Recap: bivariate statistics

How to run bivariate analyses in R

UZH
IKMZ

1 Univariate statistics

1.1 Univariate recap

UZH
IKMZ

Measures	Statistics	Definition	Variable type numeric	Variable type categorical (nominal)
distribution	*frequency*			x
	proportion			x
central tendency	*mode*	*most frequent modality of a variable: not necessarily unique, not necessarily the center*	x	x
	mean	*average (arithmetic): not robust to extreme observations*	x	
	median	*50% of data is smaller and 50% of data is larger than the median: robust*	x	(if ordinal)
dispersion	*min*	*nota bene: distinction between min./max.*	x	
	max	*theoretical vs.actually observed*	x	
	quartiles	*complete the median and divide the data in 4 groups (25%)*	x	
	variance	*mean of the sum of the squares of the deviations from the mean: can vary from 0 to infinity*	x	
	standard deviation	*square root of the variance: expressed in the same unit as the data observed (expected distance between observation of the random sample and the sample mean)*	x	

Univariate statistics

UZH
IKMZ

1.2 Fiability and validity

The Total Survey Error (TSE) framework (Biemer, 2010) accounts for the different sources of errors that occur at each stage of an investigation (e.g. coverage errors, selection errors, and measurement errors).

Two important concepts are: reliability and validity.

Reliability refers to the idea of replicability. It accounts for the degree of consistency of a measurement (by different observers, at different times).

Validity addresses the conclusions we can draw from of a measure. For instance, internal validity aims at measuring the fit between the concept and the measure.

1.3 Sample statistics

Sample statistics are estimators of population parameters. A particular point estimate cannot be expected to be exactly equal to the value of the parameter in the population. Therefore, the estimate always contains a margin of error.

Normal law distribution (or Gaussian curve): suggests that the distribution of a variable is around an average value and the other values increase and decrease in a homogeneous and symmetrical way around this average value.

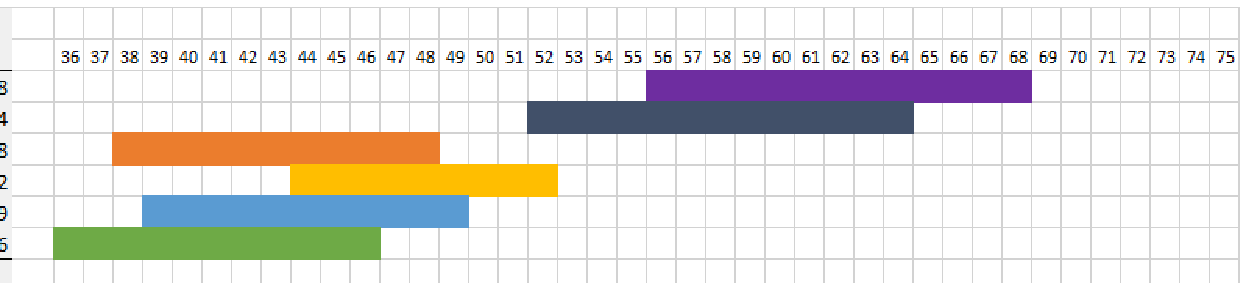
1.4 Confidence intervals

The margin of error is also called the confidence interval. It corresponds to the area where we know for a given probability that the average or the percentage of a value will be found.

For a mean: $\overline{x} \pm 1.96(\frac{\sigma(X)}{\sqrt{n}})$

For a proportion: $Z_{\alpha}\sqrt{\frac{p(1-p)}{n}}$

		Vote (%)					
age groups	#observations	yes	no	(p-(1-p))/n	standard dev.	LB	UB
18-29	288	0,62	0,38	0,0008	5,61	0,56	0,68
30-39	271	0,58	0,42	0,0009	5,88	0,52	0,64
40-49	338	0,43	0,57	0,0007	5,28	0,38	0,48
50-59	511	0,48	0,52	0,0005	4,33	0,44	0,52
60-69	413	0,44	0,56	0,0006	4,79	0,39	0,49
70+	443	0,41	0,59	0,0005	4,58	0,36	0,46
Nbr. of observ.	2264						



Confidence intervals for proportions

2 Bivariate statistics

2.1 Bivariate analyses and tests (recap)

		Independent variable		
		nominal	ordinal	interval
dependent variable	nominal ordinal	Crosstab		recode to ordinal
	interval	Mean comparison / variance analysis		Correlation

Types of bivariate analyses

Tool	Significance	Strengh
Crosstab	Chi-2 test, p-value	Cramer's V
Mean comparison / variance analysis	F test, p-value / t test, p-value	Eta
Correlation	t test, p-value	Pearson's r

Types of bivariate tests

2.2 Cross-tables: relationship between two categorical variables

- H0 (null hypothesis) states that there is no relationship
  - The Chi-2 distribution table enables us to assesses this relationship with the critical value based on:
  - Degrees of freedom (df = (row-1)*(col-1)) used to read the
  - p-value: typical threshold <0.05 (Chi-2 test should be greater that the critical value to reject H0)

2.3 Cross-tables: calculat Chi-2

For each cell of the table, we have to calculate the expected value under null hypothesis. For a given cell, the expected value is calculated as follow:

$$e = \frac{row.\,sum * col.\,sum}{grand.\,total}$$

The Chi-square statistic is calculated as follow:

$$Chi2 = \frac{\sum (o-e)^2}{e}$$

UZH
IKMZ

2.4 Cross-tables: Cramer’s V

- Cramer’s V: assesses the strength of the relation:

$$V = \sqrt{\frac{(Chi2/n)}{min(col-1, row-1)}}$$

- Chi2: Chi-square statistic
- n: total sample size
- r: number of rows
- c: Number of columns

UZH
IKMZ

2.5 Chi-2 manual example

Observed

	Passed the exam	Failed the exam	total
Participate	25	6	31
Did not participate	8	15	23
total	33	21	54

Expected: what frequencies would we expect if there was no relationship between the two variables?

	Passed the exam	Failed the exam	total
Participate	18,94	12,06	31
Did not participate	14,06	8,94	23
total	33	21	54

Chi-2

	Passed the exam	Failed the exam	total
Participate	1,94	3,04	4,98
Did not participate	2,61	4,10	6,71
total	4,54	7,14	11,7

Chi2:	11,69
df: (l-1)(c-1)	1
test for $p<0.05$:	3,84
result:	H0 rejected
	>>> the relationship between the two variables is statistically significant
Cramer's V:	0,47

Chi-2 test and Cramer's V

2.6 Chi-2 example in R

```r
data = matrix(c(7,9,12,8), nrow = 2)
data
```

```
     [,1] [,2]
[1,]    7   12
[2,]    9    8
```

```r
chisq.test(data)
```

```

	Pearson's Chi-squared test with Yates' continuity correction

data:  data
X-squared = 0.40263, df = 1, p-value = 0.5257
```

```r
rcompanion::cramerV(data)
```

```
Cramer V 
  0.1617 
```

2.7 Comparing the means of two groups

General logic: Compare the distribution of values for a quantitative variable across different groups (different categories of the categorical independent variable).

- First step: Examine the relationship between the two variables to see if they are related or independent.
  - Variation between groups: Are the different groups distinct? (central tendency: means)
  - Variation within groups: Are the different groups homogeneous? (dispersion: standard deviation)
- Second step: Determine the statistical significance of the relationship to see if it can be generalized to the population.
  - H0: The two variables are independent.
  - H1: The two variables are dependent.
  - If $p < 0.05$: Reject H0, indicating a statistically significant relationship.

2.8 t-test

- Independent Samples: Used when comparing means from two different groups.
- Paired Samples: Used when comparing means from the same group at different times or under different conditions.

Assumptions (same for ANOVA):

- Normality: Data in each group should be approximately normally distributed.
- Homogeneity of Variance: Variances in the two groups should be approximately equal.
- Independence: Observations within and between groups should be independent.

$$t = \frac{\hat{x_1} - \hat{x_2}}{\sqrt{(\frac{s_1^2}{n1}) + (\frac{s_2^2}{n_2})}}$$

where $\hat{x_1}$ and $\hat{x_2}$ are the sample means, s_1^2 and s_2^2 the sample variances, and n_1 and n_2 the sample sizes.

2.9 ANOVA: Comparing the means of more than two groups

- Unlike t-tests ANOVA can handle multiple groups simultaneously.
- It reduces the risk of Type I errors (false positives) that can occur when performing multiple t-tests.

Key Concepts:

- variance **within** samples ($S2_{within}$) reflects individual differences and error variance.
- variance **between** sample means ($S2_{between}$) measures the variation among the group means.
- F-statistic: ratio of $S2_{between}/S2_{within}$
- Eta: indicates the strength of the relation and is calculated as $\eta^2 = SS_{between}/SS_{total}$. Higher F-values indicate greater differences between group means relative to the variability within groups.

UZH
IKMZ

2.10 ANOVA: FAQ

- Question: Should I use the t-distribution table or the z-table to find the critical
- Answer:
  - If the standard deviation of the population is unknown, use t-table
  - If known, then is the sample size > 30:
    - If no, use t-distribution table
    - If yes, use z-table

2.11 Correlation

- Pearson correlation (r): measures a linear dependence between two numeric variables (x and y) and can be used only when x and y are from normal distribution
- Kendall tau and Spearman rho: are rank-based correlation coefficients (non-parametric)

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

UZH
IKMZ

2.12 Correlation test

The p-value can be determined by:

- using the correlation coefficient table for the degrees of freedom (df = n−2)
- calculating t (and determining the corresponding p-value using the t-table):

$$t = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}$$

If the p-value is < 5%, then the correlation between x and y is significant.

UZH
IKMZ

3 Quiz

3.1 Can I answer these questions?

True or false?

true	false	Statements
○	○	I use cross-tables to test a relation between two catagorical variables
○	○	Correlations measures the strength of the association between two numeric variables
○	○	The F-statistics (or Fischer test) is used to assess a relation between two numeric variables
○	○	Eta indicates the strenght of the relation between one numeric dependent variable and one categorical independent variable

Score: 0