Multiple testing – Analysis of Variance ANOVA

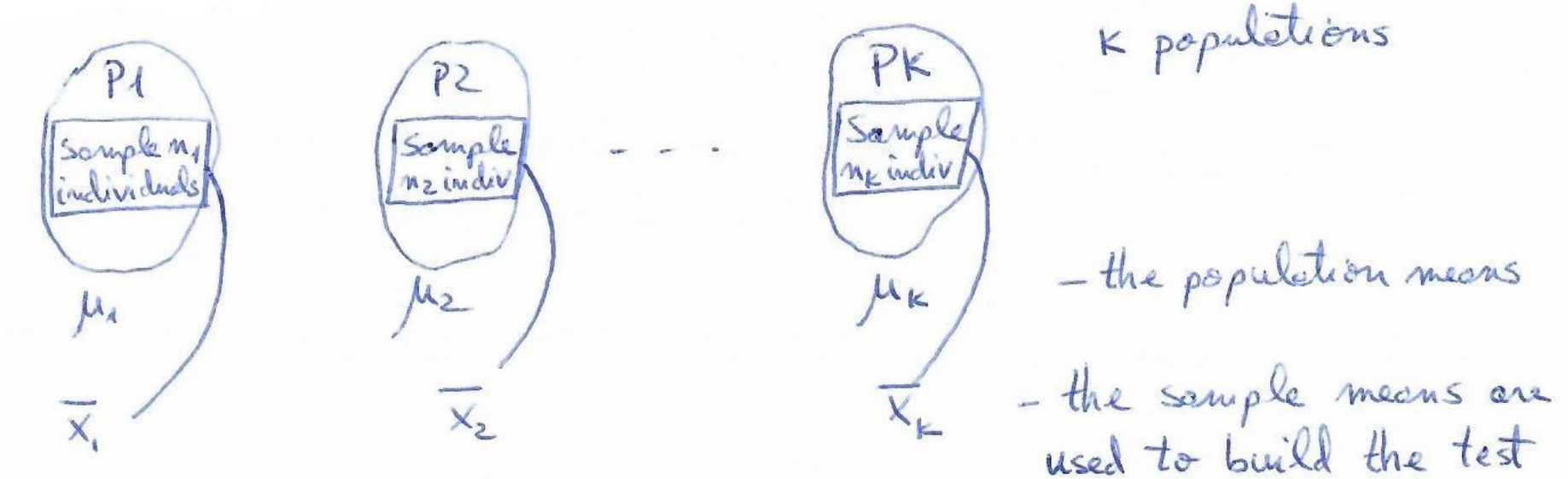
It compares 3 or more population means at once (by analyzing how these means vary from the others).

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

H_a: At least one mean differs from the others

Obs. If H_0 is rejected, the test does not tell which mean(s) is/are different – we have to do different testing to see which one(s).

Multiple testing – Analysis of Variance ANOVA



Obs. We could do instead hypothesis tests (e.g. the T-test) for every combination of 2 populations – this approach is not indicated because a lot of errors accumulate when we do a lot of tests sequential.

One-way ANOVA

It examines the influence of a factor (or independent variable) on a dependent variable.

For example, in an experiment, the factor is the soil quality (with 5 possible levels: poor, poor-medium, medium, medium-good, good) and the dependent variable is the wheat production/hectare.

One-way ANOVA

Why not compare groups with multiple t-tests?*

“Every time you conduct a t-test there is a chance that you will make a Type I error. This error is usually 5%. By running two t-tests on the same data you will have increased your chance of "making a mistake" to 10%. The formula for determining the new error rate for multiple t-tests is not as simple as multiplying 5% by the number of tests. However, if you are only making a few multiple comparisons, the results are very similar if you do. As such, three t-tests would be 15% (actually, 14.3%) and so on. These are unacceptable errors. An ANOVA controls for these errors so that the Type I error remains at 5% and you can be more confident that any statistically significant result you find is not just running lots of tests.”

*https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide-2.php

One-way ANOVA

```
g1<- round(rnorm(50,180,2))
g2<- round(rnorm(50,180,2))
g3<- round(rnorm(50,180,2))
```

One-way ANOVA(g1,g2,g3)

```
             Df Sum Sq Mean Sq F value Pr(>F)
factorgroup   2   20.3  10.167    2.52 0.0839 .
Residuals   147  593.1   4.034
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

t-test(g1,g2)

```
data:  g1 and g2
t = -1.0902, df = 98, p-value = 0.2783
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.1281362  0.3281362
sample estimates:
mean of x mean of y
   179.84    180.24
```

t-test(g1,g3)

```
data:  g1 and g3
t = 1.172, df = 98, p-value = 0.2441
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3466407  1.3466407
sample estimates:
mean of x mean of y
   179.84    179.34
```

t-test(g2,g3)

```
data:  g2 and g3
t = 2.1991, df = 98, p-value = 0.03022
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.08785633 1.71214367
sample estimates:
mean of x mean of y
   180.24    179.34
```

One-way ANOVA

A normal linear model to describe such an experiment is the following:

$$X_{ij} = \mu + \alpha_i + \xi_{ij}, \quad j = 1, \dots, n_i \quad i = 1, \dots, r \text{, where}$$

r = no. of levels of the factor

α_i = the effect of the level i^{th} of the factor over the mean μ

n_i = the sample size for the level i^{th}

$\{X_{ij}, j = 1, \dots, n_i\}$ iid observations at the level i^{th}, $i = 1, \dots, r$

$$E(X_{ij}) = \mu + \alpha_i \quad , \quad j = 1, \dots, n_i \; , \; i = 1, \dots, r$$
$$Var(X_{ij}) = \sigma^2$$

μ is the general mean for the model

$\xi_{ij} \sim N(0, \sigma^2)$ are random variables that define the random character of the studied phenomenon.

One-way ANOVA

Obs. ANOVA requires the assumption for homogeneity of variances of r samples

$$\{X_{i1}, \dots, X_{in_i}\} \text{ iid } N(\mu_i, \sigma_i^2), \quad i = 1, \dots, r$$

The Barlett test for homogeneity:

$$H_0: \{\sigma_1^2 = \dots = \sigma_r^2\}$$

$$H_a: \{\exists i, j: \sigma_i^2 \neq \sigma_j^2\}$$

One-way ANOVA

The Bartlett statistic is $$K^2 = \frac{(n-r)\ln s^2 - \sum_{i=1}^{r}(n_i-1)\ln s_i^2}{1+\frac{1}{3(r-1)}\sum_{i=1}^{r}\left(\frac{1}{n_i-1}-\frac{1}{n-r}\right)},$$

where $$s_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2, \quad i=1,\dots,r$$

$$n=\sum_{i=1}^{r} n_i, \qquad s^2 = \frac{1}{n-r}\sum_{i=1}^{r}(n_i-1)s_i^2$$

If H_0 is true, then $K^2 \sim \chi^2_{r-1}$. Thus, the acceptance region for H_0 at the significance level α is:

$$W_{n;1-\alpha} = \left\{(x_{11}\dots x_{1n_1},\dots,x_{r1},\dots,x_{rn_r}) \mid K^2 \le h_{r-1;1-\alpha}\right\},$$

where $h_{r-1;1-\alpha}$ is the $(1-\alpha)$ quantile of the χ^2_{r-1} distribution.

One-way ANOVA

Returning to the normal linear model, we denote $n = \sum_{i=1}^{r} n_i$ the total number of observations and we add the constraint

$$\sum_{i=1}^{r} n_i \alpha_i = 0$$

The model can be written in matrix form as:

$$X = Y\theta + \xi, \text{ where}$$

$$X = (x_{11}, \dots, x_{1n_1}, \dots, x_{r1}, \dots, x_{rn_r})'$$

$$\theta = (\mu, \alpha_1, \dots, \alpha_r)'$$

$$\xi = (\xi_{11}, \dots, \xi_{1,n_1}, \dots, \xi_{r1}, \dots, \xi_{rn_r})'$$

One-way ANOVA

$$Y = \begin{pmatrix} 1 & 1 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 & 1 \end{pmatrix} \begin{matrix} \} n_1 \\ \\ \\ \} n_2 \\ \\ \\ \} n_r \end{matrix}$$

$y^{(n,r)}$ of maximum rank

$\operatorname{rank} Y = r < n$

(no. of observation is greater than the no. of levels for the factor)

We want to test

$H_0: \{ \alpha_i = 0 \ , \ i = 1, \ldots, r \}$

$H_a: \{ \exists i, \ \alpha_i \neq 0 \}$

One-way ANOVA

Obs. H_0 is equivalent to the equality of the means at all levels i^{th}, $i = 1, \dots, r$

To build the test, first we have to estimate the parameters θ.

We denote $SS(X;\theta) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - \mu - \alpha_i)^2 = \|X - Y\theta\|^2$

(SS: sum of squares)

One-way ANOVA

We use the Least Squares method to estimate θ.

$\hat{\theta}_{LS}(X)$ is the estimator that, for any observed data $x = (x_{11}, \dots, x_{1n_1}, \dots, x_{r1}, \dots, x_{rn_r})'$, it has the value $\hat{\theta}_{LS}(x)$ as a solution of the minimization problem $\min_{\theta} SS(x;\theta)$.

$$\left.\begin{aligned} \frac{\partial SS(x;\theta)}{\partial \mu} = 0 \quad &\Leftrightarrow \quad -2\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \mu - \alpha_i) = 0 \\ &\text{recall that } n = \sum_{i=1}^{r} n_i, \quad \sum_{i=1}^{r} n_i \alpha_i = 0 \end{aligned}\right\} \Rightarrow$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij} \overset{\text{not}}{=} \bar{x}_{\cdot\cdot} \quad \text{(called the grand mean)}$$

One-way ANOVA

$$\frac{\partial SS(x;\theta)}{\partial \alpha_i} = 0 \iff -2\sum_{j=1}^{n_i}\left(x_{ij} - \hat{\mu} - \alpha_i\right) = 0 \,, \; i = 1,\dots,n \Rightarrow$$

$$\hat{\alpha}_i = \underbrace{\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}}_{\text{not } \overline{x}_{i\cdot}} - \overline{x}_{\cdot\cdot} = \overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot} \,, \; i = 1,\dots,n$$

$$E(X_{ij}) = \mu + \alpha_i \,, \; j = 1,\dots,n_i \,, \; i = 1,\dots,n$$

$$\widehat{E(X_{ij})} = \hat{\mu} + \hat{\alpha}_i = \overline{x}_{i\cdot}$$

↑ mean estimator

if H_0 is true, $\widehat{E_{H_0}(X_{ij})} = \hat{\mu} = \overline{x}_{\cdot\cdot}$

One-way ANOVA

We consider the following sums of squares:

$$SS_\alpha = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(\widehat{E(X_{ij})} - \widehat{E_{H_0}(X_{ij})} \right)^2 = \sum_{i=1}^{r} n_i \left(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot} \right)^2$$

"between group" variation

$$SS_r = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(X_{ij} - \widehat{E(X_{ij})} \right)^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i\cdot} \right)^2$$

"within group" variation

$$SS_t = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{\cdot\cdot} \right)^2$$ - total variability of the data

It can be shown that $SS_t = SS_\alpha + SS_r$.

One-way ANOVA

Proposition $\frac{SS_R}{\sigma^2}$ has a Chi-squared distribution with $(n-r)$ degrees of freedom χ^2_{n-r}.

Proposition If H_0 is true, then the following properties hold:

$$\frac{SS_t}{\sigma^2} \sim \chi^2_{n-1}$$

$$\frac{SS_\alpha}{\sigma^2} \sim \chi^2_{r-1}$$

SS_R and SS_α are independent.

Therefore, $$\frac{\overline{SS_\alpha}}{\overline{SS_R}} = \frac{\frac{1}{r-1}\cdot\frac{SS_\alpha}{\sigma^2}}{\frac{1}{n-r}\cdot\frac{SS_R}{\sigma^2}} \sim F_{r-1,\,n-r}$$

One-way ANOVA

The test statistic for

$H_0: \{\alpha_i = 0,\ i = 1, \dots, r\}$ against

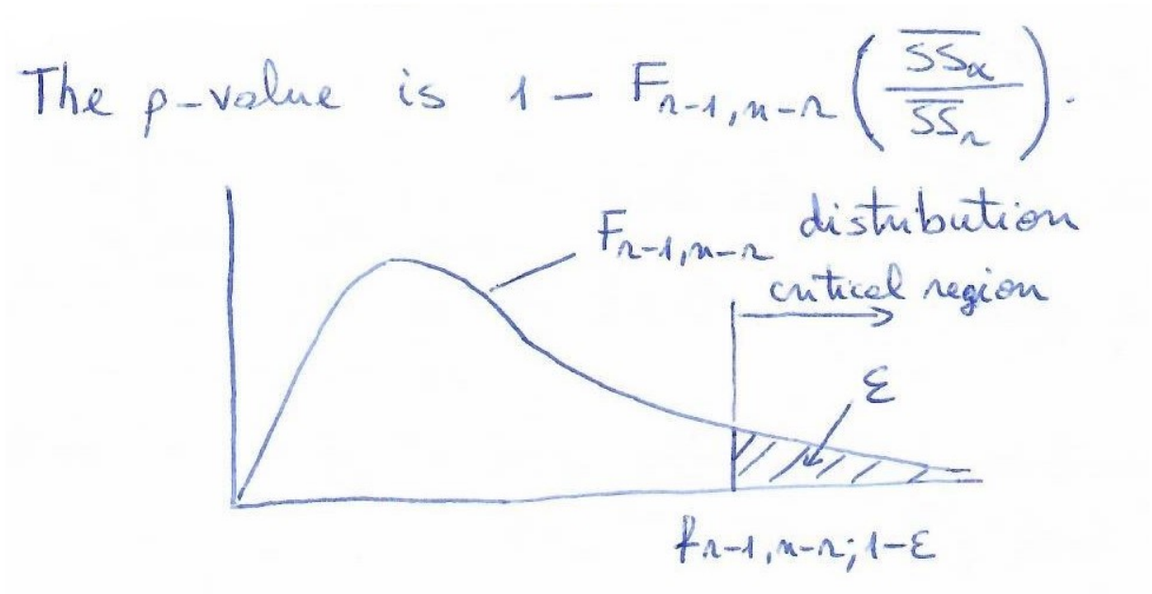
$H_a: \{\exists i,\ \alpha_i \neq 0\}$

is $\frac{\overline{SS_\alpha}}{\overline{SS_r}}$ and the acceptance region is

$$W_{n;1-\varepsilon} = \left\{ (x_{11}, \dots, x_{1n_1}, \dots, x_{r1}, \dots, x_{rn_r}) \;\middle|\; \frac{\overline{SS_\alpha}}{\overline{SS_r}} \leq f_{r-1,n-r;1-\varepsilon} \right\},$$

where ε is the significance level and $f_{r-1,n-r;1-\varepsilon}$ is the $(1-\varepsilon)$ quantile of the F distribution.

One-way ANOVA



Two-way ANOVA

It examines the influence of two factors (called independent variables) on a dependent variable.

The primary purpose of the two-way ANOVA is to find if there is an interaction between the factors on the dependent variable.

For example, we want to see whether there is an interaction between the soil quality and the temperature on wheat production/hectare. In this case, the interaction term tells us if the effect of soil quality (poor / poor-medium / medium / medium-good / good) on wheat production/hectare is influenced by temperature level (low / medium / high).

Two-way ANOVA

We consider the following normal linear model:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \xi_{ijk}, \quad \begin{array}{l} i = 1, \dots, r \\ j = 1, \dots s \\ k = 1, \dots, n \end{array}$$

r = no. of levels of the 1st factor

s = no. of levels of the 2nd factor

n = no. of i.i.d observations $\{X_{ijk}, k = 1, \dots, n\}$ in the group (i,j) – balanced experiments = the same no. of observations in each group

Two-way ANOVA

$$E(X_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \qquad i = 1, \dots, r$$
$$j = 1, \dots, s$$
$$Var(X_{ijk}) = \sigma^2 \qquad k = 1, \dots, n$$
$$\xi_{ijk} \sim N(0, \sigma^2)$$

α_i = the main effect of the level i^{th} of the 1^{st} factor over the mean μ

β_j = the main effect of the level j^{th} of the 2^{nd} factor over the mean μ

γ_{ij} = the interaction between the level i^{th} of the 1^{st} factor and the level j^{th} of the 2^{nd} factor.

Two-way ANOVA

We want to test the following hypotheses:

$H_\alpha: \{\alpha_i = 0,\ i = 1, \dots, r\}$ against $H'_\alpha: \{\exists\, i\ \alpha_i \neq 0\}$

$H_\beta: \{\beta_j = 0,\ j = 1, \dots, s\}$ against $H'_\beta: \{\exists\, j\ \beta_j \neq 0\}$

$H_\gamma: \{\gamma_{ij} = 0,\ i = 1, \dots, r,\ j = 1, \dots, s\}$ against $H'_\gamma: \{\exists\, i, j\ \gamma_{ij} \neq 0\}$

The tests are built in a similar way to one-way ANOVA: first we estimate the parameters μ, α_i, β_j and γ_{ij} using the Least Squares method; then we consider the following sums of squares:

Two-way ANOVA

SS_r = the variability within groups

SS_α = the variability generated by the 1st factor

SS_β = the variability generated by the 2nd factor

SS_γ = the variability generated by the interaction of factors

$SS_t = SS_r + SS_\alpha + SS_\beta + SS_\gamma$ - the total variability of the data

(the formulas of the estimators and of the SSs can be found in Dumitrescu & Bătătorescu, p. 209)

We denote $\overline{SS}_a = \frac{1}{\text{deg-of-freedom}} \cdot \frac{1}{T^2} \cdot SS_a$, $a = r, \alpha, \beta$ and γ

Two-way ANOVA

Proposition If the hypotheses H_α, H_β and H_γ are true, then:

$$\frac{\overline{SS_\alpha}}{\overline{SS_R}} \sim F_{r-1,\, rs(n-1)}$$

$$\frac{\overline{SS_\beta}}{\overline{SS_R}} \sim F_{s-1,\, rs(n-1)}$$

$$\frac{\overline{SS_\gamma}}{\overline{SS_R}} \sim F_{(r-1)(s-1),\, rs(n-1)}$$

$\frac{\overline{SS_\alpha}}{\overline{SS_R}}$, $\frac{\overline{SS_\beta}}{\overline{SS_R}}$, $\frac{\overline{SS_\gamma}}{\overline{SS_R}}$ are the test statistics for H_α, H_β, respectively H_γ.

Two-way ANOVA

The acceptance regions for these tests are the following:

$$W_{\alpha;1-\varepsilon} = \left\{ (x_{111}, \dots, x_{rsm}) \,\middle|\, \frac{\overline{SS_\alpha}}{\overline{SS_r}} \leq f_{r-1,\, rs(n-1);\, 1-\varepsilon} \right\}$$

$$W_{\beta;1-\varepsilon} = \left\{ (x_{111}, \dots, x_{rsm}) \,\middle|\, \frac{\overline{SS_\beta}}{\overline{SS_r}} \leq f_{s-1,\, rs(n-1);\, 1-\varepsilon} \right\}$$

$$W_{\gamma;1-\varepsilon} = \left\{ (x_{111}, \dots, x_{rsm}) \,\middle|\, \frac{\overline{SS_\gamma}}{\overline{SS_r}} \leq f_{(r-1)(s-1),\, rs(n-1);\, 1-\varepsilon} \right\}$$

where ε is the significance level and $f_{\dots;1-\varepsilon}$ are the $(1-\varepsilon)$ quantiles of the three F distributions.

For more on Analysis of Variance, please consult (Vāduva, 1970).

Two-way ANOVA

For example, a two-way ANOVA could be used to understand whether there is an interaction between gender and educational level on test anxiety amongst university students, where gender (males/females) and education level (undergraduate/postgraduate) are the independent variables, and test anxiety is the dependent variable.

Two *between-subjects* factors: each measurement is performed once for a single subject and given condition.

https://statistics.laerd.com/spss-tutorials/two-way-anova-using-spss-statistics.php

Two-way repeated measures ANOVA

A two-way repeated measures ANOVA (also known as a two-factor repeated measures ANOVA, two-factor or two-way ANOVA with repeated measures, or within-within-subjects ANOVA) compares the mean differences between groups that have been split on two within-subjects factors (also known as independent variables). A two-way repeated measures ANOVA is often used in studies where you have measured a dependent variable over two or more time points, or when subjects have undergone two or more conditions (i.e., the two factors are "time" and "conditions").

Two *within-subjects* factors: each measurement is repeatedly performed with each (the same) subject for ALL conditions.

https://statistics.laerd.com/spss-tutorials/two-way-repeated-measures-anova-using-spss-statistics.php

Mixed ANOVA

A mixed ANOVA compares the mean differences between groups that have been split on two "factors", where one factor is a "*within-subjects*" factor and the other factor is a "*between-subjects*" factor. For example, a mixed ANOVA is often used in studies where the dependent variable is measured over two or more time points or when all subjects have undergone two or more conditions (i.e., where "time" or "conditions" are the "within-subjects" factor), but also when the subjects have been assigned into two or more separate groups (e.g., based on some characteristic, such as subjects' "gender" or "educational level", or when they have undergone different interventions). These groups form the "between-subjects" factor.

A *mix* of one *between-subjects* and one *within-subjects* factor: a measurement is repeatedly performed for each level of the within-subjects factor with the same subject. For the between-subjects factor, a different group of subjects is used for each factor level.

https://statistics.laerd.com/spss-tutorials/mixed-anova-using-spss-statistics.php