VE401 Final Part2

Wang Yangyang

UM-SJTU JI

2022 Spring

Outline

1. **Comparison Tests**
   - Comparison of Two Variances
   - Comparison of Two Means
   - Non-Parametric Methods
   - Paired Test, Correlation
2. **Categorical Data**
   - Pearson Statistics and Multinomial Distribution
   - Goodness-of-Fit Test
   - Test for Independence

Outline

1. **Comparison Tests**
   - Comparison of Two Variances
   - Comparison of Two Means
   - Non-Parametric Methods
   - Paired Test, Correlation
2. **Categorical Data**
   - Pearson Statistics and Multinomial Distribution
   - Goodness-of-Fit Test
   - Test for Independence

F-Distribution

Let $\chi^2_{\gamma_1}$ and $\chi^2_{\gamma_2}$ be independent chi-squared random variables with γ_1 and γ_2 degrees of freedom, respectively. Then the random variable

$$F_{\gamma_1,\gamma_2} = \frac{\chi^2_{\gamma_1}/\gamma_1}{\chi_{\gamma_2}/\gamma_2}$$

follows a ***F-distribution*** with γ_1 and γ_2 degrees of freedom
Furthermore,

$$P\left[F_{\gamma_1,\gamma_2} < x\right] = P\left[\frac{1}{F_{\gamma_1,\gamma_2}} > \frac{1}{x}\right] = P\left[F_{\gamma_2,\gamma_1} > \frac{1}{x}\right]$$

Comparing Variances

Let S_1^2 and S_2^2 be sample variances based on independent random samples of sizes n_1 and n_2 drawn from ***normal*** populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively. The test statistic is given by

$$F_{n_1-1,n_2-1} = \frac{S_1^2}{S_2^2}.$$

We reject at significance level α

- $H_0 : \sigma_1 \leqslant \sigma_2$ if $S_1^2/S_2^2 > f_{\alpha,n_1-1,n_2-1}$,
- $H_0 : \sigma_1 \geqslant \sigma_2$ if $S_2^2/S_1^2 > f_{\alpha,n_2-1,n_1-1}$,
- $H_0 : \sigma_1 = \sigma_2$ if $S_1^2/S_2^2 > f_{\alpha/2,n_1-1,n_2-1}$ or $S_2^2/S_1^2 > f_{\alpha/2,n_2-1,n_1-1}$.

OC curve. The abscissa is defined by

$$\lambda = \frac{\sigma_1}{\sigma_2}.$$

Outline

1. **Comparison Tests**
   - Comparison of Two Variances
   - Comparison of Two Means
   - Non-Parametric Methods
   - Paired Test, Correlation

2. **Categorical Data**
   - Pearson Statistics and Multinomial Distribution
   - Goodness-of-Fit Test
   - Test for Independence

Basic Cases

For two ***Normally Distributed*** Populations:

- $X^{(1)} \sim N\left(\mu_1, \sigma_1^2\right)$
- $X^{(2)} \sim N\left(\mu_2, \sigma_2^2\right)$

Goal: compare μ_1 and μ_2.
Three Basic Cases:

- σ_1^2 and σ_2^2 are known
- σ_1^2 and σ_2^2 are unknown but $\sigma_1^2 = \sigma_2^2$
- σ_1^2 and σ_2^2 are unknown and not necessarily equal

Variance 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

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

We reject at significance level α

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

Variance Known

When testing equality of means $H_0 : \mu_1 = \mu_2$, we have $(\mu_1 - \mu_2)_0 = 0$. We can use the OC curves for normal distributions with

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

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

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

Variance Equal but Unknown

Variances equal but unknown. Let $X_1^{(i)}, \dots, X_{n_i}^{(i)}$ 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

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

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

Variance Equal but Unknown

OC curve. When testing equality of means $H_0 : \mu_1 = \mu_2$, we have $(\mu_1 - \mu_2)_0 = 0$. We can use the OC curves for the T-test in case of ***equal*** sample sizes $n = n_1 = n_2$

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

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

Variance Not Necessarily Equal and 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 ***not necessarily equal and unknown*** variances σ_1^2, σ_2^2. The test statistic is given by

$$T_\gamma = \frac{\bar{X}^{(1)} - \bar{X}^{(2)} - (\mu_1 - \mu_2)_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}, \quad \gamma = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}}$$

We reject at significance level α

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

Variance Not Necessarily Equal and Unknown

Remarks:

- Round γ down to the nearest integer.
- No simple OC curves for Welch's test.
- ***!!!*** It is not a good idea to pre-test for equal variances and then make a decision whether to use Student's or Welch's test. ***!!!*** It is fine to test for normality, equality of variances or other properties and then to gather ***new data*** for a comparison of means test. But using the ***same data*** creates serious problems.
- When variances are unknown, current recommendations are to always use Welch's test.

Outline

1. Comparison Tests
   - Comparison of Two Variances
   - Comparison of Two Means
   - Non-Parametric Methods
   - Paired Test, Correlation
2. Categorical Data
   - Pearson Statistics and Multinomial Distribution
   - Goodness-of-Fit Test
   - Test for Independence

Wilcoxon Rank-Sum Test

Let X and Y be two random samples following some continuous distributions. Decide whether to reject the null hypothesis

$$H_0 : P[X > Y] = \frac{1}{2} \quad \text{or} \quad H_0 : P[X > Y] \leqslant \frac{1}{2}$$

Procedures:

1. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, $m \leqslant n$, be random samples from X and Y and associate the rank R_i, $i = 1, \ldots, m+n$, to the R_i th smallest among the $m+n$ total observations. If ties in the rank occur, the mean of the ranks is assigned to all equal values.
2. Sum up the ranks of smaller samples. Then the test based on the statistic

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

is called the Wilcoxon rank-sum test.

Wilcoxon Rank-Sum Test

We reject $H_0 : P[X > Y] = 1/2$ at significance level α if

- for small m : W_m falls into the corresponding critical region, or
- for large $m (m \geqslant 20)$: perform a Z-test, since W_m is approximately normally distributed with

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

If there are many ties, the variance may be corrected by taking

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

where the sum is taken over all groups of t ties (not always a good way).

Outline

1. **Comparison Tests**
   - Comparison of Two Variances
   - Comparison of Two Means
   - Non-Parametric Methods
   - Paired Test, Correlation
2. **Categorical Data**
   - Pearson Statistics and Multinomial Distribution
   - Goodness-of-Fit Test
   - Test for Independence

Paired Tests for Mean

Comparing means (or the location) of two related populations X and Y. Method: Pair the samples as $D = X - Y$.

- Set the hypothesis as, i.e.,

$$H_0 : \mu_D = \mu_X - \mu_Y = (\mu_X - \mu_Y)_0 = \mu_{D0}$$

- Then use a ***T-test*** for D is called a paired T-test for X and Y

$$T_{n-1} = \frac{\bar{D} - \mu_{D_0}}{\sqrt{S_D^2/n}}$$

Paired vs. Pooled T-Tests

Assume that we have two populations of normally distributed random variables X and Y with equal variances σ^2. We want to test

$$H_0 : \mu_X - \mu_Y = (\mu_X - \mu_Y)_0$$

Then we could either perform a paired test or a pooled test. Which is more powerful? Let us compare the test statistics:

$$T_{\text{pooled}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)_0}{\sqrt{2S_p^2/n}}, \qquad \text{critical value } = t_{\alpha/2,2n-2}$$

$$T_{\text{paired}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)_0}{\sqrt{S_D^2/n}}, \qquad \text{critical value } = t_{\alpha/2,n-1}$$

Compare the two denominators, which estimate

$$\frac{2\sigma^2}{n} \quad \text{with} \quad \frac{\sigma_D^2}{n}$$

Paired vs. Pooled T-Tests

Conclusion: From $\frac{\sigma_D^2}{n} = \frac{2\sigma^2}{n}(1 - \rho_{XY})$ we see

- If $\rho_{XY} > 0$, paired T-test is more powerful. The denominator of the paired statistic will be smaller than that of the pooled statistic, leading to a larger value of the statistic.
- If ρ_{XY} is zero (or even negative), pairing is unnecessary and pooled T-test is more powerful. The reason is that it is easier to reject H_0 when comparing with $t_{\alpha/2,2n-2}$ than with $t_{\alpha/2,n-1}$.

$\Rightarrow$ *Positive correlation* makes a paired T-test more powerful.

Test for Correlation Coefficient

First, find the estimation of ρ. Since

$$\widehat{\operatorname{Var}[X]]} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

$$\widehat{\operatorname{Cov}[X, Y]} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

The natural choice for an estimator for the correlation coefficient is then

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

Test for Correlation Coefficient

- Hypothesis test: We can test $H_0 : \rho = \rho_0$, by ***Z-test***, using the test statistic

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

- Confidence interval: A $100(1-\alpha)\%$ confidence interval for ρ,

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

or

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

Outline

1. Comparison Tests
   - Comparison of Two Variances
   - Comparison of Two Means
   - Non-Parametric Methods
   - Paired Test, Correlation
2. Categorical Data
   - Pearson Statistics and Multinomial Distribution
   - Goodness-of-Fit Test
   - Test for Independence

Categorical Random Variables

A random variable X that can take on the values $1, \ldots, k$ with respective probabilities $p_1 \ldots, p_k$ as above. A random sample of size n from X is collected and the results are expressed as a *random vector*

$$(X_1, X_2, \ldots, X_k) \quad \text{with} \quad X_1 + X_2 + \cdots + X_k = n$$

The Multinomial Distribution: A random vector $\left((X_1, \ldots, X_k), f_{X_1 X_2 \cdots X_k}\right)$ where

$$f_{X_1 X_2 \cdots X_k}(x_1, \ldots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

$p_1, \ldots, p_k \in (0,1), n \in \mathbb{N} \backslash \{0\}$ is said to have a multinomial distribution with parameters n and $p_1, \ldots, p_k$

- $\mathrm{E}[X_i] = np_i, \quad i = 1, \ldots, k$
- $\operatorname{Var}[X_i] = np_i(1 - p_i), i = 1, \ldots, k$
- $\operatorname{Cov}[X_i, X_j] = -np_ip_j, 1 \leqslant i < j \leqslant k$

The Pearson Statistics

Let $\left((X_1, \ldots, X_k), f_{X_1X_2\cdots X_k}\right)$ 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 (because we have $k-1$ independent cells).
Cochran's Rule: This tell us how large n needs to be for the chi-squared distribution to be a good approximation to the true distribution of the Pearson statistic when

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

Outline

1 Comparison Tests
- Comparison of Two Variances
- Comparison of Two Means
- Non-Parametric Methods
- Paired Test, Correlation

2 Categorical Data
- Pearson Statistics and Multinomial Distribution
- Goodness-of-Fit Test
- Test for Independence

Test for Multinomial Distribution

Let $(X_1, \ldots, X_k)$ be a sample of size n from a categorical random variable with parameters $(p_1, \ldots, p_k)$. We perform the chi-squared goodness-of-fit test.
Note: In this test, we directly make assumptions on parameters p_i ***without estimation based on samples***. This may happen when we already have some prior knowledge of the distribution (e.g. PRNG).

Procedures

1. Set
$$H_0 : p_i = p_{i_0}, \quad i = 1, \ldots, k$$
2. Calculate the expected values
$$E_i = np_{i0}$$
Then test whether the ***Cochran's rule*** is satisfied.

Test for Multinomial Distribution

Procedures

3. If satisfied, calculate the Pearson statistic.

$$X_{k-1}^2 = \sum_{i=1}^{k} \frac{(X_i - np_{i0})^2}{np_{i0}}$$

which follows a chi-squared distribution with

degrees of freedom: independent cells $-m = k - 1$

independent cells = k - 1, m=0

4. We reject H_0 at significance level α if $X_{k-1}^2 > \chi_{\alpha,k-1}^2$.

Goodness-of-Fit Test for Discrete Distribution

Now, we calculate the ***estimates for parameters***, to make assumptions indirectly

Procedures

1. Suppose we guess that data ***follow some distribution***, so we set the hypothesis as

$$H_0 : \text{A specific distribution with unknown parameter } p_i$$
$$\text{e.g., } H_0 : \text{ A Poisson distribution with parameter } k.$$

2. Estimate parameters p_i from the sample based on your hypothesis. e.g., for ***Poisson distribution***, estimate k by $\widehat{k} = \bar{X}$.
   Then we can calculate p_i. Suppose we have three categories $x = 0, x = 1, x \geqslant 2$, then

$$p_0 = P[X = 0] = \frac{e^{-\widehat{k}}\widehat{k}^0}{0!}, \quad p_1 = P[X = 1] = \frac{e^{-\widehat{k}}\widehat{k}^1}{1!},$$

$$p_2 = P[X \geqslant 2] = 1 - P[X = 0] - P[X = 1]$$

Goodness-of-Fit Test for Discrete Distribution

Procedures

3. Calculate the expected values $E_i = np_i$. Then make a table by yourself as below,

Category i	Exp. Frequency E_i	Obs. Frequency O_i
0	np_0	x_0
1	np_1	x_1
2	np_2	x_2
$\dots$	$\dots$	$\dots$

4. Test whether the ***Cochran's Rule*** is satisfied. If not satisfied, then go back to procedure (ii) and (iii) and change your number of categories.

Goodness-of-Fit Test for Discrete Distribution

Procedures

5. If (iv) satisfied, calculate the ***Pearson statistic***.

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

now follows a chi-squared distribution with

$$\# \text{ independent cells } - m = k - 1 - m$$

degrees of freedom, where m is the ***number of parameters that we estimate***. e.g., for the previous Poisson distribution test with 3 categories,

$$k = 3, m = 1$$

6. We reject H_0 at significance level α if X^2 exceeds the critical value.

Outline

1. Comparison Tests
   - Comparison of Two Variances
   - Comparison of Two Means
   - Non-Parametric Methods
   - Paired Test, Correlation
2. Categorical Data
   - Pearson Statistics and Multinomial Distribution
   - Goodness-of-Fit Test
   - Test for Independence

Test for Independence

We define the marginal row and column sums

$$n_i = \sum_{j=1}^{c} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$$

For a contingency table as below,

	column 1	column 2	column 3	
row 1	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
row 2	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
row 3	n_{31}	n_{32}	n_{33}	$n_{3\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n

Test for Independence

Procedures

1. If the hypothesis is that row and column categorizations are independent, then it should be the case that

$$H_0 : p_{ij} = p_i \cdot p_j$$

2. Estimates for the row and column probabilities are $\widehat{p_i.} = \frac{n_{i.}}{n}, \widehat{p_{.j}} = \frac{n_{.j}}{n}$, so if H_0 is assumed,

$$\widehat{p_{ij}} = \widehat{p_{i.}} \cdot \widehat{p_{.j}} = \frac{n_i \cdot n_{.j}}{n^2}$$

3. Calculate the expected values

$$E_{ij} = n \cdot \widehat{p_{ij}} = \frac{n_i \cdot n \cdot j}{n}$$

Then test whether the ***Cochran's rule*** is satisfied.

Test for Independence

Procedures

4. If satisfied, calculate the Pearson statistic.

$$X^2_{(r-1)(c-1)} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

which follows a ***chi-squared distribution*** with degrees of freedom:

$$\#\text{independent cells } - m = rc - 1 - (r - 1 + c - 1) = (r-1)(c-1)$$

- $\#\text{independent cells} = rc - 1$
- $m = r - 1 + c - 1$

5. We reject H_0 if the value of $X^2_{(r-1)(c-1)}$ exceeds the critical value.

Test for Comparing Proportions

Now the row totals are fixed, rewrite the table in terms of proportions:

	column 1	column 2	column 3	
row 1	p_{11}	p_{12}	p_{13}	$p_{1.} = 1$ (fixed)
row 2	p_{21}	p_{22}	p_{23}	$p_{2.} = 1$ (fixed)
row 3	p_{31}	p_{32}	p_{33}	$p_{3.} = 1$ (fixed)
row 4	p_{41}	p_{42}	p_{43}	$p_{4} = 1$ (fixed)

We want to compare proportions from each row, so

$$H_0 : \begin{cases} p_{11} = p_{21} = p_{31} = p_{41} \\ p_{12} = p_{22} = p_{32} = p_{42} \\ p_{13} = p_{23} = p_{33} = p_{43} \end{cases}$$

Test for Comparing Proportions

Procedure

1. Supposing that H_0 is true,

$$p_j := p_{1j} = p_{2j} = p_{3j} = p_{4j}$$

where p_j is the proportion of all objects following into the j th column. If H_0 is assumed, estimates for the column proportions are

$$\widehat{p_j} = \frac{n_{.j}}{n}$$

2. Calculate the expected values

$$E_{ij} = n_i \cdot \widehat{p_{ij}} = n_i \cdot \widehat{p_j} = \frac{n_i \cdot n_{.j}}{n}$$

Then test whether the ***Cochran's rule*** is satisfied.

Test for Comparing Proportions

Procedure

1. If satisfied, calculate the Pearson statistic.

$$X^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

which follows a ***chi-squared*** distribution with degrees of freedom:

$$\#\text{ independent cells } - m = r(c-1)-(c-1) = (r-1)(c-1)$$

- $\#$independent cells $= r(c-1)$
- $m = (c-1)$

2. We reject H_0 if the value of $X^2_{(r-1)(c-1)}$ exceeds the critical value.