Statistics

Week 6: Inference for Two Samples (Chapter 8)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Information

Monday: comparing two variances; summary.

Tuesday: revision exercises. Combined slides will be available.

Homework 2 will be available.

- Show all working.
- Due 1pm, Tuesday 14 March. Submit a single file on *eDimension*.
- Take a look before the exam (and use it as practice).

Thursday: recitation (two samples) + project recitation.

Outline

Independent samples design, small sample size

Suppose $x_1, x_2, \ldots, x_n$ come from an $N(\mu_1, \sigma^2)$ distribution, and $y_1, y_2, \ldots, y_n$ come from an $N(\mu_2, \sigma^2)$ distribution; n is not necessarily large.

Note: we are assuming that the variances are the same.

Then the $(1-\alpha)$ two-sided confidence interval for $\mu_1 - \mu_2$ is:

$$\bar{x} - \bar{y} - t_{2n-2,\,1-\alpha/2}\sqrt{\frac{s_1^2 + s_2^2}{n}} \le \mu_1 - \mu_2 \le \bar{x} - \bar{y} + t_{2n-2,\,1-\alpha/2}\sqrt{\frac{s_1^2 + s_2^2}{n}}.$$

General rule for calculating the *degree of freedom*: total number of data points minus the number of contraints.

Comparing variances

Therefore it is useful to test whether two populations have the same variance.

Set up: assume that $x_1, x_2, \ldots, x_{n_1}$ come from an $N(\mu_1, \sigma_1^2)$ distribution, and $y_1, y_2, \ldots, y_{n_2}$ come from an $N(\mu_2, \sigma_2^2)$ distribution.

To test whether the variances are the same, we use the ratio σ_1^2/σ_2^2, which is estimated using s_1^2/s_2^2.

Terminology: when not all the random variables in a collection have the same variance, they are called *heteroscedastic*.

Snedecor's F distribution

Define the random variable

$$F_{u,v} = \frac{\chi_u^2/u}{\chi_v^2/v}.$$

This random variable has an F distribution with degrees of freedom u and v.

Its probability density function is

$$F_{u,v}(x) = \frac{\Gamma((u+v)/2)}{\Gamma(u/2)\Gamma(v/2)}\left(\frac{u}{v}\right)^{u/2} x^{u/2-1}\left(1+\frac{u}{v}x\right)^{-(u+v)/2}.$$

Comparing variances – CI

Recall that $\dfrac{(n_i - 1)s_i^2}{\sigma_i^2}$ is a $\chi^2_{n_i-1}$ random variable.

Therefore $\dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ is an $F_{n_1-1,\,n_2-1}$ random variable.

It follows that

$$P\left(f_{n_1-1,\,n_2-1,\,\alpha/2} \leq \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \leq f_{n_1-1,\,n_2-1,\,1-\alpha/2} \right) = 1 - \alpha,$$

where $f_{n_1-1,\,n_2-1,\,x}$ is a critical point, given by `F.INV`(x, $n_1 - 1$, $n_2 - 1$) in *Excel*.

Hence, a two-sided $(1 - \alpha)$ CI for σ_1^2/σ_2^2 is

$$\left[\frac{1}{f_{n_1-1,\,n_2-1,\,1-\alpha/2}} \frac{s_1^2}{s_2^2},\ \frac{1}{f_{n_1-1,\,n_2-1,\,\alpha/2}} \frac{s_1^2}{s_2^2} \right].$$

Comparing variances – example

Suppose $s_1 = 1.1$, $s_2 = 0.9$, and $n_1 = n_2 = 50$.

How do we test $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$, using $\alpha = 0.05$?

Outline

Non-comprehensive summary of what we covered

- Experiments and surveys: sensitive questions, sampling methods, placebo effect, control group, randomized block design, Latin square, Simpson's paradox.

  Excel: Data Analysis → `Random Number Generation`; `Sampling` (systematic; with replacement).

- Summary statistics: mean, median, standard deviation, IQR, outlier, covariance.

  Excel: Data Analysis → `Descriptive Statistics`.

- Graphical methods: bar chart, histogram, box plot, Q-Q plot, time series, EWMA.

  Excel histogram: Data Analysis → `Histogram`. EWMA: `Exponential Smoothing` (Damping factor $= 1 - \alpha$).

- Unbiased estimators: s^2, German tank problem.

- Distributions: normal, CLT, χ^2, Student t, F.

  Excel: `norm.s.inv`, `norm.s.dist`, etc.

- Confidence intervals, single sample: for mean with known and unknown σ; for variance; one and two sided.

- Confidence intervals, two samples: for mean (independent samples, matched pair); for variance.

  Excel: Data Analysis $\rightarrow$ `z-Test:  Two Sample for Means`, `t-Test:  Paired Two Sample for Means`

- Hypothesis testing: H_0 vs H_1; reject/do not reject H_0; one and two sided; for μ and σ; connection with CI; p-value; α and β; power and sample size calculation.

After the break

- Hypothesis testing involving proportions.
- Chi-squared test: how to show quantitatively if your data fits a certain distribution?
- Regression (trend lines): linear, multiple, logistic.
- ANOVA: when there are more than two treatments, how can we know if one is significantly better than the others?
- Non-parametric tests: bootstrap, permutation test, sign test.
- Test for randomness: how can you tell if some data (e. g. sequence of coin tosses) is made up?
- Maximum likelihood.