Comparing 3 or more populations via ANOVA

Gabriela Czanner PhD CStat
Department of Biostatistics
Department of Eye and Vision Science

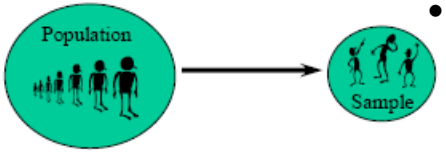
13 March 2013

MERSEY POSTGRADUATE TRAINING PROGRAMME

Workshop Series: Basic Statistics for Eye Researchers and Clinicians

Motivation: How to compare several groups?

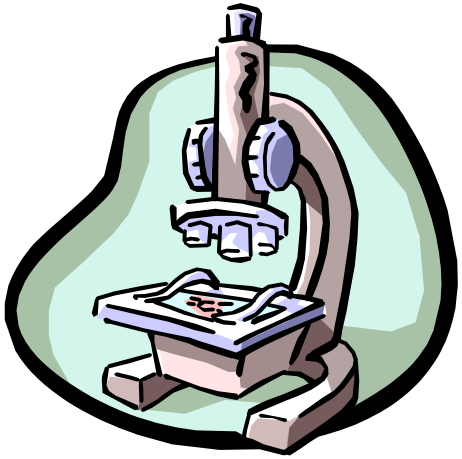
- Often the research question is:
  - Is mfERG intensity response same across five disease (e.g. diabetic retinopathy) patient groups?
  - Is visual acuity same across three treatment groups?
  - Is the degree of staining in conjuctivial tissues same across 5 different storage methods?
  - Is visual acuity different at two different time points after a treatment?



- An efficient strategy is to collect data on random independent samples in each group.
  - Then data are used to answer these questions on whole population of patients.
  - One way of comparing the groups is to compare their **means** via Analysis of Variance (ANOVA)

Outline

- One-way ANOVA
  - Introduction to the problem
  - Assumptions of ANOVA
  - How ANOVA works: The F-test
  - Post-hoc analyses i.e after ANOVA: Multiple comparisons
  - Non-parametric method of comparison
- Summary
- References
- Extra slides
  - Power and sample size calculations



One-way analysis of variance (ANOVA): Definitions & Goals

What is it used for? It allows us to compare the means of two or more independent groups analysing the different sources of variability.

Example 1:

Three groups of diabetic retinopathy (DR) were randomly selected. The groups are: #

- Healthy,
- DR with clinical signs of macular edema (DRnoCSMO),
- DR with CSMO (DRwithCSMO..

We are interested to see if the groups differ in terms of Microperimetry.

Healthy	DRnoCSMO	DRwithCSMO
83	86	71
85	82	73
99	72	61
100	89	64
98	69	75
91	94	85
95	79	86
77	65	74
100	80	82
90	79	60

Research question that can be addressed with ANOVA

Are there significant differences among the three groups of patients with regard to the MP1?

Definition of factors and levels in ANOVA

Factor (categorical variable): Any general aspect of interest we want to analyse (e.g., disease group, treatments).

Level (categories): Specific realizations of a factor (e.g., three different levels of the disease, levels of treatment).

To identify if the factors are random or fixed is relevant to determine the type of statistical analysis.

How to decide on fixed or random factors?

Random-factor: Factor whose levels may be regarded as a sample from a large population of levels. (e.g. subjects when a sample of subjects is analyzed to infer about a large population).

Fixed-factor: Factor whose levels are the only levels of interest (e.g., treatments)

E.g.:

Observers → random factor

Gender → Fixed factor (Male, Female)

Drug → Fixed factor

Most common ANOVA with <u>one</u> <u>fixed</u> factor, called one-way ANOVA

- **One-way ANOVA** involves the comparison of two or more population means $\mu_1, \mu_2, \ldots, \mu_k$. The null hypothesis of interest is:

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

And the alternative hypothesis is given by:

H_A: "<u>Not all</u> *k* population means are equal"

- If H_0 is rejected, we cannot conclude that <u>all</u> population means are different, (i.e., we cannot conclude that <u>all</u> treatments are different).

- Rejecting H_0 means that <u>at least</u> two population means have different values. (i.e., <u>at least</u> two treatments are different).

Data and means in one-way ANOVA: formally

Treatment	Sample size	Observations	Sample mean
1	n_1	$y_{11}, y_{12}, \ldots, y_{1n_1}$	$\bar{y}_1 = \frac{1}{n_1} \cdot \sum_{j=1}^{n_1} y_{1j}$
2	n_2	$y_{21}, y_{22}, \ldots, y_{2n_2}$	$\bar{y}_2 = \frac{1}{n_2} \cdot \sum_{j=1}^{n_2} y_{2j}$
⋮			
k	n_k	$y_{k1}, y_{k2}, \ldots, y_{kn_k}$	$\bar{y}_k = \frac{1}{n_k} \cdot \sum_{j=1}^{n_k} y_{kj}$

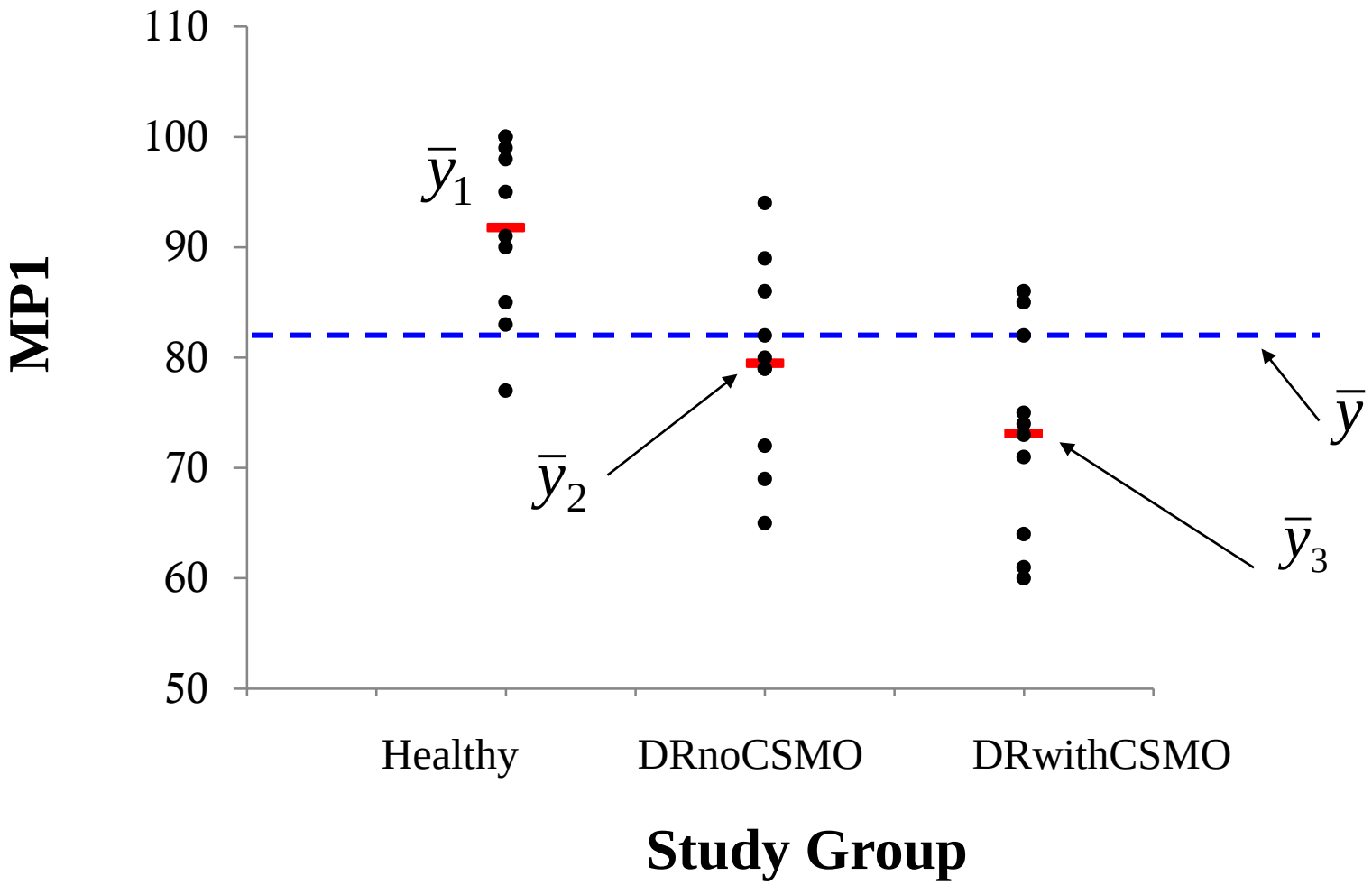
$$N = \sum_{i=1}^{k} n_i$$

$$\bar{y} = \frac{1}{k} \cdot \sum_{i=1}^{k} \bar{y}_i$$

Data and means in one-way ANOVA: visually

Overall sample mean $\bar{y}$

Sample means for each treatment $\bar{y}_i$; $i = 1,2,\ldots, k$



Assumptions in One-way ANOVA

1. **Random Samples** (i.e., individuals, animals, tissue samples, etc.) have been selected for each of the k treatments i.e. subjects selected independently from each other.

2. The variable of interest, y, is **normally distributed for <u>each</u> group** (treatment).

3. **Same amount of variability in each group**. The population variance of y (denoted as σ^2).
   - In such case, an estimator of the total variance is then obtained by pooling the individual sample variances.

NB. Always check if the ***assumptions*** of the statistical method are ***satisfied.*** If assumptions are not satisfied the results of ANOVA can be *biased*.

How ANOVA data analysis work: Informally

- The total variability in the observations can be decomposed into two components:

Total variation = variation ***between*** treatments + variation ***within*** treatments

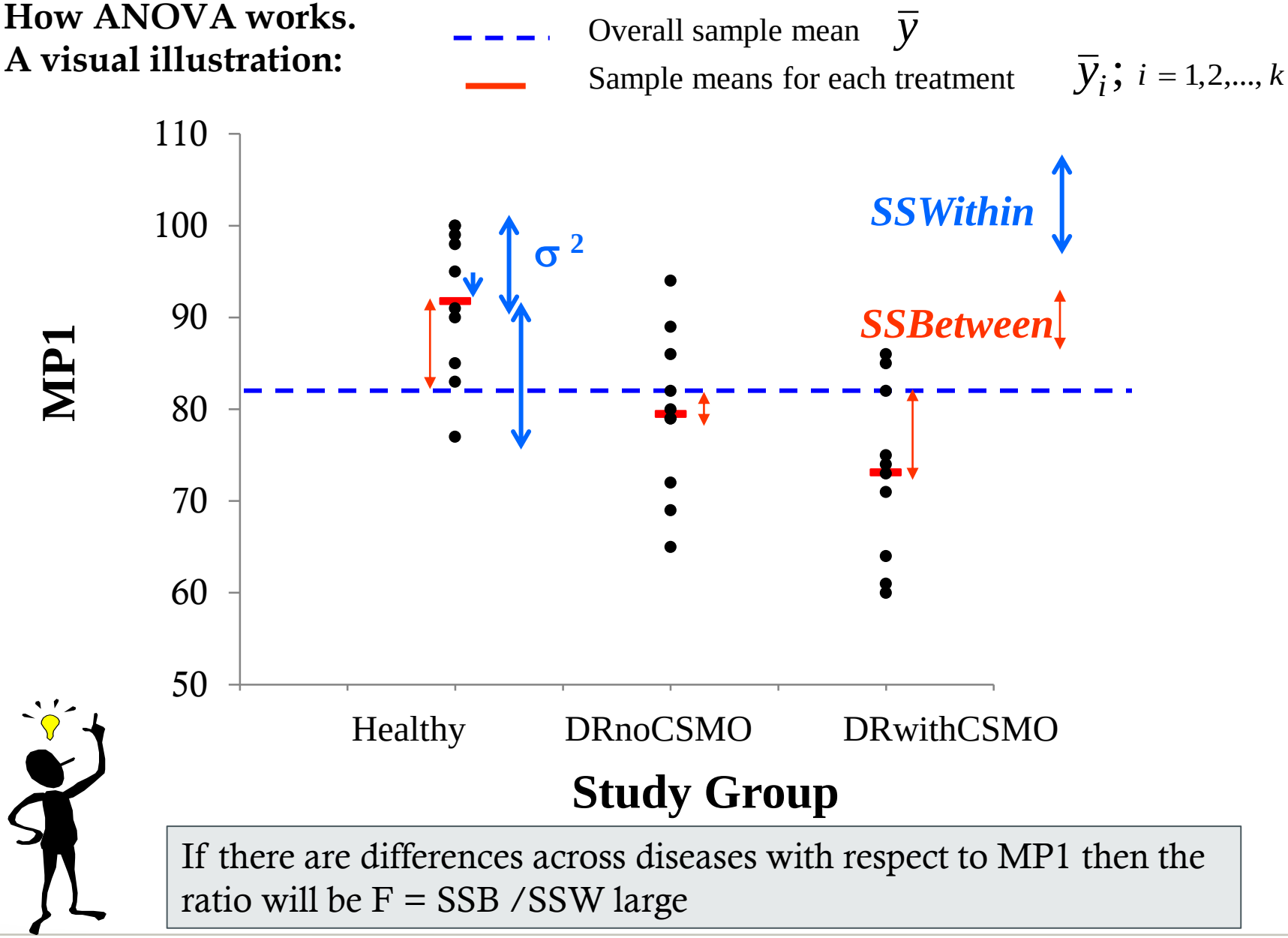
$$TSS = SSB + SSW$$

Sum squares **between** treatments

Sum squares **within** treatments

- If the patient groups are different then this should be reflected in
  - SSB large or small? **SSB large when compared with SSW**
  - SSW large or small? **SSW small when compared to SSB**
  - Ratio SSB /SSW large or small? **Ratio large. This ratio is called F and is the measure used in ANOVA.**

How ANOVA works.
A visual illustration:



If there are differences across diseases with respect to MP1 then the ratio will be F = SSB /SSW large

How ANOVA works. The formulas for the variabilities:

Analysis variance table corresponding to one-way ANOVA

Source of Variability	Degrees of freedom (df)	Sum squares	Mean squares	*F statistic*
Between	$k-1$	$SSB = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within	$N-k$	$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	$MSW = \frac{SSW}{N-k}$	
Total	$N-1$	$TSS = SSB + SSW$		

How ANOVA works. The F-test and interpretation

Hypotheses:

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

H_A: "Not all k population means are equal"

Test statistic:

$$F = \frac{MSBetween}{MSWithin}$$

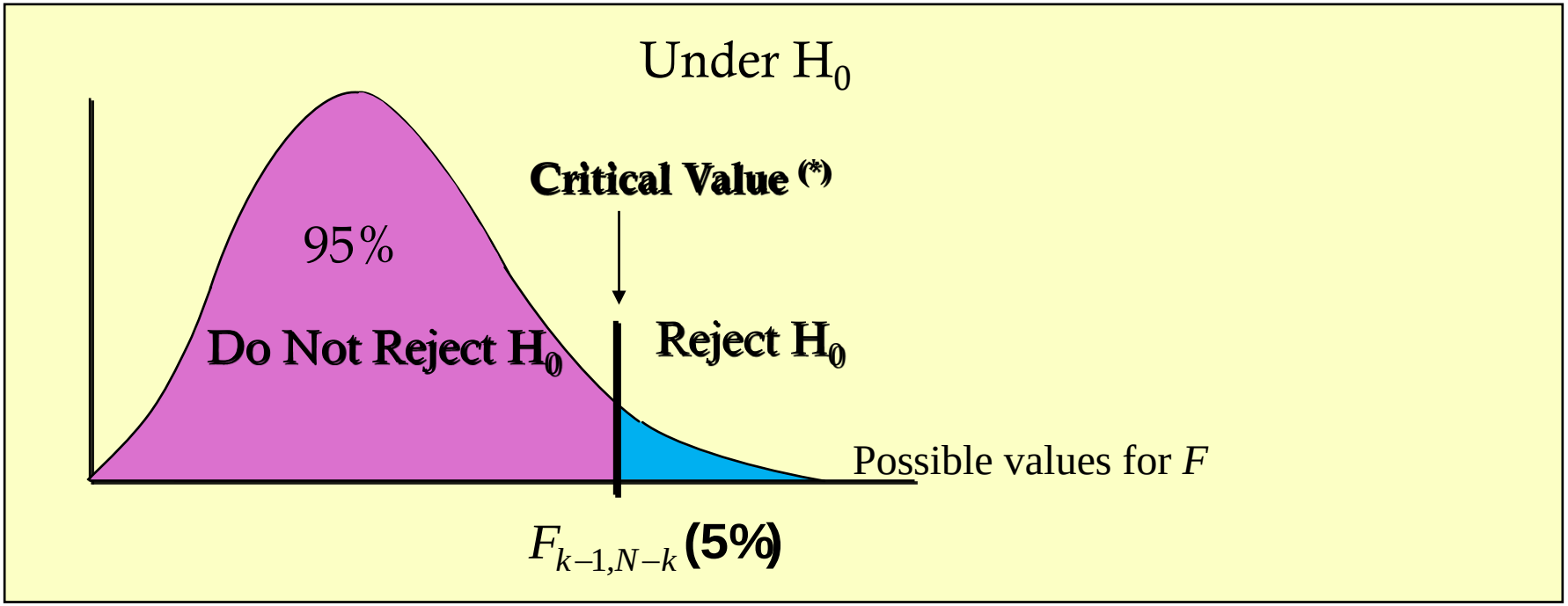
Under H0 → Estimate of σ^2

always → Estimate of σ^2

Large values of F ⟶ Total variability due mainly to differences between treatments rather than to differences within treatments.

Small values of F ⟶ Variability between treatments can not be significantly differentiated from the variability within treatments.

Under the null hypothesis H_0 (i.e., all population means are equal), the F statistic follows the F distribution with k – 1 and N – k degrees of freedom.



- If $F > F_{k-1,N-k}(5\%)$ we reject H_0 at the 5% significance level and conclude that <u>at least</u> two of the population means are different from one another.

- If $F \leq F_{k-1,N-k}(5\%)$ we don't reject H_0.

(*) Note that we could test H_0 at a different level of significance by replacing 5% with α

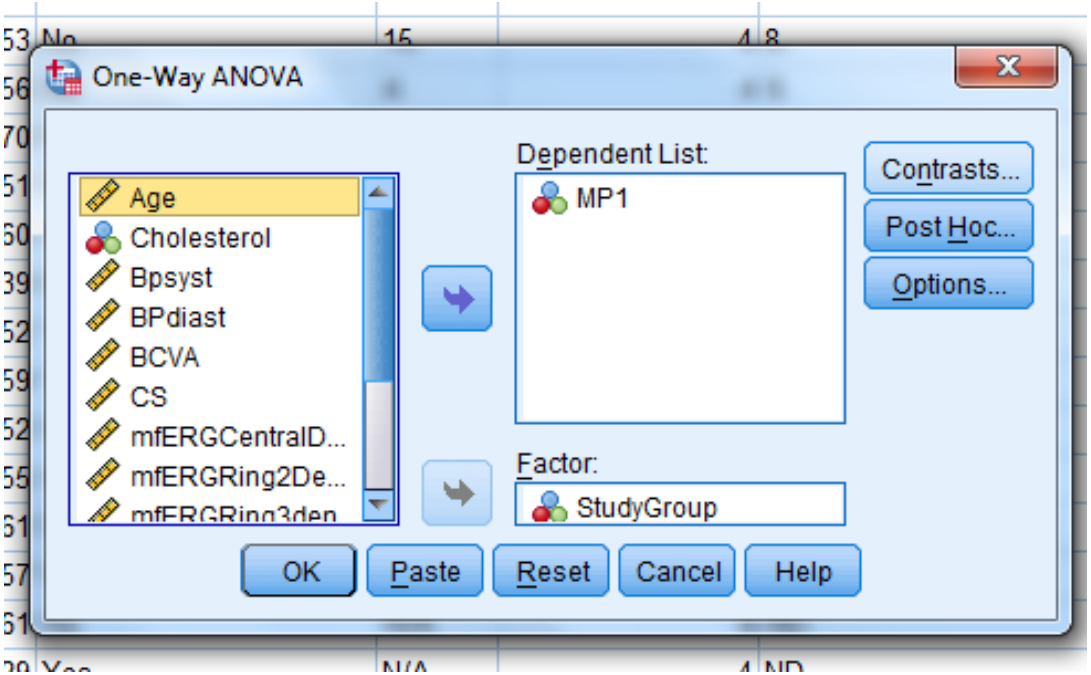
Example 1

We use ANOVA to compare 3 diabetic retinopathy groups with respect to MP1.

In SPSS Menu:

Analyze > Compare Means

> One-Way ANOVA



SPSS output:

ANOVA

MP1

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1806.467	2	903.233	11.585	.000
Within Groups	2105.000	27	77.963		
Total	3911.467	29			

Example 1

Source of Variability	Degrees of freedom (df)	Sum squares	Mean squares	*F statistic*
Between	$k - 1 = 2$	$SSB = 1086$	$MSB = 903$	$F = \frac{MSB}{MSW} = \mathbf{11.6}$
Within	$N - k = 27$	$SSW = 2105$	$MSW = 78$	
Total	$N - 1 = 29$	$TSS = 94.97$		

SPSS output:

ANOVA

MP1

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1806.467	2	903.233	11.585	.000
Within Groups	2105.000	27	77.963		
Total	3911.467	29			

Example 1

Source of Variability	Degrees of freedom (df)	Sum squares	Mean squares	*F statistic*
Between	$k-1=2$	$SSB=1086$	$MSB=903$	$F=\frac{MSB}{MSW}=\mathbf{11.6}$
Within	$N-k=27$	$SSW=2105$	$MSW=78$	
Total	$N-1=29$	$TSS=94.97$		

- Is $F > F_{k-1,N-k}(5\%)$? $F=11.6 > F_{2,27}(5\%)=2.96$
- Is $F \geq F_{k-1,N-k}(1\%)$? $F=11.6 < F_{3,27}(1\%)=4.60$

<u>Conclusion</u>: Can we reject at the 1% and 5% significance level the hypothesis that all treatments have the same effects?

Answer: Yes.

F test in one-way ANOVA as a generalization of the two-sample t test

For the simplest case involving a comparison of **two** population means, one-way ANOVA is equivalent to a t test for two independent groups with the assumption of equal population variances. Therefore, the F test yields the same P-value as the t-test. In fact, when k = 2 we have:

$$F_{1,N-2}(5\%) = t^2_{N-2}(5\%)$$

Non-parametric comparison of independent samples

Non-parametric comparison of samples should be done (i.e. ANOVA can NOT be used) if any of the following holds:

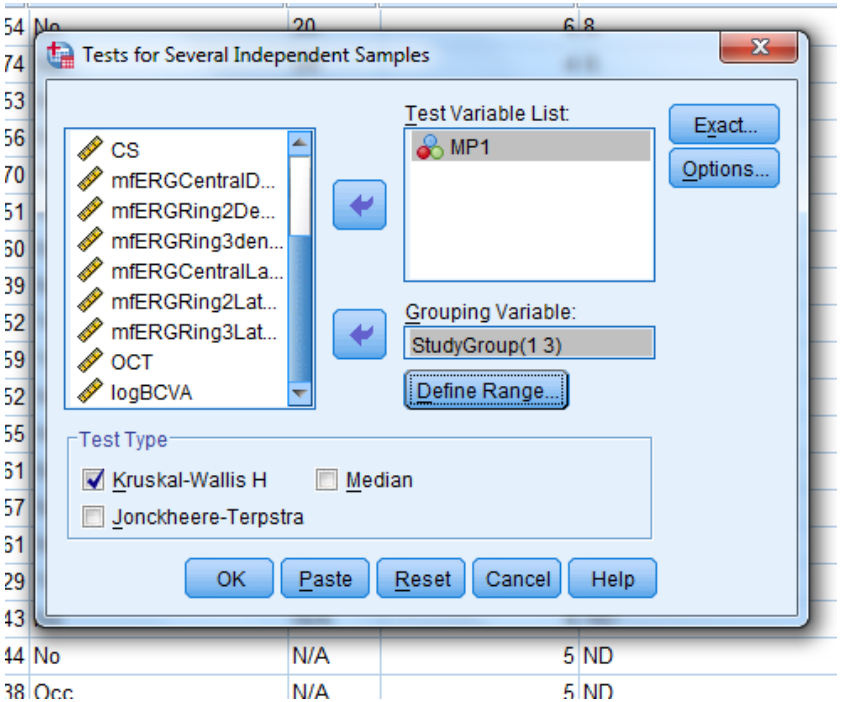
- Data are categorical
- Data are not normally distributed in one or more groups.
- Homogeneity of variance cannot be assumed.

The null hypothesis H_0: "all k population medians are equal" could be tested by applying the non-parametric **Kruskal-Wallis test** (which is a generalization of the non-parametric Mann-Whitney two-sample test).

Example 1

In SPSS menu: Analyze
> Nonparametric tests
> Legacy dialogs
> K independent samples

Result given by SPSS:



Ranks

	StudyGroup	N	Mean Rank
MP1	Healthy	10	23.35
	DRnoCSMO	10	13.80
	DRwithCSMO	10	9.35
	Total	30	

Test Statistics[a,b]

	MP1
Chi-Square	13.219
df	2
Asymp. Sig.	.001

a. Kruskal Wallis Test

b. Grouping Variable: StudyGroup

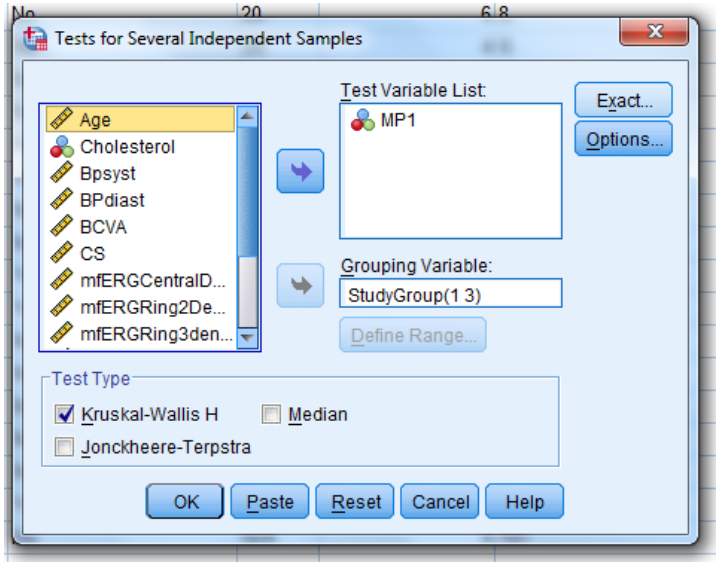
H_0 is rejected with 5% significance

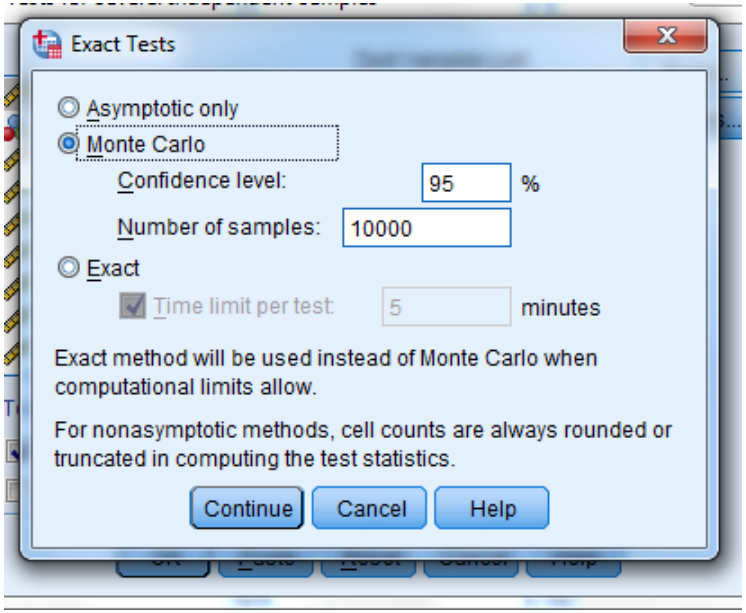
level by applying the *Kruskal-Wallis* test.st:

Note about Kruskal-Wallis test and SPSS:

By **default**, in SPSS the p-value of the Kruskal-Wallis test is calculated **approximatelly** – i.e. it is based on a simplified formula that is valid only in large samples. You can try one of other two options especially if your sample is short:

- Monte-Carlo calculated p-values,
- exact p-value formula





Example 1. Kruskal-Wallis test p-values comparison

Ranks

	StudyGroup	N	Mean Rank
MP1	Healthy	10	23.35
	DRnoCSMO	10	13.80
	DRwithCSMO	10	9.35
	Total	30	

Test Statistics[a,b]

	MP1
Chi-Square	13.219
df	2
Asymp. Sig.	.001

a. Kruskal Wallis Test

b. Grouping Variable: StudyGroup

Ranks

	StudyGroup	N	Mean Rank
MP1	Healthy	10	23.35
	DRnoCSMO	10	13.80
	DRwithCSMO	10	9.35
	Total	30	

Test Statistics[a,b]

			MP1
Chi-Square			13.219
df			2
Asymp. Sig.			.001
Monte Carlo Sig.	Sig.		.001[c]
	95% Confidence Interval	Lower Bound	.000
		Upper Bound	.002

a. Kruskal Wallis Test

b. Grouping Variable: StudyGroup

c. Based on 10000 sampled tables with starting seed 299883525.

In this example the p-values from 2 methods are same.

Post hoc analyses (i.e. after ANOVA) and problem of multiple comparisons

Question: ANOVA analysis concluded that the disease groups are different with respect to their MP1. How can we interpret this difference? Which groups are different? We need post-hoc analyses.

- One possibility is to compare each pair of means by using t-tests.
- When several t-tests are performed, each at a significant level α (Type I error), the probability of incorrectly obtaining at least one significant finding will be much larger than α and it will increase with the number of tests made.
- E.g., when $k = 5$ we can perform 10 different tests, if we choose $\alpha = 0.05$ in each of them, the probability of finding a significant difference when all population means are equal is:

$$\text{Prob} = 1 - 0.95^{10} = 40\,\%\,\textbf{!!}$$

Solutions to multiple comparisons

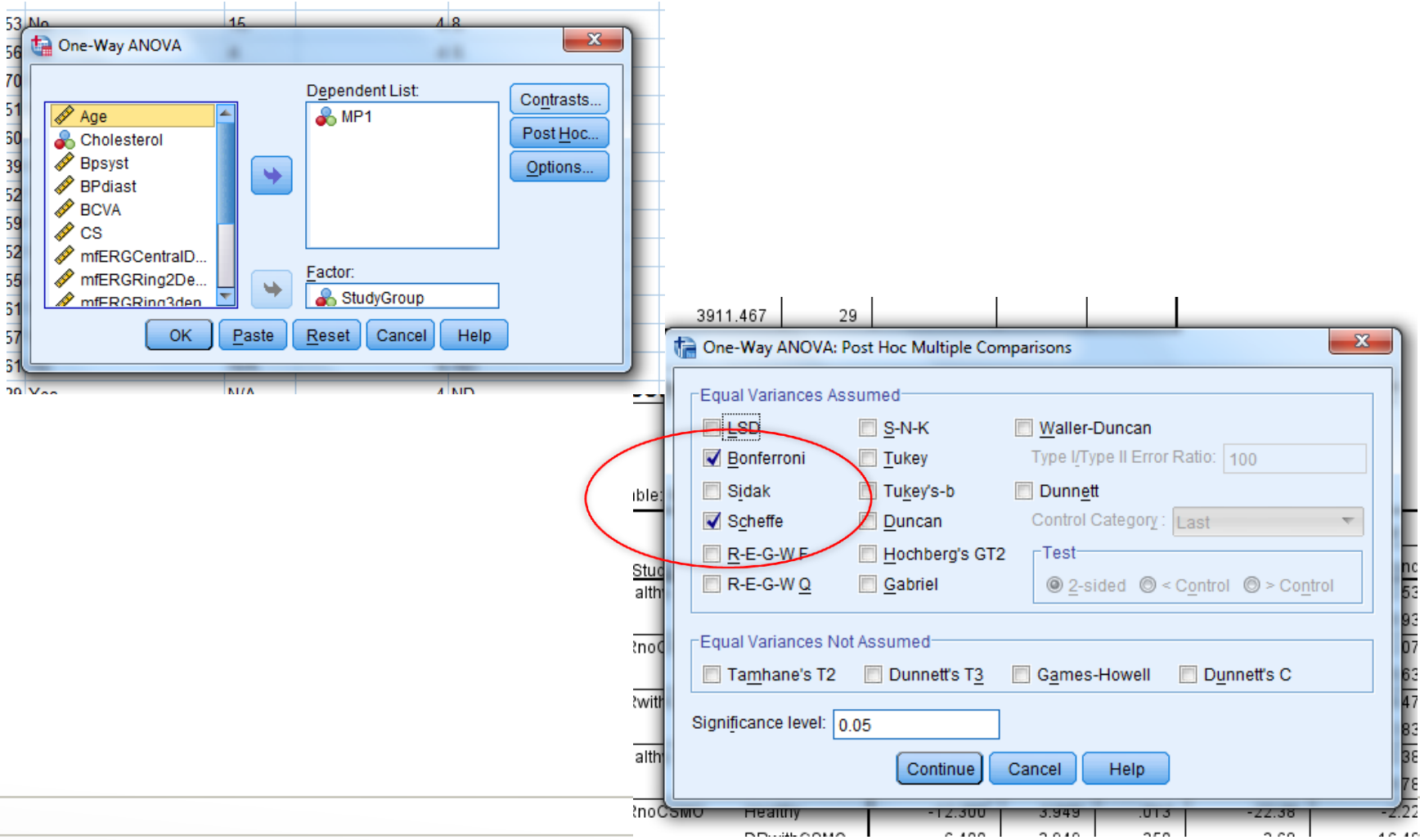
• One simple, but not optimal method, for addressing this problem, is to employ the so-called **Bonferroni correction**. If we perform *r* tests, then, to keep the *total* Type I error α small, we should use a significant level $\alpha' = \alpha / r$.

- For example, if we use Bonferroni method, if $k = 5$ (which implies $r = 10$) and we want a final Type I error equal to 0.05, the individual t-tests should be performed with a significance level $\alpha' = 0.05 / 10 = 0.005$.

• For small number of tests, Bonferroni correction is reasonable, but for large numbers the method is highly conservative (i.e., difficult to detect an existing difference) and alternative methods, such as *Duncan's multiple range* or *Scheffé's method*, are more appropriate.

Example 1

In SPSS menu: Analyze > Compare Means > One-Way ANOVA



Example 1. Results given by SPSS

Multiple Comparisons

Dependent Variable: MP1

	(I) StudyGroup	(J) StudyGroup	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Scheffe	Healthy	DRnoCSMO	12.300*	3.949	.016	2.07	22.53
		DRwithCSMO	18.700*	3.949	.000	8.47	28.93
	DRnoCSMO	Healthy	-12.300*	3.949	.016	-22.53	-2.07
		DRwithCSMO	6.400	3.949	.286	-3.83	16.63
	DRwithCSMO	Healthy	-18.700*	3.949	.000	-28.93	-8.47
		DRnoCSMO	-6.400	3.949	.286	-16.63	3.83
Bonferroni	Healthy	DRnoCSMO	12.300*	3.949	.013	2.22	22.38
		DRwithCSMO	18.700*	3.949	.000	8.62	28.78
	DRnoCSMO	Healthy	-12.300*	3.949	.013	-22.38	-2.22
		DRwithCSMO	6.400	3.949	.350	-3.68	16.48
	DRwithCSMO	Healthy	-18.700*	3.949	.000	-28.78	-8.62
		DRnoCSMO	-6.400	3.949	.350	-16.48	3.68

*. The mean difference is significant at the 0.05 level.

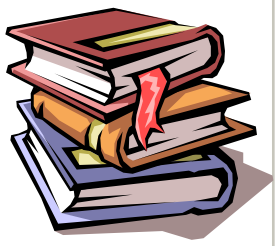
Conclusion?

With an <u>overall</u> 5% significance level, the treatments Healthy and DRnoCSMO are significantly different (p-value=0.016); also Healthy and Drwith CSMO are different (p-value<0.001)

Summary for ANOVA

- When to use ANOVA?
  - Comparison of groups with respect to their means
  - When assumptions are satisfied.
- Assumptions?
  - Random independent samples of patients
    - If samples related then you need a suitable method for related samples (Repeated measures ANOVA, Linear mixed models, Non-parameteric tests for related samples...)
  - The variable of interest (e.g. MP1) is normally distributed for each group
    - Remedy: Try log or sqrt transform if data are unimodal but skewed in each group. If this does not work use non-parametric test (e.g. Kruskal-Wallis)
    - If data are not normal because they are nominal you need to use a non-parametric method.
  - The variance (of MP1) is same for all the groups.
    - Remedy: Try log or sqrt transform if data are unimodal but skewed in each group. If this does not work use non-parametric test (e.g. Kruskal-Wallis)
- Post-hoc analyses needed after ANOVA concludes different groups.
  - Need to adjust for multiple comparisons.

Resources

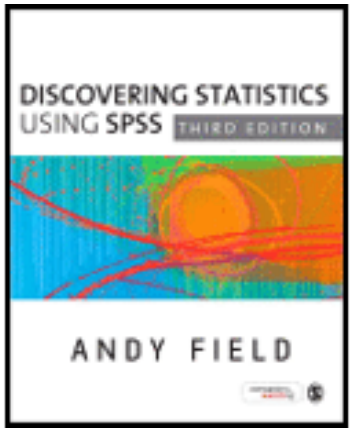


Books

- Practical statistics for medical research by Douglas G. Altman
- Medical Statistics from Scratch by David Bowers

Journals' with series on how to do statistics in clinical research

- American Journal of Ophthalmology has **Series on Statistics**
- British Medical Journal has series **Statistics Notes**



Manual for SPSS statistical software - with lots of worked-out examples

- Andy Field, Discovering statistics using SPSS

Workshops organized by Biostatistics Department, U of Liverpool

- http://www.liv.ac.uk/medstats/courses.htm,
- Design and analysis of laboratory-based studies**, 22 April 2013**
- **Statistical issues in the design and analysis of research projects 15-19 April 2013**

Thank you for your attention

These slides and worksheet can be found on: http://pcwww.liv.ac.uk/~czanner/

Planned future workshops:

- How to analyze data if they are not Normal? Nonparametric methods
- How to predict if a patient is having a disease? Classification methods. (Spt/Oct)
- How to make sense of many measured characteristics? Multivariate stats methods
- Log-odds ratios? When primary outcome is dichotomous (disease or no disease), how we measure odds of someone having a disease if he is a smoker?
- Ideas are welcome!

Statistical Clinics for ophthalmic clinicians and researchers !
Run by appointment.
Email: czanner@liv.ac.uk
Phone: +44-151-706-4019
Further information: http://pcwww.liv.ac.uk/~czanner/

Power of a test and sample size

Remember that α, known as the significance level or **Type I error**, is the **probability of rejecting the null hypothesis when it is true**.

Type II error, usually denoted as β, is the **probability of not rejecting the null hypothesis when it is false**.

Power of a test, which is denoted as $(1-\beta)$, is the **probability of rejecting the null hypothesis when it is false**.

Comparison of two independent samples

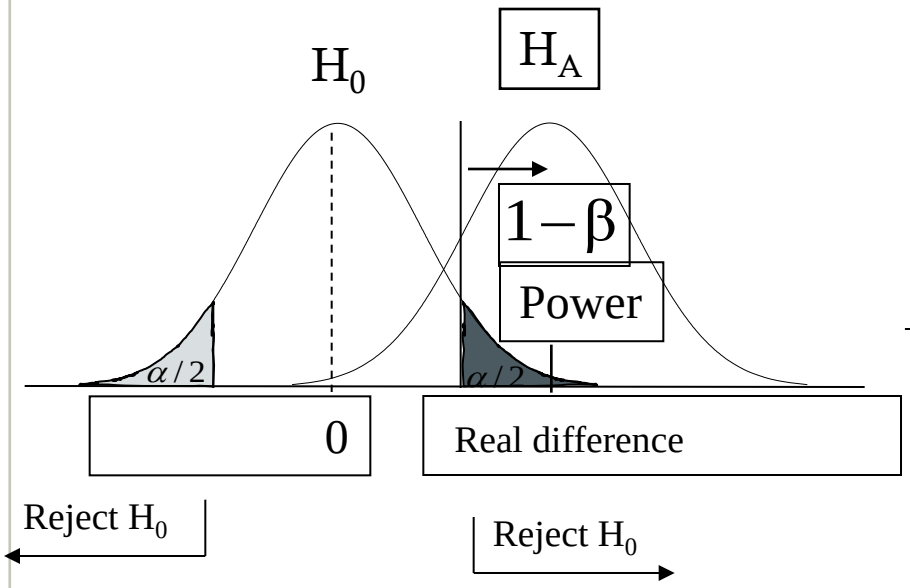


Figure.1. Illustration of the hypothesis testing

Probabilities of outcomes of hypothesis testing

Reality ↓	Conclusion: H_0 not rejected	Conclusion: H_0 rejected
H_0	$1 - \alpha$	α (Type I error))
H_A	β (Type II error)	$1 - \beta$ (Power)

How to calculate the required sample size?

Two independent samples ←

The minimum sample size required to detect an existing difference δ with significance level α and power β is given by the following formula:

$$n = (z_{1-\alpha/2} + z_{\beta})^2 \cdot \frac{2\sigma^2}{\delta^2}$$

where σ^2 is the variance within groups and $z_{1-\alpha/2}$, z_{β} are the $(1-\alpha/2)$ and β percentiles of the standard normal distribution, respectively.

Comparison of three or more independent samples

The minimum sample size required to reject H_0 depends on four different factors:

1. The real differences between population means

$\left|\mu_i - \mu_j\right|_{\,i,j=1,2,\ldots,k}$ $\uparrow$ $\longrightarrow$ Sample size decreases

2. The variability of the observations for each of the populations.

$\sigma_i^2 \;\; i=1,2,\ldots,k$ $\uparrow$ $\longrightarrow$ Sample size increases

3. Significance level (Type I error)

α $\downarrow$ $\longrightarrow$ Sample size increases

4. Power

$1-\beta$ $\uparrow$ $\longrightarrow$ Sample size increases

How to calculate the required sample size?

1. Decide the value of the significance level and power (e.g., $\alpha = 0.05$ and $1-\beta = 0.90$)

2. Calculate the parameter ϕ applying the following formula:

$$\phi = \sqrt{\frac{\sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2}{k \cdot MSE}}$$

3. Use the *Feldt–Mahmoud table* (see attached sheets) to calculate the minimum sample size per group required for a significant test with power $1-\beta$.