Statistics for EMCL week 5

Tamás Biró
Humanities Computing
University of Groningen
t.s.biro@rug.nl

This week:

- Two (or more) variables.
- Scatterplots, correlation, regression (M&M 2).
- Causation and lurking variable(s).
- Two-way tables and chi-square test (M&M 9.1).

More variables

Cases $\times$ Variables $=$ Values

- Case $=$ unit $=$ subject.
- For each case: measure/observe value of variables X_1, X_2, X_3, etc.
- Type of var: categorical/quantitative.
- x_{ij}: value of variable i for case j.

More variables

	sex	origin	father born	mother born
case 1	f	us	1948	1948
case 2	m	de	1954	1953
...				
case n	f	ch	1949	1955

More variables

- Variable 1 $vs.$ variable 2.
- **Explanatory variable** $vs.$ **response variable**: causation?
- Values of categorical variable $\longrightarrow$ separate populations (depends on experimental design).

More variables

- **Association** between variables: knowing the value of variable 1 for case i helps predict probable value of variable 2 for same case (and vice versa).
- **Independence**: distribution of variable 2 is same for all values of variable 1.

Two quantitative variables

Two quantitative variables

Scatterplot:

- X-axis: variable 1 (explanatory variable)
- Y-axis: variable 2 (response variable)
- One dot/cross/letter for each case

  (adding third, categorical variable).
- Pattern, deviations. Distinct clusters.

Two quantitative variables

Linear relationship

- **Positive association**: $increase$ in value of variable X typically together with $increase$ in value of variable Y.

- **Negative association**: $increase$ in value of var. X typically together with $decrease$ in value of var. Y (and vice versa).

Two quantitative variables

Correlation: measures direction and strength of linear relationship between two quantitative variables X and Y:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Two quantitative variables

- $-1 \leq r \leq 1$; $0 \leq r^2 \leq 1$
- Positive association: $r > 0$,
  negative association: $r, 0$,
  no linear association $r = 0$.
- $r^2 = 1$: scatterplot exactly straight line.

Two quantitative variables

Regression line: straight line fitting best data points: $y = b_0 + b_1 \cdot x$.

- b_1: slope, b_0: intercept.
- Prediction: given x_i, we predict $\hat{y} = b_0 + b_1 \cdot x_i$. Extrapolation, interpolation.
- $y_i =$ model (prediction from x_i) $+$ residual.
- *Least-squares regression*: minimize sum of squares of residuals. Use software.

Causation and lurking variables

Association $\neq$ causation

- Strong association: r^2 close to 1.
- Same r if we reverse X and Y: no information on which variable is explanatory, and which one is response.
- Often: even no (direct) causation!
- **Lurking variable**.

Association $\neq$ causation

- Time series: for each year, your age and my gray hair.
- Common cause: for each child, vocabulary size and sentence complexity.
- Confounded variables: their effect cannot be distinguished.

Smoking and lung cancer: how to establish causation?

Categorical variables

Models if variable is categorical

Var Y is categorical, k different values.

- One population, two variables (X and Y) measured on it. Test whether variables are correlated or independent.
- k populations, and measure variable X on each of them. Test if X has same distribution in each population.

Categorical vs. quantitative

X categorical, Y quantitative:
Does knowing X influence value you expect for Y?

- Each value of X: separate population.
- Expected value of Y: population mean.
- Comparing two means: two-sample t-test (M&M 7.2).
  Comparing more means: ANOVA (M&M 12.2).

Categorical vs. categorical

In assignment 1: in different languages, different distribution of 1st, 2nd and 3rd person pronouns. Can it be due to random variation, or is there a systematic difference?

- Null hypothesis: random variation.

Example: pronouns in languages

Number of pronouns found per language:

	Lang 1	Lang 2	Lang 3	Lang 4
1st p.	25	30	22	23
2nd p.	20	10	16	18
3rd p.	5	3	6	4

Two-way tables

- Row variable, column variable, cell.
- Joint distribution:
  frequency of having $X = x$ and $Y = y$.
- Marginal distribution:
  frequency of $X = x$ (whatever Y).
  frequency of $Y = y$ (whatever X).

- Conditional distribution:
  for a fixed $Y = y$, frequency of $X = x$.
  for a fixed $X = x$, frequency of $Y = y$.

- Three-way tables.

- Simpson's paradox and perils of aggregation: see example in M&M 2.5.

Example: pronouns in languages

Number of pronouns found per language:

	Lang 1	Lang 2	Lang 3	Lang 4
1st p.	25	30	22	23
2nd p.	20	10	16	18
3rd p.	5	3	6	4

Example: pronouns in languages

Marginal totals and grand total

	Lang 1	Lang 2	Lang 3	Lang 4	Total
1st p.	25	30	22	23	100
2nd p.	20	10	16	18	64
3rd p.	5	3	6	4	18
Total	50	43	44	45	182

Example: pronouns in languages

- Grand total: sample size n.
- Joint distribution: divide each cell by n.
- Marginal distributions: divide marg. totals by n:
  - Freq of 1st form among pronouns of any language.
  - Freq of Language 1 among pronouns of any language.
- Conditional distr: divide cell by row/column total:
  - Freq of 1st form among pronouns of language 2.
  - Freq of Language 1 among 3rd person pronouns.

Two-way tables

Is there association between X and Y?

If no association (X and Y independent):

- $P(X = x \text{ and } Y = y) = P(X = x) \cdot P(Y = y)$, Joint distribution = product of marginal distributions.

- $\text{expected cell count} = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$

Chi-square test

- Null hypothesis: no association.
- **Chi-square statistic**: measures divergence of observed cell counts from expected cell counts:

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

- Follows chi-square distribution with degree of freedom $df = (r-1)(c-1)$.

Example: pronouns in languages

Marginal totals and grand total

	Lang 1	Lang 2	Lang 3	Lang 4	Total
1st p.	25	30	22	23	100
2nd p.	20	10	16	18	64
3rd p.	5	3	6	4	18
Total	50	43	44	45	182

Example: pronouns in languages

Expected cell counts:

	Lang 1	Lang 2	Lang 3	Lang 4	Total
1st p.	27.5	23.6	24.2	24.7	100
2nd p.	17.6	15.1	15.4	15.8	64
3rd p.	4.9	4.3	4.4	4.5	18
Total	50	43	44	45	182

$$\chi^2 = \frac{25 - 27.5^2}{27.5} + \frac{30 - 23.6^2}{23.6} + ... + \frac{6 - 4.4^2}{4.4} + \frac{4 - 4.5^2}{4.5} = 6.6$$

Degrees of freedom:
$df = (c-1)(r-1) = (4-1) \cdot (3-1) = 6$

Critical value: $\alpha = 0.05 \rightarrow (\chi^2)^* = 12.59$

Data do not provide sufficient support to reject null hypothesis at $\alpha = 0.05$ level ($\chi^2 = 6.66$, $df = 6$, $P = ...$).

Next week:

- ANOVA

  NB: Assignment 3 notes on web.