En intro til radiologisk statistik

Erik Morre Pedersen

- Hypoteser og testning
- Statistisk signifikans
- 2 x 2 tabellen og lidt om ROC
- Inter- og intraobserver statistik
- Styrkeberegning
- Konklusion
- Litteratur

Hypoteser I

Varierer "noget" fra hinanden???
opstille hypoteser, man kan teste.

- *0-hypotesen:*
  - ➡ Der er ingen forskelle i middelværdi mellem populationen og de udtrukne (f.eks. raske vs syge).
  - ➡ En evt. forskel kan forklares ved tilfældigheder
- Den *alternative hypotese:*
  - ➡ Der er reel forskel I middelværdien mellem de to grupper
- Pointe:
  - ➡ *Det er mere realistisk at "bevise" at nulhypotesen ikke passer, end at bevise at den alternative hypotese er sand.*

Hypotesetestning

Hypotesetestning giver to mulige konklusioner:

- Det er *usandsynligt* at forskellen I middelværdi kun skyldes tilfældigheder
  - ➡ Vi forkaster nulhypotesen og konkluderer at grupperne må være forskellige
- Forskellen kan forklares ved tilfældigheder alene
  - ➡ Vi accepterer nulhypotesen – I det mindste for nuværende .

- Men: Vi kan tage fejl I begge tilfælde!

Mulige resultater af hypotesetestning

"Sandheden"

Statistisk test:	0-hypotesen sand (grupper ikke forskellige)	0-hypotesen falsk (grupper er forskellige)
Behold 0-hypotesen	Korrekt beslutning	Type II fejl (β)
Forkast 0-hypotesen	Type I fejl ()	Korrekt beslutning

Type I og type II fejl

- **Type I fejl** ($\checkmark$) : 0-hypotesen forkastes selvom den er sand
  - ➡ Vores test viser fejlagtigt at noget er forskelligt, selvom det ikke er det i virkeligheden.
    - Ex: Vi konkluderer fejlagtigt at CT er bedre end UL til at påvise galdesten.

- **Type II fejl** (β) : 0-hypotesen forkastes ikke selvom den er falsk
  - ➡ Vores test viser fejlagtigt at der ikke er forskel mellem to grupper, selvom de er forskellige
    - Ex. Vi konkluderer fejlagtigt at CT ikke er bedre end konventionel rtg til påvisning af fri luft

Begge fejl kan altid forekomme, men man kan designe sit studie,så risikoen for dette minimeres.

P-værdi

- “Sandsynligheden for at vores statistiske test antager en værdi som er mere ekstrem end den vi har observeret I vores data”
- P-værdien ligger langt fra hvad vi forventer at finde (0-hypotesen siger jo at der ikke er forskel)
- Jo mindre p-værdi, jo stærkere bevis for at vores 0-hypotese ikke er sand = grupperne er forskellige
- Mange statistik programmer beregner p-værdien direkte. Beregnes manuelt fåes en Z-fordeling der skal oversættes til en p-værdi.

Statistisk signifikans

- Hvor meget bevis behøver vi for at være sikre?
  - ➡ Afgør hvor mange ressourcer vi skal bruge
- Definition: ✔ = signifikans niveau
- Hvis P-værdien er så lille som eller mindre end ✔, siger vi at data er signifikant forskellige ved signifikansniveau ✔.
- Betyder, at det ikke er sandsynligt at resultatet skyldes tilfældigheder
- Signifikansniveau *vælges* I medicinsk forskning som regel til 5% ($p < 0,05$)

Styrkeberegning

- Styrkeberegning foretages FØR et studie for at undgå type II fejl.
- Vi repeterer:
  - ➡ Type II fejl: chancen for fejlagtigt at antage at grupperne *ikke* er forskellige selvom de er det
    - Typisk: Et studie viser ingen signifikant forskel mellem grupperne. Flere muligheder
      - har vi undersøgt for få pt til at påvise forskellen?
      - Er vores metoder for upræcise (for meget variation/støj i målingerne)?
- Styrken (power) =: 1- risiko for type 2 fejl (β)
  - ➡ http://wise.cgu.edu/powermod/power_applet.asp

POWER

Styrkeberegning

- **Faktorer, der indgår I styrkebegning (forsimplet):**
  - **Patientpopulation ("n"):** Letteste faktor at manipulere. Jo flere patienter jo større styrke/mindre type II fejl.
  - **Forskellen I middelværdi** man leder efter. Det er sværere at finde små forskelle. *Det er altså nødvendigt at have et bud på forskellen før man starter studiet!*
  - **Variationen af målingerne.** Hvis individerne/målingerne varierer meget indenfor en gruppe går styrken ned.
  - **Niveauet for P-værdien (signifikansniveauet).** Hvis niveauet sættes til 0,01 I stedet for 0,05 vil det være sværere at afvise, at grupperne er ens og styrken vil blive mindre.

Eng J, Radiology 2003; 227: 309-313

POWER

2 x 2 tabeller

- Typisk radiologisk statistisk problemstilling:
  - ➡ Vi laver en billeddiagnostisk test og ønsker at vide:
    - Hvor sikre er vi på at syge pt ikke fejlagtigt erklæres raske ? (sensitivitet)
    - Hvor sikre er vi på at raske pt ikke fejagtigt erklæres syge ? (specificitet)
    - Hvis testen er positiv, hvor stor er sandsynligheden for at pt er syg (positiv prædiktiv værdi –PPV)
    - Hvis testen er negativ, hvor sikre er vi så på at patienten virkelig er rask (negativ prædiktiv værdi – NPV)

- Disse spørgsmål kan regnes på i en 2 x 2 tabel!

2 x 2 tabel: Diagnostisk performance

Examination Result	D+	D−	Total
T+	TP	FP	TP + FP
T−	FN	TN	FN + TN
Total	TP + FN	FP + TN	*N*

- Sensitivitet (SEN)= TP / (TP + FN)
- Specificitet (SPEC)= TN / (FP + TN)
- Positiv prædiktiv værdi (PPV)= TP / (TP + FP)
- Negativ prædiktiv værdi (NPV)= TN / (FN + TN)
- TP = true positive, TN = true negative, FP= false positive, FN = false negative

Eksempel data 2 x 2 tabel

Patient Data in Experiment to Study Breast MR Imaging

MR Imaging Result	Malignant	Benign	Totals
Positive	71	28	99
Negative	3	80	83
Total	74	108	182

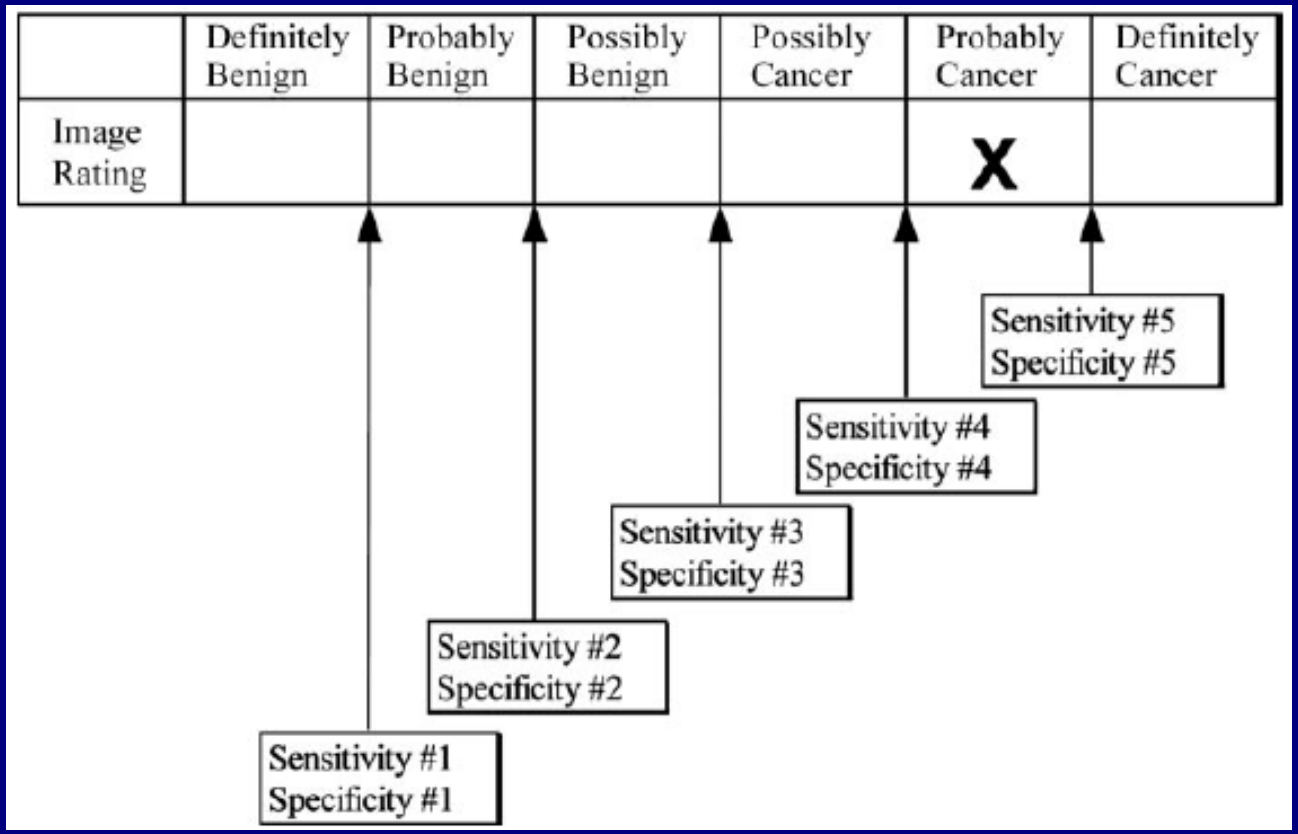
Note.—Data are numbers of women with malignant or benign breast tumors.

Langlotz CP, Radiology 2003; 228: 3-9

Receiver operating characteristics (ROC)

- Anvendelsen af begreberne sand og falsk kræver et entydigt ”cut-off”
- Ofte er dette umuligt i praksis:
  - ➡ Forskellige radiologer læser undersøgelser forskelligt
  - ➡ Det giver bedre mening i praksis at have gråzoner: ”potentielt malign” etc.
- Flytter man grænsen mellem sand og falsk vil værdierne for såvel sens som spec. ændres (indbyrdes afhængige)
- Optegning af ROC-kurver kan give et indtryk af testens kvaliteter uafhængig af disse ”cut off” settings.

ROC: mere end én ”cut-off” værdi



Langlotz CP, Radiology 2003; 228: 3-9

ROC: for hvert "cut-off" laves en 2 x 2 tabel

	Definitely Benign	Probably Benign	Possibly Benign	Possibly Cancer	Probably Cancer	Definitely Cancer	Totals
Cancer Cases	2	3	5	10	30	50	100
Non-Cancer Cases	50	30	10	5	3	2	100
Totals	52	33	15	15	33	52	100

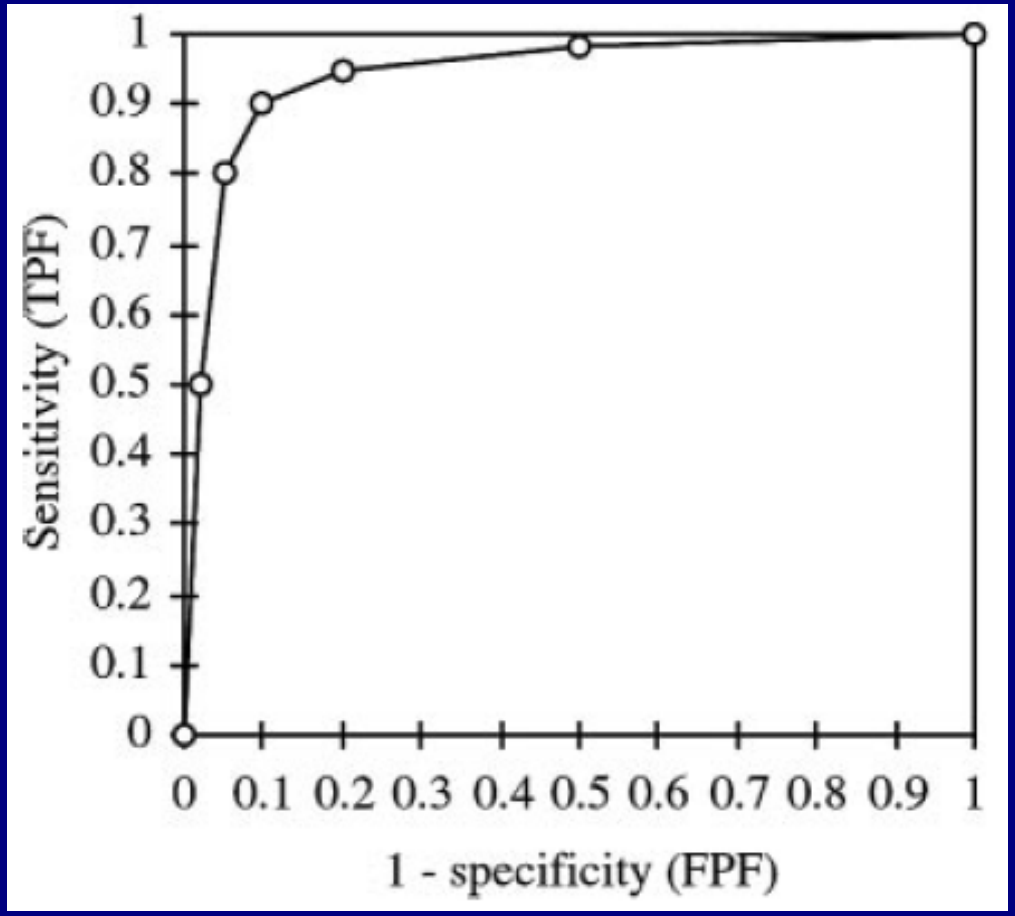
	D+	D-	Totals
T+	98	50	148
T-	2	50	52
Totals	100	100	200

Sensitivity #1 = 98%
Specificity #1 = 50%

Langlotz CP, Radiology 2003; 228: 3-9

ROC-kurven

- Jo mindre areal i øverste venstre hjørne jo bedre er undersøgelsen
- Alle punkterne repræsenterer de samme data, men med forskellige ”cut-off” for sand og falsk
- Bemærk: Dit valg af diagnostisk kriterie (”cut-off”) er afgørende for sens/spec!



Langlotz CP, Radiology 2003; 228: 3-9

Inter- og intraobserver statistik

- I forskning hvor der indgår ”imaging” er en stor del af data fremkommet ved subjektiv vurdering af billeder.
- Interessant:
  - ➡ Hvor god er den billeddiagnostiske metode
  - ➡ Hvad viser undersøgelsen med så objektive briller som muligt
- Uinteressant:
  - ➡ Hvor god er radiologen/observatøren?

Inter- og intraobserver statistik

- Afhængigheden (bias) af den vurderende observatør (”agreement”) kan studeres ved gentagen vurdering af billederne (altid i blindet form):
  - ➡ Af den samme observatør (intraobservatør afvigelse)
  - ➡ Af flere observatører (interobservatør afvigelse

God og dårlig agreement

- Agreement udregnes ofte med ”kappa” statistik (se reference nedenfor), der tager hensyn til tilfældigt sammenfald.

Kappa = 0,31

TABLE 1
Joint Judgment of Two Readers about Same 150 Images

	Second Reader		
First Reader	Positive for Disease	Negative for Disease	Total
Positive for disease	7	10	17
Negative for disease	12	121	133
Total	19	131	150

Guidelines for Strength of Agreement Indicated with κ Values

κ Value	Strength of Agreement beyond Chance
<0	Poor
0–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Kundel HL, Polansky M. Radiology 2003; 228: 303-308

Konklusion

- Statistik er en yderst vigtig del af radiologisk (og al anden!) forskning, og man er nødt til at forholde sig til det hvis man vil lave forskningsprojekter
- Statistik starter når man *planlægger* et projekt!!!!
  - ➡ Hvor mange pt skal jeg undersøge for at finde ud af det jeg vil? – og hvordan skal de undersøges?
- Kontakt en biostatistiker (findes ved alle sundhedsvidenskabelige fakulteter) før du starter dit projekt.
- Det koster måske penge og du skal sandsynligvis undersøge flere pt end du regnede med. Men sammenlign det med al den tid, besvær og penge, du bruger på projektet iøvrigt.

Litteratur

- Statistik serie i ”*Radiology*”
  - ➡ http://pubs.rsna.org/page/radiology/sections

  Vælg ”Statistical and data analysis” under ”Reviews and Commentaries” sektionen
- Power beregnings visualisering:
  - ➡ http://wise.cgu.edu/portfolio/demo-statistical-power/
- Hypotesetestnings visualisering:
  - ➡ http://wise.cgu.edu/portfolio/demo-hypothesis-testing/

Øvelse

- Beregn i 2 x 2 tabel:
  - ➡Sensitivitet, Specificitet, Positiv prædiktiv værdi og Negativ prædiktiv værdi for bryst cancer data i slide 13
- Power beregnings visualisering:

http://wise.cgu.edu/portfolio/demo-statistical-power/