MULTIPPEL TESTING

Anne Marie Fenstad
Biostatistiker
Nasjonalt kvalitets- og kompetansenettverk for leddproteser og hoftebrudd

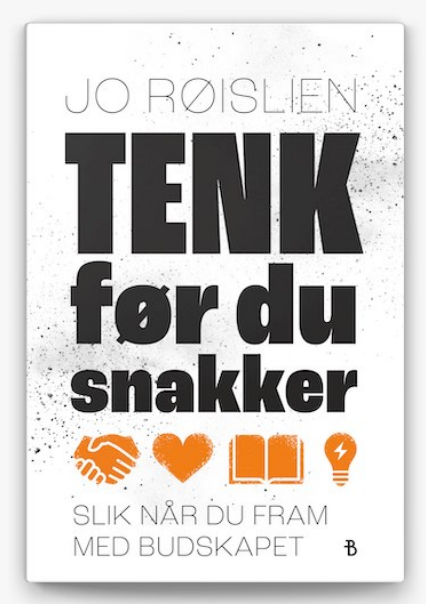
HELSE BERGEN
Haukeland universitetssjukehus

Oversikt

- Nasjonalt Register for Leddproteser
- Hvorfor er multippel testing en utfordring
- Hypotesetesting
- Mulige strategier
- Eksempler
- Oppsummering
- Litteratur

Tusen takk for inspirasjon og innspill

- Mette Langaas, institutt for matematiske fag, NTNU
- Egil Ferkingstad, University of Iceland (tidl NR)
- Ellinor Ytterstad, Matematikk og statistikk, UiT
- Stian Lydersen, Institutt for psykisk helse, NTNU
- Jo Røislien, Det helsevitenskapelige fakultet, UiS



HELSE BERGEN
Haukeland universitetssjukehus

Nasjonalt Register for Leddproteser (NRL)

- Registrere data for leddproteser i hofte, kne, skulder…
- Oppstart i 1987
- Det overordnede målet er å kvalitetssikre og forbedre behandlingsmetodene og tilbudet til pasientene
- Årsrapport og egen rapport til hvert sykehus
- Over 95 % dekningsgrad
- Mangfold av prosjekter
- Del av Nasjonalt kvalitets- og kompetansenettverk for leddproteser og hoftebrudd

Aktuelt fra NRL

- Innføring av elektronisk rapportering (MRS) og ePROM
- R-RCT kneproteser med bensement med og uten antibiotika og infeksjon innen 1 år som endepunkt
- Kvalitetsforbedringsprosjekt – eldre pasienter bør få sementert femurstamme

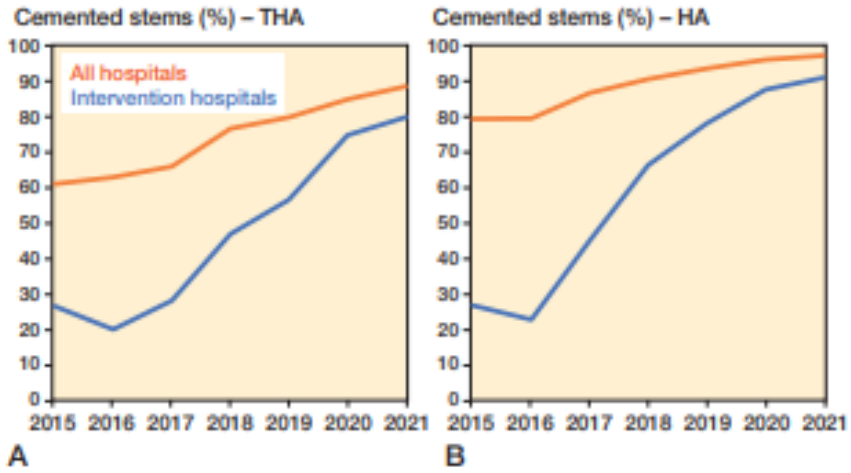
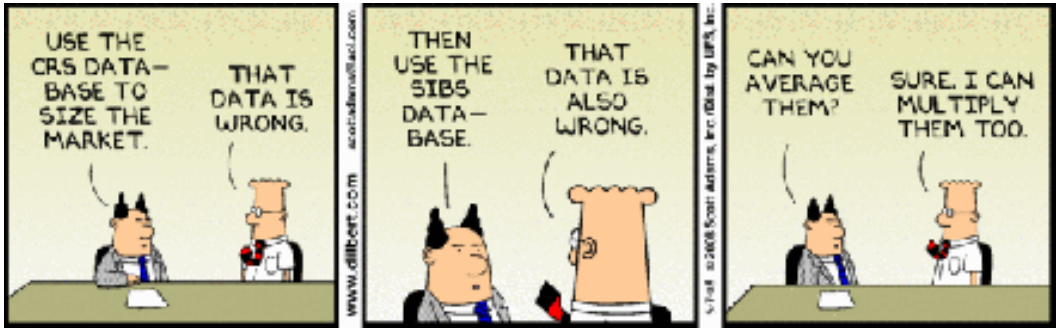


Figure 2. A. Proportion of total hip arthroplasties (THA) with cemented femoral stems in women ≥ 75 years of age reported to NAR. B. Proportion of hemiarthroplasties (HA) with cemented femoral stems in patients ≥ 70 years of age reported to NHFR.

Registerdata

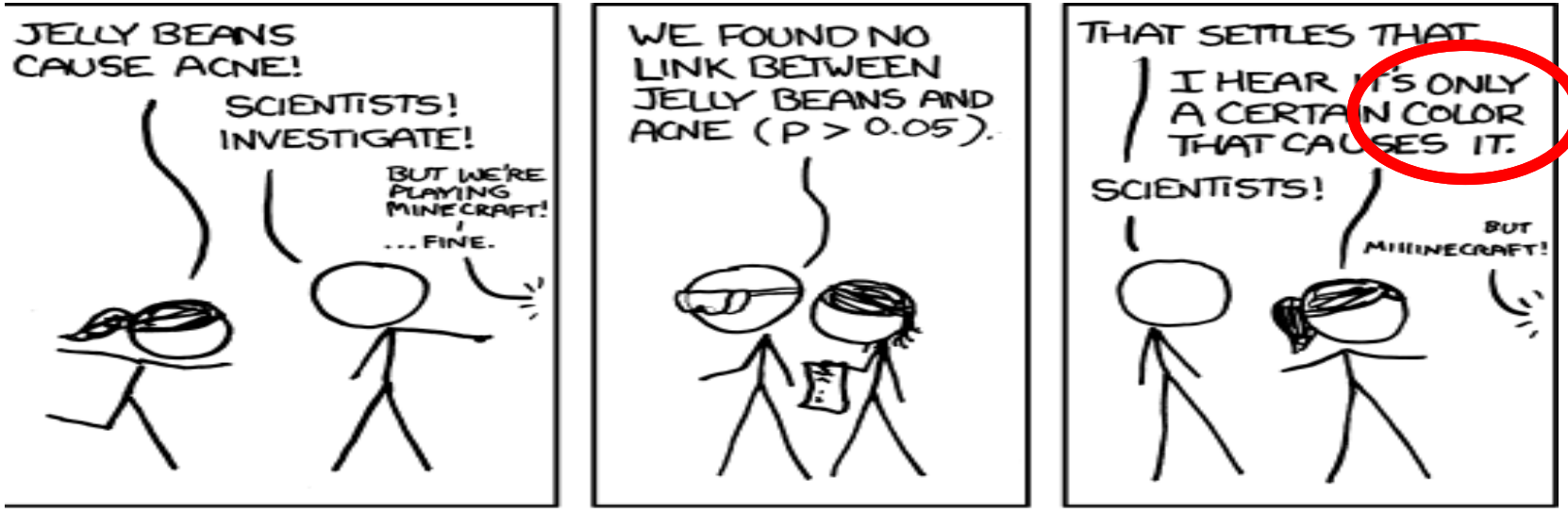


- Innsamlingsrutiner (pålagt, samtykkebasert)
- Reservasjonsrett for Hoftebruddregisteret fra 1. juli 2021
- Representativitet
- Metadata: type data med detaljnivå, definisjoner, kodelister
- Kompletthet, kvalitet og validering
- To randomiserte studier (R-RCT) med utgangspunkt i registre
  - Kneprotese festet med sement med eller uten antibiotika
  - Korsbånd – operasjon eller konservativ behandling

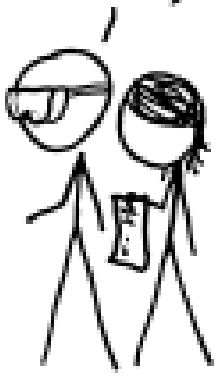
Innledning

- I hvilke situasjoner er multippel testing en utfordring
- Identifisere problem med multippel testing og beskrive metoder som korrigerer for problemene
- Kan vi unngå dette problemet med å planlegge godt før en studie
- Finnes det en metode som alltid løser problemet «best»?
- Først et eksempel fra xkcd (https://xkcd.com/882/)

Forårsaker «jelly beans» kviser?



LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).



LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).

LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).

LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).

LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

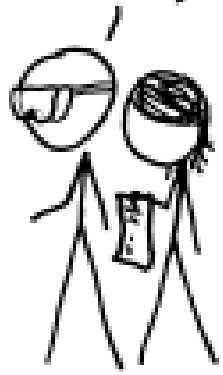
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).

WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).

WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).



WE FOUND NO

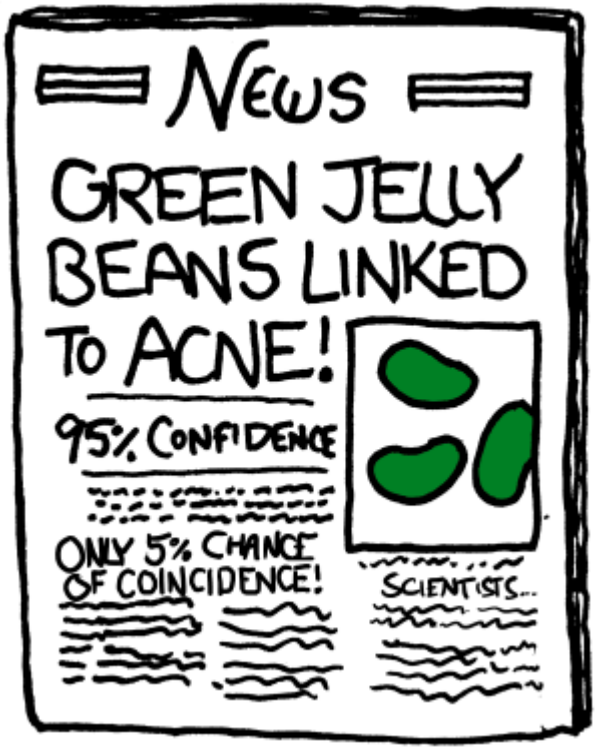
WE FOUND NO

WE FOUND NO

WE FOUND NO

WE FOUND NO

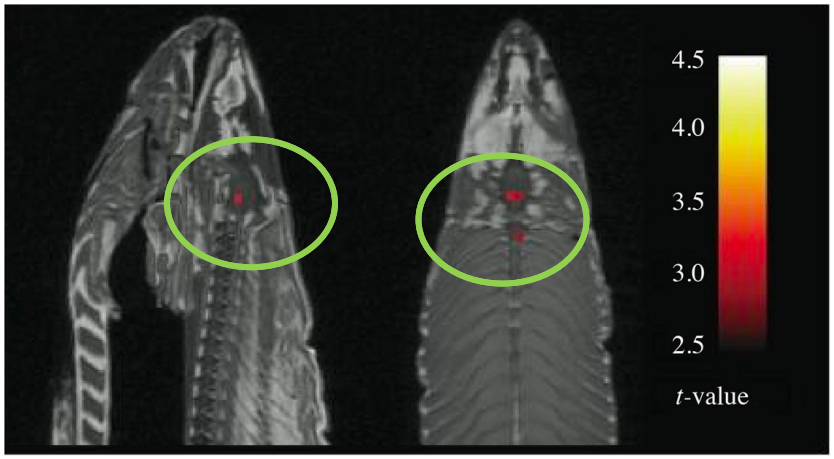
Og finner altså «bevis» for at grønne jelly beans forårsaker kviser



Statistisk lærdom fra en (død) laks

Statistics lessons from a salmon. Røislien J.Tidsskr Nor Laegeforen. 2023 Sep 22;143(13). doi: 10.4045/tidsskr.23.0471

- Undersøke beslutningstaking hos mennesker ved hjelp av funksjonell magnetisk resonanstomografi (fMRI)
- Laksen ble vist fotografier av mennesker i ulike sosiale situasjoner og ble så bedt om å avgjøre hvilke følelser menneskene i fotografiene opplevde
- Et av bildene av hjerneaktiviteten til laksen viste statistisk signifikante prikker



Statistisk lærdom fra en (død) laks forts.

- En voxel kan vise aktivitet selv om det ikke er noen. Sannsynligheten for feil i en enkelt voxel er lav, men med så mange voxler at det ikke er usannsynlig at noen av dem feilaktig viser aktivitet.

- Justering kommer med en pris: tap av statistisk styrke. Du unngår kanskje falskt positive svar, men du risikerer samtidig å ikke finne ting som faktisk er der, såkalte falskt negative.

Statistisk lærdom fra en (død) laks forts.

- I fMRI-feltet diskuteres det om det er falskt positive funn eller falskt negative som er verst.

- I **kliniske studier** er problemet mer skjult. Studier der man tester mange hypoteser og mange utfallsvariabler på de samme individene, har samme utfordring.
- Gjør man mange nok statistiske tester, vil sjansen øke for at man finner noe, og denne økte sannsynligheten for falskt positive funn må håndteres.

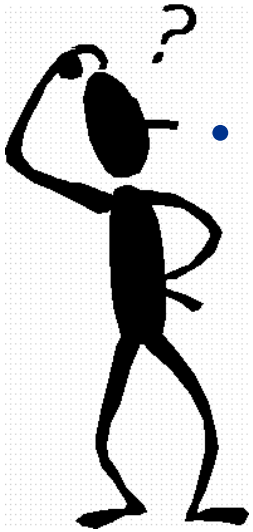
Hvorfor multippel testing

- Studere effekt på flere utfallsvariable
- Sammenligne to eller flere grupper
- Gjøre separate analyser for undergrupper

- Først litt om hypotesetesting

Hypotesetesting

- Vi setter opp en konservativ/nøytral hypotese, H_0, som vi har mistanke om at ikke stemmer
- Mot alternativ hypotese, H_A, som vi ønsker å teste



- Vi vil undersøke om våre data gir grunnlag for å påstå at mistanken er berettiget

Hypotesetesting (forts.)

- Spesifiserer en nullhypotese H_0 og alternativ hypotese H_A
- Samler inn data
- Beregner testobservator (T) basert på de aktuelle data, hvor stor T gir «bevis» mot H_0
- Vi beregner sannsynligheten (*p-verdi*) for at det vi observerer er tilfeldig
- Vi forkaster H_0 om p-verdien er liten, for eksempel mindre enn 0,05 (*signifikansnivået*)

Eksempler kan være å teste om en parameter er lik null, eller om behandling A er bedre enn behandling B

Hypotesetesting (forts.)

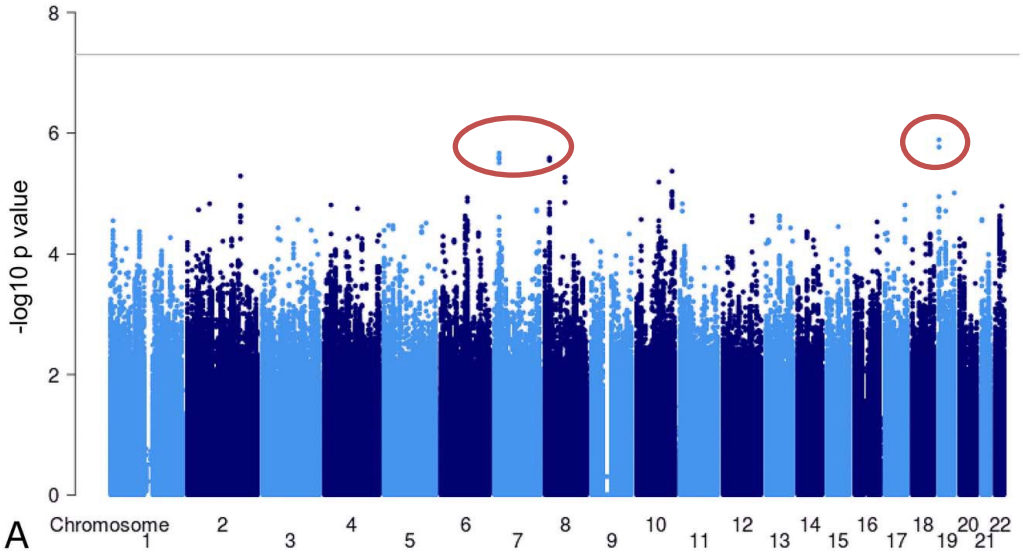
- Hva om vi ønsker å teste et sett av hypoteser simultant?
- Først prøver vi «en om gangen», for eksempel 20 hypoteser og signifikansnivå lik 0,05
- Hva er sannsynligheten for å finne minst et signifikant resultat ved en tilfeldighet?
- *P(minst ett signifikant resultat) = 1 – P(ingen signifikante resultater)*
  = 1 – (1 – 0,05)^20
  = 0,6415141
- Med 20 tester har vi altså 64 % sjanse for å finne minst ett signifikant resultat

Hypotesetesting (forts.)

- I praksis (f.eks. genforskning) har vi ofte mer enn 20
- Med 100 hypoteser får vi over 99 % sjanse for å finne minst ett signifikant resultat
- Et eksempel:
  - I registeret ønsket vi å se om vi kunne identifisere et «gen for løsning av hofteprotese»)
  - Litt forenklet: en gruppe pasienter med løsnet protese og matchet på kjønn, alder og operasjonsår til kontrollgruppen (1:5)
  - How does genome-wide variation affect ostelysis risk after THA?
    MacInnes et al Clin Orthop Relat Res (2019) 477:297-309

Manhatten plott

Den lyse grå linjen indikerer Bonferroni signifikansgrense
(Her er PBonf = 5.0×10^{-8})



Generelt om hvorfor multippel testing har betydning

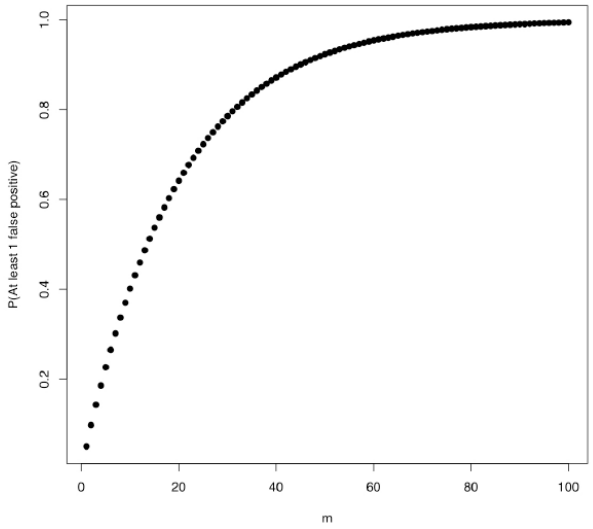
Hvis vi utfører *m* hypotesetester, hva blir sannsynligheten for en falsk positiv?

$$P(\text{Making an error}) = \alpha$$

$$P(\text{Not making an error}) = 1 - \alpha$$

$$P(\text{Not making an error in m tests}) = (1 - \alpha)^m$$

$$P(\text{Making at least 1 error in m tests}) = 1 - (1 - \alpha)^m$$

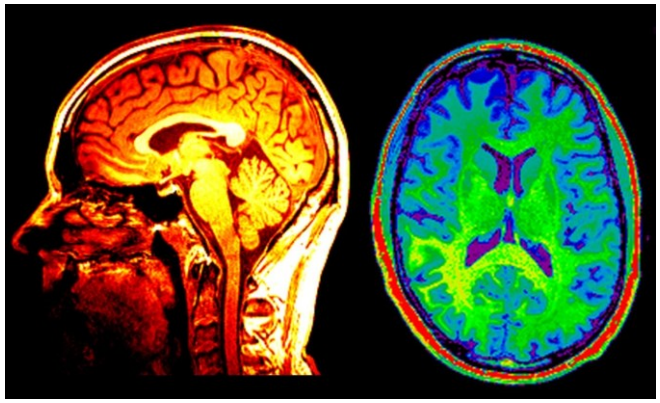


α = *alfa* = signifikansnivå
m = antall hypoteser

Sannsynligheten for minst en falsk positiv

Multippel testing er aktuelt i mange felt

- Medisinsk forskning - store spørreskjema, ernæringsstudier
- Genomikk (ofte tusenvis av tester for å forsøke å identifisere genetiske varianter som er assosiert med sykdom eller spesielle egenskaper)
- Miljøvitenskap – flere kilder til forurensing i forskjellige prøver
- Astronomi
- Finans (aksjemarkedet)
- MRI og fMRI – bilder for eksempel av hjernen eller hjernefunksjon



«Psychological symptoms in children of parents with chronic pain - The HUNT study»

Kaasbøll J, Lydersen S, Indredavik MS, Pain, 2012 May:153(5):1054-1062.

- 4 grupper ungdommer: Foreldre med kroniske smerter:
  - Ingen
  - Bare mor (M)
  - Bare far (F)
  - Både mor og far (MF)
- 2 uavhengige variable:
  - Angst/depresjon (SCL-5)
  - Atferdsproblemer
- Separate analyser for gutter og jenter

Type I og Type II feil – enkel hypotese

	Sannhet H_0 er sann	**Sannhet** H_A er sann
Beslutning Beholde H_0	Riktig konklusjon (sant negativ)	Gal konklusjon Type II-feil (falskt negativ)
Beslutning Forkaste H_0 (og påstå H_A)	Gal konklusjon Type I-feil (falskt positiv)	Riktig konklusjon (sant positiv)

Figur: Lydersen S. Type I-feil og type II-feil. Tidsskr Nor Legeforen 2021; 141. doi: 10.4045/tidsskr.21.0013

To typer feil:

- Falske positive = Type I feil = «uskyldig dømt»
- Falske negative = Type II feil = «skyldig kriminell går fri»

Multippel testing

Kjente størrelser: **R** og *m*

Ukjente størrelser: m_0, *V*, *T*, *U*, *S*

	H_0 er sann	H_A er sann	Total
Beholde H_0	*U*	*T*	m-R
Forkaste H_0	*V*	*S*	R
Alle	m_0	$m-m_0$	*m*

HELSE BERGEN
Haukeland universitetssjukehus

Definisjoner

- m er totalt antall nullhypoteser H_0
- m_0 er antallet sanne nullhypoteser
- $m - m_0$ er antallet falske nullhypoteser
- V er antall falske positive (type I feil), forkaste H_0 når H_0 er sann
- S er antall sanne positive, forkaste H_0 når H_0 er usann
- T er antall falsk negative (type II feil), ikke forkaste H_0 når H_0 er falsk
- U er antall sann negative, ikke forkaste H_0 når H_0 er sann
- $R = V + S$ er totalt antall forkastede nullhypoteser (både sanne og falske)

Ulike metoder for justering

- En enkelt hypotese: Type I feil og styrke (1 - Type II feil)
- Familievis feilrisiko (FWER): Sannsynligheten for å gjøre type I-feil i minst en av hypotesetestene
- FDR (false discovery rate) [Benjamini & Hochberg (1995)] – når vi har et høyt antall hypoteser vi undersøker (eks. genetikk)
- *pFDR (positive false discovery rate) [Storey (2002)]*

Strategier i multippel testing – Type I

- Single hypothesis: Type I feil

- FWER (Family-wise error rate), FWER = P(V>=1)
  - Bonferroni korreksjoner
  - Tukey multiple comparison test, Tukey´s method
  - The Sidak korreksjon
  - Holms step-down-korreksjon
  - Hochbergs step-up korreksjon
  - Hommel-korreksjon

- Falsk deteksjonsandel (FDR)
  - Tillater en viss andel falske positive funn

Sterk eller svak kontroll av Type I feil for multippel feilrate

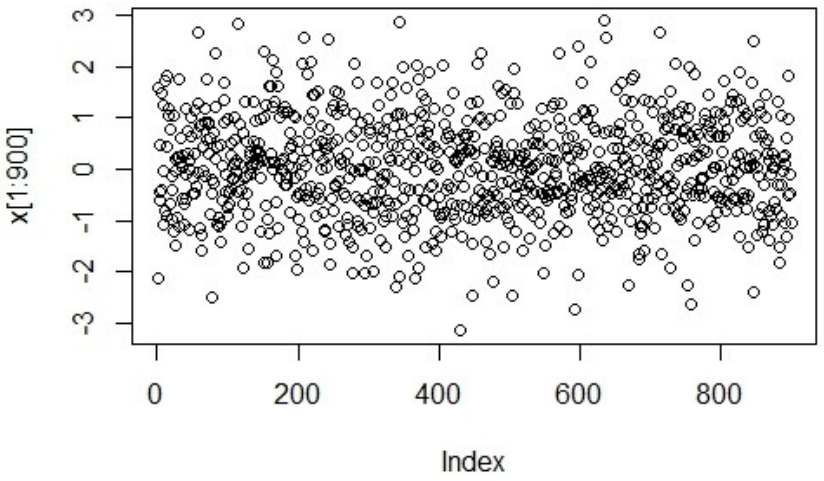
- **Sterk kontroll** vil si å ha kontroll på feilraten under enhver kombinasjon av sanne og usanne hypotese
- **Svak kontroll** vil si å ha kontroll på feilraten bare når alle nullhypotesene er sanne
- Har vi sterk kontroll vil vi også ha svak kontroll

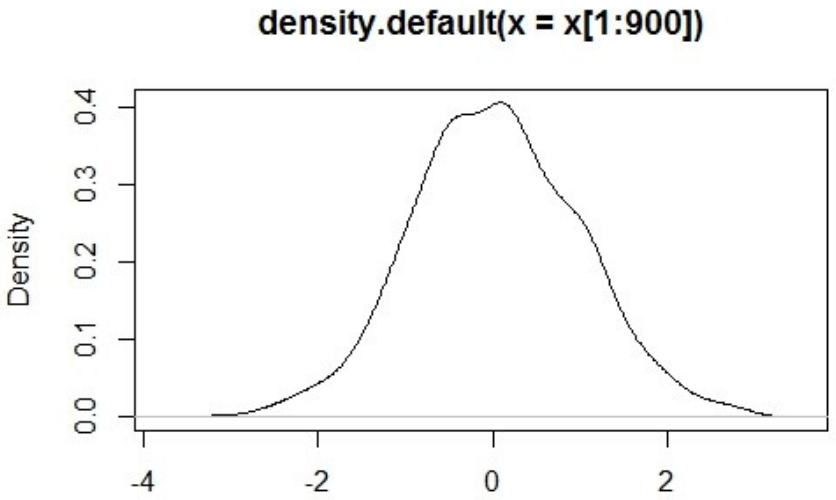
Metoder for korrigering av signifikansnivå – først Bonferroni korreksjon

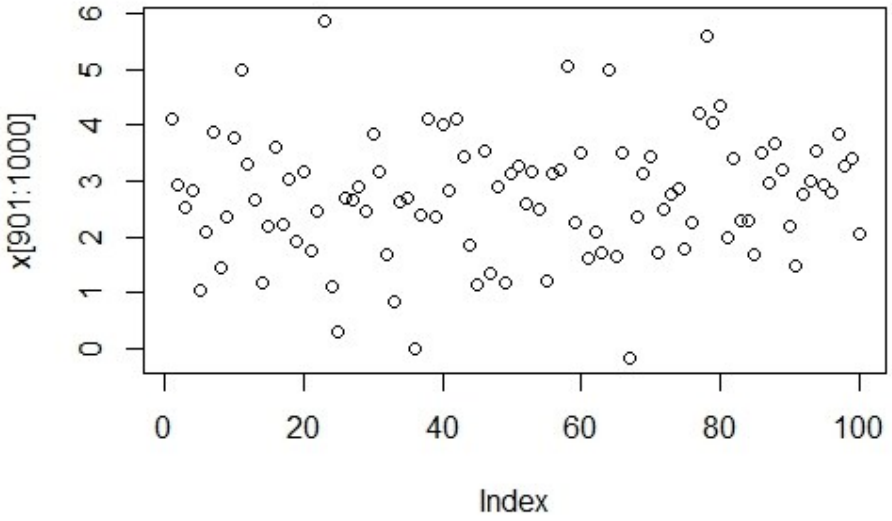
- Vi skal se på måter å justere signifikansnivået (*alfa*) på slik at sannsynligheten for å observere minst et signifikant resultat ved en tilfeldighet er under vår bestemte grense
- En metode er **Bonferroni korreksjon**, da setter man signifikansnivået lik *alfa/m,* hvor *m=antall hypoteser*
- I forrige eksempel får vi 0,05/20 = 0,0025 slik at en nullhypotese forkastes om *p-verdien* er mindre enn 0,0025
- Dette viser seg å være svært strengt (konservativt)

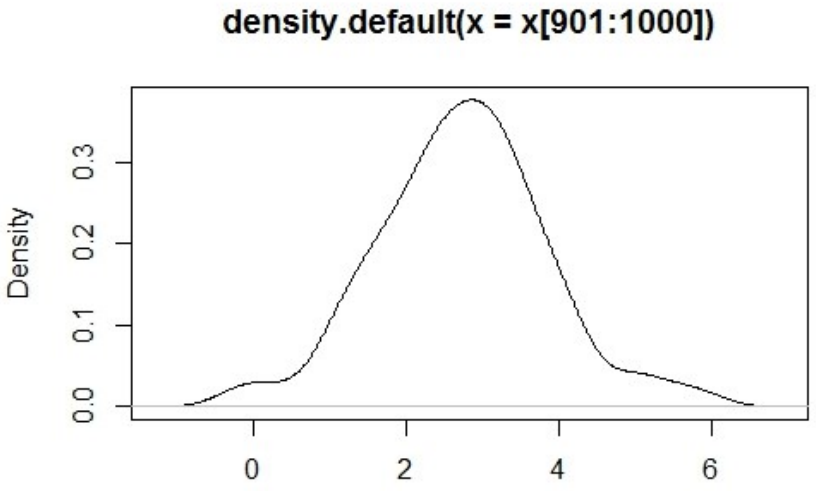
Eksempel 1

- Lager et datasett med *n=1000* tilfeldige tall som følger en normalfordeling
- De første 900 har en *N(0,1)* og de 100 siste *N(3,1)*
- Setter så opp en ensidig test for å teste om x er lik null, alternativ hypotese er at x er større enn null
- $H_0 : x = 0$ mot $H_1 : x > 0$
- Vi vet nå at de første 900 observasjonene ikke skal forkaste null hypotesen, mens de siste 100 skal gjøre det









Eksempel 1 – uten korreksjon

```
> X <- c(rnorm(900), rnorm(100,mean=3))
> p<- pnorm(X,lower.tail = F)
> test <- p> 0.05
> summary(test[1:900])
   Mode   FALSE    TRUE
logical      36     864
> summary(test[901:1000])
   Mode   FALSE    TRUE
logical      92       8
>
```

- Type I feil blir 36/900 = 0,04 (H_0 : x = 0 *er sann her*)
- Type II feil 8/100 = 0,08 (H_0 : x = 0 *er ikke sann*) *(styrke>90%)*
- Dette vet vi siden vi laget datasettet nettopp slik

Eksempel 1 – med Bonferroni korreksjon

Med nivå lik 0,05 og 1000 tester blir Bonferroni korreksjon å studere p-verdier mindre enn 0,00005

```
> bonftest <- p> 0.00005
> summary(bonftest[1:900])
   Mode   FALSE    TRUE
logical       1     899
> summary(bonftest[901:1000])
   Mode   FALSE    TRUE
logical      14      86
>
```

Type I feil blir 1/900 = 0,0011 og Type II feil 86/100 = 0,86 *(styrke<20%)*

Vi har redusert antall falske positive på bekostning av falske negative.

False Discovery Rate (FDR)
Benjamini-Hochberg (BH) threshold (en av mange metoder)

- Sett en øvre grense for *alfa*, for eksempel 0,05
- Beregne alle *m p-verdier*
- Sorter *p-verdiene* fra minst til størst: *p(1), p(2),…, p(m)*
- Plott disse mot sin rang *y* (rekkefølge)
- Estimer antall sanne H_0 ut fra plottet *(=m_0)*
- Finn den høyeste rangen *y* slik at *p(y) <= y/m_0* og forkast alle hypoteser med *p-verdi <= p(y)*

Eksempel 1 – The False Discovery Rate (FDR)

```
> psort <-sort(p)
> fdrtest <-NULL
> for (i in 1:1000)
+   fdrtest <- c(fdrtest, p[i]>match(p[i], psort) * 0.05/1000)
> summary(fdrtest[1:900])
   Mode   FALSE    TRUE
logical       6     894
> summary(fdrtest[901:1000])
   Mode   FALSE    TRUE
logical      69      31
```

Type I feil blir 6/900 = 0,00667 og Type II feil 31/100 = 0,31 *(styrke = 70%)*

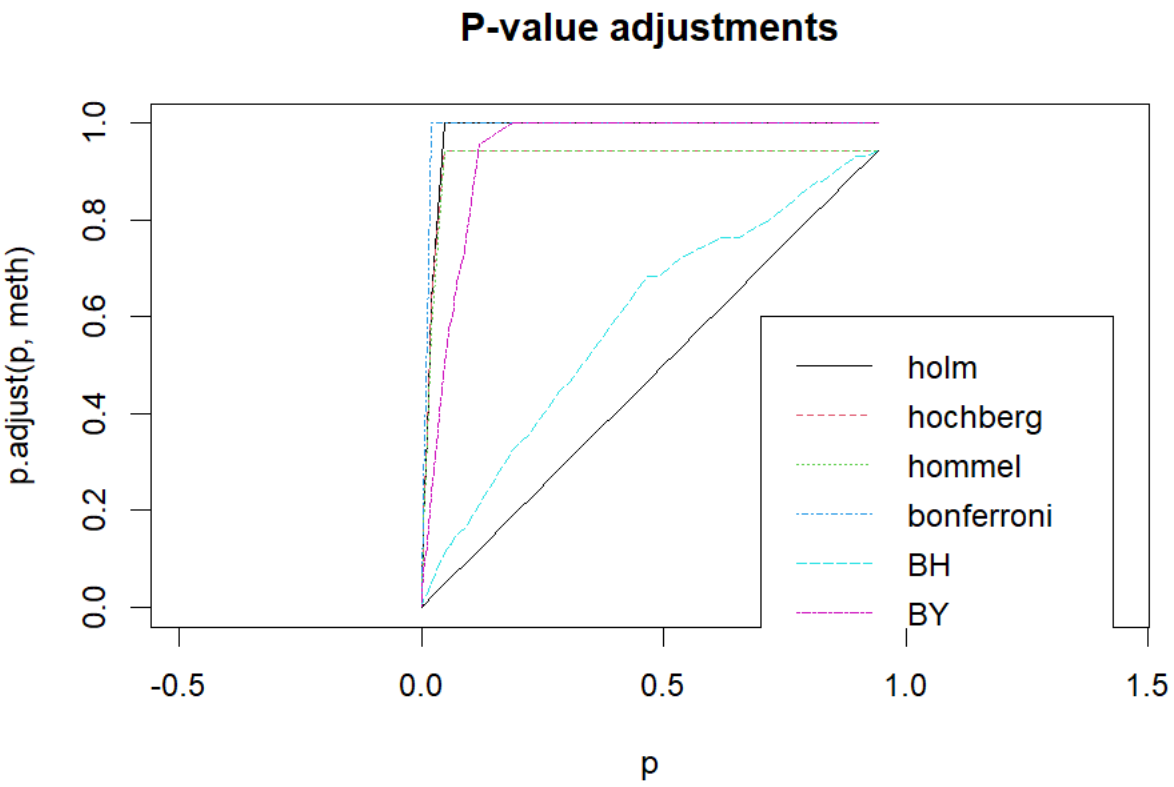
Eksempel 1 – videre

- Kan gjøre samme eksempel for andre signifikansnivåer (for eksempel 0,01, 0,025 og 0,10) og studere resultatene for Type I og Type II feil
- Husk at det vil endre seg noe når du generer data på nytt
- Programvare brukt her er R Foundation for Statistical Computing https://www.r-project.org/
- Kan også bruke SPSS (General Linear Model – Multivariate – Post Hoc)
- STATA, SAS og andre

Mer om muligheter i R

- Funksjon p.adjust() – Adjust P-values for Multiple Comparisons
- Inkluderer de mest brukte metodene
    - FWER (Family-wise error rate)
        - *Bonferroni (1979)*
        - *Holm (1979)*
        - *Hochberg (1988)*
        - *Hommel (1988)*
    - FDR (False discovery rate)
        - *Benjamini og Hochberg (1995)*
        - *Benjamini og Yekutieli (2001)*

Mer om muligheter i R



Mer om muligheter i R and Stata

https://rviews.rstudio.com/2019/10/02/multiple-hypothesis-testing/

(Roland Stevenson, 2019)

https://blogs.worldbank.org/en/impactevaluations/updated-overview-multiple-hypothesis-testing-commands-stata

(David Mckenzie, 2021)

Eksempel med ekte data

- Sammenligner pasienters BMI ved 41 sykehus i Norge
- Data fra Norsk Ryggregister, 18 390 pasienter
- Analyserer med enveis ANOVA i SPSS
- H_0 : Alle sykehus har pasienter med lik BMI
- SPSS resultater
  - F=3,422, p-verdi <<< 0,001 og H_0 forkastes
- Konklusjon: Det er minst to forskjellige sykehus

Tests of Between-Subjects Effects

Dependent Variable: BMI

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2598,541[a]	40	64,964	3,422	,000
Intercept	1987773,125	1	1987773,125	104710,180	,000
AvdNavn	2598,541	40	64,964	3,422	,000
Error	311900,070	16430	18,984		
Total	12276051.60	16471			
Corrected Total	314498,611	16470			

a. R Squared = ,008 (Adjusted R Squared = ,006)

BMI eksempel (forts.)

- Med 41 sykehus kan det kan utføres 820 parvise sammenligninger
- 146 p-verdier <= 0,05, 18 % av alle p-verdier
- 28 p-verdier <= 0,000061, Bonferroni korreksjon
- 30 p-verdier er signifikante ved å bruke Tukey metoden
- 57 p-verdier er under FDR grensen

Oppsummering

- *m* multiple tester
- Type I feil: Forkaste sann H_0
- Signifikansnivå = 0,05 -> forventer 5% Type I feil
- Family-wise error rate (FWER)
  - *FWER = P(minst en Type I feil) <= m(signifikansnivå)*
- Kontrollerer *FWER* ved **å senke signifikansnivået**
- Flere metoder for å kontrollere FWER (i R, Stata, SPSS)
- False discovery rate (FDR) **kontrollerer andelen Type I feil**

Oppsummering (forts.)

- For stor *m* (mange hypoteser): *FWER << alfa*
- Derfor ikke anbefalt å bruke når vi har tusenvis av tester
- FDR kontrollerer andelen Type I feil blant alle forkastede
- FDR = V/R hvor
    - R = totalt antall forkastede (av alle *m* hypoteser)
    - V = antall sanne H_0 som forkastes, Type I feil

Og litt til

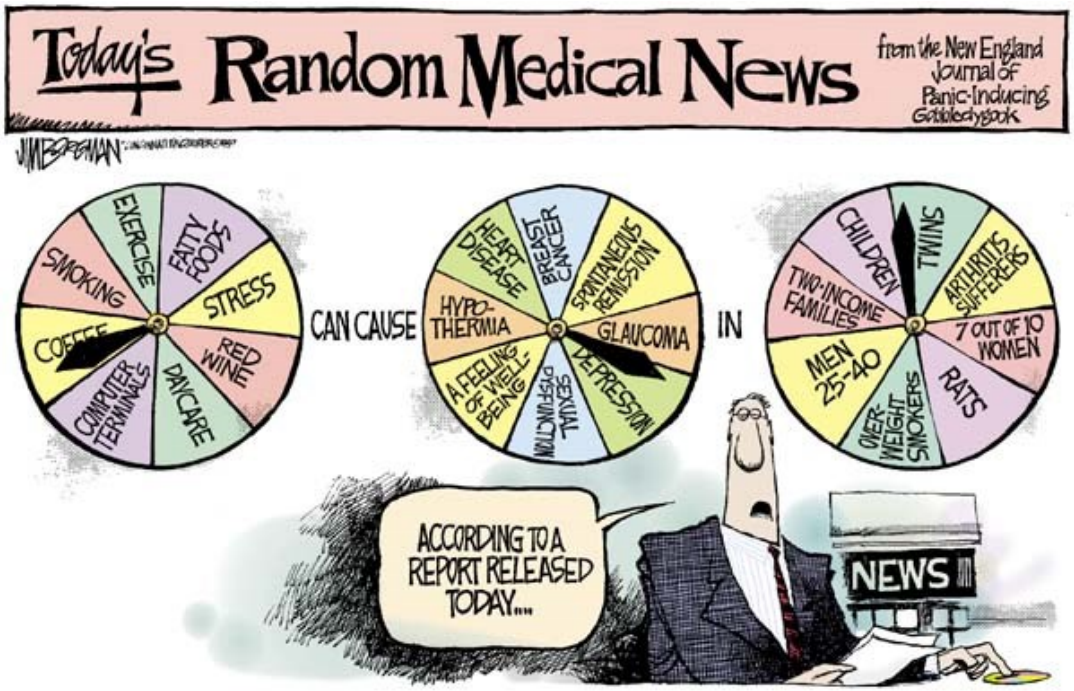
- Vær varsom når du skal utføre multippel testing
- Det minste du kan gjøre er å redusere signifikansnivået til hver test
- Vær ryddig i oppsettet
- Medfører statistisk signifikans alltid klinisk relevans?
- Kritisk blikk på hypotesetesting (Robert Coe, Durham University):
  - Effektstørrelse tas ikke hensyn til i hypotesetesting
  - Hypotesetesting tar ikke hensyn til apriorisk kunnskap
  - Restriksjoner på utvalgsstørrelse
  - Grunnløse antagelser

Skal man alltid justere for multiple hypoteser?

- Med flere utfallsvariable, definere den primære
  - Legge mindre vekt på funn ved sekundære utfallsvariable
  - Kenneth Rothman; *No adjustments are needed for multiple comparisons. Epidemiology 1990; 1: 43–6.*
- Bruke RECORD (https://www.record-statement.org/) og STROBE (https://www.strobe-statement.org/) guidelines
- Det finnes ingen allmenn konsensus om når/hvordan man bør justere for multiple hypoteser
- Valg av fremgangsmåte må spesifiseres i protokollen på forhånd, for å unngå «fisking» etter signifikante funn

Eksempler hvor det er aktuelt med multippel testing

- Forskjellige behandlingsopplegg
- Levetider for forskjellige merker av proteser
- PROMs som i spørreskjema for pasientgrupper

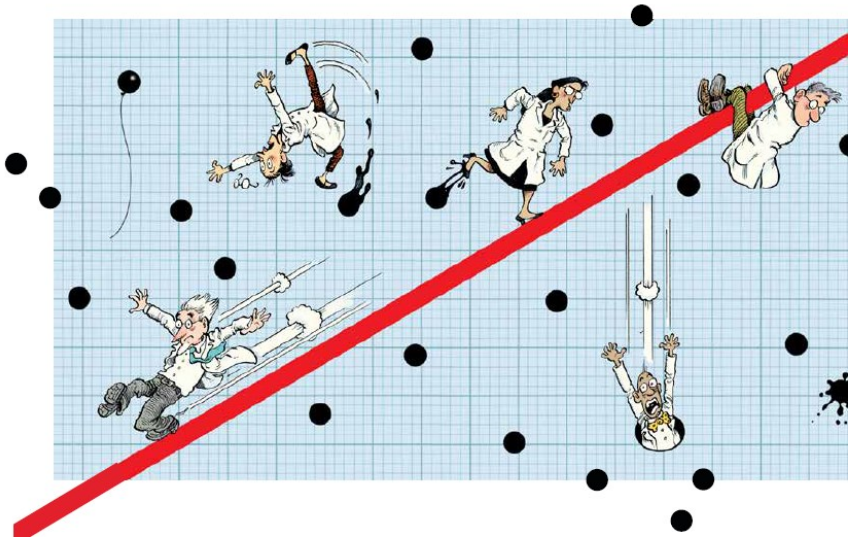


Kort oppsummert

- Mangfold (multiple problemer) har multiple perspektiver og løsninger
- Mangfold er ikke «bare et problem» som leder til mindre signifikante resultater, men også en mulighet for å oppdage nytt/mer
- Vurder forskjellige metoder som justerte p-verdier, FN, FD, FDR osv.
- Dagens metoder blir mer fleksible og gir mer informasjon

Ref: Yudi Pawitan og Arvid Sjölander, Karolinska Institutet, 2015.

Datamishandling



- Dårlig datagrunnlag
- Feil statistisk modell
- Ut på fisketur
- Blande statistisk signifikans og relevans
- Simpsons paradoks
- Blandet sammenheng og årsak
- Ufullstendig rapportering

Løsning: Ring en statistiker!

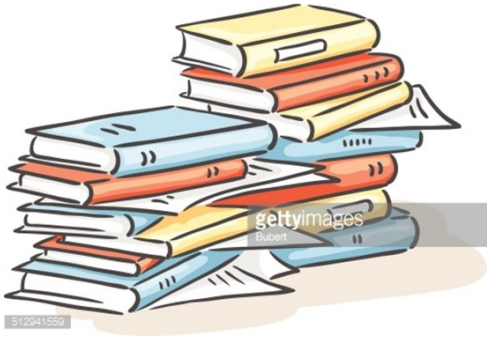
Forskningsetikk nr. 4, 2018. Tekst: Grønli KS. Illustrasjon:David Parkins for Nature

Take home message

- Always act carefully when performing multiple tests
- The least one can do is to reduce the significance level for each test

Litteraturforslag vår 2024

- Forelesningsnotater
- Lydersen S. *Justering av p-verdier ved multiple hypoteser*. Tidsskr Nor Legeforen 2021 doi:10.4045/tidsskr.21.0357
- Lydersen S. *Type I-feil og type II-feil*. Tidsskr Nor Legeforen 2021; 141. doi: 10.4045/tidsskr.21.0013



Litteraturoversikt – videre lesing

- Røislien J. Statistics lessons from a salmon. Tidsskr Nor Laegeforen. 2023 Sep 22;143(13). doi: 10.4045/tidsskr.23.0471
- Glickman, Mark E, Sowmya R Rao, Mark R Schultz. *False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies".* Journal of clinical epidemiology 2014; 67.8, s. 850-857
- Lydersen S. *Justering av p-verdier på norsk*. Tidsskr Nor Legeforen 2021; 141. doi: 10.4045/ tidsskr.21.0360
- Skovlund E. *Spør først, regn siden*. Tidsskr Nor Legefoen 2013; 133:10. doi: 10.4045/tidsskr.12.1345
- Mills JL. *Data torturing*. N Engl J Med 1993; 329: 1196-9

Takk

Nasjonalt Register for Leddproteser

HELSE BERGEN
Haukeland universitetssjukehus