Social Robotics

Albert-Ludwigs-Universität Freiburg

Felix Lindner, Laura Wächter, Bernhard Nebel

SoSe 2019

Correlation

Why We Care

- Imagine you want to test if the older people get the more they respond positively to your robot (or something like that).
- One possibility could be to bin the people into age groups (like: Young and Old).
- But doing so is often quite artificial and arbitrary.
- What you really want to investigate is if Age and Response are related in the sense that, e.g., higher Age values come with higher Response values.
- I.e., you are interested in how the variables are correlated.

Correlation and Causation: Examples

- The more firefighters the higher the damage caused by the fire.
- High risk of wood fire correlates with number of votes for the AfD party in 2017.
- Total revenue generated by arcades correlates with the number of computer science doctorates awarded in the US.
- The state of the thermometer correlates with the room temperature.

Correlation and Causation: Inferences

- Variables may correlate due to common causes.
- Variables may correlate due to a causal relation (direction not always that clear).
  - Causation allows for several inferences:
    - Prediction: If the room temperature is high, the thermometer will be high.
    - Diagnosis: If the thermometer is observed to be high, the room temperature will be high.
    - Intervention: If the thermometer is manipulated to be high, thermometer and room temperature become independent.
    - ... more on this kind of stuff in Knowledge Representation!

From Variance to Covariance

- We assume pairs (X_i, Y_i) sampled from the joint distribution of X and Y.
- The sample covariance between the two variables X, Y is defined as

$$Cov(X,Y) = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

- Variance is a special case of covariance

$$Cov(X,X) = \frac{\sum_i (X_i - \overline{X})(X_i - \overline{X})}{n-1} = \frac{\sum_i (X_i - \overline{X})^2}{n-1} = s_X^2$$

$\Rightarrow$Cf., lecture11.Rmd

From Covariance to Correlation

- $Corr(X,Y) = \frac{Cov(X,Y)}{s_X s_Y} = Cov(\frac{X-\overline{X}}{s_X}, \frac{Y-\overline{Y}}{s_Y})$
- Property: $-1 \leq Corr(X,Y) \leq 1$
- Proof: W.l.o.G. assume X, Y are already standardized (i.e., they have mean 0 and variance 1). Now consider
  - $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y) = 2 + 2Cov(X,Y) \geq 0$
  - $Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y) = 2 - 2Cov(X,Y) \geq 0$ $\square$

Pearson's Correlation Coefficient

- $r = Corr(X,Y) = \frac{Cov(X,Y)}{s_X s_Y} = Cov(\frac{X-\overline{X}}{s_X}, \frac{Y-\overline{Y}}{s_Y})$ is also called the Pearson's Correlation Coefficient or Pearson's r, the Pearson product-moment correlation coefficient, or the bivariate correlation.
- $r = \frac{Cov(X,Y)}{s_X s_Y} = \frac{\frac{\sum_i (X_i-\overline{X})(Y_i-\overline{Y})}{n-1}}{\sqrt{\frac{\sum_i (X_i-\overline{X})^2}{n-1}}\sqrt{\frac{\sum_i (Y_i-\overline{Y})^2}{n-1}}} = Cov(\frac{X-\overline{X}}{s_X}, \frac{Y-\overline{Y}}{s_Y}) = \frac{\sum_i \left(\frac{(X_i-\overline{X})}{s_X} \cdot \frac{(Y_i-\overline{Y})}{s_Y}\right)}{n-1}$
- Pearson's r can be used to test relationship hypotheses involving two interval-scaled variables. That is, we can test H0: $\rho = 0$ against its alternative.

$\Rightarrow$Cf., lecture11.Rmd

Pearson's Correlation: Example

Example

In an experiment, we measure each participant's age and the time each participant needs to complete a task in seconds. The data: $(20, 100)$, $(21, 100)$, $(30, 120)$, $(31, 130)$, $(45, 130)$, $(50, 200)$

- $\overline{Age} = 32.83$, $s_{Age} = 12.32$, $\overline{Time} = 130$, $s_{Time} = 36.88$
- Standardized Scores: $(-1.04, -.81)$, $(-.96, -.81)$, $(-.23, -.27)$, $(-.15, 0)$, $(.98, 0)$, $(1.39, 1.90)$
- $r = \frac{-1.04 \cdot -.81 + \ldots + 1.39 \cdot 1.90}{6-1} = .87$

$\Rightarrow$Cf., lecture11.Rmd

Pearson's Correlation: Undirectional

- We want to test the alternative hypothesis that there is a relationship between Age and Time, i.e., $H1 : \rho \neq 0$, $H0 : \rho = 0$ based on r.
  - If X, Y come from a normal distribution, then the inference can be done via a t-Test. The test statistics reads

$$t = \frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}} = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$$

  - Thus, to test if our $r = .87$ is statistically significant with $\alpha = .05$, we compute $t = .87\sqrt{\frac{6-2}{1-.87^2}} = 3.53$, and the corresponding $p = P(t \leq -3.53) + 1 - P(t \leq 3.53) = .024$.

Pearson's Correlation: Directional

- Test for the alternative hypothesis that higher age comes with higher times, i.e., $H1 : \rho > 0$, $H0 : \rho \leq 0$.
  - Thus, to test if our $r = .87$ requires to reject $H0$ with $\alpha = .05$, we compute $t = .87\sqrt{\frac{6-2}{1-.87^2}} = 3.53$, and the corresponding $p = 1 - P(t \leq 3.53) = .012$.
- To test the alternative hypothesis that higher age comes with lower times, i.e., $H1 : \rho < 0$, $H0 : \rho \geq 0$.
  - Thus, to test if our $r = .87$ requires to reject $H0$ with $\alpha = .05$, we compute $t = .87\sqrt{\frac{6-2}{1-.87^2}} = 3.53$, and the corresponding $p = P(t \leq 3.53) = .988$.

Alternative Test using Z

Theorem

Let r be the correlation coefficient and ρ the population correlation. The statistic:

$$W = \frac{1}{2} ln \frac{1+r}{1-r}$$

follows an approximate normal distribution with mean $E(W) = \frac{1}{2} ln \frac{1+\rho}{1-\rho}$ and variance $Var(W) = \frac{1}{n-3}$.

- Instead of a t-Test, one can also run a z-Test, Cf., lecture11.Rmd

Point-biserial Correlation

- Things can be slightly simplified when one variable is interval-scaled and the other one a two-valued categorical variable. In this case, the categorical variable gets encoded with 1 and 0. Pearson's r then can be computed as:
  - $r_{pb} = \frac{\overline{X1} - \overline{X0}}{s_X} \sqrt{\frac{n_1 n_0}{n(n-1)}}$
  - $t = r_{pb} \sqrt{\frac{n_1 + n_0 - 2}{1 - r_{pb}^2}}$ with degree of freedom $n_1 + n_2 - 2$.

    Reduction from relationship to difference: r_{pb} is significant iff the difference between the two groups is significant.

Example

We observe green and red phases of a traffic light and the number of people crossing the street during these phases. The data: $(g,5)$, $(r,2)$, $(g,10)$, $(r,1)$, $(g,8)$, $(r,3)$, $(g,9)$, $(r,2)$.

Point-biserial Correlation: Example

Example

We observe green and red phases of a traffic light and the number of people crossing the street during these phases. The data: $(g,5)$, $(r,2)$, $(g,10)$, $(r,1)$, $(g,8)$, $(r,3)$, $(g,9)$, $(r,2)$.

- The period variable gets encoded by 1 and 0: $(1,5)$, $(0,2)$, $(1,10)$, $(0,1)$, $(1,8)$, $(0,3)$, $(1,9)$, $(0,2)$
- $\overline{X1} = 32/4 = 8$, $\overline{X2} = 8/4 = 2$, $s_X = 3.55$
- $r_{pb} = \frac{8-2}{3.55}\sqrt{\frac{4 \cdot 4}{8(8-1)}} = .904$
- $t = .904\sqrt{\frac{4+4-2}{1-.904^2}} = 5.19$, $df = 6$

Point-biserial Correlation: Comparison

Example

We observe green and red phases of a traffic light and the number of people crossing the street during these phases. The data: $(g,5)$, $(r,2)$, $(g,10)$, $(r,1)$, $(g,8)$, $(r,3)$, $(g,9)$, $(r,2)$.

- The period variable gets encoded by 1 and 0: $(1,5)$, $(0,2)$, $(1,10)$, $(0,1)$, $(1,8)$, $(0,3)$, $(1,9)$, $(0,2)$
- $\overline{X1} = 32/4 = 8$, $\overline{X0} = 8/4 = 2$, $s_{X1} = 2.16$, $s_{X0} = 0.82$
- $t = \sqrt{4}\frac{8-2}{\sqrt{4.67+0.67}} = 5.19$, $df = 6$

$\Rightarrow$Cf., lecture11.Rmd

Spearman's Correlation Coefficient

- To compute correlation for pairs of observed ordinal variables, rank-based correlation coefficients have been defined. One of these is Spearman's r_s.
- The idea is straighforward: First, the data sample gets converted to ranks (see previous lecture). Then, Pearson's r is used to compute r_s.
- So, let rg_X be ranks of the sample from variable X, and rg_Y be the ranks of the sample from variable Y.

$$r_s = \frac{Cov(rg_X, rg_Y)}{s_{rg_X} s_{rg_X}}$$

Spearman's Correlation Coefficient

- In case there are not too many ties, there is an equivalent way of computing r_s:

$$r_s = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)},$$

- where $d_i = rg_{X_i} - rg_{Y_i}$ for each observed pair (X_i, Y_i), and n is the number of ranks.

Spearman's Correlation: Example

Example

The hypothesis that the more utilitarian someone is the more he dislikes the robot is to be tested. To this end, an experiment is setup that measures the participants moral view on a 5-point Likert scale ranging from 1 (weak utilitarian) to 5 (strong utilitarian), and the likability of the robot measured on a 5-point Likert scale ranging from 1 (low likability) to 5 (high likability). The data looks like this: $(2,4)$, $(1,4)$, $(5,2)$, $(4,1)$, $(3,3)$

$\Rightarrow$Cf., lecture11.Rmd

Spearman's Correlation: Example

Example

The data looks like this: $(2,4)$, $(1,4)$, $(5,2)$, $(4,1)$, $(3,3)$

- First, the data gets ranked seperately
  - Ranks: $rg_X : 1,2,3,4,5$, $rg_Y : 1,2,3,4.5,4.5$
  - Ranked Pairs: $(2,4.5),(1,4.5),(5,2),(4,1),(3,3)$
- Second, the differences get computed
  - $2-4.5=-2.5$, $1-4.5=-3.5$, $5-2=3$, $4-1=3$, $3-3=0$
- $r_s = 1 - \frac{6(-2.5^2+\ldots+0^2)}{5(5^2-1)} = -0.825$
- Finally, to test the hypothesis $H1 : r_s < 0, H0 : r_s \geq 0$:
  - $t = -0.825\sqrt{\frac{5-2}{1-(-0.825^2)}} = -2.53$, $df = 5-2=3$
  - $p = P(t \leq -2.53) = 0.042$

Reporting Correlations

- First, not every correlation is considered equal. Evans (1996) suggests for absolute values of r:
  - $.00 - .19$: Very Weak
  - $.20 - .39$: Weak
  - $.40 - .59$: Moderate
  - $.60 - .79$: Strong
  - $.80 - 1.0$: Very Strong
- Report: Based on the result of the study, a stronger utilitarian view on morality is very strongly related to the disliking of the robot, $r_s = -.825, p = .042$.

Outlook

- Today: We now can also test relationship hypotheses!
- Next time, regression models will be used to analyse the data.

Sketches

Intentionally left blank :-)