Chapter 4

Hypothesis Testing

Contents

Statistical Hypothesis

Definition 4.0.1. *A statistical hypothesis is a statement concerning the values of parameters in a population.*

- *Null hypothesis H_0: the hypothesis we want to test.*
- *Alternative hypothesis H_1: the negation of H_0.*

Hypothesis Test

Definition 4.0.2. *A procedure aimed at providing a decision rule to choose between two statistical hypotheses based on sample results.*

Chi-Square Tests

Definition 4.0.3. *Chi-square tests are based on the χ^2 statistic introduced by Karl Pearson. Their main purpose is to compare distributions. These tests can be applied to* ***qualitative*** *variables.*

There are three types of chi-square tests:

- **Goodness-of-fit test**
- **Homogeneity test**
- **Independence test**

4.1 Goodness-of-Fit Test

4.1.1 Objective

Compare an observed sample distribution with a theoretical distribution.

4.1.2 Notation

- T_i: expected (theoretical) frequencies
- O_i: observed frequencies
- n: total number of observations
- k: number of categories

4.1.3 Steps of the Test

1. **Formulate hypotheses**
   - H_0: The observed distribution matches the theoretical distribution
   - H_1: The observed distribution differs from the theoretical one
2. **Calculate expected frequencies** Validity condition: all $T_i \geq 5$
3. **Test statistic**:
$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - T_i)^2}{T_i}$$
4. **Critical value**: found in chi-square table with df $= k - 1$
$$\chi_T^2 = \chi^2_{(k-1,\ 1-\alpha)}$$
5. **Decision**:
   - If $\chi_c^2 \leq \chi_T^2$: Accept H_0
   - If $\chi_c^2 > \chi_T^2$: Reject H_0

Examples 4.1.1. *In a maternity hospital, 100 births were observed: 44 boys and 56 girls. Is this observation consistent with national statistics, which state 53% boys and 47% girls?*

Solution:

***Choice of test**: Chi-square goodness-of-fit test*

1. ***Hypotheses***
   - H_0*: The observed distribution matches the national statistics*
   - H_1*: The observed distribution differs*
2. ***Expected frequencies***

Gender	*Observed* O_i	*National %*	*Expected* T_i
Boys	*44*	*0.53*	*53*
Girls	*56*	*0.47*	*47*
Total	*100*	*1*	*100*

   All $T_i \geq 5 \longrightarrow$ *the test is valid.*

3. ***Chi-square statistic:***
$$\chi_c^2 = \frac{(44-53)^2}{53} + \frac{(56-47)^2}{47} = 3.25$$
4. ***Critical value:***
$$df = 2 - 1 = 1, \quad 1 - \alpha = 0.95, \quad \chi_T^2 = \chi^2_{(1,\ 0.95)} = 3.84$$
5. ***Decision**: Since* $\chi_c^2 < \chi_T^2$*, we accept* H_0
6. ***Conclusion**: The observed distribution is consistent with the national distribution.*

4.2 Homogeneity Test

4.2.1 Objective

Compare two or more observed distributions across different samples.

4.2.2 Notation

- r: number of rows (groups)
- c: number of columns (categories)
- n: total observations

4.2.3 Steps of the Test

1. **Formulate hypotheses**
   - H_0: The observed distributions are equivalent
   - H_1: The observed distributions are not equivalent
2. **Compute expected frequencies**

   Let:
   $$n_i = \sum_j O_{ij}, \quad n_j = \sum_i O_{ij}, \quad T_{ij} = \frac{n_i \cdot n_j}{n}$$
   Condition: All $T_{ij} \geq 5$
3. **Test statistic**:
   $$\chi_c^2 = \sum_{j=1}^{r} \sum_{i=1}^{c} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$
4. **Critical value**:
   $$\text{df} = (c-1)(r-1), \quad \chi_T^2 = \chi^2_{(\text{df},\, 1-\alpha)}$$
5. **Decision**:
   - If $\chi_c^2 \leq \chi_T^2$: Accept H_0
   - If $\chi_c^2 > \chi_T^2$: Reject H_0

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

Outcome	*A*	*B*
Symptom disappeared	*100*	*220*
Symptom persisted	*40*	*80*
Aggravation	*20*	*70*
Side effects	*30*	*40*

Can we say that both treatments have the same effects? Use $\alpha = 0.05$

Solution

1. ***Choice of test***: *Chi-square test of homogeneity*

2. ***Hypotheses***

   - H_0: *Treatments are equivalent*
   - H_1: *Treatments are not equivalent*

3. ***Expected frequencies (partial):*** *Calculation of Expected Frequencies: Contingency Table*

	Disappearance		Persistence		Worsening		Side Effect		Total
	O_{ij}	T_{ij}	O_{ij}	T_{ij}	O_{ij}	T_{ij}	O_{ij}	T_{ij}	
A	100	101.33	40	38.00	20	28.50	30	22.16	190
B	220	218.66	80	82.00	70	61.50	40	47.83	410
Total	320		120		90		70		600

   All $T_{ij} \geq 5 \rightarrow$ *the test is valid.*

4. ***Chi-square statistic***:

$$\chi^2_c = 7.94$$

5. ***Critical value***:

$$df = (4-1)(2-1) = 3, \quad \chi^2_T = \chi^2_{(3,\ 0.95)} = 7.815$$

6. ***Decision***: *Since* $\chi^2_c > \chi^2_T$, *we reject* H_0

7. ***Conclusion***: *The two treatments do not have the same effect.*

4.3 Independence Test

4.3.1 Objective

4.3.2 Steps of the Test

Study the relationship between two variables in a single sample.

The steps for conducting a Chi-square Test of Independence are as follows:

1. **Formulate Hypotheses**

   - (H_0): The two variables are independent (no association).

   - (H_1): The two variables are dependent (there is an association).

2. **Calculate Expected Frequencies**

   - The expected frequency for each cell in the table is calculated using the formula:

$$T_{ij} = \frac{(row\ total)_i \times (column\ total)_j}{grand\ total}$$

3. **the Chi-square Test Statistic** - The formula for the chi-square statistic is:

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

   - Where:

   - O_{ij} = observed frequency

   - T_{ij} = expected frequency

   - r = number of rows.

   c = number of columns.

   Condition: All $T_{ij} \geq 5$.

4. **Find the Critical Value**

   - Using a chi-square low table, find the critical value (χ_T^2) corresponding to the calculated degrees of freedom and chosen significance level (α) (typically 0.05), and : $df = (r-1)(c-1)$.

5. **Decision Rule**

   - If the test statistic χ_c^2 is greater than the critical value χ_T^2, reject the null hypothesis.

   - If χ_c^2 is less than or equal to χ_T^2, fail to reject the null hypothesis.

6. **Conclusion**

   - Reject H_0: There is a significant association between the two variables (they are dependent).

   - Fail to reject H_1: There is no significant association between the two variables (they are independent).

Example 4.3.1. *A collective food poisoning incident occurred among primary school students. A doctor was assigned to investigate and produced the following table:*

	Sick	***Healthy***
Chocolate Ice Cream	*69*	*83*
No Ice Cream	*31*	*17*

Question: *Is food poisoning related to the consumption of chocolate ice cream?*

Solution

Choice of test*: Chi-square test of independence*

1. ***Hypotheses***

   - H_0*: No association between food poisoning and ice cream consumption*

   - H_1*: There is an association*

2. ***Expected frequencies:***

   All expected $T_{ij} \geq 5 \rightarrow$ *the test is valid.*

	Sick		**Healthy**		**Row Total** n_i
	O_{ij}	T_{ij}	O_{ij}	T_{ij}	
Chocolate Ice Cream	69	76	83	76	152
No Ice Cream	31	24	17	24	48
Column Totals n_j	100	100	100	100	$n = 200$

3. ***Chi-square statistic***:

$$\chi_c^2 = 5.4$$

4. ***Critical value***:

$$df = (2-1)(2-1) = 1, \quad \chi_T^2 = \chi^2_{(1,\ 0.95)} = 3.841$$

5. ***Decision***: *Since* $\chi_c^2 > \chi_T^2$, *we reject* H_0

6. ***Conclusion***: *There is a link between food poisoning and chocolate ice cream consumption.*