Module 4. Bivariate analysis: qualitative variables

Data Science

Sabine De Vreese Lieven Smits Pieter Van Der Helst Bert Van Vreckem

2024–2025

HO
GENT

Contents

HO
GENT

Learning goals

- Dependent/independent variable
- Apply suitable analysis techniques for each combination of measurement levels
- Contingency tables and Cramér's *V*
- Visualization

HO
GENT

Bivariate analysis: overview

Independent	Dependent	Test/Metric
Qualitative	Qualitative	χ^2-test Cramér's V
Qualitative	Quantitative	two-sample t-test Cohen's d
Quantitative	Quantitative	— Regression, correlation

HO GENT

Bivariate analysis

HO
GENT

Bivariate analysis

- ...is determining whether there is an association between two stochastic variables (X and Y).
- **Association** = you can predict (to some extent) the value of Y from the value of X
  - X — Independent variable
  - Y — Dependent variable
- **Important!** Finding an association does **NOT** imply a causal relation!

Example research questions

- Is there a difference in taste preference between two beverage brands?
- Is there a difference in spending at the campus restaurant between students and staff?
- Do smokers die more often of lung cancer than non-smokers?
- Do men and women have a different opinion on a survey question?
- …

We will use `data/rlanders.csv` from the Github repo for lab assignments to explore the last question.

HO
GENT

Contingency tables

Contingency tables

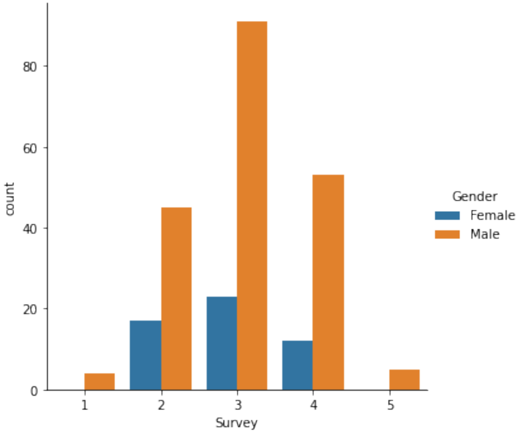
(also: crosstab)

See Python example code in `demo-4.01-chi-squared.ipynb`

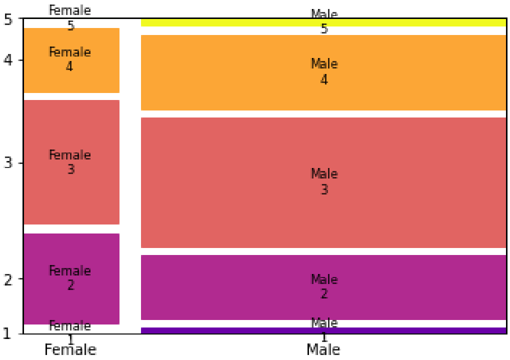
Gender Survey	Female	Male
Strongly disagree	0	4
Disagree	17	45
Neutral	23	91
Agree	12	53
Strongly agree	0	5

HO
GENT

Visualization



A clustered bar chart



A mosaic plot

Marginal totals

Gender Survey	Female	Male	Total
Strongly disagree	0	4	4
Disagree	17	45	62
Neutral	23	91	114
Agree	12	53	65
Strongly agree	0	5	5
Total	52	198	250

HO GENT

Expected values

If there is no difference (association), we expect the same ratios in each column of the table!

Gender Survey	Female	Male	Total
Strongly disagree	0.832	3.168	4
Disagree	12.896	49.104	62
Neutral	23.712	90.288	114
Agree	13.520	51.480	65
Strongly agree	1.040	3.960	5
Total	52	198	250

In each cell: (row total × column total) / *n*

Measuring dispersion

How far is the observed value o from the expected e?

$$\frac{(o - e)^2}{e}$$

Gender Survey	Female	Male
Strongly disagree	0.832	0.219
Disagree	1.306	0.343
Neutral	0.021	0.006
Agree	0.171	0.045
Strongly agree	1.040	0.273

HO GENT

The chi-squared statistic

The sum of all these values is notated:

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i} \approx 4.255$$

- χ is the Greek letter *chi*
- o_i = number of observations in the i'th cell of the contingency table
- e_i = expected frequency
- Small value ⇒ no association
- Large value ⇒ association

HO
GENT

When is χ^2 large enough?

- 2×2-table with $\chi^2 = 10$
  - Relatively large difference
  - Indicates association
- 5×5-table with $\chi^2 = 10$
  - Relatively small difference
  - Does NOT indicate association

We need a metric independent of table size!

HO
GENT

Cramér's V

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} = \sqrt{\frac{4.255}{250(2-1)}} \approx 0.130$$

with n the number of observations, $k = \min(numRows, numCols)$

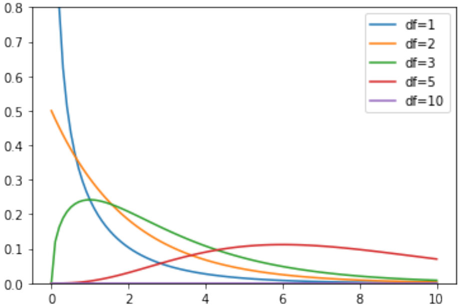
Cramér's V	Interpretation
≈ 0	no association
≈ 0.1	weak association
≈ 0.25	moderate association
≈ 0.5	strong association
≈ 0.75	very strong association
≈ 1	complete association

HO
GENT

Chi-squared test for independence

χ^2 test for independence

- = Alternative to Cramér's V to investigate association between qualitative variables.
- Value of χ^2 distributed according to the χ^2 distribution



HO GENT

χ^2-distribution in Python

Import `scipy.stats`
For a χ^2-distribution with `df` degrees of freedom:

Function	**Purpose**
`chi2.pdf(x, df=d)`	Probability density for `x`
`chi2.cdf(x, df=d)`	Left-tail probability $P(X <$ `x`$)$
`chi2.sf(x, df=d)`	Right-tail probability $P(X >$ `x`$)$
`chi2.isf(1-p, df=d)`	`p`% of observations is expected to be lower than this value

HO
GENT

Test procedure

- **Step 1.** Formulate hypotheses:
  - H_0: there is no association (χ^2 is "small")
  - H_1: there is an association (χ^2 is "large")
- **Step 2.** Choose significance level, e.g. $\alpha = 0.05$
- **Step 3.** Calculate the test statistic, $\chi^2 = 4.255$

HO
GENT

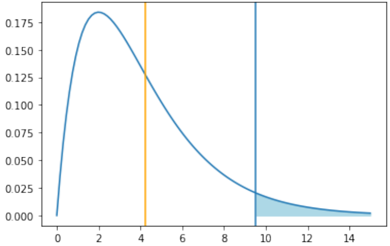
Test procedure (cont.)

- **Step 4.** Use $df = (numRow - 1) \times (numCol - 1)$ and either:
    - Determine critical value g so $P(\chi^2 > g) = \alpha$
    - Calculate the p-value
- **Step 5.** Draw conclusion:
    - $\chi^2 < g$: do not reject H_0; $\chi^2 > g$: reject H_0
    - $p > \alpha$: do not reject H_0; $p < \alpha$: reject H_0

HO
GENT

Example (Gender vs Survey)

- `g = stats.chi2.isf(0.05, df=4)` (result: 9.4877)
- `p = stats.chi2.sf(4.2555, df=4)` (result: 0.3725)



Conclusion: we do not reject the null hypothesis. Differences between expected and observed values are not significantly large.
We found no association between variables *Gender* and *Survey*

Test for independence in Python

SciPy has a function to calculate χ^2 and *p*-value from a contingency table:

```
observed = pd.crosstab(rlanders.Survey, rlanders.Gender)
chi2, p, df, expected = stats.chi2_contingency(observed)

print(”Chi-squared       : %.4f” % chi2)
print(”Degrees of freedom: %d” % df)
print(”P-value           : %.4f” % p)
```

Goodness-of-fit test

HO
GENT

Goodness-of-fit test

The χ^2 test can also be used to determine whether a sample is **representative** for the population.

Goodness-of-fit test

This test indicates to what degree a sample corresponds to a null hypothesis regarding the distribution of a qualitative variable over mutually exclusive classes.

Example

Type	# sample	# population
Mutant	127	35%
Human	75	17%
Alien	98	23%
God	27	8%
Demon	73	17%
Total	400	100%

HO
GENT

Example

Is the distribution of the sample (n = 400) representative for the full population (all superheroes)?

- What numbers would you *expect* if the sample is representative?
- How large are the differences from the *observed* numbers?
  - small ⇒ distribution is representative
  - large ⇒ distribution is **not** representative

HO
GENT

Example

Is the distribution of the sample (n = 400) representative for the full population (all superheroes)?

- What numbers would you *expect* if the sample is representative?
- How large are the differences from the *observed* numbers?
  - small ⇒ distribution is representative
  - large ⇒ distribution is **not** representative

Can you see the link with contingency tables and Cramer's V?

HO
GENT

Goodness of fit test

- Exactly representative ⇒ 35% of superheroes in the sample is a mutant
- The expected number therefore is $e = 0.35 \times 400 = 140$.

Therefore:

$$e = n \times \pi$$

If the differences $o - e$ are relatively small they can be attributed to random sampling errors.

Goodness of fit test

Consider χ^2:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

Draw a conclusion based on the value of χ^2:

- small ⇒ distribution is representative
- large ⇒ distribution is **not** representative

χ^2 measures the degree of conflict with the null hypothesis

HO GENT

Goodness of fit test

Superhero type	o	π	e	$\frac{(o-e)^2}{e}$
Mutant	127	35%	140	1.21
Human	75	17%	68	0.72
Alien	98	23%	92	0.39
God	27	8%	32	0.78
Demon	73	17%	68	0.37
			χ^2 =	3.47

HO
GENT

Goodness of fit test

- The test statistic χ^2 follows the χ^2 distribution.
- Critical value g from the χ^2 distribution: this is dependent on the number of degrees of freedom (df). In general:

  $$df = k - 1$$

  with k the number of categories.
- The critical value g for a given significance level α and number of degrees of freedom df can be calculated in Python using the function `isf()`.

$$P(\chi^2 < g) = 1 - \alpha$$

HO
GENT

Goodness of fit test

Testing Procedure

1. **Formulate hypotheses**
   - H_0: sample is representative for the population
   - H_1: sample is not representative for the population
2. **Choose significance level**: $\alpha = 0.05$

HO
GENT

Goodness of fit test

Testing Procedure

1. **Calculate test statistic**:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

   1.1 **Critical area**: Calculate g so that $P(\chi^2 < g) = 1 - \alpha$
   1.2 **Probability value**: Calculate $p = 1 - P(X < \chi^2)$

2. Conclusion (the test is always right-tailed):
   2.1 $\chi^2 < g \Rightarrow$ do not reject H_0, $\chi^2 > g \Rightarrow$ reject H_0
   2.2 $p > \alpha \Rightarrow$ do not reject H_0, $p < \alpha \Rightarrow$ reject H_0

HO
GENT

Example

- `g = stats.chi2.isf(0.05, df=4)` (result: 9.4877)
- `p = stats.chi2.sf(3.4679, df=4)` (result: 0.4828)

Conclusion. $\chi^2 \approx 3.47 < g \approx 9.4877$, so we don't reject the null hypothesis. This sample is representative for the population.

HO
GENT

Goodness-of-fit test in Python

```python
observed   = np.array([127,  75,  98,  27, 73])
expected_p = np.array([.35, .17, .23, .08, .17])
expected   = expected_p * sum(observed)
chi2, p = stats.chisquare(f_obs=observed, f_exp=expected)

print("χ² = %.4f" % chi2)
print("p  = %.4f" % p)
```

HO
GENT

Standardized residuals

Example: families

Consider all families with exactly 5 children in a given community.

Example: families

Consider all families with exactly 5 children in a given community. When we look at the number of boys/girls, there are 6 possible combinations:

1. 5 boys
2. 4 boys, 1 girl
3. 3 boys, 2 girls
4. 2 boys, 3 girls
5. 1 boy, 4 girls
6. 5 girls

A survey was conducted regarding 1022 families with exactly 5 children

Are the observed numbers in the 6 classes representative for a population in which the probability of having a boy is equal to the probability of having a girl, or more concrete 0.5?

Example

i	0	1	2	3	4	5
o_i	58	149	305	303	162	45

HO
GENT

Example

i	0	1	2	3	4	5
o_i	58	149	305	303	162	45

If the assumption is true, the probability π_i to have i boys is determined by a binomial distribution with parameters $n = 5$ and $p = 0.5$. For example, the probability to have 2 boys out of 5 children is equal to:

$$(0.5)^2 \times (1 - 0.5)^{5-2} \times \binom{5}{2}$$

In general (you don't have to know why):

$$\pi_i = \binom{5}{i} \times 0.5^i \times 0.5^{5-i} = \frac{5!}{i!(5-i)!} \times 0.5^i$$

HO
GENT

Example

i	0	1	2	3	4	5	Total
o_i	58	149	305	303	162	45	1022
π_i	0,031	0,156	0,313	0,313	0,156	0,031	1
e_i	31,9	159,7	319,4	319,4	159,7	31,9	1022
$\frac{(o-e)^2}{e}$	21,268	0,715	0,647	0,840	0,033	5,343	28,846
r_i	4,686	-0,921	-0,970	-1,105	0,199	2,348	

HO GENT

Example

1. **Formulate both hypotheses**
   - H_0: the sample is representative for the population
   - H_1: the sample is not representative for the population
2. **Determine α and n**: $\alpha = 0.01$ and $n = 1022$.
3. **Value of the test statistic in the sample**:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \approx 28.846$$

4. **Calculate and plot critical area**: The critical value is 15.0863. Our test statistic is inside the critical area, so we can reject H_0.

HO GENT

Standardized Residuals

Standardized Residuals

The standardized residuals indicate which classes make the greatest contribution to the value of χ^2.

$$r_i = \frac{o_i - n\pi_i}{\sqrt{n\pi_i(1 - \pi_i)}}$$

- $r_i \in [-2, 2] \Rightarrow$ “acceptable” values
- $r_i < -2 \Rightarrow$ underrepresented
- $r_i > 2 \Rightarrow$ overrepresented

Conclusion: families in which all children are of the same gender are overrepresented.

Cochran’s rules

Conditions

In order to apply the χ^2-test, the following conditions must be met (Rule of Cochran)

1. For all categories, the expected frequency e must be greater than 1.
2. In a maximum of 20 % of the categories, the expected frequency e may be less than 5.

HO
GENT