



# Effect estimation versus hypothesis testing

PD Dr. C. Schindler

Swiss Tropical and Public Health Institute

University of Basel

[christian.schindler@unibas.ch](mailto:christian.schindler@unibas.ch)

Annual meeting of the Swiss Societies of Neurophysiology,  
Neurology and Stroke, Lucerne, May 19<sup>th</sup> 2011

# Contents

Effect estimation (effect estimates and 95%-confidence intervals)

Hypothesis testing

95%-confidence intervals and statistical significance

Statistical errors and statistical power

Confirmatory and exploratory analyses

Multiple hypotheses and multiple testing

Beyond individual studies

**Effect estimation :**

**Effect estimates with  
95%-confidence intervals**

## Example

In a random sample of 100 patients treated with medication M, 20 subjects experienced side effects within two weeks of treatment. The observed proportion of patients with side-effects thus equalled 20% in this random sample.

The associated 95%-confidence interval is (12%, 28%)

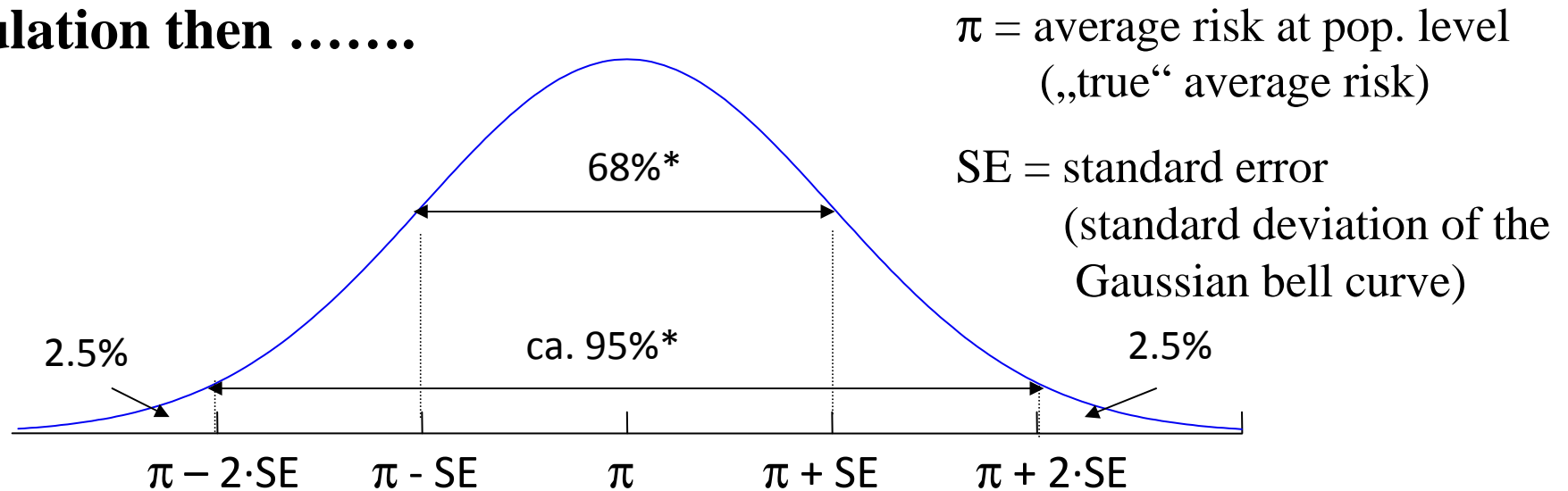
## Interpretation

We can be 95% confident that

the true average risk of patients to experience side effects within 2 weeks of taking medication M

is covered by this interval.

**If we were to draw a large number of random samples of the same size from the same population then .....**



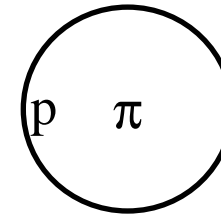
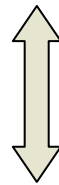
\* of all sample estimates  $p$  of  $\pi$

we would find:

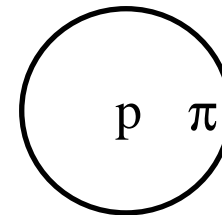
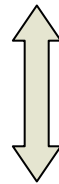
- a) Scatter of the observed proportions  $p$  of subjects with side-effects around the true average risk  $\pi$  = almost symmetrical.
- b) Frequency distribution of the observed values  $p$  close to a Gaussian bell curve with mean  $\pi$  and standard deviation SE.

## Equivalent formulations:

The observed value of  $p$  would be within 2 SE of the true average risk  $\pi$  in 95% of all samples



The true average risk  $\pi$  would be within 2 SE of the observed value  $p$  in 95% of all samples



The true average risk  $\pi$  would be in the interval  $(p - 2 \cdot \text{SE}, p + 2 \cdot \text{SE})$  in 95% of all samples

  
95%-confidence interval

## General Definition of the 95%-confidence interval

$\pi$  = any quantitative parameter of a given population P  
(e.g., the mean or rate of a certain variable),

$p$  = corresponding value observed in a random sample drawn from P  
(e.g., the sample mean or rate of the respective variable).

Then, if the sample size is large, the **95%-confidence interval** of the **parameter estimate**  $p$  may be approximated by

$$(p - 1.96 \cdot SE, p + 1.96 \cdot SE).$$

One can then be about 95%-confident that this interval covers the true value  $\pi$  of the respective parameter in the underlying population.

## A word of caution

For a confidence interval to be valid, the sample should have been drawn randomly.

Non-random samples tend to provide biased estimates of the parameters of the underlying population.

If this is the case, the coverage probability of „95%-confidence intervals“ is generally smaller than 95%.

=> we can then no longer be 95% confident that a given 95%-confidence interval will include the true value of the respective population parameter.

## The standard error

is a measure of the uncertainty inherent with estimating the true and unknown population parameter  $\pi$  by the corresponding value  $p$  observed in a given random sample.

General law: With few exceptions, the statistical uncertainty decreases with increasing sample size  $n$ ,

and is proportional to  $\frac{1}{\sqrt{n}}$  (square root of n-law)

## Simple standard error formulas

A) Standard error of a sample mean:

$$SE = \frac{\sigma}{\sqrt{n}}$$

where  $\sigma$  denotes the standard deviation of the respective variable in the underlying population. (Since  $\sigma$  is generally unknown, it must be replaced by the sample standard deviation  $s$ ).

B) Standard error of a sample rate  $p$  (expressed as fraction of 1):

$$SE = \frac{\sqrt{p \cdot (1-p)}}{\sqrt{n}}$$

## Back to introductory example

$$p = 0.2 \quad \text{and} \quad n = 100$$

$$\text{Thus,} \quad SE = \frac{\sqrt{0.2 \cdot (1 - 0.2)}}{\sqrt{100}} = 0.04$$

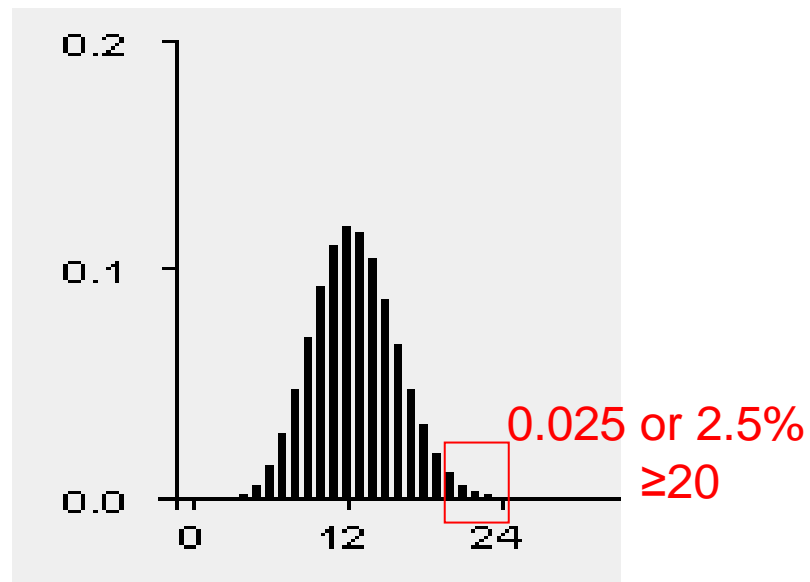
What if we only accepted a standard error of 0.02  
(i.e., half the present uncertainty) –  
How much larger would the sample size have to be?

Since the sample size  $n$  appears under the root, the sample size  $n$  would have to be 4 times as large, i.e., we would need  $n = 400$ .

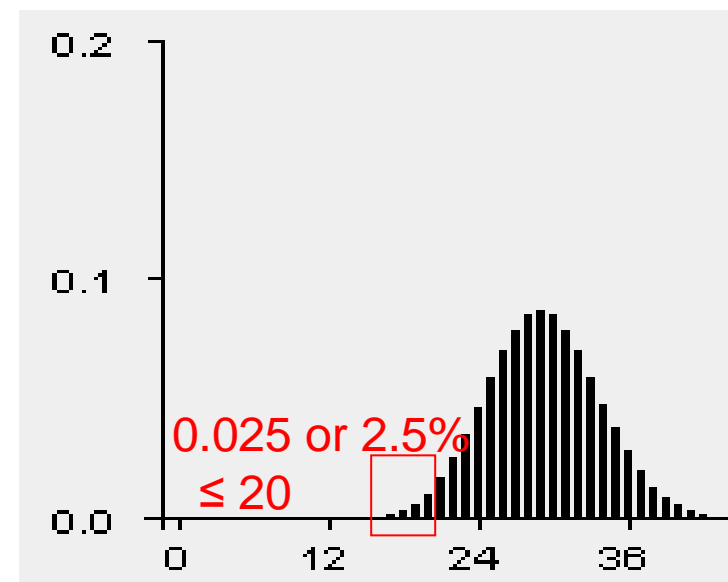
## Exact confidence interval for a proportion

The previously calculated confidence interval is only an approximation to the exact one which would be (12.67% to 29.18%)

Distribution of observed number of patients with side effects if  $n=100$  and  $\pi = 0.1267$



Distribution of observed number of patients with side effects if  $n = 100$  and  $\pi = 0.2918$



## General form of standard errors

$$SE = \frac{\text{term depending on distribution of the data but not on } n}{\sqrt{n}}$$

## Log-transformation of comparison measures

Sometimes, the computation of a standard error makes sense only for the logarithm of a comparison measure.

For instance, standard errors are not computed for odds ratios (OR) and relative risks (RR) but for their natural logarithms  $\ln(\text{OR})$  and  $\ln(\text{RR})$ .

(The distribution of sample OR's and RR's under the null hypothesis is skewed while the distribution of  $\ln(\text{OR})$  and  $\ln(\text{RR})$  is close to normal already for relative small sample sizes.)

## 95%-confidence intervals (approximative)

	with outcome	w/o outcome	
exposed	a	b	a+b
unexposed	c	d	c+d
	a+c	b+d	

$$RD \pm 1.96 \sqrt{\frac{ab}{(a+b)^3} + \frac{cd}{(c+d)^3}}$$

$$RR \cdot e^{\pm 1.96 \sqrt{\frac{1}{a} - \frac{1}{(a+b)} + \frac{1}{c} - \frac{1}{(c+d)}}}$$

$$OR \cdot e^{\pm 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$$

# Hypothesis testing

## Comparison of the side effects of two otherwise equivalent medications

	with side effects	without side effects	
medication A	48 (48%)	52	100
medication B	32 (32%)	68	100
Difference	16 (16%)	-16	200

Frequency difference = 16% (95%-confidence interval: 3% to 29%)

## Thought experiment:

We assume that the two medications are not only equivalent with respect to their main effect but do not differ either with respect to the proportion of patients in whom they cause side effects (-> Null hypothesis  $H_0$ ).

If this assumption (i.e.,  $H_0$ ) were true, then the observed difference would have no explanation other than chance.

With which probability would we then expect a difference of at least the same size in a replication study of the exact same design?

This probability is called the the **p-value** of the observed difference. (In our example, the p-value equals 0.03 or 3%.)

## Decision on the statistical significance of an observed difference / effect / association

The observed difference / effect / association is said to be **statistically significant at level  $\alpha$** , if its p-value is smaller than  $\alpha$  ( $p < \alpha$  in short).

The number  $\alpha$  is referred to as **significance level** and must be defined in advance, i.e., in the study protocol.

The usual choice for  $\alpha$  is 0.05 or 5%.

Thus, in our example, the observed difference is statistically significant at the 0.05-level since  $p = 0.03$ .

## Decision against the null hypothesis

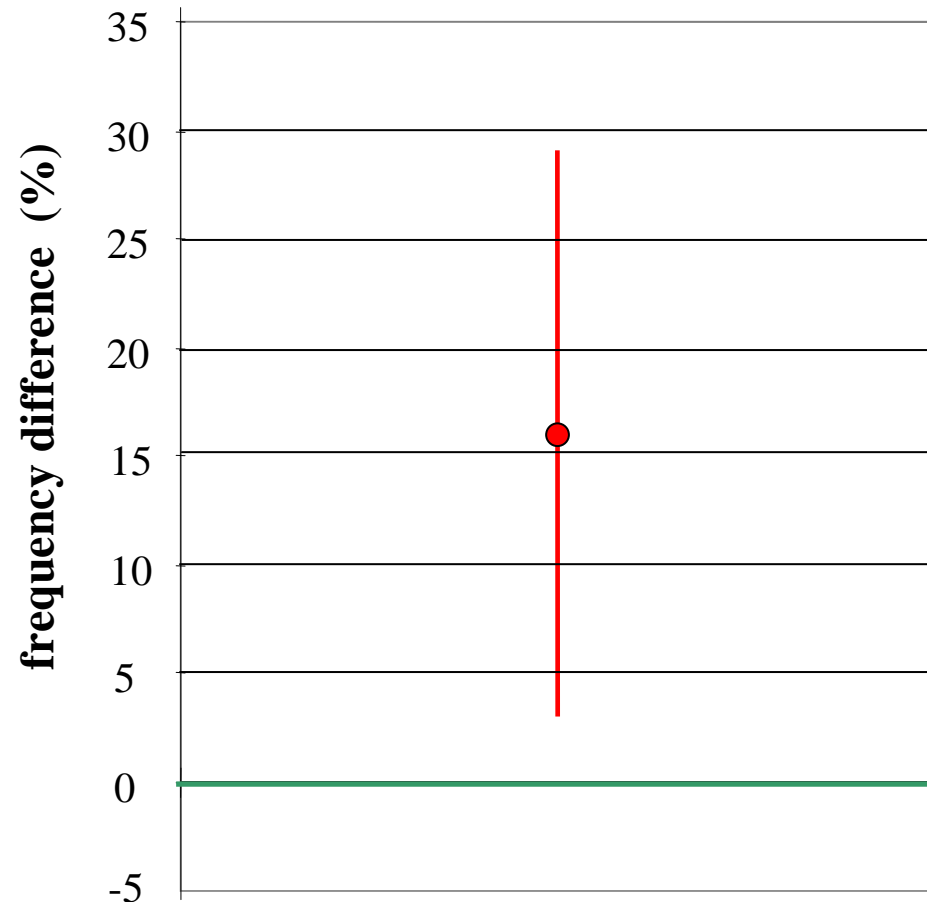
If the observed difference / effect / association is statistically significant at the previously agreed  $\alpha$ -level, then the null hypothesis is rejected.

Otherwise, we cannot decide against the null hypothesis (This is not the same as deciding for the null hypothesis!)

*„Absence of evidence  $\neq$  Evidence of absence“  
(e.g., of an effect)*

# **95%-confidence interval and statistical significance**

## Graphical representation of the frequency difference and of its 95%-confidence interval.



The 95%-confidence interval of our observed difference does not include the value 0, which we would expect to observe on average if the null hypothesis were true.

Therefore, we can conclude that the difference is statistically significant at the level of 0.05.

← Null hypothesis  
(„in reality, no difference“)

**Table 2.** Mean baseline levels and mean change from baseline to the end of supplementation in serum concentrations of ascorbic acid and uric acid levels, and estimated GFR by supplementation group\*

Measure	Supplementation group		Unadjusted difference in mean change	Adjusted difference in mean change†
	Placebo	Vitamin C		
Serum ascorbic acid, $\mu$ moles/liter				
Baseline, mean $\pm$ SD	60.0 $\pm$ 16.7	64.4 $\pm$ 14.1		
Mean change (95% CI)	1.2 (-2.1, 4.5)	21.3 (14.7, 27.9)	20.2 (13.0, 27.4)‡	21.5 (14.8, 28.2)‡
Serum uric acid, mg/dl				
Baseline, mean $\pm$ SD	5.1 $\pm$ 1.5	5.2 $\pm$ 1.4		
Mean change (95% CI)	0.09 (-0.05, 0.2)	-0.5 (-0.6, -0.3)	-0.6 (-0.8, -0.4)‡	-0.5 (-0.7, -0.3)‡
GFR, ml/min/1.73 m <sup>2</sup> §				
Baseline, mean $\pm$ SD	78.1 $\pm$ 17.1	74.7 $\pm$ 15.0		
Mean change (95% CI)	0.4 (-2.6, 3.5)	4.8 (1.5, 8.0)	4.0 (0.5, 7.4)¶	3.5 (0.3, 6.8)#

\* GFR = glomerular filtration rate; 95% CI = 95% confidence interval.

† Mean change in serum ascorbic acid was adjusted for age, sex, and baseline ascorbic acid levels. Mean change in serum uric acid was adjusted for age, sex, baseline ascorbic acid levels, and baseline uric acid levels. Mean change in GFR was adjusted for age, sex, race, baseline ascorbic acid, and estimated GFR.

‡  $P < 0.0001$ .

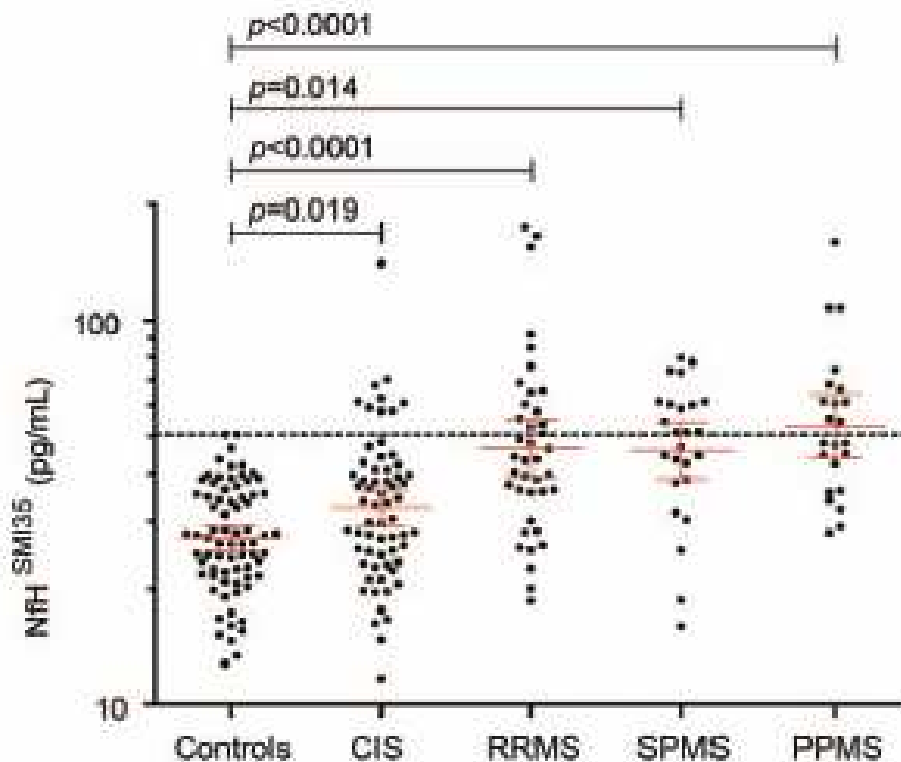
§ Estimated according to the modified Modification of Diet in Renal Disease equation.

¶  $P = 0.02$ .

#  $P = 0.03$ .

**95%-CI's do not include 0**

**Figure 1** NfH<sup>SMI35</sup> levels in the controls, patients with clinically isolated syndrome (CIS), and patients with multiple sclerosis (MS)

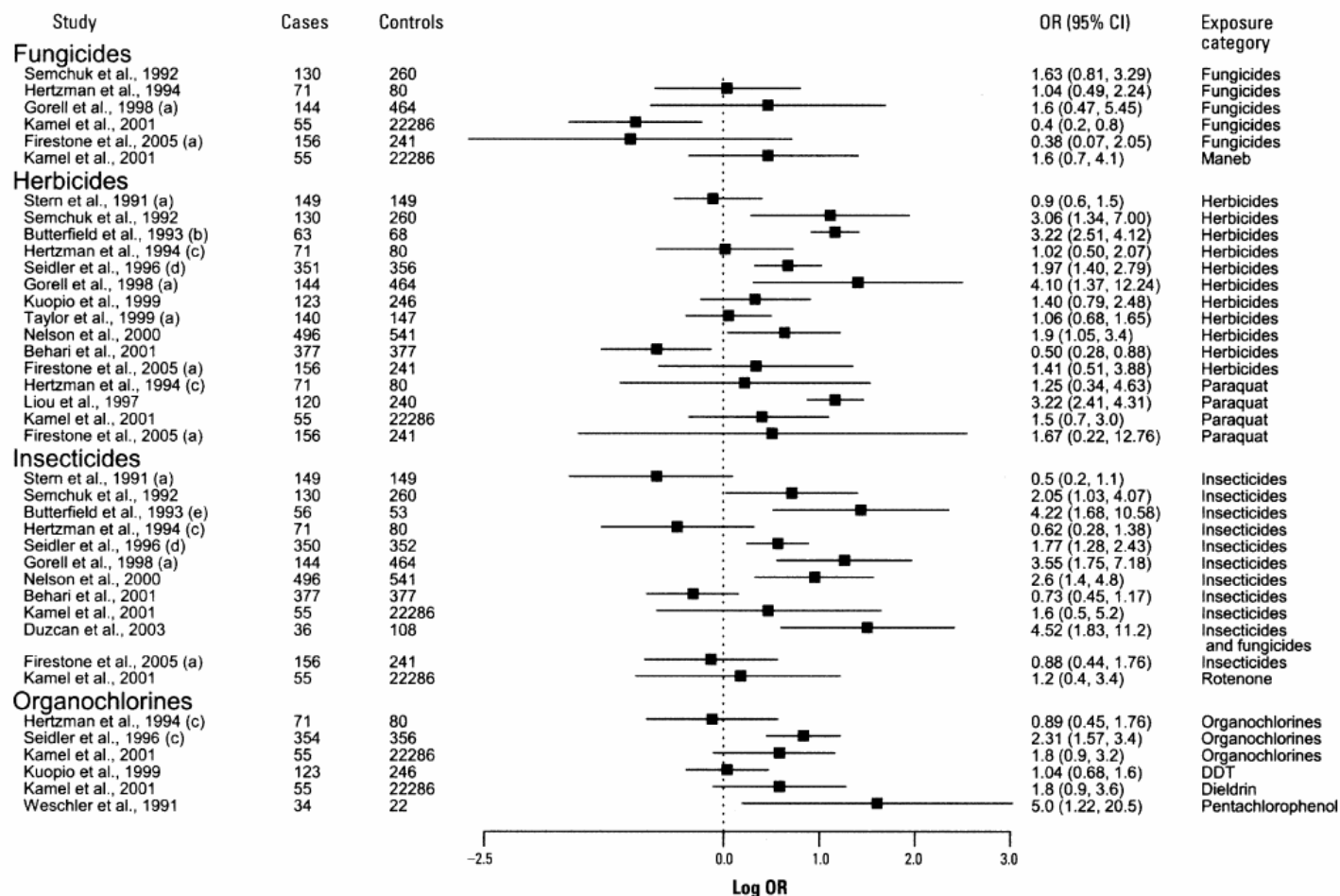


Geometric mean and 95% confidence interval are displayed. CIS (32.4 pg/mL) relapsing-remitting MS (RRMS) (46.4 pg/mL) secondary progressive MS (SPMS) (45.9 pg/mL) and primary progressive MS (PPMS) (53.0 pg/mL) showed higher CSF NfH<sup>SMI35</sup> levels than controls (27.1 pg/mL) and RRMS higher values than CIS. Dots represent individual samples. The horizontal dotted line represents the cutoff value of 50.5 pg/mL (highest value observed in the control group). p Values are adjusted for age and corrected by Bonferroni method.

Levels of neurofilament heavy chain protein in cerebrospinal fluid in 4 groups of MS-patients and in healthy controls

CIS = clin. isolated syndrome  
 RRMS = relapsing / remitting MS  
 SPMS = 2ndary progressive MS  
 PPMS = primary progressive MS

# Review of studies on Parkinson's disease and exposure to pesticides



If the 95%-confidence interval of  $\ln(\text{OR})$  is  $> 0$ , only two possibilities exist:

- the true underlying OR was indeed larger than 1.
- the true underlying OR was  $\leq 1$  nonetheless. But then this was one of the 5% „bad“ cases in which the true OR was missed by the 95%-confidence interval.

## Second interpretation of the 95%-CI

A)

95%-confidence interval of the observed difference / effect / association does not contain hypothesized reference value.



Observed difference / effect / association is significantly different from the hypothesized reference value at the 5%-level.  
=> Hypothesis can be rejected (accepting an error probability of 5%).

B)

95%-confidence interval of the observed difference / effect / association contains hypothesized reference value.



Observed difference / effect / association is not significantly different from the hypothesized reference value at the 5%-level.  
=> Hypothesis cannot be rejected (with an accepted error probability of 5%).

## Confidence intervals are more informative than p-values

	95%-confidence interval	p-value
describes statistical uncertainty of the observed D/E/A* explicitly.	+	-
tells whether observed D/E/A* is statistically significant at 5%-level	+	+
provides direct information on the level of significance of the observed D/E/A*	-	+
informs about the relevance of an observed D/E/A*	+	-
allows direct comparison with results from other studies (meta-analysis)	+	-

# **Statistical errors and statistical power**

## Type I error

If the null hypothesis is rejected, this could be a false decision.

The probability of falsely rejecting the null hypothesis equals the significance level  $\alpha$  (type-I-error)

The type-I-error is thus entirely under the control of the scientist who sets the  $\alpha$ -level.

## Type II error

Conversely, if the null hypothesis can not be rejected this doesn't mean that it is correct.

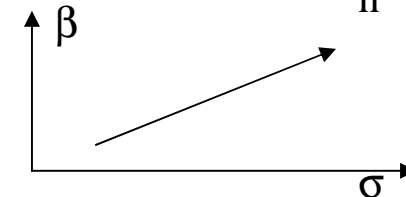
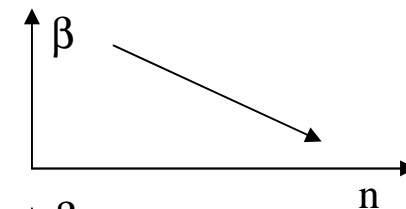
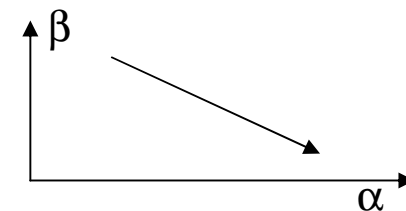
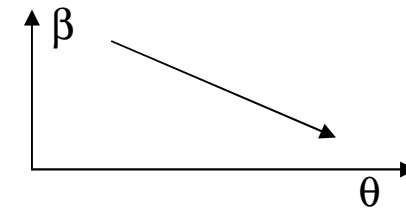
There are two alternative possibilities:

- a) The true effect is smaller than was expected when the sample size was calculated.
- b) The true effect is of the expected size but chance produced a small effect estimate failing to become statistically significant (type II error).

## Determinants of type II error

The probability  $\beta$  of a type II error can only be determined for specific effect sizes. It depends on 4 factors:

- a) The size of the true effect  $\theta$
- b) The significance level  $\alpha$
- c) The sample size  $n$
- d) The variability  $\sigma$  of the data



## Type II error and statistical power

The statistical power is the probability of not committing a type II error.

$$\text{Power} = 1 - \beta.$$

Generally, when designing a study, a power of 80% or 90% is aimed at. The required sample size is then computed under the assumption of a certain true effect  $\theta$  considered both realistic and relevant. Since  $\alpha = 0.05$  is a standard, and the variability of the data can not be greatly influenced in general, the power must essentially be tuned via the sample size.

# **Confirmatory and exploratory analyses**

## Confirmatory analyses

Confirmatory analyses are meant to result in statistical conclusions, i.e. in the rejection or non-rejection of one or more hypotheses formulated in the study protocol.

(Statements on the statistical significance of a result have – albeit not explicitly – the character of statistical conclusions.)

Confirmatory analyses should be backed up by a prior power calculation guaranteeing that, if one's expectations on the true difference / effect / association are correct, one can be 80% or 90% certain to get the hypothesis confirmed by a statistically significant result.

## Exploratory analyses

Results from exploratory analyses are primarily meant to foster scientific thoughts and not to lead to immediate statistical decisions.

These results should be reported as effect estimates with 95%-confidence intervals.

Exploratory analyses help to generate new hypotheses which can then be rigorously tested in subsequent studies.

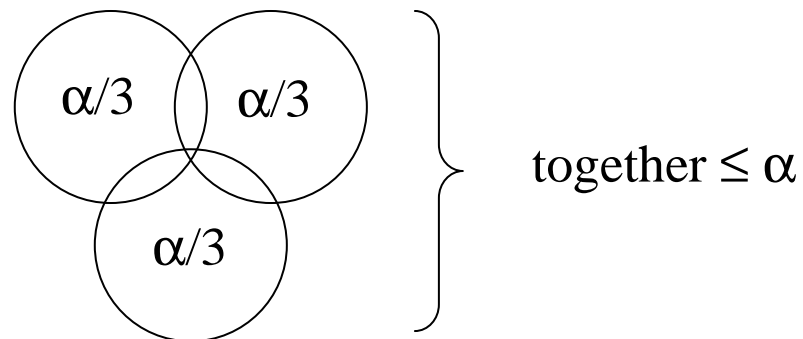
If one wants to jump to statistical decisions immediately, strict rules must be observed.

# **Multiple hypotheses and multiple testing**

## Adjusting for multiple testing

If three true null hypotheses are simultaneously tested, then the probability that at least one of the three tests will produce a p-value  $< 0.05$  may be as large as 0.15.

Modified decision rule: The joint null hypothesis is only rejected if at least one of the three p-values is smaller than  $\alpha = 0.05/3 = 0.0167$  (Bonferroni-correction of the  $\alpha$ -level).

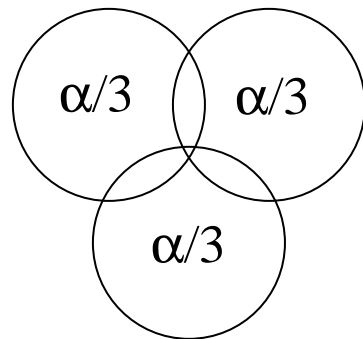


This guarantees that the joint null hypothesis is falsely rejected with a probability of no more than  $3 \cdot 0.0167 = 0.05$ .

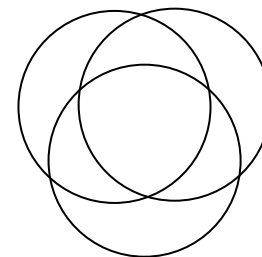
## Disadvantage of Bonferroni adjustment

Bonferroni-correction is very conservative if the parallel comparisons are „correlated“.

For instance, if the effect of an MS-treatment on EDSS and different latency parameters is tested simultaneously, then there is more overlap of the error probabilities than if unrelated outcomes were considered.



Independent tests:  
little overlap of error pr's  
-> larger combined error probability



Dependent („correlated“) tests:  
considerable overlap of error pr's  
-> smaller combined error probability

## **Necessity of adjusting for multiple testing**

Necessary, if one has a global null hypothesis stating that several differences / effects / associations are all 0.

Not necessary, if one has only one primary hypothesis and all secondary hypotheses will be addressed in exploratory analyses.

Not necessary in exploratory analyses unless one aims to step directly from an exploratory result to a statistical conclusion.

## How to do better than with Bonferroni-adjustment?

If more than two disjoint groups (e.g., treatment arms) are to be compared, then there generally exist omnibus tests enabling simultaneous comparison of all groups while testing the global null hypothesis of homogeneity across groups:

Chi<sup>2</sup>-test for qualitative outcomes,

ANOVA (parametric or non-parametric) for quantitative outcomes.

Adjustment of significance level is then only an issue in post-hoc comparisons (e.g., trying to identify pairs of treatments with significantly different effects).

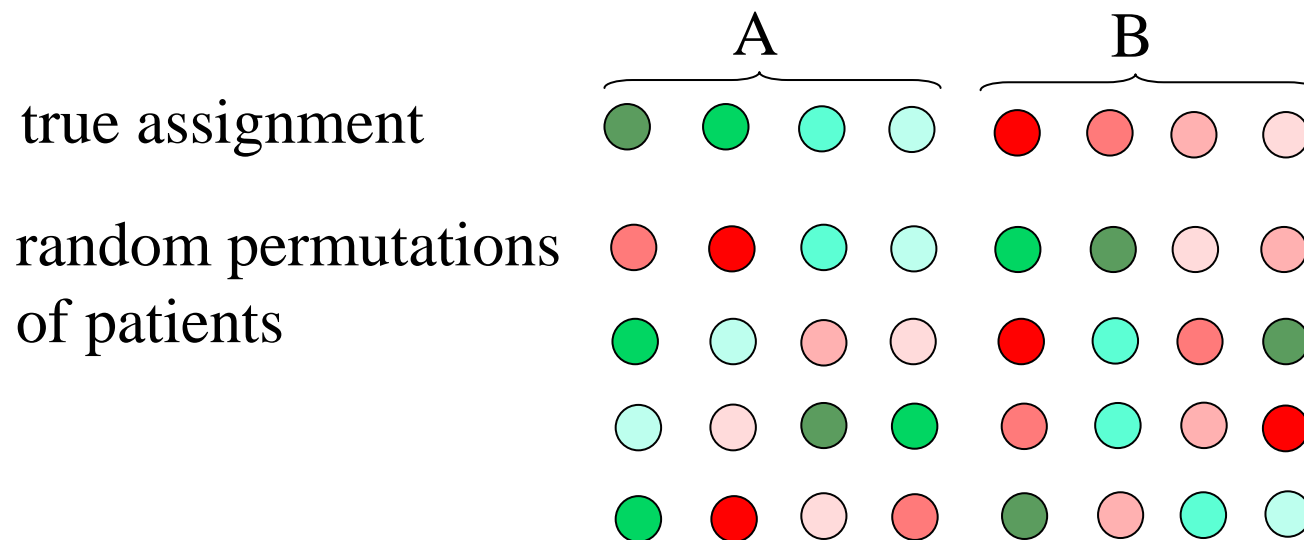
## **How to do better than with Bonferroni-adjustment (cont.)**

If several endpoints are considered simultaneously and one wishes to test the global null hypothesis that none of the endpoints is associated with a given factor of interest,

then a suitable permutation test will generally provide more power for the respective decision (unless the endpoints are uncorrelated).

## Idea of permutation test

If two treatments are completely indistinguishable, one could randomly exchange the treatment labels across patients or the patients across treatment labels without losing any meaningful information.



## How a permutation test is performed

- 1) Generate a large number of such random permutations of the patients across treatment labels
- 2) Repeat each time the comparison between treatment groups and store test results
- 3) Count how often the “new” difference between A and B becomes at least as pronounced as the one originally observed.
- 4) The proportion of such results is the p-value of the permutation test.

## How to do better than with Bonferroni-adjustment (cont.)

If several endpoints are considered simultaneously and one hypothesises that they are all positively influenced by a given intervention, then the sum of standardized effect estimates may be chosen as test statistic (**O'Brien test**).

(This may be combined with a permutation test procedure.)

E.g.,  $z_1 = \text{obs. effect on outcome 1} / \text{SE of effect estimate} = 1.63$  (->  $p = 0.10$ )  
 $z_2 = \text{obs. effect on outcome 2} / \text{SE of effect estimate} = 1.76$  (->  $p = 0.078$ )

combined test score =  $1.63 + 1.76 = 3.39$

If the two test scores  $z_1$  and  $z_2$  have a standard normal distribution under  $H_0$  and are correlated with  $r = 0.4$ , then the combined score has a variance of  $2+2r = 2.8$  and its z-value becomes  $z = 3.39 / \sqrt{2.8} = 3.39 / 1.673 = 2.03$  (->  $p = 0.042$ )

# **Beyond individual studies**

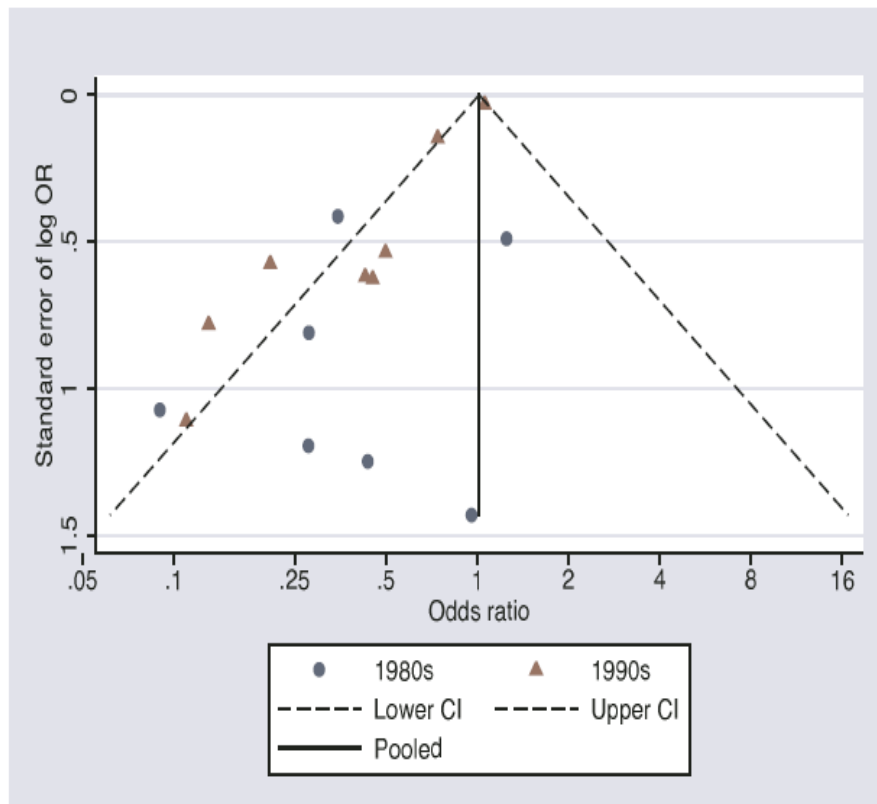
## **Statistical significance and publication bias**

Since significant results can be published more easily and in better journals than non-significant ones and since we all must use our time as efficiently as possible, non-significant results are underreported in the literature.

Thus, new scientific findings tend to be overestimates.

## Meta-analysis, funnel plots

Can Magnesium-therapy reduce mortality risk after MI?



The vertical line is the meta-analytic average of the odds ratios of the different trials. This average is very close to 1 (no effect). It is dominated by one very large trial (ISIS 4) having included almost 60'000 patients.

Smaller studies (higher standard errors) and earlier studies (circles) tended to observe stronger risk reduction.

Without publication bias, the points should scatter more symmetrically around the vertical line.

## Summary

Confidence intervals are more informative than p-values.

Meta-analyses, the main instrument for providing scientific evidence in medicine, are entirely based on confidence intervals.

P-values are important when a decision must be based on one study alone or if a hypothesis cannot be tested using a simple measure (e.g., when comparing more than 2 groups or assessing more than one endpoint or predictor in the same analysis).

Statistical significance does not imply the relevance of a result and vice versa.

The existing culture of judging results mainly by their statistical significance is one of the driving causes of publication bias.

Thank you for your attention  
again!