Formelsamling i medisinsk statistikk

Dette er formelsamling til O. O. Aalen: Innføring i statistikk med medisinske eksempler, 2. utg., Ad Notam Gyldendal, 1998. Formelsamlingen er utarbeidet i okt. 2000, med små endringer i okt. 2001.
På s. 7 er det også gitt noen *rettelser* til boken.

Gjennomsnitt

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

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.

Standardavvik

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

Grupperte data

Intervallmidtpunkter $m_1, m_2, \ldots, m_k$. Hyppigheter $f_1, f_2, \ldots, f_k$. Totalt antall observasjoner: n. Gjennomsnitt og standardavvik er gitt ved:

$$\overline{x} = \frac{1}{n}(m_1 f_1 + m_2 f_2 + \cdots + m_k f_k) = \frac{1}{n}\sum_{j=1}^{k} m_j f_j$$

$$s = \sqrt{\frac{1}{n-1}\{\sum_{j=1}^{k}(m_j - \overline{x})^2 f_j\}}$$

Median og *fraktiler* for grupperte data finnes ved lineær interpolasjon.

Insidens og prevalens

Prevalens angir andelen i befolkningen som har en viss sykdom.
Insidensraten måler forekomsten av *nye* tilfeller av sykdommen over et gitt tidsrom.

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*

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

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*

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

Bayes' lov

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

Diagnostiske tester

Sensitivitet: Sannsynlighet for positiv test gitt sykdom.
Spesifisitet: Sannsynlighet for negativ test gitt at det ikke foreligger sykdom.
Positiv prediktiv verdi: Sannsynlighet for at det foreligger sykdom gitt postiv test.
Negativ prediktiv verdi: Sannsynlighet for at det ikke foreligger sykdom gitt negativ test.

Kombinatorikk

Trekning av s kuler fra en boks med n kuler.
Antall *ordnede* utvalg *med tilbakelegging*

$$n^s$$

Antall *ordnede* utvalg *uten tilbakelegging*

$$n(n-1)(n-2)\cdots(n-s+1)$$

Antall *ikke-ordnede* utvalg *uten tilbakelegging*

$$\binom{n}{s} = \frac{n(n-1)(n-2)\cdots(n-s+1)}{s!}$$

Forventning og varians for teoretisk fordeling

$$\mathrm{E}(X) = \sum_{\text{alle } x_i} x_i \, P(X = x_i)$$

$$\mathrm{Var}(X) = \sum_{\text{alle } x_i} (x_i - \mathrm{E}(X))^2 P(X = x_i)$$

Regneregler for forventning og varians

$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b,$$

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X), \qquad \mathrm{SD}(aX + b) = |a| \mathrm{SD}(X)$$

$$\mathrm{E}(X_1 + X_2 + \cdots + X_n) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n)$$

Hvis $X_1, X_2, \ldots, X_n$ er parvis *stokastisk uavhengige* has:

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)$$

Binomisk fordeling

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

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

Forventning og varians i binomisk fordeling er gitt ved:

$$\mathrm{E}(X) = np, \qquad \mathrm{Var}(X) = np(1 - p)$$

Poissonfordeling

Sannsynligheten for x forekomster, når forventning er lik λ, er gitt ved:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad \text{for} \quad x = 0, 1, 2, \ldots.$$

Forventning og varians er gitt ved:

$$\mathrm{E}(X) = \lambda \qquad \text{og} \qquad \mathrm{Var}(X) = \lambda$$

Poissonfordelingen anvendes også ved Poissonprosesser.

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). Sannsynlighetstettheten til normalfordelingen er gitt ved følgende formel:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

der $\exp(a)$ er det samme som eksponensialfunksjonen e^a.

Formler for gjennomsnitt

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

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

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

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

og denne er Studentfordelt med $n-1$ frihetsgrader når H_0 gjelder.

Sammenlikning av to gjennomsnitt

Samme forutsetninger som over. Forøvrig antas gjennomsnittene å komme fra to uavhengige utvalg. Følgende testørrelse er Studentfordelt med $n_1 + n_2 - 2$ frihetsgrader når H_0 gjelder

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

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

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

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

Poissonfordeling som tilnærming til binomisk fordeling

Binomisk fordeling kan tilnærmes med en Poissonfordeling hvis:

$$(1) \quad p \leq 0.05 \qquad \text{og} \qquad (2) \quad n \geq 50$$

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

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

Normalfordeling som tilnærming til Poissonfordeling

Når λ i en Poissonfordeling er minst lik 5, vil Poissonfordelingen likne mye på en normalfordeling med parametre

$$\mu = \lambda, \qquad \sigma = \sqrt{\lambda}$$

Estimering av sannsynlighet

Hvis det er observert X forekomster ved n binomiske forsøk, er estimatet for sannsynligheten p gitt ved p^*, mens estimert standardfeil er gitt ved s_p

$$p^* = X/n, \qquad s_p = \sqrt{\frac{p^*(1-p^*)}{n}}$$

Fordelingen til 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 p^* er gitt ved

$$p^* \pm 2s_p$$

Teststørrelse for sammenlikning av to sannsynligheter

$$Y = \frac{p_1^* - p_2^*}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\overline{p}(1-\overline{p})}}$$

Teststørrelse for sammenlikning av to Poissonvariabler

$$Y = \frac{X_1 - X_2}{\sqrt{X_1 + X_2}}$$

Konfidensintervall for relativ risiko

Relativ risiko:

$$RR = \frac{X_1/n_1}{X_2/n_2}$$

Hjelpestørrelse:

$$s_{RR} = \sqrt{\frac{1}{X_1} + \frac{1}{X_2} - \frac{1}{n_1} - \frac{1}{n_2}}$$

95% konfidensintervall for RR:

$$(RR \times \exp(-1.96\, s_{RR}),\ RR \times \exp(1.96\, s_{RR}))$$

Regresjonsanalyse

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

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

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

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

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:

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

Bestemmelse av forsøksstørrelse

Størrelse på et binomisk forsøk for å oppnå en sikkerhet på a i estimatet for sannsynligheten:

$$n = \frac{4p(1-p)}{a^2}$$

Størrelse på et forsøk med måledata for å oppnå en sikkerhet på a i gjennomsnittet:

$$n = 4(\sigma/a)^2$$

Antallet som kreves i hver gruppe når to binomiske sannsynligheter skal sammenliknes med siginifikansnivå α og teststyrke $1-\beta$:

$$n = \frac{p_1(1-p_1)+p_2(1-p_2)}{(p_2-p_1)^2} \times f(\alpha,\beta)$$

Antallet som kreves i hver gruppe når to gjennomsnitt skal sammenliknes med siginifikansnivå α og teststyrke $1-\beta$:

$$n = 2(\sigma/\Delta)^2 \times f(\alpha,\beta)$$

Størrelsen $f(\alpha,\beta)$ er definert i tabellen nedenfor.

		β (type II feil)			
		0.05	0.10	0.20	0.50
	0.10	10.8	8.6	6.2	2.7
α (type I feil)	0.05	13.0	10.5	7.9	3.8
	0.02	15.8	13.0	10.0	5.4
	0.01	17.8	14.9	11.7	6.6

Tallene i tabellen er hentet fra Pocock (1983).

Table 1: *Tabell over funksjonen* $f(\alpha,\beta)$

Rettelser til boken

Nedenfor er det gitt noen rettelser til O. O. Aalen: Innføring i statistikk med medisinske eksempler, 2. utg., Ad Notam Gyldendal, 1998.

- side 114, linje 6: Følgende passus fjernes: "(dvs. en hvert fjerde år)"
- side 147, linje 4: Ordet "pleierne" erstattes med "graviditetene"
- side 147, siste linje: Tallet 6.09 erstattes med 4.14
- side 283, siste linje: Det skal stå 0.06%
- side 285, tabell: Tallet 167 skal erstattes med 197
- side 302, fasit til B76: Tallet -1.46 skal erstattes med -1.47