Inference for Two-way tables (chi squared test)

Randomization

Prof. Dr. Jan Kirenz

Sell a used iPod - get as much money as possible

You participate in an experiment: sell a used iPod (with defects)

You get 5% cut of the sale on top of $10 for participating

(n = 219)

You get one of three questions from a potential buyer:

- **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***

What is the **actual disclosure** per question?

If the questions makes no difference (**H_0**):

We would expect that **27.85%** of sellers will **disclose** the problem (regardless of the question)

27.85% disclosed the problem ← **61/219 = 27.85%**

158/219 = 72.15%

Question	**Disclose problem**		***Total***
General	2 (20.33)	**27.85%** x **73**	***73***
Positive assumption	23 (20.33)	**27.85%** x **73**	***73***
Negative assumption	36 (20.33)	**27.85%** x **73**	***73***
Total	***61***		***219***

61/219 = 27.85%

What is the **expected disclosure** per question?

If the questions makes no difference (**H_0**):

We would expect that **27.85%** of sellers will **disclose** the problem (regardless of the question)

Question	Disclose problem		*Total*
General	2 (20.33)	27.85% x **73**	***73***
Positive assumption	23 (20.33)	27.85% x **73**	***73***
Negative assumption	36 (20.33)	27.85% x **73**	***73***
Total	***61***		***219***

61/219 = 27.85%

What is the **expected disclosure** per question?

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

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

Question	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***

What is the **expected** amount of participants who **hide** the problem (per question)?

61/219 = 27.85%

158/219 = 72.15%

Computing expected counts in a two-way table.

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

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

Question	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***

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

$$\begin{array}{ll} \text{Row 1, Col 1} & \dfrac{(2-20.33)^2}{20.33} = 16.53 \\ \text{Row 2, Col 1} & \dfrac{(23-20.33)^2}{20.33} = 0.35 \\ \vdots & \vdots \\ \text{Row 2, Col 3} & \dfrac{(37-52.67)^2}{52.67} = 4.66 \end{array}$$

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

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

Randomization

Variability of the statistic

Null hypothesis:

individuals will disclose or hide the problems **regardless** of the **question** they are given

We can **randomize** the data by reassigning the **61 disclosed** problems and **158 hidden** problems to the three groups at random

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

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

Shows a possible randomization of the observed data under the condition that the null hypothesis is true

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

Row 1, Col 1 $\dfrac{(29 - 20.33)^2}{20.33} = 3.7$

Row 2, Col 1 $\dfrac{(15 - 20.33)^2}{20.33} = 1.4$

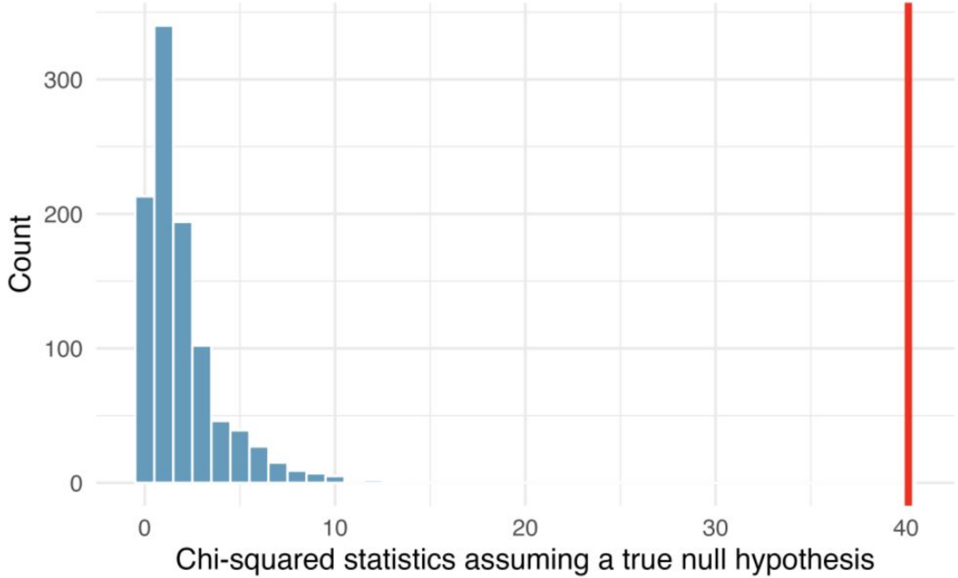
$\vdots$ $\vdots$

Row 3, Col 2 $\dfrac{(56 - 52.67)^2}{52.67} = 0.211$

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

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

1,000 chi-squared statistics generated under the null

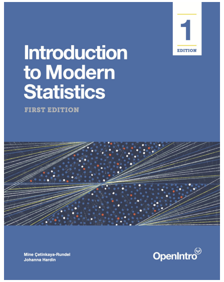


We can see that the observed value is so far from the null statistics that the simulated **p-value is zero**.

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.

Resources



The content of this presentation is mainly based on the excellent book “Introduction to Modern Statistics” by Mine Çetinkaya-Rundel and Johanna Hardin (2021).

The online version of the book can be accessed for free:

https://openintro-ims.netlify.app/index.html