Inference

two-way tables (chi squared test)

Prof. Dr. Jan Kirenz
HdM Stuttgart

Randomization test of independence

Used iPod experiment:

n = 219

- **General**: What can you tell me about it?
- **Positive** Assumption: It does not have any problems, does it?
- **Negative** Assumption: What problems does it have?

Question	Disclose problem	Hide problem	Total
General	2	71	73
Positive assumption	23	50	73
Negative assumption	36	37	73
Total	61	158	219

Table 18.1: Summary of the iPod study, where a question was posed to the study participant who acted.

Question	Disclose problem	Hide problem	Total
General			73
Positive assumption			73
Negative assumption			73
Total	61	158	**219**

61/219 = 27.85%

158/219 = 72.15%

	Disclose problem	Hide problem	Total
General	2 *(20.33)*	71 *(52.67)*	73
Positive assumption	23 *(20.33)*	50 *(52.67)*	73
Negative assumption	36 *(20.33)*	37 *(52.67)*	73
Total	61	158	219

Table 18.2: The observed counts and the expected counts for the iPod experiment.

$$0.2785 \times (\text{row 1 total}) = 20.33$$
$$0.2785 \times (\text{row 2 total}) = 20.33$$
$$0.2785 \times (\text{row 3 total}) = 20.33$$

$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 1 total}) = 20.33$$
$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 2 total}) = 20.33$$
$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 3 total}) = 20.33$$

Computing expected counts in a two-way table.

To calculate the expected count for the i^{th} row and j^{th} column, compute

$$\text{Expected Count}_{\text{row } i, \text{ col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

The observed **chi-squared** statistic

General formula $$\frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

	Disclose problem	Hide problem	Total
General	2 *(20.33)*	71 *(52.67)*	73
Positive assumption	23 *(20.33)*	50 *(52.67)*	73
Negative assumption	36 *(20.33)*	37 *(52.67)*	73
Total	61	158	219

Table 18.2: The observed counts and the expected counts for the iPod experiment.

The observed **chi-squared** statistic

General formula	$\dfrac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$
Row 1, Col 1	$\dfrac{(2-20.33)^2}{20.33} = 16.53$
Row 2, Col 1	$\dfrac{(23-20.33)^2}{20.33} = 0.35$
$\vdots$	$\vdots$
Row 2, Col 3	$\dfrac{(37-52.67)^2}{52.67} = 4.66$

Adding the computed value for each cell gives the chi-squared test statistic X^2 :

$$X^2 = 16.53 + 0.35 + \cdots + 4.66 = 40.13$$

	Disclose problem	Hide problem	Total
General	2 *(20.33)*	71 *(52.67)*	73
Positive assumption	23 *(20.33)*	50 *(52.67)*	73
Negative assumption	36 *(20.33)*	37 *(52.67)*	73
Total	61	158	219

Table 18.2: The observed counts and the expected counts for the iPod experiment.

Randomization: Variability of the statistic

Question	Disclose problem	Hide problem	Total
General	29	44	73
Positive assumption	15	58	73
Negative assumption	17	56	73
Total	61	158	219

Table 18.3: Summary of the iPod study, where a question was posed to the study participant who acted.

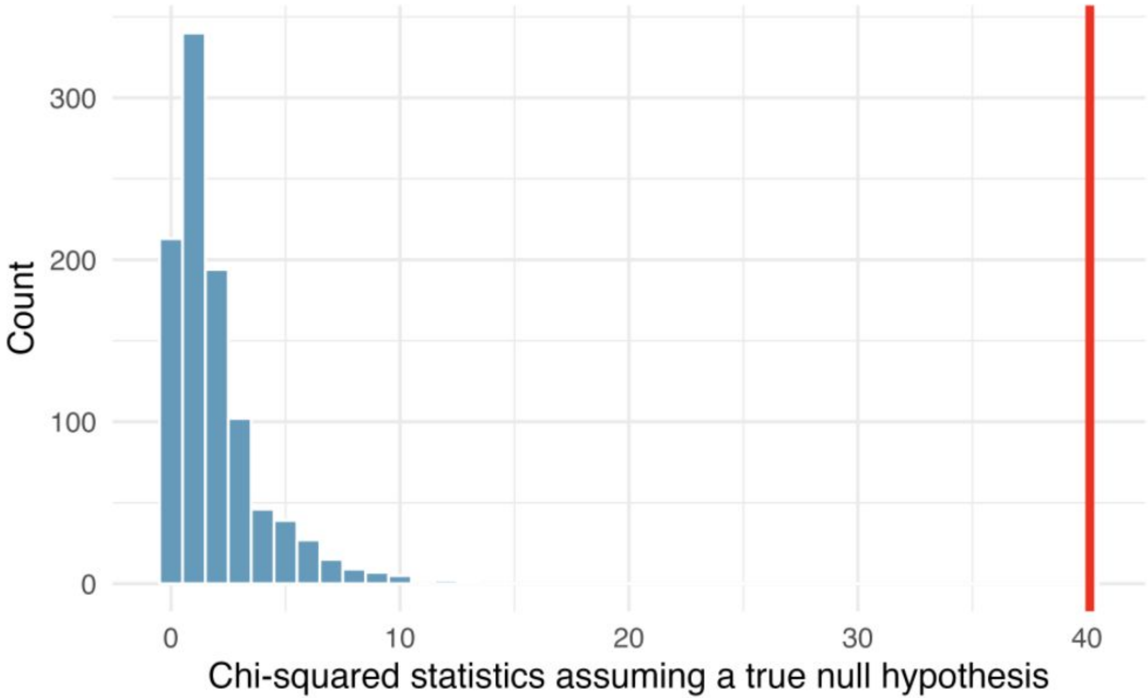
Table 18.3 shows a possible randomization of the observed data under the condition that the null hypothesis is true (in contrast to the original observed data in Table 18.1

$$
\begin{array}{ll}
\text{General formula} & \dfrac{(\text{observed count } - \text{expected count})^2}{\text{expected count}} \\
\text{Row 1, Col 1} & \dfrac{(29-20.33)^2}{20.33} = 3.7 \\
\text{Row 2, Col 1} & \dfrac{(15-20.33)^2}{20.33} = 1.4 \\
\vdots & \vdots \\
\text{Row 3, Col 2} & \dfrac{(56-52.67)^2}{52.67} = 0.211
\end{array}
$$

Adding the computed value for each cell gives the chi-squared test statistic X^2 :

$$X^2 = 3.7 + 1.4 + \cdots + 0.211 = 8$$

Observed statistic vs. null chi-squared statistics

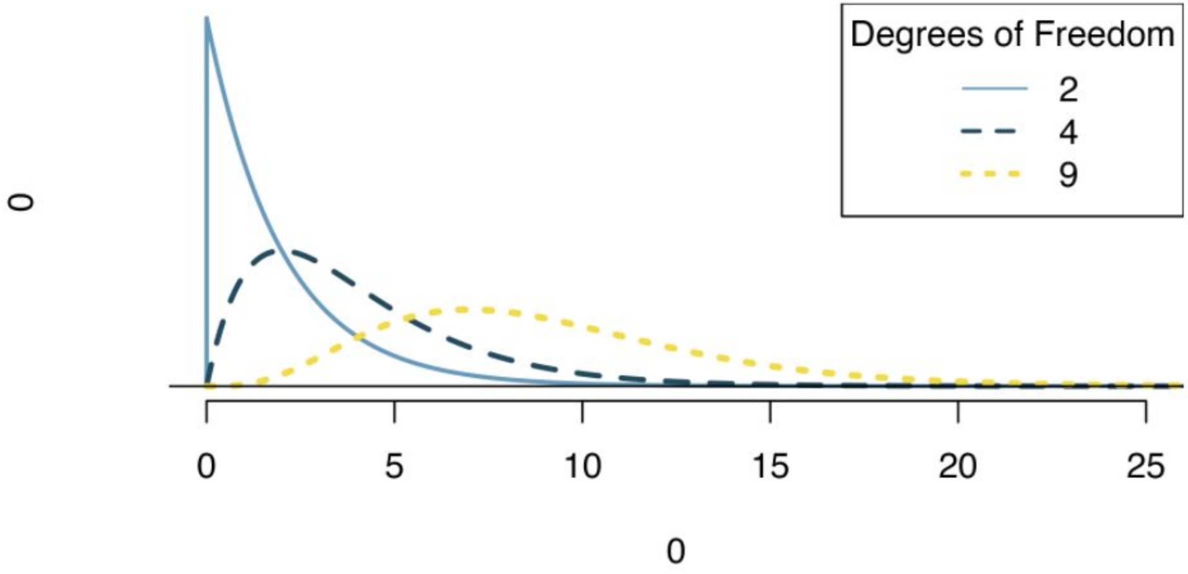


Note that with a chi-squared test, we only know that the two variables (question_class and response) are related (i.e., not independent).

We are not able to claim which type of question causes which type of response.

Mathematical model for test of independence

The chi-squared test of independence



The chi-squared distribution for differing degrees of freedom.

Variability of the chi-squared statistic

- The chi-squared test statistic follows a Chi-squared distribution when the null hypothesis is true.

- For two way tables, the degrees of freedom is equal to:
- df= (number of rows minus 1) × (number of columns minus 1)
- In our example, the degrees of freedom parameter is df = (2-1) * (3-1) = 2

The test statistic for assessing the independence between two categorical variables is a X^2.

The X^2 statistic is a ratio of how the observed counts vary from the expected counts as compared to the expected counts (which are a measure of how large the sample size is).

$$X^2 = \sum_{i,j} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

When the null hypothesis is true and the conditions are met, X^2 has a Chi-squared distribution with $df = (r-1) \times (c-1)$.

Conditions:

- Independent observations
- Large samples: 5 expected counts in each cell

Computing degrees of freedom for a two-way table.

When applying the chi-squared test to a two-way table, we use $df = (R-1) \times (C-1)$ where R is the number of rows in the table and C is the number of columns.

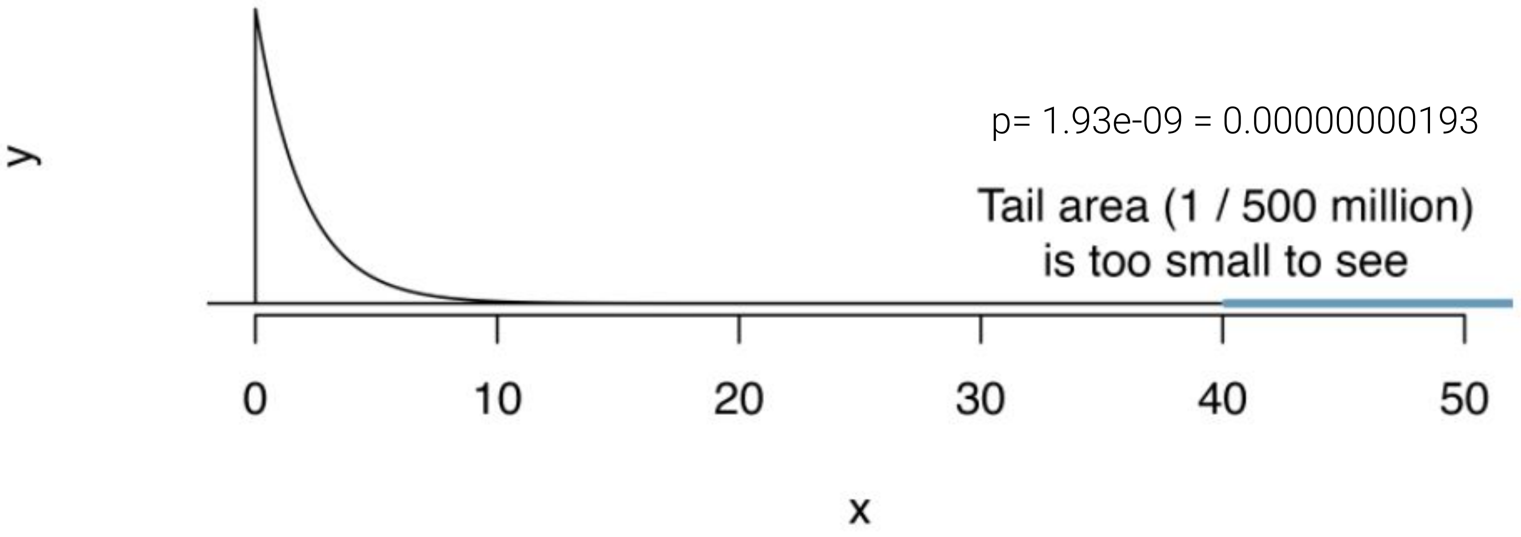


Figure 18.3: Visualization of the p-value for $X^2 = 40.13$ when $df = 2$.

Terms you should know

Chi-squared distribution	expected counts
chi-squared statistic	independence