Nov 14, 2022

Theoretical Hypothesis Testing

Week 6: Analysis of Variance (ANOVA)

Introduction

Analysis of variance (ANOVA) in its simplest form analyzes if **the mean of a normal distributed random variable differs in a number of groups**.

The groups are given by applying different treatments to subjects, for example, in designed experiments in technical applications or in clinical studies.

The problem can thereby be seen as comparing group means, which extends the t-test to **more than two groups**.

The assumptions of an ANOVA is **the homogeneity of variance** within all groups and **the normal distributed variables**.

1. One way ANOVA

Assume that $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$, $\quad i \in 1,2,...,k$ be k independent samples of independent **normal distributed random variables** with the **same variance** but possibly different group means.

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

$H_A : \mu_i \neq \mu_j$ for at least one $i \neq j$.

1. One way ANOVA

$$T_F = \frac{\sum_{i=1}^{k} n_i(\hat{Y}_{i.} - \hat{Y})/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \hat{Y}_i)^2/(n-1)}$$ where $n = \sum_{i=1}^{k} n_i$

Reject H_0 if the observed value of the test statistic $T_F > f_{(1-\alpha;k-1;n-1)}$

Application

To test if the means of the harvest in kilograms of tomatoes in three different greenhouses differ. The dataset contains observations from five fields in each greenhouse. Assume that the variance of the three different greenhouses are homogeneous.

house	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
fertilizer	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
kg	0.51	0.25	0.64	0.22	1.05	0.99	0.40	0.94	0.06	0.42	1.13	0.43	0.22	0.25	1.81

Application

Hypotheses:

$H_0 : \mu_1 = \mu_2 = \mu_3$ (The **mean** of the harvest in kilograms of tomatoes in three different greenhouses is equal.)

$H_A : \mu_i \neq \mu_j$ for at least one $i \neq j$ (At least one of the **mean** of the harvest in kilograms of tomatoes in three different greenhouses is different.)

Decision:

Interpretation:

1.1. Testing of variance heterogeneity

Bartlett test is used for testing the variances of k **normal distributed populations** differ from each other.

$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ (The variances of the k populations are homogeneous)

$H_A : \sigma_i^2 \neq \sigma_j^2$ for at least one $i \neq j$. (At least one of the variances of the k populations is not homogenous or heterogenous.)

Application

To test the variance of the three different greenhouses are homogeneous.

<u>**Hypotheses:**</u>

$H_0 :$

$H_A :$

Application

<u>Decision:</u>

<u>Interpretation:</u>

1.2. Testing of normality

Shapiro-Wilk test is used for testing the normality of k populations.

Hypotheses:

$H_0 : X \sim N(\mu, \sigma^2)$

$H_A : X \nsim N(\mu, \sigma^2)$

2. Kruskal-Wallis test

Kruskal-Wallis test is the non-parametric counterpart of the ANOVA test, and also the extension of the Wilcoxon rank-sum test for more than two independent populations.

If one of the assumptions of the ANOVA test is not met, it is less riskier to use Kruskal-Wallis test.

$$H_0 : M_1 = M_2 = \ldots = M_k$$

$H_A : M_i \neq M_j$ for at least one $i \neq j$.

Application

To test if the **medians** of the harvest in kilograms of tomatoes in three different greenhouses differ.

Hypotheses:

$H_0 : M_1 = M_2 = M_3$ (The **median** of the harvest in kilograms of tomatoes in three different greenhouses is equal.)

$H_A : M_i \neq M_j$ for at least one $i \neq j$ (At least one of the **median** of the harvest in kilograms of tomatoes in three different greenhouses is different.)

Decision:

Interpretation: