Handout: Statistical Methods: Lecture 11, Overview: Lectures 5–10

Dennis Dobler

Vrije Universiteit Amsterdam

December 11, 2018

Overview of statistical methods

	Categorical data	Numerical data
Inference about one population	Confidence interval for p Z test for one proportion p Goodness-of-fit test	Confidence interval for μ t test for mean
Inference about two populations	Confidence interval for $p_1 - p_2$ Z test for two proportions	t test for matched pairs t test for independent samples
Relationship between two variables	Chi-square test of independence Fisher's exact test	t test of correlation Simple linear regression
Comparing ≥ 2 populations	Chi-square test for homogeneity	

The following assertions are "approximately" true:

- Z test: test statistic has the standard normal distribution under H_0
- t test: test statistic has a t-distribution under H_0
- Chi-square test: test statistic has a chi-square distribution under H_0

Chapter 6 Estimates and Sample Sizes

Confidence interval (CI)

- General form: point estimate $\pm$ margin of error.
- $1-\alpha$ CI for p: $\hat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$.
- $1-\alpha$ CI for μ (σ unknown): $\overline{x} \pm t_{n-1,\alpha/2} \cdot \frac{s_n}{\sqrt{n}}$.
- $1-\alpha$ CI for μ (σ known): $\overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$.
- Critical values in Table 2 ($z_{\alpha/2}$) or 3 ($t_{n-1,\alpha/2}$).
- Interpretation: If we were to select many different samples of size n and construct corresponding CIs, then on average $1-\alpha$ of these CIs would contain the true unknown population parameter.
- Requirements: $n > 30$ (p and μ) or population is normally distributed (only μ).

Finding sample size

E desired margin of error, then required sample size n is (round to next integer)

- Proportion: $n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}(1-\hat{p})$ ($\hat{p}$ earlier estimate) or $n = \left(\frac{z_{\alpha/2}}{2E}\right)^2$
- Mean: $n = \left(\frac{\sigma z_{\alpha/2}}{E}\right)^2$.

Recipe for hypothesis testing: P-value method

First (Step 0), identify population parameter of interest.

1. Formulate H_0 and H_a. Choose significance level α.
2. Collect data

3a. Choose test statistic and identify its distribution under H_0.

3b. Compute observed value of test statistic.

3c. Compute P-value, i.e. probability of getting a value of the test statistic which is at least as extreme as observed value.
How to compute it depends on two-, right- or left-tailed test as well!

4. If P-value $\leq \alpha$ reject H_0. Otherwise fail to reject H_0.

Finally:

- Formulate non-technical conclusion: there is (not) sufficient evidence to . . .

Recipe for hypothesis testing: critical value method

First (Step 0), identify population parameter of interest.

1. Formulate H_0 and H_a. Choose significance level α.
2. Collect data

3a. Choose test statistic and identify its distribution under H_0.

3b. Compute observed value of test statistic.

3c. Find the appropriate critical value(s): take test statistic, α, sample size and whether test is two-, left- or right-tailed into account.

4. If observed value is more extreme than critical value(s) reject H_0. Otherwise fail to reject H_0.

Finally:

- Formulate non-technical conclusion: there is (not) sufficient evidence to ...

Chapter 7 Hypothesis Testing

Errors

- Type I error: rejecting H_0 when H_0 is actually true. $P(\text{Type I error}) = \alpha$.
- Type II error: failing to reject H_0 when H_0 is false. $P(\text{Type II error}) = \beta$.

Caution!

- P-value $\neq$ probability that H_0 is true.
- Failing to reject $H_0 \neq$ accept H_0.
- Type I error $\neq$ Type II error.
- α is independent of n
- β decreases with n.

Hypothesis tests for inference about one population

- Z test for proportion: if $n > 30$, the test statistic

$$Z = \frac{\hat{P}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

has under H_0 (i.e. $H_0 : p = p_0$) approximately a $N(0,1)$-distribution.

- Z test for mean (σ known): if $n > 30$ or sample is from normally distributed population, the test statistic

$$Z = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

has under H_0 (i.e. $H_0 : \mu = \mu_0$) approximately a $N(0,1)$-distribution.

- Critical values in Table 2 of Appendix.

Hypothesis tests for inference about one population

- t test for mean μ (with σ unknown): if $n > 30$ or sample is from normally distributed population, the test statistic

$$T = \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}}$$

  has under H_0 (i.e. $H_0 : \mu = \mu_0$) approximately a t-distribution with $n-1$ degrees of freedom.
- Critical values in Table 3 of Appendix. If degree of freedom is not in table: use next lowest degree of freedom.

Section 8.2 Two proportions

Independent vs dependent samples

Two samples are independent if sample values from one population are not related to samples values from the other population.
Two samples are dependent if the samples values are matched pairs, i.e. when there is a relationship between the two values (e.g. measurements from the same subject)

Inference about difference of two population proportions

If $x_1 \geq 5$, $x_2 \geq 5$, $n_1 - x_1 \geq 5$, $n_2 - x_2 \geq 5$ and the two samples are independent then

- $1-\alpha$ CI for $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$ (here $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$);
- the test statistic

$$Z_p = \frac{(\hat{P}_1 - \hat{P}_2)}{\sqrt{\bar{P}(1-\bar{P})/n_1 + \bar{P}(1-\bar{P})/n_2}}$$

approximately has a $N(0,1)$-distribution under $H_0 : p_1 = p_2$. Here $\bar{P} = (X_1 + X_2)/(n_1 + n_2)$ is the pooled sample proportion.

Section 8.3 Two Means: Independent Samples

Inference about difference of two population means

If two samples are independent, choose appropriate test statistic, depending on whether or not σ_1, σ_2 are equal.
Requirements for all statistics: Both samples should be from a normal distribution or $n_1 > 30$ and $n_2 > 30$.

Inference about $\mu_1 - \mu_2$ when σ_1, σ_2 unknown and $\sigma_1 \neq \sigma_2$ (realistic!)

Test statistic: $T_2 = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$ has approximately a t-distribution with approximately $\tilde{n}$ degrees of freedom under H_0 (usually $H_0 : \mu_1 = \mu_2$), where $\tilde{n} = \min\{n_1 - 1, n_2 - 1\}$ if no technology available.

$1 - \alpha$ CI for $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2 \pm t_{\tilde{n}, \alpha/2} \sqrt{s_1^2/n_1 + s_2^2/n_2}$

Inference about $\mu_1 - \mu_2$ when σ_1, σ_2 unknown but $\sigma_1 = \sigma_2$

$T_2^{\text{eq}} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_p^2/n_1 + S_p^2/n_2}}$ has under H_0 a t-distribution with $df = n_1 + n_2 - 2$.

Here S_p^2 is the pooled sample variance given by $S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$.

$1 - \alpha$ CI for $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2 \pm t_{n_1 + n_2 - 2, \alpha/2} \sqrt{s_p^2/n_1 + s_p^2/n_2}$

Section 8.4 Two Dependent Samples (Matched Pairs)

Inference about difference of means, based on dependent samples

$\overline{D}$ is sample mean of differences $X_1 - Y_1, \ldots, X_n - Y_n$ of the matched pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, and S_d is the sample standard deviation of the differences. If $n > 30$ or differences are from a normal distribution, then

- $1 - \alpha$ CI for $\mu_1 - \mu_2$ is $\overline{d} \pm t_{n-1,\alpha/2} \cdot \frac{s_d}{\sqrt{n}}$.
- The test statistic $T_d = \frac{\overline{D} - (\mu_1 - \mu_2)}{S_d/\sqrt{n}}$ has under H_0 (usually $H_0 : \mu_1 - \mu_2 = 0$) approximately a t-distribution with $n - 1$ degrees of freedom.

Section 9.2 Correlation

(Linear) Correlation

Correlation between two variables — the values of two variables are somehow associated with each other.
Linear correlation — the relationship is approximately a straight line. Construct a scatterplot for this.
A measure for linear relationship between variables x and y is given by the sample linear correlation coefficient r:

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

Also, r is an estimate of the population linear correlation coefficient ρ.

Testing $\rho = 0$

If the data follow approximately a straight line and there no outliers, then the test statistic $T_\rho = \frac{R}{\sqrt{\frac{1-R^2}{n-2}}}$, has under $H_0 : \rho = 0$ a t-distribution with $n-2$ degrees of freedom.

Section 9.3 Regression

Simple linear regression

The simple linear regression model is given by $y_i = \beta_0 + \beta_1 x_i + \text{error}_i$.
The regression equation is $\hat{y} = b_0 + b_1 x$ where b_0 and b_1 are least-squares estimates of β_0 and β_1.

Least-squares estimates

As a result of the least-squares approach we obtain

$$b_1 = r\frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \overline{y} - b_1\overline{x},$$

where s_x is the sample standard deviation of x and $\overline{x}$ is the sample mean of x (s_y and $\overline{y}$ are corresponding sample statistics for y).

Coefficient of determination (Section 9.4, page 567)

r^2 is called the coefficient of determination and equals the proportion of variation in y explained by the regression equation

Section 9.3 Regression

Is linear regression model a good model?

- Straight line is reasonable fit: check scatterplot
- Coefficient of determination r^2 is high, close to 1;
- b_1 is significantly different from 0: test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$

Testing β_1

If the points follow approximately a straight line and the errors are independent and from normal distribution with fixed standard deviation, the score $t_\beta = \frac{b_1 - \beta_1}{s_{b_1}}$ is a realization of the test statistic T_β that under H_0 has a t-distribution with $n-2$ degrees of freedom.

Checking assumptions about errors

Consider a normal QQ plot of the residuals $y_i - \hat{y}_i$ and a residual plot (scatterplot of the residuals against the x values).

Section 10.2 Goodness-of-Fit

Goodness-of-fit test

- Suppose there are k different categories and a random sample of size n is conducted.
- H_0: frequency counts agree with the claimed distribution $p_1 = \langle\text{value}\rangle, \ldots$
  H_a: frequency counts do not agree with the claimed distribution.
- Let O_i be the observed frequency count of category i.
  Expected frequency E_i is computed by $E_i = n \cdot p_i$.
- Requirements: all $E_i \geq 5$.
- If the requirements are met, then the test statistic

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

  has approximately a chi-square distribution with $k - 1$ degrees of freedom.
- H_0 is rejected for large values of the observed value χ^2: reject H_0 if $\chi^2 > \chi^2_{k-1,\alpha}$, where $\chi^2_{k-1,\alpha}$ can be found in Table 4 of Appendix.

10.3 Contingency Tables

Test of independence

- row variable has r categories, column variable has c categories.
- H_0: row and column variable are independent;
  H_a: row and column variable are dependent;
- $E_{ij} = \frac{(i\text{-th row total}) \cdot (j\text{-th column total})}{\text{grand total}}$.
- Requirements (E_{ij} is expected frequency count in cell (i, j) under H_0)
  - 2×2: all $E_{ij} \geq 5$.
  - larger tables: all $E_{ij} \geq 1$ and 80% of $E_{ij} \geq 5$.
- If the requirements are met, the test statistic $X^2 = \sum \frac{(O-E)^2}{E} = \sum_{(i,j)} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ has under H_0 approximately a chi-square distribution with $(r-1)(c-1)$ degrees of freedom.
- Reject H_0 if observed value $\chi^2 > \chi^2_{(r-1)(c-1),\alpha}$.

Test of homogeneity

Same procedure as test of independence, but null hypothesis is that *different populations* have the same proportions of some characteristics. Data is obtained by multiple samples from the different populations.

10.3 Contingency Tables

If requirements for chi-square test are not met or if we want to test a directed claim for 2×2 contingency table: use Fisher exact test.

Fisher exact test for 2×2 contingency table

- H_0: row and column variable are independent;
  H_1: occurrence of "first column category" is more common
  in group of "first row category" than in group of "second row category".
  Significance level α.
- Test statistic: frequency count in cell $(1, 1)$ has under H_0 and given marginals a hypergeometric distribution with parameters
  $n = \langle \text{first row total} \rangle, N = \langle \text{grand total} \rangle$ and $k = \langle \text{first column total} \rangle$.
- For this alternative hypothesis, H_0 is rejected for large values of the test statistic.

December 18 – Final exam at Emergohal (be there on time, bring ID!)

Good luck!

(Merry Christmas and a Happy New Year!)