Chapter 4

Hypothesis Testing

Contents

Definition 4.0.1. *(Statistical Hypothesis)*

A statistical hypothesis is a statement about the values of parameters of a population.

Null Hypothesis H_0: The hypothesis we aim to test..

Alternative Hypothesis H_1: The negation of H_0.

Definition 4.0.2. *(Hypothesis Test)*

A process intended to provide a decision rule allowing us to choose between two statistical hypotheses based on sample results.

Definition 4.0.3. *(Chi-Square Tests)*

Chi-square tests are based on the χ^2 statistic proposed by Karl Pearson. Their main purpose is to compare distributions. These tests can be applied to qualitative or quantitative variables.

There are three types of chi-square tests:

Goodness-of-fit Test, Homogeneity Test, Independence Test.

4.1 Goodness-of-Fit Test

Objective:

Compare an observed sample distribution to a theoretical distribution.

Notation:

T_i: theoretical frequencies

O_i: observed frequencies

n: total number of observations.

k: number of categories.

Steps:

❶ Formulate hypotheses

H_0: The observed distribution matches the theoretical distribution.

H_1: The observed distribution differs from the theoretical distribution.

❷ Compute theoretical frequencies

Validity condition: all $T_i \geq 5$.

❸ Test statistic:

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - T_i)^2}{T_i}$$

❹ Critical value

Read from the chi-square distribution table with degrees of freedom $df = k - 1$

$$\chi_T^2 = \chi^2_{(df,1-\alpha)}$$

❺ Decision

If $\chi_c^2 \leq \chi_T^2$ accept H_0

If $\chi_c^2 > \chi_T^2$ reject H_0

Example 4.1.1.

In a maternity ward, out of 100 *births,* 44 *boys and* 56 *girls were recorded. Is this observation consistent with national statistics indicating proportions of* 53% *boys and* 47% *girls?*

Solution

Test: Chi-square goodness-of-fit.

❶ *Formulate hypotheses*

H_0 *: The observed distribution matches the national distribution.*

H_1 *: The observed distribution is different.*

❷ *Compute theoretical frequencies*

sex	O_i	T_i
boys	*44*	*53*
girls	*56*	*47*
total	*100*	*100*

all $T_i \geq 5$ *so the test is valid.*

❸ *Test statistic:*

$$\chi^2_c = \sum_{i=1}^{k} \frac{(O_i - T_i)^2}{T_i} = \frac{(44-53)^2}{53} + \frac{(56-47)^2}{47} = 3.25$$

❹ *Critical value :*

$$\begin{cases} df = k - 1 = 2 - 1 = 1 \\ 1 - \alpha = 1 - 0.05 = 0.95 \end{cases} \quad so \quad \chi^2_T = \chi^2_{(1;0.95)} = 3.84$$

❺ *Decision*

$\chi^2_c < \chi^2_T$ *Accept* H_0

Conclusion: The observed distribution matches the national one.

4.2 Homogeneity Test

Objective:

Compare two or more observed distributions from different samples.

Notation:

l: number of rows

c: number of columns

Steps:

❶ Formulate hypotheses

H_0: The distributions are equivalent.

H_1: The distributions are not equivalent.

❷ Compute theoretical frequencies

$$n_i = \sum_{j=1}^{c} O_{ij} \quad n_j = \sum_{i=1}^{l} O_{ij} \quad T_{ij} = \frac{n_i n_j}{n}$$

Validity condition: all $T_{ij} \geq 5$.

❸ Test statistic:

$$\chi_c^2 = \sum_{j=1}^{c} \sum_{i=1}^{l} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$

❹ Critical value:

Read from the chi-square distribution table with degrees of freedom $df = (c-1)(l-1)$

$$\chi_T^2 = \chi^2_{(df, 1-\alpha)}$$

❺ Decision

If $\chi_c^2 \leq \chi_T^2$ accept H_0

If $\chi_c^2 > \chi_T^2$ reject H_0

Example 4.2.1.
Two drugs, A and B, were tested on two patient groups. The results were:

	Symptom Disappeared	Symptoms Persisted	Worsened	Side Effects
A	100	40	20	30
B	220	80	70	40

Do the two treatments have the same effect? We will take $\alpha = 5\%$

Solution

Test: Chi-square homogeneity test

❶ *Formulate hypotheses*

H_0: *The two treatments are equivalent .*

H_1: *The two treatments are not equivalent.*

❷ *Compute theoretical frequencies: (contingency table)*

	Symptom Disappeared		Symptoms Persisted		Worsened		Side Effects		total n_i
	O_{ij}	T_{ij}	O_{ij}	T_{ij}	O_{ij}	T_{ij}	O_{ij}	T_{ij}	
A	100	101.33	40	38	20	28.5	30	22.16	190
B	220	218.66	80	82	70	61.5	40	47.83	410
total n_j	320		120		90		70		n=600

all $T_i \geq 5$ *so the test is valid.*

❸ *Test statistic:*

$$\chi^2_c = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{(O_{ij}-T_{ij})^2}{T_{ij}} = \sum_{j=1}^{4}\sum_{i=1}^{2}\frac{(O_{ij}-T_{ij})^2}{T_{ij}} = 7.94$$

❹ *Critical value:*

$$\begin{cases} df = (c-1)(l-1) = (4-1)(2-1) = 3 \\ 1-\alpha = 1-0.05 = 0.95 \end{cases} \quad so \quad \chi^2_T = \chi^2_{(3;0.95)} = 7.815$$

❺ *Decision:* $\chi^2_c > \chi^2_T$ *reject* H_0

Conclusion: The two treatments are not equivalent.

4.3 Independence Test

Objective: study the relationship between two variables in the same sample.

Example 4.3.1.

A food poisoning occurred in a primary school. A doctor recorded the following data:

	Sick	*Healthy*
Ate chocolate ice cream	*69*	*83*
Did not eat ice cream	*31*	*17*

Is food poisoning linked to eating chocolate ice cream?We will take $\alpha = 5\%$

Solution

Test: Chi-square independence test

❶ Formulate hypotheses

H_0 : No association between food poisoning and ice cream consumption.

H_1 : There is an association between food poisoning and chocolate ice cream consumption.

❷ Compute theoretical frequencies: (contingency table)

	Sick		Healthy		total n_i
	O_{ij}	T_{ij}	O_{ij}	T_{ij}	
Ate chocolate ice cream	69	76	83	76	152
Did not eat ice cream	31	24	17	24	48
total n_j	100		100		n=200

all $T_{ij} \geq 5$, so the test is valid.

❸ Test statistic:

$$\chi_c^2 = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{(O_{ij} - T_{ij})^2}{T_{ij}} = \sum_{j=1}^{2} \sum_{i=1}^{2} \frac{(O_{ij} - T_{ij})^2}{T_{ij}} = 5.4$$

❹ Critical value:

$$\begin{cases} df = (c-1)(l-1) = (2-1)(2-1) = 1 \\ 1 - \alpha = 1 - 0.05 = 0.95 \end{cases} \quad so \quad \chi_T^2 = \chi^2_{(1;0.95)} = 3.841$$

❺ Decision: $\chi_c^2 > \chi_T^2$ rejcet H_0

Conclusion: There is an association between food poisoning and chocolate ice cream consumption.