Methodological skills rMA linguistics, week 12

> Tamás Biró ACLC University of Amsterdam t.s.biro@uva.nl





Types of the explanatory variables

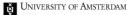
 \times type of the dependent variable

Scale of the	categorical	quantitative
explanatory	(nominal, ordinal)	(interval, ratio,
variable(s) is		logarithmic)
Dependent variable	crosstabs	logistic regression
with categorical scale		
Dependent variable	t-test,	correlation,
with quantitative scale	ANOVA	regression



Null hypotheses and tests



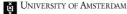






What is the mean of one population?







 H_0 : population $\mu = m$

- Population σ known: one-sample z-test.
- Calculate *z*-statistic: $z := \overline{z} = \frac{\overline{x} m}{\sigma/\sqrt{n}}$.
- Due to Central Limit Theorem, if H₀ true, then the sampling distribution of the z-statistic is/approaches standard Normal distribution N(0,1).
- Use standard Normal table to calculate p-value.
 http://bcs.whfreeman.com/ips6e/content/cat_050/ips6e_table-a.pdf



H_0 : population $\mu = m$ One-tailed or two-tailed?

p-value: probability to draw a sample whose statistic is at least as extreme as the one of our sample, provided that H_0 is true.

What does it mean "at least as extreme"?

If the alternative hypothesis is

- H_a : $\mu \neq m$, then two-tailed.
- H_a : $\mu > m$, then one-tailed. H_a : $\mu < m$, then one-tailed.



 H_0 : population $\mu = m$

- Population σ unknown: one-sample *t*-test.
- Approximate pop. σ with sample s (using n-1).

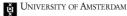
• Calculate *t*-statistic:
$$t := \frac{\overline{x} - m}{s/\sqrt{n}}$$
.

If H₀ is true, then the sampling distribution of the t-statistic for sample size n is/approaches Student's t-distribution with degree of freedom df = n - 1.
 http://bcs.whfreeman.com/ips6e/content/cat_050/ips6e_table-d.pdf



Assumptions (1)

• Simple Random Sample: cases are independent.





Assumptions (2)

- The population follows a Normal distribution, or
- the population distribution is close to a Normal distribution, and n is large, or
- the population distribution is not too far from a Normal distribution, and n is very large.
- Normal quantile plot: test whether sample distribution is close to what you would expect from a sample drawn from a population with a reasonably Normal distribution.

(Or use other Normality tests.)

Non-parametric tests when Normality assumption does not hold.



Are the means of two populations equal?





H_0 : population $\mu_1 = \mu_2$

- Population σ_1 and σ_2 known: two-sample *z*-test.
- Population σ_1 and σ_2 unknown: two-sample *t*-test.
- *df* depends on:
 - Do we know equality of the variances: $\sigma_1 = \sigma_2$, or
 - at least

do the samples make equality of the variances likely?

- Is
$$n_1 = n_2$$
?



Assumptions (1)

• Simple Random Sample: cases are independent.

Paired samples *t***-test**:

Data points come in pair, and so independence does not hold. Perform one-sample *t*-test on differences ("improvement").



Assumptions (2)

- The population follows a Normal distribution, or
- the population distribution is close to a Normal distribution, and n is large, or
- the population distribution is not too far from a Normal distribution, and n is very large.
- Normal quantile plot and Normality tests.

Non-parametric tests when Normality assumption does not hold.



 H_0 : more populations $\mu_1 = \mu_2 = \mu_3$ etc.

ANOVA: Analysis of Variance.

(next week)





Dichotomous data (dichotomous output/dependent variable): yes/no, success/failure, left/right, dead/alive, pregnant/not pregnant.

Null-hypotheses:

•
$$P(\text{success}) = P(\text{failure}) = 0.5.$$

- For some p predicted by theory, P(success) = p.
- P(success|condition1) = P(success|condition2)



More than two outputs: red/blue/green, S/NP/VP/PP, etc. Null-hypotheses:

- P(red) = P(green).P(red|condition1) = P(red|condition2)
- For some p_r predicted by theory, $P(red) = p_r$.
- For some p_r , p_b and p_g predicted by theory $(p_r+p_b+p_g=1)$, $P(\text{red}) = p_r$ and $P(\text{blue}) = p_b$ and $P(\text{green}) = p_g$.



- Pearson's chi-squared test (df = number of outputs -1).
- *z*-test for proportions.
- http://budling.nytud.hu/~birot/prop.php
 http://www.quantitativeskills.com/sisa/statistics/ t-test.htm.



Association of variables





Association of variables

Knowing the value of the explanatory variable(s), we can guess the value of the dependent variable:

Scale of the	categorical	quantitative
explanatory	(nominal, ordinal)	(interval, ratio,
variable(s) is		logarithmic)
Dependent variable	crosstabs	logistic regression
with categorical scale		
Dependent variable	t-test,	correlation,
with quantitative scale	ANOVA	regression

Association of variables

Knowing the value of the explanatory variable(s), we can guess the value of the dependent variable:

When categorical explanatory var and scale dependent var:

Dependent var = mean (explanatory var) + noise

 H_0 : no association: the mean is the same for all levels of the explanatory variable, that is, $\mu_1 = \mu_2 = ...$

(and noise is also the same for all levels of the explanatory variable, that is, $\sigma_1=\sigma_2=...$)



Contingency table = cross tabulation

X: categorical explanatory variable, with n levels. Y: categorical dependent variable, with m levels.

 H_0 : variables X and Y are independent: $P(X = x \& Y = y) = P(X = x) \cdot P(Y = y)$

• Pearson's chi-squared test.

Calculate X²-statistic: $X^2 := \sum_{i=1}^{n \times m} \frac{(O_i - E_i)^2}{E_i}$

- The sampling distribution of the X^2 -statistic asymptotically approaches a χ^2 distribution with $df = (n - 1) \times (m - 1)$. http://bcs.whfreeman.com/ips6e/content/cat_050/ips6e_table-f.pdf
- If small sample sizes: Fisher's exact test, etc.



Next week

- 1. Article presentations?
- 2. Student project presentations?



See you next week!



