Big Data Fundamentals and Applications

Statistical Analysis (III)

Hypothesis Testing

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

Introduction
Road Map of Statistical Analysis
Hypothesis Testing
Type I and Type II Errors
Reliability & Validity Analyses
Inferential Statistics
Test of Normality
Differences between Parametric and Nonparametric Statistics
Parametric Statistics
Nonparametric Statistics
Correlation Analysis

Introduction

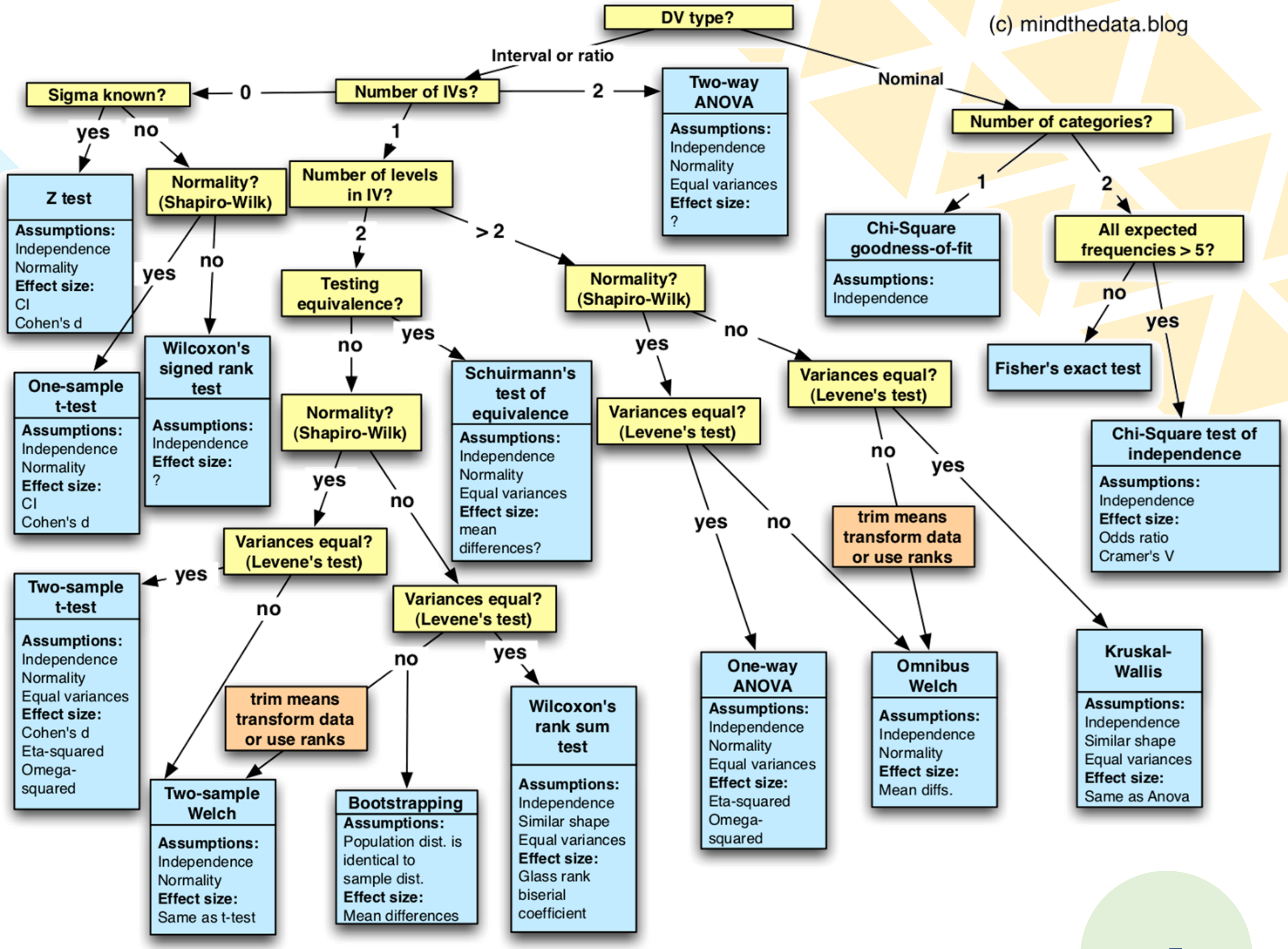
Introduction

- Before we introduce the inferential statistics, it is important to understand the fundamental concept of statistics, parametric and nonparametric, type I and type II error, reliability and validity, normality and independence, homoscedasticity and heteroscedasticity.
- Here, inferential statistics cover three section: parametric statistics, nonparametric statistics, and correlation analysis.
- The following three slides demonstrate the road map/ cheat sheet/ decision tree of inferential statistics.

Introduction
Road Map of Statistical Analysis
Hypothesis Testing
Type I and Type II Errors
Reliability & Validity Analyses
Inferential Statistics
Test of Normality
Differences between Parametric and Nonparametric Statistics
Parametric Statistics
Nonparametric Statistics
Correlation Analysis

Road Map of Statistical Analysis

Decision Tree



Source: https://mindthedata.blog/2020/05/14/statistics-cheat-sheet-part-1/

Choice of Statistical Tests

Table 1 Choice of statistical test from paired or matched observation

Variable	Test
Nominal	McNemar's Test
Ordinal (ordered categories)	Wilcoxon
Quantitative (discrete or non-normal)	Wilcoxon
Quantitative (normal)	Paired t-test

Table 2 Parametric and nonparametric tests for comparing two or more groups

Parametric Test	Situation	Nonparametric Test
t-test	Two independent population	Wilconxon rank sum test
t-test		Mann-Whitney U test
One way analysis of variance	Three or more populations	Kruskal Wallis test
Paired t-test	Paired population	Sign test
		Wilconxon rank sign test
Pearson correlation	Correlation	Spearman correlation

Source: https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/parametric-nonparametric-tests

Choice of Statistical Tests

Table 3 Choice of statistical test for independent observations

		Outcome variable					
Input variable		**Nominal**	**Categorical (>2)**	**Ordinal**	**Quantitative Discrete**	**Quantitative Non-normal**	**Quantitative Normal**
	Nominal	χ^2 or Fisher's	χ^2	χ^2-trend or Mann-Whitney	Mann-Whitney	Mann-Whitney or log-rank[a]	T-test
	Categorical (>2)	χ^2	χ^2	Kruskal-Wallis[b]	Kruskal-Wallis[b]	Kruskal-Wallis[b]	ANOVA[c]
	Ordinal	χ^2-trend or Mann-Whitney	[e]	Spearman rank	Spearman rank	Spearman rank	Spearman rank or Linear regression[d]
	Quantitative Discrete	Logistic regression	[e]	[e]	Spearman rank	Spearman rank	Spearman rank or Linear regression[d]
	Quantitative Non-normal	Logistic regression	[e]	[e]	[e]	Plot data and Pearson or Spearman rank	Plot data and Pearson or Spearman rank and Linear regression
	Quantitative Normal	Logistic regression	[e]	[e]	[e]	Linear regression[d]	Pearson or Linear regression

[a] If data are censored. [b] The Kruskal-Wallis test is used for comparing ordinal or non-Normal variables for more than two groups, and is a generalisation of the Mann-Whitney U test. [c] Analysis of variance is a general technique, and one version (one way analysis of variance) is used to compare Normally distributed variables for more than two groups, and is the parametric equivalent of the Kruskal-Wallistest. [d] If the outcome variable is the dependent variable, then provided the residuals (the differences between the observed values and the predicted responses from regression) are plausibly Normally distributed, then the distribution of the independent variable is not important. [e] There are a number of more advanced techniques, such as Poisson regression, for dealing with these situations. However, they require certain assumptions and it is often easier to either dichotomise the outcome variable or treat it as continuous.

Source: https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/parametric-nonparametric-tests

Cheat Sheet

<table>
<tr><th colspan="3" rowspan="3"></th><th colspan="4">Criterion / Measure / Dependent Variable</th></tr>
<tr><th colspan="2">Categorical</th><th colspan="2">Continuous</th></tr>
<tr><th>1 Variable, 2 Categories</th><th>1 Variable, >2 Categories</th><th>1 Variable</th><th>>1 Variable</th></tr>
<tr><td rowspan="7">Predictor / Covariate / Independent Variable</td><td rowspan="6">Categorical</td><td>1 Variable
2 Categories
Between-subjects</td><td colspan="2" rowspan="4">χ² Test
(Crosstabs → Statistics → ☑ Chi-square)</td><td>Independent t Test
(Compare Means → Independent-Samples)</td><td rowspan="2"></td></tr>
<tr><td>1 Variable
2 Categories
Within-subjects</td><td>Paired t Test
(Compare Means → Paired Samples)</td></tr>
<tr><td>1 Variable
>2 Categories
Between-subjects</td><td>One-Way ANOVA
(Compare Means → One-way ANOVA)</td><td>One-Way MANOVA
(General Linear Model → Multivariate → Add Dependent Variables)</td></tr>
<tr><td>1 Variable
>2 Categories
Within-subjects</td><td>Repeated Measures ANOVA
(General Linear Model → Repeated Measures → Add Within-Sbj Factors)</td><td>Repeated Measures MANOVA
(Repeated Measures ANOVA → Add Measures)</td></tr>
<tr><td>>1 Variable
All Categorical
Between-subjects</td><td rowspan="3">Binomial
Logistic Regression
with Categorical Predictors
(Regression → Binary Logistic → Categorical Covariates)</td><td rowspan="3">Multinomial
Logistic Regression
(Regression → Multinomial Logistic)</td><td>Factorial ANOVA
(General Linear Model → Univariate → Add Fixed Factors)</td><td>Factorial MANOVA
(One-Way MANOVA → Add Fixed Factors)</td></tr>
<tr><td>>1 Variable
All Categorical
Mixed Within- &
Between-subjects</td><td>Mixed-Design ANOVA
(Repeated Measures ANOVA → Add Between-Sbj Factors)</td><td>Mixed-Design MANOVA
(Mixed-Design ANOVA → Add Measures)</td></tr>
<tr><td></td><td>>1 Variable
Mixed Categorical
& Continuous</td><td>One-Way ANCOVA
(One-Way ANOVA → Add Covariates)</td><td>One-Way MANCOVA
(One-Way MANOVA → Add Covariates)</td></tr>
</table>

Source: https://towardsdatascience.com/demystifying-statistical-analysis-1-a-handy-cheat-sheet-b6229bf992cf

Introduction
Road Map of Statistical Analysis
Hypothesis Testing
Type I and Type II Errors
Reliability & Validity Analyses
Inferential Statistics
Test of Normality
Differences between Parametric and Nonparametric Statistics
Parametric Statistics
Nonparametric Statistics
Correlation Analysis

Hypothesis Testing

Hypothesis Testing

- A **statistical hypothesis test** is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis. Hypothesis testing allows us to make probabilistic statements about population parameters.
- The usual line of reasoning is as follows:
  1. There is an initial research hypothesis of which the truth is unknown.
  2. The first step is to state the relevant **null** and **alternative hypotheses**. This is important, as mis-stating the hypotheses will muddy the rest of the process.
  3. The second step is to consider the **statistical assumptions** being made about the sample in doing the test; for example, assumptions about the **statistical independence** or about **the form of the distributions** of the observations. This is equally important as invalid assumptions will mean that the results of the test are invalid.

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

Hypothesis Testing

4. Decide which test is appropriate, and state the relevant **test statistic** *T*.
5. Derive the distribution of the test statistic under the null hypothesis from the assumptions. In standard cases this will be a well-known result. For example, the test statistic might follow a Student's t distribution with known degrees of freedom, or a normal distribution with known mean and variance. If the distribution of the test statistic is completely fixed by the null hypothesis we call the hypothesis simple, otherwise it is called composite.
6. Select a significance level (*α*), a probability threshold below which the null hypothesis will be rejected. Common values are 5% and 1%.

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

Hypothesis Testing

7. The distribution of the test statistic under the null hypothesis partitions the possible values of T into those for which the null hypothesis is rejected—the so-called *critical region*—and those for which it is not. The probability of the critical region is α. In the case of a composite null hypothesis, the maximal probability of the critical region is α.
8. Compute from the observations the observed value t_{obs} of the t-test statistic.
9. Decide to either reject the null hypothesis in favor of the alternative or not reject it. The decision rule is to reject the null hypothesis H_0 if the observed value t_{obs} is in the critical region, and not to reject the null hypothesis otherwise.

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

Hypothesis Testing

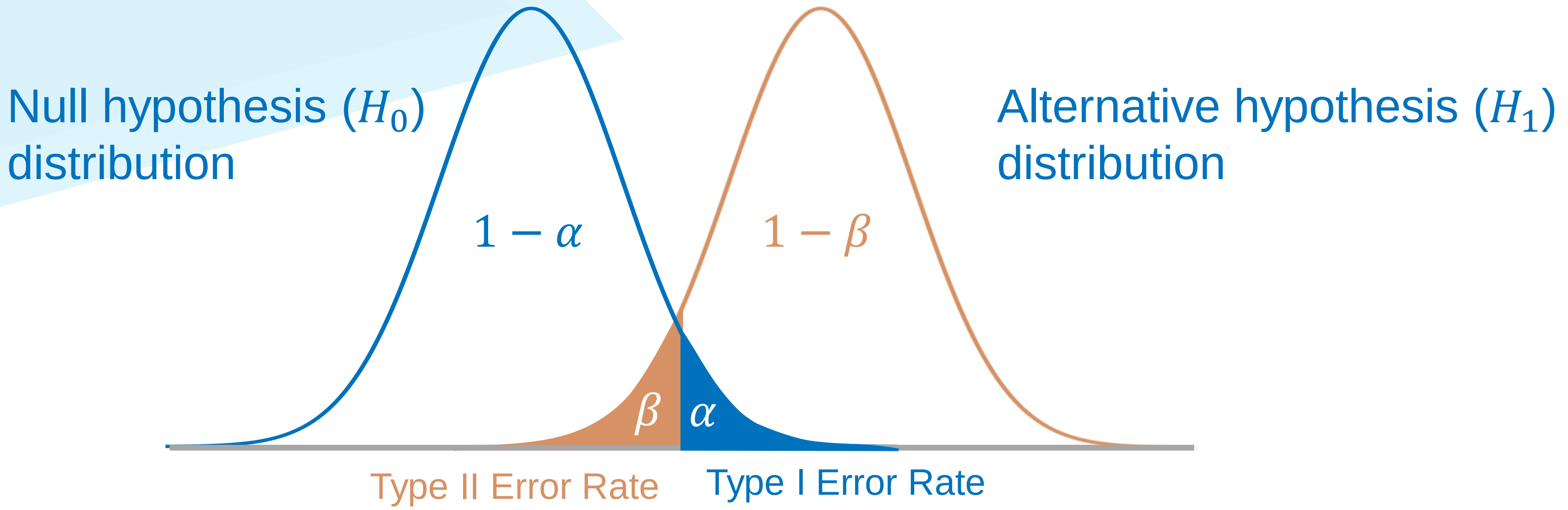
- A common alternative formulation of this process goes as follows:
  1. Compute from the observations the observed value t_{obs} of the test statistic *T*.
  2. Calculate the *P* value. This is the probability, under the null hypothesis, of sampling a test statistic at least as extreme as that which was observed (the maximal probability of that event, if the hypothesis is composite).
  3. Reject the null hypothesis, in favor of the alternative hypothesis, if and only if the *p*-value is less than (or equal to) the significance level (the selected probability) threshold (α), for example 0.05 or 0.01.

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

Introduction
Road Map of Statistical Analysis
Hypothesis Testing
Type I and Type II Errors
Reliability & Validity Analyses
Inferential Statistics
Test of Normality
Differences between Parametric and Nonparametric Statistics
Parametric Statistics
Nonparametric Statistics
Correlation Analysis

Type I and Type 2 Errors

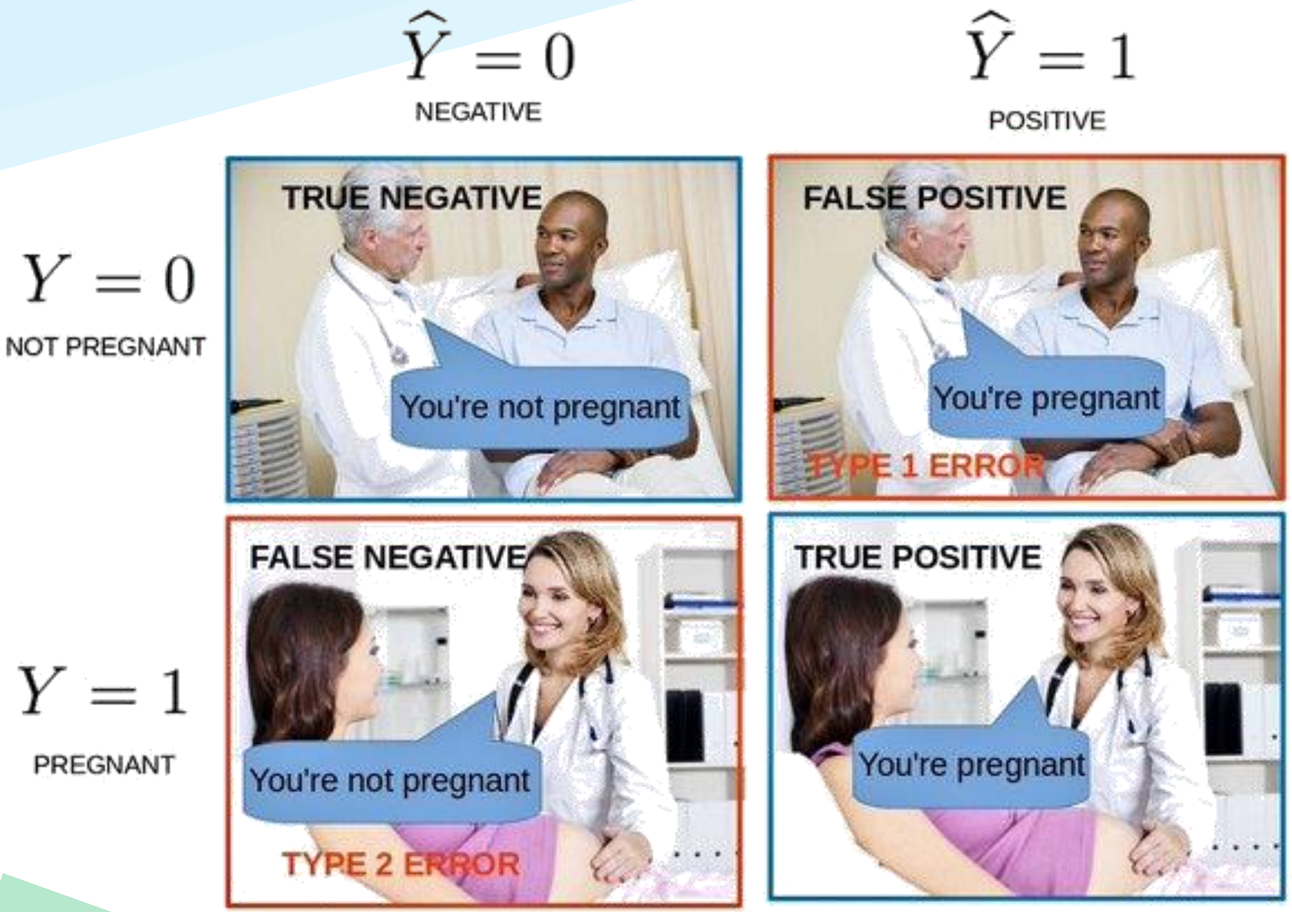
Type I and Type II Errors

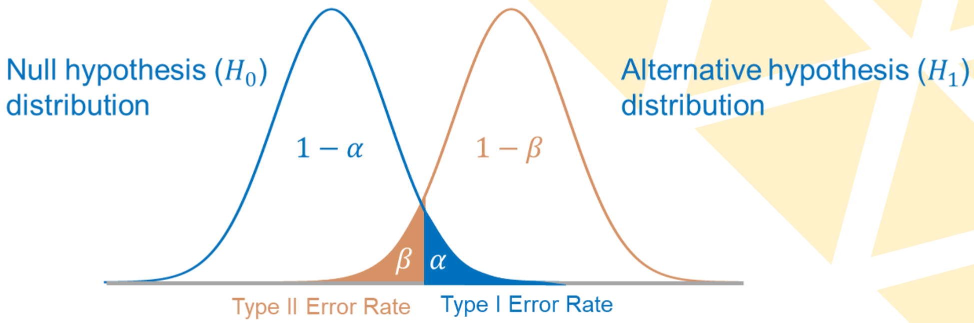


		Actual	
	Null hypothesis …	**H_0 is False**	**H_0 is True**
Prediction	**Rejected H_0**	Correct Decision **True Positive** $Probability = 1-\beta$	**Type I Error** **False Positive** $Probability = \alpha$
	Accepted H_0	**Type II Error** **False Negative** $Probability = \beta$	Correct Decision **True Negative** $Probability = 1-\alpha$

Type I and Type II Errors

		Actual	
	Null hypothesis …	H_0 is False	H_0 is True
Prediction	Rejected H_0	Correct Decision True Positive $Probability = 1 - \beta$	Type I Error False Positive $Probability = \alpha$
	Accepted H_0	Type II Error False Negative $Probability = \beta$	Correct Decision True Negative $Probability = 1 - \alpha$





Hint: This is the "confusion matrix". We will use this matrix to demonstrate the fit performance of the model. The matrix calculates several metrics, such as recall, precision, F1 score, and precision.

Photo source: https://dzone.com/articles/understanding-the-confusion-matrix

Question Time

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

The End

Thank you for your attention))