Experimental Data Analysis
in ©MATLAB

Lecture 6:
Type I & Type II errors, multiple comparisons, sample size estimation

Jan Rusz

Czech Technical University in Prague

Motivation: Type I & Type II Errors

Imagine a situation at the court: Miss Victoria, daughter of a wealthy businessman, has been assassinated. We have only one suspect, a young gardener Geoff.

"Ei incumbit probatio qui dicit, non qui negat"
(The burden of proof is on the one who declares, not on one who denies)

Presumption of innocence = Null Hypothesis (H_0)
Accordingly to *Charter of Fundamental Rights and Freedoms*.

We are looking for evidence that Geoff is guilty (to reject H_0).

Type I error
We did compile too many unreliable evidences of guilt, such as: "We have three witnesses Charlie (drunkard), Joe (drug addict) and Frank (hustler); who have seen Geoff walking into Victoria's house that day" (i.e. fishing). We did increase the chance that court will adjudge Geoff as guilty even when he is maybe innocent.

Type II error
We did compile not enough evidence of guilt, such as: "Geoff did worked all that rainy day in the garden and we have found only some unidentifiable footsteps around the venue" (i.e. small sample size). We did increase the chance that the court will not adjudge Geoff as innocent even when he is likely guilty.
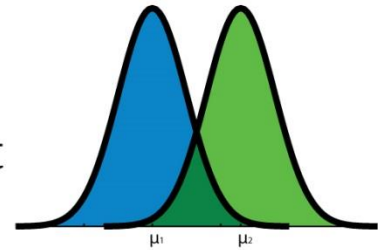
Type I & Type II Errors

Null Hypothesis (H_0): The accused is *not guilty*

Decision about H_0	Null Hypothesis (H_0) is	
	true = not guilty	false = guilty
Reject H_0: guilty	Type I error	Correct
Fail to reject H_0: not guilty	Correct	Type II error

Type I & Type II Errors

Null Hypothesis (H_0): There is no positive effect of treatment

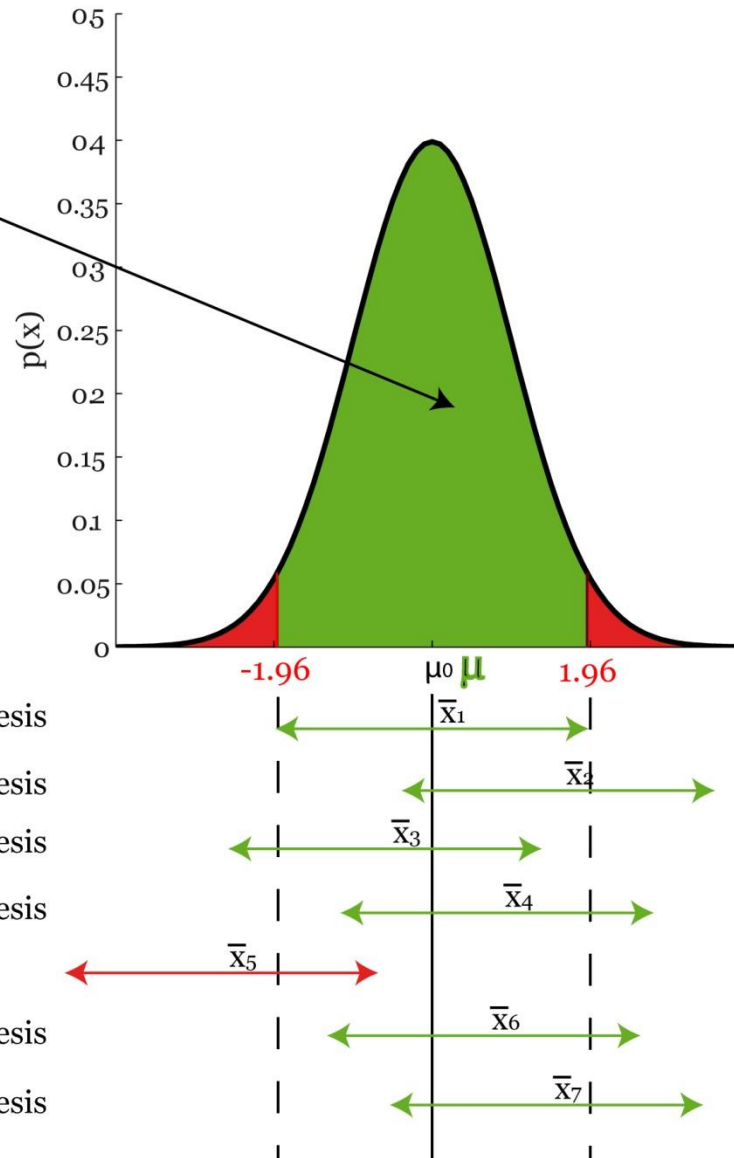|	Null Hypothesis (H_0) is	
Decision about H_0	true = no effect	false = positive effect
Reject H_0: no effect	Type I error	Correct
Fail to reject H_0: positive effect	Correct	Type II error

Type I & Type II Errors

Decision about H_0	Null Hypothesis (H_0) is	
	true	**false**
Reject H_0	Type I error (False Positive)	Correct (True Positive)
Fail to reject H_0	Correct (True Negative)	Type II error (False Negative)

α = 0.05

95% of all sample means ($\bar{x}$) are hypothesized to be in this region.

H_0: μ = $μ_0$
H_a: μ ≠ $μ_0$

p(x)

-1.96 $μ_0$ μ   1.96

Fail to reject null hypothesis $\bar{x}_1$

Fail to reject null hypothesis $\bar{x}_2$

Fail to reject null hypothesis $\bar{x}_3$

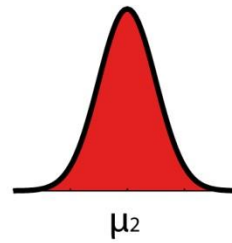Fail to reject null hypothesis $\bar{x}_4$

Reject null hypothesis $\bar{x}_5$

Fail to reject null hypothesis $\bar{x}_6$

Fail to reject null hypothesis $\bar{x}_7$

If we took a sample, and it was by chance like $\bar{x}_5$, we would incorrectly reject the null hypothesis.

Type I error

α here is our level of tolerance for making a Type I error.
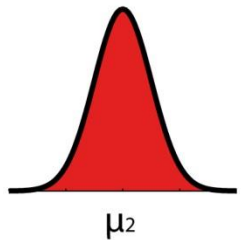
Multiple tests
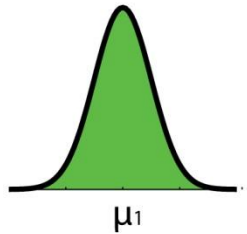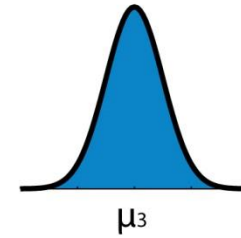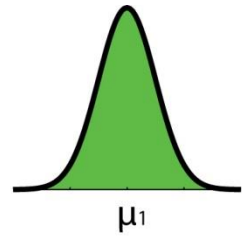
$H_0: \mu_1 = \mu_2; \alpha = 0.05$ $H_0: \mu_1 = \mu_3; \alpha = 0.05$

Pairwise comparison means three tests, each with $\alpha = 0.05$ Type I error rate at 95% confidence.

$H_0: \mu_2 = \mu_3; \alpha = 0.05$

However, error COMPOUNDS with each test:
$(0.95)*(0.95)*(0.95) = 0.857$

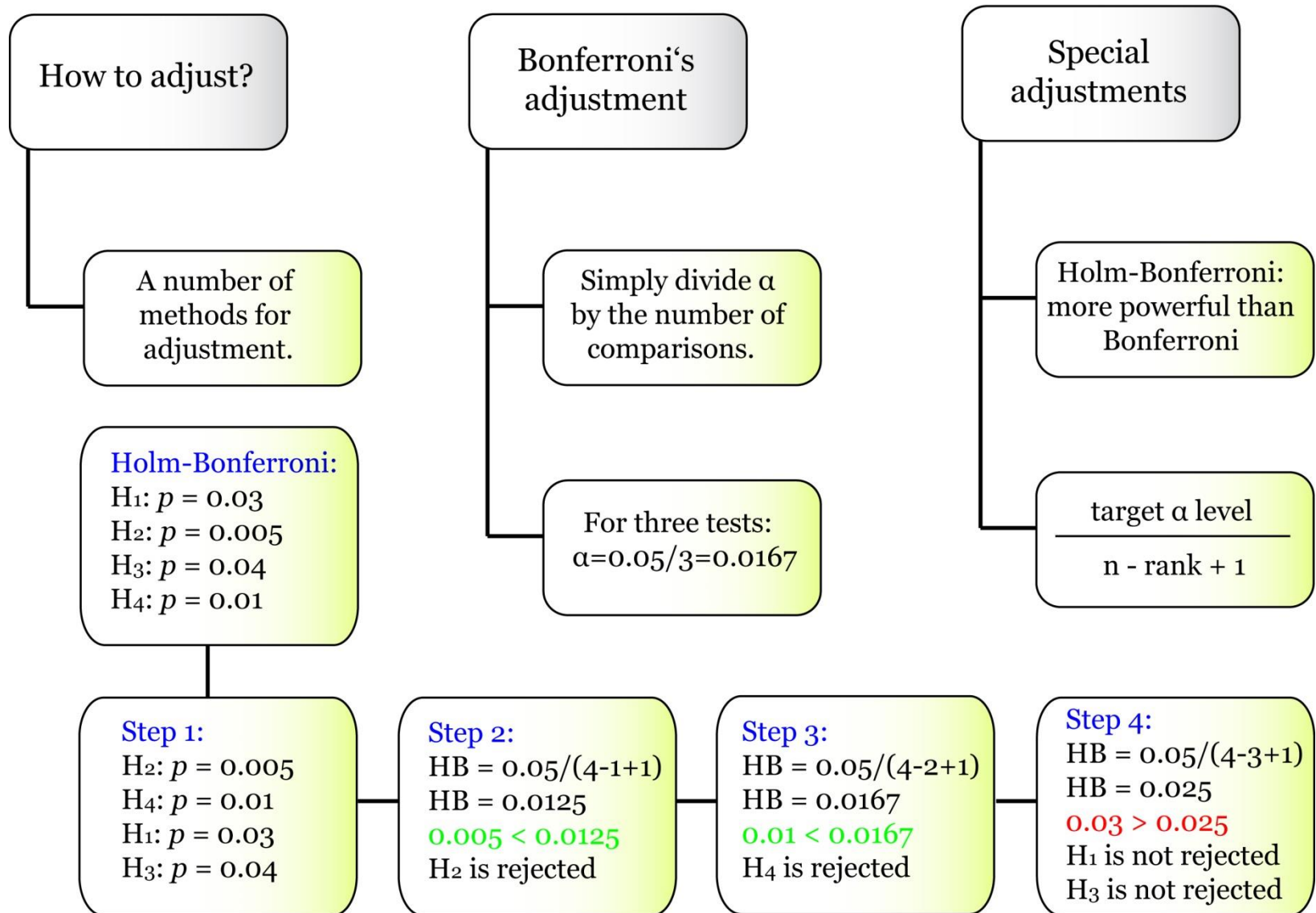$\alpha = 1 - 0.857 = 0.143$!
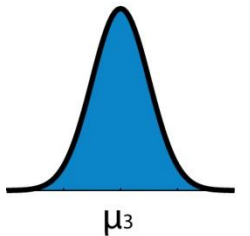
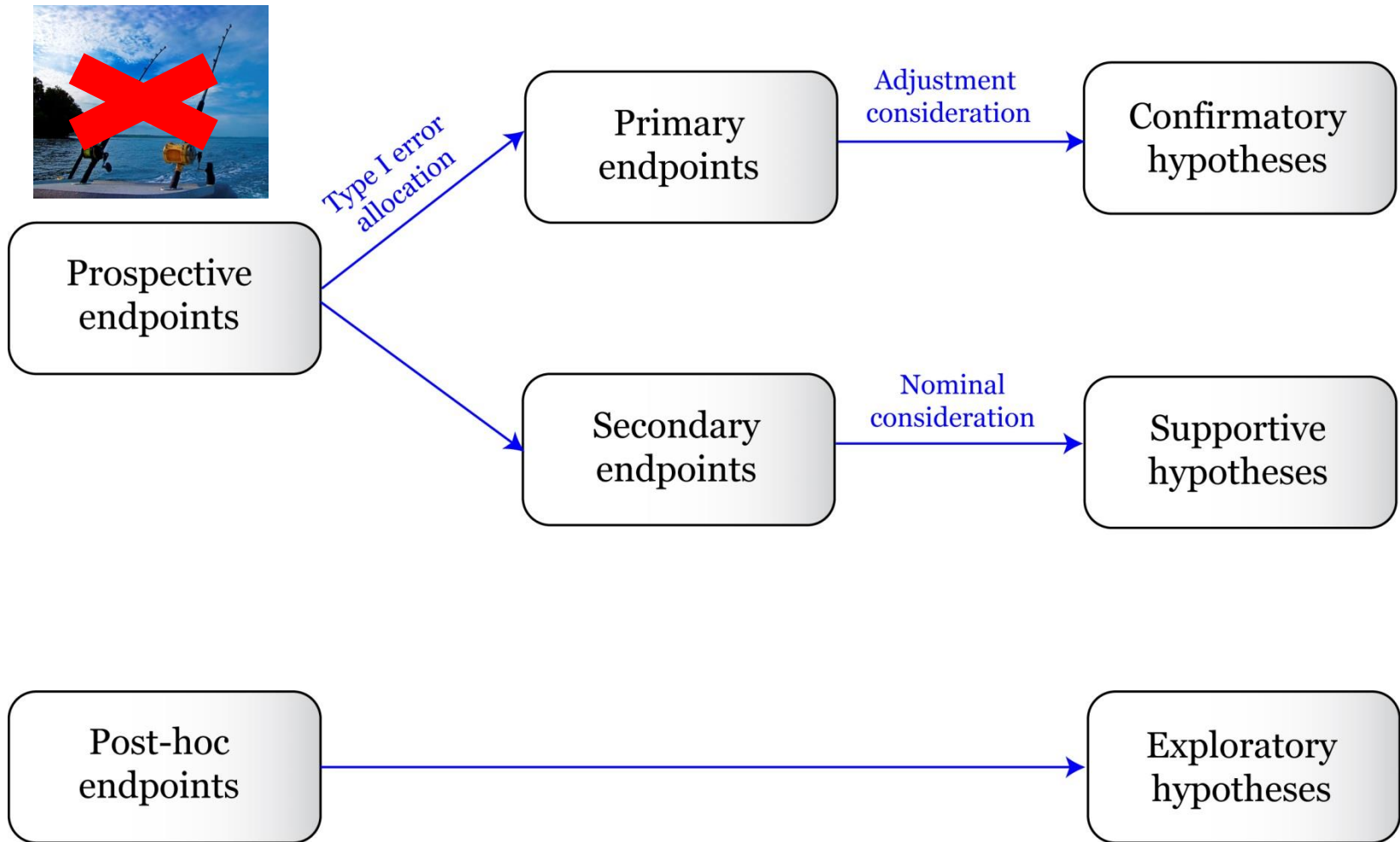# Inflation of Type I Error

Number of tests performed	2	3	10
1 or more „accepted" by chance	10%	14%	40%
α to ensure α = 0.05	0.025	0.017	0.005

Controlling Type I Errors

How to adjust?

A number of methods for adjustment.

Holm-Bonferroni:
H_1: $p = 0.03$
H_2: $p = 0.005$
H_3: $p = 0.04$
H_4: $p = 0.01$

Bonferroni's adjustment

Simply divide α by the number of comparisons.

For three tests:
$α = 0.05/3 = 0.0167$

Special adjustments

Holm-Bonferroni: more powerful than Bonferroni

$$\frac{\text{target } α \text{ level}}{n - \text{rank} + 1}$$

Step 1:
H_2: $p = 0.005$
H_4: $p = 0.01$
H_1: $p = 0.03$
H_3: $p = 0.04$

Step 2:
HB = $0.05/(4-1+1)$
HB = 0.0125
$0.005 < 0.0125$
H_2 is rejected

Step 3:
HB = $0.05/(4-2+1)$
HB = 0.0167
$0.01 < 0.0167$
H_4 is rejected

Step 4:
HB = $0.05/(4-3+1)$
HB = 0.025
$0.03 > 0.025$
H_1 is not rejected
H_3 is not rejected

Disciplined approach

Prospective endpoints → (Type I error allocation) → Primary endpoints → (Adjustment consideration) → Confirmatory hypotheses

Prospective endpoints → Secondary endpoints → (Nominal consideration) → Supportive hypotheses

Post-hoc endpoints → Exploratory hypotheses
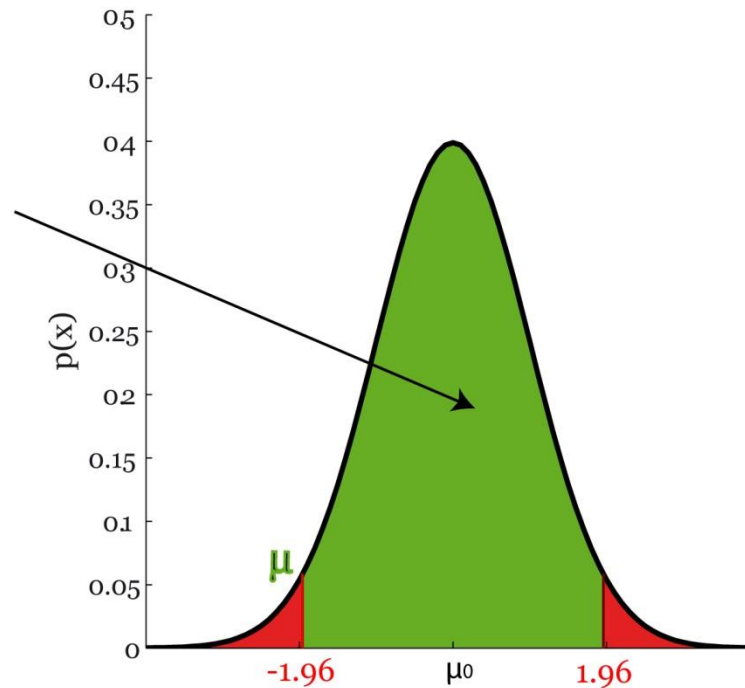
$\alpha = 0.05$

95% of all sample means ($\bar{x}$) are hypothesized to be in this region.

$H_0: \mu = \mu_0$
$H_a: \mu \neq \mu_0$

If we took a sample, and it was by chance like $\bar{x}_6$, we would incorrectly **accept** the null hypothesis.

Type II error

β here is the probability of committing a Type II error.

Reject null hypothesis | $\bar{x}_1$

Reject null hypothesis | $\bar{x}_2$

Reject null hypothesis | $\bar{x}_3$

Reject null hypothesis | $\bar{x}_4$

Reject null hypothesis | $\bar{x}_5$

Fail to reject null hypothesis | $\bar{x}_6$

Reject null hypothesis | $\bar{x}_7$

Sample size: our answer to Type II error

Margin of error

Point estimate ± Margin of error

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$
$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$
$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$z_{\alpha/2}$ = z-boundary of interval probability

$\frac{\sigma}{\sqrt{n}}$ = standard error

1. We choose E, our margin of error
2. We choose our confidence probability boundary $z_{\alpha/2}$
3. We are given or we estimate population standard deviation σ
4. Solve for n

Margin of error rearrangement

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$$

Now we can solve for the sample size n using the other three measures.

Standard deviation σ

Solving the sample size requires the population standard deviation σ. Most often we do not know it, thus we have to use an estimate:
1. Estimate σ from previous studies using the same population of interest.
2. Conduct a pilot study using a preliminary sample and calculate σ.
3. Use the judgment or "best guess" for σ. A common "guess" is the data range (high – low) divided by 4.
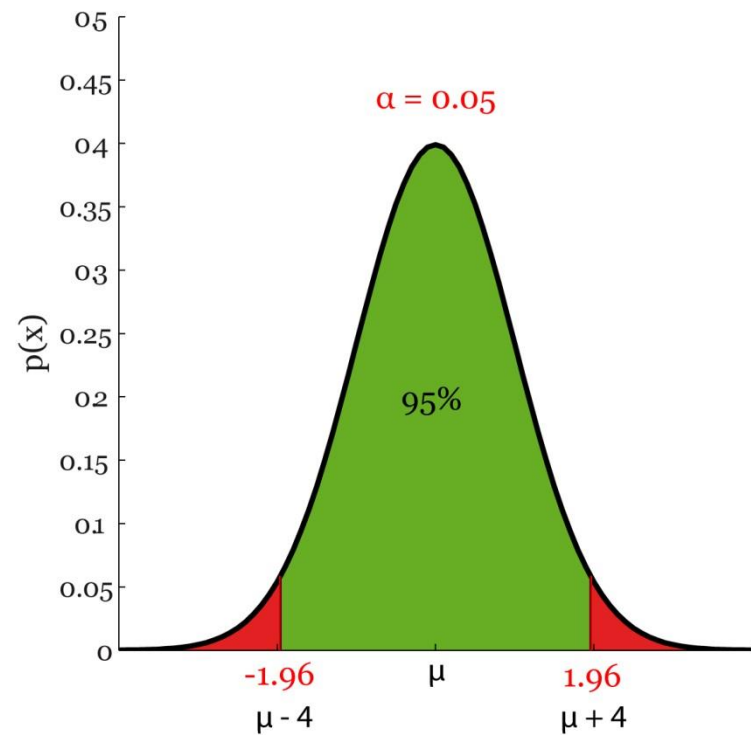
Sample size calculation example

How large a sample should be selected to provide a 95% confidence interval with a margin of error (E) of 4? Assuming that the population standard deviation σ is 16.
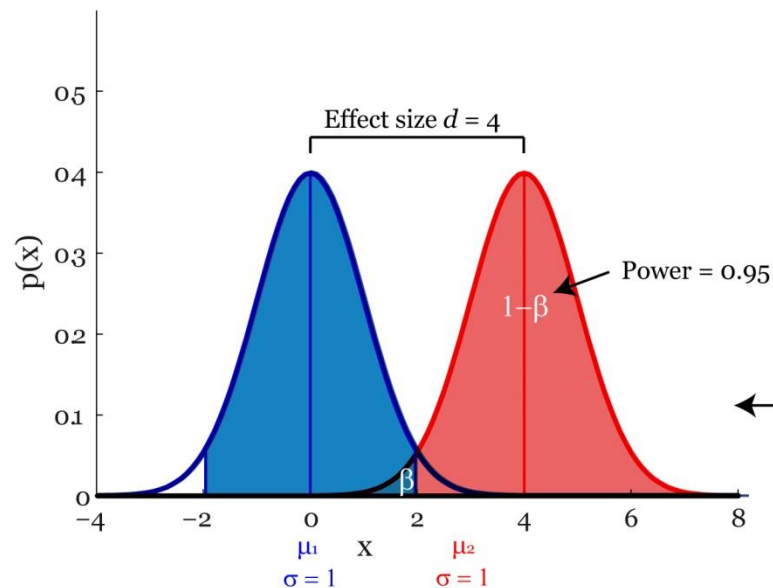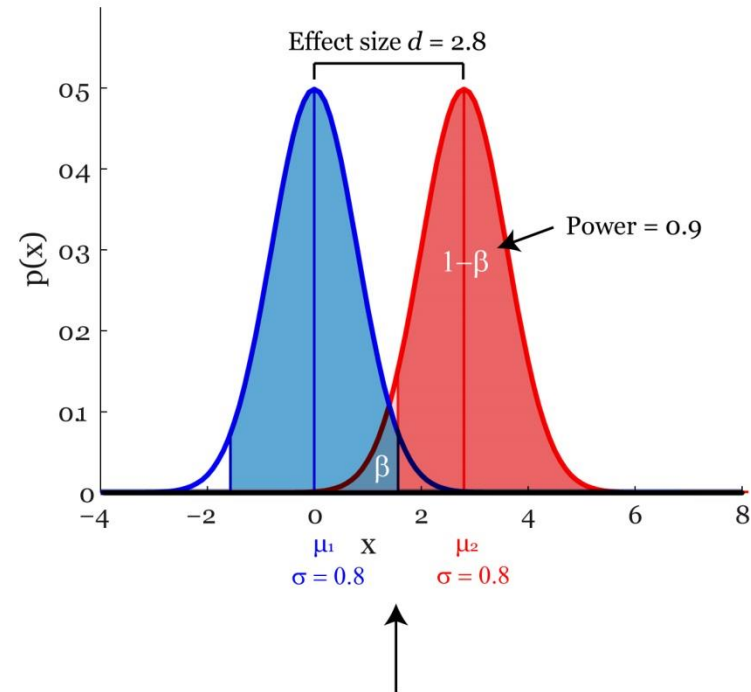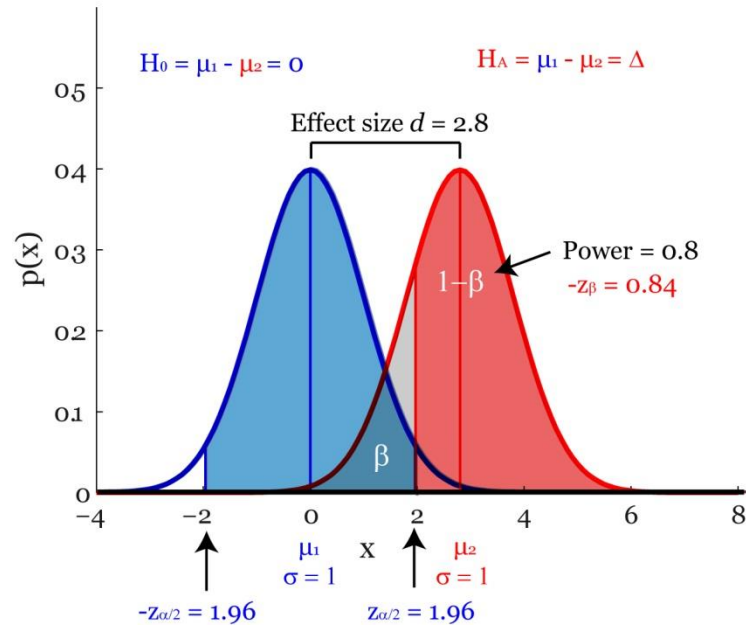
$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$$

$$n = \frac{(1.96)^2 16^2}{4^2} = 61.47$$

Interpretation: To have 95% of our sample means contain μ, we need a sample size of 61.

Sample size, power

Effect size $d = 2.8$

$H_0 = \mu_1 - \mu_2 = 0$ $H_A = \mu_1 - \mu_2 = \Delta$

Power = 0.8
$-z_\beta = 0.84$

$1-\beta$

β

μ_1 X μ_2
$\sigma = 1$ $\sigma = 1$

$-z_{\alpha/2} = 1.96$ $z_{\alpha/2} = 1.96$

Effect size $d = 2.8$

Power = 0.9

$1-\beta$

β

μ_1 X μ_2
$\sigma = 0.8$ $\sigma = 0.8$

Effect size $d = 4$

Power = 0.95

$1-\beta$

β

μ_1 X μ_2
$\sigma = 1$ $\sigma = 1$

Inreasing power by increasing sample size n leading to decrement of variability in the population

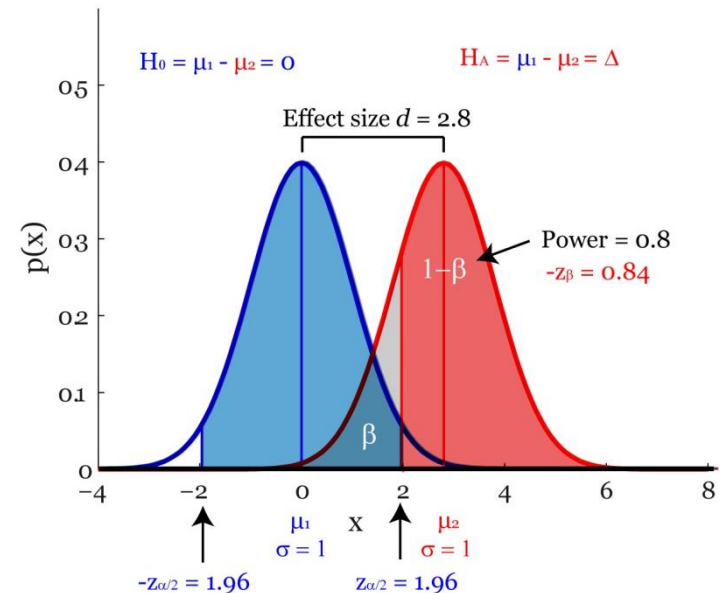Inreasing power by increasing size d of the effect in the population

Choices of α

- z_α may be one-sided or two-sided
- usually α = 0.05 and z_α = 1.645 or $z_{\alpha/2}$ = 1.96

Choices of β

- z_β is always one-sided
- usually β = 0.20 (power = 0.80) and z_β = 0.84
or β = 0.10 (power = 0.90) and z_β = 1.28

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 (\sigma_1^2 + \sigma_2^2)}{\Delta^2}$$

Sample size calculation example: two groups

We would like to determine the sample size required to detect difference of 5 points in average motor score of the Unified Parkinson's Disease Rating Scale between individuals receiving placebo versus dopaminergic drug.

- Assume a significant level (α) of 0.05 and power (β-1) of 0.80.
- Assume $\sigma_1 = \sigma_2 = 12$.
- Assume equal sample size in both groups.

$$n = \frac{(z_{\alpha/2} + z_{\beta})^2 \, (\sigma_1^2 + \sigma_2^2)}{\Delta^2}$$

$$n = \frac{(1.96 + 0.84)^2 + (12^2 + 12^2)}{5^2} = 90.3$$

Interpretation: We would need a sample size of 90 Parkinson's patients in order to capture effect of dopaminergic therapy.

G*Power

G*Power is a tool to compute statistical power analyses for many different *t* tests, *F* tests, χ² tests, *z* tests and some exact tests. G*Power can also be used to compute effect sizes and to display graphically the results of power analyses.

http://www.gpower.hhu.de/en.html

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.
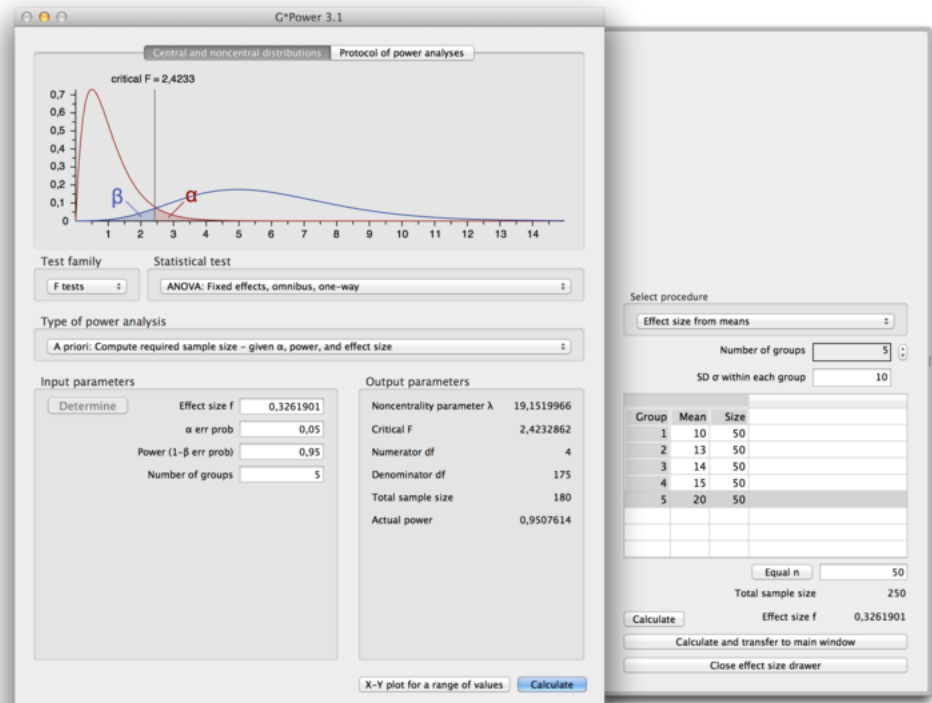
Relationship between Eta-squared, Cohen's f, and Cohen's d

$$\eta^2 = \frac{f^2}{(1+f^2)}$$

$$f = \sqrt{\frac{\eta^2}{(1-\eta^2)}}$$

Two groups (k = 2):

$$d = 2 \times f$$

Three and more groups (k = 3+):

$$d = f\sqrt{2 \times k}$$