Biostatistics

Hypothesis testing

Burkhardt Seifert & Alois Tschopp

Biostatistics Unit
University of Zurich

Testing of hypotheses

What does a test do?

Due to sample $\longrightarrow$ never certainty about facts:
by chance or not?

Statistical tests $\longrightarrow$ decision rules with specified probabilities

Introducing **examples**:

- Confirmation that therapy A is better than therapy B (difference is not by chance).
- Aetiologic confirmation for diseases (asbestos, smoking as risk factors for diseases).
- Evidence for theories, e.g. "emotionally disturbed childhood may lead to mental illness"

Randomised trial of safety and efficacy of immediate postoperative enteral feeding in patients undergoing gastrointestinal resection

Cornelia S Carr, K D Eddie Ling, Paul Boulos, Mervyn Singer

Abstract

***Objectives*—To assess whether immediate postoperative enteral feeding in patients who have undergone gastrointestinal resection is safe and effective.**

***Design*—Randomised trial of immediate postoperative enteral feeding through a nasojejunal tube *v* conventional postoperative intravenous fluids until the reintroduction of normal diet.**

***Setting*—Teaching hospitals in London.**

***Subjects*—30 patients under the care of the participating consultant surgeon who were undergoing elective laparotomies with a view to gastrointestinal resection for quiescent, chronic gastrointestinal disease. Two patients did not proceed to resection.**

***Main outcome measures*—Nutritional state, nutritional intake and nitrogen balance, gut mucosal permeability measured by lactulose-mannitol differential sugar absorption test, complications, and outcome.**

***Results*—Successful immediate enteral feeding was established in all 14 patients, with a mean (SD) daily intake of 6·78 (1·57) MJ (1622 (375) kcal before reintroduction of oral diet compared with 1·58 (0·14) MJ (377 (34) kcal) for those on intravenous fluids ($P<0·0001$). Urinary nitrogen balance on the first postoperative day was negative in those on intravenous fluids but positive in all 14 enterally fed patients (mean (SD) −13·2 (11·6) g *v* 5·3 (2·7) g; $P<0·005$). There was no difference by day 5. There was no change in gut mucosal permeability in the enterally fed group but a significant increase from the test ratios seen before the operation in those on intravenous fluids (0·11 (0·06) *v* 0·15 (0·12); $P<0·005$). There were also fewer postoperative complications in the enterally fed group ($P<0·005$).**

***Conclusions*—Immediate postoperative enteral feeding in patients undergoing intestinal resection seems to be safe, prevents an increase in gut mucosal permeability, and produces a positive nitrogen balance.**

STATISTICAL ANALYSIS

We used Student's *t* test and Cox's proportional hazards model for analysis. All analyses were stratified by the number of measurements of serum cholesterol for each subject (three to five).

Table 1—*Relative risks (95% confidence interval) of suicide among 6393 men by average serum cholesterol concentration and change in cholesterol concentration*

	No of subjects	No of suicides	Adjusted relative risk (95% confidence interval)*	P value
Average serum cholesterol concentration (mmol/l)†				
<4.78	827	10	3.16 (1.38 to 7.22)	0.007
4.78-6.21	3600	13	1.00	
>6.21	1966	9	1.28 (0.55 to 3.01)	0.56
Change in serum cholesterol concentration (mmol/l a year)‡				
Decline >0.13	1143	11	2.17 (0.97 to 4.84)	0.056
Change ≤0.13	2795	13	1.00	
Increase >0.13	2455	8	0.72 (0.30 to 1.72)	0.46

*Relative risks for average cholesterol concentration were adjusted, using Cox's proportional hazards model, for age, smoking habits (never, former, or current), and mean corpuscular volume at first examination. Relative risks for change in cholesterol concentration were adjusted as above and for average serum cholesterol concentration.

†Mean of serum cholesterol concentrations from all examinations.

‡Estimated using within person linear regression method (0.13 mmol/l equivalent to 5 mg/dl).

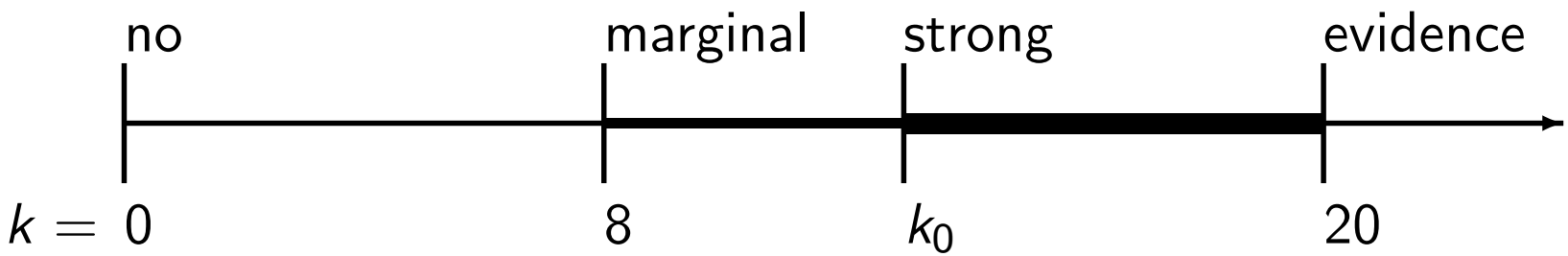
Example

Standard drug is effective in 40% of all cases ($p = 0.4$).
Is a new drug better?

Sample $n = 20$ patients

If equally good $\longrightarrow$ on average $k = 8$ patients are cured

Evidence that $p_{\text{new}} > 0.4$:



k = number of cured patients

Example

Question: How likely is $k \geq k_0$, if $p_{\text{new}} = 0.4$?

$\longrightarrow$ k binomial distributed with $p = 0.4$

$$\begin{array}{lrll} \longrightarrow & \mathrm{P}(k \geq 11) & = & 0.128 \quad \text{from table} \\ & \mathrm{P}(k \geq 12) & = & 0.057 \\ & \mathrm{P}(k \geq 13) & = & 0.021 \\ & \mathrm{P}(k \geq 14) & = & 0.006 \end{array}$$

Logic:
If one observes $k \geq 13$, then $p_{\text{new}} = 0.4$ is unlikely
and one concludes $p_{\text{new}} > 0.4$

General formalization

H_1: Scientific hypothesis or **alternative hypothesis**

Example: $\mathrm{H}_1: \; p_{\text{new}} > 0.4$

Originates e.g. from scientific or clinical experience

H_0: Statistical hypothesis or **null hypothesis**

Example: $\mathrm{H}_0: \; p_{\text{new}} = 0.4 \quad \text{or} \quad (p_{\text{new}} - 0.4) = 0$

Pay attention:

Both hypotheses refer to population parameters and not sample realizations.

Statistical test

- Testing the null hypothesis
- If null hypothesis is implausible based on data (example: $k \geq 13$)
  $\longrightarrow$ Decide in favour of the scientific hypothesis H_1; reject H_0.
- If null hypothesis is plausible (example: $k < 13$)
  $\longrightarrow$ Keep the null hypothesis (e.g. old therapy);
  H_1 is not proven

Possible errors made in a decision:

		Truth	
		H_0 is true	H_0 is not true
Decision	Do not reject H_0	true	type II error "β"
	Reject H_0	type I error "α"	true

Wrongly rejecting H_0 is in general worse than wrongly not rejecting H_0 ("conservative").

$\longrightarrow$ Keep type I error (α-error) small!

Analogy

	Lawsuit	Hypothesis testing
Strong evidence required	conviction	accept new hypothesis
Null hypothesis H_0	not guilty	old theory true
Alternative hypothesis H_1	guilty	new theory true
Position	plead not guilty **without** strong evidence	keep null hypothesis unless it is very implausible

Further analogy: diagnostics (sensitivity, specificity)

Role of a statistical test

Control the probability of a wrong decision.
Certainty does not exist.

Definition: Level of significance of a test α

$=$ maximal probability of a type I error
$=$ probability to consider a new therapy or theory as better even though the old one is equivalent

Usually $\alpha = 0.05$ is specified

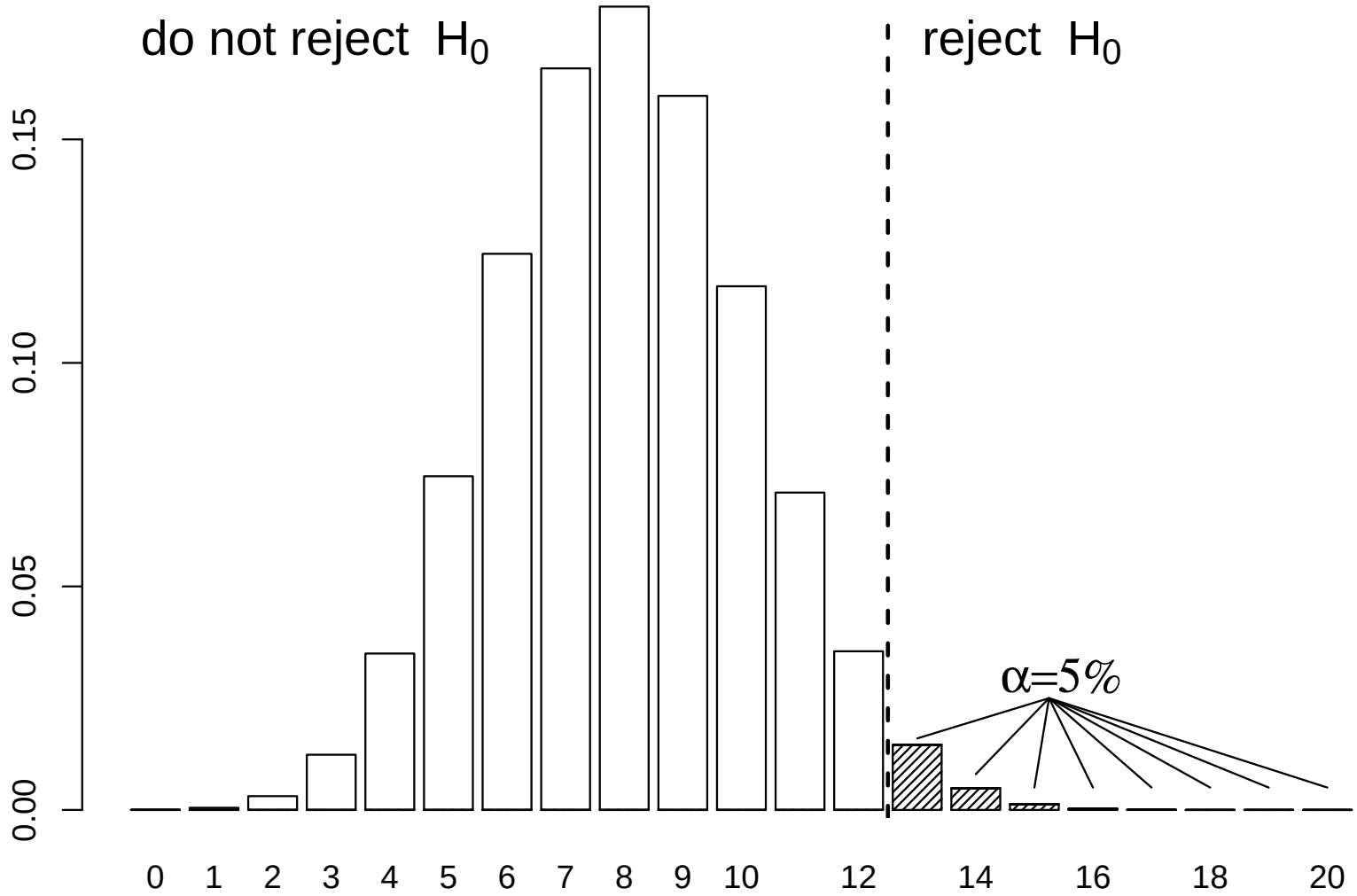
Definition: *p*-value of a test

$p =$ probability, given the null hypothesis is true, of observing a result at least as extreme as the test statistic computed from data.

Illustration with drug example ($\alpha = 5\%$)

- If result $k = 13$
  $\longrightarrow$ $p = \mathrm{P}(k \geq 13) = 0.021$ ("p-value")
  Compare p and α: $p \leq \alpha$
  Decision: Reject H_0, accept H_1
  $\longrightarrow$ "new drug better"
- If result $k = 14$
  $\longrightarrow$ $p = \mathrm{P}(k \geq 14) = 0.006$
  Compare p and α: again $p \leq \alpha$
  Decision: Reject H_0, accept H_1
- If result $k = 12$
  $\longrightarrow$ $p = \mathrm{P}(k \geq 12) = 0.057$
  Compare p and α: $p > \alpha$
  Decision: Do not reject H_0
  $\longrightarrow$ "superiority of the new drug could not be proven"

Illustration with drug example ($\alpha = 5\%$)



Power of a test

Definition: Power of a test $1-\beta$

$= 1-$ probability of a type II error
$=$ probability to prove a new theory which is true

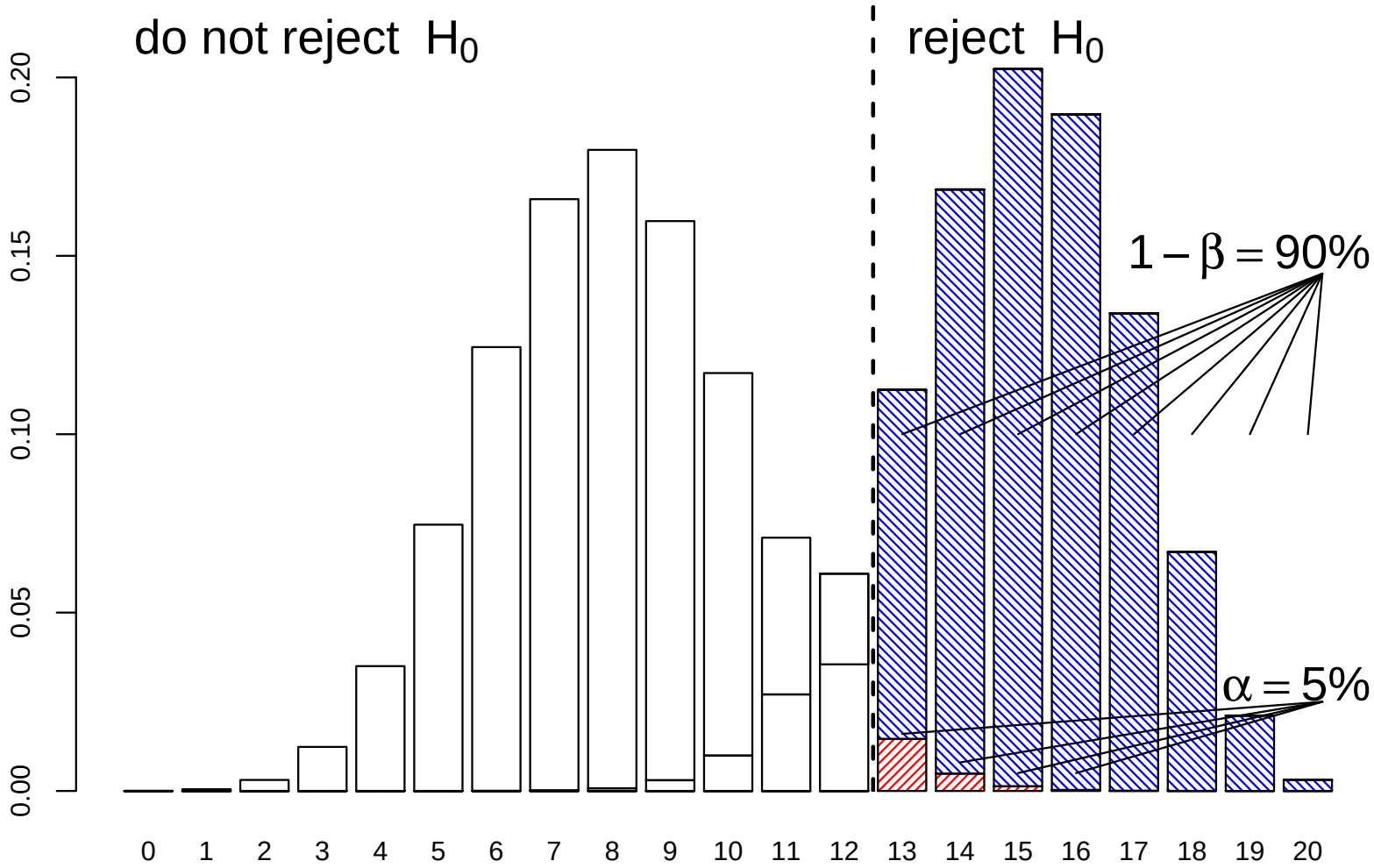
- depends on the sample size n and the **effect size**

In the drug example:

Effect size $= (p_{\text{new}} - 0.4)$
If $p_{new} = 0.4 \longrightarrow 1 - \beta = \mathrm{P}(k \geq 13) = 0.02$

p_{new}	0.4	0.5	0.6	0.7	0.75	0.8	0.9
Power	0.02	0.13	0.42	0.77	0.90	0.97	0.99

Illustration with drug example ($\alpha = 5\%$)



Example for the construction of a test

Question: Is the first learning to walk delayed with cardiac children?

- Norm for first learning to walk

  $\mu_0 = 12$ months (population average)
  $\sigma_0 = 1.8$ months (population variation)

- Scientific hypothesis: Children with congenital heart disease learn to walk later in life (average μ).

  $\mu > \mu_0$ (one-sided hypothesis; otherwise $\mu \neq \mu_0$)

- Statistical (null-) hypothesis:

  $\mu = \mu_0$

- Empirical study with $n = 10, 20, 40, 80$ cardiac children

  Average age for first learning to walk: $\bar{x} = 12.8$ months
  Furthermore, let $\sigma = \sigma_0 = 1.8$ months

Statistical test

1. Is the difference $(\bar{x} - \mu)$ large?
2. Large in relation to the standard error $\sigma_0/\sqrt{n}$

$\longrightarrow$ Test statistic:

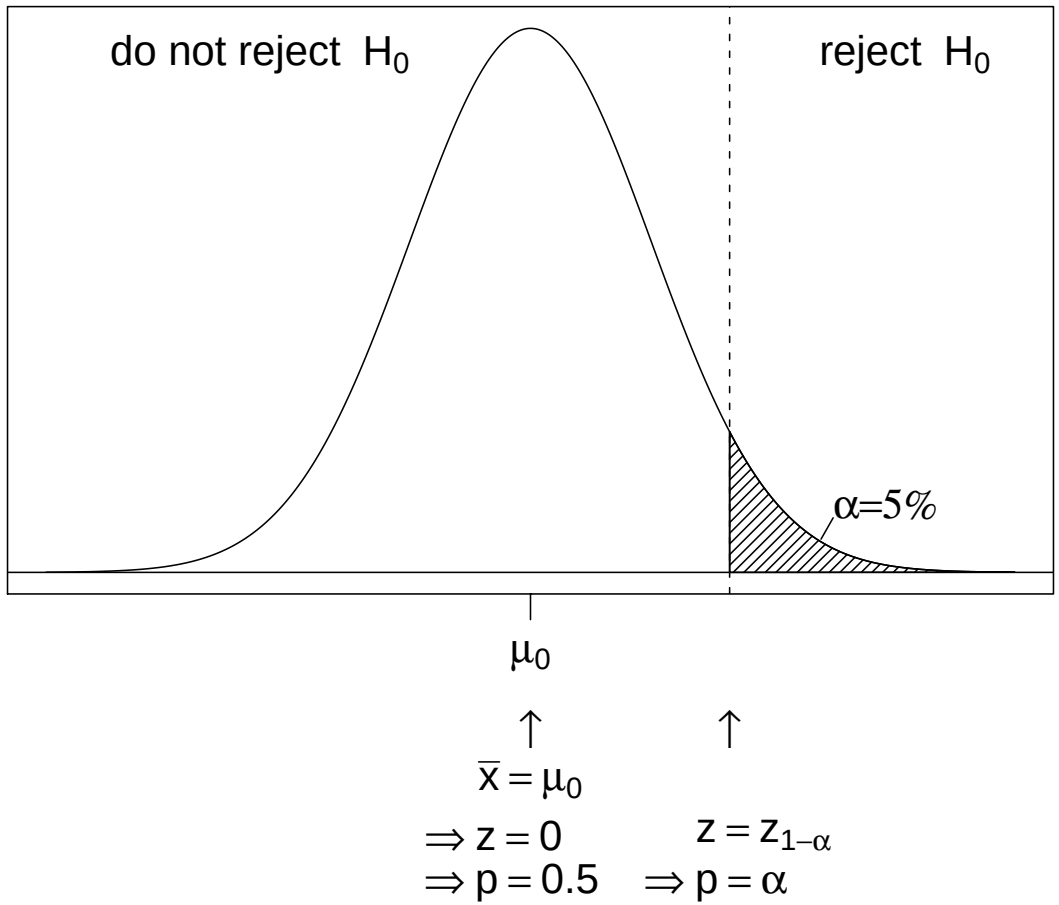
$$z = \frac{\bar{x} - \mu_0}{\sigma_0 \,/\, \sqrt{n}} = \frac{0.8}{1.8 \,/\, \sqrt{20}} = 1.99$$

Assumption: data are normally distributed $\longrightarrow$ z normally distributed

p = probability to obtain by chance (under the null hypothesis)
a value at least as large as z.

Statistical test

p = probability to obtain by chance (under the null hypothesis) a value at least as large as z.



If $\alpha = 5\% \Rightarrow z_{0.95} = 95\%$ percentile of the normal distribution

Statistical test

If $p \leq \alpha$: difference is statistically significant at significance level α

n	10	20	40	80
z	1.41	1.99	2.81	3.98
p	.079	.023	.0025	.0003

z grows with $\sqrt{n}$

Thus: with $\alpha = 0.05$ result significant for $n \geq 20$
with $\alpha = 0.01$ result significant for $n \geq 40$

Thus: Larger n
$\longrightarrow$ significance more likely
$\longrightarrow$ better power

Also: Larger difference $(\mu - \mu_0)$,
smaller σ_0 (better measuring accuracy, homogeneous sample)
$\longrightarrow$ better power

$$\text{effect size} = \frac{(\mu - \mu_0)}{\sigma_0}$$

General procedure for tests of significance

- Formulate hypotheses H_0, H_1
  (related to population characteristics!).
- Decide on a significance level α.
- Define a test statistic $T(x_1, \ldots, x_n)$

  Desired properties:
  - sensitive to H_1
  - distribution of T mathematically computable (under H_0)

  Typical form of T
  - with one-sided alternative hypothesis:

$$T = \frac{\text{observed value} - \text{hypothetical value}}{\text{standard error observed value}}$$

  - with two-sided alternative hypothesis

$$T = \left| \frac{\text{observed value} - \text{hypothetical value}}{\text{standard error observed value}} \right|$$

  Example "Learning to walk": $T = \dfrac{\bar{x} - \mu_0}{\sigma_0 / \sqrt{n}}$.

General procedure for tests of significance

- Compute test statistic for $x_1, \ldots, x_n \longrightarrow T_0$
- Let the distribution $F_T(x)$ of T under the null hypothesis H_0 be known
- Calculate the p-value for the observed T_0

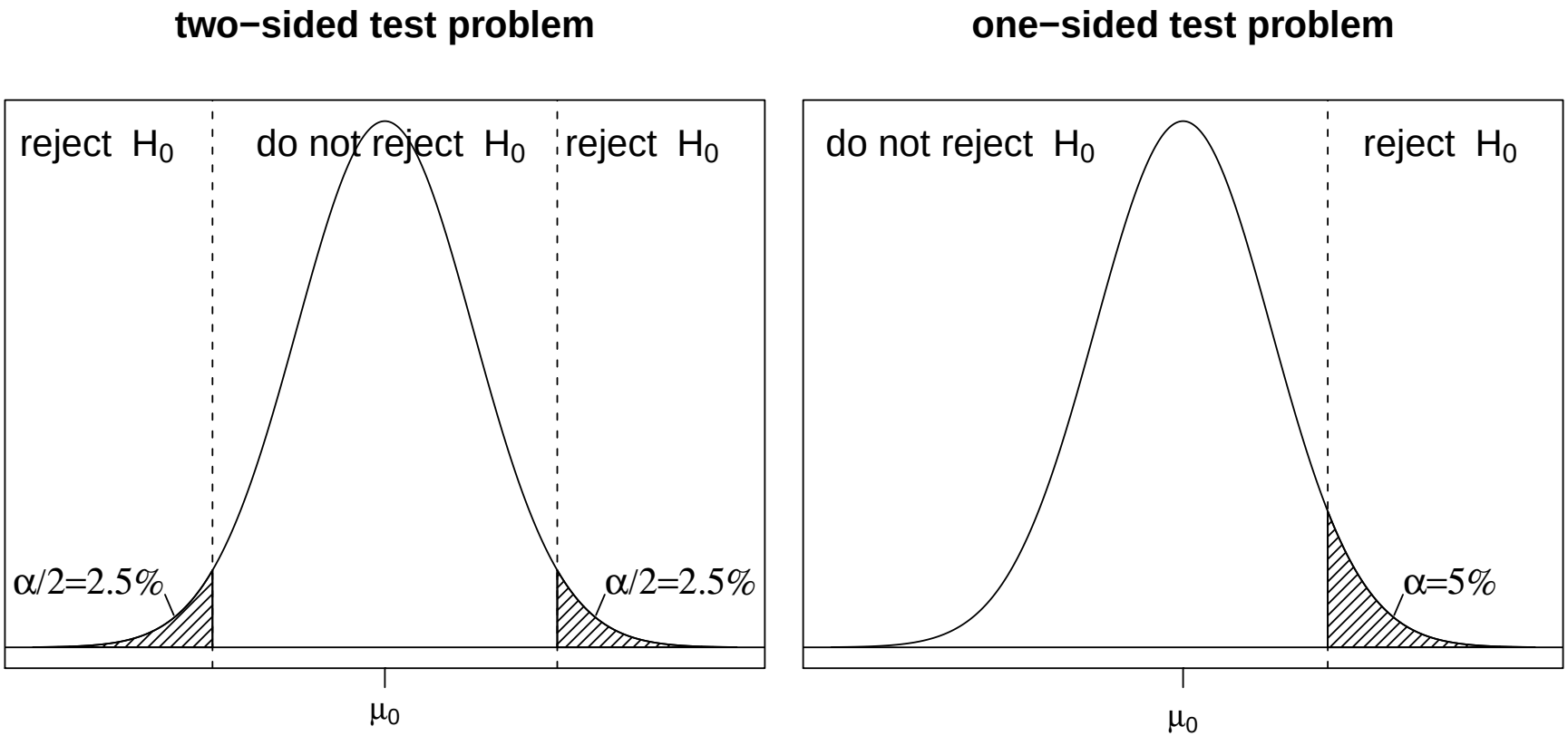
$$p = 1 - F_T(T_0)$$

  Is T_0 likely or unlikely for the null hypothesis?

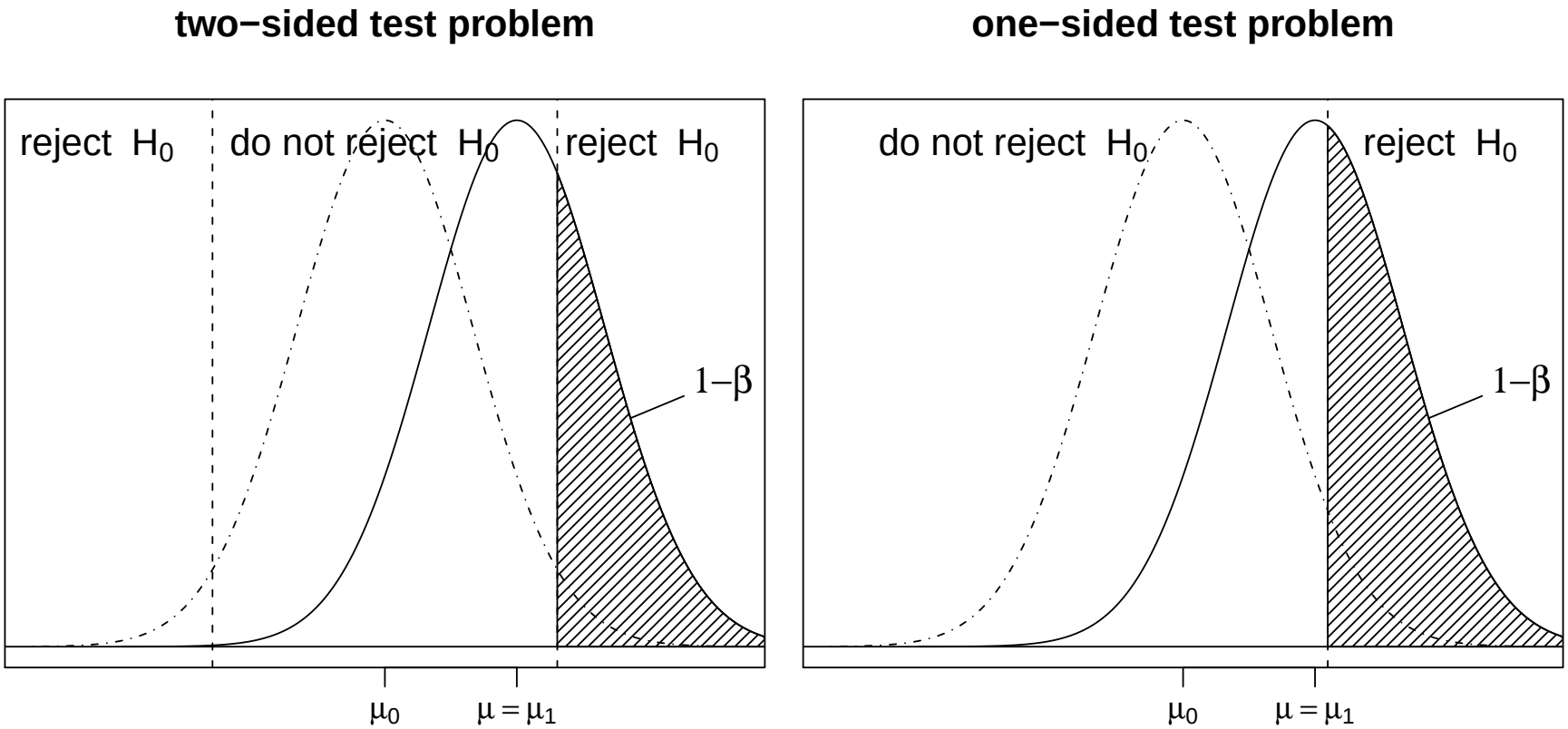
- Decide:
  If $p \leq \alpha \longrightarrow$ reject H_0
  If $p > \alpha \longrightarrow$ do not reject H_0

For different questions $\longrightarrow$ multitude of tests

Type I error


two−sided test problem
reject H0
do not reject H0
reject H0
α/2=2.5%
α/2=2.5%
μ0
one−sided test problem
do not reject H0
reject H0
α=5%
μ0

Type II error



Power of a test

- Optimal tests are defined to have maximal power for predetermined α
  (Example: t–test in the case of a normal distribution).
- The power decreases when α gets smaller ("uncertainty principle": if one error gets smaller, the other error gets larger).
- The power increases when the variability gets smaller.
  This means that homogeneous groups or better measurement techniques are advantageous.
- The power is larger for one-sided tests.
- In an experimental design, the sample size n can be chosen such that e.g. $\beta = 0.20$ or 0.10 is obtained (i.e. given power of 80% or 90%). Thus a clear decision regarding the null or the alternative hypothesis is possible ("power analysis").

♣ Sample size calculation

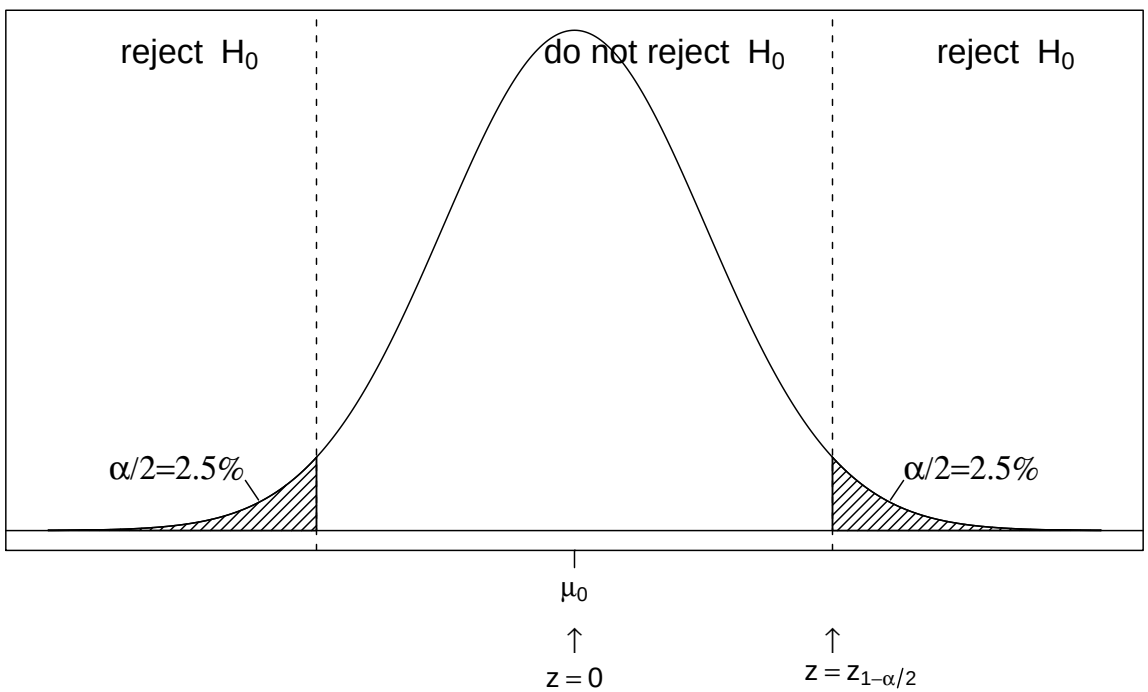
Example: Confirm difference in mean to given μ_0, with known σ_0^2 and independently normally distributed data $x_1, \ldots, x_n$

Test statistic

$$z = \sqrt{n}\,\frac{\bar{x} - \mu_0}{\sigma_0}$$

$\mathrm{H}_0 : \mu = \mu_0 \longrightarrow z \sim \mathcal{N}(0,1)$

$\longrightarrow$ reject H_0, if $|z| > z_{1-\alpha/2}$:



♣ Sample size calculation

If $\mu = \mu_1 > \mu_0$

$$\longrightarrow \quad z = \sqrt{n}\,\frac{\bar{x} - \mu_1}{\sigma_0} + \sqrt{n}\,\frac{\mu_1 - \mu_0}{\sigma_0} \sim \mathcal{N}(\sqrt{n}\delta, 1)$$

Effect size: $\delta = \dfrac{\mu_1 - \mu_0}{\sigma_0}$

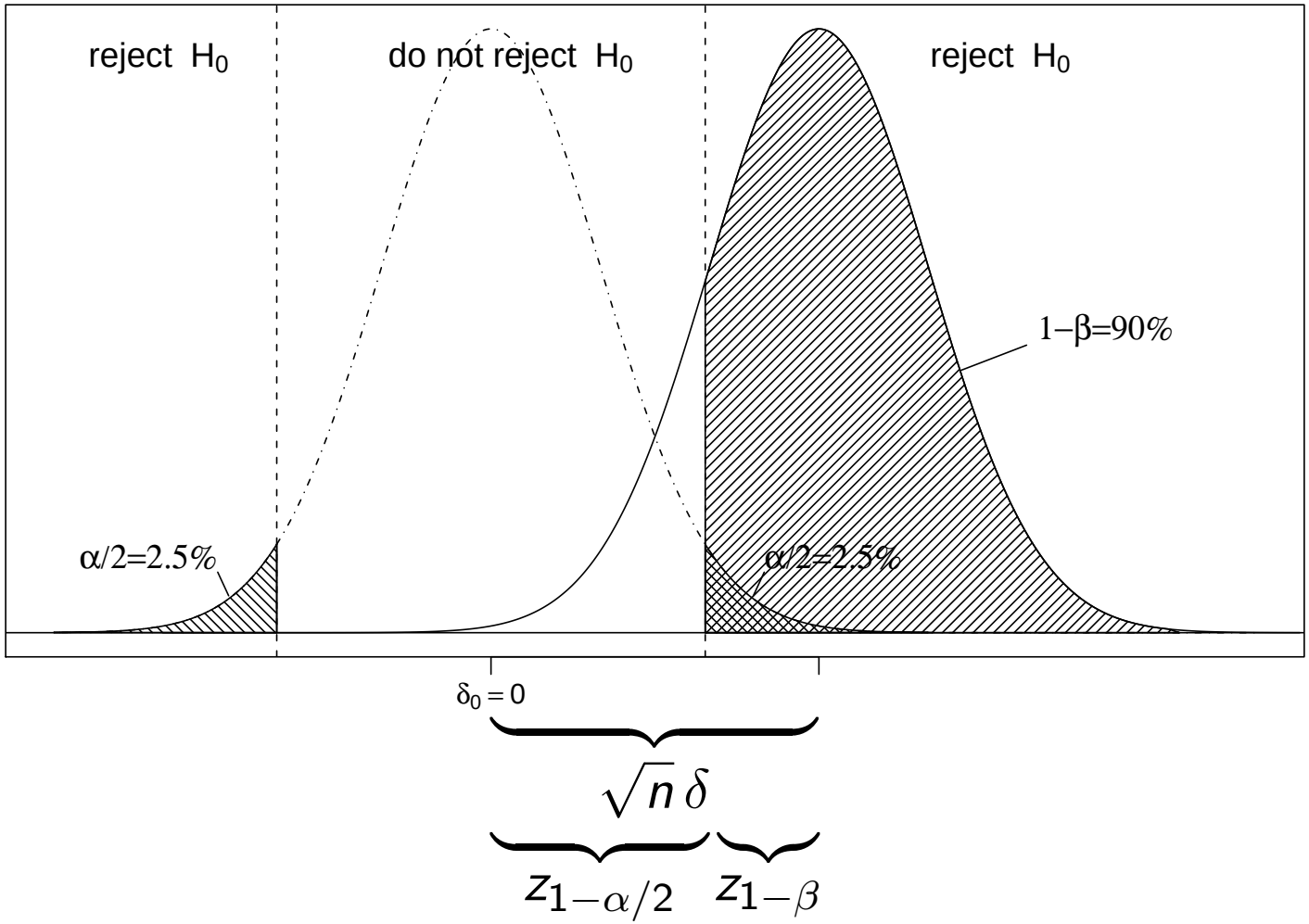
Power $1 - \beta$ is obtained from

$$1 - \beta = \mathrm{P}_1\left[z < z_{\alpha/2}\right] + \mathrm{P}_1\left[z > z_{1-\alpha/2}\right]$$

- left area is negligible

♣ Sample size calculation

- left area is negligible



$$\longrightarrow \quad \sqrt{n}\delta = z_{1-\alpha/2} + z_{1-\beta} \quad \longrightarrow \quad n = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\delta^2}$$

♣ Sample size calculation

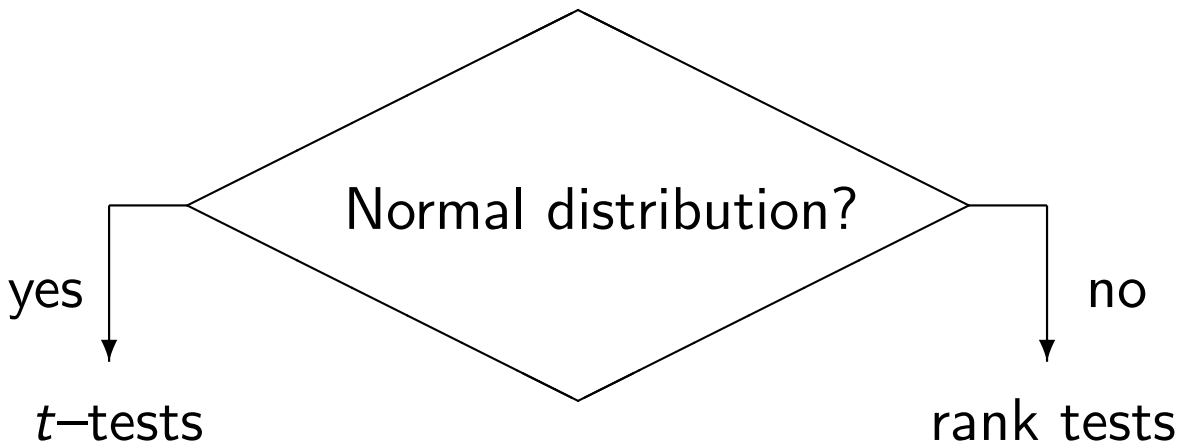
$$n = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\delta^2}$$

- $n \propto \sigma_0^2$
- $n \propto 1/(\mu_1 - \mu_0)^2$
- n grows with decrease of α (non-linear)
- n grows with $1 - \beta$ (non-linear)

Testing differences between means

Comparison of means

1. Comparison with known value (one-sample test)
2. Comparison of 2 **independent** samples (unpaired two-sample test)
3. Comparison of **paired** samples (paired two-sample test)



Possibly transformation necessary!

Haemodilution tolerance in patients with mitral regurgitation

D. R. Spahn,[1] B. Seifert,[2] T. Pasch[1] and E. R. Schmid[1]

1 Institute of Anaesthesiology, University Hospital, University of Zürich, Rämistrasse 100, CH-8091 Zürich, Switzerland
2 Department of Biostatistics, University of Zürich, CH-8091 Zürich, Switzerland

Summary

Haemodynamic parameters and oxygen consumption were determined in 20 patients with mitral regurgitation before and after a $12\,ml.kg^{-1}$ isovolaemic exchange of blood for 6% hydroxyethyl starch. During haemodilution, mean (SEM) haemoglobin concentration decreased from 13.0 (0.4) to 10.3 (0.4) $g.dl^{-1}$ ($p = 0.001$). With cardiac filling pressures maintained at predilution levels, cardiac index increased from 1.84 (0.08) to 1.94 (0.08) $l.min^{-1}.m^{-2}$ ($p = 0.025$) while systemic vascular resistance decreased from 1556 (86) to 1425 (83) $dyne.s.cm^{-5}$ ($p = 0.002$) and oxygen extraction increased from 31.7 (1.1) to 37.3 (1.4)% ($p = 0.001$) resulting in an unchanged oxygen consumption. The haemodynamic response to haemodilution was not affected by the patients' cardiac rhythm, i.e. whether it was sinus rhythm or atrial fibrillation. In conclusion, isovolaemic haemodilution to a haemoglobin of $10.3\,g.dl^{-1}$ is well tolerated in patients with mitral regurgitation. Compensatory mechanisms include both an increase in cardiac index and an increase in oxygen extraction.

Anaesthesia, 1998, **53**, p. 20–24

Haemodilution tolerance in patients with mitral regurgitation

Changes during haemodilution were analysed using paired t-tests. Patients were divided into two groups for analysis: those patients in sinus rhythm ($n = 10$) and those in atrial fibrillation ($n = 10$). Patient characteristics between these two groups were compared using unpaired t-tests. The effect of sinus rhythm and atrial fibrillation on changes due to haemodilution were analysed using repeated measures analysis of variance. Fisher's exact test was used to compare frequencies between patients in sinus rhythm and patients in atrial fibrillation. A probability value of less than 0.05 was considered to be statistically significant. Data are presented as mean (SEM).

Haemodilution tolerance in patients with mitral regurgitation

Table 1 Demographic and pre-operative data. Values are given as mean (SEM) where appropriate.

	All patients	Patients in sinus rhythm	Patients in atrial fibrillation	p value
Number	20	10	10	
Age; years	63.1 (2.7)	61.5 (3.4)	64.7 (4.4)	0.572
Weight; kg	69.7 (2.5)	70.9 (3.7)	68.4 (3.7)	0.635
Height; cm	170 (2)	171 (3)	169 (4)	0.682
Body surface area; m^2	1.8 (0.1)	1.8 (0.1)	1.8 (0.1)	0.569
Sex ratio; F: M	5: 15	2: 8	3: 7	0.652
ASA Grade; III: IV	2: 18	0: 10	2: 8	0.237
Left ventricular ejection fraction; %	61.3 (2.5)	63.0 (3.4)	59.6 (3.9)	0.516
Left ventricular end-diastolic pressure; mmHg	9.4 (1.3)	10.0 (1.7)	8.7 (5.7)	0.640
Pre-operative haemoglobin; $g.dl^{-1}$	14.2 (0.3)	14.4 (0.4)	14.1 (0.6)	0.686
Cardiac medication				
Diuretics; *n*	14	6	8	0.629
ACE inhibitors; *n*	12	5	7	0.410
Digoxin; *n*	10	3	7	0.101
β-blockers; *n*	6	3	3	0.999
Amiodarone; *n*	1	1	0	0.500
Calcium channel blocker; *n*	1	0	1	0.500
Nitrates; *n*	1	0	1	0.500

ACE: angiotensin converting enzyme.

One-sample t–test

Statistical comparison of a mean $\bar{x}$ with a hypothetical value μ_0.

Example: Learning to walk of babies

$x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma^2)$
$\mathrm{H}_0 : \mu = \mu_0$

Up to now $\sigma^2 = \sigma_0^2$ known.

$$\longrightarrow z = \frac{\bar{x} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}}$$

Under the null hypothesis normally distributed $\mathcal{N}(0, 1)$

If σ^2 is unknown?

Replace $\sigma \longrightarrow s$

One-sample t–test

Test statistic: **one-sample t–test**

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

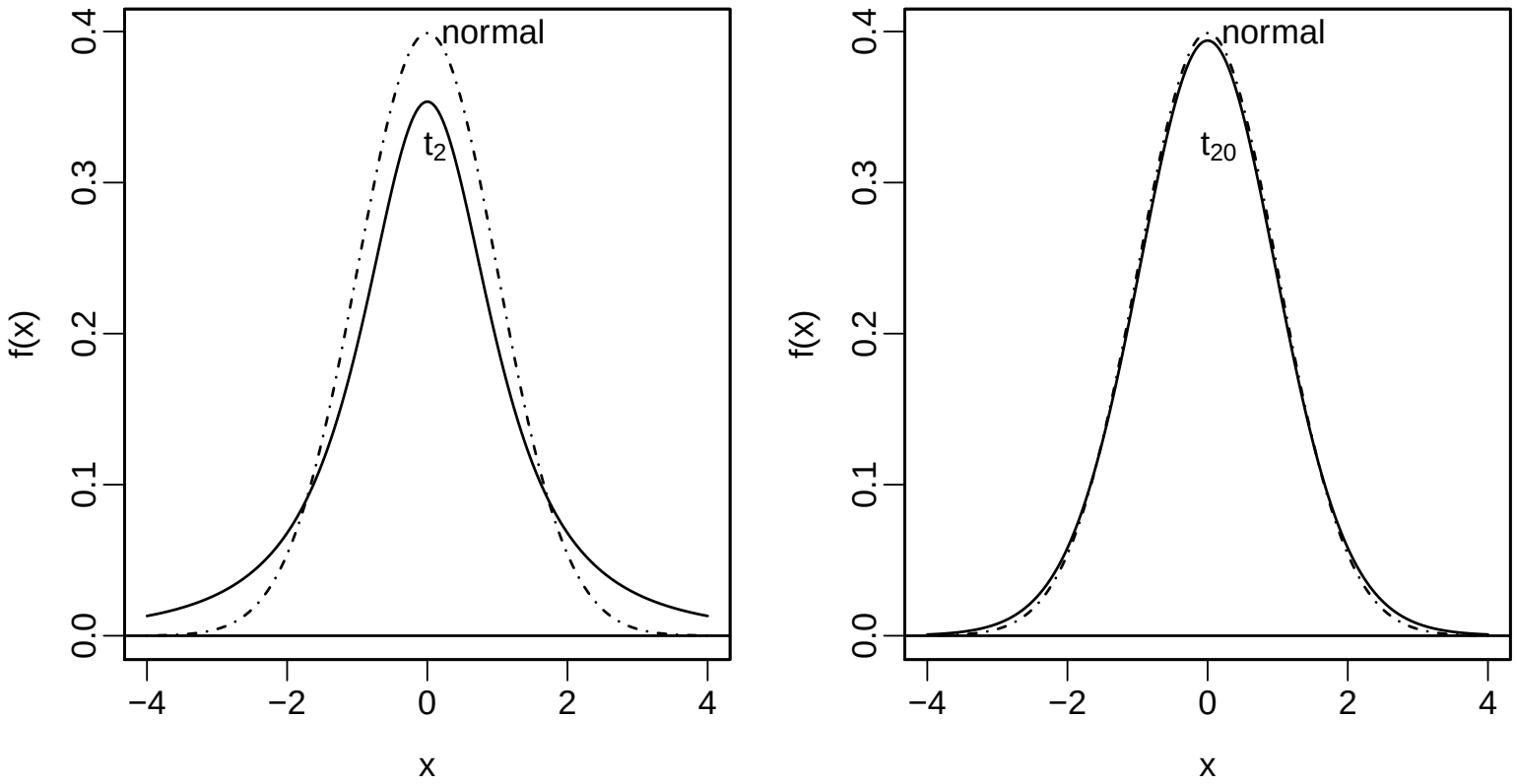
- Test statistic t is t–distributed with $(n-1)$ degrees of freedom

Definition: t–**distribution**

$X_1, \ldots, X_n$ independent $\mathcal{N}(0, 1)$

$$\longrightarrow t = \frac{\bar{x}}{\frac{s}{\sqrt{n}}}$$ t–distributed with $(n-1)$ degrees of freedom

Comparison $t \longrightarrow \mathcal{N}$



- s not fixed $\longrightarrow$ more probability "outside"

0.975–quantiles of the t_n–distribution:

n	5	10	15	20	30	60	120	∞
$t_{.975}$	2.78	2.26	2.14	2.09	2.05	2.00	1.98	1.96

One-sample t–test

Assumption: $s = 1.8$

n	10	20	40	80
t	1.41	1.99	2.81	3.98
p	.096	.031	.0039	.0008

Thus: with $\alpha = 0.05$ result significant for $n \geq 20$
with $\alpha = 0.01$ result significant for $n \geq 40$

p–values are a little larger than with z–test

Two-sample t–test

Statistical comparison of the means in two groups.

Example: Comparison of the logarithmised number of T_4–cells for Hodgkin– and non-Hodgkin–patients

Group 1 (Hodgkin):	$n = 20, \bar{x} = 6.49, s_x = 0.71$
Group 2 (non–Hodgkin):	$m = 20, \bar{y} = 6.09, s_y = 0.63$

Scientific hypothesis: number of T_4–cells with Hodgkin raised also after remission

Null hypothesis:

$$\mathrm{H}_0 : \mu_x = \mu_y \longrightarrow \mu_x - \mu_y = 0$$

Scientific or alternative hypothesis:

$$\mathrm{H}_1 : \mu_x > \mu_y \text{ (one-sided)}$$
$$\mu_x \neq \mu_y \text{ (two-sided)}$$

Construction of the test statistic

Observed $-$ Expected under $\mathrm{H}_0 = \bar{x} - \bar{y} = 0.4$

Large or close to 0 ?

Divide by standard error of the difference:

$$\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$$

$\sigma \longrightarrow s$, since σ unknown

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$$

Estimated standard error made up from both samples

Two-sample t–test

Test statistic: **two-sample t–test**

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Assumptions:

1. independent, **normally distributed** quantities
   $x_1, \ldots, x_n,\ y_1, \ldots, y_m$
2. equal variance in both populations: $\sigma_x^2 = \sigma_y^2$

Then: The test statistic t is t–distributed with $n + m - 2$ degrees of freedom.

Example: log T_4–cells, $s = 0.67$

$$t = \frac{0.4}{0.67\sqrt{\frac{1}{20} + \frac{1}{20}}} = 1.88$$

P($t \geq 1.88$) = one-sided $p = 0.034$

$\longrightarrow$ $p \leq \alpha = 0.05$

$\longrightarrow$ Hodgkin–patients have a significantly larger number of T_4–cells

P($t \leq -1.88$ oder $t \geq 1.88$) = two-sided $p = 0.068$

$\longrightarrow$ $p > \alpha = 0.05$

$\longrightarrow$ no significant difference with two-sided test

General: One-sided tests have more power $(1 - \beta)$

Paired two-sample t–test

Up to now: Comparison of two independent samples.

Examples for the use of paired samples:

- pre-post-comparisons of therapy studies
- repeated measurements for the same patient
- comparison of EEG for left and right brain hemisphere

Example: heart rate of $n = 8$ diabetic patients with poor or good metabolic control

$$H_0: \quad \mu_x = \mu_y$$
$$H_1: \quad \mu_x > \mu_y$$

$x_1, \ldots, x_n$: data at time-point 1
$y_1, \ldots, y_n$: data at time-point 2

Scientific question (H_1): Do values improve with good metabolic control (good compliance)?

Increased myocardial contractility in short-term Type 1 diabetic patients: an echocardiographic study

L. Thuesen, J. Sandahl Christiansen, N. Falstie-Jensen, C. K. Christensen, K. Hermansen, C. E. Mogensen and P. Henningsen

II University Clinic of Internal Medicine and University Department of Cardiology, Aarhus Kommunehospital, Aarhus, Denmark

Summary. Cardiac function was investigated by echocardiography in 24 short-term Type 1 diabetic patients with a mean diabetes duration of 7 years (range 4–14 years) during conditions of ordinary metabolic control. Compared to 24 age and sex matched normal control subjects, measurements of myocardial contractility as left ventricular fractional shortening and mean circumferential shortening velocity were increased by 12% and 20% respectively. Another 8 Type 1 diabetic patients were examined during conditions of poor (hyperglycaemia and ketosis) and good metabolic control. Following improved glycaemic control, left ventricular fractional shortening and mean circumferential shortening velocity decreased by 16% and 24% respectively. Our findings show that short-term Type 1 diabetes is associated with increased myocardial contractility. Furthermore, this condition is related to the state of metabolic control.

Key words: Echocardiography, left ventricular function, Type 1 diabetes, metabolic control, diabetic cardiopathy.

Diabetologia, 1985, **28**, p. 822–826

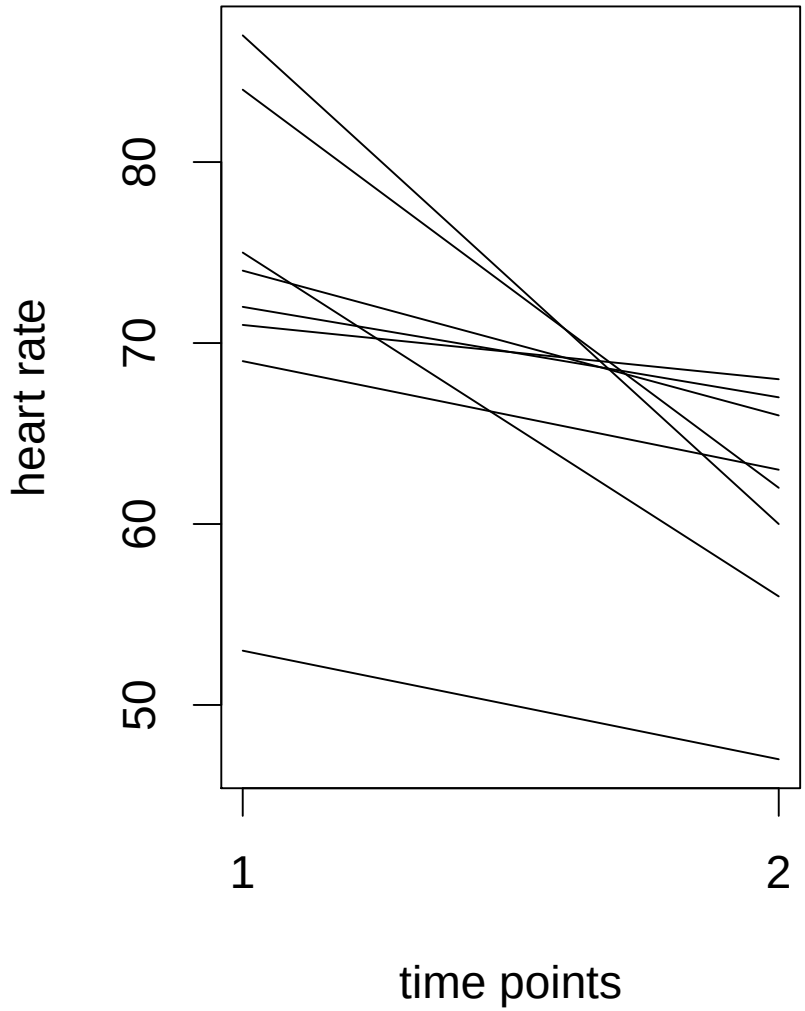
Example: heart rate of diabetic patients

Improvements: $d_i = y_i - x_i$

$\mathrm{H}_0 : \delta = \mu_y - \mu_x = 0$

$\mathrm{H}_1 : \delta < 0$

Id	x	y	d
1	74	66	-8
2	72	67	-5
3	84	62	-22
4	53	47	-6
5	75	56	-19
6	87	60	-27
7	69	63	-6
8	71	68	-3
mean	73	61	-12
s	10	7	9.2



Mean difference large? – large with respect to standard error?

Example: heart rate of diabetic patients

Test statistic: **paired two-sample t–test**

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

Assumption: d_i normally distributed

$\longrightarrow$ t is t–distributed with $(n-1)$ degrees of freedom, if H_0 valid.

$$t = \frac{-12}{9.2/\sqrt{8}} = -3.7$$

If H_0 valid: t–distributed with 7 degrees of freedom

$\longrightarrow$ P($t \leq -3.7$ or $t \geq 3.7$) = two-sided $p = 0.008$

$\longrightarrow$ $p < 0.05$

$\longrightarrow$ Improvement significant with good "compliance"

The use of the usual two-sample t–test would be **wrong** (**not independent!**)

Rank tests: Mann-Whitney and Wilcoxon test

- Testing without normal distribution
- Idea: Use only the ranking of the data,
  similar to median, interquartile range, etc.
- Comparison of 2 independent groups
  (analogy to two-sample t–test):
  Mann-Whitney U test (Wilcoxon rank-sum test)
- Pre-post comparisons (analogy to paired two-sample t–test):
  Wilcoxon matched pairs test (Wilcoxon signed-rank test)

Rank tests: Mann-Whitney and Wilcoxon test

Pros:

- valid without assumption of normality
- robust towards outliers and extreme data
- applicable to ordinal data
- good power also with normality (efficiency 96%)

Cons:

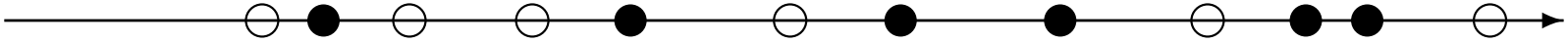
- not applicable to complex problems
- problematic with small sample sizes

Example: rank tests

Compare two independent samples $x_1, \ldots, x_6$ and $y_1, \ldots, y_6$ in a joint ranking:

Situation 1: $\mu_x \approx \mu_y$

Data	x	y	x	x	y	x	y	y	x	y	y	x
Ranks	1	2	3	4	5	6	7	8	9	10	11	12



The average rank of the x_i is 5.8
The average rank of the y_i is 7.2

Situation 2: $\mu_y > \mu_x$

Data	x	x	y	x	x	y	x	x	y	y	y	y
Ranks	1	2	3	4	5	6	7	8	9	10	11	12

The average rank of the x_i is 4.5
The average rank of the y_i is 8.5

Large discrepancy !

Procedure for **Mann-Whitney U test**

1. Build a **joint** ranking of $x_1, \ldots, x_n$, $y_1, \ldots, y_m$
2. Compute separate average ranks or rank sums R_x, R_y
3. Compute $U_x = nm + \frac{n(n+1)}{2} - R_x$ as well as U_y
4. Choose the smaller value of U_x, U_y as test statistic ("U–test")
5. tabulated p–values, approximately $\mathcal{N}$ if $n, m > 10$

Example: T_4–cells with Hodgkin and non-Hodgkin patients

Number of T_4–cells not normally distributed!

Here is a selection of the ranked numbers:

Group	nH	nH	H	nH	nH	H	$\cdots$
T_4–cells	116	151	171	192	208	257	$\cdots$
Rank	1	2	3	4	5	6	$\cdots$

Non-Hodgkin strongly represented by small numbers!

One obtains rank sums: $R_H = 475 \quad R_{nH} = 345$

and: $U_H = 20 \times 20 + \dfrac{20 \times 21}{2} - 475 = 135 \quad U_{nH} = 265$

Thus $U = U_H$ is our test statistic

$\alpha = 5\%$ (one-sided)

$\longrightarrow$ Reject null hypothesis, if $U \leq 138$

$\longrightarrow$ Deviation significant, p–value $= 4.0\%$

Example: T_4–cells with Hodgkin and non-Hodgkin patients

Approximation by $\mathcal{N}$:

$$\mathrm{E}[U] = \frac{nm}{2} \qquad \mathrm{Var}(U) = \frac{1}{12} nm(n+m+1)$$

$$\frac{U - \mathrm{E}[U]}{\sqrt{\mathrm{Var}(U)}} = -1.758$$

$\longrightarrow$ Reject null hypothesis, if value $< z_\alpha = -1.64$
$\longrightarrow$ Deviation significant, p–value $= 3.8\%$

Tests of proportions or probabilities

One wants to statistically compare

1. an observed frequency with a hypothetical frequency

   Example: Observed frequency of male newborns

   $p_{\text{obs}} = \frac{n_1}{n} = 0.51$

   Did deviation from the hypothetical value 0.5 occur by chance?

2. two observed frequencies

   Example: Is the prevalence of stomach cancer in Japan significantly higher than in Europe?

One sample situation

Let p_0 be a known probability.

$$\mathrm{H}_0\colon\ p = p_0, \qquad \mathrm{H}_1\colon \begin{array}{lll} p & \begin{matrix} > \\ < \end{matrix} & p_0 \quad \text{one-sided} \\ p & \neq & p_0 \quad \text{two-sided} \end{array}$$

Test via binomial distribution.

Examples:

1. Treatment with standard drug cures 40% ($p_0 = 0.4$)
   New drug $p_{\text{new}} > p_0$?
2. Male newborns $p = 0.5\,(= p_0)$ or $p \neq p_0$?

(Unpaired) two sample situation

Comparison of two empirical relative frequencies $\hat{p}_x$, $\hat{p}_y$

Example: Isolation of influenza antibodies in
34 out of 113 tested boys and 54 out of 139 tested girls.
Gender-related difference?

$$\mathrm{H}_1 : p_y \neq p_x$$

Test statistic:

- One can directly compare the empirical proportions
  $\hat{p}_x\,(= 34/113)$ and $\hat{p}_y\,(= 54/139)$
  (appropriate standardization, approximate normal distribution)
- More simple is the test of homogeneity in a 2×2 table
  via a χ^2–test.

Pay attention: Paired samples have to be tested differently!
(Example: Frequency of pain before and after treatment,
same patient)
$\longrightarrow$ McNemar test

The χ^2–test

Suited for answering various questions in the case of categorical data.

Example: Comparison of drug A with drug B in $n = 150$ patients.
Clinical evaluation of the state of health: very good, good, poor

Data:

name	drug	clin_eval
A.A.	A	good
C.A.	A	poor
M.C.	A	very good
...	...	...
R.B.	B	good
B.C.	B	good
M.F.	B	very good
...	...	...

The χ^2–test

Contingency table (frequency table, cross-table)

	very good	good	poor	n
A	37	24	19	80
B	17	33	20	70
Total	54	57	39	150

- 2×3 "cells"
- observed cell frequency = number per cell
- cell (A, good) contains the number of patients treated with drug A, and whose state of health was evaluated good

χ^2 goodness-of-fit test

Aim: testing the distribution of categorical data.

Example: Genotypes A, B and C with **model-based** relative frequencies $1/4, 1/2, 1/4$.

100 plants are grown:

3 cells:

A	B	C
18	55	27

Question: In agreement with model?

$$\mathrm{H}_0 : p_A = 1/4, \quad p_B = 1/2, \quad p_C = 1/4$$
$$\longrightarrow \text{ expected frequencies } 25, 50, 25$$

χ^2 goodness-of-fit test

Idea: Compare observed (Obs) and expected (Exp) cell frequencies:

$$(18-25)^2+(55-50)^2+(27-25)^2$$

$$X^2=\frac{(18-25)^2}{25}+\frac{(55-50)^2}{50}+\frac{(27-25)^2}{25}=2.62$$

General:

- k cells with n_i observations and hypothetical probabilities $p_i(0)$ $(i=1,\ldots,k)$
- H_0: $p_1=p_1(0),\ldots,p_k=p_k(0)$

Test statistic: χ^2 goodness-of-fit test

$$X^2=\sum_{i=1}^{k}\frac{(n_i-n\,p_i(0))^2}{n\,p_i(0)}=\sum_{\text{cells}}\frac{(\text{Obs}-\text{Exp})^2}{\text{Exp}}$$

χ^2 goodness-of-fit test

Test distribution:
χ^2–distribution with $(k-1)$ degrees of freedom (approximate!)

Example: $X^2 = 2.62 \longrightarrow \chi^2_2$ distributed
5% quantile of χ^2_2: 5.99
$\longrightarrow$ not significant, as $2.62 < 5.99$

Here: Testing the goodness-of-fit of discrete probabilities for categorical data

Similar approach for continuous variables:
Classify data and compare the observed relative frequencies with the respective hypothetical values for the classes.

Application: Goodness-of-fit tests for distributions

Testing for differences in contingency tables
χ^2 test of homogeneity

Aim: Comparison of the empirical frequencies of two or more groups

Example: Comparison of drug A with drug B in $n = 150$ patients.
Clinical evaluation of the state of health: very good, good, poor

80 patients randomized to receive drug A
70 patients randomized to receive drug B

Contingency table (frequency table, cross-table)

	very good	good	poor	n
A	37	24	19	80
B	17	33	20	70
Total	54	57	39	150

Alternative hypothesis H_1: Effects of drug A and B are different

Null hypothesis H_0: Both A and B have the same effect, i.e.

$$p_{A_1} = p_{B_1} = p_1, \quad p_{A_2} = p_{B_2} = p_2, \quad p_{A_3} = p_{B_3} = p_3$$

χ^2 test of homogeneity

Remark: problem similar to two-sample problem with continuous data

Testing principle: Compare in each cell the number of observed to the number of expected

Test statistic: χ^2 test of homogeneity

$$X^2 = \sum_{\text{cells}} \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$$

Example: health status – drug A vs. drug B

	very good	good	poor	n
A	37 (28.8)	24 (30.4)	19 (20.8)	80
B	17 (25.2)	33 (26.6)	20 (18.2)	70
Total	54	57	39	150

( ) expected, if homogeneous, no differences between groups

$\longrightarrow$ Test statistic $X^2 = 8.22 \sim \chi^2_2$
$p = \mathrm{P}(X^2 \geq 8.22) = 0.016 < 0.05$

$\longrightarrow$ A significantly different from B with significance level $\alpha = 0.05$

χ^2 test of homogeneity

Keep in mind:

- p–values are only approximately valid (depending on n)
  $\longrightarrow$ **Fisher's exact test**

- If A, B paired ("pre-post comparisons")
  $\longrightarrow$ **McNemar test**

  Since: if post = pre

  $\longrightarrow$ pre

	post	
	n_1	0
	0	n_2

  not homogeneous $\longrightarrow$ significant
  But: no improvement

General formulation: $r \times c$ table

Test distribution: χ^2 with $(r-1)(c-1)$ degrees of freedom

r = number of rows in cross-table

c = number of columns in cross-table

$(r \times c)$–contingency table

	1	...	c	
1	n_{11}	...	n_{1c}	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	n_{r1}	...	n_{rc}	$n_{r.}$
	$n_{.1}$	...	$n_{.c}$	$n_{..}$

$n_{i.}, n_{.j}$ = marginal sums, $n_{..} = n$

$$\text{Obs}(i,j) = n_{ij}; \qquad \text{Exp}(i,j) = \frac{n_{i.} n_{.j}}{n}$$

General null hypothesis of homogeneity:
probabilities (across all c columns) in all r rows identical.

Test of independence of two variables

(For continuous data: see test of correlation)

Problem: Two discrete variables are surveyed in a sample of size n and tabulated in a contingency table. Are the variables independent?

$\mathrm{H}_0 : p_{ij} = p_i p_j$ for all i, j

Example: The handedness of 400 children and the handedness of their parents is determined.

Scientific hypothesis H_1: Handedness is genetically passed down, i.e. $p_{ij} \neq p_i\, p_j$.

	Handedness child		
Father × mother	right	left	total
right, right	303 (295.8)	37 (44.2)	340
right, left	29 (33.1)	9 (4.9)	38
left, left	16 (19.1)	6 (2.9)	22
total	348	52	400

( ) = expected, if independent

Test of independence of two variables

H_0 : no dependency ("not genetically passed down")

Test: Formally identical to test of homogeneity,
test statistic X^2 also $\chi^2_{(r-1)\,(c-1)}$ distributed.

In the example:

$X^2 = 9.15, \quad p = \mathrm{P}(\chi^2_2 \geq 9.15) = 0.010 < \alpha$
$\longrightarrow$ reject H_0
$\longrightarrow$ handedness is to some extent genetically passed down.

Multiple testing

A statistical test is valid (i.e. significance level correct) for **one** statistical hypothesis. For multiple hypotheses the significance level increases.

Example:

- Study with 4 diagnostic groups
- 20 variables surveyed.

$\longrightarrow$ 120($= 60 \times 20$) pairwise comparisons possible

$\longrightarrow$ 120 statistical tests possible

H_0: No difference at all

H_1: Difference in at least one variable $\alpha = 0.05$

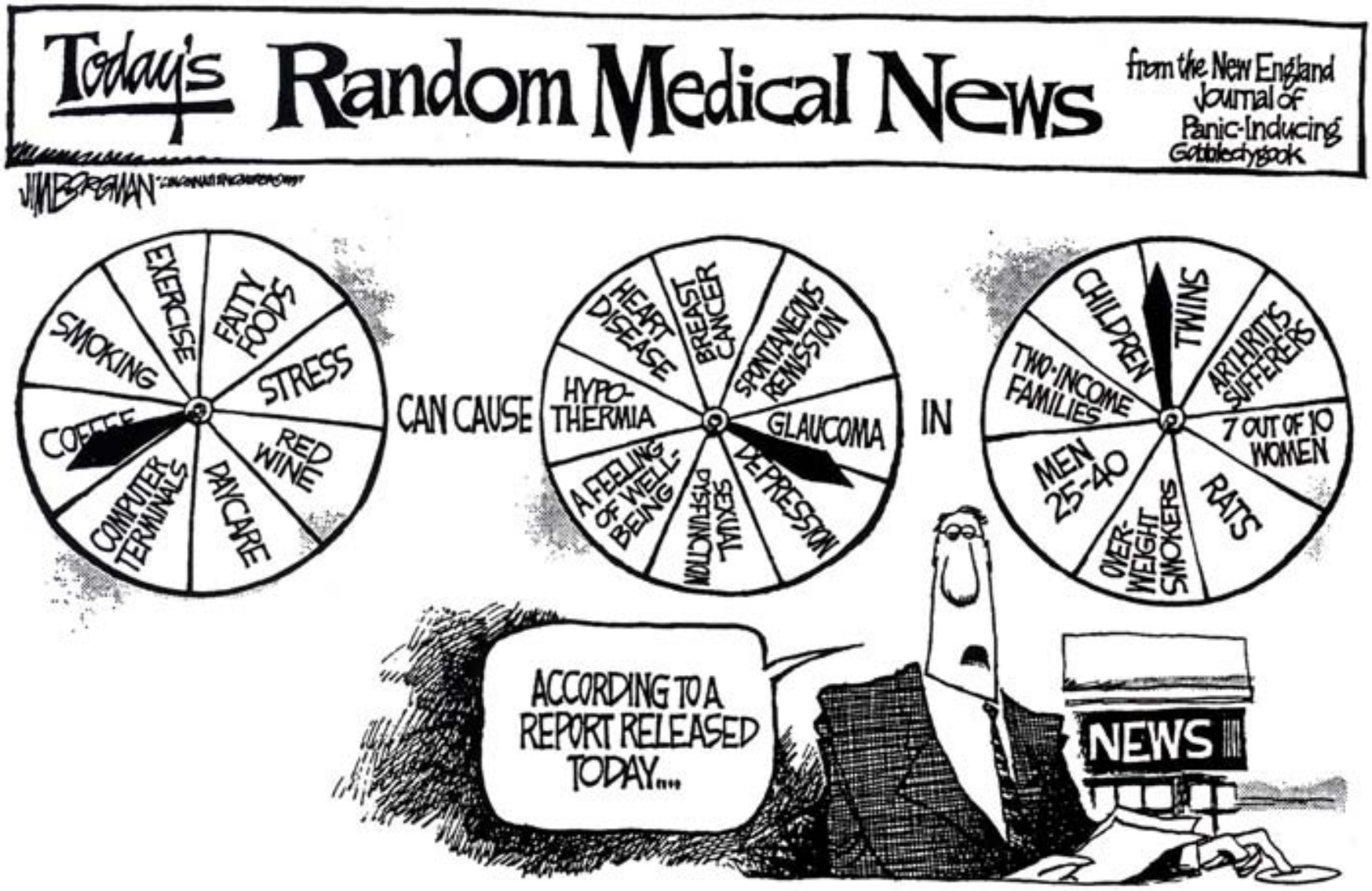
If H_0 is valid: Nevertheless $0.05 \times 120 = 6$ rejections on average.

Multiple testing

In general: k tests on nominal 5% level

k	nominal α	effective α
1	0.05	0.05
2	0.05	0.10
3	0.05	0.14
5	0.05	0.23
10	0.05	0.40
20	0.05	0.64
50	0.05	0.92

Inflates α-error!


Today's Random Medical News
from the New England Journal of Panic-Inducing Gobbledygook
JIM BORGMAN
SMOKING
EXERCISE
FATTY FOODS
STRESS
RED WINE
DAYCARE
COMPUTER TERMINALS
COFFEE
CAN CAUSE
HEART DISEASE
BREAST CANCER
SPONTANEOUS REMISSION
GLAUCOMA
DEPRESSION
SEXUAL DYSFUNCTION
A FEELING OF WELL-BEING
HYPO-THERMIA
IN
CHILDREN
TWINS
ARTHRITIS SUFFERERS
7 OUT OF 10 WOMEN
RATS
OVER-WEIGHT SMOKERS
MEN 25-40
TWO-INCOME FAMILIES
ACCORDING TO A REPORT RELEASED TODAY...
NEWS

Multiple testing

Solutions:

(a) multivariate statistical methods, for example variance analysis (overall-α)

(b) Bonferroni-correction (for small k!)

$$\text{Bonferroni inequality: } \mathrm{P}\left(\sum_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} \mathrm{P}(A_i)$$

$\sum_{i=1}^{k} A_i$ = any rejection of H_0 in k tests

$$\mathrm{P}\left(\sum_{i=1}^{k} A_i\right) = 0.05 \leq k\ \mathrm{P}[A_i]$$

$$\longrightarrow \quad \boxed{\mathrm{P}(\text{single test}) = \frac{0.05}{k}}$$

is conservative

(c) Design of experiments

$\longrightarrow$ few stringent hypotheses for testing, if not analyse hypotheses descriptively.

Confidence interval (credibility region)

- When repeating a study we get different statistical quantities. This can be explained by the different samples, which necessarily lead to a random effect. There is need to quantify this random effect in the statistical quantities.
- Since the true quantity θ (for example $\theta = \mu, p$) is unknown and the estimation contains a statistical inaccuracy: Exists an interval which contains θ with high probability? ("Quantification of the inaccuracy")

Definition:
The 95%–**confidence interval** $[\hat{\theta}_l, \hat{\theta}_u]$ is a random interval that contains the unknown, true value θ with a probability of 95%.

In formulas: $P(\hat{\theta}_l \leq \theta \leq \hat{\theta}_u) \geq 0.95$

Confidence interval (credibility region)

- It is also possible to define $(1-\alpha)\times 100\%$ confidence intervals in general. Conventionally $\alpha = 0.05$.
- For repeating experiments you are mistaken in $\alpha \times 100\%$ of the cases.
- It is obvious that confidence intervals are related to the concept of significance tests, so that we introduce them here.

Summary

In 1985 an overview of clinical trials confirmed that patients treated within 6 h of the onset of symptoms of myocardial infarction benefited from thrombolytic therapy. Doubt remained about treatment later than this and this uncertainty prompted further randomised studies. The South American multicentre trial EMERAS is one of these.

4534 patients entering hospital up to 24 h after the onset of suspected acute myocardial infarction were randomised between intravenous streptokinase (SK) 1·5 MU and placebo, during the period January, 1988, to January, 1991. Once the results of ISIS-2 were known, only patients presenting more than 6 h after symptom onset were randomised. There was no significant difference in mortality during the hospital stay (269/2257 [11·9%] deaths among SK patients *vs* 282/2277 [12·4%] in controls). Among the 2080 patients presenting 7–12 h from symptom onset there was a non-significant trend towards fewer deaths with SK (11·7% SK *vs* 13·2% control; 14% [SD 12] reduction with 95% confidence interval [CI] of 33% reduction to 12% increase), whereas there was little difference among the 1791 patients presenting after 13–24 h (11·4% *vs* 10·7%; 8% [16] increase with a 95% CI of 20% reduction to 45% increase). These 95% CIs are wide and are consistent with the results of previous studies among patients presenting late after symptom onset.

The EMERAS results, though not conclusive on their own, do contribute substantially to accumulating evidence on the question of whether fibrinolytic therapy really does produce any worthwhile improvement in survival among such patients.

Lancet 1993; **342:** 767–72

Confidence interval (credibility region)

- In the previous study no difference in mortality after heart attack between the groups with and without thrombolytic therapy could be detected ("no significant difference").
  This does not mean, that there is no difference.
- The confidence intervals show, that it is possible, that the therapy results in improvements of up to 33%.
- However, this needs to be confirmed with new studies, as the other limit of the confidence interval (impairment of 12%) is possible as well.
- Relation to hypothesis testing: a result is significant with $\alpha = 5\%$, if the value of the null hypothesis is not within the 95%–confidence interval.

Confidence interval for μ with known σ^2

- Measurement instrument with **known** dispersion $\sigma^2 = \sigma_0^2$
- Measurements $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma_0^2)$
- 95%–confidence interval for μ?

(i) $\bar{x}$ distributed with $\mathcal{N}(\mu, \frac{\sigma_0^2}{n})$

(ii) $\dfrac{\bar{x} - \mu}{\sigma_o/\sqrt{n}}$ distributed with $\mathcal{N}(0, 1)$

(iii)

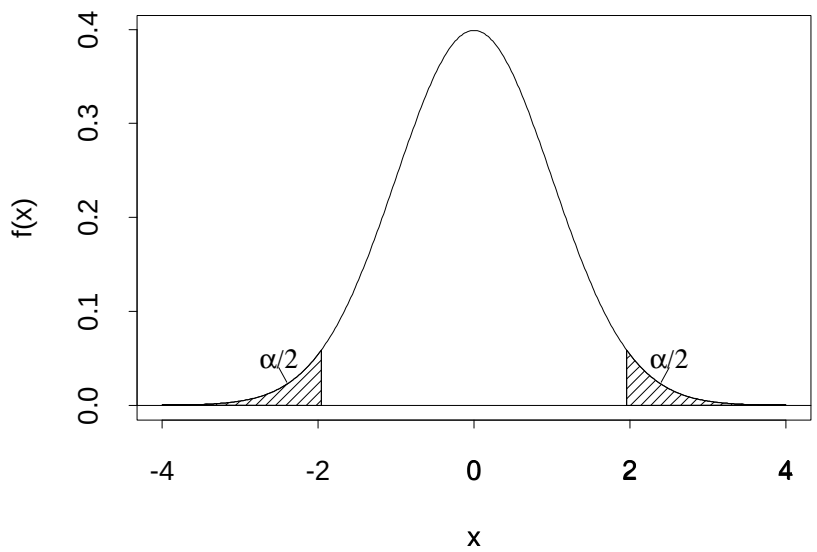
95%–confidence interval for mean μ with known σ_0:

$$\bar{x} - z_{0.975} \cdot \frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.975} \cdot \frac{\sigma_0}{\sqrt{n}}$$

$z_{0.975}$ = 97.5%–percentile of the normal distribution = 1.96

Confidence interval for μ with known σ^2

Motivation:



- $\alpha = 0.05 \rightarrow z_{\alpha/2} = -1.96, z_{1-\alpha/2} = 1.96$
- by definition $\mathrm{P}\left(z_{\alpha/2} \leq \dfrac{\bar{x} - \mu}{\sigma_0/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1 - \alpha$
- as $\mathcal{N}(0,1)$ symmetric: $z_{\alpha/2} = -z_{1-\alpha/2}$
- Solving for μ: $(1-\alpha)$–confidence interval

$$-z_{1-\alpha/2}\ \frac{\sigma_0}{\sqrt{n}} \leq \bar{x} - \mu \leq z_{1-\alpha/2}\ \frac{\sigma_0}{\sqrt{n}}$$

$$\Longrightarrow \bar{x} - z_{1-\alpha/2}\ \frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\ \frac{\sigma_0}{\sqrt{n}}$$

Confidence interval for μ with known σ^2

Comments:

- symmetric around $\bar{x}$, width determined by n, σ_0, α
- random as consequence of position at $\bar{x}$
- known σ_0 is not realistic

Numerical example: $\bar{x} = 0.2, \sigma_0 = 0.1$
Illustration of the dependence of α, n:

	α		
n	0.05	0.01	0.001
10	[0.14, 0.26]	[0.12, 0.28]	[0.10, 0.30]
50	[0.17, 0.23]	[0.16, 0.24]	[0.15, 0.25]
200	[0.19, 0.21]	[0.18, 0.22]	[0.18, 0.22]

"uncertainty relation"

Confidence interval for μ with unknown σ^2

Random variable $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

Example: Mean μ of the number of T_4-cells, $n = 20$ Hodgkin–patients.

Problem: Data right-skewed, obviously not normally distributed.

Solution:
1. take the logarithm
2. assume the log. data to be approximatively normally distributed

log T_4: $\bar{x} = 6.49$, $s = 0.71$

Idea: Standardise $\bar{x}$: $t = \frac{\bar{x}-\mu}{s/\sqrt{(n)}}$

Reason: t would be standard normally distributed if σ and not s is in the denominator

Consequence: t-distributed

otherwise as confidence interval for normal distribution with known variance

Confidence interval for μ with unknown σ^2

95%–confidence interval for μ with unknown σ

$$\bar{x} - t_{0.975} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{0.975} \cdot \frac{s}{\sqrt{n}}$$

- $t_{0.975}$ is the 97.5%–percentile of the t–distribution with $n-1$ degrees of freedom
- Interval symmetric around $\bar{x}$, width depends on n, s, α

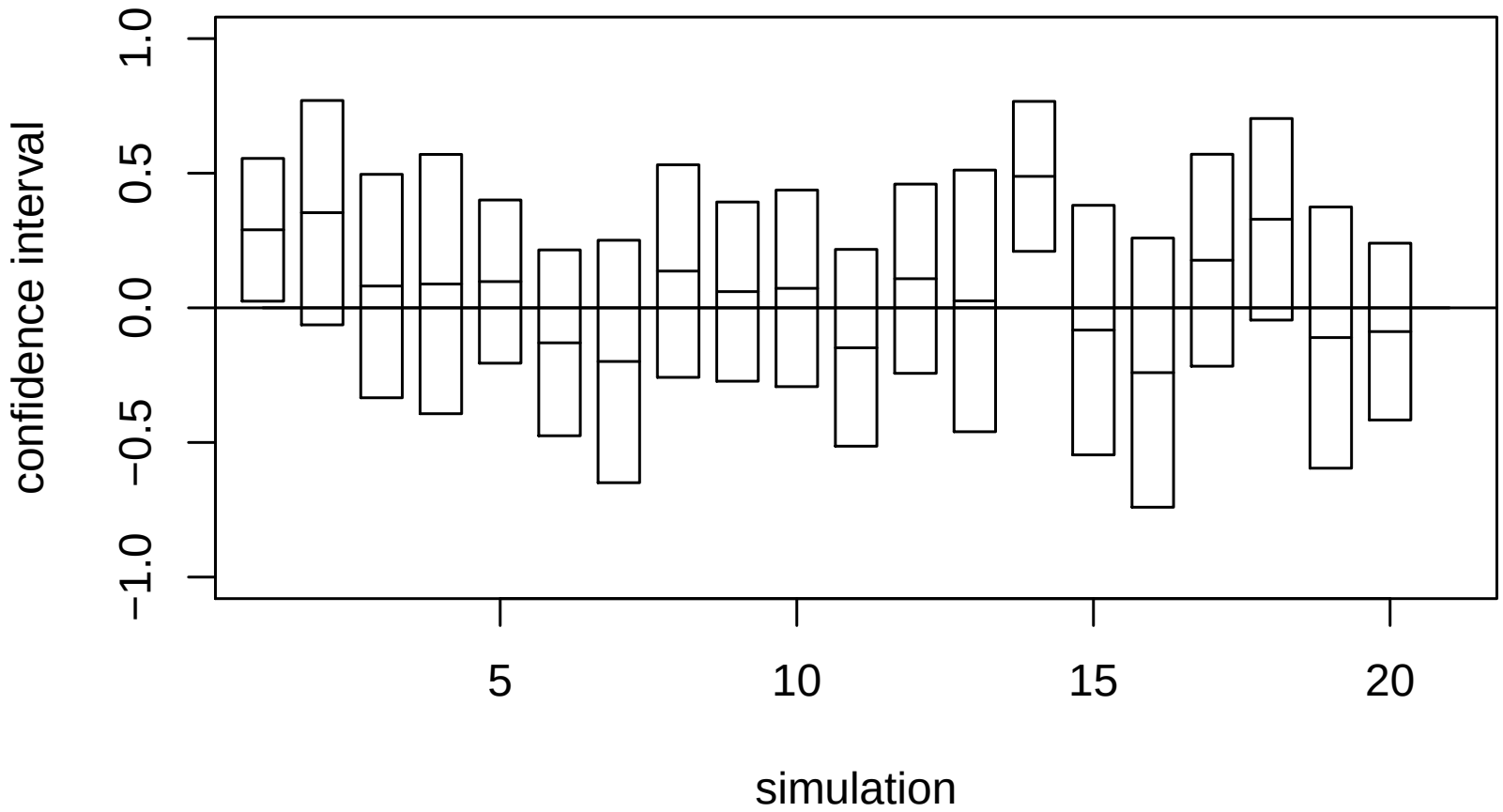
log T_4–cells:

$$\begin{aligned} \alpha = 0.05: &\quad 6.14 \leq \mu \leq 6.84 \\ \alpha = 0.01: &\quad 5.90 \leq \mu \leq 6.99 \\ \alpha = 0.001: &\quad 5.76 \leq \mu \leq 7.22 \end{aligned}$$

Variability of confidence intervals

$$x_1, \ldots, x_{25} \sim \mathcal{N}(0, 1)$$

Therefrom we consider 20 samples and calculate the 95%-confidence intervals for μ.



Confidence interval for relative frequency p

- For n persons a disease is observed k times.
- Relative frequency p estimated: $\hat{p} = k/n$
- $X_1, \ldots, X_n$ independent binary $(0, 1)$ variables with parameter p
- $\sum X_i$ binomial distributed with parameter p

$(1 - \alpha)$–confidence interval for true p?

Confidence interval for relative frequency p

Approximate calculation (without / with computer)

$$z = \frac{k - np}{\sqrt{n\,p(1-p)}}$$

approximative $\mathcal{N}(0,1)$, if n large (central limit theorem)

$\longrightarrow$ approximate 95%–confidence interval for p:

$$\hat{p} - z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Confidence interval for relative frequency p

More precise: Wilson confidence interval

$$A = 2k + z_{0.975}^2 \quad B = z_{0.975}\sqrt{z_{0.975}^2 + 4k(1-\hat{p})} \quad C = 2(n + z_{0.975}^2)$$

$$\hat{p}_l = (A - B)/C \quad \hat{p}_u = (A + B)/C$$

$\longrightarrow$ Wilson 95%–confidence interval for p:

$$\boxed{\hat{p}_l \leq p \leq \hat{p}_u}$$

Confidence interval for relative frequency p

Example: $n = 20$ births, $7\times$ boy $\longrightarrow \hat{p} = 7/20 = 0.35$

95%–confidence interval ?

approximate CI:	(0.14, 0.56)
Wilson–CI:	(0.18, 0.57)

i.e. :

- Credible region is **wide** (n too small)
- Credible region includes 0.5 (fair coin)

Confidence interval for relative frequency p

Real example: Frequency of male and female newborns

1950-1970: 1 944 700 births in CH, therefrom 997 600 males

$$\hat{p} = 0.5130 \quad (\neq 0.5 \text{ by choice?})$$

99%–confidence interval:

$$(0.5121, 0.5139)$$

i.e. :

- Credible region is **narrow**
- Credible region does **not** include 0.5 (unfair coin)