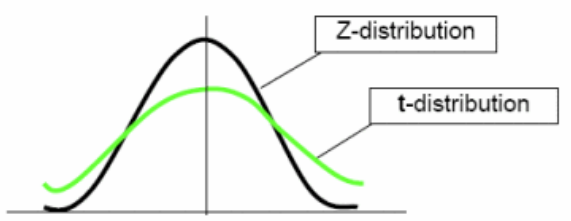
**Since calculated $X^2$ (9.7581) > critical value (3.84) so success rate of the multiple pressure method is significantly greater than the scratch method.**

**Practice questions: Using the data in the following table to test if there is a relation between consanguinity and congenital anomalies.**

| Consanguinity | Congenital anomalies | | Total |
|---|---|---|---|
| | Yes | No | |
| Yes | 30 | 10 | 40 |
| No | 20 | 40 | 60 |
| Total | 50 | 50 | 100 |

**Critical value at 5% level of significance and 1 d.f = 3.84.**
**1)$H_0$ : No – between Consanguinity &Congenital anomalies**
**2)$H_A$ : There is a relation between Consanguinity &Congenital anomalies**
**3) suitable test → chi-square .**
**4) E1=40X50/100= 20**
**E2=60X50/100=30**
**E3=40X50/100=20**
**E4=60X50/100=30**

**$X^2$ = (30-20)/20 + (20-30)/30 + (10-20)/20 + (40-30)/20 = 16.667**
**5)d.f = 1 → t=3.84**
**7) reject $H_0$ and accept $H_A$.**




**Again, the t-distribution approaches the normal distribution as n approaches infinity.**
**Example:**

| Consanguinity | Congenital anomalies | | Total |
|---|---|---|---|
| | Yes | No | |
| Yes | 10 | 15 | 25 |
| No | 12 | 14 | 26 |
| Total | 22 | 29 | 51 |

**E1= 25X22/51=10.8**
**E2=26X22/51=11.2**
**E3=25X29/51=14.2**
**E4=26X29/51=14.8**
**$X^2$= (10-10.8)/10.8 + (12-11.2)/11.2 + (15-14.2)/14.2 + (14-14.8)/14.8 = 0.147**
**→ Accept the null hypothesis.**
**2. Statistical tests for Continuous data:**

## 2.4. Testing continuous, non-normal data

**2-All of the above tests assume that your data are normally or approximately normally distributed, or your sample size is large enough to apply the properties of the central limit theorem.**
**But sometimes your data are not normal and your sample size is relatively small.**

**You can try to mathematically transform the data into a normal distribution (for example by taking the square root, or the logarithmof all the values).**
**If you can make them normal, you can use the t-tests or ANOVA.**

**2-If the data are still not normally distributed, we use a different class of tests known as "non-parametric" tests, i.e. the Mann Whitney U test.**
**These tests are based on the ranking or ordering of the data, rather than their numerical values.**

## *2.5. Statistical Test for Nominal Data:*

**Categorical or nominal data is usually tested with the Chi-square test statistic. Here's an example:**

**1) Null hypothesis: Cigarette use does not affect the risk of lung cancer in men;**
**or Proportion of smokers who get lung Ca = Proportion of nonsmokers who get lung Ca.**
**2) Alternative hypothesis: The two proportions are not equal (two-sided test).**
**Set alpha = 0.05.**
**Study Design: 20-year cohort study of 210 men, ages 30-50 living. After 20 years, we OBSERVE:**

| | Lung Cancer | No Cancer | Total |
|---|---|---|---|
| Smokers | 25 (A) | 75 (B) | 100 |
| Nonsmokers | 17 (C) | 93 (D) | 110 |
| Total | 42 | 168 | 210 |

**→ Smokers and nonsmokers are the two groups being compared.**
**The data of interest is the rate of lung cancer, which is a categorical variable (yes/no).**
**This is a 2x2 table; it has 4 cells; each is arbitrarily named A-D. For categorical data, use a Chi-square test statistic:**

$$X^2 = \sum \frac{(\text{Observed}-\text{Expected})^2}{\text{Expected}}$$

**We calculate EXPECTED values under the null hypothesis of no difference between the two groups using the following**

$$\text{Expected for each cell} = \frac{(\text{Column total} * \text{Row total})}{\text{Grand total}}$$

| | Lung Cancer | No Cancer | |
|---|---|---|---|
| Smokers | 100 x 42 /210 = 20 | 100 x 168 /210 = 80 | 100 |
| Nonsmokers | 110 x 42/210 = 22 | 110 x 168/210= 88 | 110 |
| | 42 | 168 | 210 |

**Then, we can calculate the chi-square test statistic:**

| Cell | Observed | Expected | O-E | $(O-E)^2$ | $(O-E)^2/E$ |
|---|---|---|---|---|---|
| A | 25 | 20 | 5 | 25 | 25/20 = 1.25 |
| B | 75 | 80 | -5 | 25 | 25/80 = 0.31 |
| C | 17 | 22 | -5 | 25 | 25/22 = 1.14 |
| D | 93 | 88 | 5 | 25 | 25/88 = 0.28 |
| | | | | $X^2$= 2.98 | |

→ Getting a p-value: calculate the "degrees of freedom" (df) = (# rows - 1) * (# columns - 1),
For example, a 2x2 table always has: (2 - 1) * (2 - 1)= 1*1 = 1 df.
The probability has been calculated for seeing any particular chi-square value with any number of degrees of freedom by chance alone, under the chi-square distribution.
These probabilities can be found in $X^2$ tables or computer programs. So, we look up the probability of getting this value of 2.98 (or one more extreme) with 1 degree of freedom by chance alone…p=0.09. P>alpha, so we cannot reject our null hypothesis.
Conclude: The difference we observed between cigarette smokers and non-smokers in the rate of lung cancer could have occurred by chance alone.
Note: If there are too few data in a single cell of an r x c table (less than 5 observations per cell), the chi-square test is not accurate.
You then need to use a special test, called the Fisher's Exact test.
*Greater chi-square value , greater relationship.*

---

# *Chi square*

→Chi square is a non-parametric test of statistical significance for bivariate tabular analysis (also known as cross-breaks).

→Any appropriately performed test of statistical significance lets you know the degree of confidence you can have in accepting or rejecting a null hypothesis.

→Typically, the hypothesis tested with chi square is whether or not two different samples (of people, texts, whatever) are different enough in some characteristic or aspect of their behavior that we can generalize from our samples that the populations from which our samples are drawn are also different in the behavior or characteristic

| | Normal | Chronic bronchitis |
|---|---|---|
| Smoking | 20 | 80 |
| Nonsmoking | 80 | 20 |

Table: Relation between smoking and chronic bronchitis

→Bivariate tabular analysis is good for asking the following kinds of questions:

1-Is there a relationship between any two variables IN THE DATA?

2-How strong is the relationship IN THE DATA?

**3-What is the direction and shape of the relationship IN THE DATA?**

**Requirements (assumptions):**

**1-The sample must be randomly drawn from the population.**

**2-Data must be reported in raw frequencies (not percentages);**

**3-Measured variables must be independent;**

**4-Values/categories on independent and dependent variables must be mutually exclusive and exhaustive;**

**5-Observed frequencies cannot be too small.**

## Student t-test

**** The t-test assesses whether the means of two groups are *statistically* different from each other.**

**** This analysis is appropriate whenever you want to compare the means of two groups,**

***Requirements (assumptions):***

**1- A normal (Gaussian) distribution for the populations of the random errors,**

**2- that there is no significant difference between the standard deviations of both population samples.**