Big Data Fundamentals and Applications

Statistical Analysis (V)
Test of Normality

Asst. Prof. Chan, Chun-Hsiang

Master program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Applied Artificial Intelligence, Chung Yuan Christian University, Taoyuan, Taiwan

Outlines

1. Introduction
2. Road Map of Statistical Analysis
3. Hypothesis Testing
4. Type I and Type II Errors
5. Reliability & Validity Analyses
6. Inferential Statistics
7. Test of Normality
8. Differences between Parametric and Nonparametric Statistics
9. Parametric Statistics
10. Nonparametric Statistics
11. Correlation Analysis
12. Question Time

Inferential Statistics

Inferential Statistics

- After a series of data checking, we finally are able to compare one feature to another, or do a comparison between several features.
- First, there are two major parts in the statistical tests : **categorical** and **continuous** data.
- Second, we will introduce **parametric** and **nonparametric** statistical tests.

Test of Normality

Test of Normality

- For some statistical analyses, the assumption of normality is necessary; therefore, here, we will introduce statistical analyses for normality before you proceed with further analysis.
- **Shapiro–Wilk test**
- **Kolmogorov-Smirnov test**
- **Pearson chi-squared test**
- Lilliefors test

Shapiro–Wilk Test

- The Shapiro–Wilk test is a test of normality in frequentist statistics. **The Shapiro–Wilk test tests the null hypothesis that a sample $x_1, \ldots, x_n$ came from a normally distributed population.** The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, where\ x_i$$

(with parentheses enclosing the subscript index i; not to be confused with x_i) is the ith order statistic, i.e., the ith-smallest number in the sample; $\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n), 8 \leq n \leq 50$ is the sample mean.

Source: https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

Source: https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test
Source: http://blog.excelmasterseries.com/2015/05/how-to-create-completely-automated_4.html

Shapiro–Wilk Test

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$ is the sample mean.

- The coefficient a_i are given by: $(a_1, \dots, a_n) = \frac{m^T V^{-1}}{C}$, where C is a vector norm: $C = \|V^T m\| = \sqrt{m^T V^{-1} V^{-1} m}$ and the vector m, $m = (m_1, \dots, m_n)^T$ is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution; finally, V is the covariance matrix of those normal order statistics.

Source: https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test
Source: http://blog.excelmasterseries.com/2015/05/how-to-create-completely-automated_4.html

Shapiro–Wilk Test

$\alpha = 0.05$			$n = 15$			Data Pairs					
Raw Data		Sorted Data		a value		Upper value		Lower value		Difference	a*Difference
1	20	1	18	a1	0.5150	#15	22	#01	18	4	2.0600
2	19	2	18	a2	0.3306	#14	21	#02	18	3	0.9918
3	18	3	18	a3	0.2495	#13	21	#03	18	3	0.7485
4	19	4	18	a4	0.1878	#12	21	#04	18	3	0.5634
5	22	5	19	a5	0.1353	#11	20	#05	19	1	0.1353
6	18	6	19	a6	0.0880	#10	20	#06	19	1	0.0880
7	21	7	19	a7	0.0433	#09	19	#07	19	0	0.0000
8	19	8	19								
9	21	9	19								
10	18	10	20								
11	18	11	20								
12	19	12	21								
13	20	13	21								
14	21	14	21								
15	19	15	22								

$\sum_{i=1}^{n} a_i x_{(i)}$	4.59	$\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2$	21.0681
$\sum_{i=1}^{n} (x_i - \bar{x})^2$	23.733	W	0.886541
		W critical	0.881

Kolmogorov-Smirnov Test

- The Kolmogorov–Smirnov test (K-S test or KS test) is a **nonparametric test** of the equality of continuous (or discontinuous), one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test), where n is larger than 50.
- **The null distribution of this statistic is calculated under the null hypothesis that the sample is drawn from the reference distribution (in the one-sample case) or that the samples are drawn from the same distribution (in the two-sample case).**

Source: https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

Kolmogorov-Smirnov Test

- The two-sample K–S test is one of the most useful and general **nonparametric** methods for comparing two samples, as it is **sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples**.

$$F_n = \frac{number\ of\ (elements\ in\ the\ sample\ \leq x)}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i)\,,$$

$$where 1_{(-\infty,x]}(X_i)\ is\ the\ indicator\ function, equal\ to\ 1\ if X_i \leq 1\ and\ qual\ to\ 0\ otherwise.$$

- The Kolmogorov-Smirnov statistic for a given cumulative distribution function $F(x)$ is

$$D_n = \sup_x |F_n(x) - F(x)|\,,$$

$$where\ \sup_x\ is\ the\ supermum\ of\ the\ set\ of\ distances.$$

- Intuitively, the statistic takes the largest absolute difference between the two distribution functions across all x values.

Pearson Chi-squared Test

- Suppose that n observations in a random sample from a population are classified into k mutually exclusive classes with respective observed numbers x_i (for $i = 1,2,\dots,k$), and a null hypothesis gives the probability p_i that an observation falls into the ith class. So we have the expected numbers $m_i = np_i$ for all i, where

$$\sum_{i=1}^{k} p_i = 1, \sum_{i=1}^{k} m_i = n \sum_{i=1}^{k} p_i = n$$

Source: https://en.wikipedia.org/wiki/Chi-squared_test

Pearson Chi-squared Test

- Pearson proposed that, under the circumstance of the null hypothesis being correct, as $n \to \infty$ the limiting distribution of the quantity given below is the χ^2 distribution.

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i} = \sum_{i=1}^{k} \frac{x_i^2}{m_i} - n$$

Source: https://en.wikipedia.org/wiki/Chi-squared_test

Reading

Nonparametric Correlation Techniques: Techniques for Correlating Nominal & Ordinal Variables
https://staff.blog.ui.ac.id/r-suti/files/2010/05/noparcorelationtechniq.pdf
Parametric and Non-parametric tests for comparing two or more groups
https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/parametric-nonparametric-tests
多重比較分析檢定
http://amebse.nchu.edu.tw/new_page_534.htm
單向 ANOVA：事後檢定
https://www.ibm.com/docs/zh-tw/spss-statistics/beta?topic=anova-one-way-post-hoc-tests

Question Time

If you have any questions, please do not hesitate to ask me.

The End

Thank you for your attention))