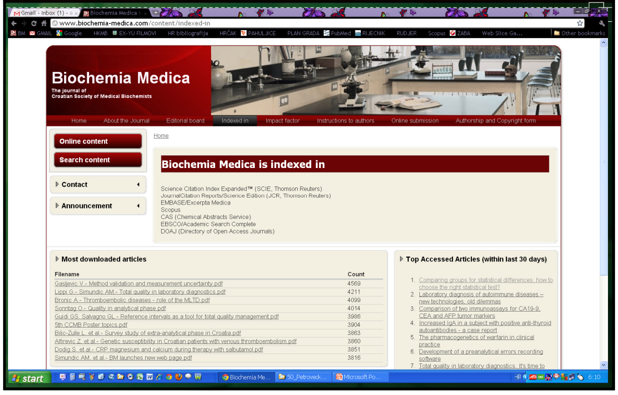
TransMedRi Workshop in Biostatistics:
Critical evaluation of statistical analysis in scientific paper

Errors in statistical analysis and data presentation

Ana-Maria Šimundić
Editor-in-chief, Biochemia Medica

Clinical Institute of Chemistry
Emergency Laboratory Department
University Hospital Center Sestre milosrdnice
Zagreb, CROATIA

http://www.biochemia-medica.com/



Biochemia Medica
The journal of Croatian Society of Medical Biochemists

Biochemia Medica is indexed in

Science Citation Index Expanded™ (SCIE, Thomson Reuters)
JournalCitation Reports/Science Edition (JCR, Thomson Reuters)
EMBASE/Excerpta Medica
Scopus
CAS (Chemical Abstracts Service)
EBSCO/Academic Search Complete
DOAJ (Directory of Open Access Journals)

Most downloaded articles

Filename	Count
Gasljevic V.- Method validation and measurement uncertainty.pdf	4569
Lippi G.- Simundic AM.- Total quality in laboratory diagnostics.pdf	4211
Bronic A.- Thromboembolic diseases - role of the MLTD.pdf	4099
Sonntag O.- Quality in analytical phase.pdf	4014
Guidi GS, Salvagno GL.- Reference intervals as a tool for total quality management.pdf	3986
5th CCMB Poster topics.pdf	3904
Bilic-Zulle L. et al.- Survey study of extra-analytical phase in Croatia.pdf	3863
Alfirevic Z. et al.- Genetic susceptibility in Croatian patients with venous thromboembolism.pdf	3860
Dodig S. et al.- CRP magnesium and calcium during therapy with salbutamol.pdf	3851
Simundic AM. et al.- BM launches new web page.pdf	3816

Top Accessed Articles (within last 30 days)

1. Comparing groups for statistical differences: how to choose the right statistical test?
2. Laboratory diagnosis of autoimmune diseases – new technologies, old dilemmas
3. Comparison of two immunoassays for CA19-9, CEA and AFP tumor markers
4. Increased IgA in a subject with positive anti-thyroid autoantibodies - a case report
5. The pharmacogenetics of warfarin in clinical practice
6. Development of a preanalytical errors recording software
7. Total quality in laboratory diagnostics. It's time to

Statistics in research

- widely accepted as a powerful tool
- increase in the use of stat methods

A great number of published medical research contains statistical errors!

Some examples…

- McKinney WP, et al. The inexact use of Fisher's Exact Test in six major medical journals. JAMA. 1989;261:3430-3.
- Kanter MH, Taylor JR. Accuracy of statistical methods in Transfusion: a review of articles from July/August 1992 through June 1993. Transfusion. 1994;34:697-701.
- Kuo YH. Extrapolation of correlation between 2 variables in 4 general medical journals. JAMA. 2002 Jun 5;287(21):2815-7.
- Nagele P. Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. Br J Anaesth. 2003 Apr;90(4):514-6.
- Simundic AM, Nikolac N, Topic E. Methodological issues in genetic association studies of inherited thrombophilia. Clin Appl Thromb Hemost. 2009;15(3):327-33.
- Simundic AM, Nikolac N. Most common statistical errors of articles submitted to Biochem Med. 2009;19(3):294-300.

Why worry?

Inappropriate use of statistical analysis
(concerns basic statistics, not advanced)

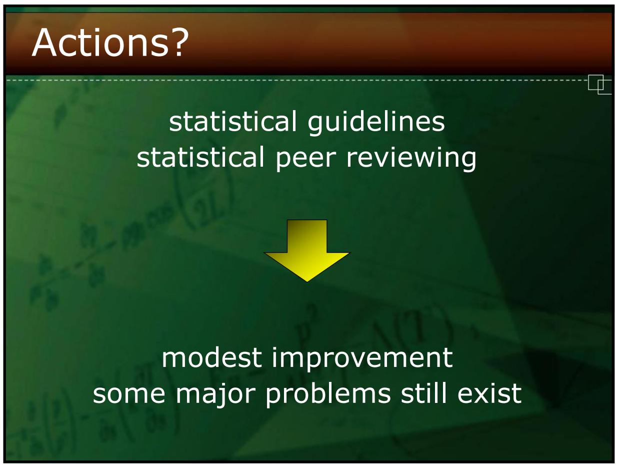
- serious problem,
- may lead to:
  - distorted results, incorrect conclusions,
  - waste of valuable resources,
- unethical,
- can have serious clinical consequences.

Errare humanum est…

'to err is human'

> " *It appears that misuse arises from various sources: degrees of competence in statistical theory and methods, honest error in the application of methods, egregious negligence, and deliberate deception (misconduct.).*"

Gardenier JS, Resnik DB. The misuse of statistics: concepts, tools, and a research agenda. Account Res. 2002;9(2):65-74.

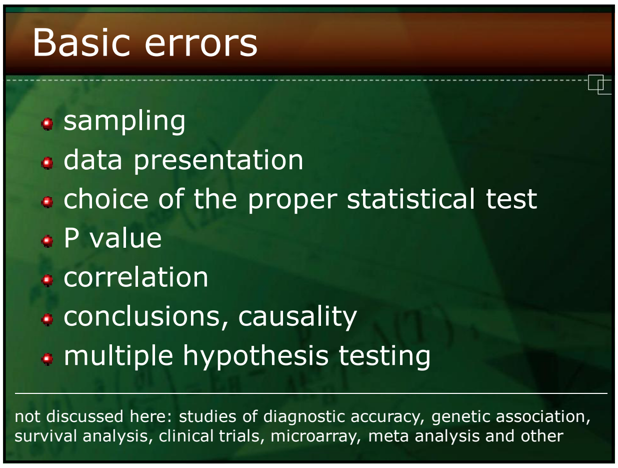


Actions?

statistical guidelines
statistical peer reviewing

↓

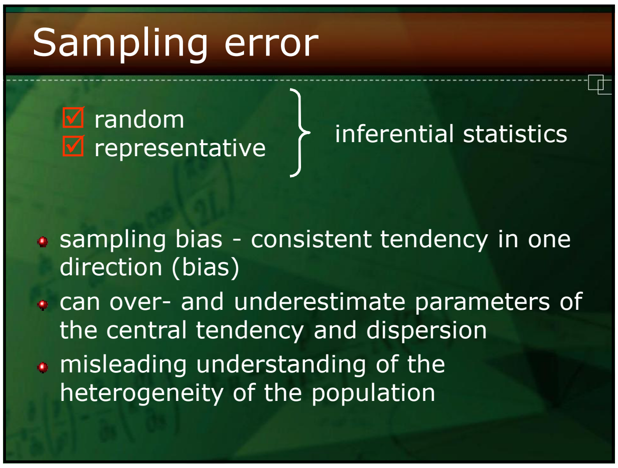
modest improvement
some major problems still exist



Basic errors

- sampling
- data presentation
- choice of the proper statistical test
- P value
- correlation
- conclusions, causality
- multiple hypothesis testing

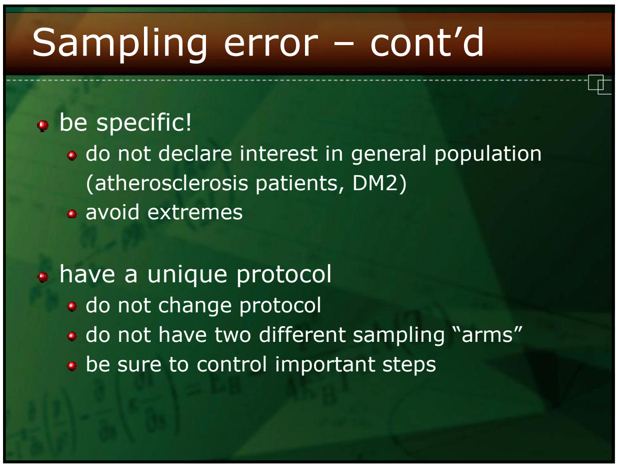
not discussed here: studies of diagnostic accuracy, genetic association, survival analysis, clinical trials, microarray, meta analysis and other



Sampling error

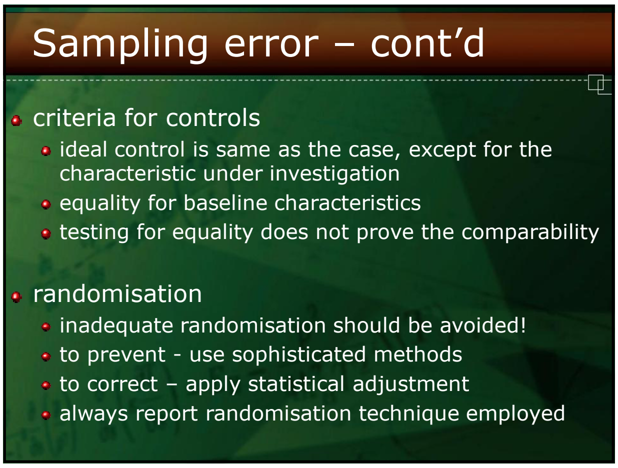
☑ random
☑ representative } inferential statistics

- sampling bias - consistent tendency in one direction (bias)
- can over- and underestimate parameters of the central tendency and dispersion
- misleading understanding of the heterogeneity of the population



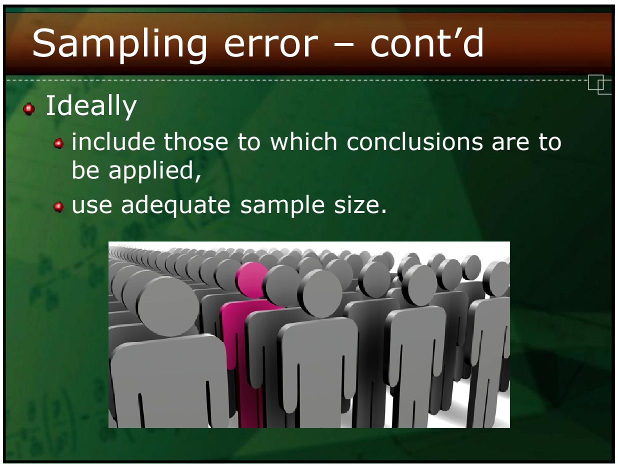
Sampling error – cont’d

- be specific!
  - do not declare interest in general population (atherosclerosis patients, DM2)
  - avoid extremes
- have a unique protocol
  - do not change protocol
  - do not have two different sampling “arms”
  - be sure to control important steps



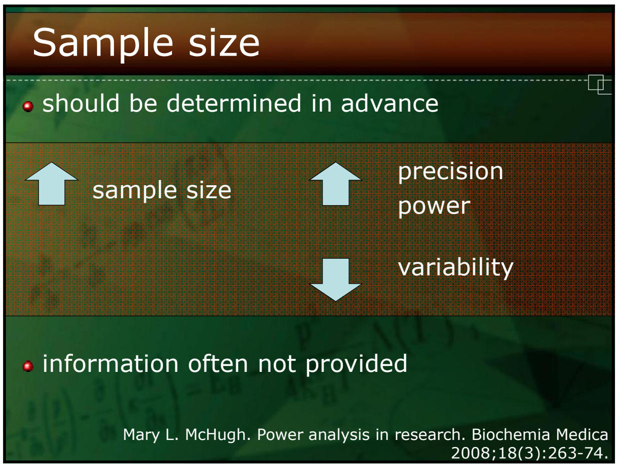
Sampling error – cont’d

- criteria for controls
  - ideal control is same as the case, except for the characteristic under investigation
  - equality for baseline characteristics
  - testing for equality does not prove the comparability
- randomisation
  - inadequate randomisation should be avoided!
  - to prevent - use sophisticated methods
  - to correct – apply statistical adjustment
  - always report randomisation technique employed



Sampling error – cont’d

- Ideally
  - include those to which conclusions are to be applied,
  - use adequate sample size.



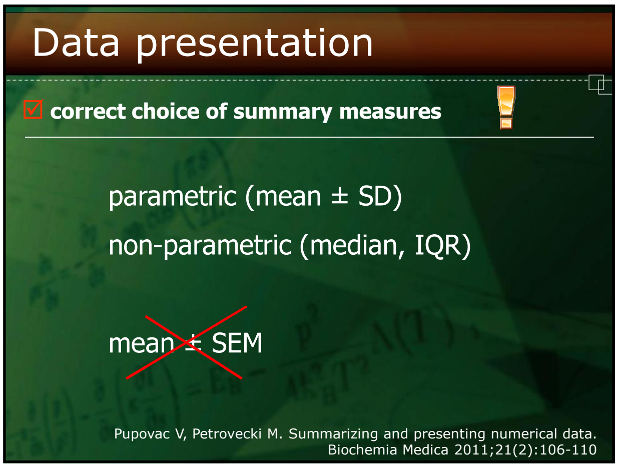
Sample size

- should be determined in advance

↑ sample size ↑ precision, power; ↓ variability

- information often not provided

Mary L. McHugh. Power analysis in research. Biochemia Medica 2008;18(3):263-74.



Data presentation

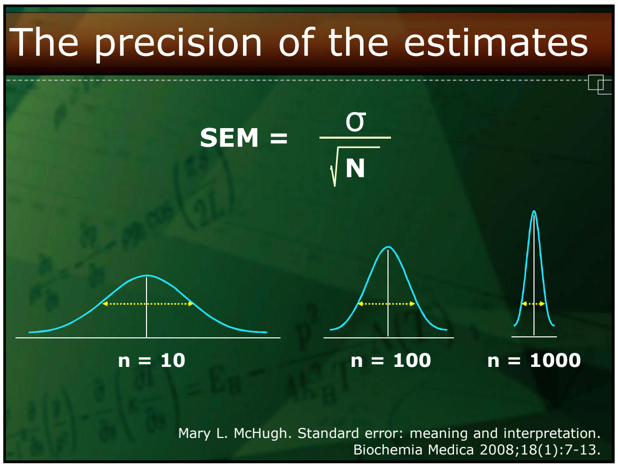
☑ **correct choice of summary measures**

parametric (mean ± SD)

non-parametric (median, IQR)

~~mean ± SEM~~

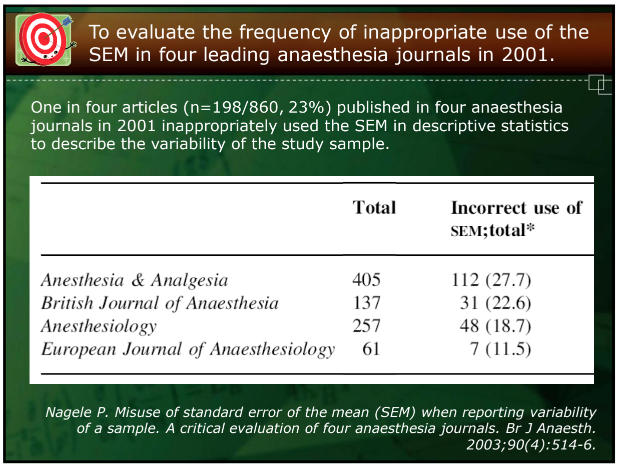
Pupovac V, Petrovecki M. Summarizing and presenting numerical data. Biochemia Medica 2011;21(2):106-110



The precision of the estimates

$$\textbf{SEM} = \frac{\sigma}{\sqrt{N}}$$

Mary L. McHugh. Standard error: meaning and interpretation. Biochemia Medica 2008;18(1):7-13.

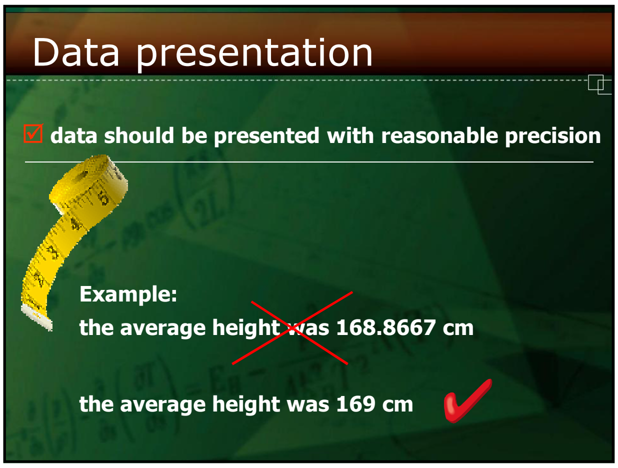


To evaluate the frequency of inappropriate use of the SEM in four leading anaesthesia journals in 2001.

One in four articles (n=198/860, 23%) published in four anaesthesia journals in 2001 inappropriately used the SEM in descriptive statistics to describe the variability of the study sample.

	Total	**Incorrect use of SEM;total***
Anesthesia & Analgesia	405	112 (27.7)
British Journal of Anaesthesia	137	31 (22.6)
Anesthesiology	257	48 (18.7)
European Journal of Anaesthesiology	61	7 (11.5)

Nagele P. Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. Br J Anaesth. 2003;90(4):514-6.



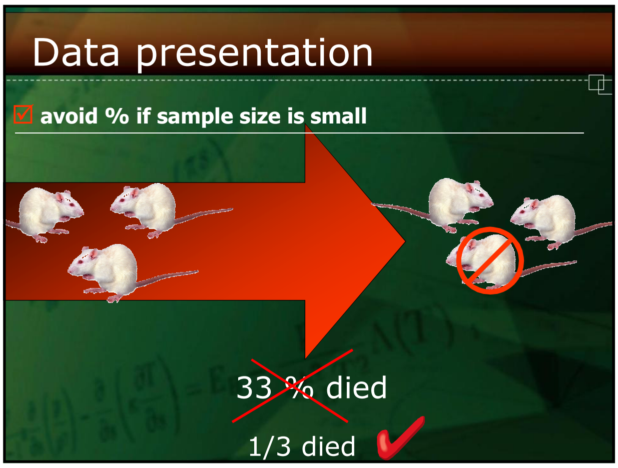
Data presentation

☑ **data should be presented with reasonable precision**

Example:

the average height ~~was 168.8667 cm~~

the average height was 169 cm ✓

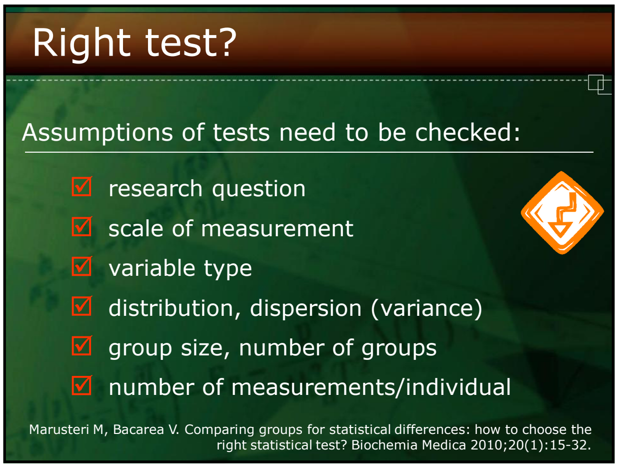


Data presentation

☑ **avoid % if sample size is small**

~~33 % died~~

1/3 died ✓

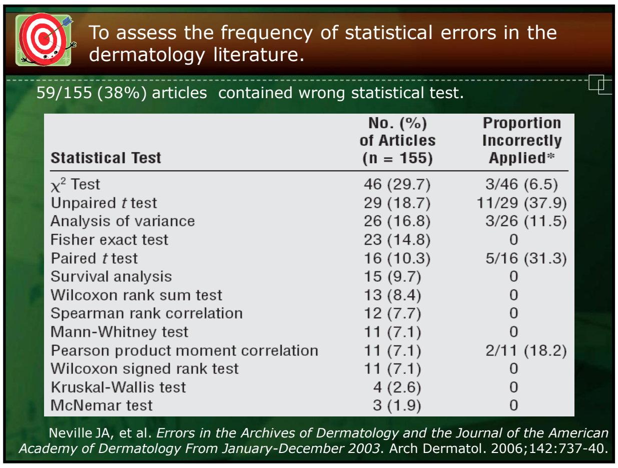


Right test?

Assumptions of tests need to be checked:

- ☑ research question
- ☑ scale of measurement
- ☑ variable type
- ☑ distribution, dispersion (variance)
- ☑ group size, number of groups
- ☑ number of measurements/individual

Marusteri M, Bacarea V. Comparing groups for statistical differences: how to choose the right statistical test? Biochemia Medica 2010;20(1):15-32.

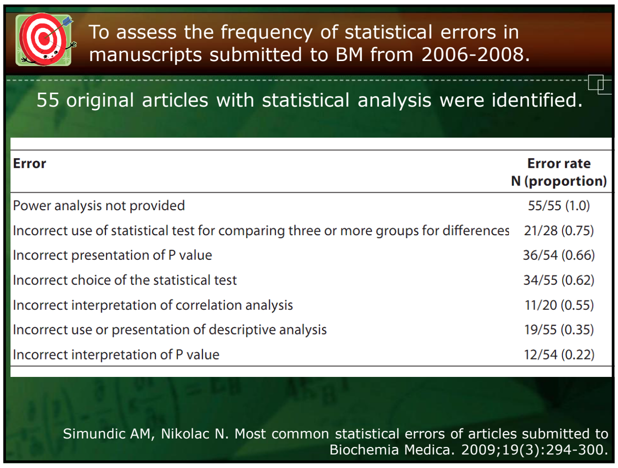


To assess the frequency of statistical errors in the dermatology literature.

59/155 (38%) articles contained wrong statistical test.

Statistical Test	No. (%) of Articles (n = 155)	Proportion Incorrectly Applied*
χ^2 Test	46 (29.7)	3/46 (6.5)
Unpaired *t* test	29 (18.7)	11/29 (37.9)
Analysis of variance	26 (16.8)	3/26 (11.5)
Fisher exact test	23 (14.8)	0
Paired *t* test	16 (10.3)	5/16 (31.3)
Survival analysis	15 (9.7)	0
Wilcoxon rank sum test	13 (8.4)	0
Spearman rank correlation	12 (7.7)	0
Mann-Whitney test	11 (7.1)	0
Pearson product moment correlation	11 (7.1)	2/11 (18.2)
Wilcoxon signed rank test	11 (7.1)	0
Kruskal-Wallis test	4 (2.6)	0
McNemar test	3 (1.9)	0

Neville JA, et al. *Errors in the Archives of Dermatology and the Journal of the American Academy of Dermatology From January-December 2003*. Arch Dermatol. 2006;142:737-40.

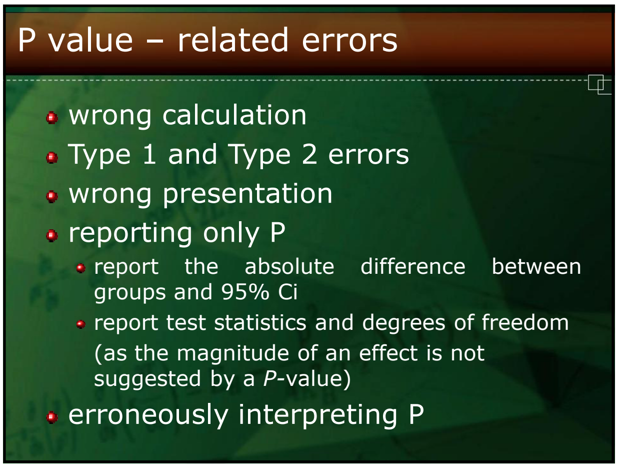


To assess the frequency of statistical errors in manuscripts submitted to BM from 2006-2008.

55 original articles with statistical analysis were identified.

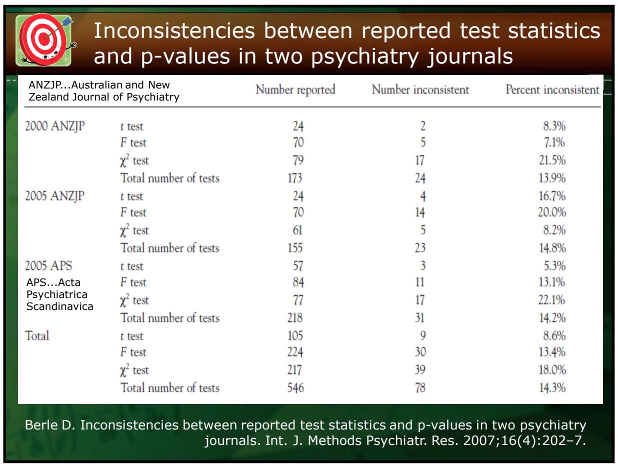
Error	Error rate N (proportion)
Power analysis not provided	55/55 (1.0)
Incorrect use of statistical test for comparing three or more groups for differences	21/28 (0.75)
Incorrect presentation of P value	36/54 (0.66)
Incorrect choice of the statistical test	34/55 (0.62)
Incorrect interpretation of correlation analysis	11/20 (0.55)
Incorrect use or presentation of descriptive analysis	19/55 (0.35)
Incorrect interpretation of P value	12/54 (0.22)

Simundic AM, Nikolac N. Most common statistical errors of articles submitted to Biochemia Medica. 2009;19(3):294-300.



P value – related errors

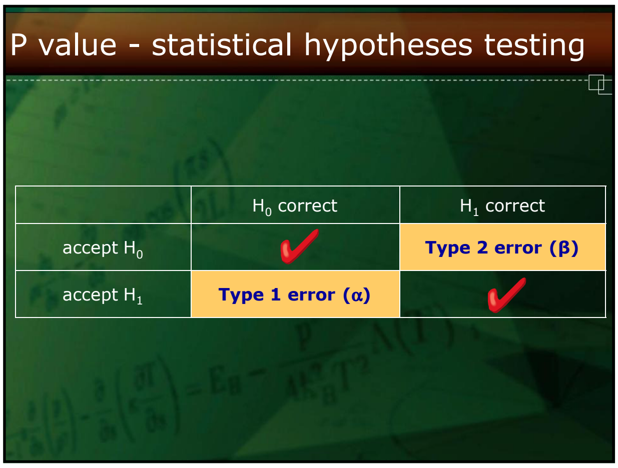
- wrong calculation
- Type 1 and Type 2 errors
- wrong presentation
- reporting only P
  - report the absolute difference between groups and 95% Ci
  - report test statistics and degrees of freedom (as the magnitude of an effect is not suggested by a *P*-value)
- erroneously interpreting P



Inconsistencies between reported test statistics and p-values in two psychiatry journals

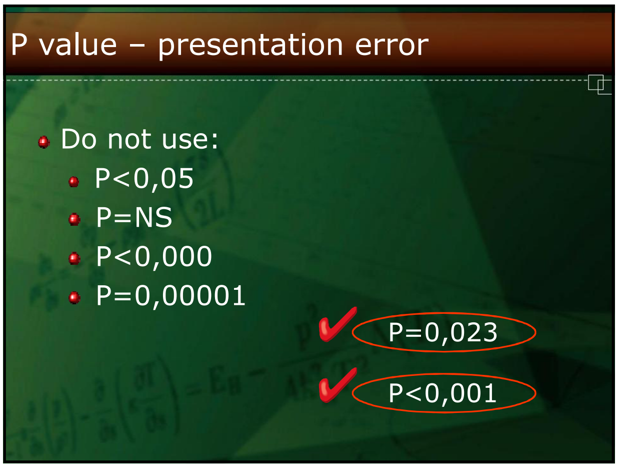
ANZJP...Australian and New Zealand Journal of Psychiatry		Number reported	Number inconsistent	Percent inconsistent
2000 ANZJP	*t* test	24	2	8.3%
	F test	70	5	7.1%
	χ^2 test	79	17	21.5%
	Total number of tests	173	24	13.9%
2005 ANZJP	*t* test	24	4	16.7%
	F test	70	14	20.0%
	χ^2 test	61	5	8.2%
	Total number of tests	155	23	14.8%
2005 APS	*t* test	57	3	5.3%
APS...Acta Psychiatrica Scandinavica	*F* test	84	11	13.1%
	χ^2 test	77	17	22.1%
	Total number of tests	218	31	14.2%
Total	*t* test	105	9	8.6%
	F test	224	30	13.4%
	χ^2 test	217	39	18.0%
	Total number of tests	546	78	14.3%

Berle D. Inconsistencies between reported test statistics and p-values in two psychiatry journals. Int. J. Methods Psychiatr. Res. 2007;16(4):202–7.



P value - statistical hypotheses testing

	H_0 correct	H_1 correct
accept H_0	✓	**Type 2 error (β)**
accept H_1	**Type 1 error (α)**	✓

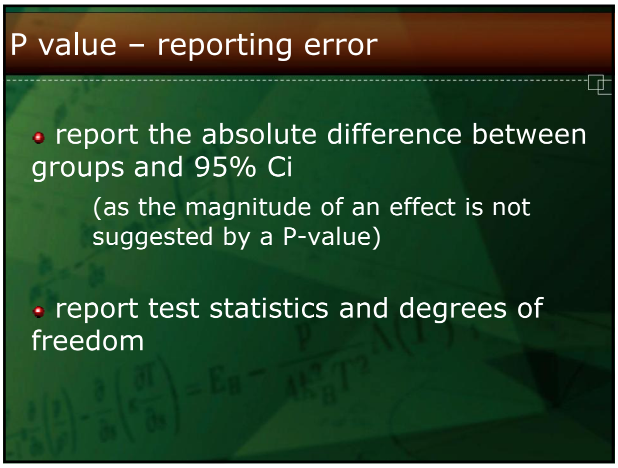


P value – presentation error

- Do not use:
  - P<0,05
  - P=NS
  - P<0,000
  - P=0,00001

✓ P=0,023

✓ P<0,001

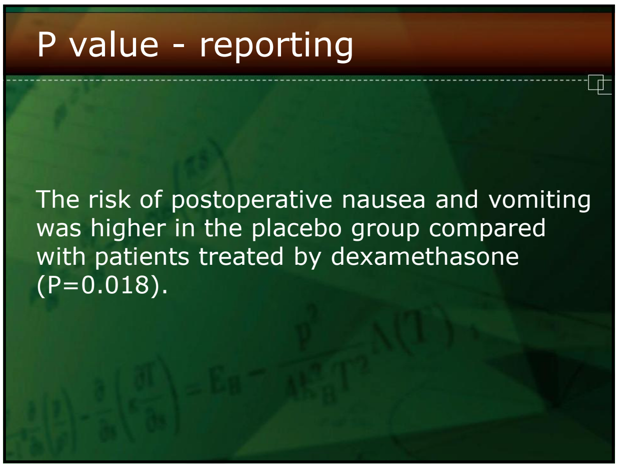


P value – reporting error

- report the absolute difference between groups and 95% Ci

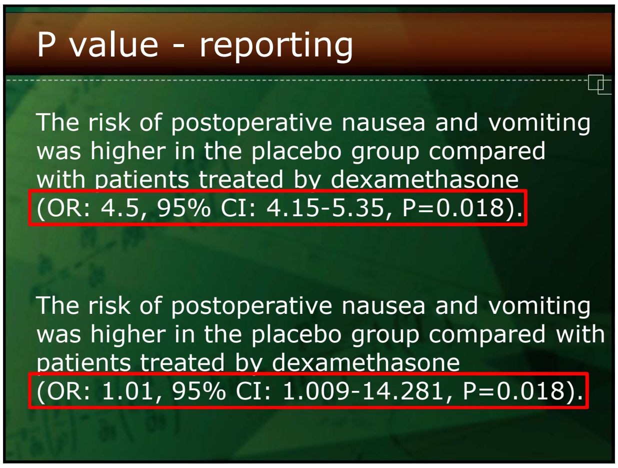
  (as the magnitude of an effect is not suggested by a P-value)

- report test statistics and degrees of freedom



P value - reporting

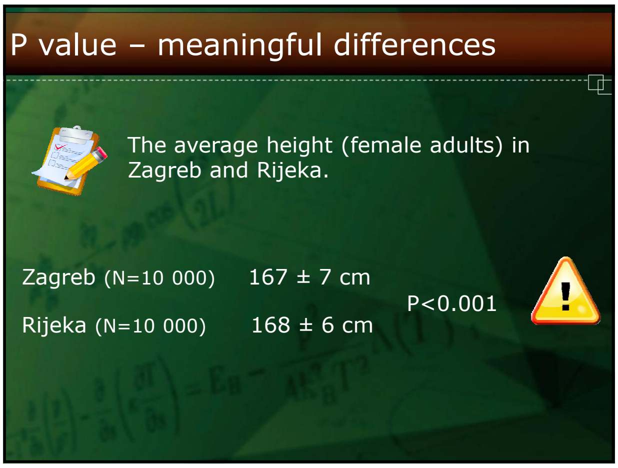
The risk of postoperative nausea and vomiting was higher in the placebo group compared with patients treated by dexamethasone (P=0.018).



P value - reporting

The risk of postoperative nausea and vomiting was higher in the placebo group compared with patients treated by dexamethasone (OR: 4.5, 95% CI: 4.15-5.35, P=0.018).

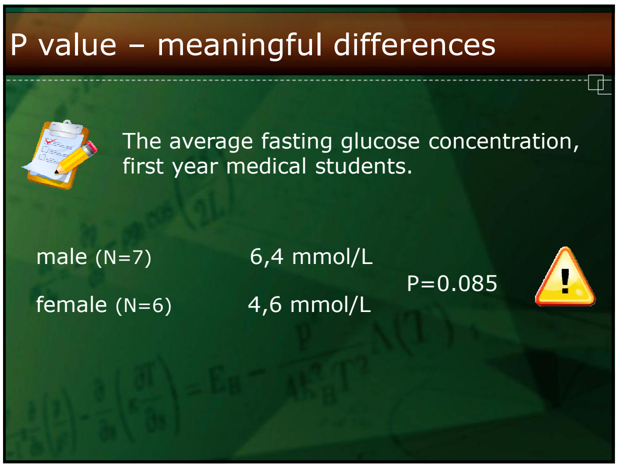
The risk of postoperative nausea and vomiting was higher in the placebo group compared with patients treated by dexamethasone (OR: 1.01, 95% CI: 1.009-14.281, P=0.018).



P value – meaningful differences

The average height (female adults) in Zagreb and Rijeka.

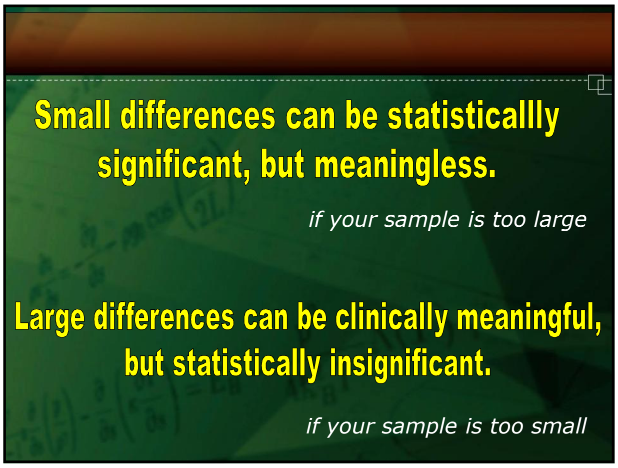
Zagreb (N=10 000)	167 ± 7 cm	P<0.001 ⚠
Rijeka (N=10 000)	168 ± 6 cm	



P value – meaningful differences

The average fasting glucose concentration, first year medical students.

male (N=7)	6,4 mmol/L	P=0.085 ⚠
female (N=6)	4,6 mmol/L	

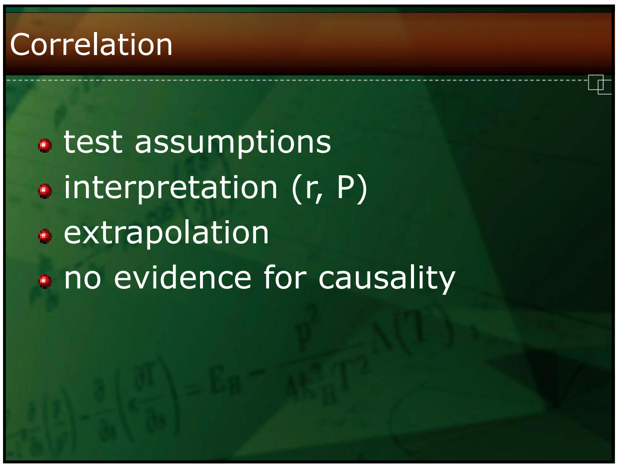


Small differences can be statisticallly significant, but meaningless.

if your sample is too large

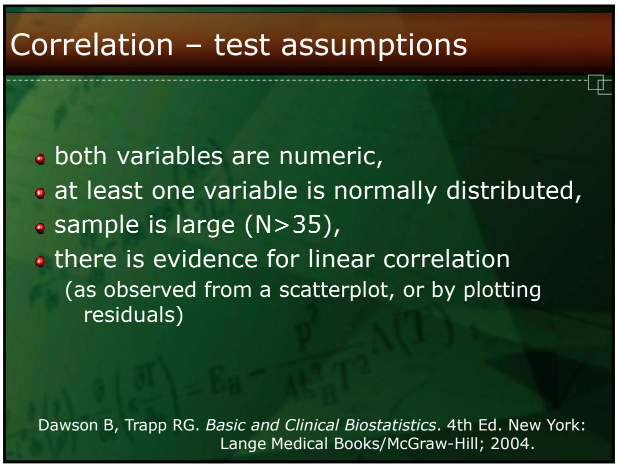
Large differences can be clinically meaningful, but statistically insignificant.

if your sample is too small



Correlation

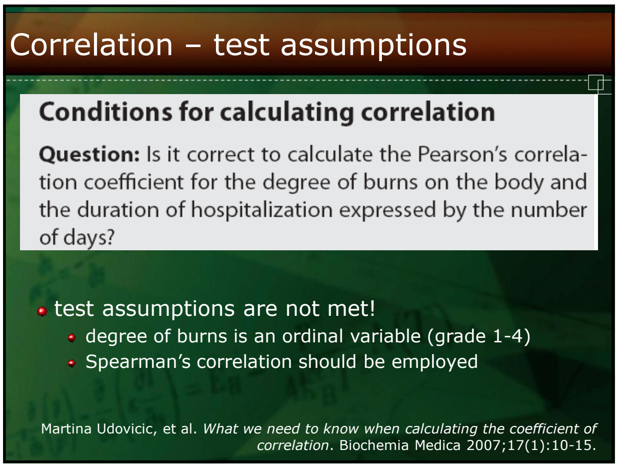
- test assumptions
- interpretation (r, P)
- extrapolation
- no evidence for causality



Correlation – test assumptions

- both variables are numeric,
- at least one variable is normally distributed,
- sample is large (N>35),
- there is evidence for linear correlation
  (as observed from a scatterplot, or by plotting residuals)

Dawson B, Trapp RG. *Basic and Clinical Biostatistics*. 4th Ed. New York: Lange Medical Books/McGraw-Hill; 2004.



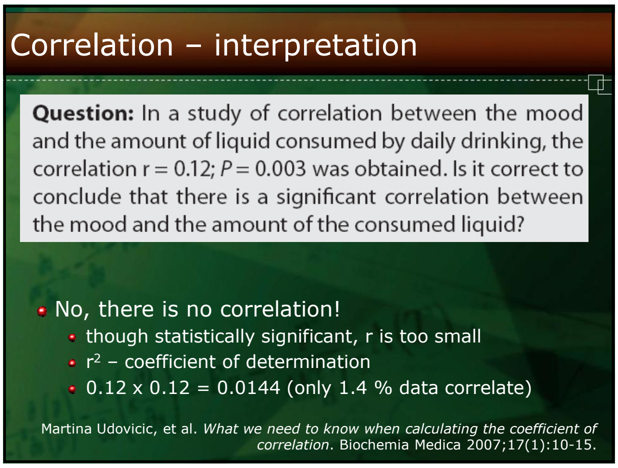
Correlation – test assumptions

Conditions for calculating correlation

Question: Is it correct to calculate the Pearson's correlation coefficient for the degree of burns on the body and the duration of hospitalization expressed by the number of days?

- test assumptions are not met!
  - degree of burns is an ordinal variable (grade 1-4)
  - Spearman's correlation should be employed

Martina Udovicic, et al. *What we need to know when calculating the coefficient of correlation*. Biochemia Medica 2007;17(1):10-15.

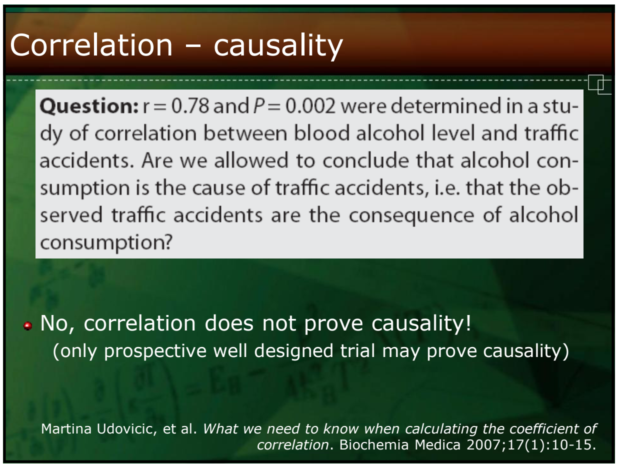


Correlation – interpretation

Question: In a study of correlation between the mood and the amount of liquid consumed by daily drinking, the correlation $r = 0.12$; $P = 0.003$ was obtained. Is it correct to conclude that there is a significant correlation between the mood and the amount of the consumed liquid?

- No, there is no correlation!
  - though statistically significant, r is too small
  - r^2 – coefficient of determination
  - 0.12 x 0.12 = 0.0144 (only 1.4 % data correlate)

Martina Udovicic, et al. *What we need to know when calculating the coefficient of correlation*. Biochemia Medica 2007;17(1):10-15.

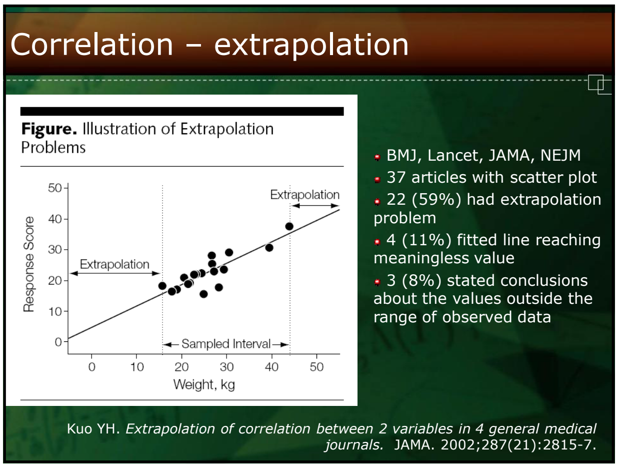


Correlation – causality

Question: $r = 0.78$ and $P = 0.002$ were determined in a study of correlation between blood alcohol level and traffic accidents. Are we allowed to conclude that alcohol consumption is the cause of traffic accidents, i.e. that the observed traffic accidents are the consequence of alcohol consumption?

- No, correlation does not prove causality!
  (only prospective well designed trial may prove causality)

Martina Udovicic, et al. *What we need to know when calculating the coefficient of correlation*. Biochemia Medica 2007;17(1):10-15.

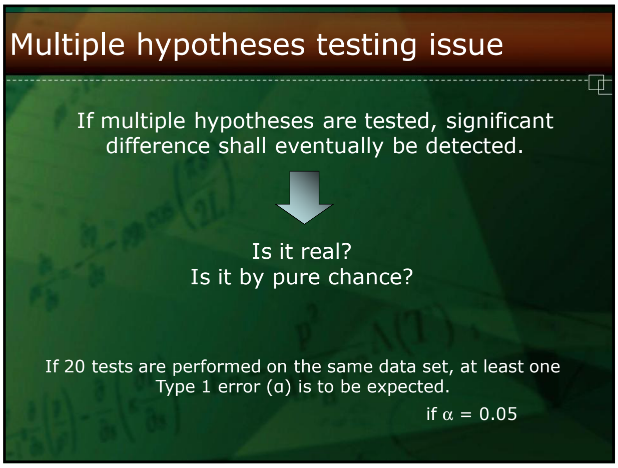


Correlation – extrapolation

Figure. Illustration of Extrapolation Problems

- BMJ, Lancet, JAMA, NEJM
- 37 articles with scatter plot
- 22 (59%) had extrapolation problem
- 4 (11%) fitted line reaching meaningless value
- 3 (8%) stated conclusions about the values outside the range of observed data

Kuo YH. *Extrapolation of correlation between 2 variables in 4 general medical journals.* JAMA. 2002;287(21):2815-7.



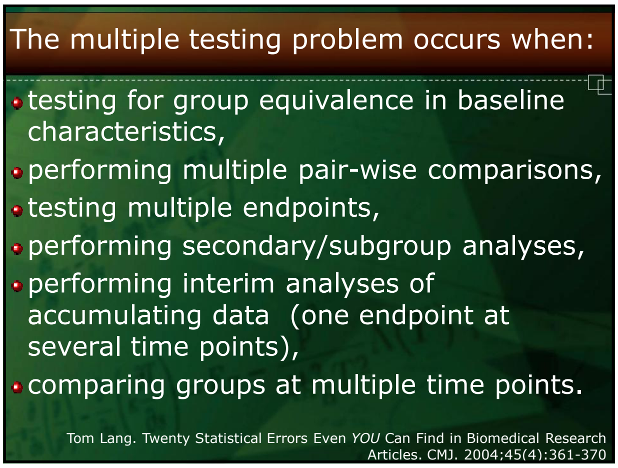
Multiple hypotheses testing issue

If multiple hypotheses are tested, significant difference shall eventually be detected.

Is it real?
Is it by pure chance?

If 20 tests are performed on the same data set, at least one Type 1 error (α) is to be expected.

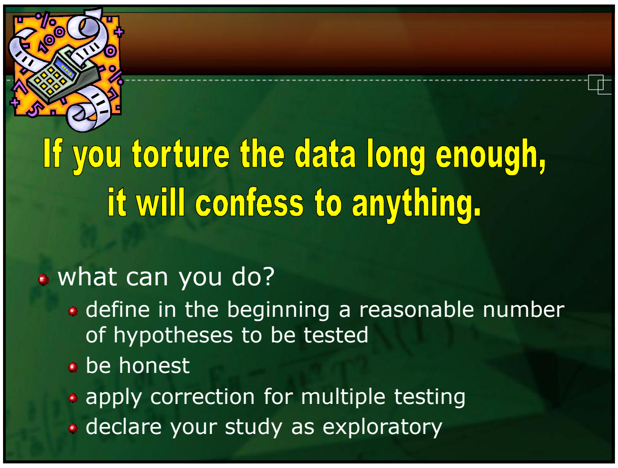
if α = 0.05



The multiple testing problem occurs when:

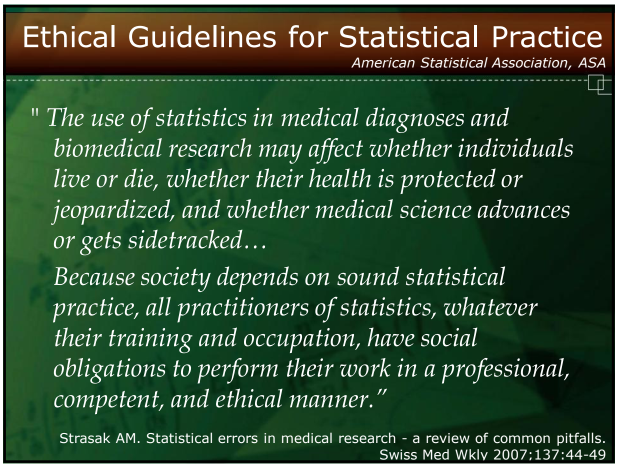
- testing for group equivalence in baseline characteristics,
- performing multiple pair-wise comparisons,
- testing multiple endpoints,
- performing secondary/subgroup analyses,
- performing interim analyses of accumulating data (one endpoint at several time points),
- comparing groups at multiple time points.

Tom Lang. Twenty Statistical Errors Even *YOU* Can Find in Biomedical Research Articles. CMJ. 2004;45(4):361-370



If you torture the data long enough, it will confess to anything.

- what can you do?
  - define in the beginning a reasonable number of hypotheses to be tested
  - be honest
  - apply correction for multiple testing
  - declare your study as exploratory



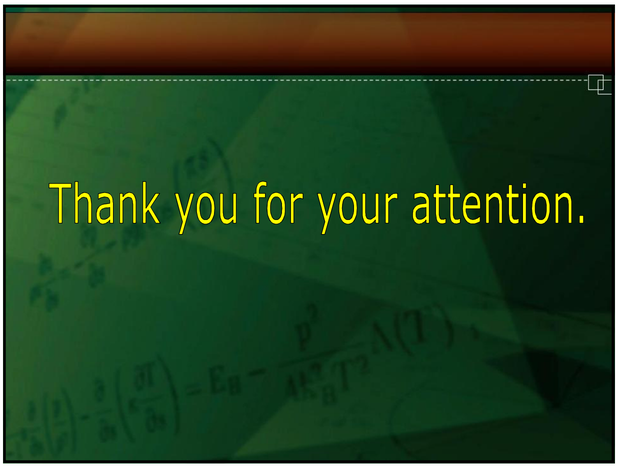
Ethical Guidelines for Statistical Practice

American Statistical Association, ASA

" The use of statistics in medical diagnoses and biomedical research may affect whether individuals live or die, whether their health is protected or jeopardized, and whether medical science advances or gets sidetracked…

Because society depends on sound statistical practice, all practitioners of statistics, whatever their training and occupation, have social obligations to perform their work in a professional, competent, and ethical manner."

Strasak AM. Statistical errors in medical research - a review of common pitfalls. Swiss Med Wkly 2007;137:44-49



Thank you for your attention.