

# TEACHING MATERIALS 2

## INTRODUCTION TO ADVANCED STATISTICS

### AIMS FOR THIS CLASS

- Hypothesis Testing and Statistical Test Selection
- Regression Analysis
- PCA

### HYPOTHESIS TESTING AND STATISTICAL TEST SELECTION

Testing of Statistical hypotheses is a remarkable aspect of statistical theory, which helps us to make decisions where there is a lack of uncertainty. There are many real-life situations where we would like to take a decision for further action. Further, there are some problems, for which we would like to determine whether the claims are acceptable or not. The statistical tests for testing the significance of the difference between the hypothesized value and the actual value of the sample statistic or the difference between any set of sample statistics are called tests of significance.

### TEST SELECTION

In Figure 1 one can follow a decision tree towards finding a suited basic statistical test. Before conducting a test, usually check the general assumptions, so that your data do not violate any of these.

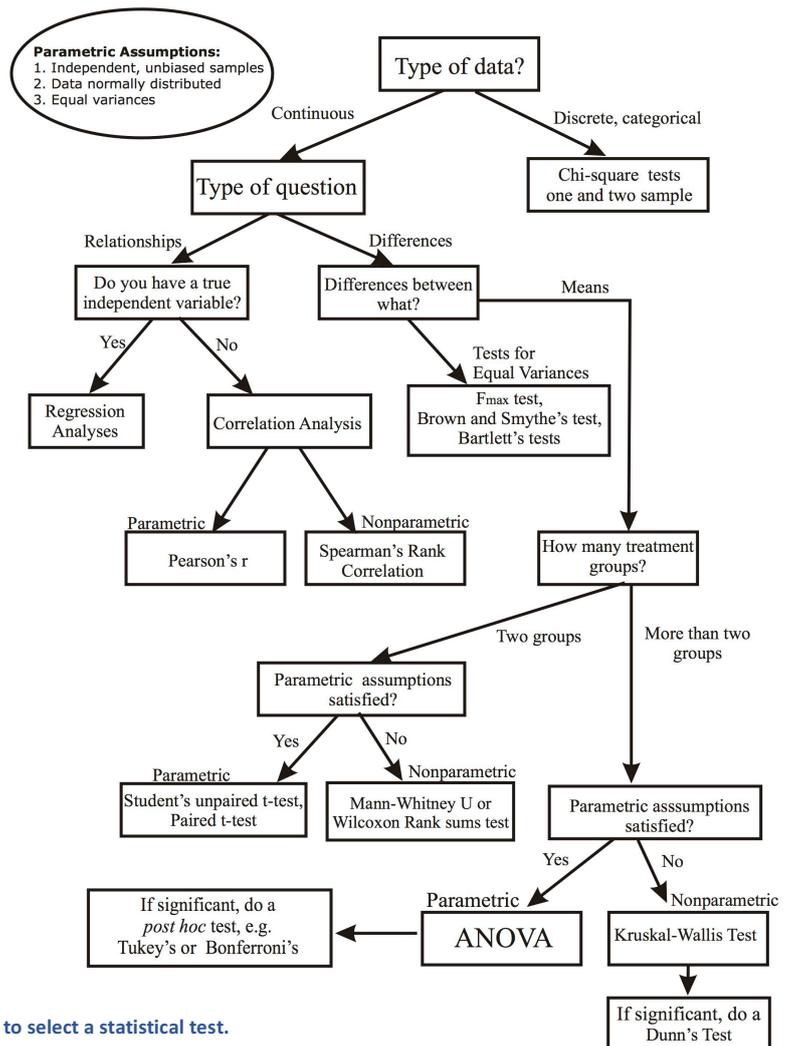


Figure 1: One of the many possible ways how to select a statistical test.

## TUTORIALS FOR TEST SELECTION

---

- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116565/>
- <http://imsear.li.mahidol.ac.th/bitstream/123456789/159737/1/ach2014v2n2p54.pdf>

## REGRESSION ANALYSIS IN R

Regression analysis is used for explaining or modelling the relationship between a single variable  $Y$ , called the response, output or dependent variable, and one or more predictor, input, independent or explanatory variables,  $X_1 \dots X_p$ . When  $p=1$ , it is called simple regression but when  $p>1$  it is called multiple regression or sometimes multivariate regression. When there is more than one  $Y$ , then it is called multivariate multiple regression which we won't be covering here.

## LITERATURE AND TUTORIALS ON LINEAR REGRESSION

---

- [https://books.google.at/books?id=YDgoDAAAQBAJ&dq=linear+regression+R&lr=&hl=de&source=gbs\\_navlinks\\_s](https://books.google.at/books?id=YDgoDAAAQBAJ&dq=linear+regression+R&lr=&hl=de&source=gbs_navlinks_s)
- <https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf>
- <http://www.cyclismo.org/tutorial/R/linearLeastSquares.html>

### EXAMPLES:

*# General formula of the influence of a predictor variable on the dependent variable*

```
fit <- lm(formula, data)
```

*The formula in R is typically written as  $Y \sim x_1 + x_2 + \dots + x_k$*

*where  $\sim$  separates the dependent variable( $y$ ) on the left from independent variables( $x_1, x_2, \dots, x_k$ ) from right, and the independent variables are separated by  $+$  signs.*

## PCA IN R

Principal component analysis (PCA) is a dimensionality reduction technique that is widely used in data analysis. Reducing the dimensionality of a dataset can be useful in different ways. For example, our ability to visualize data is limited to 2 or 3 dimensions. Lower dimension can sometimes significantly reduce the computational time of some numerical algorithms. Besides, many statistical models suffer from high correlation between covariates, and PCA can be used to produce linear combinations of the covariates that are uncorrelated between each other.

## LITERATURE AND TUTORIALS ON PCA

---

- <https://tgmstat.wordpress.com/2013/11/28/computing-and-visualizing-pca-in-r/>
- Venables, W. N., Ripley, B. D. R. Modern applied statistics with S-PLUS. Springer-Verlag. (Section 11.1)
- [https://cran.r-project.org/web/packages/HSAUR/vignettes/Ch\\_principal\\_components\\_analysis.pdf](https://cran.r-project.org/web/packages/HSAUR/vignettes/Ch_principal_components_analysis.pdf)
- <https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/>

### EXAMPLES:

*# General formula*

```
prcomp(x, scale = FALSE)
```

```
princomp(x, cor = FALSE, scores = TRUE)
```

*Arguments for prcomp():*

*x : a numeric matrix or data frame*

## 180148 VO+UE Tools in Cognitive Science II: Basic Statistics for Cognitive Scientists (2017S)

*scale* : a logical value indicating whether the variables should be scaled to have unit variance before the analysis takes place

*Arguments for princomp()* :

*x* : a numeric matrix or data frame

*cor* : a logical value. If TRUE, the data will be centered and scaled before the analysis

*scores* : a logical value. If TRUE, the coordinates on each principal component are calculated