Statistical methods in medical research

BMJ, 1954

Report from a meeting in the Royal Statistical Society:

“Medicine was an art, statistics a science …… when it came to mixing science and art, statistics was out of place as a skillet in a Crown Derby tea-service.”

Use of statistics in medical journals (Sept. 2009).

		Statistics
The Lancet	Paper 1	Repeated measures, Survival (Cox)
	Paper 2	Repeated measures
	Paper 3	Survival (Cox), logistic
	Paper 4	Survival (Cox)
NEJM	Paper 1	Survival (Cox)
	Paper 2	Survival (Cox)
	Paper 3	More simple statistics
	Paper 4	Survival (Cox)

JAMA	Paper 1	Simple statistics
	Paper 2	Simple statistics
	Paper 3	Repeated measures
	Paper 4	Clustered logistic, negative binomial
	Paper 5	Simple statistics
Nature Medicine	Paper 1	Linear regression
	Paper 2	Mann-Whitney, Kaplan-Meier
	Paper 3	Mann-Whitney
	Paper 4	T-test, ANOVA
	Paper 5	T-test, chi-square

The American Statistician
2007, vol. 61, no. 1, pp. 47 - 55

The Use of Statistics in Medical Research: A Comparison of *The New England Journal of Medicine* and *Nature Medicine*

Alexander M. Strasak; Qamruz Zaman; Gerhard Marinell; Karl P. Pfeiffer; Hanno Ulmer

Abstract

There is widespread evidence of the extensive use of statistical methods in medical research. Just the same, standards are generally low and a growing body of literature points to statistical errors in most medical journals. However, there is no comprehensive study contrasting the top medical journals of basic and clinical science for recent practice in their use of statistics.

All original research articles in Volume 10, Numbers 1-6 of *Nature Medicine* (*Nat Med*) and Volume 350, Numbers 1-26 of *The New England Journal of Medicine* (*NEJM*) were screened for their statistical content. Types, frequencies, and complexity of applied statistical methods were systematically recorded. A 46-item checklist was used to evaluate statistical quality for a subgroup of papers.

94.5 percent (95% CI 87.6-98.2) of *NEJM* articles and 82.4 percent (95% CI 65.5-93.2) of *Nat Med* articles contained inferential statistics. *NEJM* papers were significantly more likely to use advanced statistical methods ($p < 0.0001$). Statistical errors were identified in a considerable proportion of articles, although not always serious in nature. Documentation of applied statistical methods was generally poor and insufficient, particularly in *Nat Med*. Compared to 1983, a vast increase in usage and complexity of statistical methods could be observed for *NEJM* papers. This does not necessarily hold true for *Nat Med* papers, as the results of the study indicate that basic science sticks with basic analysis. As statistical errors seem to remain common in medical literature, closer attention to statistical methodology should be seriously considered to raise standards.

Why are statistics so important in medical research?

Two obvious reasons:

- Mechanistic understanding is (still) limited. Must trust observations / data

- We produce an enormous amount of data

Back to the introductory lecture day 1.

- Scurvy. James Lind set up a controlled trial with lime. No knowledge of Vitamine C (1747).

- John Snow – transmission of cholera (1854).

- Up to our days: Cigarette smoking causes lung cancer (1950). Still no real mechanistic understanding.

So, statistics is important, as one would believe that it can contribute to causal understanding.

We will use these lectures to present and discuss statistics as a methodological tool, or methodological platform, common to most types of medical research.

We need to go into some basics.

Hypothesis testing

Remember from lectures in science theory, a crucial point in the Hypothetico-deductive model is to set up some hypothesis, and to try to falsify this.

In statistics, we have built up a machinery for doing this.

We will start out by looking at a randomized study (typically clinical study), as this is obviously closest to the idea of showing causality.

Assume a simple study, set up to test the effect of some cholesterol reducing drug. We are interested in the average level of cholesterol after three months of treatment.

This is a randomized controlled trial, meaning we randomize patients to one treatment group and one control group (could be more groups).

What is our hypothesis, the one that we want to falsify?

Our null hypothesis is that there is no difference between the active treatment and the control treatment, in the population.

We measure effect by the average level of cholesterol after three months of treatment.

Let μ_1 denote the average population level based on the active treatment, while μ_2 denotes the corresponding level based on the control treatment.

Our formal hypothesis is

$$H_0: \mu_1 = \mu_2$$

We want to test this against the alternative: $H_1: \mu_1 \neq \mu_2$.

In our data we observe the average levels $\overline{X}_1$ and $\overline{X}_2$ in the two groups. Say we observe 4.5 mmol/l vs. 5.3 mmol/l.

Question: Based on our observations, can we reject H_0?

Could this observed difference be due to chance, or is it a substantial effect of treatment?

T-test:

$$\frac{\overline{X}_1 - \overline{X}_2}{s_p \sqrt{1/n_1 + 1/n_2}}$$

where s_p denotes the standard deviation, assumed to be equal in the two groups.

It can be shown that this statistic is t-distributed, which means we can use the t-distribution to calculate p-values:

If we believe in our hypothesis (H_0); what is the probability of observing our data?

If there is really no effect of the treatment; what is the probability of observing as large difference as we did $(\overline{X}_1 - \overline{X}_2)$ just by chance?

This is the p-value.

A low p-value means there is little chance of observing this difference if there is no real effect of treatment.

So, low p-values (typically <0.05) is taken as an indication of effect, and we reject H_0.

This also means we accept a 5% chance of rejecting H_0 even if it is true!

Assume the common SD is 1.1 mmol/l, and that we have 100 patients in each group. This gives a test statistic of 5.1, and the p-value can be shown to be <0.001.

Assume instead we have the same SD, but only 20 patients in each group. This gives a t-statistic = 2.3, and p = 0.03.

P-value obviously a function of the number of observations!

Implications for planning of studies.

”To consult a statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.”
(R. A. Fisher)

We can clearly re-write H_0 as

$$H_0: (\mu_1 - \mu_2) = 0$$

Going back to the t-statistic, this can be written

$$\frac{(\bar{X}_1 - \bar{X}_2) - 0}{SE(\bar{X}_1 - \bar{X}_2)},$$

so it is the difference between what we observe and what we assume under H_0, scaled by a measure of uncertainty (or noise) in data / estimate.

Effect measure

We will typically be interested in saying something about the size of the effect of the new treatment.

We want some measure of effect.

In our example, we compared the two treatments with regard to average level of cholesterol, so the typical effect measure will be the difference between the mean values, $\overline{X}_1 - \overline{X}_2$.

We will present this, together with a measure of uncertainty, typically a 95% confidence interval.

Based on $\overline{X}_1$ and $\overline{X}_2$ equal 4.5 mmol/l and 5.3 mmol/l, respectively, SD = 1.1 in both groups and sample sizes 20 vs. 100, we calculate the following intervals for (μ_1 - μ_2):

0.8 (0.5 – 1.1) for n = 100
0.8 (0.1 – 1.5) for n = 20

Confidence interval given by

$$(\overline{X}_1 - \overline{X}_2) \pm c \times s_p \sqrt{1/n_1 + 1/n_2}$$

Assume we instead of measuring level of cholesterol after thee months, follow these patients for a longer period and count patients with e.g. heart disease.

What will then be the natural effect measure?

Let $\hat{p}_1, \hat{p}_2$ be the observed proportions of patients with heart disease in the two groups.

Two effect measures typically used are

$\hat{p}_1 - \hat{p}_2$ estimated risk difference

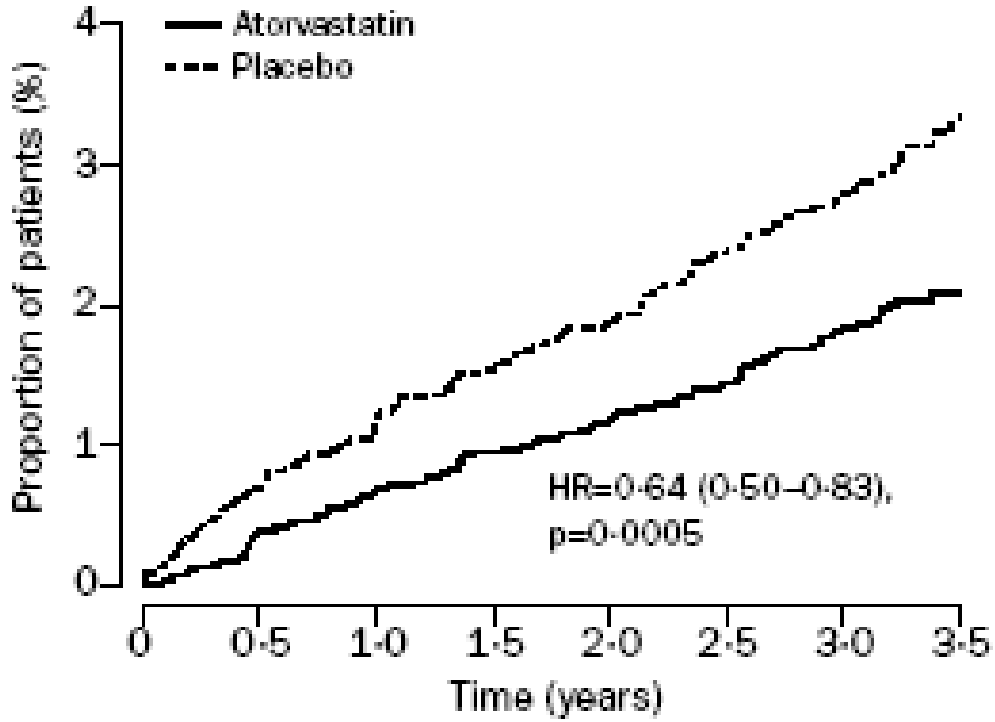
$\hat{p}_1 / \hat{p}_2$ estimated relative risk (RR)

Interpretation of effect estimates

The ASCOT study (lipid lowering arm),
Lancet, 2003

Compares a statin and placebo with regard to risk of heart disease.

Hazard ratio (RR) = 0.64 (0.50 – 0.83)



Risk is reduced by 36%.

What about absolute risk (risk difference)?

Statin: 100 events in 5168 persons.

Placebo: 154 events in 5137 persons.

Absolute risk: 2% vs. 3%

Risk difference: 1%

Oppgave

ORIGINAL ARTICLE

Intensive Lipid Lowering with Simvastatin and Ezetimibe in Aortic Stenosis

Anne B. Rossebø, M.D., Terje R. Pedersen, M.D., Ph.D., Kurt Boman, M.D., Ph.D., Philippe Brudi, M.D., John B. Chambers, M.D., Kenneth Egstrup, M.D., Ph.D., Eva Gerdts, M.D., Ph.D., Christa Gohlke-Bärwolf, M.D., Ingar Holme, Ph.D., Y. Antero Kesäniemi, M.D., Ph.D., William Malbecq, Ph.D., Christoph A. Nienaber, M.D., Ph.D., Simon Ray, M.D., Terje Skjærpe, M.D., Ph.D., Kristian Wachtell, M.D., Ph.D., and Ronnie Willenheimer, M.D., Ph.D., for the SEAS Investigators*

You need to know that the hazard ratio is to be interpreted as a relative risk.

Questions for discussion

- This study is based on a seemingly well founded theory about the anticipated effect of treatment. However, the data do not seem to support this theory. What is the basis for the theory, and how strong is the evidence against it?

- The study ends up showing an increased risk of cancer in the treatment group. How should we relate to this finding?

- Randomized trials are in some sense the gold standard when we talk about evidence-based medicine. However, there are problems also in such studies. Discuss this based on figure 1. What does “Discontinued placebo, followed per protocol” mean?

- There are at least two levels of effect in this study. What types of effect measures are used?

- How do you read the figures 2 and 3?

- How do you interpret p-values in table 1 in light of how we define the p-value?

Nytt tema

P-values

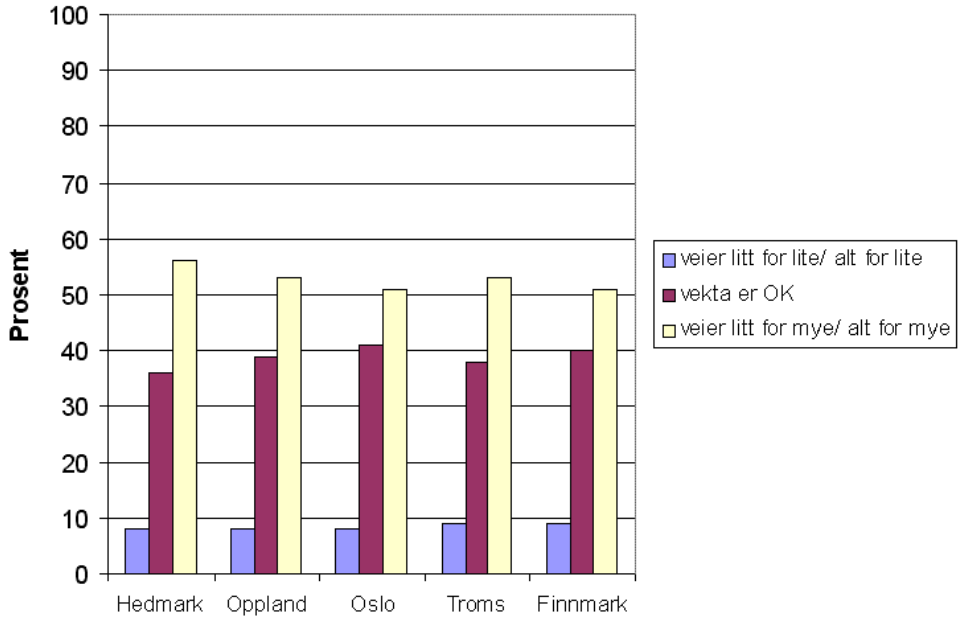
The medical area has a tradition for presenting results from a study in terms of significance / non-significance, or in terms of p-values

This is somewhat problematic!

Dagbladet, 2004 – survey about 15-16 year olds:

’I think my own bodyweight is okay’

Oslo:	41%
Oppland:	39%
Hedmark:	36%
Troms:	38%
Finmark:	40%



”Forskjellene vi fant er signifikante, selv om prosentforskjellene på gjennomsnittet ikke er så tydelige.”

BMJ, 1999, Letter to editor

Meta analysis of clozapine vs. ”typical” antipsychotic drugs.
Compares sponsored and non-sponsored studies.

Relapse:

Significant

OR = 0.5 (0.3 – 0.7) sponsored
OR = 0.4 (0.1 – 1.4) non-sponsored

Leaving the study early:

Significant

OR = 0.5 (0.4 – 0.7) sponsored
OR = 0.6 (0.3 – 1.2) non-sponsored

” The observation that drug industry sponsorship is associated with more-favourable outcomes is of concern.”

K. Rothman (1998)

”When writing for *Epidemiology*, you can also enhance your prospects if you omit tests of statistical significance. ….. In *Epidemiology*, we do not publish them at all.”

Nytt tema

Microarray studies

February, 2001

Microarray data and gene expressions

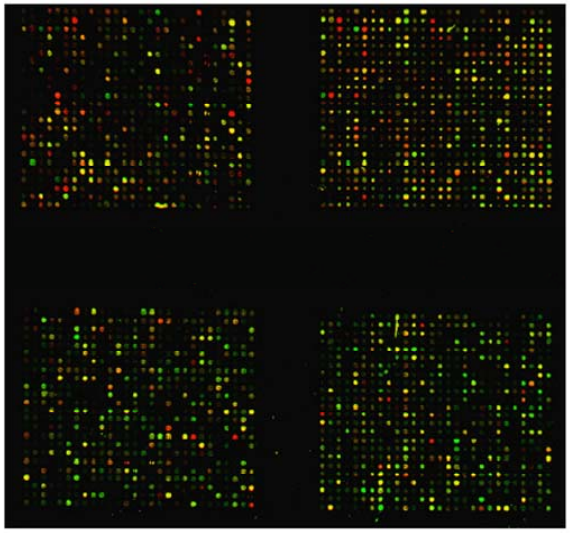
DNA

↓ transcription

mRNA

↓ translation

protein → cell phenotype → organism phenotype



- Microarrays measure gene expression at the transcription level
- Gene expression is a measure of how much a gene transcribes
- Gene expressions tell how much a gene might contribute to biological dynamics

Microarrays enable monitoring of expression levels for thousands of genes simultaneously.

The human genome consists of ~35 000 genes.

Will typically be interested in identifying so-called differentially expressed genes between cases and controls, or between intervention group and control group.

Study set up to investigate the effect of a certain diet, rich in antioxidants, on gene expression.

10 persons in intervention, 10 persons in control.

Measurement from 35 000 genes, before and after the intervention!

What are we interested in?

- Identify single genes that respond to the intervention

- Group genes with a similar pattern of behaviour

- Derive a biological pathway, a network of genes jointly responsible for a biological dynamics

Identify single genes.

For each single gene we can test the hypothesis H_0: $\mu_1 = \mu_2$, e.g. by a t-test.

This test leads to a p-value, we reject H_0 if p<significance level.

With a significance level of 5% this means there is a 5% chance that we claim an effect even if there is no.

We are about to produce 35 000 such p-values!

Multiple testing

If we assume all genes act independently of each other and there is truly no effect of the treatment, how many “false positive” findings will we expect?

$0.05 \times 35\,000 = 1\,750$!!!

We obviously have to do something about this!

Illustration

We create a simulated data set, where nothing is differentially expressed, and then we compute the t statistics and the p-values. No gene should be found as differentially expressed.

Simulation of 6,000 genes with 8 treatments and 8 controls: All the gene expression values were simulated *i.i.d* from a $N(0,1)$ distribution, i.e. NOTHING is differentially expressed in our simulation.

Ten smallest p-values:

"Gene"-index	t-statistic	p-value
2271	4.93	2×10^{-4}
5709	4.82	3×10^{-4}
5622	-4.62	4×10^{-4}
4221	4.34	7×10^{-4}
3156	-4.31	7×10^{-4}
5898	-4.29	7×10^{-4}
2164	-3.98	1.4×10^{-3}
5930	3.91	1.6×10^{-3}
2427	-3.90	1.6×10^{-3}
5694	-3.88	1.7×10^{-3}

Two common ways to deal with multiple testing problems

Control family-wise error rate

P(# rejected $H_0 \geq 1 \mid H_0$ true) < significance level

Typical example: Bonferroni correction.

Control false discovery rate (FDR)

The expected proportion of false rejections among all rejections < significance level

	# declared non-significant	# declared significant	Total
# true null hypotheses	U	V *(false pos.)*	m_0
# non-true null hypotheses	T	S	$m - m_0$
Total	$m - R$	R	m

FWER = $P(V \geq 1 \mid H_0)$ < typically 0.05

FDR = $E(V/R)$ < typically 0.05

FDR is less conservative, and preferred in these situations.

Problem: How to handle varying (and unknown) dependencies among genes?

Back to the intervention study

Under FDR < 5% we found 44 transcripts to be differentially expressed in the intervention group as compared to the control group when comparing changes from before to after intervention.

Unfortunately, not all of these have a known function.

Maybe more interesting:

Identify differentially regulated gene sets

Databases with pre-defined gene sets exist.

One option:

Rank your genes / transcripts according to some measure of effect (e.g. the t-statistic).

Test whether pre-defined sets of genes cluster in one or the other end of your list.

As of today: 5452 gene sets defined in one of these databases.

So, again multiple testing problems. Solved by use of FDR corrections.

In our data, we found a number of gene sets to be up- or down regulated. These are typically related to DNA repair and defence responses.

But, where are the effect estimates …..??

Nytt tema

Observational studies

Epidemiology is mainly based upon observational studies.

Main problem: Confounding – an observed association between an exposure and an outcome is really caused by another factor.

Epidemiology, 2005

Folate Supplementation and Twin Pregnancies

Stein Emil Vollset,[†] Håkon K. Gjessing,[‡] Anne Tandberg,[§] Thorbjørn Rønning,[†] Lorentz M. Irgens,[*†] Valborg Baste,[†] Roy M. Nilsen,[†] and Anne Kjersti Daltveit[*†]*

Background

Pregnant women and women who plan to become pregnant are advised to increase their intake of folate to prevent neural tube defects.

Several studies have reported an association between use of folate and multiple births.

Folic acid has also been used to increase litter size in the swine industri.

This possible association with multiple births has been used as an argument against fortification of foods.

Medical Birth Registry, births 1998 – 2001.

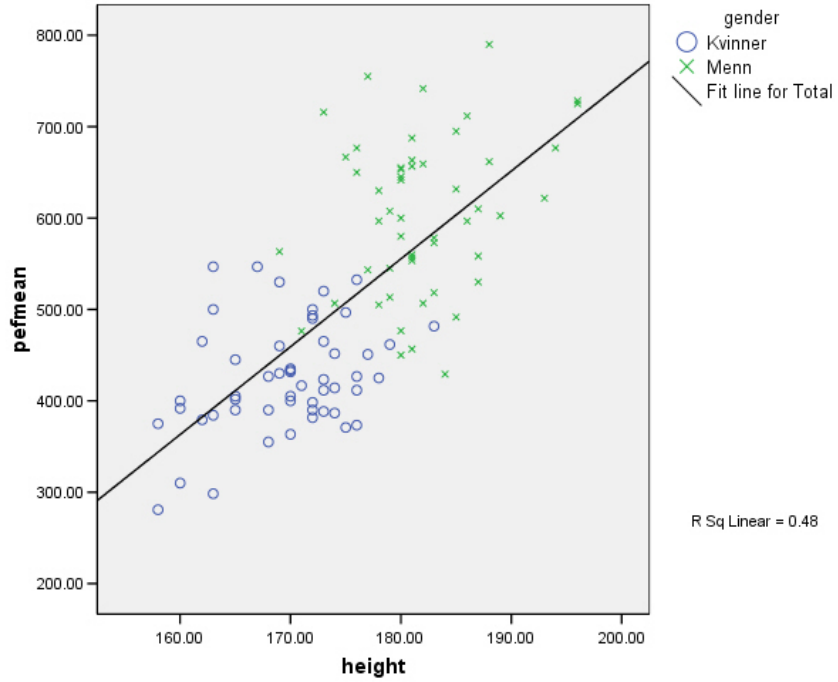
Observed association between folate use and twin births:
OR = 1.76 (1.57 – 1.97).

Logistic regression, adjustment for maternal age and parity:
OR = 1.59 (1.41 – 1.78).

Further adjustment for in vitro fertilization:
OR = 1.04 (0.91 – 1.18).

Linear regression example:

Association between height and lung function.

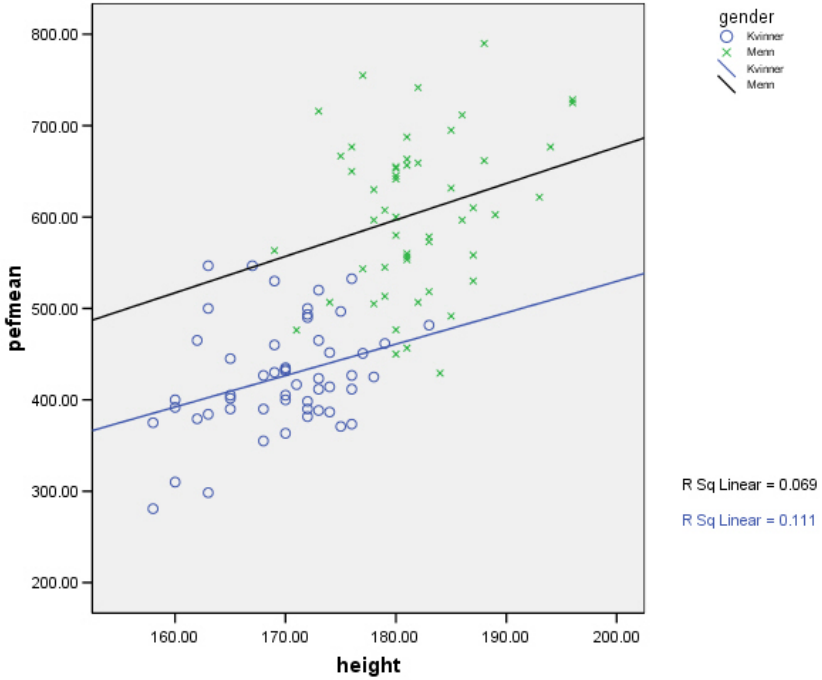


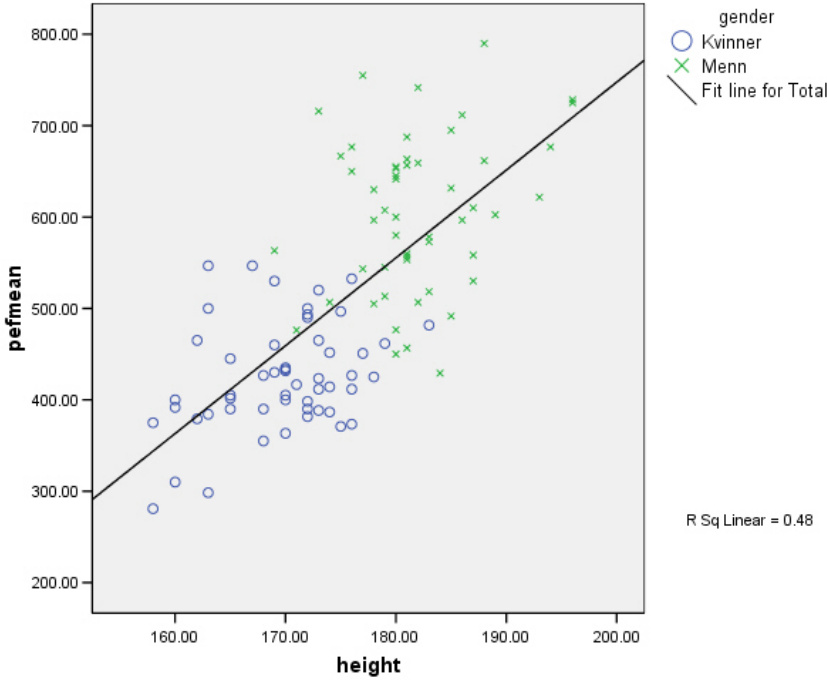
Regression equation

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Pefmean = -1174.9 + 9.61×height + ε

What about the effect of gender?


gender
Kvinner
Menn
Kvinner
Menn
800.00
700.00
600.00
500.00
400.00
300.00
200.00
pefmean
160.00
170.00
180.00
190.00
200.00
height
R Sq Linear = 0.069
R Sq Linear = 0.111


gender
Kvinner
Menn
Fit line for Total
800.00
700.00
600.00
500.00
400.00
300.00
200.00
pefmean
160.00
170.00
180.00
190.00
200.00
height
R Sq Linear = 0.48

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Pefmean = -337.3 + 3.7×height + 133.7×gender + ε

For given gender, the effect of height is estimated to be an increase in lung function of 37 litres with a 10 cm increased height.

Opposite: For a given height, the effect of gender is estimated to be 133.7 litres.

Remember we are now assuming a model for the data.

Back to the logistic example, what does this model look like?

Assume

$p = \Pr(\text{twin})$

Model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \text{folate} + \beta_2 \times \text{IVF}$$

Linear and additive on log-odds scale!

Nytt tema

Non-parametric tests (and small samples)

Lab studies are typically performed on very small samples, and also typically analyzed by non-parametric (rank) tests.

These tests are based on the ranking of the observations, and throw away the actual observed data.

It makes sense not to trust t-tests in these small-sample situations, as the normality assumption may be doubtful.

However, also the non-parametric tests are based on some assumptions.

Study from Rikshospitalet.

GAD65 IgG autoantibodies in stiff person syndrome: clonality, avidity and persistence

Compares binding capacity of GAD65 in serum and cerebrospinal fluid (CSF). Data from five patients.

Serum	CSF	Diff
184.00	2.60	181.49
1.60	.02	1.58
9.00	.40	8.60
2640.00	16.50	2623.50
880.00	4.50	875.50

Wilcoxons signed rank test

Idea:

- Compute the differences
- Rank the absolute values of the differences
- Summarize the ranks of the positive (and the negative) differences.

If no difference, these sums should be about equal.

Binding capacity B_{max} (nM, mean ± SEM)[c]	
Serum	CSF
184 ± 19	2.6 ± 0.1
1.6 ± 0.2	0.021 ± 0.00
9.0 ± 0.2	0.4 ± 0.08
2640 ± 300	16.5 ± 3.4
880 ± 180	4.5 ± 0.5

are shown in Fig. 2. All SPS patients displayed higher GAD65 binding capacities in serum than in CSF, but this difference was not statistically significant ($P = 0.06$). In patients SPS 2 and 3, the avidity of the

In this situation (all differences in the same direction), this test reduces to calculating the probability of having five (out of five) differences in the same direction.

Under H_0, this is $0.5^5 = 0.03$, and a two-tailed test then gives a p-value of $2 \times 0.03 = 0.06$, no matter how large the difference is!

This test also builds on an assumption about a symmetric distribution of the differences.

Impossible to assess in such small samples.

Further, the Mann-Whitney test is often seen and used as a distribution-free alternative to the two-sample t-test, used to compare mean values in situations with non-normal data.

Is that what it does?

A key question – what does nonparametric mean?

What is the alternative?

What does parametric mean?

Assume a variable X normally distributed: $X \sim N(\mu,\sigma)$, where μ denotes the expected value and σ denotes the standard deviation.

(X may be cholesterol level in the population, (μ,σ) is then the population mean and standard deviation, respectively).

μ, σ are called parameters.

Parametric statistics does inference about such parameters.

We may for instance test the hypothesis

$H_0: \mu = 0$

Typical situation:

Compare mean values in two independent groups – two-sample t-test vs. Mann-Whitney test.

$H_0: \mu_1 = \mu_2$

“The t-test assumes the data to be normally distributed.”

True only in small samples!

This means that the Mann-Whitney test is an interesting alternative in small sample situations mainly.

BUT, the Mann-Whitney test is not testing the same hypothesis, as it is a non-parametric test.

SPSS …….

We will often be interested in some estimated quantity, in addition to the hypothesis test. In that respect, the nonparametric tests are somewhat difficult to interpret. They are seldom estimating anything we are interested in.

In general, the Mann-Whitney test is testing whether the two distributions differ.

“An assumption behind the Mann-Whitney test is that the two distributions have the same shape”.

This is really not an assumption behind the test, but it is a necessary condition if you want to test the hypothesis about mean values.

Another situation where nonparametric tests are of interest is when we have data that are pure rankings. People may be asked to rank different treatments according to preference. In these situations it may not be meaningful to talk about mean values, and the nonparametric test is a good alternative.

To summarize:

Statistics is an important research tool in most types of medical research …….

……. BUT, you should know what you are doing! There are many difficulties!