VE401 Probabilistic Methods in Eng.
RC 6

CHEN Xiwen

UM-SJTU Joint Institute

April 16, 2020

Table of contents

Basic Statistic

Suppose sample means $\overline{X}^{(1)}$ and $\overline{X}^{(2)}$ are calculated from samples of sizes n_1 and n_2 respectively from normal populations with means μ_1, μ_2 and variances σ_1, σ_2. Then since

$$\overline{X}^{(1)} \sim \mathrm{N}(\mu_1, \sigma_1^2/n_1), \qquad \overline{X}^{(2)} \sim \mathrm{N}(\mu_2, \sigma_2^2/n_2),$$

the statistic

$$Z = \frac{\overline{X}^{(1)} - \overline{X}^{(2)} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

follows a standard normal distribution.

Variances Known

Variances known. Let $X_1^{(i)}, \ldots, X_{n_i}^{(i)}$ with $i = 1, 2$ be samples of sizes n_1 and n_2 from normal distributions with unknown means μ_1, μ_2 and ***known*** variances σ_1^2, σ_2^2. Then the test statistic is given by

$$Z = \frac{\overline{X}^{(1)} - \overline{X}^{(2)} - (\mu_1 - \mu_2)_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

We reject at significance level α

- $H_0 : \mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$ if $|Z| > z_{\alpha/2}$,
- $H_0 : \mu_1 - \mu_2 \leq (\mu_1 - \mu_2)_0$ if $Z > z_\alpha$,
- $H_0 : \mu_1 - \mu_2 \geq (\mu_1 - \mu_2)_0$ if $Z < -z_\alpha$.

Variances Known

OC curve. We can use the OC curves for normal distributions with

$$d = \frac{|(\mu_1 - \mu_2) - (\mu_1 - \mu_2)_0|}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

with $n = n_1 = n_2$. When $n_1 \neq n_2$, we use the equivalent sample size

$$n = \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2/n_1 + \sigma_2^2/n_2}.$$

Variances Equal but Unknown — Student's T-Test

Variances equal but unknown. Let $X_1^{(i)}, \ldots, X_{n_i}^{(i)}$ with $i = 1, 2$ be samples of sizes n_1 and n_2 from normal distributions with unknown means μ_1, μ_2 and ***equal*** but ***unknown*** variances $\sigma^2 = \sigma_1^2 = \sigma_2^2$. Then the test statistic is given by

$$T_{n_1+n_2-2} = \frac{\overline{X}^{(1)} - \overline{X}^{(2)} - (\mu_1 - \mu_2)_0}{\sqrt{S_p^2(1/n_1 + 1/n_2)}},$$

with ***pooled estimator for variance***

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

We reject at significance level α

- $H_0 : \mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$ if $|T_{n_1+n_2-2}| > t_{\alpha/2, n_1+n_2-2}$,
- $H_0 : \mu_1 - \mu_2 \leq (\mu_1 - \mu_2)_0$ if $T_{n_1+n_2-2} > t_{\alpha, n_1+n_2-2}$,
- $H_0 : \mu_1 - \mu_2 \geq (\mu_1 - \mu_2)_0$ if $T_{n_1+n_2-2} < -t_{\alpha, n_1+n_2-2}$.

Variances Equal but Unknown — Student's T-Test

OC curve. We use the OC curves for the T-test in case of equal sample sizes $n = n_1 = n_2$

$$d = \frac{|(\mu_1 - \mu_2) - (\mu_1 - \mu_2)_0|}{2\sigma}.$$

When reading the charts, we must use the modified sample size $n^* = 2n - 1$.

Variances Unequal and Unknown — Welch's T-test

Welch-Satterthwaite Relation. Let $X^{(1)}, \ldots, X^{(k)}$ be k independent normally distributed random variables with variances $\sigma_1^2, \ldots, \sigma_k^2$. Let $s_1^2, \ldots, s_k^2$ be sample variances based on samples of sizes $n_1, \ldots, n_k$ from the k populations, respectively. Let $\lambda_1, \ldots, \lambda_k > 0$ be positive real numbers and define

$$\gamma := \frac{(\lambda_1 s_1^2 + \cdots + \lambda_k s_k^2)^2}{\sum_{i=1}^{k} \frac{(\lambda_i s_i^2)^2}{n_i - 1}}.$$

Then

$$\gamma \cdot \frac{\lambda_1 s_1^2 + \cdots + \lambda_k s_k^2}{\lambda_1 \sigma_1^2 + \cdots + \lambda_k \sigma_k^2}$$

follows approximately a chi-squared distribution with γ degrees of freedom, where we round γ down to the nearest integer.

Variances Unequal and Unknown — Welch's T-test

Welch's T-test. Let $X_1^{(i)}, \ldots, X_{n_i}^{(i)}$ with $i = 1, 2$ be samples of sizes n_1 and n_2 from normal distributions with unknown means μ_1, μ_2 and ***unequal*** and ***unknown*** variances σ_1^2, σ_2^2. The test statistic is given by

$$T_\gamma = \frac{\overline{X}^{(1)} - \overline{X}^{(2)} - (\mu_1 - \mu_2)_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}, \qquad \gamma = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\dfrac{(S_1^2/n_1)^2}{n_1 - 1} + \dfrac{(S_2^2/n_2)^2}{n_2 - 1}}$$

We reject at significance level α

- $H_0 : \mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$ if $T_\gamma > t_{\alpha/2,\gamma}$,
- $H_0 : \mu_1 - \mu_2 \leq (\mu_1 - \mu_2)_0$ if $T_\gamma > t_{\alpha,\gamma}$,
- $H_0 : \mu_1 - \mu_2 \geq (\mu_1 - \mu_2)_0$ if $T_\gamma < -t_{\alpha,\gamma}$.

Test for Statistics

- Comparison of Two Means
- Non-parametric Comparisons
- Paired Tests
- Correlation Coefficient

Categorical Data

Wilcoxon Rank-Sum Test

Wilcoxon rank-sum test. Let X and Y be two random populations following some continuous distributions.

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, where $m \leq n$, be random samples from X and Y and associate the rank $R_i, i = 1, \ldots, m+n$, to the R_ith smallest among the $m+n$ total observations. If ties in the rank occur, the mean of the ranks is assigned to all equal values. The test statistic is given by

$$W_m = \text{sum of the ranks of } X_1, \ldots, X_m$$

We reject $H_0 : P[X > Y]$ at significance level α if W_m falls into the corresponding critical region.

Wilcoxon Rank-Sum Test

Wilcoxon rank-sum test. For large values of $m (m \geq 20)$, W_m is approximated normally distributed with

$$\mathrm{E}[W_m] = \frac{m(m+n+1)}{2}, \qquad \mathrm{Var}[W_m] = \frac{mn(m+n+1)}{12}.$$

In case of ties, the variance may be corrected by taking

$$\mathrm{Var}[W_m] = \frac{mn(m+n+1)}{12 - \sum\limits_{\text{groups}} \dfrac{t^3+t}{12}},$$

where the sum is taken over all groups of t ties.

Paired T-Test

Paired T-test. Let $X_1^{(i)}, \ldots, X_{n_i}^{(i)}$ with $i = 1, 2$ be samples of size $n = n_1 = n_2$ from normal distributions with unknown means μ_1, μ_2 and ***equal*** but ***unknown*** variances $\sigma^2 = \sigma_1^2 = \sigma_2^2$. Then $D_i = X_i - Y_i$ follows normal distributions. Then the test statistic is given by

$$T_{n-1} = \frac{\overline{D} - \mu_0}{\sqrt{S_D^2/n}}.$$

We reject at significance level α

- $H_0 : \mu_D = \mu_0$ if $|T_{n-1}| > t_{\alpha/2,n-1}$,
- $H_0 : \mu_D \leq \mu_0$ if $T_{n-1} > t_{\alpha,n-1}$,
- $H_0 : \mu_D \geq \mu_0$ if $T_{n-1} < -t_{\alpha,n-1}$.

Paired vs. Pooled T-Tests

With two populations X and Y with equal variances σ^2, we want to test $H_0 : \mu_X = \mu_Y$ using samples of equal size n. Then the statistics are

$$T_{\text{pooled}} = \frac{\overline{X} - \overline{Y}}{\sqrt{2S_p^2/n}}, \qquad \text{critical value} = t_{\alpha/2, 2n-2},$$

$$T_{\text{paired}} = \frac{\overline{X} - \overline{Y}}{\sqrt{S_D^2/n}}, \qquad \text{critical value} = t_{\alpha/2, n-1}.$$

Preferring a more powerful test, we consider the following.

- $t_{\alpha/2,2n-2} < t_{\alpha/2,n-1}$, smaller critical values $\Rightarrow$ easier to reject.
- $2S_p^2/n$ estimates $2\sigma^2/n$, while S_D^2/n estimates $\sigma_D^2/n = \sigma_{\overline{D}}^2$, where

$$\sigma_{\overline{D}}^2 = \frac{2\sigma^2}{n}(1 - \rho_{\overline{XY}}) = \frac{2\sigma^2}{n}(1 - \rho_{XY}).$$

When $\rho_{XY} > 0$, paired T-test would be more powerful.

Non-parametric Paired Test

Comparison of medians. Let X and Y be two independent random variables that follow the same distribution but differ only in their location, i.e., $X' := X - \delta$ and Y are independent and identically distributed. Then $D = X - Y$ and $2\delta - D$ follow the same distribution. Therefore, D is symmetric about δ.

$$f_D(d - \delta) = f_D(\delta - d).$$

Then we can perform the Wilcoxon signed-rank test on D.

Test for Statistics

Comparison of Two Means

Non-parametric Comparisons

Paired Tests

Correlation Coefficient

Categorical Data

Estimating Correlation

Estimator for correlation. The unbiased estimators for variance and covariance are given by

$$\widehat{\mathrm{Var}[X]} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

$$\widehat{\mathrm{Var}[Y]} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

$$\widehat{\mathrm{Cov}[X, Y]} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}),$$

giving

$$R := \widehat{\rho} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2}\sqrt{\sum(Y_i - \overline{Y})^2}}.$$

Hypothesis Tests for the Correlation Coefficient

Distribution. Suppose (X, Y) follows a bivariate normal distribution with relation coefficient $\rho \in (-1, 1)$. For large sample size n, the Fisher transformation of R

$$\frac{1}{2} \ln \left(\frac{1+R}{1-R} \right) = \mathrm{Artanh}(R)$$

is approximately normal with

$$\mu = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right) = \mathrm{Artanh}(\rho), \qquad \sigma^2 = \frac{1}{n-3}.$$

Hypothesis Tests for the Correlation Coefficient

Confidence interval. A $100(1-\alpha)\%$ confidence interval for ρ is given by

$$\left[\frac{1+R-(1-R)e^{2z_{\alpha/2}/\sqrt{n-3}}}{1+R+(1-R)e^{2z_{\alpha/2}/\sqrt{n-3}}}, \frac{1+R-(1-R)e^{-2z_{\alpha/2}/\sqrt{n-3}}}{1+R+(1-R)e^{-2z_{\alpha/2}/\sqrt{n-3}}}\right]$$

or

$$\tanh\left(\mathrm{Artanh}(R) \pm \frac{z_{\alpha/2}}{\sqrt{n-3}}\right).$$

Hypothesis Tests for the Correlation Coefficient

Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are samples of size n from X and Y, where (X, Y) follows a bivariate normal distribution with relation coefficient $\rho \in (-1, 1)$. The test statistic is given by

$$\begin{aligned} Z &= \frac{\sqrt{n-3}}{2}\left(\ln\left(\frac{1+R}{1-R}\right) - \ln\left(\frac{1+\rho_0}{1-\rho_0}\right)\right) \\ &= \sqrt{n-3}(\mathrm{Artanh}(R) - \mathrm{Artanh}(\rho_0)). \end{aligned}$$

We reject at significance level α

- $H_0 : \rho = \rho_0$ if $|Z| > z_{\alpha/2}$,
- $H_0 : \rho \leq \rho_0$ if $Z > z_\alpha$,
- $H_0 : \rho \geq \rho_0$ if $Z < -z_\alpha$.

The Multinomial Distribution

Definition. A random vector $((X_1, \dots, X_k), f_{X_1 X_2 \cdots X_k})$ with

$$(X_1, \dots, X_k) : S \to \Omega = \{0, 1, 2, \dots, n\}^k$$

and joint distribution function

$$f_{X_1 X_2 \cdots X_k} : \Omega \to \mathbb{R}, \qquad f_{X_1 X_2 \cdots X_k}(x_1, \dots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdot p_k^{x_k},$$

$p_1, \dots, p_k \in (0, 1), n \in \mathbb{N} \setminus \{0\}$ is said to have a ***multinomial distribution*** with parameters n and $p_1, \dots, p_k$. For $i = 1, \dots, k$ and $1 \le i < j \le k$,

$$\mathrm{E}[X_i] = np_i, \quad \mathrm{Var}[X_i] = np_i(1 - p_i), \quad \mathrm{Cov}[X_i, X_j] = -np_i p_j.$$

The Pearson Statistic

Theorem. Let $((X_1, \ldots, X_k), f_{X_1 X_2 \cdots X_k})$ be a multinomial random variable with parameters n and $p_1, \ldots, p_k$. For large n the ***Pearson statistic***

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

follows an approximate chi-squared distribution with $k-1$ degrees of freedom, where O_i are observed values and E_i are expected values.
Cochran's rule. For good approximation, we require

$$\begin{aligned} \mathrm{E}[X_i] &= np_i \geq 1, && \text{for all } i = 1, \ldots, k, \\ \mathrm{E}[X_i] &= np_i \geq 5, && \text{for 80\% of all } i = 1, \ldots, k. \end{aligned}$$

Test for Multinomial Distribution

Pearson's chi-squared goodness-of-fit test. Let $(X_1, \ldots, X_k)$ be a sample of size n from a categorical random variable with parameters $p_1, \ldots, p_k$ satisfying Cochran's Rule. Let $(p_{1_0}, \ldots, p_{k_0})$ be a vector of null values. We want to test

$$H_0 : p_i = p_{i_0}, \qquad i = 1, \ldots, k.$$

based on the test statistic

$$X^2_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_{i_0})^2}{np_{i_0}}.$$

We reject H_0 at significance level α if $X^2_{k-1} > \chi^2_{\alpha,k-1}$.

Goodness-of-Fit Test for a Discrete Distribution

Goodness-of-fit test. Dividing data into k categories to estimate m parameters of distributions, we have the statistic

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

which follows a chi-squared distribution with $k - 1 - m$ degrees of freedom.

Independence of Categorizations

Thanks for your attention!