

# Biostatistik og Epidemiologi på modul 4

Jacob Krabbe Pedersen (5 forelæsninger om statistik)  
Epidemiologi, Biostatistik, og Biodemografi, IST, SDU

## Basis for formelsamling:

- A) Formler fra forelæsninger
- B) Placeret under relevant scenarie (/model)

## "Indholdsfortegnelse":

- Overblik: Statistiske scenarier/modeller Slide 2
- Overblik: Sandsynlighedsteoretiske scenarier/emner Slide 3
- Formler: Én numerisk stikprøve Slide 4-5
- Formler: To numeriske stikprøver Slide 6-7
- Formler: Én binær stikprøve' Slide 8
- Formler: To numeriske stikprøver Slide 9-10
- Formler: Association i en  $r \times s$ -tabel Slide 11
- Formler: Diagnostik og screeningsredskaber Slide 12

# Statistikforelæsningerne

## Én numerisk stikprøve

- *Deskriptiv statistik for en (numerisk) stikprøve.*
- *Konfidensinterval for middelværdien  $\mu$*
- *Prædiktionsinterval*

## To numeriske stikprøver

- *Uparret t-test for  $H_0: \mu_1 = \mu_2$*
- *Konfidensinterval for differensen mellem middelværdierne*

## Én binær stikprøve

- *Konfidensinterval for en proportion (sandsynlighedsparameter)*

## To binære stikprøver:

- $\chi^2$ -test for  $H_0: \pi_1 = \pi_0$ , eller (mere generelt) ingen sammenhæng mellem række- og søjleinddeling
- Større tabeller:  $\chi^2$ -test i  $r \times s$ -tabel for  $H_0$ : ingen sammenhæng mellem række- og søjleinddeling

**NB:** I alle 4 scenarier/modeller ovenfor indgår fælles elementer, f.eks. *frihedsgrader* og *standardafvigelse* (se), men formlerne for beregning af disse (og fortolkning) er forskellig og afhænger af det specifikke scenarie.

# Opsummering af sandsynlighedsteori ved statistikforelæsningerne:

## Bayes formel (generelt)

- (formaliserede) sandsynligheder  $P(\cdot)$ .
- Betingede sandsynligheder
- Bayes formel

---

## Anvendelser af Bayes formel

- Genetik: monogenetiske, autosomale sygdomme (dominant eller recessivt nedarvet)
- Diagnostik: positiv prædiktiv værdi (PPV) på baggrund af et screeningsredskabs sensitivitet og specificitet, samt på prævalensen i den population der screenes.

---

## Relevans af ovenstående ved eksamen på M4: biostatistik og epidemiologi

- |                                  |   |
|----------------------------------|---|
| ➤ Formaliserede sandsynligheder: | der eksamineres ikke selvstændigt i dette emne  |
| ➤ Betingede sandsynligheder:     | der eksamineres ikke selvstændigt i dette emne  |
| ➤ Bayes formel i genetik:        | der eksamineres ikke selvstændigt i dette emne  |
| ➤ Bayes formel i diagnostik:     | indgår i eksamen via den specifikt afledte version af Bayes formel til dette scenarie |

---

**NB:** Forståelsen af sandsynlighedsbegrebet er grundlæggende for at kunne forstå og anvende Bayes formel. Anvendelse af Bayes formel i genetik er en del af dette kursus, fordi genetikkurset har bedt os om at gennemgå denne formel i relation til genetiske anvendelser.

## Én numerisk stikprøve - Prædiktionsinterval

### ➤ **Beregning:**

$$95\% - PI = \bar{x} \pm 1,96 \cdot SD$$

### ➤ **Fortolkning:**

- 95% af alle observationer i populationen ligger indenfor prædiktionsintervallets grænser
- I klinikken: et (passende) prædiktionsinterval angiver de målinger der ikke anses for unormale/ekstreme
- F.eks. normal vs. unormal fødselsvægt

### ➤ **Forudsætninger for formel:**

- Uafhængige, normalfordelte observationer med samme måleusikkerhed (se slides til anden statistikforelæsning, slides 2-3)

### ➤ **I K&S:** også kaldet "reference range" (s. 47-49, 56-57)

### **Forudsætninger for formel:**

- Uafhængige, normalfordelte observationer med samme måleusikkerhed

## En numerisk stikprøve - Konfidensinterval for middelværdien $\mu$

NB: Der findes to forskellige formler: En approksimativ og en eksakt. Den eksakte må altid benyttes, men den approksimative kan bruges når  $n \geq 60$  (hvor approksimationen er ok)

### Beregning:

Et eksakt 95%-konfidensinterval for middelværdien  $\mu$  er givet ved

$$\bar{x} - t' \cdot se ; \bar{x} + t' \cdot se$$

hvor  $t'$  er det tosidede 5%-punkt i t-fordelingen med  $n - 1$  frihedsgrader

(i KS: side 54-55;  $t'$  findes ved opslag i Tabel A3)

Et (approksimativt) 95%-konfidensinterval for middelværdien  $\mu$  er givet ved

$$\bar{x} - 1,96 \cdot se ; \bar{x} + 1,96 \cdot se$$

I begge formler er  $se = se(\bar{x}) = SD/\sqrt{n}$

### Fortolkning af konfidensinterval

- Hvis vi trak 100 nye stikprøver og beregnede tilsvarende 100 konfidensintervaller for hver af disse, så ville vi forvente at 95 af dem indeholdt den sande middelværdi i populationen (jf. i afs. 6.3 og Fig. 6.2. i bogen).
- M.a.o.: der er 95% sandsynlighed for at et konfidensinterval omfavner den sande værdi  $\mu$  for middelværdien i hele populationen

### Forudsætninger for formel:

- Uafhængige, normalfordelte observationer med samme måleusikkerhed

## To numeriske stikprøver - det uparrede t-test

Nulhypotese:

- $H_0: \mu_1 = \mu_2$ , dvs. ingen forskel i middelværdien af udfaldet i de to grupper

Teststørrelse:

- $t(x) = \frac{\bar{x}_1 - \bar{x}_2}{se(\bar{x}_1 - \bar{x}_2)}$ , hvor  $se(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot SD_{fælles}$  og

$$SD_{fælles} = \sqrt{\frac{(n_1-1) \cdot SD_1^2 + (n_2-1) \cdot SD_2^2}{n_1 + n_2 - 2}}$$

Frihedsgrader:

$$df = n_1 + n_2 - 2$$

P-værdi:

- Findes ved opslag af  $|t(x)|$  (den numeriske værdi af  $t(x)$ ) i tabellen for T-fordelingen (Tabel A3) i rækken med det passende antal frihedsgrader

Konklusion:

- Ved en p-værdi under 5% forkastes nulhypotesen. Men finder altså en signifikant sammenhæng mellem eksponering og udfald.
- Ved en p-værdi over 5% kan nulhypotesen ikke forkastes. Man kan altså ikke afvise at der ikke er nogen sammenhæng mellem eksponering og udfald

## To numeriske stikprøver – konfidensintervallet for forskellen på middelværdierne

NB: Der findes to forskellige formler: En approksimativ og en eksakt. Den eksakte må altid benyttes, men den approksimative kan bruges når  $\min(n_1, n_2) \geq 30$  (hvor approksimationen er ok)

Eksakt formel:

- $95\% - CI(\mu_1 - \mu_2) = \bar{x}_1 - \bar{x}_2 \pm t' \cdot se(\bar{x}_1 - \bar{x}_2)$ ,  
hvor  $t'$  er det tosidede 5%-punkt i T-fordelingen med  $df = n_1 + n_2 - 2$  frihedsgrader.

Approksimativ formel:

- $95\% - CI(\mu_1 - \mu_2) = \bar{x}_1 - \bar{x}_2 \pm 1,96 \cdot se(\bar{x}_1 - \bar{x}_2)$ ,

I begge formler er  $se(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot SD_{fælles}$  og  $SD_{fælles} = \sqrt{\frac{(n_1-1) \cdot SD_1^2 + (n_2-1) \cdot SD_2^2}{n_1 + n_2 - 2}}$

Konklusion:

- Ingen forskel på middelværdierne svarer til en differens på 0
- Hvis konfidensintervallet derfor ikke indeholder værdien 0, er der en signifikant forskel på middelværdierne, og derfor en sammenhæng mellem eksponering og udfald (en difference på 0 er ikke plausibel på baggrund af studiet).
- Hvis konfidensintervallet indeholder værdien 0, er der ikke en signifikant sammenhæng mellem eksponering og udfald (en difference på 0 er plausibel på baggrund af studiet).
- Mere generelt indeholder konfidensintervallet de værdier (for differencen), der er plausible på baggrund af studiet.

## En binær stikprøve - Konfidensinterval for proportionen $\pi$

NB: Der præsenteres en formel, som er approksimativ. Den kan bruges, når både  $n \cdot p \geq 10$  og  $n \cdot (1 - p) \geq 10$ , hvor  $n$  er antal observationer og  $p$  angiver andelen af cases

### Beregning:

Et (approksimativt) 95%-konfidensinterval for proportionen  $\pi$  er givet ved

$$p - 1,96 \cdot se ; p + 1,96 \cdot se$$

hvor  $se = se(p) = \sqrt{\frac{p \cdot (1-p)}{n}}$

### Fortolkning af konfidensinterval

- Hvis vi trak 100 nye stikprøver og beregnede tilsvarende 100 konfidensintervaller for hver af disse, så ville vi forvente at 95 af dem indeholdt den sande proportion i populationen.
- M.a.o.: der er 95% sandsynlighed for at et konfidensinterval omfavner den sande værdi  $\pi$  for proportionen i hele populationen

### Forudsætninger for formel:

- Uafhængige observationer fra en binær population



**To binære stikprøver -  $\chi^2$ -testet for  $H_0: \pi_1 = \pi_0$ ,**  
eller (mere generelt) ingen sammenhæng mellem række- og søjleinddeling i en  $2 \times 2$ -tabel

<u>Observeret tabel:</u>	Syge	Raske	Rækketotal
<b>Eksponerede</b>	$d_1$	$h_1$	$n_1$
<b>Ueksponerede</b>	$d_0$	$h_0$	$n_0$
<b>Søjletotal</b>	$d$	$h$	$n$

Nulhypotese:

- $H_0: \pi_1 = \pi_0$ , dvs. ingen forskel i risikoen i de to grupper (eller: ingen sammenhæng mellem eksponering og udfald)

Teststørrelse – version 1 (hurtigformel):

➤ 
$$X^2 = n \cdot \frac{(d_1 \cdot h_0 - d_0 \cdot h_1)^2}{d \cdot h \cdot n_1 \cdot n_0}$$

Frihedsgrader ( $2 \times 2$ -tabel):

➤ 
$$df = (2 - 1) \cdot (2 - 1) = 1$$

P-værdi:

- Findes ved opslag af  $X^2$  i tabellen for  $\chi^2$ -fordelingen (Tabel A5) i rækken med det passende antal frihedsgrader


Konklusion:

- Ved en p-værdi under 5% forkastes nulhypotesen. Men finder altså en signifikant sammenhæng mellem eksponering og udfald.
- Ved en p-værdi over 5% kan nulhypotesen ikke forkastes. Man kan altså ikke afvise at der ikke er nogen sammenhæng mellem eksponering og udfald

NB: Der præsenteres en formel, som er approksimativ. Den kan bruges, når alle forventede antal i  $2 \times 2$ -tabellen er på mindst 5

**To binære stikprøver -  $\chi^2$ -testet for  $H_0: \pi_1 = \pi_0$ ,**  
 eller (mere generelt) ingen sammenhæng mellem række- og søjleinddeling i en  $2 \times 2$ -tabel

<u>Observeret tabel:</u>	Syge	Raske	Rækketotal	<u>Forventede antal</u>	Syge	Raske	Rækketotal
<b>Eksponerede</b>	$d_1$	$h_1$	$n_1$	<b>Eksponerede</b>	$e_{11}$	$e_{12}$	$n_1$
<b>Ueksponerede</b>	$d_0$	$h_0$	$n_0$	<b>Ueksponerede</b>	$e_{21}$	$e_{22}$	$n_0$
<b>Søjletotal</b>	$d$	$h$	$n$	<b>Søjletotal</b>	$d$	$h$	$n$

"forventede" =  $\frac{\text{rækketotal} \cdot \text{søjletotal}}{\text{total}}$  " 

Det forventede antal  $e_{ij}$  i cellen i række i, søjle j bliver:

$$e_{ij} = \frac{(\text{række } i\text{'s total}) \cdot (\text{søjle } j\text{'s total})}{n}$$

Nulhypotese:

- $H_0: \pi_1 = \pi_0$ , dvs. ingen forskel i risikoen i de to grupper (eller: ingen sammenhæng mellem eksponering og udfald)

Teststørrelse – version 2 (observerede vs forventede):

➤ 
$$X^2 = \sum_{\text{alle 4 celler}} \frac{(\text{observeret antal} - \text{forventet antal})^2}{\text{forventet}}$$
 (nb: "Σ" står for "sum over" (dvs., alle 4 celler))

Frihedsgrader ( $2 \times 2$ -tabel):

➤  $df = (2 - 1) \cdot (2 - 1) = 1$

P-værdi:

- Findes ved opslag af  $X^2$  i tabellen for  $\chi^2$ -fordelingen (Tabel A5) i rækken med det passende antal frihedsgrader

Konklusion:

- Ved en p-værdi under 5% forkastes nulhypotesen. Men finder altså en signifikant sammenhæng mellem eksponering og udfald.
- Ved en p-værdi over 5% kan nulhypotesen ikke forkastes. Man kan altså ikke afvise at der ikke er nogen sammenhæng mellem eksponering og udfald

NB: Der præsenteres en formel, som er approksimativ. Den kan bruges, når alle forventede antal i  $2 \times 2$ -tabellen er på mindst 5

# Opsummering af sammenhæng mellem kategoriske variable:

## To binære stikprøver (en $2 \times 2$ -tabel):

- $\chi^2$ -test for  $H_0: \pi_1 = \pi_0$ , eller (mere generelt) ingen sammenhæng mellem række- og søjleinddeling
- Frihedsgrader:  $df = (r - 1) \cdot (s - 1) = (2 - 1) \cdot (2 - 1) = 1$
- P-værdi findes ved opslag af den beregnede  $X^2$  i tabellen for  $\chi^2$ -fordelingen med 1 frihedsgrad

## $r$ eksponeringer med $s$ udfald, eller, $r \times s$ eksponeringer (en $r \times s$ -tabel):

- $\chi^2$ -test i  $r \times s$ -tabel for  $H_0$ : ingen sammenhæng mellem række- og søjleinddeling
- Frihedsgrader:  $df = (r - 1) \cdot (s - 1)$
- P-værdi findes ved opslag af den beregnede  $X^2$  i tabellen for  $\chi^2$ -fordelingen med  $(r - 1) \cdot (s - 1)$  frihedsgrader

# Bayes formel (generelt)

For et sandsynlighedsmaal  $P(\cdot)$  og to hændelser  $A$  og  $B$  gælder Bayes formel:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A^C) \cdot P(A^C)}$$

(Teknisk betingelse: gælder så længe,  $0 < P(A) < 1$  og  $P(B) > 0$ ).

## Anvendelse af Bayes formel i diagnostik

- Ved screening for en sygdom (f.eks. COVID-19) er det velkendt at man kan teste "falsk positiv". Det er derfor af interesse at vide, i fald man tester positiv, hvor sandsynligt det er at man faktisk er syg.
- Denne sandsynlighed kaldes den positive prædiktive værdi (PPV), og den afhænger dels af den risikovurdering man har før man screener (altså, prævalensen af sygdommen i den population der screenes), men den afhænger også af selve screeningsredskabets egenskaber, dvs. dets evne til korrekt at identificere de syge (sensitiviteten) og dets evne til korrekt at identificere de raske (specificiteten).

Givet en specifik prævalens, sensitivitet og specificitet kan PPV beregnes ved formelen:

$$PPV = \frac{\textit{sensitivitet} \cdot \textit{prævalens}}{\textit{sensitivitet} \cdot \textit{prævalens} + (1 - \textit{specificitet}) \cdot (1 - \textit{prævalens})}$$