Statistics

Lecture 11

ANOVA:
Analysis of Variance
(for a single factor)

SILESIAN UNIVERSITY
SCHOOL OF BUSINESS ADMINISTRATION IN KARVINA

David Bartl
Statistics
INM/BASTA

Outline of the lecture

- Analysis of Variance (ANOVA) for a single factor
- Bartlett's test

Two-sample *t*-test for the diff. of the pop. means // $\sigma_X = \sigma_Y$

First of all, recall the two-sample *t*-test for the difference of the population means (assuming the same variance):

Let $X \sim \mathcal{N}(\mu_X, \sigma^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$ be two unknown random variables.

We assume that both random variables X and Y are normally distributed,

but we do not know their population means μ_X and μ_Y nor their variance,

but we do assume that the variance σ^2 of both variables X and Y is the same.

We sample the variable X m-times, so we have the sample $x_1, x_2, \ldots, x_m$.

We sample the variable Y n-times, so we have the sample $y_1, y_2, \ldots, y_n$.

Two-sample *t*-test for the diff. of the pop. means // $\sigma_X = \sigma_Y$

Having the m observations $x_1, x_2, \dots, x_m$ of the random variable $X \sim \mathcal{N}(\mu_X, \sigma^2)$ and having the n observations $y_1, y_2, \dots, y_n$ of the random variable $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$, we test the null hypothesis that both population means are the same (H_0: $\mu_X = \mu_Y$) against the two-sided alternative hypothesis (H_1: $\mu_X \neq \mu_Y$).

Calculate the statistic

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}} \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

Two-sample *t*-test for the diff. of the pop. means // $\sigma_X=\sigma_Y$

Finish the two-sample *t*-test for the difference of the population means as follows:

- choose **the level of significance**, a small number $\alpha > 0$, a very popular value is $\alpha = 5\ \%$, other popular values are 10 % or 1 % or 0.1 % etc.
- find the **critical value** $c > 0$ so that

$$\int_{-\infty}^{-c} f(x)\,\mathrm{d}x + \int_{+c}^{+\infty} f(x)\,\mathrm{d}x = \alpha$$

  where f is the density of the *t*-distribution with $m+n-2$ degrees of freedom
- if $T \in (-\infty, -c] \cup [+c, +\infty)$, **the critical region**, then **<u>reject</u>** the null hypothesis
- if $T \in (-c, +c)$, then **<u>do not reject</u>** (or <u>fail to reject</u>) the null hypothesis

One-Way ANOVA

One-way ANOVA

Motivation:

Let $Y_1 \sim \mathcal{N}(\mu_1, \sigma^2)$, $Y_2 \sim \mathcal{N}(\mu_2, \sigma^2)$, etc., $Y_k \sim \mathcal{N}(\mu_k, \sigma^2)$ be unknown random variables.

We assume that the random variables $Y_1, Y_2, \dots, Y_k$ are normally distributed, but we do not know their population means $\mu_1, \mu_2, \dots, \mu_k$ nor their variance, but we do assume that the variance σ^2 of all variables $Y_1, Y_2, \dots, Y_k$ is the same.

One-way ANOVA

We sample the variable Y_1 n_1-times, so we have the sample

$$y_{11}, y_{12}, y_{13}, \dots, y_{1n_1}$$

We sample the variable Y_2 n_2-times, so we have the sample

$$y_{21}, y_{22}, \dots, y_{2n_2}$$

We sample the variable Y_3 n_3-times, so we have the sample

$$y_{31}, y_{32}, y_{33}, y_{34}, y_{35}, \dots, y_{3n_3}$$

Etc.

We sample the variable Y_k n_k-times, so we have the sample

$$y_{k1}, y_{k2}, \dots, y_{kn_k}$$

One-way ANOVA

Having the samples $y_{11}, y_{12}, \dots, y_{1n_1}$, $y_{21}, y_{22}, \dots, y_{2n_2}$, etc., $y_{k1}, y_{k2}, \dots, y_{kn_k}$ of the random variables $Y_1 \sim \mathcal{N}(\mu_1, \sigma^2)$, $Y_2 \sim \mathcal{N}(\mu_2, \sigma^2)$, etc., $Y_k \sim \mathcal{N}(\mu_k, \sigma^2)$, respectively, we formulate the **null hypothesis**:

all samples come from the same population:
the values of the population means are the same

$$H_0: \quad \mu_1 = \mu_2 = \dots = \mu_k$$

Recall that we do not know the true population means $\mu_1, \mu_2, \dots, \mu_k$.

We only test the hypothesis by means of the samples of the measurements.

One-way ANOVA

Example I:

We have got a gross sample of n patients cured for some disease.

The patients were divided into k groups of sizes $n_1, n_2, \dots, n_k$ so that

$$n = n_1 + n_2 + \cdots + n_k$$

The 1st group has been treated by the 1st method.

The 2nd group has been treated by the 2nd method.

Etc.

The k^{th} group has been treated by the k^{th} method.

One-way ANOVA

Example I:

Then $y_{11}, y_{12}, \dots, y_{1n_1}$, $y_{21}, y_{22}, \dots, y_{2n_2}$, etc., $y_{k1}, y_{k2}, \dots, y_{kn_k}$

are the results of a medical test after the treatment.

Based on the samples, we test the null hypothesis that

the results of all the treatments are (on average) the same.

One-way ANOVA

Example II:

We test k distinct cars. We test the 1st car n_1 times, we test the 2nd car n_2 times, etc., and we test the k^{th} car n_k times for mileage.

Then $y_{11}, y_{12}, \dots, y_{1n_1}$, $y_{21}, y_{22}, \dots, y_{2n_2}$, etc., $y_{k1}, y_{k2}, \dots, y_{kn_k}$ are the results of the measurements, i.e. the mileages.

We test the null hypothesis that the average mileage of each car is the same.

One-way ANOVA

Remark:

If $k=2$, then we can equivalently use the two-sample t-test for the difference of the means (with the assumption of the same variance) to test the null hypothesis.

If the number of the groups is larger $(k>2)$ and we apply the two sample t-test to all the pairs of the groups $(1–2, 1–3, \dots, 1–k,\ 2–3, \dots, 2–k,$ etc., $(k-1)–k)$ separately, then the probability of the error cumulates and is then much larger than the originally prescribed α=5% !!!

One-way ANOVA

We have got k groups of observations of a quantitative (numerical) data item:

The values in the 1st group are: $y_{11}, y_{12}, y_{13}, \dots, y_{1n_1}$

The values in the 2nd group are: $y_{21}, y_{22}, \dots, y_{2n_2}$

The values in the 3rd group are: $y_{31}, y_{32}, y_{33}, y_{34}, y_{35}, \dots, y_{3n_3}$

Etc.

The values in the kth group are: $y_{k1}, y_{k2}, \dots, y_{kn_k}$

Recall our assumption that the samples come from normally distributed random variables $Y_1, Y_2, \dots, Y_k$ with the same variance σ^2.

One-way ANOVA

The one-way analysis of variance (ANOVA) proceeds as follows:

Having the samples $y_{11}, y_{12}, \dots, y_{1n_1}$, $y_{21}, y_{22}, \dots, y_{2n_2}$, etc., $y_{k1}, y_{k2}, \dots, y_{kn_k}$, calculate the

- **sample variance between the groups**
- **sample variance within the groups**

One-way ANOVA

Sample variance (mean squares) <u>between</u> the groups:

sum of squares (between)

mean squares (between)

degrees of freedom (between)

$$\mathrm{MS_B} = \frac{\mathrm{SS_B}}{\mathrm{DF_B}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2}{k-1}$$

where

- n_i is the size of the i-th group
- $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ is the sample mean of the i-th group
- $\bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$ is the grand sample mean
- $n = \sum_{i=1}^{k} n_i$ is the size of the grand sample

One-way ANOVA

Sample variance (mean squares) between the groups:

The traditional ANOVA terminology is used:

Sum of Squares (between):

$$SS_B = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{k} n_i \times (\bar{y}_i - \bar{y})^2$$

Degrees of Freedom (between):

$$DF_B = k - 1$$

One-way ANOVA

Sample variance (mean squares) between the groups:

The traditional ANOVA terminology is used:

Mean Squares (between):

$$\mathrm{MS_B} = \frac{\mathrm{SS_B}}{\mathrm{DF_B}} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2}{k-1}$$

Observe intuitively:

The more the null hypothesis $(\mu_1 = \mu_2 = \cdots = \mu_k)$ holds true,

the more the mean squares $\mathrm{MS_B}$ tend to zero: $\mathrm{MS_B} \longrightarrow 0$

One-way ANOVA

Sample variance (mean squares) <u>within</u> the groups:

sum of squares (within)

mean squares (within)

degrees of freedom (within)

$$MS_W = \frac{SS_W}{DF_W} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2}{n - k}$$

where

- n_i is the size of the i-th group
- $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ is the sample mean of the i-th group
- $n = \sum_{i=1}^{k} n_i$ is the size of the grand sample

One-way ANOVA

Sample variance (mean squares) within the groups:

The traditional ANOVA terminology is used:

Sum of Squares (within):

$$SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

Degrees of Freedom (within):

$$DF_W = \sum_{i=1}^{k} (n_i - 1) = n - k$$

One-way ANOVA

Sample variance (mean squares) within the groups:

The traditional ANOVA terminology is used:

Mean Squares (within):

$$MS_W = \frac{SS_W}{DF_W} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2}{n-k}$$

Observe intuitively:

The more the mean squares MS_W tend to zero $(MS_W \longrightarrow 0)$,

the less the null hypothesis $(\mu_1 = \mu_2 = \cdots = \mu_k)$ holds true.

One-way ANOVA

<u>**Theorem:**</u>

If $Y_{11}, Y_{12}, \dots, Y_{1n_1}, \; Y_{21}, Y_{22}, \dots, Y_{2n_2}, \; \dots, \; Y_{k1}, Y_{k2}, \dots, Y_{kn_k} \sim \mathcal{N}(\mu, \sigma^2)$ are independent,

then

$$\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2} \Big/ \frac{k-1}{n-k} \sim F_{k-1,\, n-k}$$

where $F_{k-1,\, n-k}$ denotes Fisher's *F*-distribution with $k-1$ and $n-k$ d.f. (degrees of freedom).

One-way ANOVA: the test

The one-way ANOVA test proceeds as follows:

- Given the samples $y_{11}, y_{12}, \dots, y_{1n_1}$, $y_{21}, y_{22}, \dots, y_{2n_2}$, etc., $y_{k1}, y_{k2}, \dots, y_{kn_k}$ of the random variables $Y_1 \sim \mathcal{N}(\mu_1, \sigma^2)$, $Y_2 \sim \mathcal{N}(\mu_2, \sigma^2)$, etc., $Y_k \sim \mathcal{N}(\mu_k, \sigma^2)$, respectively, formulate the null hypothesis:

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k$$

- The alternative hypothesis is $H_1: \neg H_0$, i.e. $\mu_{i'} \neq \mu_{i''}$ for some $i' \neq i''$

One-way ANOVA: the test

- Calculate the statistic

$$F = \frac{\mathrm{MS_B}}{\mathrm{MS_W}} = \frac{\mathrm{SS_B}}{\mathrm{SS_W}} \Big/ \frac{\mathrm{DF_B}}{\mathrm{DF_W}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2} \Big/ \frac{k-1}{n-k}$$

- If the null hypothesis is true, then we have by the Theorem

$$F \sim F_{k-1,\,n-k}$$

- Choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$, other popular values are $\alpha = 10\,\%$ or $\alpha = 1\,\%$ or $\alpha = 0.1\,\%$ etc.

One-way ANOVA: the test

- find the **critical value** $c > 0$ so that

$$\int_c^{+\infty} f(x)\,\mathrm{d}x = \alpha$$

  where f is the density of the *F*-distribution with $k-1$ and $n-1$ d.f.

- if $F \in [c, +\infty)$, **the critical region,** then **<u>reject</u>** the null hypothesis
- if $F \in [0, c)$, then **<u>do not reject</u>** (or <u>fail to reject</u>) the null hypothesis

<u>Remark:</u>
If $k = 2$, then the two-sample *t*-test for the difference of the means is equivalent.

One-way ANOVA: total variation

Total sample variance (mean squares):

sum of squares (total)

mean squares (total)

degrees of freedom (total)

$$\mathrm{MS_T} = \frac{\mathrm{SS_T}}{\mathrm{DF_T}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}\right)^2}{n-1}$$

where

- n_i is the size of the i-th group
- $\bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$ is the grand sample mean
- $n = \sum_{i=1}^{k} n_i$ is the size of the grand sample

One-way ANOVA: total variation

Total sample variance (mean squares):

The traditional ANOVA terminology is used:

Sum of Squares (total):

$$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

Degrees of Freedom (total):

$$DF_T = n - 1$$

One-way ANOVA: total variation

<u>Total sample variance (mean squares):</u>

The traditional ANOVA terminology is used:

Mean Squares (<u>total</u>):

$$\mathrm{MS_T} = \frac{SS_\mathrm{T}}{\mathrm{DF_T}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}\right)^2}{n-1}$$

One-way ANOVA: total variation

Notice that it holds:

$$SS_T = SS_W + SS_B$$

or

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2$$

One-way ANOVA: total variation

Notice first:

$$\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)=\sum_{j=1}^{n_i}y_{ij}-\sum_{j=1}^{n_i}\bar{y}_i=\sum_{j=1}^{n_i}y_{ij}-n_i\times\bar{y}_i=$$

$$=\sum_{j=1}^{n_i}y_{ij}-n_i\frac{\sum_{j=1}^{n_i}y_{ij}}{n_i}=$$

$$=\sum_{j=1}^{n_i}y_{ij}-\sum_{j=1}^{n_i}y_{ij}=0$$

One-way ANOVA: total variation

Notice now:

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\left(y_{ij}-\bar{y}_i\right)+\left(\bar{y}_i-\bar{y}\right)\right)^2 =$$

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

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

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{y}_i-\bar{y}\right)^2 = SS_W + SS_B$$

Bartlett’s test

Bartlett's test

Motivation: Recall the assumptions of the one-way ANOVA method:

Given random variables $Y_1 \sim \mathcal{N}(\mu_1, \sigma^2)$, $Y_2 \sim \mathcal{N}(\mu_2, \sigma^2)$, etc., $Y_k \sim \mathcal{N}(\mu_k, \sigma^2)$, we assume that

- the random variables $Y_1, Y_2, \dots, Y_k$ are normally distributed
- the variance σ^2 of all variables $Y_1, Y_2, \dots, Y_k$ is the same

Given the samples $y_{11}, y_{12}, \dots, y_{1n_1}$, $y_{21}, y_{22}, \dots, y_{2n_2}$, $\dots$, $y_{k1}, y_{k2}, \dots, y_{kn_k}$ of the random variables $Y_1, Y_2, \dots, Y_k$, respectively,

Bartlett's test

<u>Theorem:</u> If

$Y_{11}, Y_{12}, \dots, Y_{1n_1} \sim \mathcal{N}(\mu_1, \sigma^2),\ Y_{21}, Y_{22}, \dots, Y_{2n_2} \sim \mathcal{N}(\mu_2, \sigma^2),\ \dots,$

$Y_{k1}, Y_{k2}, \dots, Y_{kn_k} \sim \mathcal{N}(\mu_k, \sigma^2)$ are independent, then

$$\frac{(n-k)\ln\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i\right)^2}{n-k} - \sum_{i=1}^{k}(n_i-1)\ln\frac{\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i\right)^2}{n_i-1}}{1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1}-\frac{1}{n-k}\right)} \sim \chi^2_{k-1}$$

approximately

(if all $n_i \geq 7$)

Bartlett's test

Given unknown random variables

$$Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2), \quad \ldots, \quad Y_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$$

Sampling the variable Y_1 n_1-times yields the sample $y_{11}, y_{12}, y_{13}, \ldots, y_{1n_1}$

Sampling the variable Y_2 n_2-times yields the sample $y_{21}, y_{22}, \ldots, y_{2n_2}$

Sampling the variable Y_3 n_3-times yields the sample $y_{31}, y_{32}, y_{33}, y_{34}, y_{35}, \ldots, y_{3n_3}$

Etc.

Sampling the variable Y_k n_k-times yields the sample $y_{k1}, y_{k2}, \ldots, y_{kn_k}$

Null hypothesis:

$$H_0: \quad \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

Bartlett's test

- Calculate the statistic

$$X^2 = \frac{(n-k)\ln \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i\right)^2}{n-k} - \sum_{i=1}^{k}(n_i-1)\ln \frac{\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i\right)^2}{n_i-1}}{1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1}-\frac{1}{n-k}\right)}$$

- If the null hypothesis is true, then we have by the Theorem

$$X^2 \sim \chi^2_{k-1} \qquad \textit{approximately}$$

- Choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$, other popular values are $\alpha = 10\,\%$ or $\alpha = 1\,\%$ or $\alpha = 0.1\,\%$ etc.

Bartlett's test

- find the **critical value** $c > 0$ so that

$$\int_c^{+\infty} f(x)\,\mathrm{d}x = \alpha$$

  where f is the density of the χ^2-distribution with $k-1$ degrees of freedom

- if $X^2 \in [c, +\infty)$, **the critical region**, then **reject** the null hypothesis (the ANOVA should not be used)
- if $X^2 \in [0, c)$, then **do not reject** (or fail to reject) the null hypothesis