

Kun je met statistiek werkelijk alles bewijzen ?

Geert Verbeke

Biostatistisch Centrum, K.U.Leuven

International Institute for Biostatistics and statistical Bioinformatics

`geert.verbeke@med.kuleuven.be`

`http://perswww.kuleuven.be/geert_verbeke`

25 februari 2008

INLEIDING

Het belang van de statistiek

- Steeds meer data verzameld,
over steeds meer mensen,
in steeds meer contexten

- Inzicht krijgen in de data
- Statistiek wint aan belang:
 - ▷ Opleidingen
 - ▷ Vakliteratuur

	1	2	3	4	5	6	7	8	9	10
	PATID	Geslacht	BMI	Leeftijd	Roker	HEMA0	HEMA1	CARDIO	REJECT	FALE
1	1	man	17	30	Ex	33	33	Nee	Ja	Nee
2	2	vrouw	20	27	Ex	28	46	Nee	Nee	Nee
3	3	man	26	39	Ja	22	38	Nee	Ja	Nee
4	4	man	25	48	Ja	30	32	Nee	Ja	Nee
5	5	vrouw	22	32	Ja	32	49	Nee	Ja	Nee
6	6	vrouw	23	57	Ja	31	46	Nee	Nee	Nee
7	7	man	28	49	Ja	41	47	Nee	Nee	Nee
8	8	man	23	25	Nee	25	36	Nee	Ja	Nee
9	9	man	24	26	Ja	28	31	Nee	Ja	Ja
10	10	man	21	19	Ja	31	37	Nee	Ja	Ja
11	11	vrouw	21	22	Ex	40	38	Nee	Nee	Nee
12	12	man	26	55	Ja	41	36	Nee	Ja	Nee
13	13	vrouw	20	27	Ex	30	36	Nee	Nee	Ja
14	14	man	21	41	Ja	31	41	Nee	Ja	Ja
15	15	vrouw	22	62	Ja	32	34	Nee	Nee	Nee
16	16	vrouw	20	44	Ex	18	43	Nee	Ja	Nee
17	17	vrouw	28	66	Ja	29	41	Ja	Nee	Nee
18	18	vrouw	21	49	Ex	34	41	Nee	Nee	Nee
19	19	man	30	41	Ja	40	38	Nee	Nee	Nee
20	20	man	23	26	Ex	20	44	Nee	Nee	Nee
21	21	man	26	30	Ja	37	37	Nee	Nee	Ja
22	22	vrouw	22	30	Ja	33	36	Ja	Nee	Nee
23	23	man	18	43	Ja	24	30	Nee	Nee	Nee
24	24	vrouw	24	38	Ja	32	47	Nee	Nee	Nee
25	25	man	25	58	Ja	35	33	Ja	Ja	Nee
26	26	man	25	54	Ja	37	38	Nee	Nee	Nee
27	27	man	25	56	Ja	28	44	Nee	Nee	Nee
28	28	vrouw	23	55	Ja	35	60	Ja	Ja	Ja
29	29	vrouw	26	56	Ja	33	39	Nee	Nee	Nee
30	30	man	23	61	Ja	29	44	Nee	Nee	Nee
31	31	man	24	47	Ja	27	31	Nee	Nee	Nee
32	32	vrouw	23	37	Ex	27	40	Nee	Ja	Ja
33	33	vrouw	23	35	Ja	40	42	Nee	Nee	Nee

Gebruik en misbruik van de statistiek

- García-Berthou & Alcaraz (Med. Res. Meth. 2004):

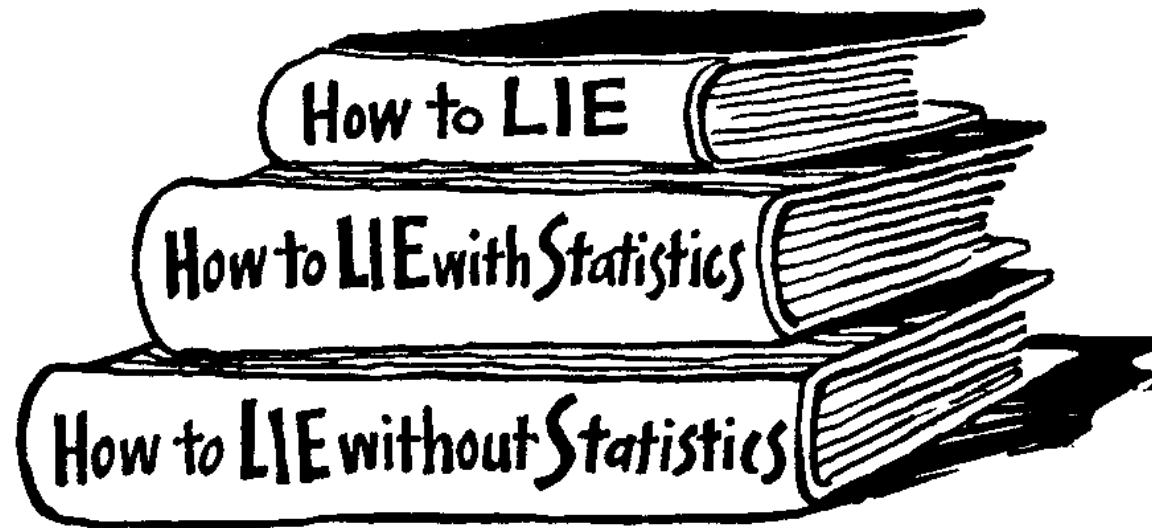
- ▷ Jaargang 2001 van Nature en British Medical Journal
- ▷ 38% en 25% van artikelen bevat een statistische fout
- ▷ 11%: fouten tegen interpretatie
- ▷ 4%: besluit spreekt evidentie tegen

- Gevolg:

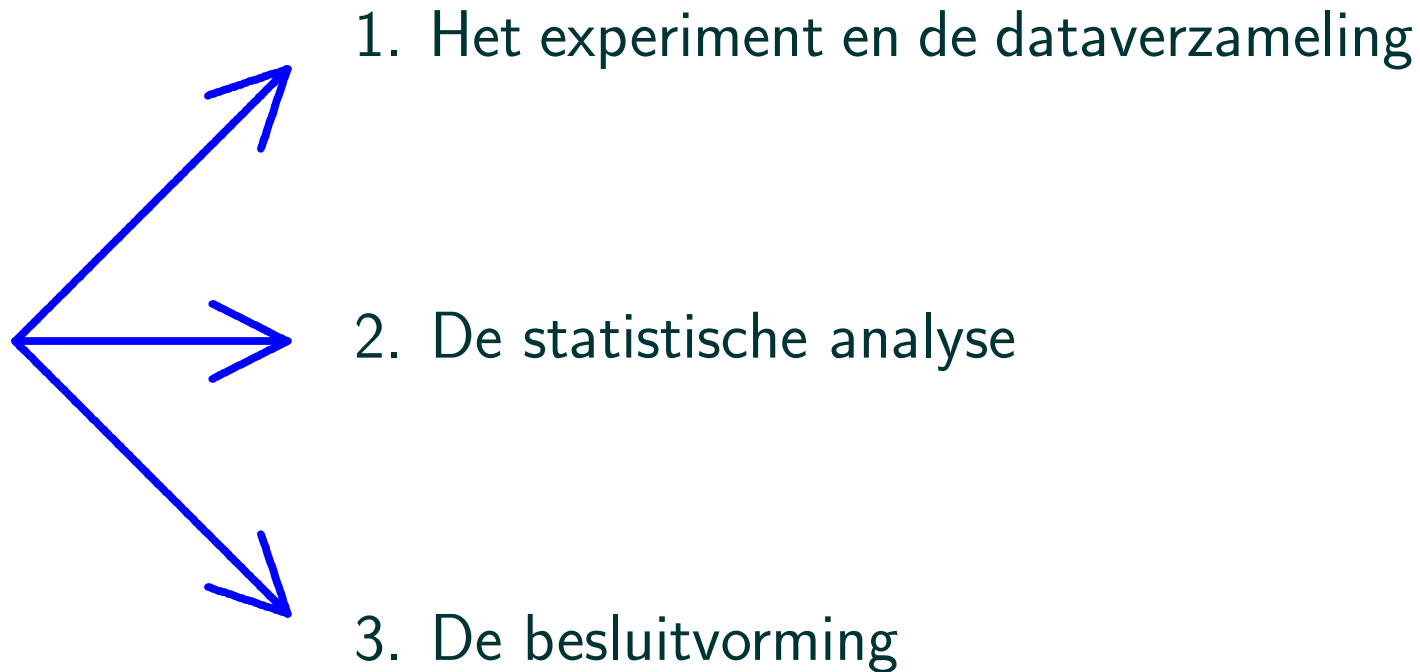
- ▷ Niet alle gepubliceerde resultaten zijn correct
- ▷ Tegenstrijdige en niet-reproduceerbare resultaten

Implicatie

“Met statistiek kun je alles bewijzen”



Stappen in het verzamelen van evidentie



DEEL 1

Het experiment en de dataverzameling

- ▷ Populatie versus steekproef
- ▷ Placebo-gecontroleerde studies
- ▷ Blinde en dubbelblinde experimenten
- ▷ Randomisatie en causaliteit

Populatie versus steekproef

- Aantonen dat een nieuwe behandeling werkt
- Ideaal:

Ganse populatie behandelen

- Praktisch haalbaar:

Een 'kleine' steekproef van patiënten behandelen



RANDOM



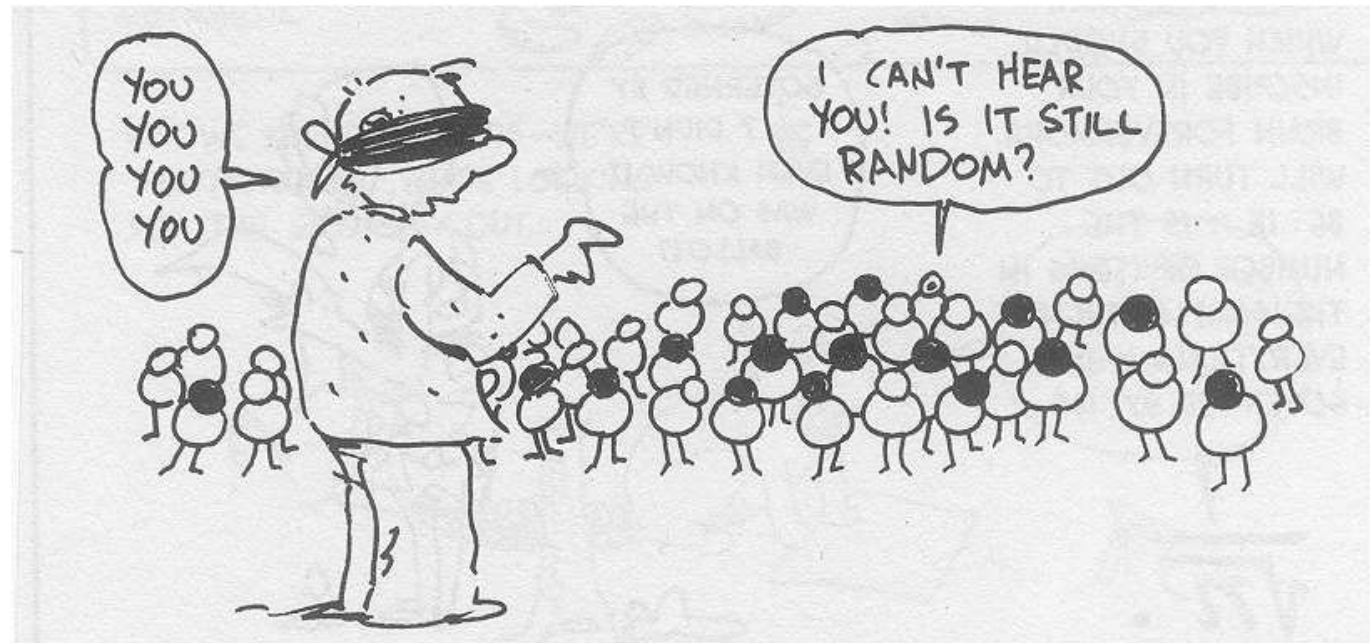
The word "RANDOM" is written in green. To its right is a large green arrow pointing downwards, indicating a transition or selection process from the population above to the sample below.



- De steekproef moet een afspiegeling zijn van de populatie en moet dus willekeurig (random) gekozen worden:

- ▷ Vergelijkbare leeftijdsverdeling
- ▷ Vergelijkbare geslachtsverdeling

▷ ...



- Een volledig random steekproef samenstellen is vaak zeer moeilijk

Placebo-gecontroleerde studies

- Een geobserveerd effect na behandeling is niet steeds toe te schrijven aan de behandeling:
 - ▷ Natuurlijke, gunstige evolutie ?
 - ▷ Psychologische effecten ?
 - ▷ ...
- Placebo:
 - ▷ Niet actief
 - ▷ Zelfde uitzicht, smaak, toediening,...
- Ethisch niet steeds verantwoord (bvb. standaard behandeling)



Blinde en dubbelblinde experimenten

- Patiënten met standaard therapie kunnen effect over-/onderschatten
Patiënten met nieuwe therapie kunnen effect over-/onderschatten
- Soms is het effect moeilijk objectief te meten

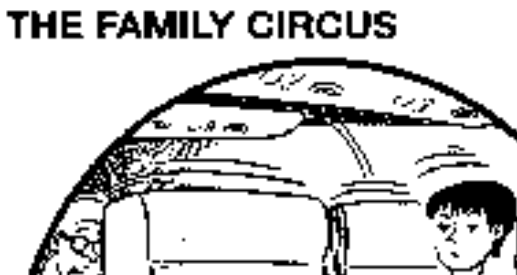


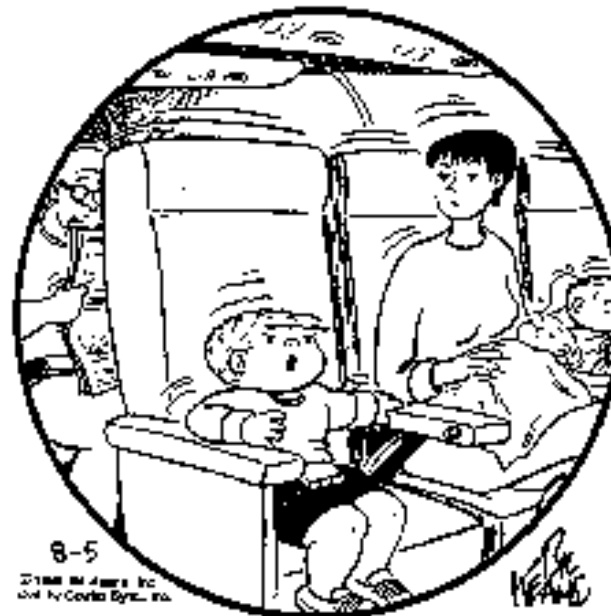
Blind: patiënt onwetend

Dubbelblind: patiënt en arts onwetend

- Niet steeds mogelijk (radio- vs. chemotherapie)

Randomisatie en causaliteit

- Een geobserveerd verschil is niet steeds oorzakelijk
 - Wat indien behandelingsgroepen verschillen wat betreft
 - ▷ leeftijdsverdeling ?
 - ▷ geslachtsverdeling ?
 - ▷?
- 
- A cartoon illustration of a man with dark hair looking out of a train window. Above the window, the text "THE FAMILY CIRCUS" is written in a bold, sans-serif font. The train is moving to the right, as indicated by the motion lines behind it. The man's expression is one of surprise or concern.



"I wish they didn't turn on that seatbelt sign so much! Every time they do, it gets bumpy."

- Causaliteit kan alleen maar besloten worden indien groepen gelijk zijn voor alle gekende **en ongekende** karakteristieken
- Oplossing:

Randomisatie:
Behandelingen willekeurig toewijzen aan patiënten

- Randomisatie niet steeds mogelijk (bvb. roken)



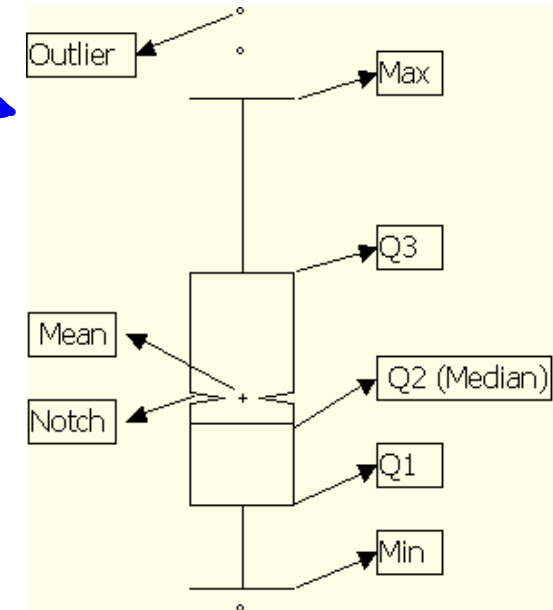
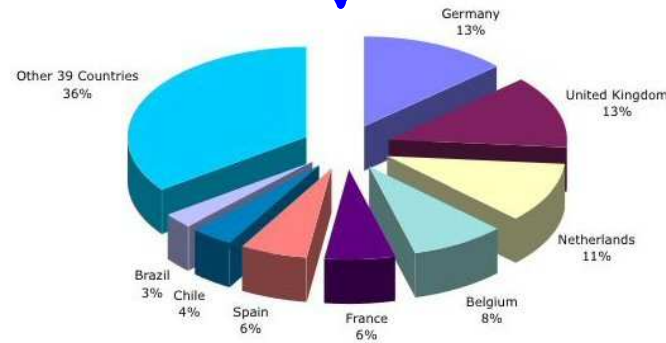
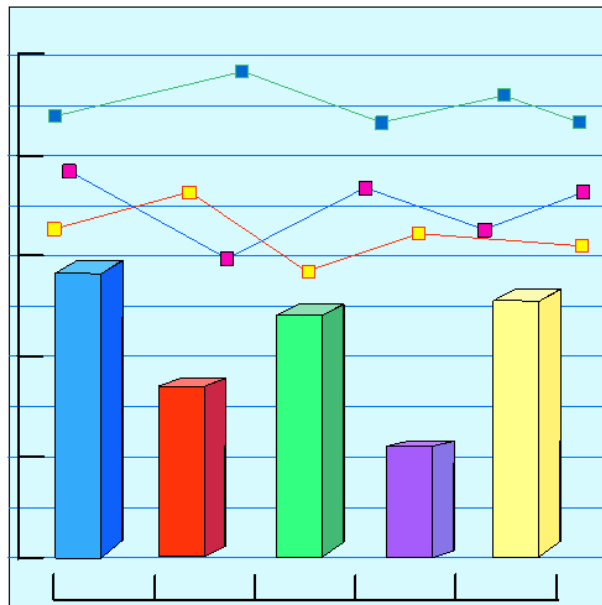
DEEL 2

De statistische analyse

- ▷ Beschrijvende statistiek
- ▷ Inferentiële statistiek
- ▷ p -waarde en significantie
- ▷ Een statistisch kookboek ?

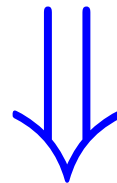
Beschrijvende statistiek

	1	2	3	4	5	6	7	8	9	10
	PATID	Geslacht	BMI	Leeftijd	Roker	HEMA1	HEMA2	CARDIO	REJECT	FALE
1	1	man	17	30	EX	33	33	Nee	Ja	Nee
2	2	vrouw	20	27	EX	28	48	Nee	Nee	Nee
3	3	man	28	35	Ja	22	38	Nee	Ja	Nee
4	4	man	25	48	Ja	30	32	Nee	Ja	Nee
5	5	vrouw	22	32	Ja	32	48	Nee	Ja	Nee
6	6	vrouw	23	57	Ja	31	48	Nee	Nee	Nee
7	7	man	28	48	Ja	41	47	Nee	Nee	Nee
8	8	man	23	25	Nee	25	36	Nee	Ja	Nee
9	9	man	24	26	Ja	28	31	Nee	Ja	Ja
10	10	man	21	19	Ja	31	37	Nee	Ja	Ja
11	11	vrouw	21	22	EX	40	36	Nee	Nee	Nee
12	12	man	26	55	Ja	41	36	Nee	Ja	Nee
13	13	vrouw	20	27	EX	30	36	Nee	Nee	Ja
14	14	man	21	41	Ja	31	41	Nee	Ja	Ja
15	15	vrouw	22	62	Ja	32	34	Nee	Nee	Nee
16	16	vrouw	20	44	EX	18	43	Nee	Ja	Nee
17	17	vrouw	28	66	Ja	23	41	Ja	Nee	Nee
18	18	vrouw	21	48	EX	34	41	Nee	Nee	Nee
19	19	man	30	41	Ja	40	38	Nee	Nee	Nee
20	20	man	23	26	EX	20	44	Nee	Nee	Nee
21	21	man	28	30	Ja	37	37	Nee	Nee	Ja
22	22	vrouw	22	30	Ja	33	36	Ja	Nee	Nee
23	23	man	18	43	Ja	24	30	Nee	Nee	Nee
24	24	vrouw	28	38	Ja	32	47	Nee	Nee	Nee
25	25	man	25	58	Ja	36	33	Ja	Ja	Nee
26	26	man	26	54	Ja	37	38	Nee	Nee	Nee
27	27	man	25	55	Ja	28	44	Nee	Nee	Nee
28	28	vrouw	23	55	Ja	35	60	Ja	Ja	Ja
29	29	vrouw	28	55	Ja	33	39	Nee	Nee	Nee
30	30	man	23	61	Ja	29	44	Nee	Nee	Nee
31	31	man	28	47	Ja	27	31	Nee	Nee	Nee
32	32	vrouw	23	37	EX	27	40	Nee	Ja	Ja
33	33	vrouw	23	35	Ja	40	42	Nee	Nee	Nee



- De geobserveerde gegevens worden samengevat
- De vorm van statistiek waar het brede publiek mee vertrouwd is:
 - ▷ Gemiddelden, percentages, ...
 - ▷ Grafieken en tabellen
 - ▷ Kranten, TV, ...

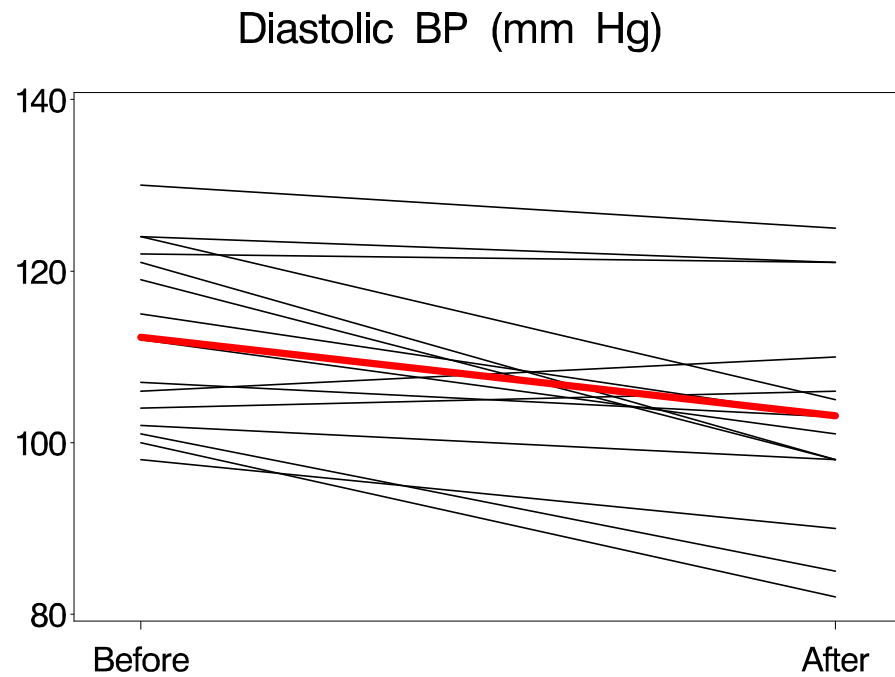
Dit laat geen veralgemening naar populatie toe



Inferentiële statistiek

Inferentiële statistiek

- Effect van een bloeddrukverlagende behandeling in 15 proefpersonen:

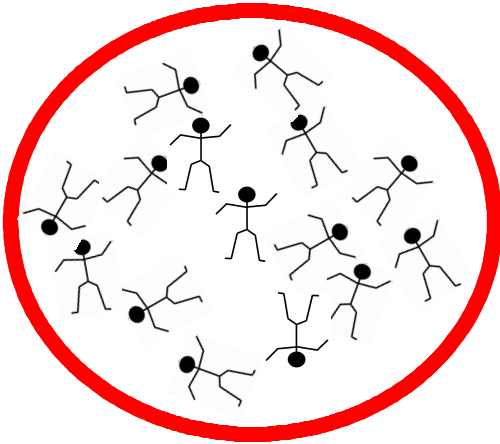


Gemiddelde (mm Hg)	
Voor:	112.3
Nadien:	103.1

- Herhaling van het experiment zou leiden tot nieuwe patiënten, en dus tot nieuwe observaties:
 - ▷ Groter effect ?
 - ▷ Kleiner effect ?
- Het geobserveerde effect zou dus toeval kunnen geweest zijn

**Is het geobserveerde effect voldoende evidentie
om te besluiten dat de behandeling effect heeft
bij toekomstige patiënten, m.a.w.,
in de totale populatie ?**

POPULATIE



Effect in totale populatie ?

RANDOM



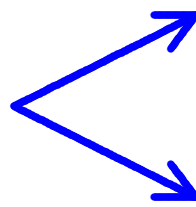
STEEKPROEF



Effect van 9mmHg
in 15 personen

p -waarde en significantie

Hoe waarschijnlijk is een verschil van 9mmHg
in een experiment van 15 personen,
indien de behandeling geen effect zou hebben (toeval) ?



Heel waarschijnlijk \Rightarrow “weinig evidentie voor effect”

Weinig waarschijnlijk \Rightarrow “sterke evidentie voor effect”

**De p -waarde is de kans om door toeval
een verschil van 9mmHg te observeren**

- In ons voorbeeld is er $p = 0.001 = 0.1\%$ kans om door puur toeval een verschil van 9mmHg te observeren.
- Significantieniveau α :
 $p \geq \alpha \implies$ “geen significant effect”
 $p < \alpha \implies$ “een significant effect”

- Standaard wordt $\alpha = 0.05 = 5\%$ gekozen
- Een significant effect is dan een effect dat minder dan 5% kans heeft om voor te vallen door puur toeval



Een statistisch kookboek ?

- Berekening van p afhankelijk van:
 - ▷ Type gegevens
 - ▷ Studie opzet
 - ▷ Specifieke vraagstelling
 - ▷ Modelveronderstellingen
 - ▷ ...
- Foute keuzes leiden tot foute resultaten
- Keuze van correcte techniek vereist kennis en ervaring, en kan dus niet geautomatiseerd worden.

DEEL 3

De besluitvorming

- ▷ Evidentie versus bewijs
- ▷ Fouten
- ▷ Significantie versus relevantie
- ▷ Equivalentie
- ▷ Herhaaldelijk testen

Evidentie versus bewijs

- **Significantie is geen bewijs van een aanwezig effect**
- Bvb., $p < 0.0001$ sluit niet uit dat het geobserveerd effect puur toeval is
- **Niet-significantie is geen bewijs van afwezigheid van effect**
- Bvb., $p = 0.99$ sluit niet uit dat een reëel effect niet gedetecteerd werd

**Statistiek kwantificeert de evidentie in de observaties
voor de aan- of afwezigheid van een effect in de populatie**

Fouten

		Realiteit	
		Geen effect	Wel effect
Test resultaat	Geen effect	OK	Type II fout
	Wel effect	Type I fout	OK

- **Type I fout:** Onterecht beslissen dat er effect is
- **Type II fout:** Onterecht beslissen dat er geen effect is

Fouten zijn niet uit te sluiten



**Fouten kunnen onder controle gehouden worden,
alleen door veel data te verzamelen**

Significantie versus relevantie

- De kans om een aanwezig effect te missen is klein bij grote studies
- Een aanwezig effect, hoe klein ook, zal dus vroeg of laat gedetecteerd worden, indien de studie voldoende groot is.
- Bvb., een bloeddrukverlagend produkt, uitgetest op 10000 patiënten kan leiden tot een significant effect ($p < 0.0001$) van gemiddeld 0.1mmHg
- Een gemiddelde daling van 0.1mmHg is klinisch niet relevant

Statistische significantie



Klinische relevantie

Equivalentie

- Soms wil men aantonen dat twee behandelingen **'even effectief'** zijn:
 - ▷ Minder nevenwerkingen
 - ▷ Gemakkelijker toe te dienen
 - ▷ Goedkoper
 - ▷ ...
- Vaak wordt **'equivalentie'** besloten op basis van niet-significantie
- Klassieke testen zoeken naar evidentie in de data voor aanwezigheid van effecten

- Men hoopt dan dat, indien er verschillen zijn, men die niet ontdekt, m.a.w. dat een type II fout gemaakt wordt.
- Dit impliceert dat equivalentie het best zou kunnen aangetoond worden door zo weinig mogelijk gegevens te verzamelen
- Besluit:

Niet-significantie \neq **equivalentie**

- **Equivalentietesten** zijn ontwikkeld om evidentie te zoeken dat effecten voldoende klein zijn, om ze als **klinisch verwaarloosbaar** te kunnen beschouwen.

Voorbeeld: Shatari et al. (Col.Dis. 2004)

Original article	
------------------	--

Long strictureplasty is as safe and effective as short strictureplasty in small-bowel Crohn's disease

T. Shatari*, M. A. Clark*, T. Yamamoto*, A. Menon*, C. Keh*, J. Alexander-Williams* and M. Keighley*

*Department of Surgery, Queen Elizabeth Hospital, Edgbaston, United Kingdom

Received 17 November 2003; accepted 13 March 2004

	Long strictureplasty	Short strictureplasty	<i>P</i> -value
Number of patients	21	41	
Male:Female	8 : 13	18 : 23	0.669
Age (years) (median (range))	35 (12-67)	40 (14-77)	0.548
Smoking habit	9 (42.8%)	17 (45.9%)	0.820
Previous operation	17 (80.9%)	36 (87.8%)	0.468
Previous small bowel resection	16 (76.1%)	29 (70.7%)	0.648
Time since previous operation (months)(:median (range))	81 (8-305)	82 (3-402)	0.609
Abscess at operation	2 (9.5%)	6 (14.6%)	0.550
Fistula at operation	1 (4.7%)	7 (17.0%)	0.171
Site of strictureplasty			
Jejunum	5	7	0.525
Ileum	10	15	0.400
Ileal-caecal lesion	6	19	0.171
Simultaneous small bowel resection	9 (42.8%)	16 (39.0%)	0.770
Post-operative complication			
Stay (days) (median (range))	10 (4-23)	9 (4-18)	0.720
Leakage rate	0 (0%)	1 (2.4%)	0.470
Intra-abdominal abscess rate	2 (9.5%)	1 (2.4%)	0.218
Fistula	2 (9.5%)	1 (2.4%)	0.218
Disease-free rate (%)			0.702
3 years	80.4	62.1	
5 years	55.2	49.8	
10 years	49.1	33.5	

**Geen
significanties !**

Conclusions These data indicate that long strictureplasty is safe and produces equivalent results to conventional (short) strictureplasty.

Fout !

Herhaaldelijk testen

- Omdat fouten nooit helemaal uitgesloten kunnen worden zullen ze zich vroeg of laat voordoen
- Bij herhaaldelijk testen wordt dus vroeg of laat een type I gemaakt
- Implicatie:

Hoe meer testen worden uitgevoerd, hoe groter de kans dat iets gevonden wordt wat berust op puur toeval

- Dit probleem van **herhaaldelijk testen** komt voor onder verschillende vormen

Voorbeeld: Het aula experiment

- Verdeel het auditorium in twee delen: Links en rechts
- Test voor verschillen in gewicht, lengte, leeftijd, geslacht, haarkleur, lengte pink, dikte linker grote teen, ...
- Vroeg of laat worden significante verschillen ontdekt, bvb. voor de grootte van de neus

Zitten lange neuzen systematisch meer rechts ???

Voorbeeld: Krantenartikel

- Volgende 'wetenschappelijke bevinding' was in Belgische krant te lezen:



- Er werd verder vermeld dat diegenen die opstaan voor 7u21 significant hogere stress waarden hebben dan diegenen die na 7u21 opstaan

Herhaaldelijk testen: Implicaties

- Significante resultaten bekomen door herhaaldelijk te testen worden overgeïnterpreteerd
- Indien het aantal testen gerapporteerd wordt, dan kan 'de lezer' het belang van de resultaten inschatten
- Problematisch indien enkel significante resultaten gerapporteerd worden, zonder vermelding van het aantal testen
- Dit leidt tot resultaten die niet reproduceerbaar zijn

- Bvb., een nieuw aula experiment zou geen evidentie meer bieden voor een verschil in neuslengte, maar misschien wel voor een verschil in ■ ■ ■ ?
- Bvb., een nieuw slaap experiment toont misschien geen relatie tussen stress en moment van ontwaken, of misschien vindt men enkel significantie indien men ontwaakt voor ■ ■ ■ ?

**Waren gerapporteerde effecten
a priori te verwachten ?**

**Hoe aanneembaar is het dat het experiment werd
opgezet om dit specifiek effect te bestuderen ?**

BESLUIT

~~“Met statistiek kun je alles bewijzen”~~

- ▷ Statistiek kwantificeert enkel evidentie gevonden in de data
- ▷ Geen formeel bewijs
- ▷ Resultaten mogelijks sterk afhankelijk van assumpties

**“Zelfs met correcte statistiek
kun je niets bewijzen”**



The End !