Silesian University of Technology

STATISTICAL ANALYSES OF EXPERIMENTAL RESULTS FROM LOW AND HIGH THROUGHPUT APPROACHES IN RADIATION RESEARCH

Joanna Polańska

CELET: Cellular effects of high and low LET ionising radiation – introduction to radiation biology

RadoNorm
Managing risks from radon and NORM

High throughput screening

- High Throughput Screening (HTS) is a method that uses automation and large data set processing to quickly assay the biological or biochemical activity of a large number of compounds, proteins, and genes.

- HTS is an approach that has gained widespread popularity over the last two decades.

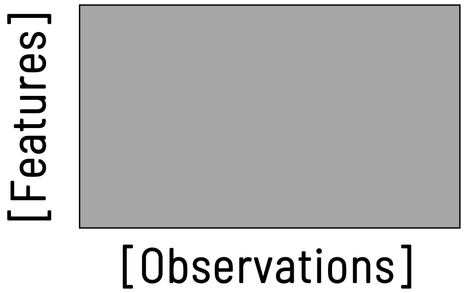
HTS – dimensionality issue

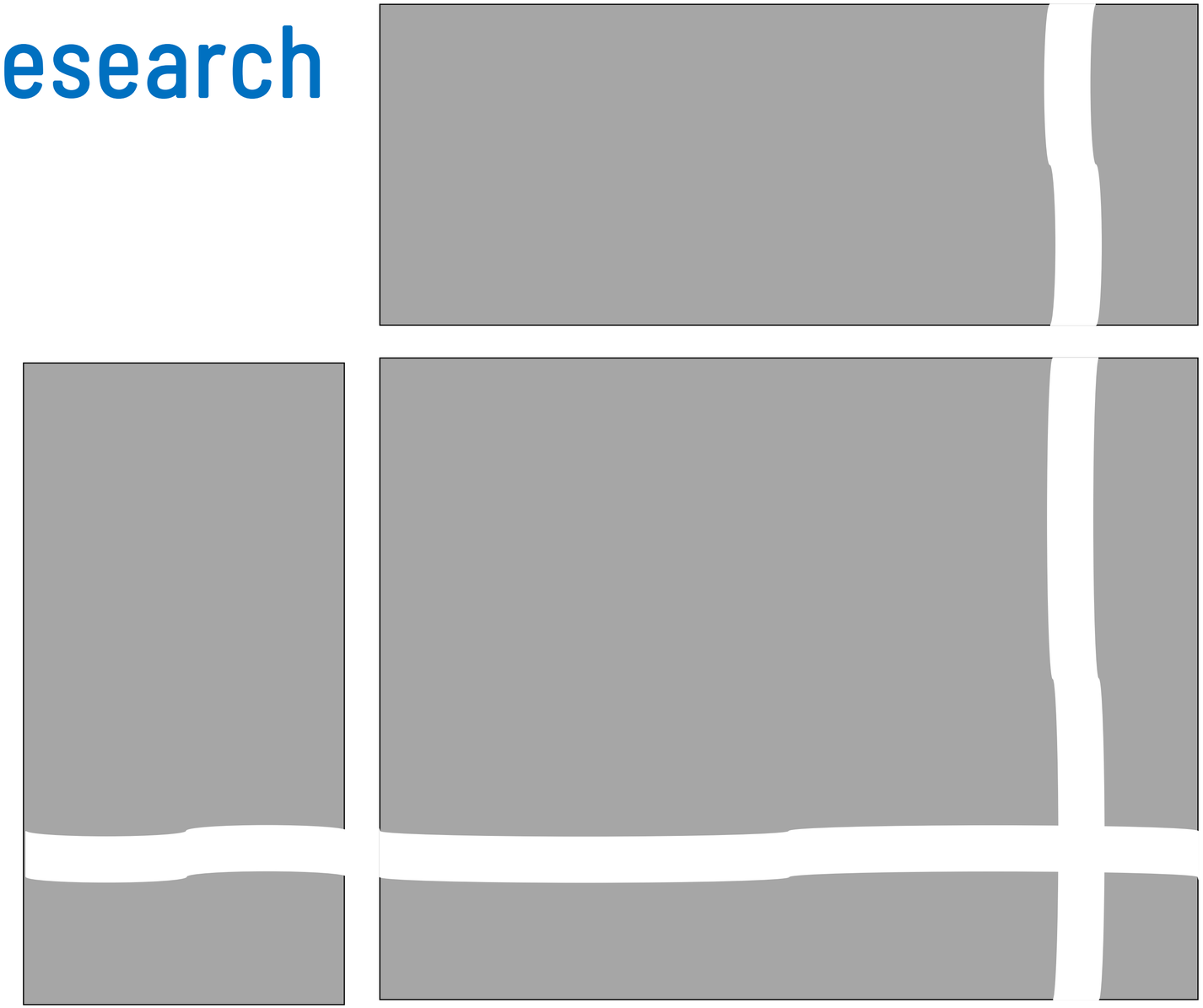
Platform	Number of observations (up to)	Number of features (up to)
PCR	$10^2 - 10^3$	$10^1 - 10^2$
RNA microarrays	$10^2 - 10^3$	10^4
Tiling arrays	$10^2 - 10^3$	$10^6 - 10^7$
RNA sequencing	$10^2 - 10^3$	$10^6 - 10^7$
SNP microarrays	$10^2 - 10^3$	$10^5 - 10^6$
CNV arrays	$10^2 - 10^3$	10^6
Methylation arrays	$10^2 - 10^3$	$10^8 - 10^9$
Whole genome sequencing	$10^2 - 10^3$	10^9

High throughput screening

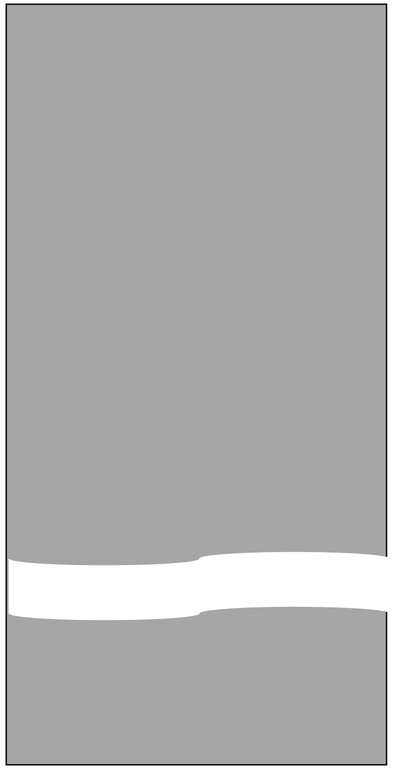
Advantages	Disadvantages
Miniaturization (low sample volume, less waste)	High cost
Automation (high speed, unattended operation, better data reproducibility)	Relatively low data quality (accuracy, precision, especially for weak/low signals)
Relatively low background signal	Data analysis (false positive discoveries)

Data in radiation research





CASE A: A LOT OF FEATURES



Statistical analysis

Significance tests

- Formal procedure for comparing observed data with a hypothesis whose truth we want to assess.
    - Hypothesis: statement about the parameters in a population or model
- Results of the test are expressed in terms of a probability that measures how well the data and the hypothesis agree.

Test results

KIAED
KATEDRA INŻYNIERII I ANALIZY EKSPLORACYJNEJ DANYCH

Decision / Grand true	# not enough evidence to reject null hypothesis	# rejection of the null hypothesis	∑
# null hypothesis is true	TN True negative	FP False positive Type I error	n_0
# null hypothesis is false	FN False negative	TP True positive	n_1
∑	n-R ujemne	R dodatnie	n

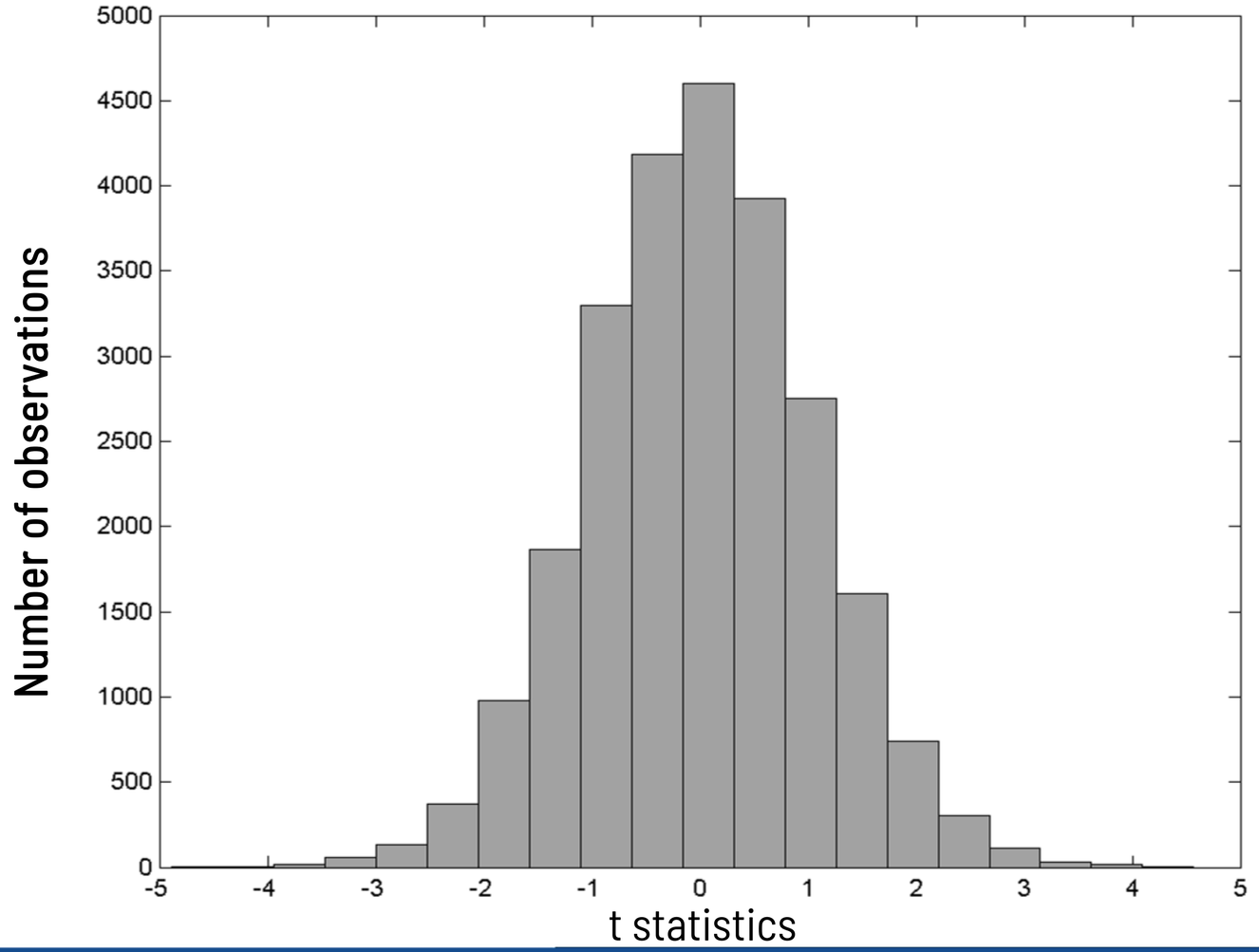
Multiple testing

- Statistical data analysis usually requires the verification of multiple statistical hypotheses.
- In the case of a single test, it is possible to determine *a priori* α level to control the first type of error (i.e. rejection of the null hypothesis when it is correct; false positive FP).
- For example, $\alpha = 0.05$ means that on average 1 in 20 tests will give a false positive result. If we run 100 tests with $\alpha = 0.05$ for each test, we can expect an average of 5 false positive results.

Example

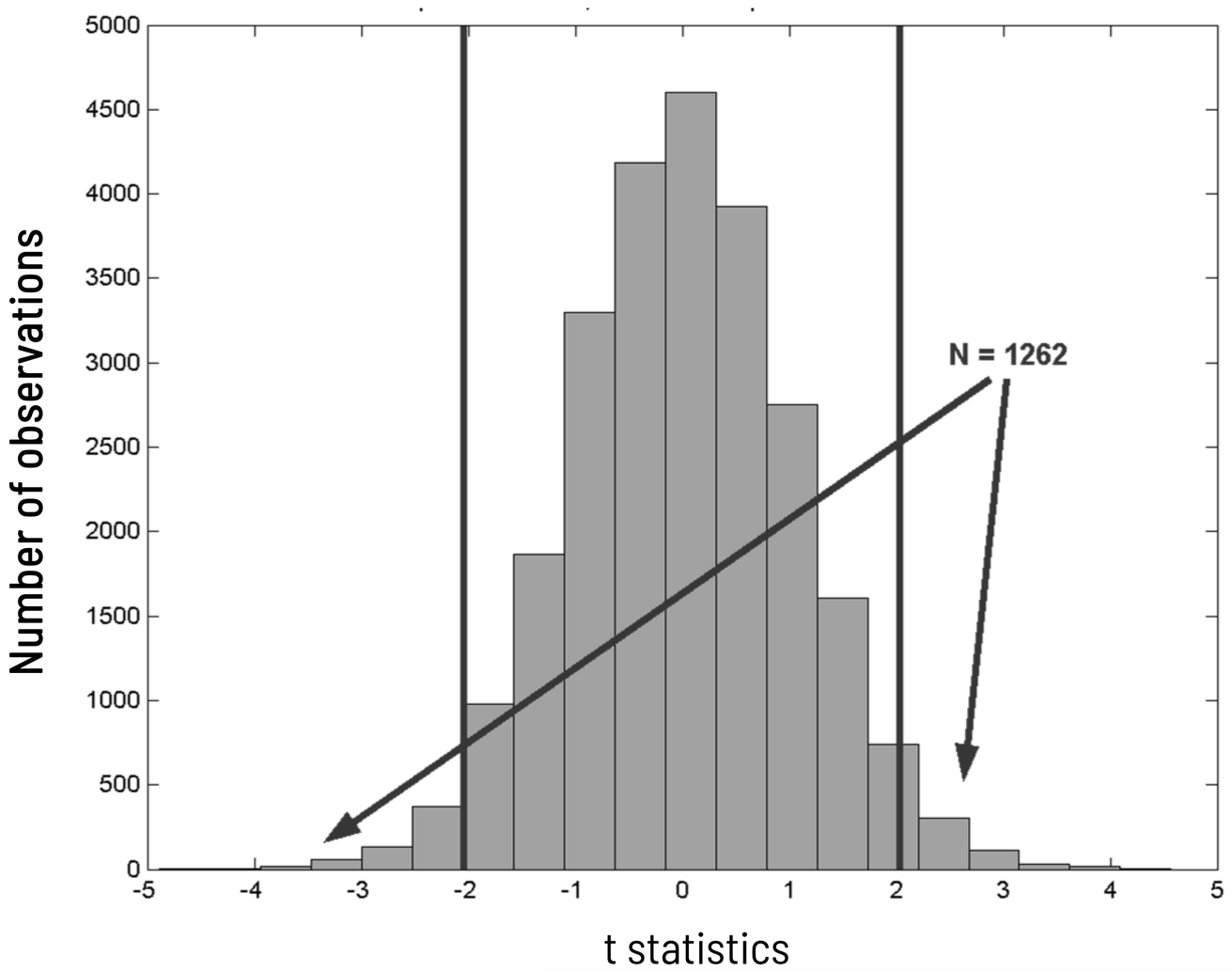
- Let's generate two series of 16 measurements that are observations of random variables with the distribution N (0,1).
- We will calculate the values of the test statistics and the associated p-values of the test verifying the hypothesis about the equality of the mean values of both distributions (t test).
- We will repeat the experiment n = 25 000 times.

Distribution of the test statistics



False positives

For α=0.05 and two-sided test $t_{critical}$= 2.04



False discoveries ...

How can we deal with FPs and FNs?

- **experimental validation** – same platform, different platforms (p-value integration);
- **statistical techniques** - multiple testing correction (statistical analyses), multiple random validation (machine learning: classification, clusterisation, and so on);
- **literature search;**
- **functional analysis;**
- **estimation of the effect size.**

Correction for multiple testing

- In the case of multiple testing of statistical hypotheses, the goal is to control the number of false positive results not only at the level of a single test but the whole series (test family).

Error control methods

A number of type I error control indicators have been proposed:

1. PCER *(per-comparison-error-rate)* - equal to the expected number of false positive test results (rejection of the null hypothesis if it is true) related to the number of tests performed → $E(FP)/n = \frac{1}{n}\sum_{i=1}^{n} \alpha_i$
2. PFER (*per-family-error-rate)* - equal to the expected number of false positive test results → $E(FP) = \sum_{i=1}^{n} \alpha_i$
3. FWER *(family-wise-error-rate)* - equal to the probability of rejecting at least one true null hypothesis → $P(FP \geq 1)$
4. FDR *(false-discovery-rate)* - defined as the expected value of the ratio of false positive (FP) results in the positive (R) group → $E\left(\frac{FP}{R}\middle| R > 0\right)$

FWER (family-wise-error-rate)

- The FWER index is equal to the probability of rejecting at least one true null hypothesis:

$$FWER = \Pr(FP \geq 1)$$

- The most popular methods to control FWER were proposed by Bonferroni (1936) and Dunn-Šidák (1958, 1967).

FWER

- By using Bonferroni's inequality

$$\Pr\left(\bigcup_{k=1}^{n} E_k\right) \leq \sum_{k=1}^{n} \Pr(E_k)$$

a correction of the significance level of a single test is given that allows FWER control to not greater than π:

$$FWER \leq \pi \Rightarrow \alpha = \frac{\pi}{n}$$

This procedure is called the Bonferroni correction (1936).

FWER – example

i	1	2	3	4	5	6	7	8	9	10
p(i)	0.0020	0.0045	0.0060	0.0080	0.0085	0.0090	0.0175	0.0250	0.1055	0.5350

Required error level π = 0.05.

Bonferroni's correction means verifying a single hypothesis at 0.005.

We reject hypothesis 1 and 2, accepting the others.

FWER

- Assuming the independence of tests, the FWER index can be estimated as equal to:

$$FWER = \Pr(FP \geq 1) = 1 - \Pr(FP < 1) = 1 - \prod_{j=1}^{n}(1 - \alpha_j)$$

- If we assume that each test will be carried out with the same level of significance α, then for a fixed value of the FWER indicator

$$FWER = \pi = 1 - (1 - \alpha)^n \Rightarrow \alpha = 1 - (1 - \pi)^{1/n}$$

- This method is called the **Dunn-Šidák method** (1958, 1967).

FWER – example

i	1	2	3	4	5	6	7	8	9	10
p(i)	0.0020	0.0045	0.0060	0.0080	0.0085	0.0090	0.0175	0.0250	0.1055	0.5350

Required error level π = 0.05.

Dunn-Sidak's correction means verifying a single hypothesis at $\alpha = 1 - (1 - 0.05)^{\frac{1}{10}} = 0.0051$.

We reject hypothesis 1 and 2, accepting the others.

FWER - stepwise methods

- Testing power is significantly reduced by using Bonferroni or Dunn-Šidák corrections.
- In the stepwise methods, the α level is corrected with each subsequent test, taking into account only the remaining tests to be corrected.

FWER – stepwise methods

The most popular method is the **Holm method** (1979).

1. Sort the results (p values) of individual tests from the smallest to the highest value $p(1) \leq p(2) \leq \ldots \leq p(n)$ and denote by H(i) the hypothesis associated with the value of p(i).
2. If $p(1) > \pi/n$ – no H(i) hypothesis will be rejected
3. If $p(1) \leq \pi/n$ – there is an evidence to reject the hypothesis H(1);
4. If $p(2) > \pi/(n-1)$ – acceptance of all H(i) hypotheses for i = 2, ..., n
5. If $p(2) \leq \pi/(n-1)$ – there is an evidence to reject the hypothesis H(2);
6. ...

FWER – stepwise methods

i	1	2	3	4	5	6	7	8	9	10
p(i)	0.0020	0.0045	0.0060	0.0080	0.0085	0.0090	0.0175	0.0250	0.1055	0.5350
$\frac{\pi}{n-i+1}$	0.0050	0.0056	0.0063	0.0071	0.0083	0.0100	0.0125	0.0167	0.0250	0.0500

Required error level π = 0.05.

1. Holm method: we start from p(1); reject if $p(i) \leq \pi/(n-i+1)$
2. For our data $i \leq 3$.

FWER - stepwise methods

An alternative to the Holm procedure is the Simes-Hochberg procedure (1986, 1988), which starts from the highest value p(n)

1. If $p(n) \leq \pi$ then all H(i) hypothese are rejected, i = 1, ..., n;
2. If not, there is no evidence to reject H(n), and H(n-1) is analyzed
3. If $p(n-1) \leq \pi/2$ then all H(i) hypothese for $i \leq n-1$ are rejected;
4. If not, then H(n-2) is analysed with the significance level $\pi/3$ and so on

The Simes-Hochberg correction procedure is more powerful than the Holm method, but it can be used when the individual tests are independent of each other. The Holm method does not have such a limitation.

FWER – stepwise methods

i	1	2	3	4	5	6	7	8	9	10
p(i)	0.0020	0.0045	0.0060	0.0080	0.0085	0.0090	0.0175	0.0250	0.1055	0.5350
$\frac{\pi}{N-i+1}$	0.0050	0.0056	0.0063	0.0071	0.0083	0.0100	0.0125	0.0167	0.0250	0.0500

Required error level π = 0.05.

1. Simes-Hochberg method: we start from p(n); we accept the null hypothesis H(i) if p(i) > **π**/(n-i+1)
2. For our data it means i ≥ 7.
3. Finally we reject H(1)...H(6).

FWER – stepwise methods

- The stepwise correction method proposed by **Hommel** (1989) is slightly more complicated but does not reduce as much as the other test power.
- Hypotheses for which the value of p is less than or equal to π/k are rejected, where

$$k = \max_i p(n - i + j) > \pi \frac{j}{i} \quad dla \quad j = 1,\ldots,i$$

FWER – stepwise methods

i	1	2	3	4	5	6	7	8	9	10
p(i)	0.0020	0.0045	0.0060	0.0080	0.0085	0.0090	0.0175	0.0250	0.1055	0.5350

i	j	j/i	**(π = 0.05)* j/i**	
1	1	1	0.05	p(10) > 0.05
2	1; 2	0.5; 1	0.025; 0.05	p(9) > 0.025; p(10) > 0.05
3	1; 2; 3	0.33; 0.67; 1	0.0167; 0.033; 0.05	p(8) > 0.0167; p(9) > 0.033; p(10) > 0.05
4	1; 2; 3; 4	0.25; 0.5; 0.75; 1	0.0125; 0.025; 0.0375; 0.05	p(7) > 0.0125; p(8)=0.025; p(9) > 0.0375; p(10) > 0.05

Calculations carried out by the **Hommel method** give the value of the coefficient k = 3, so for a single test we use the level $\alpha = 0.05 / 3 = 0.0167$.

We reject the hypotheses H(1), ..., H(6) because p(i) < 0.0167.

FDR (false-discovery-rate)

- The FDR is defined as the expected value of the ratio of false positive (FP) results in the positive (R) group. The authors differ only in the definition of this indicator for the case when R = 0. Benjamini and Hochberg (1995) propose to take the value 0.

$$FDR = E\left(\frac{FP}{R}\middle| R > 0\right)\Pr(R > 0)$$

- Storey (2002) proposes to set the indicator only when R> 0. Then

$$pFDR = E\left(\frac{FP}{R}\middle| R > 0\right)$$

FDR

Benjamini and Hochberg procedure.

The steps of the algorithm are as follows:

- Sort the p values of individual tests from the smallest to the highest value p(1) ≤ p(2) ≤ ... ≤ p(n) and denote by H(i) the hypothesis associated with the value of p(i).
- Rejection of null H (i) hypotheses such that $p(i) \leq \frac{i\pi}{n}$ guarantees $FDR \leq \frac{n_0\pi}{N} \leq \pi$

FDR

i	1	2	3	4	5	6	7	8	9	10
p(i)	0.0020	0.0045	0.0060	0.0080	0.0085	0.0090	0.0175	0.0250	0.1055	0.5350
$\frac{i\pi}{n}$	0.005	0.010	0.015	0.020	0.025	0.030	0.035	0.040	0.045	0.050
$FDR = p(i) \cdot \frac{n}{i}$	0.0020 *10 = 0.020	0.0045 *10/2 = 0.0225	0.0060 *10/3 = 0.020	0.0080 *10/4 = 0.020	0.0085 *10/5 = 0.017	0.0090 *10/6 = 0.015	0.0175* 10/7 = 0.025	0.0250 *10/8 = 0.0313	0.1055* 10/9 = 0.1172	0.5350* 10/10 = 0.5350

Assuming FDR = 0.05 for Benjamini-Hochberg method, we reject the hypotheses H(1), ..., H(8)

CASE B: A LOT OF OBSERVATIONS

Mass spectrometry imaging (MSI)

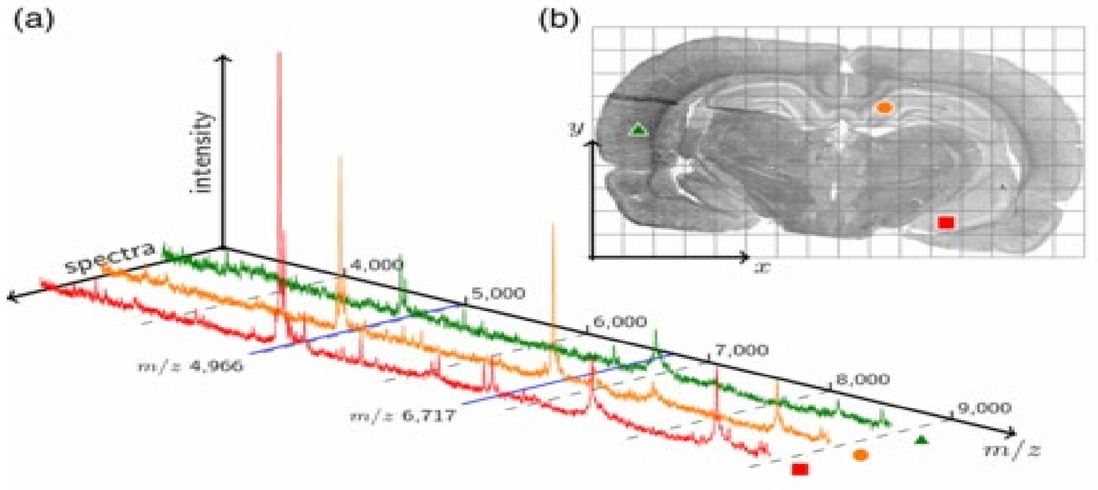
KIAED
KATEDRA INŻYNIERII I ANALIZY EKSPLORACYJNEJ DANYCH

- Over the last few years, mass spectrometry MS techniques have undergone important improvements, enabling the exploration of proteins along a wide range of molecular weights in biological samples.

- Mass spectrometers have been a revolution to the field of proteomics, allowing the researchers to build "signatures" or proteomic patterns specific to different conditions or pathological states.

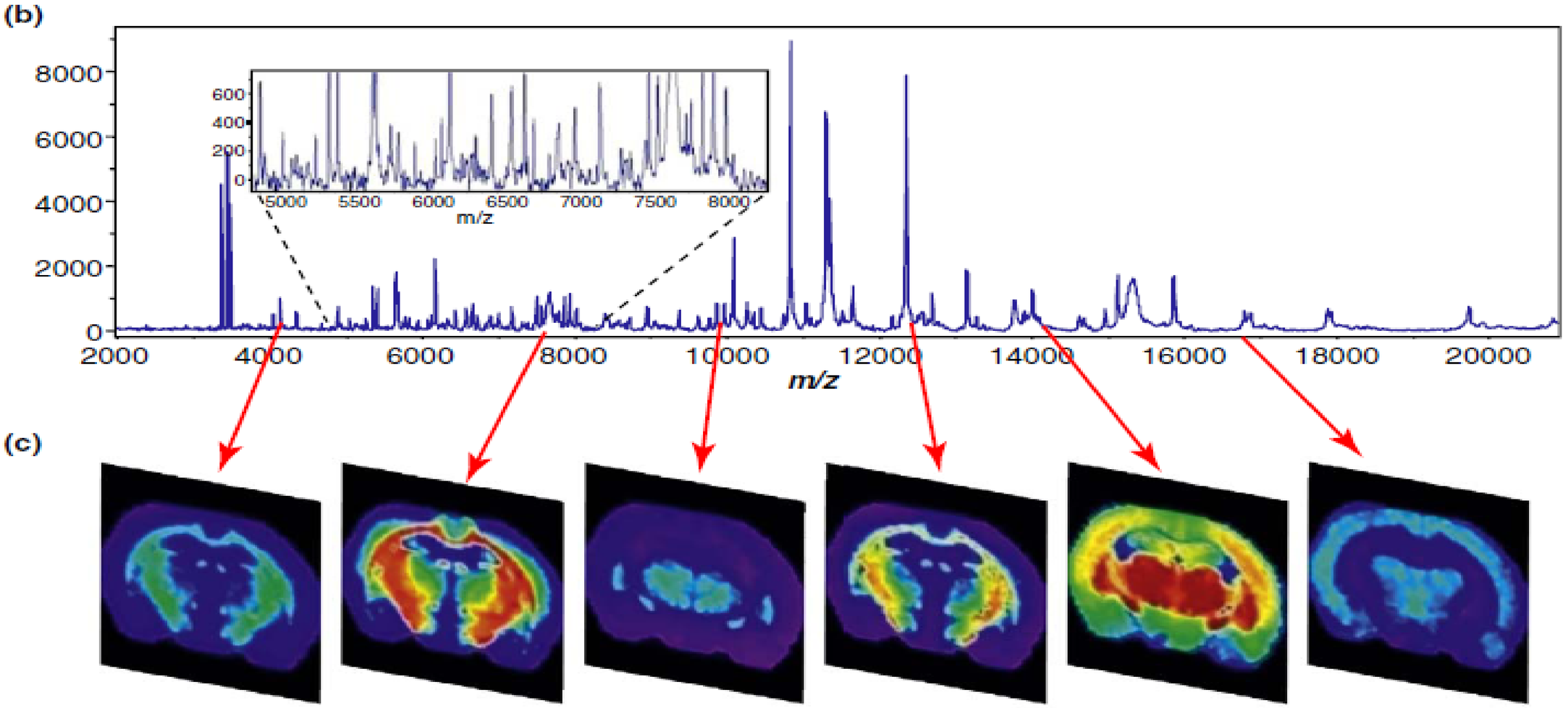
Mass spectrometry imaging (MSI)

- The application of MALDI-MSI in cancer research allows for the spatial identification of molecular profiles and their heterogeneity within the tumour, but leads to the creation of highly complicated datasets of great volume.



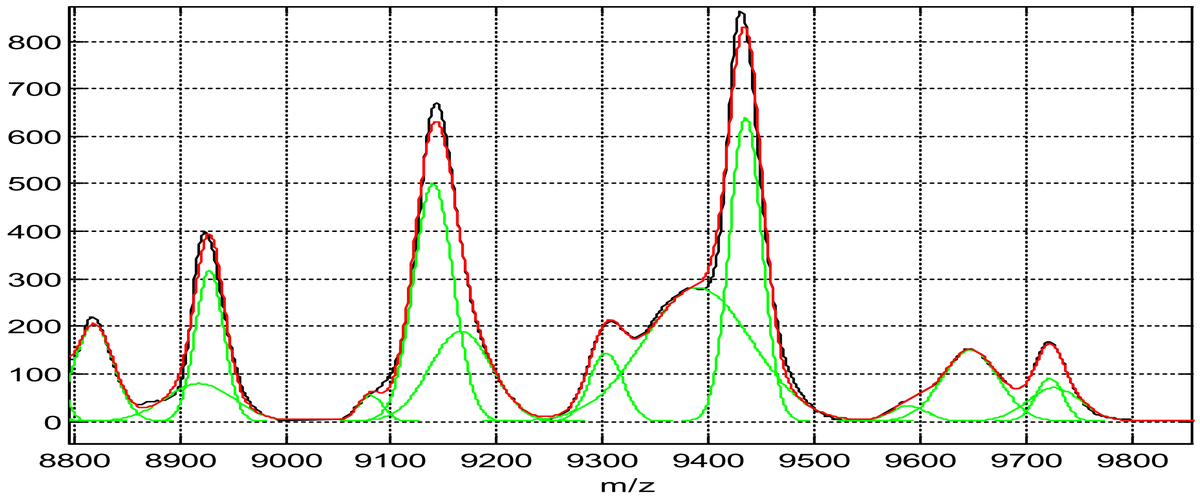
Schwamborn, Kristina; Caprioli, Richard M: M**olecular maging by mass spectrometry – looking beyond classical histology.** NATURE REVIEWS CANCER, 10(9):639-646, Sep. 2010

Mass spectrometry imaging (MSI)



Mass spectrometry imaging (MSI)

- ~ 100,000 mass channels per spectrum
- ~ 10,000 spectra per sample.
- ~ 2 billion numerical values per sample.
- Raw data ~ 8 GB per one sample.



Dimensionality reduction is needed

From dimensionality reduction technique we require to remove the data redundancy as much as possible and to loose the crucial information as less as possible. Gaussian mixture modelling (GMM) of the MALDI spectrum allows for efficient feature extraction (peak picking) in MSI data.

1. Polanski A, Marczyk M, Pietrowska M, Widlak P, Polanska J: *Signal Partitioning Algorithm for Highly Efficient Gaussian Mixture Modeling in Mass Spectrometry*. Plos ONE, 2015, 10(7): e0134256
2. Polanski A, Marczyk M, Pietrowska M, Widlak P, Polanska J*: Initializing EM algorithm for univariate Gaussian, multi-component, heteroscedastic mixture models by dynamic programming partitions*, International Journal of Computational Methods, 2018, 15(3): 1850012

Example

109,568 mass channels ranged from 800 ÷ 3,500 Da

3 tissue samples

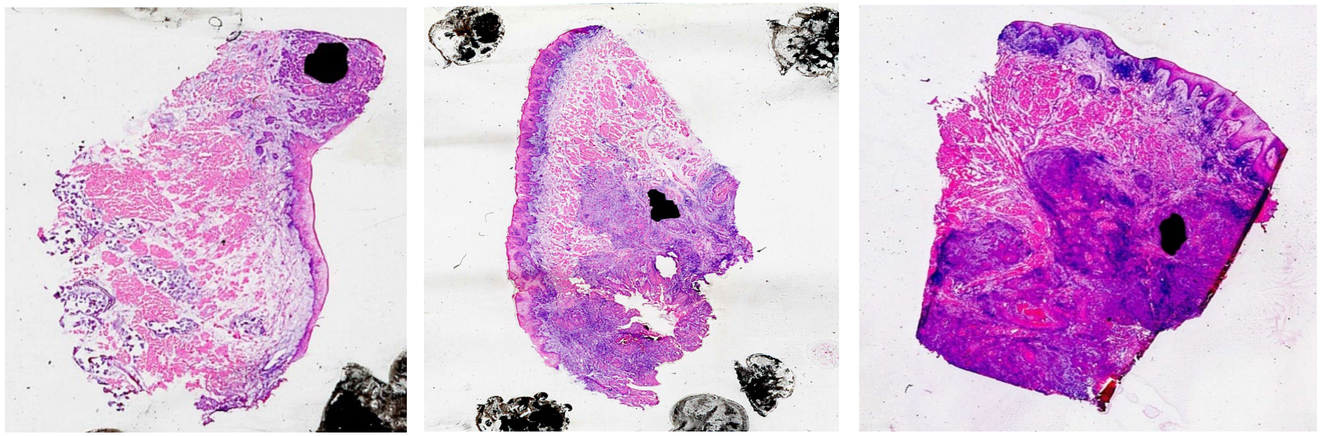
30,157 mass spectra

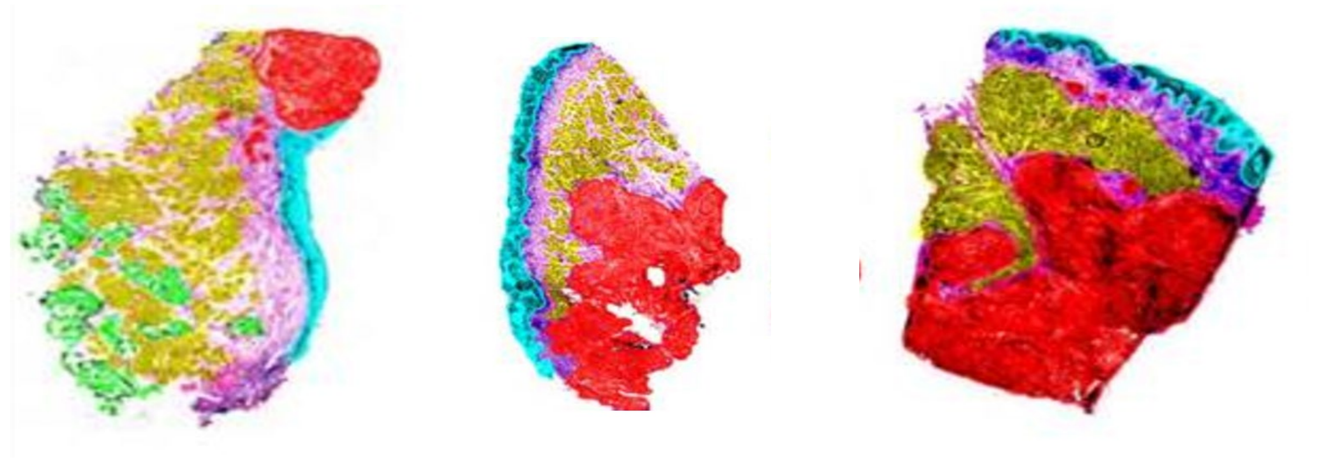
11,551 from the cancer region

1,535 from the healthy epithelium

6,216 Gaussian components for the complete model

2435 features in the final model





Statistical testing

KIAED
KATEDRA INŻYNIERII I ANALIZY EKSPLORACYJNEJ DANYCH

Cancer tissue versus healthy epithelium – signature identification

- Nonparametric Mann-Whitney U test : p-value ≤ 0.05; n_1 = **2386** (97.99%)
- Bonferroni correction: p-value ≤ 0.05/2435 = 0.00002; n_2 = **2316** (95.11%)

Is p-value enough?

Motivation

- „Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude – not just, does a treatment affect people, but how much does it affect them."

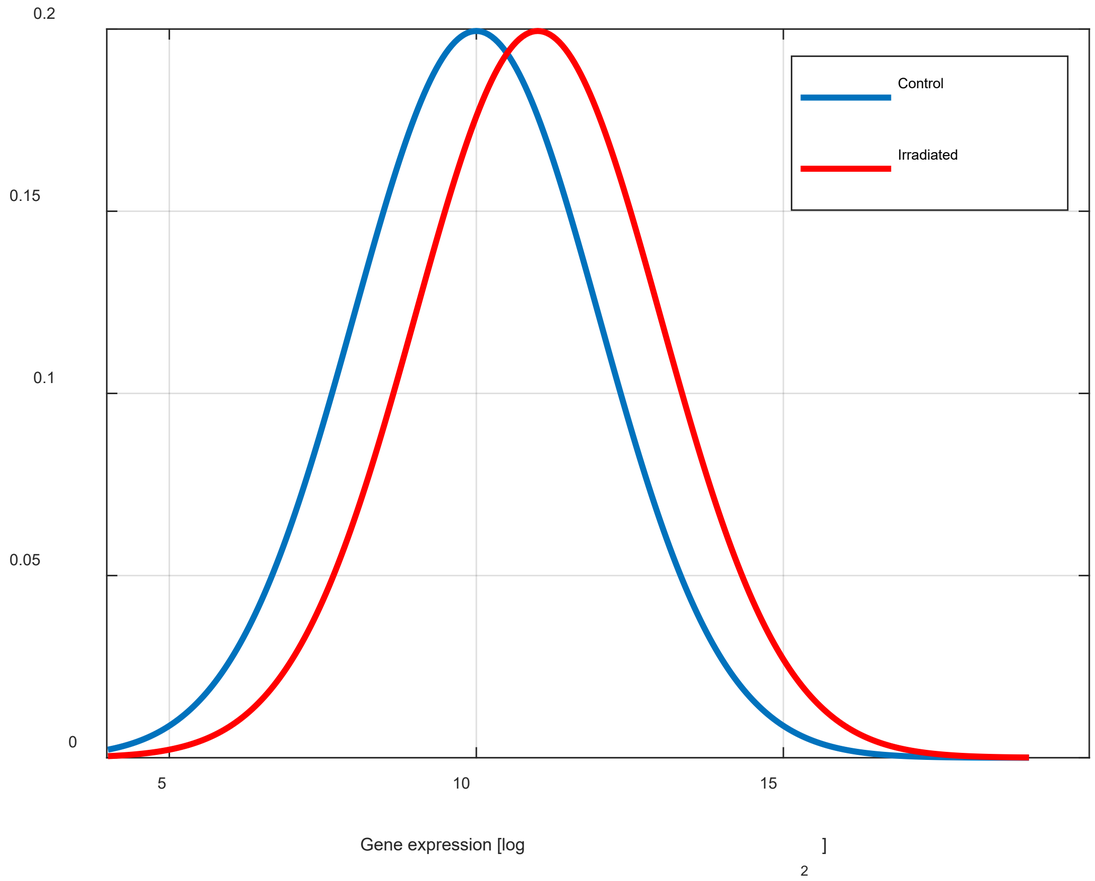
- Gene V. Glass (born June 19, 1940) - an American statistician and researcher working in educational psychology and the social sciences who introduced the term "meta-analysis" and illustrated its first use in his presidential address to the American Educational Research Association in San Francisco in April, 1976.

Why p-value is not enough?

- P-value answers the question **whether an effect exists**, not how great it is.
- The p-value directly depends on the size of the sample.
- With a sufficiently large sample, the statistical test will almost always indicate a significant difference, as long as these differences are not equal to zero. Hence, very small differences, even if significant, will be irrelevant.

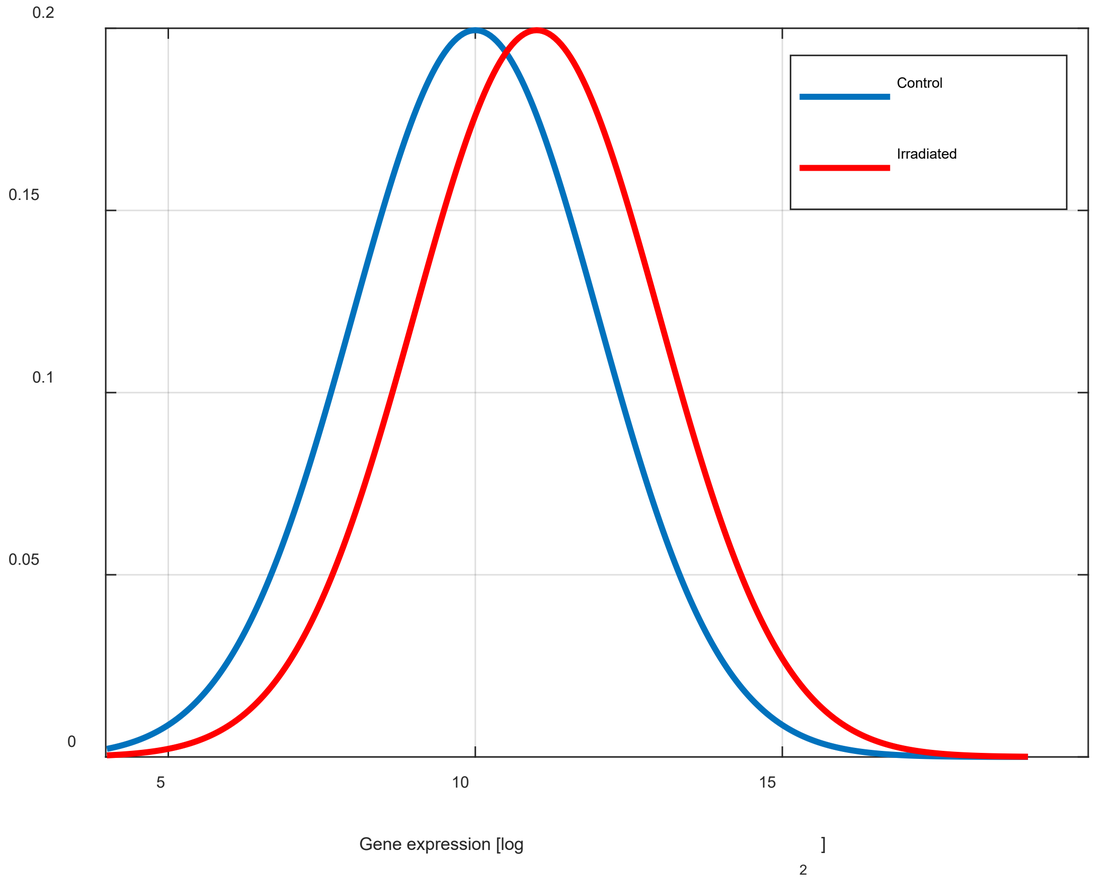
Example

- $\mu_1 = 10, \mu_2 = 11, \sigma_1 = \sigma_2 = 2$
- $H_0: \mu_1 = \mu_2 \qquad H_1: \mu_1 \neq \mu_2$
- $n_1 = n_2 = n = 10$
- $\overline{x_1} = 10.2;\ \overline{x_2} = 11.1;$
- $s_1 = 1.95;\quad s_2 = 2.3;$
- $s = \sqrt{\frac{{s_1}^2 + {s_2}^2}{2}} = 2.13$
- $t_1 = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s^2}{n} + \frac{s^2}{n}}} = \frac{10.2 - 11.1}{2.13\sqrt{2/10}} = -0.9448$ $\qquad p_1 = 0.3573$



Example

- $\mu_1 = 10, \mu_2 = 11, \sigma_1 = \sigma_2 = 2$
- $H_0: \mu_1 = \mu_2 \qquad H_1: \mu_1 \neq \mu_2$
- $n_1 = n_2 = n = 50$
- $\overline{x_1} = 10.2;\ \overline{x_2} = 11.1;$
- $s_1 = 1.95;\quad s_2 = 2.3;$
- $s = \sqrt{\frac{{s_1}^2 + {s_2}^2}{2}} = 2.13$
- $t_2 = \frac{\overline{x_1} - \overline{x_2}}{s \cdot \sqrt{\frac{2}{n}}} = \frac{10.2 - 11.1}{2.13\sqrt{2/50}} = -2.1127 \qquad p_2 = 0.0489$



Measure of magnitude

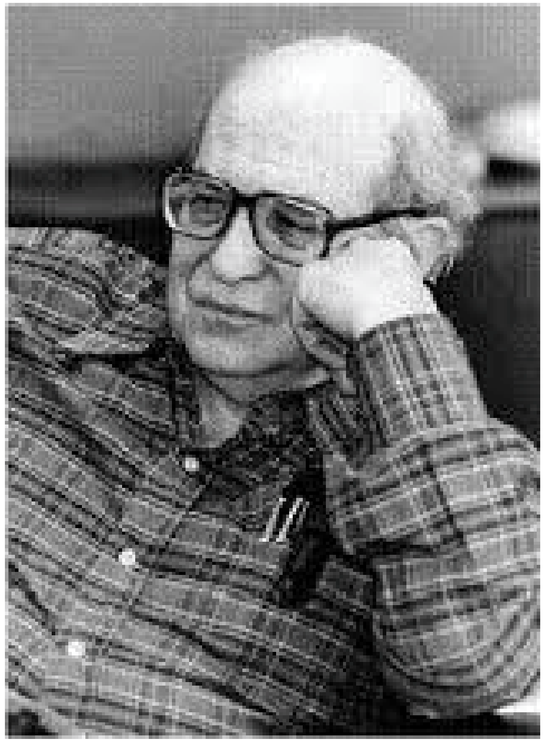
- **Effect size** - a quantitative measure of the strength of a phenomenon calculated on the basis of data;
- The measure of how great the effect of one variable exerts on the other variable;
- The effect size metric must be independent of the sample size.

Effect size - definition

- Jacob Cohen defined the concept of the magnitude of the effect as a degree to which the phenomenon exists.
- Cohen's d is defined as the difference between means divided by the standard deviation in the sample for two independent samples of the same size and equal but unknown variances.

$$d = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{(s_1^2 + s_2^2)/2}}$$

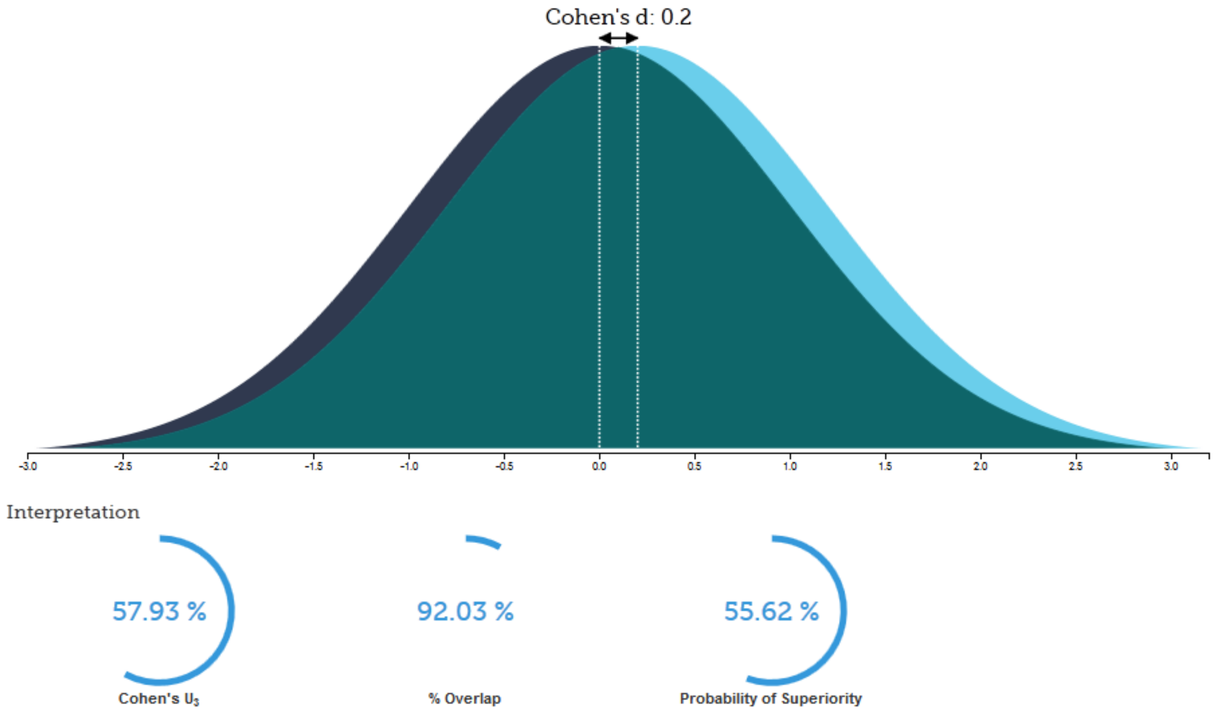
- At the moment there are about 100 different measures of the size of the effect.



$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{(s_1^2 + s_2^2)/(2n)}} \rightarrow d = \frac{t}{\sqrt{n}}$$

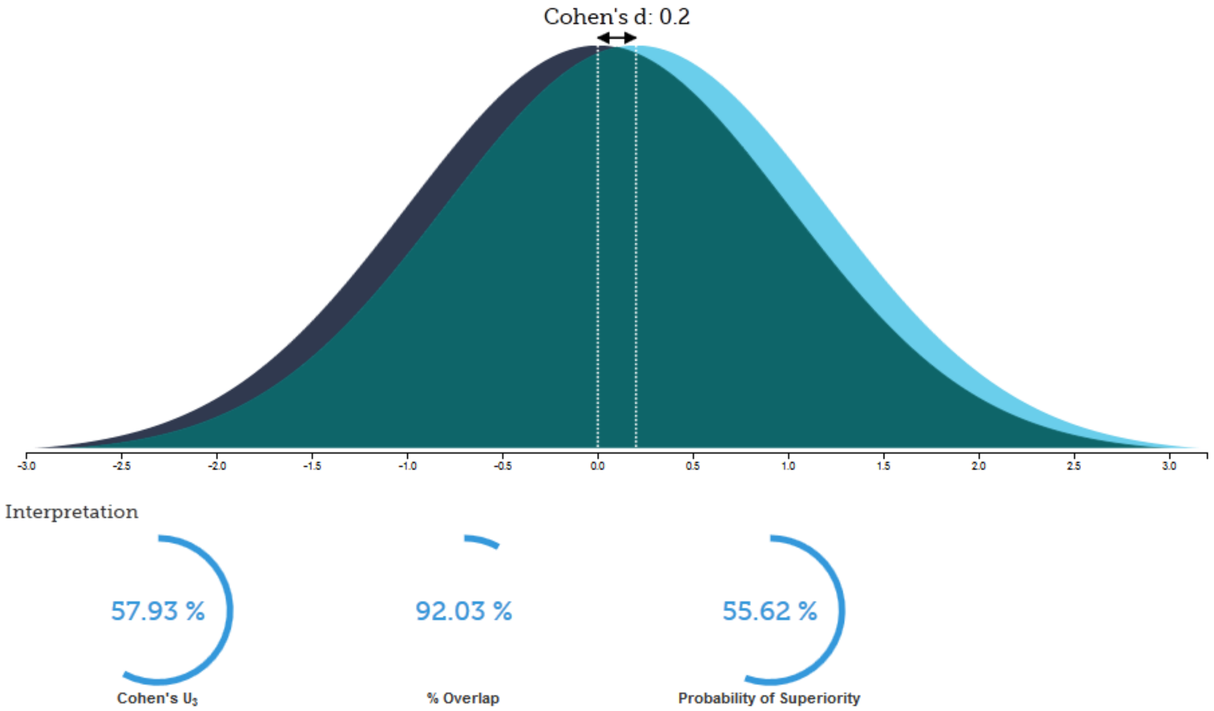
Jacob Cohen: Statistical Power Analysis for the Behavioral Sciences, Routledge, 1988

How do we interpret Cohen's d?

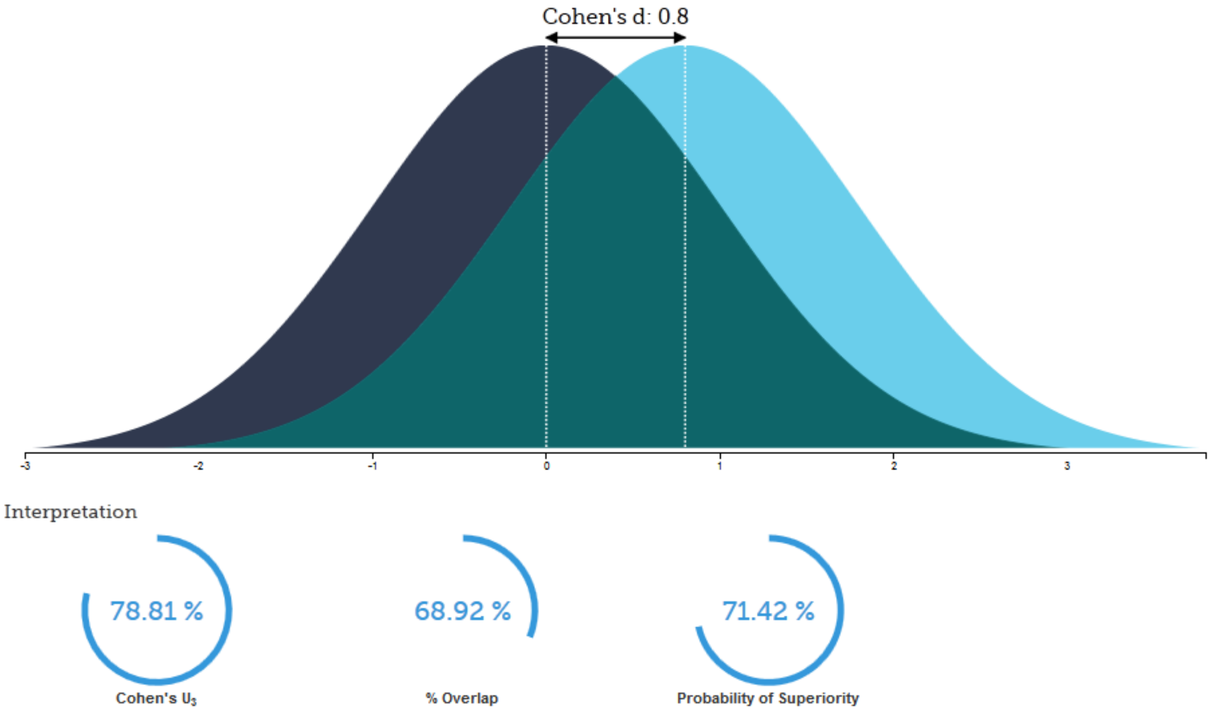


- With a Cohen's d of 0.2, ~58% of the treatment group will be above the mean of the control group (Cohen's U3), 92% of the two groups will overlap, and there is a 56% chance that a person picked at random from the treatment group will have a higher score than a person picked at random from the control group (probability of superiority).

How do we interpret Cohen's d?

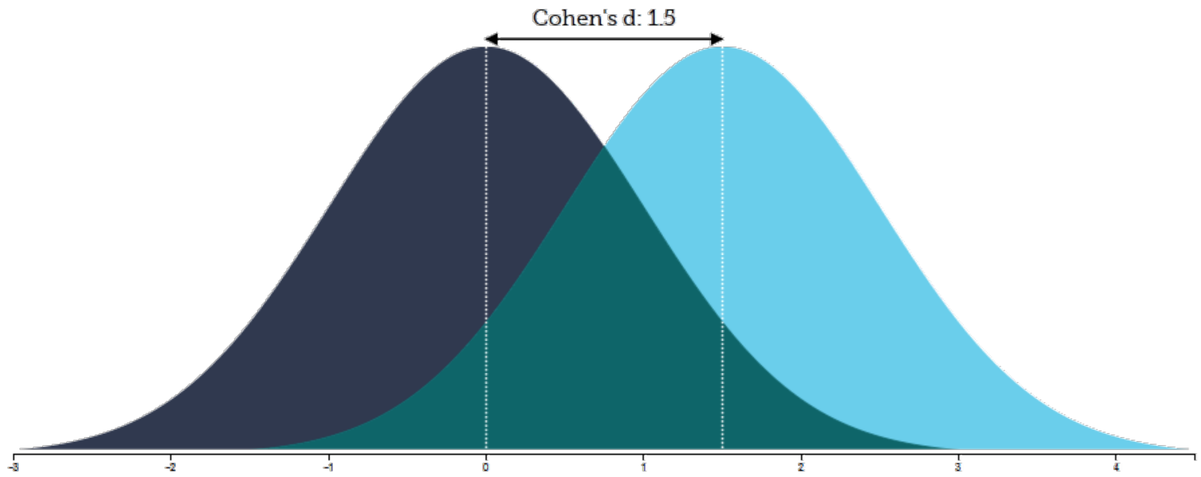


% of treatment above the control mean



% chance that a person picked at random from the treatment group will have a higher score than a person picked at random from the control group

How do we interpret Cohen's d?

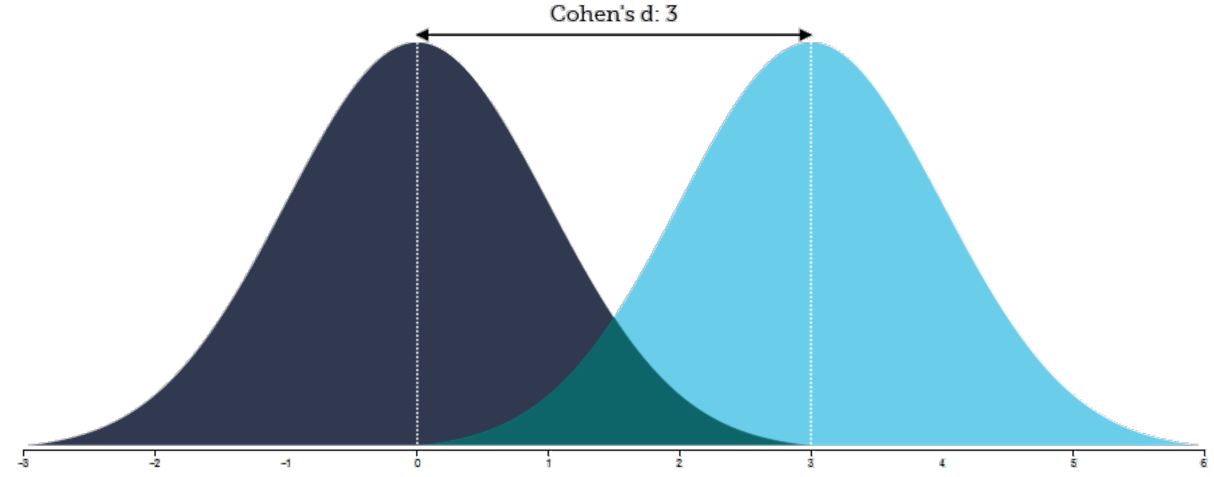


Interpretation

Cohen's U_3

% Overlap

Probability of Superiority



Interpretation

Cohen's U_3

% Overlap

Probability of Superiority

How do we interpret Cohen's d?

KIAED
KATEDRA INŻYNIERII I ANALIZY EKSPLORACYJNEJ DANYCH

Effect size (d-value)	Cohen's U_3 (% of treatment above the control mean)	% of non-overlap
0	50	0
0.2	58	15
0.5	69	33
0.8	79	47
1.0	84	55
1.5	93	71
2.0	97	81
3.0	99.9	87

Effect size

Depending on the type of research, the effect size metrics can be divided into two groups:

- Indicating differences between groups (*d* family): risk difference, risk ratio, odds ratio, Cohen's d, Glass's delta, Hedges' g, the probability of superiority, ω^2;

- Estimating measure of similarity between variables (*r* family): the correlation coefficient r, R^2, Spearman's rho, Kendall's tau, phi coefficient, Cramer's V, Cohen's f, η^2

Cohen’s d - one-sample and paired tests

KIAED
KATEDRA INŻYNIERII I ANALIZY EKSPLORACYJNEJ DANYCH

- For a one-sample t-test Cohen's d for n subjects in the group.

$$d = \frac{\bar{x} - x_0}{s} = \frac{t}{\sqrt{n}}$$

- In a paired t-test, Cohen's d equals (Rosenthal, 1991) $d = \frac{t}{\sqrt{n}}$
- Dunlap *et al.* (1996) suggest using an alternative estimator

$$d = t \cdot \sqrt{\frac{2(1-r)}{n}}$$

for n subjects and a correlation r between the paired responses.

Hedge's g – unequal sizes and roughly equal variances

- Cohen's d relies on the average standard deviation (the denominator of equation) to standardize the measure of the ES; it assumes the groups having (roughly) equal size and variance. When deviation from this assumption is not negligible (e.g. one group doubles the other) it is possible to account for it using the Bessel's correction for the biased estimation of sample standard deviation.
- This gives rise to the Hedge's g, which is a standardized mean difference corrected by the pooled weighted standard deviation.

Hedge's g – unequal sizes and roughly equal variances

- Hedge's g is defined as the difference between means divided by the pooled standard deviation in the sample:

$$g = \frac{|\overline{x_1} - \overline{x_2}|}{s_{pooled}} \quad s_{pooled} = \sqrt{\frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2}}$$

  for independent samples.

- Hedge's g shall be used in case of the small unequal sample sizes and/or slightly unequal variances.

Glass's Δ – treatment group versus control

KIAED
KATEDRA INŻYNIERII I ANALIZY EKSPLORACYJNEJ DANYCH

- A particular case of ES estimation involves experiments in which one of the two groups acts as a control.
- In that we presume that any measure on control is untainted by the effect, we can use its standard deviation to standardize the difference between averages in order to minimize the bias, as it is done in the Glass's Δ:

$$\Delta = \frac{|\overline{x_1} - \overline{x_2}|}{s_{control}}$$

Interpretation of the effect size

TABLE 1: Thresholds for interpreting effect size

Test	Relevant effect size	Effect size threshold			
		Small	Medium	Large	Very large
Standardized mean difference	d, Δ, Hedges' g	.20	.50	.80	1.30

Notes: The rationale for these benchmarks can be found in Cohen (1988) at the following pages: d (p.40) and r (pp.79-80). Supplementing Cohen's (1988) original small, medium and large effect sizes, Rosenthal (1996) added a classification of very large, defined as being equivalent to, or greater than d = 1.30

Example - MSI

Effect size	Number of features	Min p-value
Small [0.2, 0.5)	1771	2e-75
Medium [0.5, 0.8)	116	5e-183
Large [0.8, 1.3)	1	5e-194
Huge ≥ 1.3	0	

One-way ANOVA

- It is possible to extend the framework of difference family also to more than two groups, correcting the overall difference (difference of each observation from the average of all observations) by the number of groups considered.

- Under a formal point of view this corresponds to the omnibus effect of a 1 factor analysis of variance design with fixed effect (One-way ANOVA).

One-way ANOVA – η^2

- For a one-way analysis of variance, expresses the proportion of the total variance that can be assigned to an independent variable ranging from 0 to 1.

$$\eta^2 = \frac{SS_{between}}{SS_{total}} = \frac{SS_{factor}}{SS_{total}}$$

Effect size	η^2
Small	0.01
Medium	0.06
Large	0.14

One-way ANOVA – ε^2 and ω^2

- η^2 was developed as a descriptive index while ε^2 and ω^2 are intended for inferential purposes and are constructed by substituting the variance component parameters of η^2 with its bias-corrected sample estimators.

$$\varepsilon^2 = \frac{SS_{factor} - SS_{error}\frac{k-1}{N-k}}{SS_{total}}$$

$$\omega^2 = \frac{SS_{factor} - \frac{k-1}{N-k}SS_{error}}{SS_{total} + SS_{error}/(N-k)}$$

Effect size	η^2
Small	0.01
Medium	0.06
Large	0.14

Okada, K. (2013). *Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA*. Behaviormetrika, 40(2), 129-147.

Linear correlation – Pearson's r

- In the association-based family the effect is measured as the size of variation between two (or more) variables observed in the same or in several different samples. Within this family it is possible to do a further distinction, based on the way the variability is described.
- In the first sub-family, variability is shown as a joint variation of the variables considered. Under a formal point of view it is nothing but the concept which resides in the Pearson's product moment correlation coefficient, which is indeed the progenitor of this group.

Linear correlation – Pearson's r

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2}\sqrt{\sum(x - \bar{x})^2}}$$

TABLE 1: Thresholds for interpreting effect size

Test	Relevant effect size	Effect size threshold			
		Small	Medium	Large	Very large
Standardized mean difference	*d*, Δ, Hedges' *g*	.20	.50	.80	1.30
Correlation	*r*	.10	.30	.50	.70

Notes: The rationale for these benchmarks can be found in Cohen (1988) at the following pages: *d* (p.40) and *r* (pp.79-80). Supplementing Cohen's (1988) original small, medium and large effect sizes, Rosenthal (1996) added a classification of very large, defined as being equivalent to, or greater than *d* = 1.30 or *r* = .70.

Nonparametric significance tests

KIAED
KATEDRA INŻYNIERII I ANALIZY EKSPLORACYJNEJ DANYCH

For Mann-Whitney U test the effect size is measured by rank biserial correlation coefficient.

The most popular Wendt's formula is as follows:

$$r_{bc} = 1 - 2U/(n_1 \cdot n_2)$$

where U stands for Mann-Whitney U statistics (in the case of two-sided test $U = \min(U_1, U_2)$).

Effect size	Small	Medium	Large	Very large
r_{bc}	0.10	0.30	0.50	0.70

Nonlinear association

- When a nonlinear association is thought to be present, or the continuous variable were discretized into ranks, it is possible to use the Spearman's rho (ρ) instead.
- Alternatively, for those variable naturally nominal, if a two-by-two (2 x 2) table is used, it is possible to calculate the ES through the coefficient phi.
- In case of unequal number of rows and columns, the Cramer's V can be used, in which a correction factor for the unequal ranks is used, similarly to what is done with the difference family.

Binary association – Pearsons' phi

- Pearson's correlation coefficient for binary data, Φ, is an effect size measure for the 2x2 table given by

$$\Phi = \frac{a - n_1 m_1}{\sqrt{n_1 m_1 n_0 m_0}}$$

	Y0	Y1	
X0	a	b	n_0
X1	c	d	n_1
	m_0	m_1	n

- With the thresholds proposed by Cohen

Effect size	Small	Medium	Large
Pearson's Φ	0.10	0.30	0.50

Nominal association - Cramér's *V*

- It is used for contingency tables to measure the correlation for data consisting of two categorical variables that have two or more than two levels. It ranges from 0 to 1.

	Y0	Y1	
X0	n_{11}	n_{12}	r_1
X1	n_{21}	n_{22}	r_2
	c_1	c_2	n

$$V = \frac{\chi^2}{n(m-1)}$$

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \frac{r_i * c_j}{n})^2}{\frac{r_i * c_j}{n}}$$

m-1= min(r-1,c-1)	**Small**	**Medium**	**Large**
1	0.10	0.30	0.50
2	0.07	0.21	0.35
3	0.06	0.17	0.29

Hays, W. L. 1981. Statistics for the social sciences, 3rd ed. New York: Holt, Rinehart, and Winston
Cohen, Jacob (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

2x2 contingency tables

- The odds ratio (OR) can be regarded as a peculiar kind of ES measure because is suits both 2 x 2 contingency tables as well as non-linear regression models like logistic regression.
- In general, OR can be tought as a special kind of association family ES for dicothomous (binary) variables. The OR represents the likelihood that an event occurs due to a certain factor against the probability that it arises just by chance (that is when the factor is absent).

Odds ratio

- For 2 x 2 tables the OR can be easily calculated using the cross product of cells frequency. OR can be also estimated by means of logistic regression.

$$\mathrm{OR} = \frac{(x_1 y_1)(x_0 y_0)}{(x_1 y_0)(x_0 y_1)} = \frac{ad}{bc}$$

$$\mathrm{OR} = e^{\beta} \quad \text{for logistic regression}$$

	Y0	Y1	
X0	a	b	n_0
X1	c	d	n_1
	m_0	m_1	n

Effect size for balanced design	Small	Medium	Large
OR (2x2 table) by Cohen	1.49	3.45	9.0

Relative risk and risk difference

- Let

  p_1 = risk of disease among Group 1 (exposed)

  p_2 = risk of disease among Group 2 (unexposed)

- Then it is reasonable to estimate

$$\hat{p}_1 = \frac{a}{n_0} = \frac{a}{a+b} \qquad RD = \hat{p}_1 - \hat{p}_2$$

$$\hat{p}_2 = \frac{c}{n_1} = \frac{c}{c+d} \qquad RR = \frac{\hat{p}_1}{\hat{p}_2}$$

Risk difference for 2x2 tables

- The effect size mesure h is defined as a difference between the arcsine transformation values of proportions p_1 and p_2: $h = \arcsin(p_1) - \arcsin(p_2)$.
- If we restrict ourselves to the part of the proportion p scale between 0.05 and 0.95, the range of RD is tolerably small. Thus, we do not have to pay a large price in consistency of interpretation of effect size h in terms of $p_1 - p_2$.

Effect size	Small	Medium	Large
Difference in arcsines	0.20	0.50	0.80

Risk ratio (relative risk) for 2x2 table

- Assuming allocation ratio m_1/N as Δ equal to 0.5 (balanced design) recommendations for corrected a for small (a=0.1), medium (a=0.3), and large (a=0.5) effect size are:

Risk measure	For particular **Φ** threshold a and allocation ratio **Δ**	Relative effect (for **Δ**=0.5)
Relative risk, odds ratio (rare events), hazard ratio, incidence ratio, standardized mortality ratio	$RR_a = 1 + \frac{a}{(1-a)\Delta}$	Small: 1.22 Medium: 1.86 Large: 3.0
Odds ratio (no-rare events)	$OR_a = 1 - \frac{ln\big((1-\Delta)(1-a)\big)}{\Delta(1-a)+a}(RR_a - 1)$	Small: 1.36 Medium: 2.38 Large:4.70

Jake Olivier, Warren L. May & Melanie L. Bell (2017) Relative Effect Sizes for Measures of Risk. Communications in Statistics - Theory and Methods, 46(14):6774-6781

Example – OR

- During a 5-year study, among the 9 trials in the analysis, 50,868 subjects were treated with aspirin and 49,170 received placebo or control. More than 22,000 patients identified the association of aspirin with a decreased number of myocardial infractions (MI).

End Point	Odds Ratio	95% Confidence Interval	p Value
Total CHD	0.854	0.688–1.061	0.154
Nonfatal MI	0.813	0.667–0.992	0.042
Total CV events	0.865	0.804–0.930	0.001
Stroke	0.919	0.828–1.021	0.116
Cardiovascular mortality	0.956	0.799–1.143	0.619
All-cause mortality	0.945	0.881–1.014	0.115

Example – OR

- The outcome of the study: OR=0.813 (1/OR=1.23 for the „risk" factor) p-value=0.042. For allocation Δ=50,868/(50,868+49,170)=50.85% the corrected small effect threshold (a=0.1) equals

$$RR_{0.1} = 1 + \frac{0.1}{(1-0.1)0.5085} = 1.2185$$

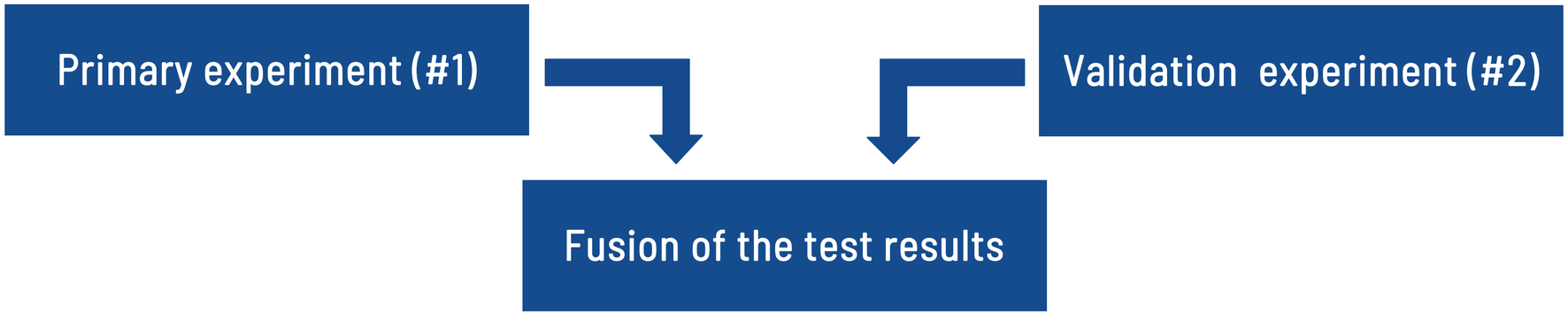
$$OR_{0.1} = 1 - \frac{\ln\big((1-0.5085)(1-0.1)\big)}{0.5085(1-0.1)+0.1}(1.2185-1) = 1 - \frac{\ln(0.4424)}{0.5577}0.2185 = 1.3195$$

- Since the obtained risk factor OR = $\frac{1}{0.813}$ = 1.23 < $OR_{0.1}$ suggests very small effect size, the recommendation to use aspirin for preventive purposes was discontinued. Further research indicated an even lower amount of effect.

Alfred A. Bartolucci, Michal Tendera, George Howard, *Meta-Analysis of Multiple Primary Prevention Trials of Cardiovascular Events Using Aspirin*, The American Journal of Cardiology, 2011, 107(12):1796-1801

CASE C: EXPERIMENTAL VALIDATION

Problem formulation



Goal: Compare the results of the different analyses, which may be dependent or independent designs.

One of the options is p-value integration.

What is a p-value?

- The probability that the phenomenon observed in some sample measurements could occur by chance, due to random variability of the samples, in a situation where such a phenomenon does not occur at all in the population.

or

- The probability, calculated under the assumption that H0 is true, that the test statistic would be as extreme or more extreme than what is actually observed is called the p-value of the test.

P-value tells us on what level of significance (noted as α) the null hypothesis can be rejected.

Types of statistical integration

- **Direct / independent integration** - the same experiment performed on a different set of data (we measure the same feature on an independent set of input data);
- **Indirect / dependent integration -** we use the same input data for another trial (we measure the same phenomenom by other means).

INDEPENDENT VALIDATION

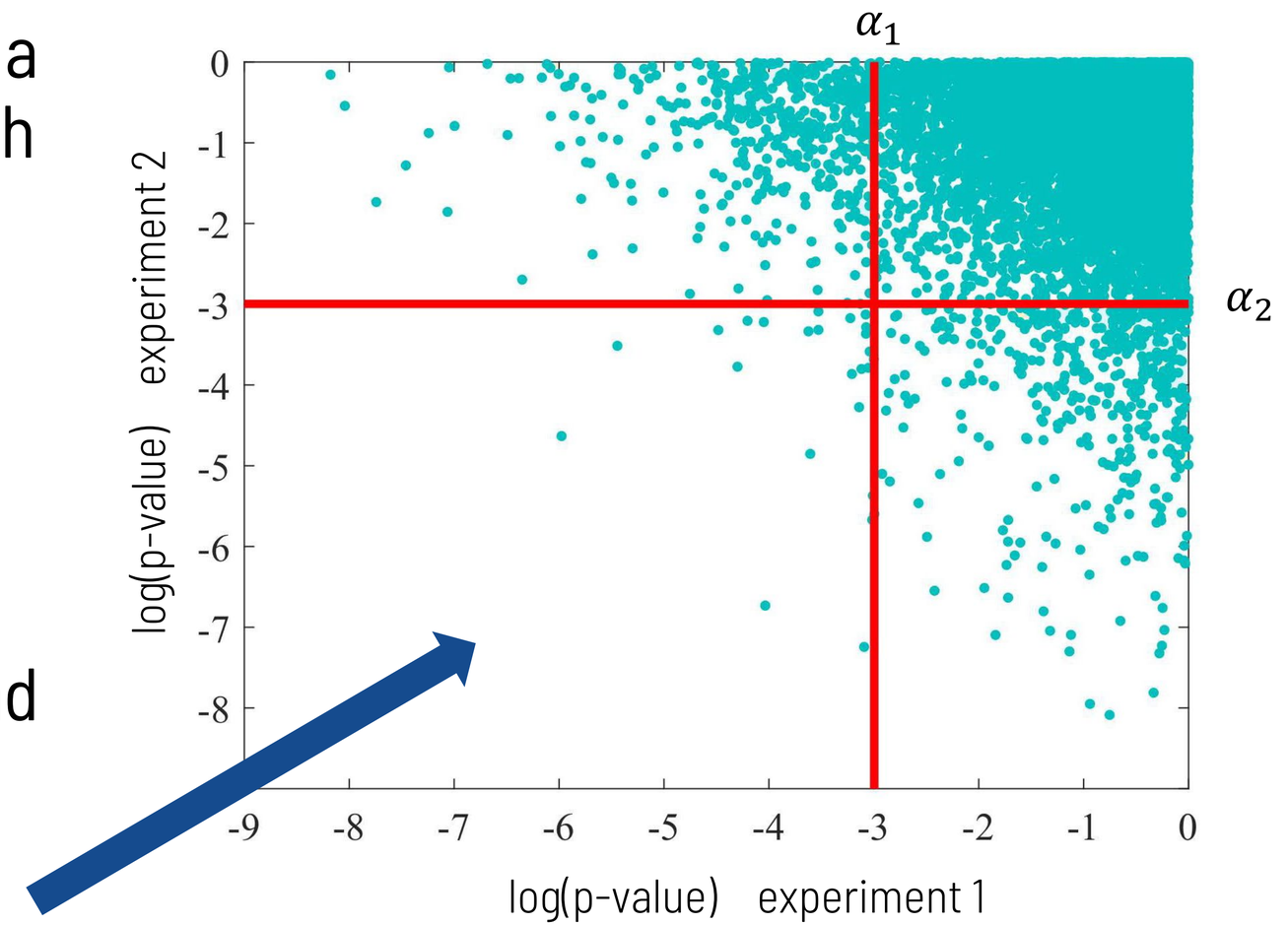
Restricted approach

Select only those results that show a significant signal difference for both experiments at the assumed significance level.

It means we look for test results i that:

$$p_{1,i} \leq \alpha_1 \; and \; p_{2,i} \leq \alpha_2$$

Assuming $\alpha = 0.05$, pair p_1=0.049 and p_2=0.049 is ok, while p_1=0.049 and p_2=0.051 is not ok.



Product approach

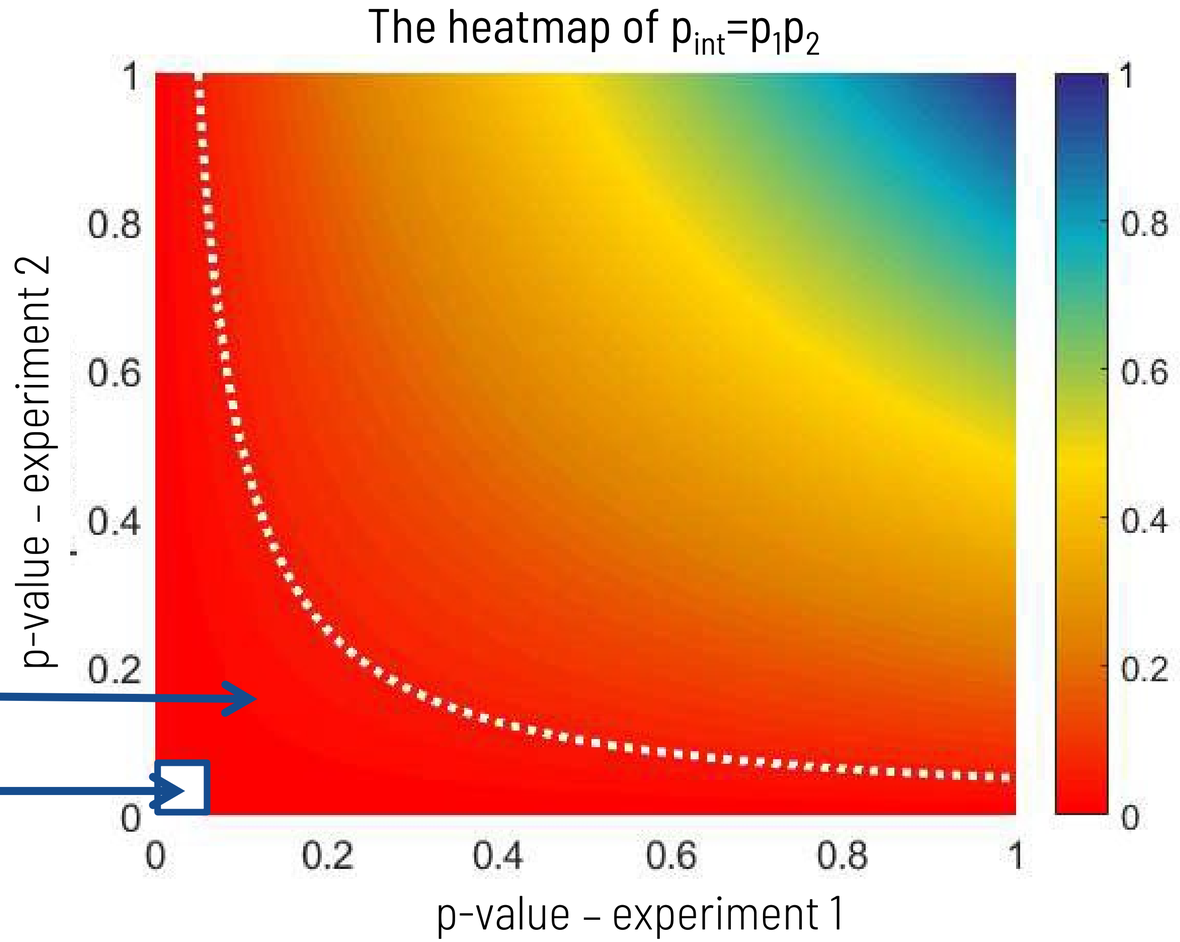
- We have two independently collected data sets (D_1 and D_2).
- For both data, the same trait was measured in a similar way (e.g. expression of the same gene, could be different platform).
- Statistical tests were performed and two p-values (p_1 and p_2) were obtained, one for each trait.
- The product method uses the multiplication of p_1 and p_2 to make the final conclusion.
- What critical value for p_1p_2 should be used for thresholding between validated and non-validated results?

Product approach

$$p_{int} = \prod_{i=1}^{k} p_i$$

Area with $p_{int} \leq 0.05$ (limited by the white dotted line).

The restricted approach



Product approach – an issue

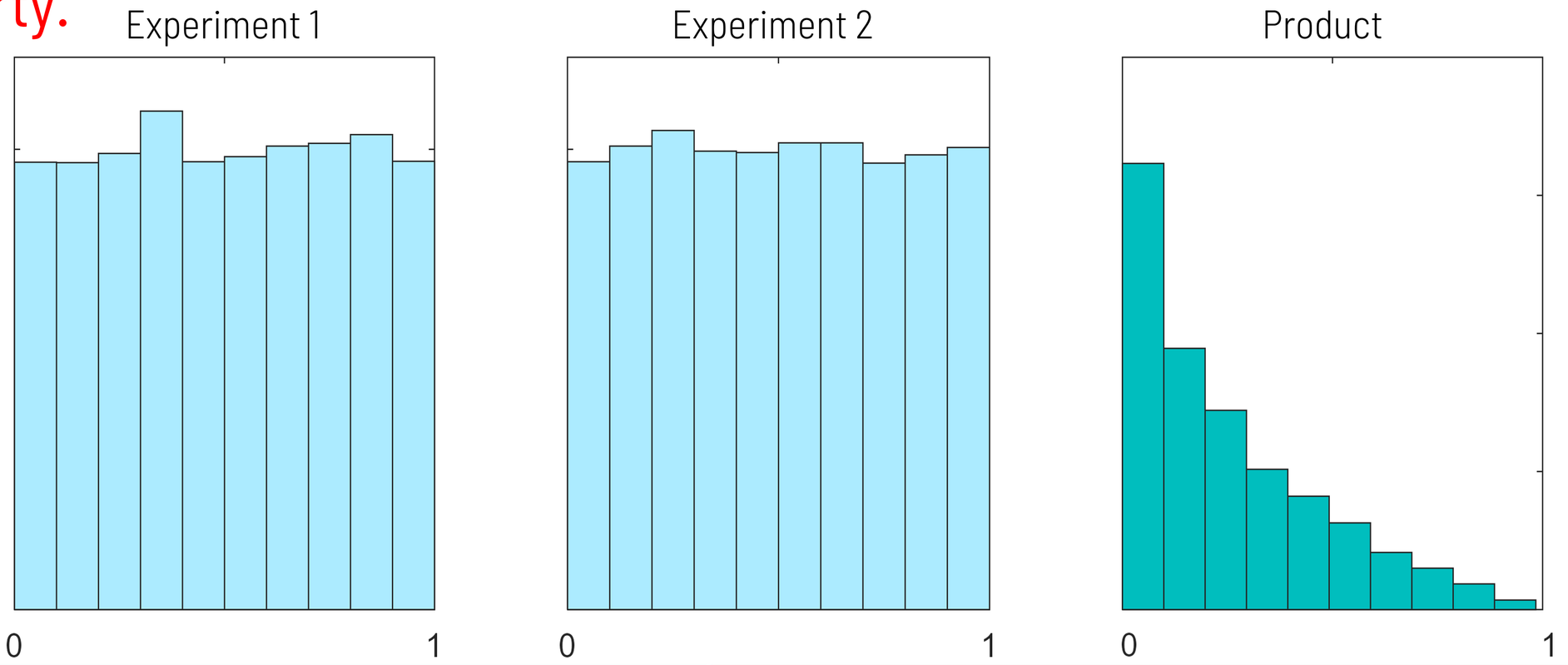
- Imagine you have data from 5 independent experiments. In each of them, a test was performed to investigate the same hypothesis.
- For each of them, the p-value was 0.5 (not significant at 0.05 significance level).
- What is the product value?

$$0.5^5 = 0.03125 < 0.05$$

- We conclude that direct comparison of the p-value product the the original signifance level is improper and can lead to false discoveries.

Why is that?

- If the null hypothesis is true, then the p-value can be described by a uniform distribution over the interval [0,1]. The product of p-values has no longer this property.



How to solve the problem?

Property 1: The logarithm of a uniformly distributed random variable follows the exponential distribution with λ=1.

Property 2: The sum of a number of values of χ^2 is itself distributed in the χ^2 distribution with the appropriate number of degrees of freedom.

Property 3: The exponential distribution is approximately equal to the χ^2 distribution with 2 degrees of freedom (d.f.) when λ=1/2.

Fisher method

We start with calculation of the F statistics

$$F = -2\ln\left(\prod_{i=1}^{k} p_i\right) = -2\sum_{i=1}^{k} ln(p_i) \sim \chi^2_{(2k)}$$

and then the p-value is obtained from the proper χ^2 distribution.

Fisher method – our example

- Reminder – we have data from 5 independent experiments, in each of them, a test was performed to investigate the same hypothesis.
- For each of them, the p-value was 0.5 (not significant at 0.05 significance level). What is the F-value?

$$F = -2\ln\left(\prod_{i=1}^{5} 0.5\right) = -2\sum_{i=1}^{5} ln(0.5) = 6.93 \sim \chi^2_{(10\ d.f.)}$$

$$p = 1 - chi2cdf(6.93,10) = 0.7320$$

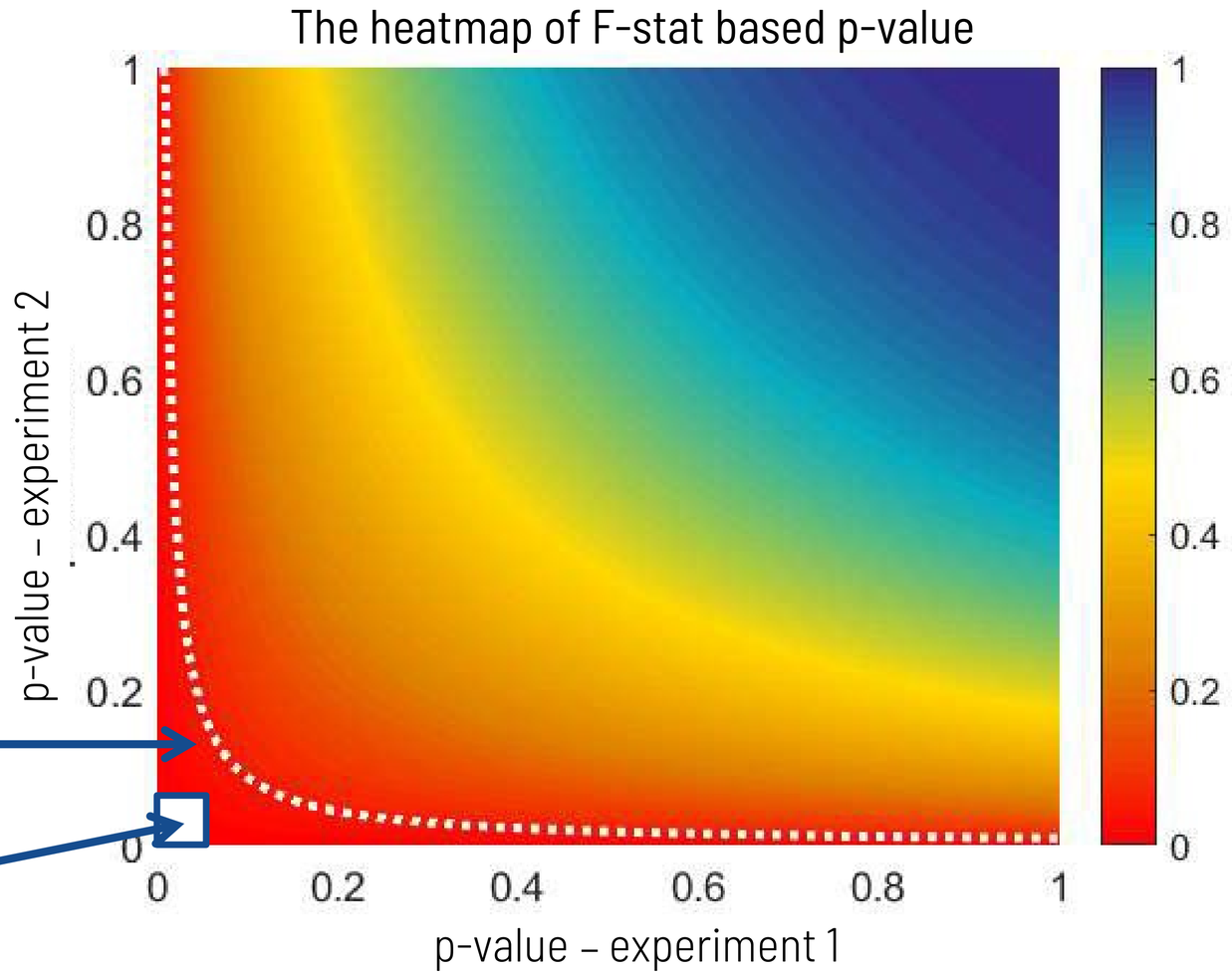
- It can be concluded that the Fisher method represents a more appropriate approach for direct integration of the p-value than the product approach..

Fisher method

$$F = -2\ln\left(\prod_{i=1}^{k} p_i\right) = -2\sum_{i=1}^{k} ln(p_i) \sim \chi^2_{(2k)}$$

Area with F-stat based integrated p-value below 0.05 (limited by the white dotted line)

The restricted approach



Limitation

- It would be inadvisable to employ Fisher's method in instances where the p-values in question exhibit a significant degree of disparity.

$$p_1 = 0.01 \quad p_2 = 0.01 \Rightarrow \quad p = 0.001$$

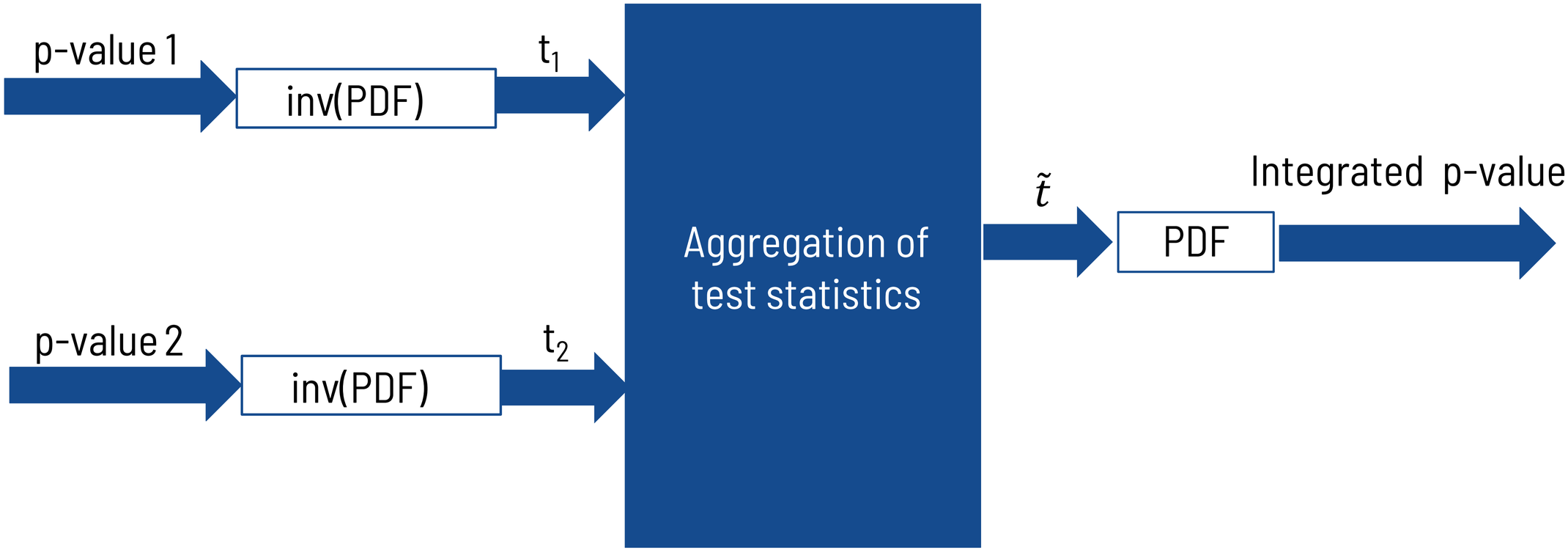
$$p_1 = 0.99 \quad p_2 = 0.99 \Rightarrow \quad p = 0.9998$$

$$p_1 = 0.001 \quad p_2 = 0.999 \Rightarrow \quad p = 0.0079$$

Transformation-based approach

KIAED
KATEDRA INŻYNIERII I ANALIZY EKSPLORACYJNEJ DANYCH

General idea



Stouffer method

It assumes test statistics follow approximatelly Gaussian distribution.

First step: $p_i \rightarrow z_i = \phi^{-1}(p_i)$

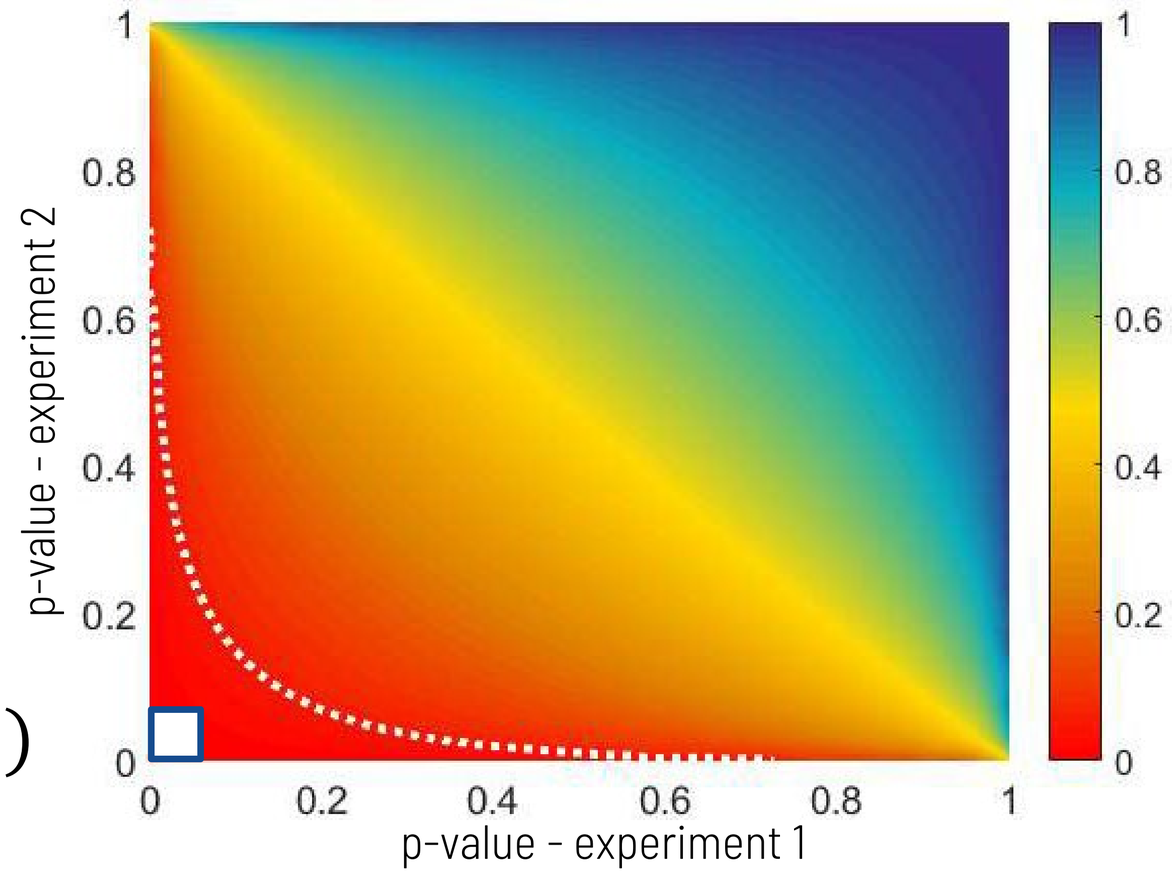
Second step: $Z = \frac{\sum_{i=1}^{k} z_i}{\sqrt{k}} \sim N(0,1)$

Third step for one side test

$$\text{integrated p} = \Phi^{-1}(Z)$$

For two side tests use $abs(z_i)$ instead of z_i and the following formula for p

$$\text{integrated p} = 2 \cdot (1 - \Phi^{-1}(Z))$$



Stouffer, S., DeVinney, L. & Suchmen, E. 1949. The American soldier: Adjustment during army life, vol. 1. Princeton University Press, US.

Weighted Z-transform

- At the stage of Z-statistic aggregation, the weighted average is employed in lieu of the conventional arithmetic mean value.

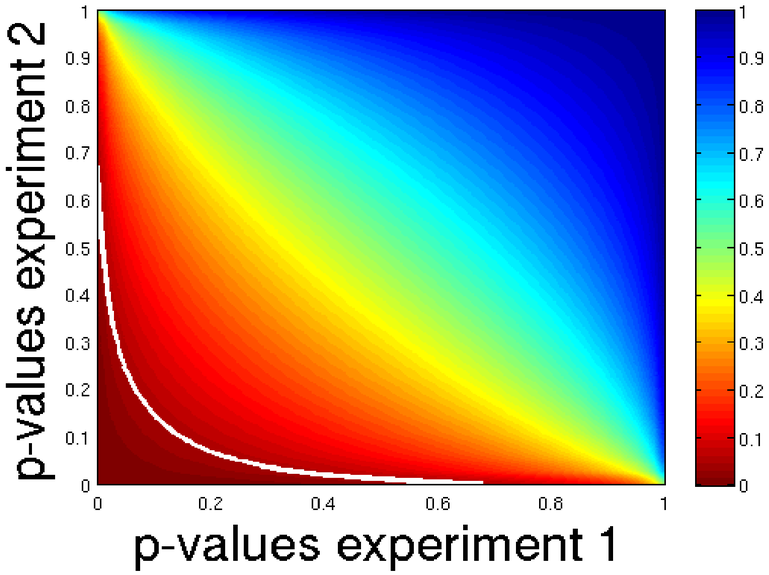
$$Z = \frac{\sum_{i=1}^{k} w_i \cdot z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}}$$

- There are a number of possible approaches to setting the weights. If each weight is equal to one, the method is that of Stouffer.
- Liptak conducted a comprehensive comparative analysis of the various weighting techniques and recommends the particular approaches:

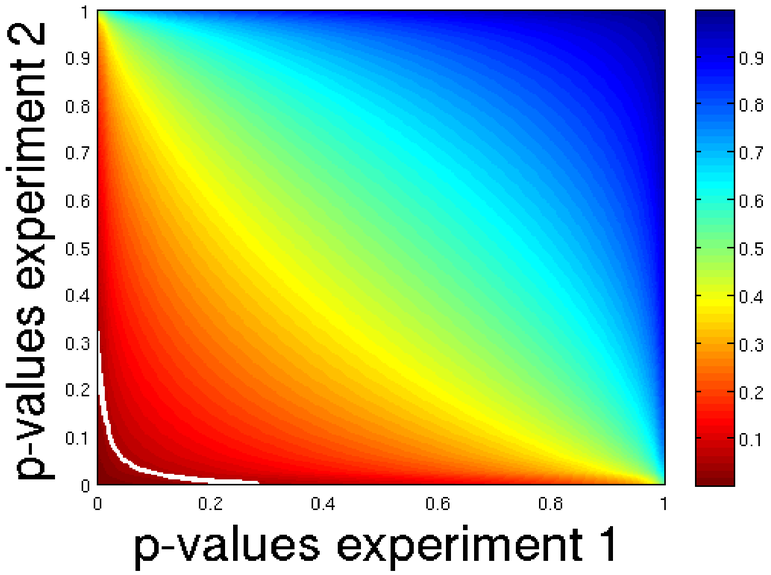
$$w_i = \sqrt{n_i} \quad \text{or} \quad w_i = \frac{1}{SE_i} = \frac{\sqrt{n_i}}{s_i}$$

Liptak, T. (1958). *On the combination of independent tests.* Magyar Tud. Akad. Mat. Kutato Int. Kozl. 3: 171-197.

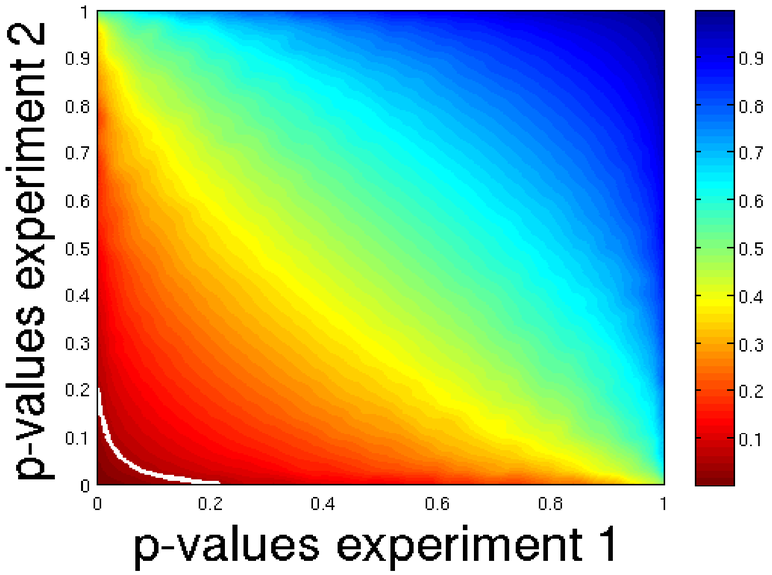
Weighted z-transformation



$$w = 1$$



$$w = \sqrt{n}$$



$$w = \frac{1}{SE_i} = \frac{\sqrt{n_i}}{s_i}$$

DEPENDENT VALIDATION

Problem formulation

KIAED
KATEDRA INŻYNIERII I ANALIZY EKSPLORACYJNEJ DANYCH

- The p-value integration methods are not solely employed for the consolidation of findings from primary and validation experiments.
- In many instances, the secondary experiment is conducted using the same dataset but with a different variable being measured. This could be the abundance of a particular protein fraction in MS/MS analyses, the expression of an another gene in transcriptomics, and so forth.
- It is frequently the case that these features are highly correlated (for example, as a result of the involvement of the same signalling pathway or the manifestation of the same protein). Consequently, the test results are not independent.
- It is not possible to employ the aforementioned methods, as this would result in a biased outcome.

Some exemplary solutions

- Brown MB, *A method for combining non-independent, one-sided tests of significance.* Biometrics, 1975, 31(4):987-992.
- Kost JT, McDermott MP, *Combining dependent p-values.* Statistics & Probability Letters 2002, 60(2):183-190
- Poole W, Gibbs DL, Shmulevich I, Bernard B, Knijnenburg TA, *Combining dependent P-values with an empirical adaptation of Brown's method,* Bioinformatics, 2016, 32(17):i430–i436

Take home message

1) When performing multiple tests, it is essential to choose a less conservative method, particularly when there are a large number of tests involved. Do not confuse the issues of multiple testing (HTS data) with those of multiple comparisons (ANOVA).

2) The p-value causes problems for both too small and too large sample sizes. The p-value is useful, but it should be accompanied by the estimator of the effect size. Always look for the proper effect size measure to fit into your experimental design.

3) Data integration at the p-value level is an effective way to validate experiment results. It increases the power of statistical inference by indirectly increasing the population sample size.

joanna.polanska@polsl.pl