Statistical Computation
Math 475

Jimin Ding

Department of Mathematics
Washington University in St. Louis
www.math.wustl.edu/ jmding/math475/index.html

October 8, 2013

Part III

Two Continuous Variables

Correlation

Pearson Correlation Coefficient: (population version)

$$\rho_{xy} = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}.$$

Correlation

Pearson Correlation Coefficient: (population version)

$$\rho_{xy} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}.$$

Sample Pearson correlation: (an estimate of ρ)

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Correlation

Pearson Correlation Coefficient: (population version)

$$\rho_{xy} = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}.$$

Sample Pearson correlation: (an estimate of ρ)

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- $|\rho| \leq 1$.
- ρ only describes linear association.
- $\rho > 0$: positive association.
- different from regression, x and y are exchangable.

Paired t-test:

A t-test on the difference between X_i and Y_i. This test is only meaningful when X_i and Y_i are from the same subjects and the difference is of interest. The sample sizes of X's and Y's have to be same.

$$H_0 : E(X_i - Y_i) = 0, \quad v.s. \quad H_a : E(X_i - Y_i) \neq 0.$$

Paired t-test:

A t-test on the difference between X_i and Y_i. This test is only meaningful when X_i and Y_i are from the same subjects and the difference is of interest. The sample sizes of X's and Y's have to be same.

$$H_0 : E(X_i - Y_i) = 0, \quad v.s. \quad H_a : E(X_i - Y_i) \neq 0.$$

Example: P20 1.2; weights before and after a diet program;

midterm and final scores; salary increase.

Paired t-test:

A t-test on the difference between X_i and Y_i. This test is only meaningful when X_i and Y_i are from the same subjects and the difference is of interest. The sample sizes of X's and Y's have to be same.

$$H_0 : E(X_i - Y_i) = 0, \quad v.s. \quad H_a : E(X_i - Y_i) \neq 0.$$

Example: P20 1.2; weights before and after a diet program;

midterm and final scores; salary increase.

Idea: Test the mean of the observed difference.

Two-sample t-test:

A simple but common comparison of averages of two groups.

$$H_0 : E(X_i) = E(Y_i), \quad v.s. \quad H_a : E(X_i) \neq E(Y_i).$$

Two-sample t-test:

A simple but common comparison of averages of two groups.

$$H_0 : E(X_i) = E(Y_i), \quad v.s. \quad H_a : E(X_i) \neq E(Y_i).$$

Example: compare results between treatment and control group in a clinical trial; grades of males and females.

Two-sample t-test:

A simple but common comparison of averages of two groups.

$$H_0 : E(X_i) = E(Y_i), \quad v.s. \quad H_a : E(X_i) \neq E(Y_i).$$

Example: compare results between treatment and control group in a clinical trial; grades of males and females.

Hypothesis testing procedure:

- Setup a statistical model (with/without parameters);
- Setup null and alternative hypothesis;
- Calculate the test statistic;
- Find the p-value based the distribution of the test statistic;
- Draw the conclusion.

Classical two-sample t-test:

- Model: We assume $X_i \overset{iid}{\sim} N(\mu_X, \sigma^2), i = 1, 2, \cdots, m$ and $Y_j \overset{iid}{\sim} N(\mu_Y, \sigma^2), j = 1, 2, \cdots, n$. Note, the same variance!

Classical two-sample t-test:

- Model: We assume $X_i \overset{iid}{\sim} N(\mu_X, \sigma^2), i = 1, 2, \cdots, m$ and $Y_j \overset{iid}{\sim} N(\mu_Y, \sigma^2), j = 1, 2, \cdots, n$. Note, the same variance!
- $H_0 : \mu_X = \mu_Y, \quad v.s. \quad H_a : \mu_X \neq \mu_Y.$

Classical two-sample t-test:

- Model: We assume $X_i \overset{iid}{\sim} N(\mu_X, \sigma^2), i = 1, 2, \cdots, m$ and $Y_j \overset{iid}{\sim} N(\mu_Y, \sigma^2), j = 1, 2, \cdots, n$. Note, the same variance!
- $H_0 : \mu_X = \mu_Y, \quad v.s. \quad H_a : \mu_X \neq \mu_Y$.
- Test statistics:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}} \sqrt{\frac{1}{n} + \frac{1}{m}}},$$

has t_{m+n-2} distribution.

Classical two-sample t-test:

- Model: We assume $X_i \overset{iid}{\sim} N(\mu_X, \sigma^2), i = 1, 2, \cdots, m$ and $Y_j \overset{iid}{\sim} N(\mu_Y, \sigma^2), j = 1, 2, \cdots, n$. Note, the same variance!
- $H_0 : \mu_X = \mu_Y, \quad v.s. \quad H_a : \mu_X \neq \mu_Y$.
- Test statistics:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}}\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

  has t_{m+n-2} distribution.
- p-value$= P(|T| > t | H_0$ is true), where $T \sim t_{m+n-2}$.

Classical two-sample t-test:

- Decision:
  A. When p-value < 0.05 (or the level of significance α), we conclude that there is a significant evidence in the data to reject H_0, which means two groups have significantly different means.
  B. When p-value ≥ 0.05 or α, we conclude that there is no significant evidence in the data to reject H_0.
  (0.01: highly significant.)

Satterthwaite t-test:

Prospoesd by Satterthwaite (1946)

- Model: We assume $X_i \stackrel{iid}{\sim} N(\mu_X, \sigma_X^2), i = 1, 2, \cdots, m$ and $Y_j \stackrel{iid}{\sim} N(\mu_Y, \sigma_Y^2), j = 1, 2, \cdots, n$. Note, different variance!

Satterthwaite t-test:

Prospposed by Satterthwaite (1946)

- Model: We assume $X_i \overset{iid}{\sim} N(\mu_X, \sigma_X^2), i = 1, 2, \cdots, m$ and $Y_j \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2), j = 1, 2, \cdots, n$. Note, different variance!
- $H_0 : \mu_X = \mu_Y, \quad v.s. \quad H_a : \mu_X \neq \mu_Y.$

Satterthwaite t-test:

Prosposed by Satterthwaite (1946)

- Model: We assume $X_i \overset{iid}{\sim} N(\mu_X, \sigma_X^2), i = 1, 2, \cdots, m$ and $Y_j \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2), j = 1, 2, \cdots, n$. Note, different variance!
- $H_0 : \mu_X = \mu_Y, \quad v.s. \quad H_a : \mu_X \neq \mu_Y.$
- Test statistics:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{s_X^2/m + s_Y^2/n}},$$

APPROXIMATELY has t_{df} distribution, where

$$df = \frac{(W_1 + W_2)^2}{\frac{W_1^2}{m-1} + \frac{W_2^2}{n-1}}$$

and $W_1 = s_X^2/m$, $W_2 = s_Y^2/n$.

Satterthwaite t-test:

Prosposed by Satterthwaite (1946)

- Model: We assume $X_i \overset{iid}{\sim} N(\mu_X, \sigma_X^2), i = 1, 2, \cdots, m$ and $Y_j \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2), j = 1, 2, \cdots, n$. Note, different variance!
- $H_0 : \mu_X = \mu_Y, \quad v.s. \quad H_a : \mu_X \neq \mu_Y$.
- Test statistics:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{s_X^2/m + s_Y^2/n}},$$

  APPROXIMATELY has t_{df} distribution, where

$$df = \frac{(W_1 + W_2)^2}{\frac{W_1^2}{m-1} + \frac{W_2^2}{n-1}}$$

  and $W_1 = s_X^2/m$, $W_2 = s_Y^2/n$.
- p-value

Satterthwaite t-test:

Prosposed by Satterthwaite (1946)

- Model: We assume $X_i \stackrel{iid}{\sim} N(\mu_X, \sigma_X^2), i = 1, 2, \cdots, m$ and $Y_j \stackrel{iid}{\sim} N(\mu_Y, \sigma_Y^2), j = 1, 2, \cdots, n$. Note, different variance!
- $H_0 : \mu_X = \mu_Y, \quad v.s. \quad H_a : \mu_X \neq \mu_Y.$
- Test statistics:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{s_X^2/m + s_Y^2/n}},$$

  APPROXIMATELY has t_{df} distribution, where

$$df = \frac{(W_1 + W_2)^2}{\frac{W_1^2}{m-1} + \frac{W_2^2}{n-1}}$$

  and $W_1 = s_X^2/m$, $W_2 = s_Y^2/n$.
- p-value
- Decision

When to Use Which?

- First test: $H_0 : \sigma_X^2 = \sigma_Y^2 \;\; v.s. H_a : \sigma_X^2 \neq \sigma_Y^2$. If we reject H_0, then we should use Satterthwaite t-test. Otherwise, we can use pooled-variance (classical) t-test.
- Folded Form F test can be used for homogeneity test.
- Always use Satterthwaite t-test to be conservative.
- Approximation is good when n and m are not small.

More tests:

Q: What happens if the normality assumptions fail?

More tests:

Q: What happens if the normality assumptions fail?

A: If the sample size is big enough, for example more than 30, then above tests are still approporate.

More tests:

Q: What happens if the normality assumptions fail?

A: If the sample size is big enough, for example more than 30, then above tests are still approporate.

Q: How about if sample size is small and normallities fail?

Nonparametric tests:

- Wilcoxon Rank Sum Test:
  Assume equal variance (homogeneity of variance, homoscedastic).
  Order all observations from both groups (A and B) and sum the ranks for observations from group A. The test statistic is based on this sum of ranks, whose distribution can be approximated by t distribution or normal distribution.
  Use "EXACT" statement when sample size is small.
- Kruskal-Wallis Test:
  Allow nonequal variance (heteroscedastic).
  It is based on median of ranks of two groups. This test, which extend Mann-Whitney U test to more than 2 groups, is commonly used in one-way ANOVA. The test statistic is also based on ranks and approximately has χ^2 distribution.

SAS Program

- PROC TTEST;
- PROC NPAR1WAY.

Reading Assignment

Textbook: Applied Statistics and the SAS Programming Language,
Chap 5 A-D: P159-P164
Chap 6: P183-P196