UNIVERSITÉ DU LUXEMBOURG

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE
Multiomics Data Science

Correction for Multiple Testing.

ANOVA

from BSc course Biostatistics (UL)

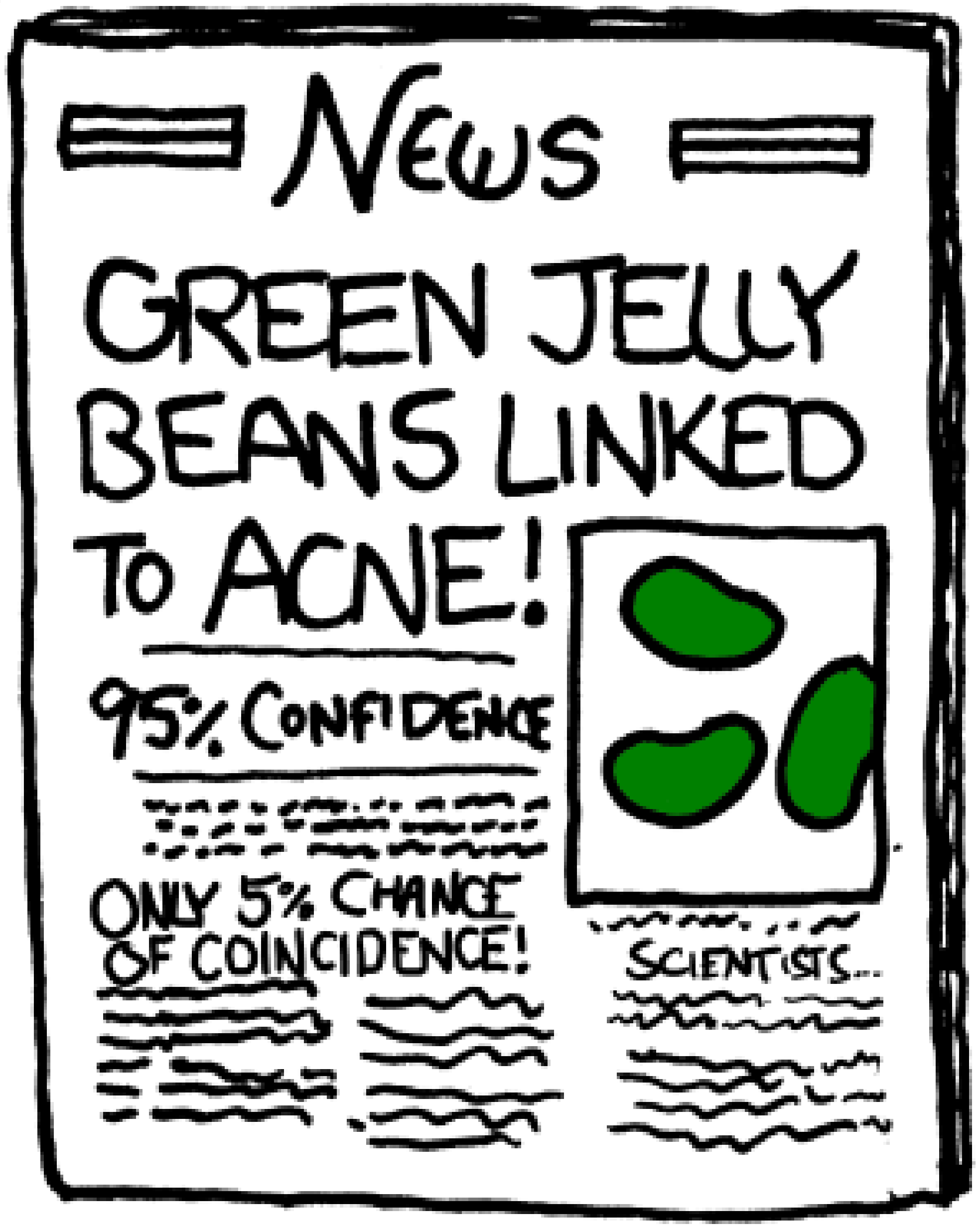
dr. Petr Nazarov

petr.nazarov@lih.lu

2020

Multiple Testing

Example



http://www.xkcd.com/882/

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

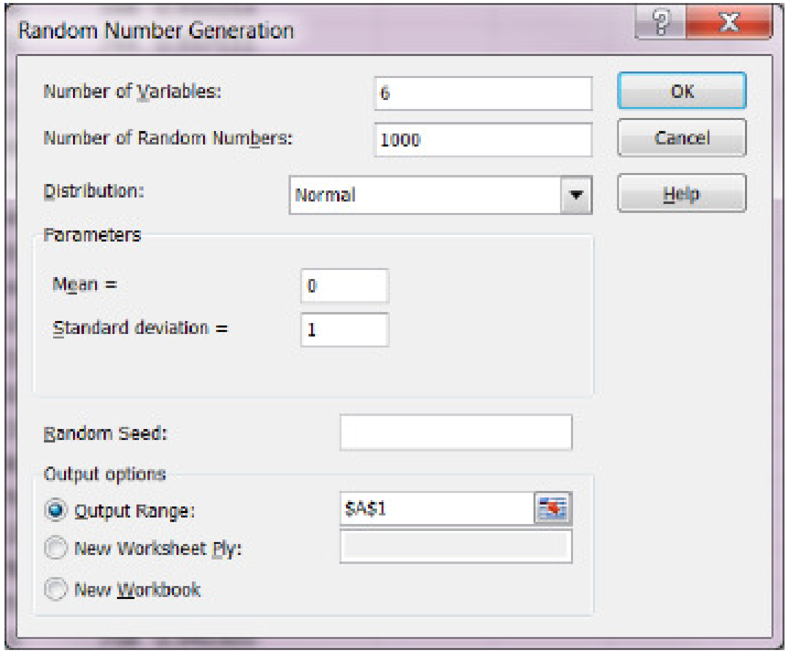
Multiple Testing

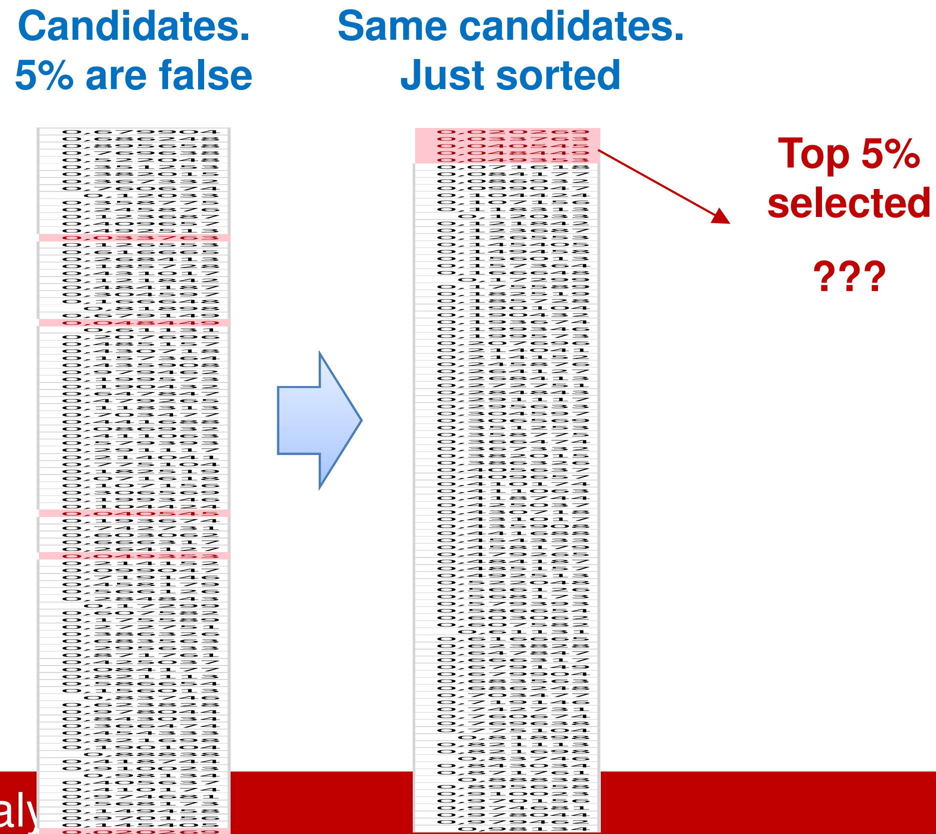
Why is it so important...

UNIVERSITÉ DU LUXEMBOURG

Let's generate a completely random experiment (Excel or R)

- Generate 6 columns of normal random variables (1000 candidate "genes" in each).
- Consider the first 3 columns as "treatment", and the next 3 columns as "control".
- Using t-test calculate p-values b/w "treatment" and "control" group. How many candidates have p-value<0.05 ?
- Calculate FDR. How many candidates you have now?





Multiple Testing

Hypotheses

		Population Condition	
		H_0 True	H_a True
Conclusion	Accept H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion

False Negative, β error (Type II Error)

False Positive, α error (Type I Error)

Probability of an error in a multiple test:

$$1-(0.95)^{\text{number of comparisons}}$$

Multiple Testing

False Discovery Rate

False discovery rate (FDR)

FDR control is a statistical method used in multiple hypothesis testing to correct for multiple comparisons. In a list of rejected hypotheses, FDR controls the expected proportion of incorrectly rejected null hypotheses (type I errors).

Population Condition

Conclusion		H_0 is TRUE	H_0 is FALSE	Total
	Accept H_0 (non-significant)	U	T	$m - R$
	Reject H_0 (significant)	V	S	R
	Total	m_0	$m - m_0$	$\boldsymbol{m}$

$$FDR = E\left(\frac{V}{V+S}\right)$$

Multiple Testing

FDR and FWER

Assume we need to perform *m* = 100 comparisons, and select maximum **FDR = α = 0.05**

$$FDR = E\left(\frac{V}{V+S}\right)$$

Expected value for FDR < α if $P_{(k)} < \frac{k}{m}\alpha$ → $\frac{mP_{(k)}}{k} < \alpha$

```
p.adjust(pv, method="fdr")
```

Theoretically, the sign should be “≤”.
But for practical reasons it is replaced by “<“

Familywise Error Rate (FWER)

Bonferroni – simple, but too stringent, not recommended

$$mP_{(k)} < \alpha$$

Holm-Bonferroni – a more powerful, less stringent but still universal FWER

$$(m+1-k)P_{(k)} < \alpha$$

```
p.adjust(pv)
```

ANOVA

Linear Models

Many conditions

We have measurements for 5 conditions. Are the means for these conditions equal?

Many factors

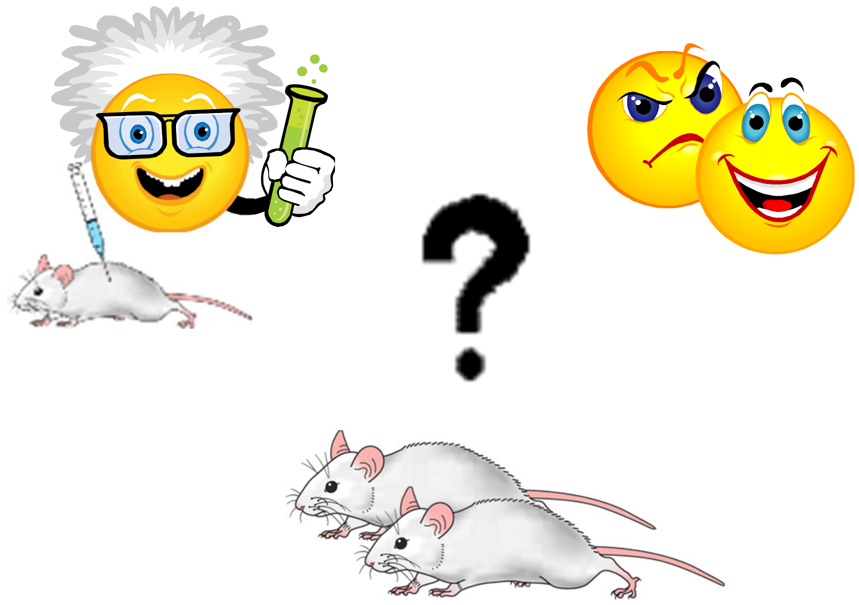
We assume that we have several factors affecting our data. Which factors are most significant? Which can be neglected?

ANOVA
example from Partek™

If we would use pairwise comparisons, what will be the probability of getting error?

Number of comparisons: $C_2^5 = \frac{5!}{2!3!} = 10$

Probability of an error: $1-(0.95)^{10} = 0.4$



http://easylink.playstream.com/affymetrix/ambsymposium/partek_08.wvx

ANOVA

Linear Models

As part of a long-term study of individuals 65 years of age or older, sociologists and physicians at the Wentworth Medical Center in upstate New York investigated the relationship between geographic location and depression. A sample of 60 individuals, all in reasonably good health, was selected; 20 individuals were residents of Florida, 20 were residents of New York, and 20 were residents of North Carolina. Each of the individuals sampled was given a standardized test to measure depression. The data collected follow; higher test scores indicate higher levels of depression.

Q: Is the depression level same in all 3 locations?

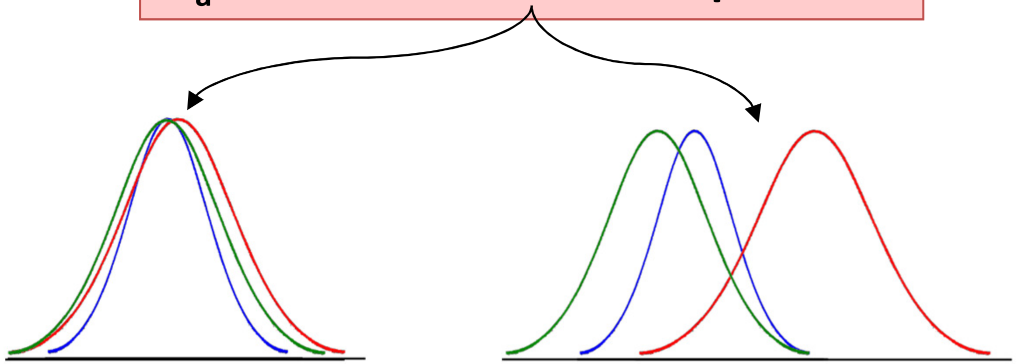
depression.txt

1. Good health respondents

Florida	New York	N. Carolina
3	8	10
7	11	7
7	9	3
3	7	5
8	8	11
8	7	8
...	...	...

H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: not all 3 means are equal

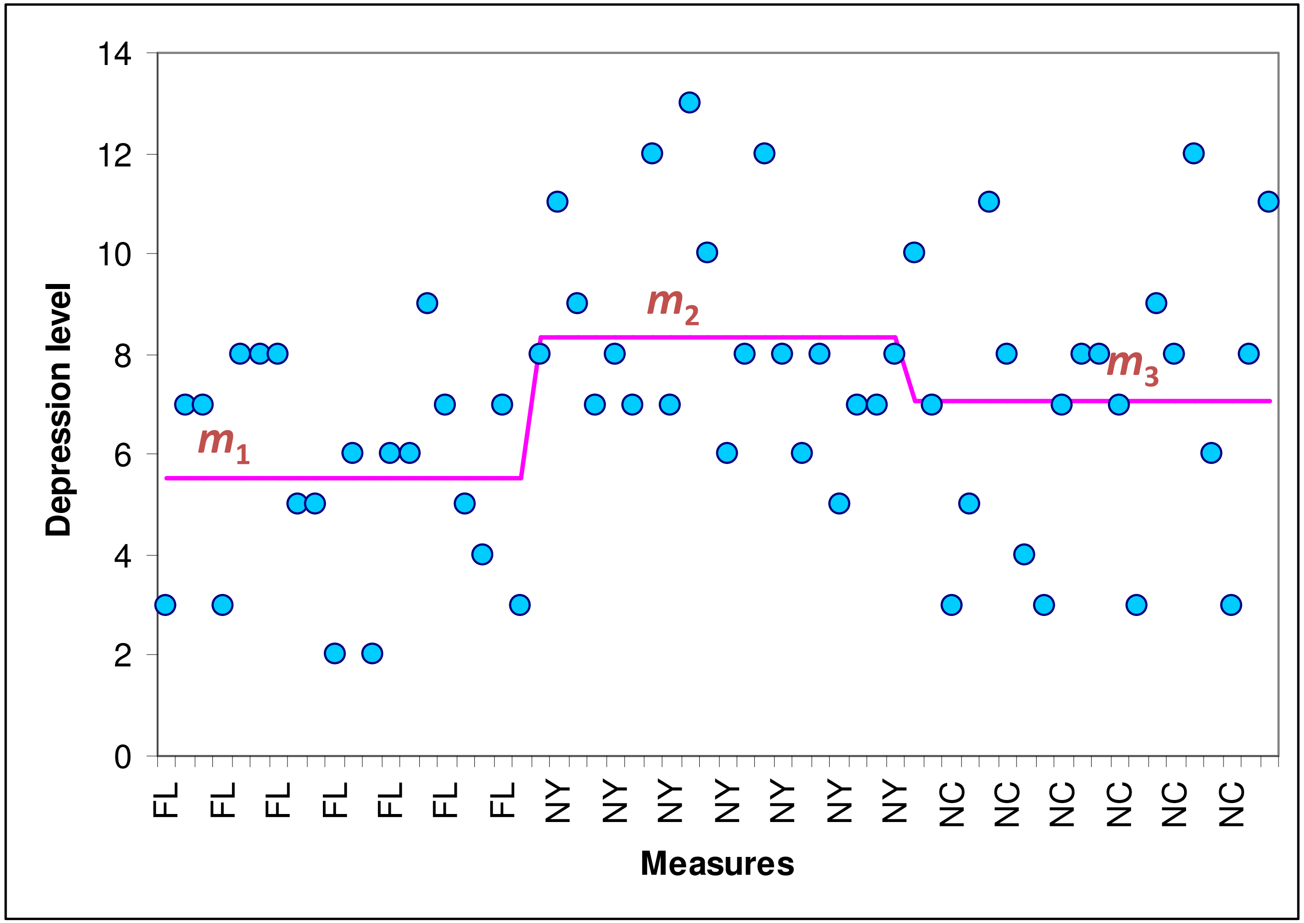


ANOVA

Linear Models

H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: not all 3 means are equal



ANOVA

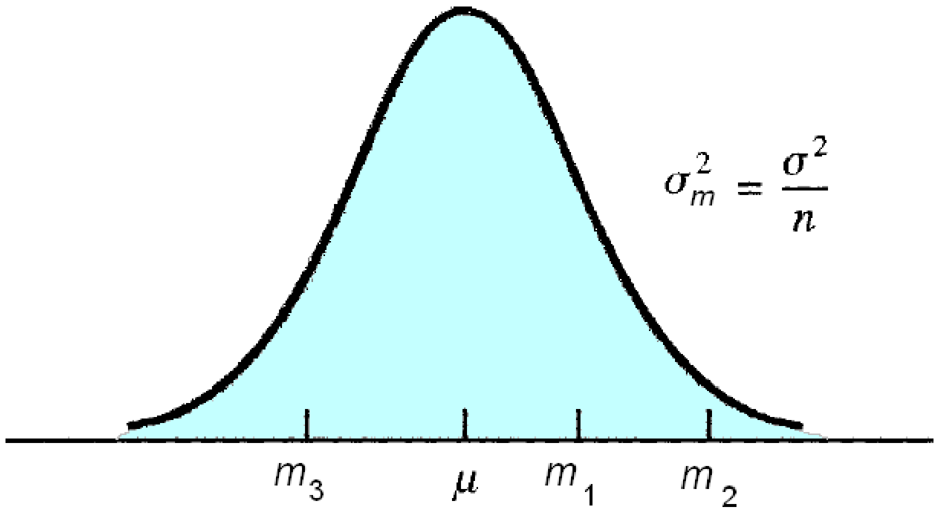
Assumptions for ANOVA

Assumptions for Analysis of Variance

1. For each population, the response variable is **normally distributed**

2. The variance of the respond variable, denoted as **σ^2 is the same** for all of the populations.

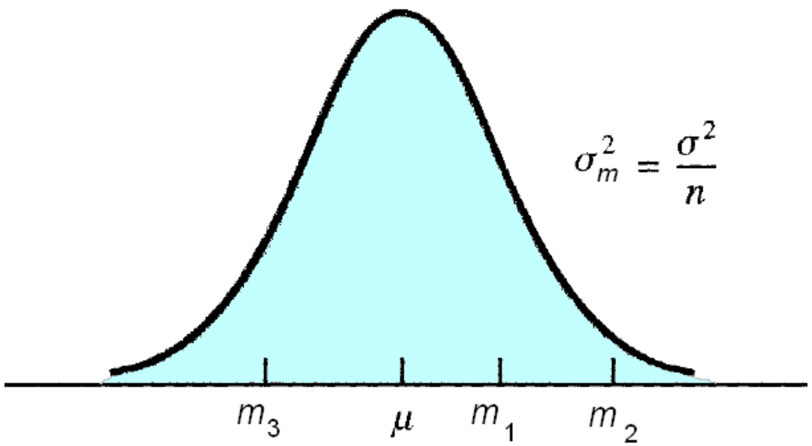
3. The observations must be **independent.**



ANOVA

Some Calculations

Parameter	Florida	New York	N. Carolina
m=	5.55	8.35	7.05
overall mean=	6.98333		
var=	4.5763	4.7658	8.0500



Let's estimate the variance of sampling distribution. If H_0 is true, then all m_i belong to the same distribution

$$\sigma_m^2 = \frac{\sum_{i=1}^{k}(m_i - \overline{m})^2}{k-1} = \frac{(5.55-6.98)^2 + (8.35-6.98)^2 + (7.05-6.98)^2}{3-1} = 1.96$$

$$\sigma^2 = n\sigma_m^2 = 20 \times 1.96 = 39.27$$

– this is called between-treatment estimate, works only at H_0

At the same time, we can estimate the variance just by averaging out variances for each populations:

– this is called within-treatment estimate

$$\sigma^2 = \frac{\sum_{i=1}^{k}\sigma_i^2}{k} = \frac{4.58 + 4.77 + 8.05}{3} = 5.8$$

Do between-treatment estimate and within-treatment estimate give variances of the same "population"?

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

ANOVA

Some definitions

UNIVERSITÉ DU LUXEMBOURG

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

H_a: not all k means are equal

Means for treatments

$$m_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

Variances treatments

$$s_j^2 = \frac{\sum_{i=1}^{n_j} \left(x_{ij} - m_j\right)^2}{n_j - 1}$$

Total mean

$$\overline{m} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n_T}$$

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

due to treatment

Sum squares

$$SSTR = \sum_{j=1}^{k} n_j \left(m_j - \overline{m}\right)^2$$

Mean squares, $\sigma_{beetween}^2$

$$MSTR = \frac{SSTR}{k-1}$$

due to error

Sum squares

$$SSE = \sum_{j=1}^{k} \left(n_j - 1\right) s_j^2$$

Mean squares, σ_{within}^2

$$MSE = \frac{SSE}{n_T - k}$$

Test of variance equality

$$F = \frac{MSTR}{MSE}$$

p-value for the treatment effect

$$p - value$$

ANOVA

The Main Equation

Total sum squares

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \overline{m}\right)^2$$

SS due to treatment

$$SSTR = \sum_{j=1}^{k} n_j\left(m_j - \overline{m}\right)^2$$

$$SST = SSTR + SSE$$

SS due to error

$$SSE = \sum_{j=1}^{k}\left(n_j - 1\right)s_j^2$$

Total variability of the data include variability due to treatment and variability due to error

$$d.f.(SST) = d.f.(SSTR) + d.f.(SSE)$$
$$n_T - 1 = (k-1) + (n_T - k)$$

Partitioning
The process of allocating the total sum of squares and degrees of freedom to the various components.

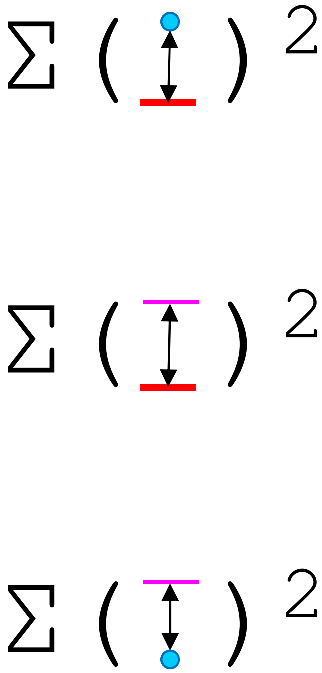
ANOVA

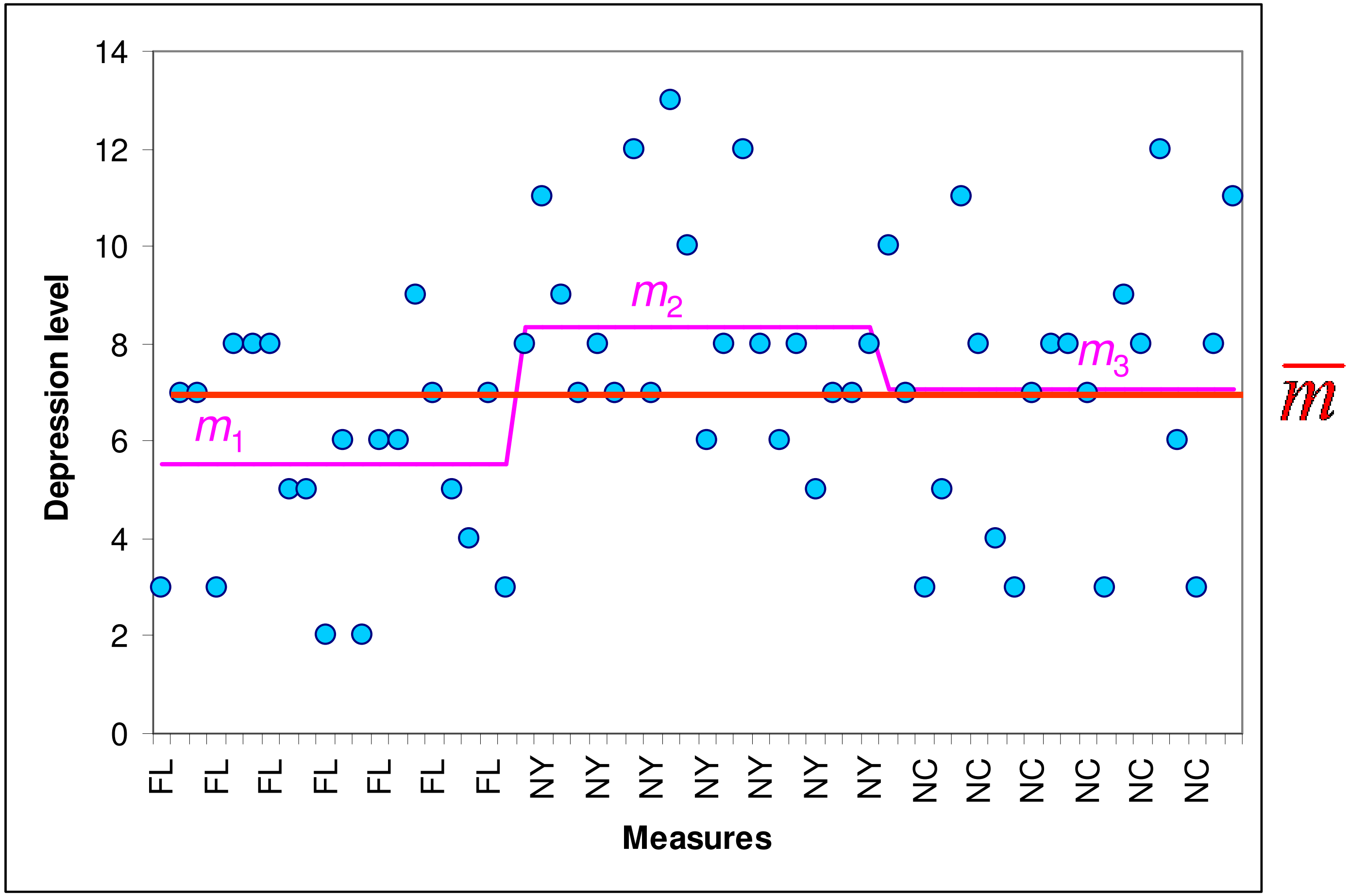
The Main Equation

SST:

SSTR:

SSE:



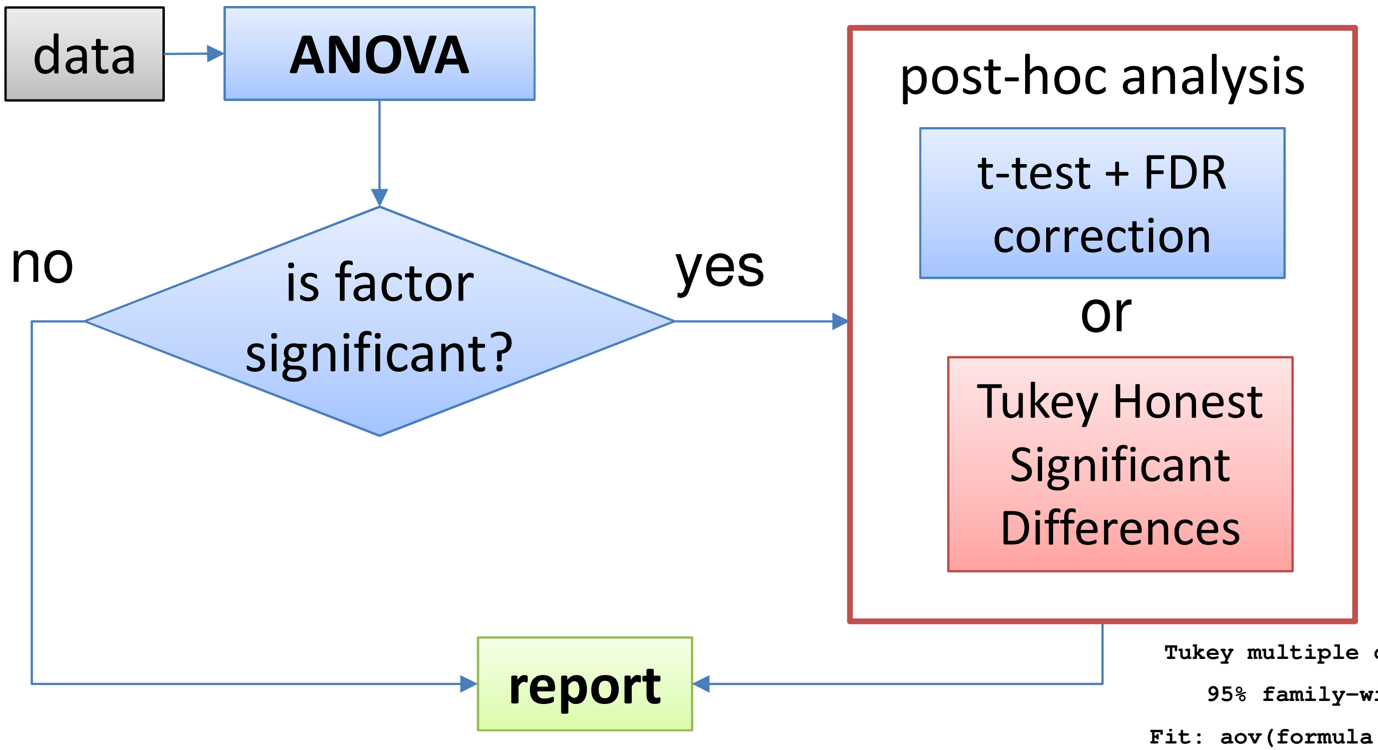


$$SST = SSTR + SSE$$

ANOVA

Post-hoc Analysis

Post-hoc analysis
allows for additional exploration of significant differences in the data, when significant effect of the factor was already confirmed (for example, by ANOVA).



$$FDR = \frac{pval * nComparisons}{k}$$

`TukeyHSD()`

Group1	Group2	p-value	k	FDR
Florida	New York	0.00021	1	0.00063
Florida	North Carolina	0.0667	2	0.10005
New York	North Carolina	0.11264	3	0.11264

```
 Tukey multiple comparisons of means
   95% family-wise confidence level

Fit: aov(formula = Depression ~ Location, data = DepGH)

$Location
                         diff        lwr       upr     p adj
New York-Florida          2.8  0.9677423 4.6322577 0.0014973
North Carolina-Florida    1.5 -0.3322577 3.3322577 0.1289579
North Carolina-New York -1.3 -3.1322577 0.5322577 0.2112612
```

ANOVA

ANOVA Table

1-way ANOVA

factor | degree of freedom | sum squares | variances | F=vf/vr | p-value | error

```
           Df Sum Sq Mean Sq F value Pr(>F)
Location    2   78.5   39.27   6.773 0.0023 **
Residuals  57  330.4    5.80
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2-way ANOVA

factors | degree of freedom | sum squares | variances | F=vf/vr | p-value

```
                 Df Sum Sq Mean Sq F value  Pr(>F)
Location          2   73.8    36.9   4.290   0.016 *
Health            1 1748.0  1748.0 203.094 <2e-16 ***
Location:Health   2   26.1    13.1   1.517   0.224
Residuals       114  981.2     8.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA

Linear Models for Transcriptomics

$$Y_{ij} = \mu_i + A_j + B_j + A_j{*}B_j + \epsilon_{ij}$$

i – gene index

j – sample index

Y_{ij} – expression of *i-th* gene in *j*-th sample

μ_i – mean expression of *i-th* gene

A_j , B_j – factors

$A_j{*}B_j$ – interaction: effect which cannot be explained by superposition A and B

limma – R package for DEA in microarrays based on linear models.

It is similar to t-test / ANOVA but using all available data for variance estimation, thus it has higher power when number of replicates is limited

edgeR – R package for DEA in RNA-Seq, based on linear models and negative binomial distribution of counts.

Better noise model results in higher power detecting differentially expressed genes

negative binomial process – number of tries before success: rolling a die until you get 6