

American University of Beirut
STAT 210: Elementary Statistics for Sciences
2022–2023 Fall

Siamak Taati

Chapter 11
Chi-squared tests

Tests on categorical variables

Tests on categorical variables

In the past couple of weeks, we learned how to test hypotheses about

- ▶ mean of a numerical variable in one or two populations
- ▶ proportion of individuals with a certain characteristic (e.g., wear glasses) in one or two population

Tests on categorical variables

In the past couple of weeks, we learned how to test hypotheses about

- ▶ mean of a numerical variable in one or two populations
- ▶ proportion of individuals with a certain characteristic (e.g., wear glasses) in one or two population

The proportion of individuals who wear glasses identifies the distribution of a categorical variable with two possible values

1 (wears glasses)

0 (does not wear glasses)

Tests on categorical variables

In the past couple of weeks, we learned how to test hypotheses about

- ▶ mean of a numerical variable in one or two populations
- ▶ proportion of individuals with a certain characteristic (e.g., wear glasses) in one or two population

The proportion of individuals who wear glasses identifies the distribution of a categorical variable with two possible values

1 (wears glasses)

0 (does not wear glasses)

Now, we are going to learn how to perform two/three other types of tests concerning categorical variables.

Tests on categorical variables

We will consider the following tests:

Tests on categorical variables

We will consider the following tests:

- ① **Goodness-of-fit** tests
 - Ⓚ Does a categorical variable have the claimed distribution?

Tests on categorical variables

We will consider the following tests:

- ① **Goodness-of-fit** tests
 - Ⓚ Does a categorical variable have the claimed distribution?
- ② Tests of **independence**
 - Ⓚ Are two categorical variables independent?

Tests on categorical variables

We will consider the following tests:

- ① **Goodness-of-fit** tests
 - Ⓚ Does a categorical variable have the claimed distribution?
- ② Tests of **independence**
 - Ⓚ Are two categorical variables independent?
- ③ Tests of **homogeneity**
 - Ⓚ Does a categorical variable have the same distribution in two or more populations?

Tests on categorical variables

We will consider the following tests:

① **Goodness-of-fit** tests

Q Does a categorical variable have the claimed distribution?

② Tests of **independence**

Q Are two categorical variables independent?

③ Tests of **homogeneity**

Q Does a categorical variable have the same distribution in two or more populations?

The strategy will be as before:

We use a statistic T and compare

1. The observed value of T ,
2. The distribution of T as suggested by H_0 .

We ask ourselves:

Q Is the observed value of T too extreme for H_0 to be plausible?

Tests on categorical variables

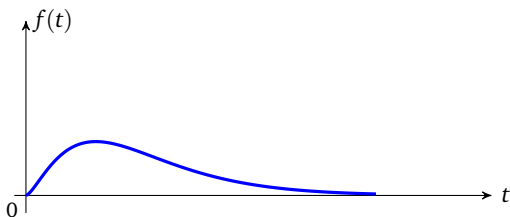
We will consider the following tests:

- ① **Goodness-of-fit** tests
 - Ⓚ Does a categorical variable have the claimed distribution?
- ② Tests of **independence**
 - Ⓚ Are two categorical variables independent?
- ③ Tests of **homogeneity**
 - Ⓚ Does a categorical variable have the same distribution in two or more populations?

It turns out, in the above three scenarios, we can use very similar test statistics which all have the so-called **chi-squared** distribution under the null hypothesis.

The chi-squared distribution

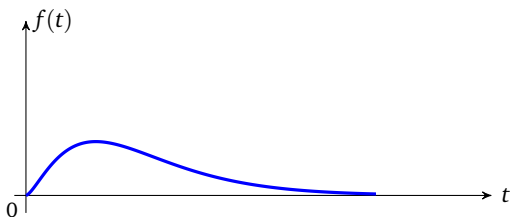
The **chi-squared distribution** with parameter df (# of **degrees of freedom**) is a specific probability density function:



A continuous RV with a chi-squared distribution is called a **chi-squared RV**.

The chi-squared distribution

The **chi-squared distribution** with parameter df (# of **degrees of freedom**) is a specific probability density function:

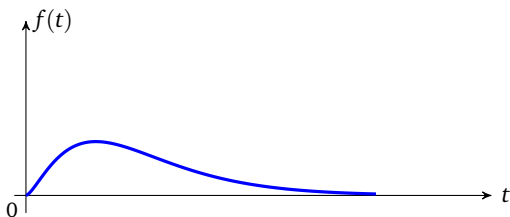


A continuous RV with a chi-squared distribution is called a **chi-squared RV**.

The chi-squared distribution is unimodal and right-skewed. A chi-squared RV can only take non-negative values.

The chi-squared distribution

The **chi-squared distribution** with parameter df (# of **degrees of freedom**) is a specific probability density function:



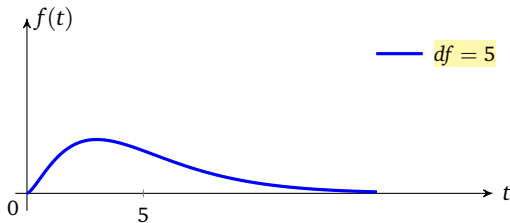
A continuous RV with a chi-squared distribution is called a **chi-squared RV**.

The chi-squared distribution is unimodal and right-skewed. A chi-squared RV can only take non-negative values.

The possible values of a chi-squared RV are all non-negative numbers $0 \leq t < +\infty$.

The chi-squared distribution

The **chi-squared distribution** with parameter df (# of **degrees of freedom**) is a specific probability density function:



Remark 1

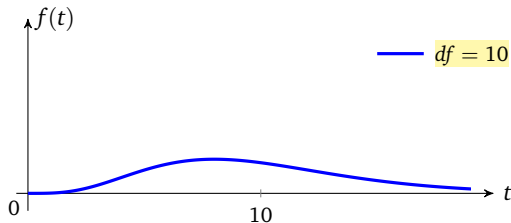
The **larger** df , the more spread is the chi-squared distribution, and the more shifted towards right.

The **mean** of a chi-squared RV is df .

The **standard deviation** of a chi-squared RV is $2 df$.

The chi-squared distribution

The **chi-squared distribution** with parameter df (# of **degrees of freedom**) is a specific probability density function:



Remark 1

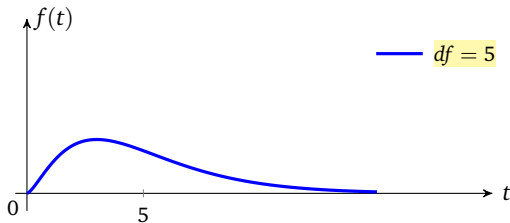
The **larger df** , the more spread is the chi-squared distribution, and the more shifted towards right.

The **mean** of a chi-squared RV is df .

The **standard deviation** of a chi-squared RV is $2 df$.

The chi-squared distribution

The **chi-squared distribution** with parameter df (# of **degrees of freedom**) is a specific probability density function:



Remark 1

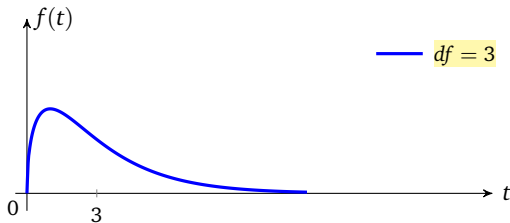
The **larger** df , the more spread is the chi-squared distribution, and the more shifted towards right.

The **mean** of a chi-squared RV is df .

The **standard deviation** of a chi-squared RV is $2 df$.

The chi-squared distribution

The **chi-squared distribution** with parameter df (# of **degrees of freedom**) is a specific probability density function:



Remark 1

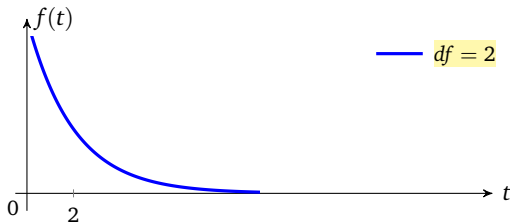
The **larger** df , the more spread is the chi-squared distribution, and the more shifted towards right.

The **mean** of a chi-squared RV is df .

The **standard deviation** of a chi-squared RV is $2 df$.

The chi-squared distribution

The **chi-squared distribution** with parameter df (# of **degrees of freedom**) is a specific probability density function:



Remark 1

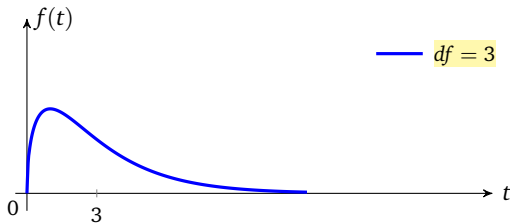
The **larger** df , the more spread is the chi-squared distribution, and the more shifted towards right.

The **mean** of a chi-squared RV is df .

The **standard deviation** of a chi-squared RV is $2 df$.

The chi-squared distribution

The **chi-squared distribution** with parameter df (# of **degrees of freedom**) is a specific probability density function:



Remark 1

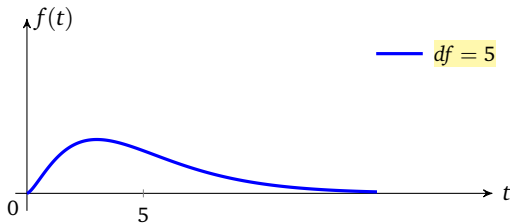
The **larger** df , the more spread is the chi-squared distribution, and the more shifted towards right.

The **mean** of a chi-squared RV is df .

The **standard deviation** of a chi-squared RV is $2 df$.

The chi-squared distribution

The **chi-squared distribution** with parameter df (# of **degrees of freedom**) is a specific probability density function:



Remark 1

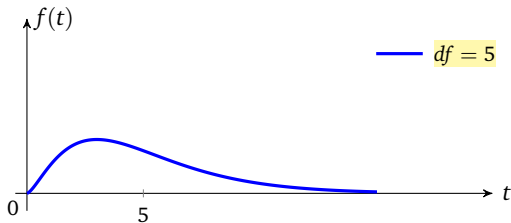
The **larger df** , the more spread is the chi-squared distribution, and the more shifted towards right.

The **mean** of a chi-squared RV is df .

The **standard deviation** of a chi-squared RV is $2 df$.

The chi-squared distribution

The **chi-squared distribution** with parameter df (# of **degrees of freedom**) is a specific probability density function:



Remark 2 (for those curious)

The chi-squared distribution is inter-connected with the normal distribution.

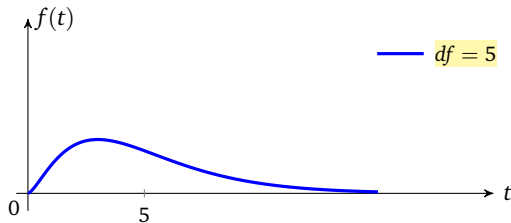
If Z_1, Z_2, Z_3, Z_4, Z_5 are independent standard normal RVs, then

$$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2$$

has the **chi-squared distribution** with $df = 5$ degrees of freedom.

The chi-squared distribution

The **chi-squared distribution** with parameter df (# of **degrees of freedom**) is a specific probability density function:



Remark 2 (for those curious)

The chi-squared distribution is inter-connected with the normal distribution.

If $Z_1, Z_2, \dots, Z_k$ are independent standard normal RVs, then

$$Z_1^2 + Z_2^2 + \dots + Z_k^2$$

has the **chi-squared distribution** with $df = k$ degrees of freedom.

Goodness of fit tests

Example (Blood types in Lebanon)

There are four major human blood types: O, A, B, and AB.

Goodness of fit tests

Example (Blood types in Lebanon)

There are four major human blood types: O, A, B, and AB.

According to a page on Wikipedia, the distribution of the blood types among the Lebanese population is the following:

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

[https://en.wikipedia.org/wiki/Blood_type_distribution_by_country (retrieved 2021-11-25)]

Goodness of fit tests

Example (Blood types in Lebanon)

There are four major human blood types: O, A, B, and AB.

According to a page on Wikipedia, the distribution of the blood types among the Lebanese population is the following:

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

[https://en.wikipedia.org/wiki/Blood_type_distribution_by_country (retrieved 2021-11-25)]

Since the Wikipedia page does not cite its source, researchers at the AUBMC are skeptical about the credibility of the information. They decide to test this by taking a random sample of size 500 from the population of Lebanon comparing the distribution of blood types among them with the distribution claimed by the Wikipedia page.

Goodness of fit tests

Example (Blood types in Lebanon)

There are four major human blood types: O, A, B, and AB.

According to a page on Wikipedia, the distribution of the blood types among the Lebanese population is the following:

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

[https://en.wikipedia.org/wiki/Blood_type_distribution_by_country (retrieved 2021-11-25)]

Q What are the competing hypotheses?

Goodness of fit tests

Example (Blood types in Lebanon)

There are four major human blood types: O, A, B, and AB.

According to a page on Wikipedia, the distribution of the blood types among the Lebanese population is the following:

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

[https://en.wikipedia.org/wiki/Blood_type_distribution_by_country (retrieved 2021-11-25)]

Q What are the competing hypotheses?

- A H_0 The proportions mentioned on Wikipedia match reality.
 H_1 The proportions mentioned on Wikipedia do not match reality.

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

The following table summarizes the observed frequencies from the experiment:

blood type	O	A	B	AB	total
frequency	224	197	61	18	500

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

The following table summarizes the observed frequencies from the experiment:

blood type	O	A	B	AB	total
frequency	224	197	61	18	500

Q What are the expected frequencies for a sample of size 500 according to the null hypothesis?

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

The following table summarizes the observed frequencies from the experiment:

blood type	O	A	B	AB	total
frequency	224	197	61	18	500

Q What are the expected frequencies for a sample of size 500 according to the null hypothesis?

A

blood type	O	A	B	AB	total
frequency	230.5	194.0	56.0	19.5	500

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct.

H_1 The claimed distribution is not correct.

observed frequencies

O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0

O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct.

H_1 The claimed distribution is not correct.

observed frequencies

O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0

O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q What is a suitable test statistic?

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct.

H_1 The claimed distribution is not correct.

observed frequencies

O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0

O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q What is a suitable test statistic?

A The following statistic

$$\chi^2 := \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

measures the overall deviation of the observed frequencies from the expected frequencies according to the null hypothesis.

[χ is the Greek letter *chi*.]

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct.

H_1 The claimed distribution is not correct.

observed frequencies

O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0

O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q What is the observed value of χ^2 ?

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct.

H_1 The claimed distribution is not correct.

observed frequencies

O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0

O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q What is the observed value of χ^2 ?

A We have

$$\chi^2 := \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

observed frequencies				
O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0				
O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q What is the observed value of χ^2 ?

A We have

$$\begin{aligned}\chi^2 &:= \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(224 - 230.5)^2}{230.5} + \frac{(197 - 194.0)^2}{194.0} + \frac{(61 - 56.0)^2}{56.0} + \frac{(18 - 19.5)^2}{19.5}\end{aligned}$$

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

observed frequencies				
O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0				
O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q What is the observed value of χ^2 ?

A We have

$$\begin{aligned}\chi^2 &:= \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(224 - 230.5)^2}{230.5} + \frac{(197 - 194.0)^2}{194.0} + \frac{(61 - 56.0)^2}{56.0} + \frac{(18 - 19.5)^2}{19.5}\end{aligned}$$

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

observed frequencies				
O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0				
O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q What is the observed value of χ^2 ?

A We have

$$\begin{aligned}\chi^2 &:= \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(224 - 230.5)^2}{230.5} + \frac{(197 - 194.0)^2}{194.0} + \frac{(61 - 56.0)^2}{56.0} + \frac{(18 - 19.5)^2}{19.5}\end{aligned}$$

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

observed frequencies				
O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0				
O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q What is the observed value of χ^2 ?

A We have

$$\begin{aligned}\chi^2 &:= \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(224 - 230.5)^2}{230.5} + \frac{(197 - 194.0)^2}{194.0} + \frac{(61 - 56.0)^2}{56.0} + \frac{(18 - 19.5)^2}{19.5}\end{aligned}$$

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

observed frequencies				
O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0				
O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q What is the observed value of χ^2 ?

A We have

$$\begin{aligned}\chi^2 &:= \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(224 - 230.5)^2}{230.5} + \frac{(197 - 194.0)^2}{194.0} + \frac{(61 - 56.0)^2}{56.0} + \frac{(18 - 19.5)^2}{19.5}\end{aligned}$$

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

observed frequencies				
O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0				
O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q What is the observed value of χ^2 ?

A We have

$$\begin{aligned}\chi^2 &:= \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(224 - 230.5)^2}{230.5} + \frac{(197 - 194.0)^2}{194.0} + \frac{(61 - 56.0)^2}{56.0} + \frac{(18 - 19.5)^2}{19.5} \\ &\approx 0.7915021\end{aligned}$$

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

observed frequencies

O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0

O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q What is the distribution of χ^2 assuming H_0 ?

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

observed frequencies

O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0

O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q What is the distribution of χ^2 assuming H_0 ?

A It turns out that χ^2 approximately has the **chi-squared distribution** with $df = 4 - 1 = 3$. [A mathematical theorem]

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

observed frequencies

O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0

O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q What is the distribution of χ^2 assuming H_0 ?

A It turns out that χ^2 approximately has the **chi-squared distribution** with $df = 4 - 1 = 3$. [A mathematical theorem]

Intuitive reason: since the sample is pretty large and the population proportions according to H_0 are not too small, the sample proportions are approximately normal.

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

observed frequencies

O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0

O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q At significance level 5%, how should the researchers conclude?

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

observed frequencies

O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0

O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q At significance level 5%, how should the researchers conclude?

A We have

$$\text{p-value} = \mathbb{P}(\chi^2 \geq 0.7915021 \mid H_0) \approx$$

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

observed frequencies

O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0

O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q At significance level 5%, how should the researchers conclude?

A We have

$$\text{p-value} = \mathbb{P}(\chi^2 \geq 0.7915021 \mid H_0) \approx 0.8515$$

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

observed frequencies

O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0

O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q At significance level 5%, how should the researchers conclude?

A We have

$$\text{p-value} = \mathbb{P}(\chi^2 \geq 0.7915021 \mid H_0) \approx 0.8515$$

Since $\text{p-value} > \alpha$, the researchers cannot reject the null hypothesis.

Goodness of fit tests

Example (Blood types in Lebanon)

population distribution according to Wikipedia

blood type	O	A	B	AB
proportion	0.461	0.388	0.112	0.039

H_0 The claimed distribution is correct. H_1 The claimed distribution is not correct.

observed frequencies

O	A	B	AB	total
224	197	61	18	500

expected frequencies according to H_0

O	A	B	AB	total
230.5	194.0	56.0	19.5	500

Q At significance level 5%, how should the researchers conclude?

A We have

$$\text{p-value} = \mathbb{P}(\chi^2 \geq 0.7915021 \mid H_0) \approx 0.8515$$

Since $\text{p-value} > \alpha$, the researchers cannot reject the null hypothesis.

Conclusion: The evidence against the distribution claimed on Wikipedia is not statistically significant.

Goodness of fit tests

Suppose X is a categorical variable with k possible values.

Goodness of fit tests

Suppose X is a categorical variable with k possible values.

Competing hypotheses

- H_0 The population distribution of X is given by $p_1, p_2, p_3, \dots, p_k$.
- H_1 The population distribution of X is not given by $p_1, p_2, p_3, \dots, p_k$.

Goodness of fit tests

Suppose X is a categorical variable with k possible values.

Competing hypotheses

- H_0 The population distribution of X is given by $p_1, p_2, p_3, \dots, p_k$.
- H_1 The population distribution of X is not given by $p_1, p_2, p_3, \dots, p_k$.

Evidence

The values of X for a sample of size n .

Goodness of fit tests

Suppose X is a categorical variable with k possible values.

Competing hypotheses

(H_0) The population distribution of X is given by $p_1, p_2, p_3, \dots, p_k$.

(H_1) The population distribution of X is not given by $p_1, p_2, p_3, \dots, p_k$.

Evidence

The values of X for a sample of size n .

Test statistic

$$\chi^2 := \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i=1}^k \frac{(O_i - np_i)^2}{np_i}$$

where O_i is the observed frequency of the i -th category within the sample.

Goodness of fit tests

Suppose X is a categorical variable with k possible values.

Competing hypotheses

- H_0 The population distribution of X is given by $p_1, p_2, p_3, \dots, p_k$.
- H_1 The population distribution of X is not given by $p_1, p_2, p_3, \dots, p_k$.

Covered scenario

The **sample is large**, the claimed population proportions are not too small.

[Rule of thumb: $np_i \geq 5$ for each category i]

Goodness of fit tests

Suppose X is a categorical variable with k possible values.

Competing hypotheses

- H_0 The population distribution of X is given by $p_1, p_2, p_3, \dots, p_k$.
- H_1 The population distribution of X is not given by $p_1, p_2, p_3, \dots, p_k$.

Covered scenario

The **sample is large**, the claimed population proportions are not too small. [Rule of thumb: $np_i \geq 5$ for each category i]

Strategy

We compare the observed value of χ^2 against the **chi-squared distribution** with $df = k - 1$. We can follow either the rejection region approach or the p-value approach.

Goodness of fit tests

Suppose X is a categorical variable with k possible values.

Competing hypotheses

H_0 The population distribution of X is given by $p_1, p_2, p_3, \dots, p_k$.

H_1 The population distribution of X is not given by $p_1, p_2, p_3, \dots, p_k$.

Covered scenario

The **sample is large**, the claimed population proportions are not too small. [Rule of thumb: $np_i \geq 5$ for each category i]

Strategy

We compare the observed value of χ^2 against the **chi-squared distribution** with $df = k - 1$. We can follow either the rejection region approach or the p-value approach.

Remark

We always* use a **right-tailed** test.

Tests of independence/homogeneity

Example (Gene and allergy)

Researchers suspect that allergy to a certain type of food is associated to a certain gene. The gene has three alleles A , B and C .

Tests of independence/homogeneity

Example (Gene and allergy)

Researchers suspect that allergy to a certain type of food is associated to a certain gene. The gene has three alleles A , B and C .

To test this, they examine a random sample of 1000 individuals, identify their alleles and test whether they are allergic to that food.

Tests of independence/homogeneity

Example (Gene and allergy)

Researchers suspect that allergy to a certain type of food is associated to a certain gene. The gene has three alleles A , B and C .

To test this, they examine a random sample of 1000 individuals, identify their alleles and test whether they are allergic to that food.

① What are the competing hypotheses?

Tests of independence/homogeneity

Example (Gene and allergy)

Researchers suspect that allergy to a certain type of food is associated to a certain gene. The gene has three alleles A , B and C .

To test this, they examine a random sample of 1000 individuals, identify their alleles and test whether they are allergic to that food.

Q What are the competing hypotheses?

A H_0 The gene and the allergy are independent.

H_1 The gene and the allergy are dependent.

Tests of independence/homogeneity

Example (Gene and allergy)

Researchers suspect that allergy to a certain type of food is associated to a certain gene. The gene has three alleles A , B and C .

To test this, they examine a random sample of 1000 individuals, identify their alleles and test whether they are allergic to that food.

The following table summarizes the gathered data:

	A	B	C	row total
allergic	82	85	53	220
not allergic	332	158	290	780
column total	414	243	343	1000

Tests of independence/homogeneity

Example (Gene and allergy)

Researchers suspect that allergy to a certain type of food is associated to a certain gene. The gene has three alleles A , B and C .

To test this, they examine a random sample of 1000 individuals, identify their alleles and test whether they are allergic to that food.

The following table summarizes the gathered data:

	A	B	C	row total
allergic	82	85	53	220
not allergic	332	158	290	780
column total	414	243	343	1000

This is an example of a **contingency table**. It describes the joint distribution of two variables within a data set.

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

	A	B	C	row total
allergic	82	85	53	220
not allergic	332	158	290	780
column total	414	243	343	1000

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

	A	B	C	row total
allergic	82	85	53	220
not allergic	332	158	290	780
column total	414	243	343	1000

Note that if the two variables are independent (i.e., if H_0 is true), we expect the frequency of sampled cases with gene A and allergy to be roughly

$$1000 \times \frac{414}{1000} \times \frac{220}{1000} = 91.08 .$$

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

	A	B	C	row total
allergic	82	85	53	220
not allergic	332	158	290	780
column total	414	243	343	1000

Note that if the two variables are independent (i.e., if H_0 is true), we expect the frequency of sampled cases with gene A and allergy to be roughly

$$1000 \times \frac{414}{1000} \times \frac{220}{1000} = 91.08 .$$

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

	A	B	C	row total
allergic	82	85	53	220
not allergic	332	158	290	780
column total	414	243	343	1000

Note that if the two variables are independent (i.e., if H_0 is true), we expect the frequency of sampled cases with gene A and allergy to be roughly

$$1000 \times \frac{414}{1000} \times \frac{220}{1000} = 91.08 .$$

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

	A	B	C	row total
allergic	82	85	53	220
not allergic	332	158	290	780
column total	414	243	343	1000

Note that if the two variables are independent (i.e., if H_0 is true), we expect the frequency of sampled cases with gene A and allergy to be roughly

$$1000 \times \frac{414}{1000} \times \frac{220}{1000} = 91.08.$$

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

	A	B	C	row total
allergic	82	85	53	220
not allergic	332	158	290	780
column total	414	243	343	1000

Note that if the two variables are independent (i.e., if H_0 is true), we expect the frequency of sampled cases with gene *A* and allergy to be roughly

$$1000 \times \frac{414}{1000} \times \frac{220}{1000} = 91.08 .$$

Similarly, we expect the frequency of sampled cases with gene *B* and allergy to be roughly

$$1000 \times \frac{243}{1000} \times \frac{220}{1000} = 53.46 .$$

and so on.

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

	A	B	C	row total
allergic	82	85	53	220
not allergic	332	158	290	780
column total	414	243	343	1000

Note that if the two variables are independent (i.e., if H_0 is true), we expect the frequency of sampled cases with gene *A* and allergy to be roughly

$$1000 \times \frac{414}{1000} \times \frac{220}{1000} = 91.08 .$$

Similarly, we expect the frequency of sampled cases with gene *B* and allergy to be roughly

$$1000 \times \frac{243}{1000} \times \frac{220}{1000} = 53.46 .$$

and so on.

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

	A	B	C	row total
allergic	82	85	53	220
not allergic	332	158	290	780
column total	414	243	343	1000

Note that if the two variables are independent (i.e., if H_0 is true), we expect the frequency of sampled cases with gene *A* and allergy to be roughly

$$1000 \times \frac{414}{1000} \times \frac{220}{1000} = 91.08 .$$

Similarly, we expect the frequency of sampled cases with gene *B* and allergy to be roughly

$$1000 \times \frac{243}{1000} \times \frac{220}{1000} = 53.46 .$$

and so on.

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

	A	B	C	row total
allergic	82	85	53	220
not allergic	332	158	290	780
column total	414	243	343	1000

Note that if the two variables are independent (i.e., if H_0 is true), we expect the frequency of sampled cases with gene *A* and allergy to be roughly

$$1000 \times \frac{414}{1000} \times \frac{220}{1000} = 91.08 .$$

Similarly, we expect the frequency of sampled cases with gene *B* and allergy to be roughly

$$1000 \times \frac{243}{1000} \times \frac{220}{1000} = 53.46 .$$

and so on.

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent.

H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

"expected" frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

"expected" frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q What is a suitable test statistic?

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent.

H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

“expected” frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q What is a suitable test statistic?

A As before,

$$\chi^2 := \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

measures the overall deviation of the observed frequencies from the “expected” frequencies according to the null hypothesis.

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

"expected" frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q What is the observed value of χ^2 ?

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

“expected” frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q What is the observed value of χ^2 ?

A We have

$$\chi^2 := \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

"expected" frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q What is the observed value of χ^2 ?

A We have

$$\begin{aligned}\chi^2 &:= \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(82 - 91.08)^2}{91.08} + \frac{(85 - 53.46)^2}{53.6} + \dots + \frac{(290 - 267.54)^2}{267.54}\end{aligned}$$

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

"expected" frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q What is the observed value of χ^2 ?

A We have

$$\begin{aligned}\chi^2 &:= \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(82 - 91.08)^2}{91.08} + \frac{(85 - 53.46)^2}{53.46} + \dots + \frac{(290 - 267.54)^2}{267.54}\end{aligned}$$

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

"expected" frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q What is the observed value of χ^2 ?

A We have

$$\begin{aligned}\chi^2 &:= \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(82 - 91.08)^2}{91.08} + \frac{(85 - 53.46)^2}{53.46} + \dots + \frac{(290 - 267.54)^2}{267.54}\end{aligned}$$

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

"expected" frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q What is the observed value of χ^2 ?

A We have

$$\begin{aligned}\chi^2 &:= \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(82 - 91.08)^2}{91.08} + \frac{(85 - 53.46)^2}{53.6} + \dots + \frac{(290 - 267.54)^2}{267.54}\end{aligned}$$

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

"expected" frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q What is the observed value of χ^2 ?

A We have

$$\begin{aligned}\chi^2 &:= \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(82 - 91.08)^2}{91.08} + \frac{(85 - 53.46)^2}{53.6} + \dots + \frac{(290 - 267.54)^2}{267.54} \\ &\approx \boxed{33.58718}\end{aligned}$$

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

"expected" frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q What is the distribution of χ^2 assuming H_0 ?

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

“expected” frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q What is the distribution of χ^2 assuming H_0 ?

A It turns out that χ^2 approximately has the **chi-squared distribution** with **$df = (2 - 1) \times (3 - 1) = 2$** .

[A mathematical theorem]

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent.

H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

"expected" frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q At significance level 5%, how should the researchers conclude?

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent.

H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

"expected" frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q At significance level 5%, how should the researchers conclude?

A We have

$$\text{p-value} = \mathbb{P}(\chi^2 \geq 33.58718 \mid H_0) \approx$$

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

“expected” frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q At significance level 5%, how should the researchers conclude?

A We have

$$\text{p-value} = \mathbb{P}(\chi^2 \geq 33.58718 \mid H_0) \approx 5.1 \times 10^{-8}$$

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

“expected” frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q At significance level 5%, how should the researchers conclude?

A We have

$$\text{p-value} = \mathbb{P}(\chi^2 \geq 33.58718 \mid H_0) \approx 5.1 \times 10^{-8}$$

Since $\text{p-value} \leq \alpha$, the researchers should reject the null hypothesis.

Tests of independence/homogeneity

Example (Gene and allergy)

H_0 gene and allergy are independent. H_1 gene and allergy are dependent.

observed frequencies

	A	B	C	row tot.
allergic	82	85	53	220
not allergic	332	158	290	780
column tot.	414	243	343	1000

"expected" frequencies according to H_0

	A	B	C	row tot.
allergic	91.08	53.46	75.46	220
not allergic	322.92	189.54	267.54	780
column tot.	414	243	343	1000

Q At significance level 5%, how should the researchers conclude?

A We have

$$\text{p-value} = \mathbb{P}(\chi^2 \geq 33.58718 \mid H_0) \approx 5.1 \times 10^{-8}$$

Since $\text{p-value} \leq \alpha$, the researchers should reject the null hypothesis.

Conclusion: There is a statistically significant evidence (with $\text{p-value} \approx 5.1 \times 10^{-8}$) showing that the gene and the allergy are dependent.

Tests of independence

Suppose X is a categorical variable with k possible values.

Suppose Y is a categorical variable with ℓ possible values.

Tests of independence

Suppose X is a categorical variable with k possible values.

Suppose Y is a categorical variable with ℓ possible values.

Competing hypotheses

H_0 X and Y are independent.

H_1 X and Y are dependent.

Tests of independence

Suppose X is a categorical variable with k possible values.
Suppose Y is a categorical variable with ℓ possible values.

Competing hypotheses

H_0 X and Y are independent.

H_1 X and Y are dependent.

Evidence

The values of X and Y for a sample of size n .

Tests of independence

Suppose X is a categorical variable with k possible values.
Suppose Y is a categorical variable with ℓ possible values.

Competing hypotheses

H_0 X and Y are independent.

H_1 X and Y are dependent.

Evidence

The values of X and Y for a sample of size n .

Test statistic

$$\chi^2 := \sum_{\text{all pairs of values}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Tests of independence

Suppose X is a categorical variable with k possible values.

Suppose Y is a categorical variable with ℓ possible values.

Competing hypotheses

(H_0) X and Y are independent.

(H_1) X and Y are dependent.

Evidence

The values of X and Y for a sample of size n .

Test statistic

$$\chi^2 := \sum_{i,j} \frac{\left(O_{i,j} - n \times \frac{O_{i,-}}{n} \times \frac{O_{-,j}}{n}\right)^2}{n \times \frac{O_{i,-}}{n} \times \frac{O_{-,j}}{n}}$$

where

- $O_{i,j}$ is the observed frequency of the cases with $X = i$ and $Y = j$,
- $O_{i,-}$ is the observed frequency of the cases with $X = i$,
- $O_{-,j}$ is the observed frequency of the cases with $Y = j$.

Tests of independence

Suppose X is a categorical variable with k possible values.

Suppose Y is a categorical variable with ℓ possible values.

Competing hypotheses

H_0 X and Y are independent.

H_1 X and Y are dependent.

Covered scenario

The **sample is large**, the population proportions are not too small.

Tests of independence

Suppose X is a categorical variable with k possible values.

Suppose Y is a categorical variable with ℓ possible values.

Competing hypotheses

H_0 X and Y are independent.

H_1 X and Y are dependent.

Covered scenario

The **sample is large**, the population proportions are not too small.

Strategy

We compare the observed value of χ^2 against the **chi-squared distribution** with $df = (k - 1)(\ell - 1)$. We can follow either the rejection region approach or the p-value approach.

Tests of independence

Suppose X is a categorical variable with k possible values.

Suppose Y is a categorical variable with ℓ possible values.

Competing hypotheses

H_0 X and Y are independent.

H_1 X and Y are dependent.

Covered scenario

The **sample is large**, the population proportions are not too small.

Strategy

We compare the observed value of χ^2 against the **chi-squared distribution** with $df = (k - 1)(\ell - 1)$. We can follow either the rejection region approach or the p-value approach.

Remark

We always* use a **right-tailed** test.

Tests of homogeneity

Suppose X is a categorical variable with k possible values.
We have ℓ distinct populations.

Tests of homogeneity

Suppose X is a categorical variable with k possible values.
We have ℓ distinct populations.

Competing hypotheses

- H_0 The distribution of X is the same among all ℓ populations.
- H_1 The distribution of X is not the same among all ℓ populations.

Tests of homogeneity

Suppose X is a categorical variable with k possible values.
We have ℓ distinct populations.

Competing hypotheses

- H_0 The distribution of X is the same among all ℓ populations.
- H_1 The distribution of X is not the same among all ℓ populations.

Evidence

- ▶ The values of X in a sample of size n_1 from population #1.
- ▶ The values of X in a sample of size n_2 from population #2.
- ▶ ...
- ▶ The values of X in a sample of size n_ℓ from population # ℓ .

The ℓ samples are **independent**.

Tests of homogeneity

Suppose X is a categorical variable with k possible values.
We have ℓ distinct populations.

Competing hypotheses

- H_0 The distribution of X is the same among all ℓ populations.
- H_1 The distribution of X is not the same among all ℓ populations.

Evidence

- ▶ The values of X in a sample of size n_1 from population #1.
- ▶ The values of X in a sample of size n_2 from population #2.
- ▶ ...
- ▶ The values of X in a sample of size n_ℓ from population # ℓ .

The ℓ samples are **independent**.

Strategy

Put together all the samples. Consider the population from which a case comes from as the value of a second categorical variable Y (i.e., $Y = j$ means the case comes from population # j).

Tests of homogeneity

Suppose X is a categorical variable with k possible values.
We have ℓ distinct populations.

Competing hypotheses

- H_0 The distribution of X is the same among all ℓ populations.
- H_1 The distribution of X is not the same among all ℓ populations.

Evidence

- ▶ The values of X in a sample of size n_1 from population #1.
- ▶ The values of X in a sample of size n_2 from population #2.
- ▶ ...
- ▶ The values of X in a sample of size n_ℓ from population # ℓ .

The ℓ samples are **independent**.

Strategy

Put together all the samples. Consider the population from which a case comes from as the value of a second categorical variable Y (i.e., $Y = j$ means the case comes from population # j). The test becomes equivalent to testing whether X and Y are independent or dependent.