Two-sample inferences

Testing $H_0 : \mu_X = \mu_Y$ when $\sigma_X = \sigma_Y$

Bruce M. Boghosian

Tufts University School of Arts and Sciences

Department of Mathematics

Tufts University

Tufts UNIVERSITY

Outline

Two sample tests

- Sometimes, instead of comparing a sample mean to a known value, we wish to compare *two sample means*:
  - **Two sources:** Farm X and Farm Y each send 10 cases of barley. For both shipments, we quantify the quality of each case. We would like to compare μ_X to μ_Y. Note this is different from comparing μ_X to a hypothesized μ_0. We might wish to do hypothesis testing on $H_0 : \mu_X = \mu_Y$, etc.
  - **Two treatments:** Farm sends two shipments, X and Y, of barley, each consisting of 10 cases. We malt (soak in water) the barley of shipment X for 8 hours before roasting it over a peat fire, and that of shipment Y for 12 hours before roasting it over a peat fire. Then we quantify the quality of the malted and roasted barley in both cases, and we compare μ_X and μ_Y. Again, we might also wish to do hypothesis testing on $H_0 : \mu_X = \mu_Y$, etc.
- Likewise, we might wish to compare *two sample variances*, e.g., $H_0 : \sigma_X^2 = \sigma_Y^2$.

Testing $H_0 : \mu_X = \mu_Y$ when $\sigma_X = \sigma_Y$

- **Thm:**
  - Let $X_1, \ldots, X_n$ be first random sample from $N(\mu_X, \sigma)$
  - Let $Y_1, \ldots, Y_m$ be second random sample from $N(\mu_Y, \sigma)$
  - Let S_X^2 and S_Y^2 be the two sample variances
  - Let S_p^2 be the *pooled variance*,

$$S_p^2 = \frac{\sum_{j=1}^{n}(X_j - \overline{X})^2 + \sum_{j=1}^{m}(Y_j - \overline{Y})^2}{n+m-2} = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

  - Then the quantity

$$T_{n+m-2} = \frac{\left(\overline{X} - \mu_x\right) - \left(\overline{Y} - \mu_Y\right)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

has Student T distribution with $n + m - 2$ df.

- **Pf:**
  - First note that we can write

$$T_{n+m-2} = \frac{\left(\overline{X} - \mu_x\right) - \left(\overline{Y} - \mu_Y\right)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{\sqrt{\frac{S_p^2}{\sigma^2}}}$$

$$= \frac{\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{\sqrt{\frac{1}{n+m-2}\left[\sum_{j=1}^{n}\left(\frac{X_j-\overline{X}}{\sigma}\right)^2 + \sum_{j=1}^{m}\left(\frac{Y_j-\overline{Y}}{\sigma}\right)^2\right]}}$$

  - $\mathrm{Var}\left(\overline{X} + \overline{Y}\right) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)$
  - So numerator of above is distributed as a standard normal

Testing $H_0 : \mu_X = \mu_Y$ when $\sigma_X = \sigma_Y$ (continued)

- **Pf:**
  - Turning our attention to the denominator of

$$T_{n+m-2} = \frac{\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{\sqrt{\frac{1}{n+m-2}\left[\sum_{j=1}^{n}\left(\frac{X_j-\overline{X}}{\sigma}\right)^2+\sum_{j=1}^{m}\left(\frac{Y_j-\overline{Y}}{\sigma}\right)^2\right]}}$$

  - We see that $\sum_{j=1}^{n}\left(\frac{X_j-\overline{X}}{\sigma}\right)^2$ and $\sum_{j=1}^{m}\left(\frac{Y_j-\overline{Y}}{\sigma}\right)^2$ are independent χ^2 r.v.s with $n-1$ and $m-1$ df, respectively.
  - Hence their sum U is χ^2 distributed with $n+m-2$ df.
  - Also numerator and denominator above are independent.
  - $\therefore\ T_{n+m-2} = \frac{Z}{\sqrt{\frac{U}{n+m-2}}}$ is t distributed with $n+m-2$ df. $\square$

Testing $H_0 : \mu_X = \mu_Y$ when $\sigma_X = \sigma_Y$ (continued)

Bruce M. Boghosian

- **Thm.:** Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ be independent random samples from normal distributions with means μ_X and μ_Y, respectively, and with the same standard deviation σ.
- Since H_0 is $\mu_X = \mu_Y$, define the quantity $t = \frac{\overline{x} - \overline{y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$.
  - To test $H_0 : \mu_X = \mu_Y$ versus $H_1 : \mu_X > \mu_Y$ at the α level of significance, reject H_0 if $t \geq +t_{\alpha, n+m-2}$.
  - To test $H_0 : \mu_X = \mu_Y$ versus $H_1 : \mu_X < \mu_Y$ at the α level of significance, reject H_0 if $t \leq -t_{\alpha, n+m-2}$.
  - To test $H_0 : \mu_X = \mu_Y$ versus $H_1 : \mu_X \neq \mu_Y$ at the α level of significance, reject H_0 if either (a) $t \leq -t_{\alpha/2, n+m-2}$ or (b) $t \geq +t_{\alpha/2, n+m-2}$.

Example

- Were Mark Twain and Quintus Curtius Snodgrass the same person?
- Proportion of 3-letter words used in $n = 8$ writings of MT and $m = 10$ of QCS.

MT	0.225	0.262	0.217	0.240	0.230	0.229	0.235	0.217		
QCS	0.209	0.205	0.196	0.210	0.202	0.207	0.224	0.223	0.220	0.201

- Find $\overline{x} = \frac{1}{8}\sum_{j=1}^{8} x_j = 0.2319$ & $\overline{y} = \frac{1}{10}\sum_{j=1}^{10} y_j = 0.2097$.
- Is this close enough to conclude that $\mu_X = \mu_Y$?

Example (continued)

- Find $\overline{x} = \frac{1}{8}\sum_{j=1}^{8} x_j = 0.2319$ & $\overline{y} = \frac{1}{10}\sum_{j=1}^{10} y_j = 0.2097$.
- Is this close enough to conclude that $\mu_X = \mu_Y$?
- Hypothesis test with $H_0 : \mu_X = \mu_Y$ and $H_1 : \mu_X \neq \mu_Y$.
- Calculate $s_X^2 = 0.0002103$ and $s_Y^2 = 0.0000955$
- Then calculate

$$s_p = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}} = 0.0121.$$

Example (continued)

- Then calculate

$$t = \frac{\overline{x} - \overline{y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{0.2319 - 0.2097}{0.0121 \sqrt{\frac{1}{8} + \frac{1}{10}}} = 3.88$$

- Take $\alpha = 0.01$, reject H_0 if $t \leq -t_{0.005,16} = -2.9208$ or $t \geq t_{0.005,16} = 2.9208$.
- Hence we reject H_0.
- MT & QCS not same person with 99% confidence.

Tufts UNIVERSITY

Summary

- We have defined two-sample tests.
- We have tested $H_0 : \mu_X = \mu_Y$.
- We have provided an example.