

Formelsamling i medisinsk statistikk

- Dette er en formelsamling til O. O. Aalen (red.): Statistiske metoder i medisin og helsefag, Gyldendal, 2006.
- Merk at boken har en nettside der det er lagt ut rettelser og supplerende stoff, se <http://www.med.uio.no/imb/stat/statbok/>

Gjennomsnitt

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n)$$

Median

Alle observasjoner ordnes i stigende rekkefølge. Ved *ulike* antall observasjoner, er medianen definert som den midterste av dem. Ved *like* antall, er medianen definert som gjennomsnittet av de to midterste.

Empirisk standardavvik

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Regneregler for sannsynlighet

Hvis begivenhetene A og B er disjunkte has

$$P(A \cup B) = P(A) + P(B)$$

For alle begivenheter A og B has

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Definisjon av *betinget sannsynlighet*

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

Begivenhetene A og B er *uavhengige* hvis

$$P(A \cap B) = P(A) \cdot P(B)$$

En tilsvarende produktregel er gyldig om vi har flere uavhengige begivenheter.

Regelen om *total sannsynlighet*

$$P(A) = P(A | B) \cdot P(B) + P(A | \overline{B}) \cdot P(\overline{B})$$

Bayes' lov

$$P(B | A) = \frac{P(B)P(A | B)}{P(B)P(A | B) + P(\overline{B})P(A | \overline{B})}$$

Binomisk fordeling

Sannsynligheten for at en begivenhet A inntreffer x ganger i løpet av n binomiske forsøk, er

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n$$

Forventning og varians i binomisk fordeling er gitt ved:

$$E(X) = np, \quad \text{Var}(X) = np(1-p)$$

Normalfordeling

En stokastisk variabel X sies å være normal (μ, σ) hvis den følger en normalfordeling med forventning (sentrum) μ og standardavvik (spredning) σ . Den standardiserte variable $Y = (X - \mu)/\sigma$ er normal $(0,1)$.

Formler for gjennomsnitt

La $\overline{X}$ være gjennomsnittet av de uavhengige variablene $X_1, X_2, \dots, X_n$. Da gjelder:

$$E(\overline{X}) = \mu, \quad \text{Var}(\overline{X}) = \frac{\sigma^2}{n}, \quad \text{SD}(\overline{X}) = \frac{\sigma}{\sqrt{n}}, \quad s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

Hvis variablene også er normalfordelte, vil et konfidensintervall være gitt ved

$$\overline{X} \pm c s_{\overline{X}}$$

der c bestemmes ut fra Studentfordelingen med $n - 1$ frihetsgrader.

En teststørrelse er gitt ved

$$t = \frac{\overline{X} - a}{s_{\overline{X}}} = \frac{\overline{X} - a}{s} \sqrt{n}$$

og denne er Studentfordelt med $n - 1$ frihetsgrader når $H_0: \mu = a$ gjelder.

Sammenlikning av pardata

Man tar differansen innenfor hvert par og bruker konfidensintervallet og teststørrelsen over med $a = 0$. Forutsetningen er at differansene er uavhengige og normalfordelte.

Sammenlikning av to gjennomsnitt

Vi forutsetter uavhengige og normalfordelte observasjoner. Forøvrig antas gjennomsnittene å komme fra to uavhengige utvalg. Følgende teststørrelse er Studentfordelt med $n_1 + n_2 - 2$ frihetsgrader når H_0 gjelder

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_f \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

der s_f er definert ved

$$s_f = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Et konfidensintervall er gitt ved

$$\bar{X}_1 - \bar{X}_2 \pm c s_f \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

der c er bestemt av Studentfordelingen med $n_1 + n_2 - 2$ frihetsgrader.

Normalfordeling som tilnærming til binomisk fordeling

Når n i en binomisk fordeling er så stor at $np \geq 5$ og $n(1 - p) \geq 5$, vil den binomiske fordelingen likne mye på en normalfordeling med parametre

$$\mu = np, \quad \sigma = \sqrt{np(1 - p)}$$

Estimering av sannsynlighet (andel)

Hvis det er observert X forekomster ved n binomiske forsøk, er estimatet for sannsynligheten p gitt ved $\hat{p}$, mens estimert standardfeil er gitt ved s_p

$$\hat{p} = X/n, \quad s_p = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Fordelingen til $\hat{p}$ er tilnærmet normalfordelt under de samme forutsetninger som for binomisk fordeling, med $\mu = p$ og $\sigma = \sqrt{\frac{p(1-p)}{n}}$.

Et 95% konfidensintervall for $\hat{p}$ er gitt ved

$$\hat{p} \pm 1.96s_p$$

Konfidensintervall for differanse mellom to andeler

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Krysstabell for eksponering og sykdom

	Ekspontert	Ikke ekspontert
Syk	a	b
Ikke syk	c	d

Konfidensintervall for relativ risiko

Estimat for relativ risiko:

$$\widehat{RR} = \frac{a/(a+c)}{b/(b+d)}$$

Hjelpestørrelse:

$$s' = \sqrt{\frac{1}{a} + \frac{1}{b} - \frac{1}{a+c} - \frac{1}{b+d}}$$

95% konfidensintervall for estimert relativ risiko:

$$(\widehat{RR} \times e^{-1.96 s'}, \widehat{RR} \times e^{1.96 s'})$$

Konfidensintervall for odds-ratio

Estimert odds-ratio:

$$\widehat{OR} = \frac{a/c}{b/d} = \frac{a \cdot d}{b \cdot c}$$

Hjelpestørrelse:

$$s'' = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

95% konfidensintervall for estimert odds ratio:

$$(\widehat{OR} \times e^{-1.96 s''}, \widehat{OR} \times e^{1.96 s''})$$

Kji-kvadrattest

Kji-kvadrattesten for en 2×2 -tabell kan beregnes ut fra følgende formel:

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)},$$

der $N = a + b + c + d$. Formelen er ikke gitt i boken, men gir samme svar som beregningen av størrelsen S på s. 138.

Regresjonsanalyse

Helningskoeffisienten, b , og skjæringspunktet med y -aksen, a , for minste-kvadraters-linjen er gitt ved

$$\hat{b} = s_{xy}/s_x^2, \quad \hat{a} = \bar{y} - \hat{b}\bar{x}$$

der s_x og s_y er standardavvikene til henholdsvis x - og y -verdiene, mens s_{xy} er definert ved

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Minste kvadratsum er gitt ved

$$\text{rest} = \sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i)^2 = (n-1)(s_y^2 - \hat{b}s_{xy})$$

Standardavvik som måler variasjonen i punktene rundt den beste linjen:

$$s_{\text{reg}} = \sqrt{\frac{\text{rest}}{n-2}}$$

Konfidensintervall for $\hat{b}$ bestemmes ut fra formelen:

$$\hat{b} \pm c \frac{s_{\text{reg}}}{\sqrt{(n-1)s_x^2}}$$

der c bestemmes ut fra en studentfordeling med $n-2$ frihetsgrader.

Korrelasjon

Korrelasjonskoeffisienten er definert på følgende måte:

$$r = \frac{s_{xy}}{s_x s_y}$$

Utvalgsstørrelse

Parallellgruppestudie – målevariabler:

$$n = 2 \left(\frac{\sigma}{\Delta} \right)^2 \cdot k$$

Overkrysningsstudie – målevariabler:

$$n = \left(\frac{\sigma_d}{\Delta} \right)^2 \cdot k$$

Utvalgsstørrelse – binomisk responsvariabel:

$$n = \frac{p_1(1-p_1) + p_2(1-p_2)}{(p_2 - p_1)^2} \cdot k$$

k bestemmes av tabellen:

		Teststyrke		
		0.80	0.90	0.95
Siginifikans-	0.10	6.2	8.6	10.8
nivå	0.05	7.9	10.5	13.0
(tosidig)	0.01	11.7	14.9	17.8

Utvalgsstørrelse basert på presisjon i estimer

Binomisk respons:

$$n = \left(\frac{1.96}{d} \right)^2 p(1-p)$$

Kontinuerlig respons:

$$n = \left(\frac{1.96 \cdot \sigma}{d} \right)^2$$

Infeksjonsmodellering

Kritisk vaksinasjonsdekning for å unngå en epidemi:

$$p_{\text{crit}} = 1 - 1/R_0$$