Big Data Fundamentals and Applications

Statistical Analysis (VI)

Parametric Statistics

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

Differences between Parametric and Nonparametric Statistics

Differences between Parametric and Nonparametric statistics

- **Parametric** statistics are based on assumptions about the distribution of population from which the sample was taken. **Nonparametric** statistics are not based on assumptions, that is, the data can be collected from a sample that does not follow a specific distribution.
- Common parametric statistics are, for example, the Student's t-tests. Common nonparametric statistics are, for example, the Mann-Whitney-Wilcoxon (MWW) test or the Wilcoxon test.

Source: https://www.ibm.com/docs/en/db2woc?topic=procedures-statistics-parametric-nonparametric

Parametric Statistics

F-Test

https://www.cuemath.com/data/inferential-statistics/

- **Goal:** test if there is a difference between the variances of two samples or populations
- **Assumption:** normal distribution and sample size ≥ 30.
- **Null hypothesis (H_0):** $\sigma_1^2 = \sigma_2^2$
- **Alternative hypothesis (H_1):** $\sigma_1^2 > \sigma_2^2$
- **Test statistics:** $f = \frac{\sigma_1^2}{\sigma_2^2}$, where σ_1^2 and σ_2^2 are the variance of the first and second population, respectively.
- **Decision Criteria:** f statistic > f critical value

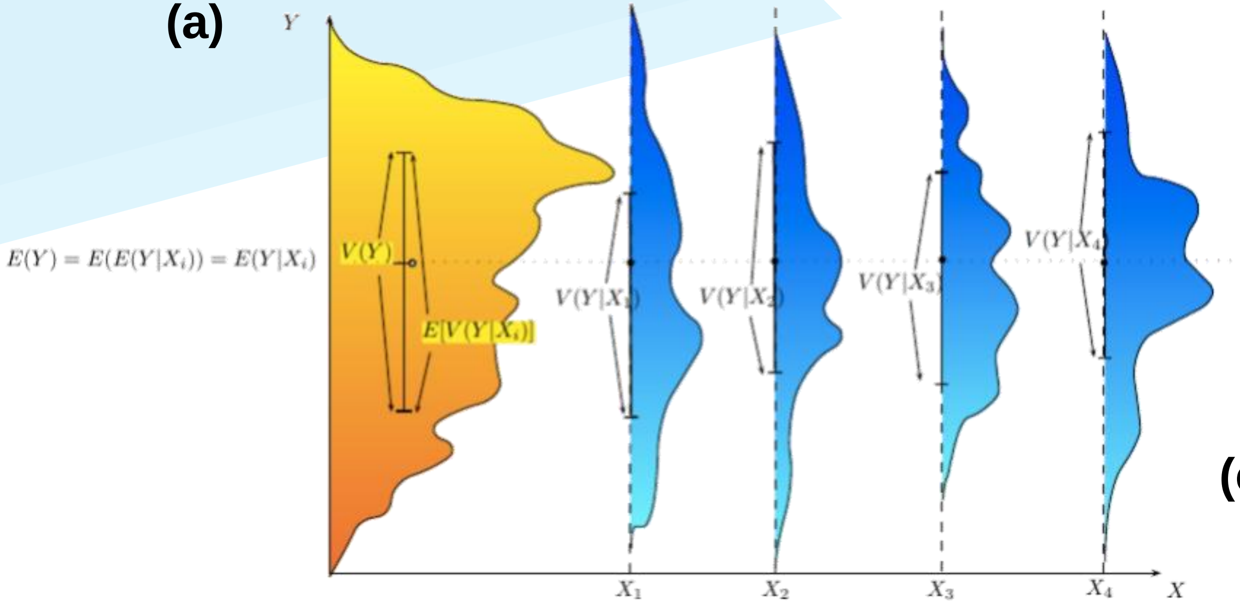
ANOVA

- **Analysis of variance** (**ANOVA**) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among means.
- ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the *t*-test beyond two means. In other words, the ANOVA is used to test the difference between two or more means.
- **Goal:** test if means of each group are equal ($\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$)
- **Null hypothesis ($\boldsymbol{H_0}$):** $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$
- **Alternative hypothesis ($\boldsymbol{H_1}$):** at least one μ_i is different

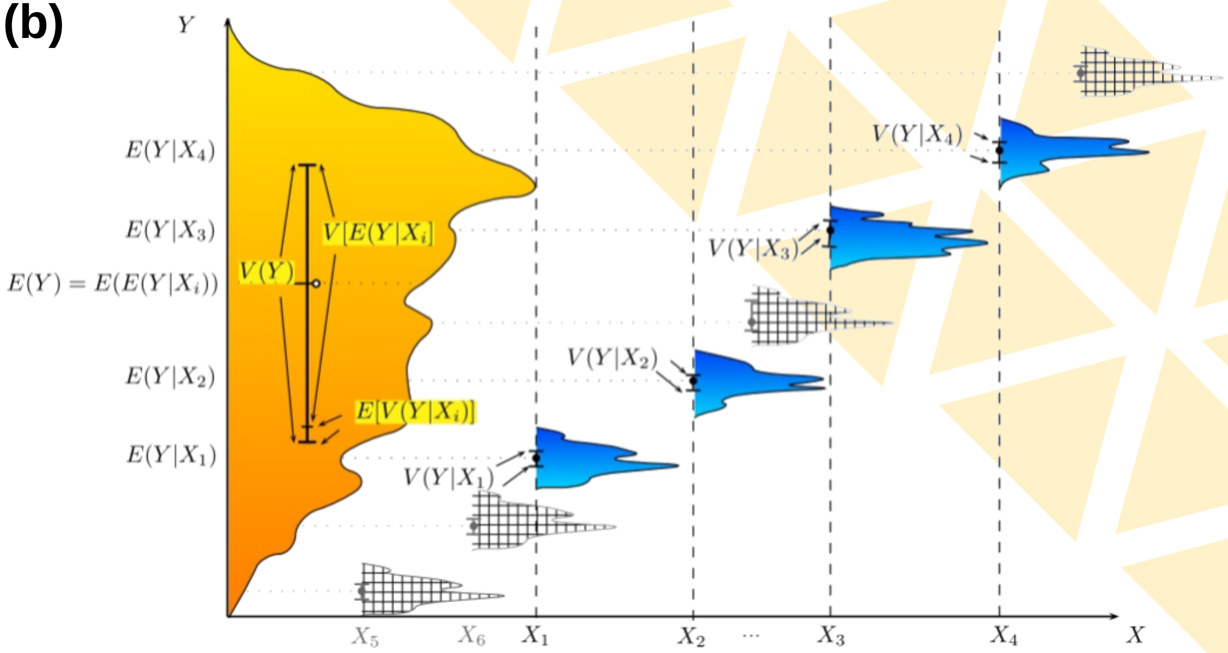
Source: https://en.wikipedia.org/wiki/Analysis_of_variance

ANOVA

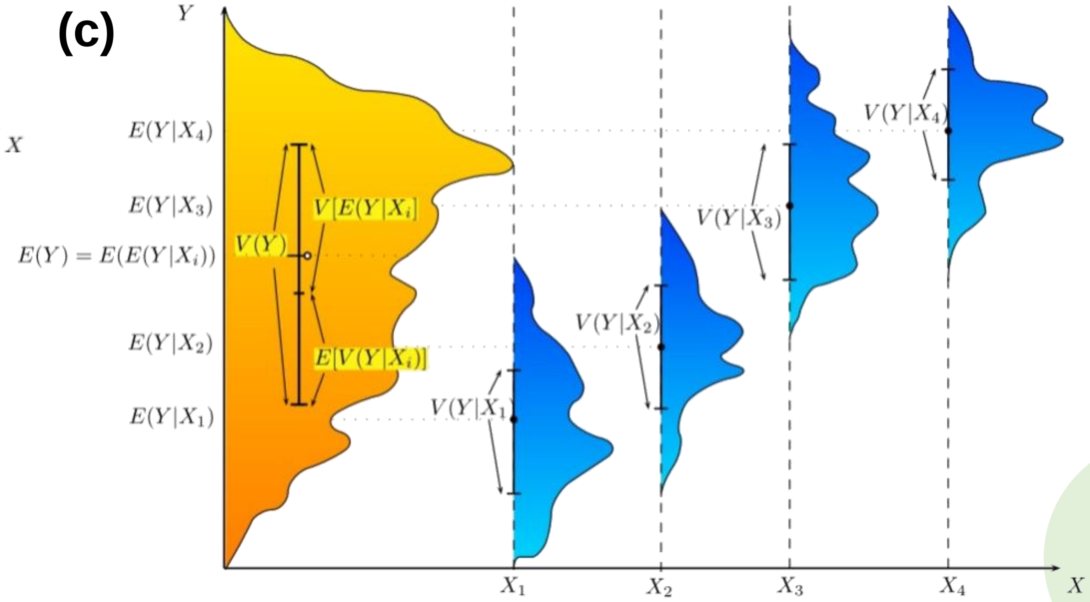
(a)



(b)



(c)



Source: https://en.wikipedia.org/wiki/Analysis_of_variance

ANOVA

Fixed-effects models	The fixed-effects model (class I) of analysis of variance applies to situations in which the experimenter applies one or more treatments to the subjects of the experiment to see whether the response variable values change. This allows the experimenter to estimate the ranges of response variable values that the treatment would generate in the population as a whole.
Random-effects models	Random-effects model (class II) is used when the treatments are not fixed. This occurs when the various factor levels are sampled from a larger population. Because the levels themselves are random variables, some assumptions and the method of contrasting the treatments (a multi-variable generalization of simple differences) differ from the fixed-effects model.
Mixed-effects models	A mixed-effects model (class III) contains experimental factors of both fixed and random-effects types, with appropriately different interpretations and analysis for the two types.

ANOVA – Summary of Assumptions

- The normal-model based ANOVA analysis assumes the independence, normality, and homogeneity of variances of the residuals.
- The randomization-based analysis assumes only the homogeneity of the variances of the residuals (as a consequence of unit-treatment additivity) and uses the randomization procedure of the experiment. Both these analyses require homoscedasticity, as an assumption for the normal-model analysis and as a consequence of randomization and additivity for the randomization-based analysis.

ANOVA – Partitioning of the Sum of Squares

- ANOVA uses traditional standardized terminology. The definitional equation of sample variance is

$$s^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$$

- The fundamental technique is a partitioning of the total sum of squares SS into components related to the effects used in the model.

$$SS_{Total} = SS_{treatments} + SS_{Error}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}\dots\right)^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}\dots)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y_{i\cdot}}\right)^2$$

ANOVA – F-Test

- The *F*-test is used for comparing the factors of the total deviation. For example, in one-way, or single-factor ANOVA, statistical significance is tested for by comparing the F test statistic.

$$F = \frac{variance\ between\ treatments}{variance\ within\ treatments}$$

$$F = \frac{MS_{Treatments}}{MS_{Error}} = \frac{\frac{SS_{Treatments}}{I-1}}{\frac{SS_{Error}}{n_T - 1}}$$

where MS is mean square, I is the number of treatments and n_T is the total number of cases.

ANOVA – Result Table

	Sum of Square Error	Degree of Freedom	Mean of Square Error	F-test
Between Treatments	$SS_{Teaetments}$	k-1	$SS_{Teaetments}/(k-1)$	$MS_{Treatment}/MS_{Error}$
Within Treatments	SS_{Error}	N-k	$SS_{Error}/(N-k)$	
Total	SS	N-1		

ANOVA – Multiple-Comparison Procedure

- ANOVA only tells us there is at least one mean value is different from the others; however, we cannot retrieve that group-to-group differences in this analysis.
- Therefore, multiple-comparison procedure could statistically demonstrate the relationship between two groups.
- A very straightforward thinking, using t-test to compare each two groups that demonstrate the multiple comparison; however, it will arise multiple Type I error, an increase in α, and the results cannot be believed.

ANOVA – Multiple-Comparison Procedure

Fisher's **L**east **S**ignificant **D**ifference	**Bonferroni**	**Tukey H**onestly **S**ignificant **D**ifference
• Uses *t* tests to perform all pairwise comparisons between group means. • No adjustment is made to the error rate for multiple comparisons. • Easy to occur Type I error	• Uses t tests to perform pairwise comparisons between group means, but controls overall error rate by setting the error rate for each test to the experimentwise error rate divided by the total number of tests. • The observed significance level is adjusted for the fact that multiple comparisons are being made. • Modified Type I error • Divide the raw significance level by the number of tests	• Sorting groups by their mean in an ascending order. • Uses the Studentized range statistic to make all of the pairwise comparisons between groups. • Sets the experimentwise error rate at the error rate for the collection for all pairwise comparisons. • Equal sample sizes • Observations are independent • Mean is from normal distribution • Equal variation across observations

ANOVA – Multiple-Comparison Procedure

Scheffe	Duncan	Dunnett
• Performs simultaneous joint pairwise comparisons for all possible pairwise combinations of means. • Uses the F sampling distribution. • Can be used to examine all possible linear combinations of group means, not just pairwise comparisons. • Highest threshold • Difficult to occur Type II error • Suitable for groups with different numbers of samples, and samples with non-normality	• Makes pairwise comparisons using a stepwise order of comparisons identical to the order used by the Student-Newman-Keuls test, but sets a protection level for the error rate for the collection of tests, rather than an error rate for individual tests. • Uses the Studentized range statistic.	• Compare control and treatment groups. • Pairwise multiple comparison *t* test that compares a set of treatments against a single control mean. The last category is the default control category. Alternatively, you can choose the first category. • **2-sided** tests that the mean at any level (except the control category) of the factor is not equal to that of the control category. • **< Control** tests if the mean at any level of the factor is smaller than that of the control category. **>** • **Control** tests if the mean at any level of the factor is greater than that of the control category.

Z Test

- **Goal:** test if sample mean and population mean are equal when population variance is known
- **Assumption:** normal distribution and sample size ≥ 30.
- **Null hypothesis (H_0):** $\mu = \mu_0$
- **Alternative hypothesis (H_1):** $\mu > \mu_0$
- **Test statistics:** $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$, where $\bar{x}$ is the sample mean, μ is the population mean, σ is the population standard deviation, and n is the sample size.
- **Decision Criteria:** Z statistic > Z critical value

One-sample T Test

- **Goal:** test if sample mean and population mean are equal when population variance is **un**known
- **Assumption:** student t distribution and sample size < 30.
- **Null hypothesis (H_0):** $\mu = \mu_0$
- **Alternative hypothesis (H_1):** $\mu > \mu_0$
- **Test statistics:** $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$, where $\bar{x}$ is the sample mean, μ is the population mean, s is the sample standard deviation, and n is the sample size.
- **Decision Criteria:** t statistic > t critical value

T-test

- In addition to one-sample t-test, there are three types of t-test, including paired t-test, two-sample independent t-test (assume that the variance of two samples or populations are [not] equal).
- In the following slides, we will give some examples to show the differences between them.

Question X

How we determine the variances between two samples or populations are equal or not?

Paired T-test

- If the two samples or populations are from a matched or paired sources, or from a replicated measurement, then you need to select paired t-test.

$$t = \frac{\overline{X_D} - \mu_0}{\frac{s_D}{\sqrt{n}}}$$

$\overline{X_D}$ *and* s_D are the average and standard deviation of the differences between all pairs, the constant μ_0 is zero if we want to test whether the average of the difference is significantly different, and n is the number of pairs.

Source: https://en.wikipedia.org/wiki/Student%27s_t-test
Source: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired-sample-t-test/
Source: https://www.omnicalculator.com/statistics/t-test#p-value-from-t-test

Two-sample Independent T-test

- If the variance of two samples or populations are equal (or very similar).

$$t = \frac{\overline{x_1} - \overline{x_2} - \mu_0}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

s_p is the pooled standard, defined by

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Source: https://en.wikipedia.org/wiki/Student%27s_t-test
Source: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired-sample-t-test/
Source: https://www.omnicalculator.com/statistics/t-test#p-value-from-t-test

Two-sample Independent T-test

- If the variance of two samples or populations are **unequal** (or very similar), referring to Welch's t-test.

$$t = \frac{\overline{x_1} - \overline{x_2} - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Source: https://en.wikipedia.org/wiki/Student%27s_t-test
Source: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired-sample-t-test/
Source: https://www.omnicalculator.com/statistics/t-test#p-value-from-t-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
第二章 多重比較的方法
https://ah.nccu.edu.tw/bitstream/140.119/33900/6/35400806.pdf
多重比較 Multiple comparisons
https://researcher20.com/2010/05/27/%E5%A4%9A%E9%87%8D%E6%AF%94%E8%BC%83-multiple-comparisons/
One-way ANOVA: Post hoc tests
https://www.ibm.com/docs/en/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))