MASB11 – Biostatistisk grundkurs
Formelsamling

Beskrivande statistik

Stickprov

$$\bar{x} = \frac{\sum x_i}{n} \qquad s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} \qquad \text{Variationskoefficient} = \frac{s}{\bar{x}}$$

Population

$$\mu = E(X) = \sum x \cdot f(x) \qquad \sigma^2 = V(X) = Var(X) = \sum (x-\mu)^2 f(x) = \sum x^2 f(x) - \mu^2$$

Binomialfördelning *Bin(n,p)*

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad , \quad x = 0, \ldots, n \qquad E(X) = np \qquad Var(X) = np(1-p)$$

Approximativt gäller:

Bin(*n*,*p*) ≈ *N(np, np(1-p))* om *np(1-p)* ≥ 10

Bin(n,p) ≈ *Po(np)* om $p \leq 0{,}1$ och $n \geq 10$

Poissonfördelning *Po(λ)*

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad , \quad x = 0,1,\ldots \qquad E(X) = \lambda \qquad Var(X) = \lambda$$

Approximativt gäller: *Po(λ)* ≈ *N(λ,λ)* om $\lambda \geq 15$

Geometrisk fördelning Geo*(p)*

$$f(x) = p(1-p)^{x-1} \quad , \quad x = 0,1,\ldots \qquad E(X) = \frac{1}{p} \qquad Var(X) = \frac{1-p}{p^2}$$

MASB11 – Biostatistisk grundkurs
Formelsamling

Hypergeometrisk fördelning Hyp(*n,p,N)*

$$f(x)=\frac{\binom{M}{x}\cdot\binom{N-M}{n-x}}{\binom{N}{n}} \quad , \quad x=0,1,\ldots,n \qquad E(X)=np \qquad Var(X)=np(1-p)\left(\frac{N-n}{N-1}\right)$$

Normalfördelning *N(μ,σ²)*

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad E(X)=\mu \qquad Var(X)=\sigma^2$$

Antag att $X \in N(\mu,\sigma^2)$. Då gäller att

$$Z=\frac{X-\mu}{\sigma} \quad \in N(0,1)$$

Om $X_i \in N(\mu,\sigma^2)$ i=1, … , n oberoende, så blir $\overline{X} \in N\left(\mu,\frac{\sigma^2}{n}\right)$

Om $X_i \in N(\mu,\sigma^2)$ i=1, … , n oberoende, så blir $\sum X_i \in N\left(n\mu,n\sigma^2\right)$

Bivariat sannolikhetsfördelning

$$\sigma_{XY}=Cov(X,Y)=\sum x_i y_j p_{ij}-E(X)\cdot E(Y)$$

$$\rho_{XY}=\frac{\sigma_{XY}}{\sigma_X\cdot\sigma_Y}=\frac{Cov(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}}$$

Linjära kombinationer av slumpvariabler

$$Z=a+bX+cY$$
$$E(Z)=a+bE(X)+cE(Y)$$
$$V(Z)=b^2V(X)+c^2V(Y)+2bc\cdot Cov(X,Y)$$

MASB11 – Biostatistisk grundkurs
Formelsamling

Ett stickprov (även matchade stickprov efter bildandet av differens)

	Konfidensintervall för μ	Testfunktion
Population nf σ är känd	$\bar{x} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}$	$z = \frac{\bar{x}-\mu}{\sqrt{\sigma^2/n}}$
Population nf σ är okänd, *n* litet	$\bar{x} \pm t_{1-\alpha/2;\,n-1}\sqrt{\frac{s^2}{n}}$	$t = \frac{\bar{x}-\mu}{\sqrt{s^2/n}}$
Population ej nf σ är okänd, *n* stort	$\bar{x} \pm z_{1-\alpha/2}\sqrt{\frac{s^2}{n}}$	$z = \frac{\bar{x}-\mu}{\sqrt{s^2/n}}$

Tvåsidigt konfidensintervall för σ^2

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right)$$

Testfunktion

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Två stickprov

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

f.g = $(n_1-1)+(n_2-1) = n_1+n_2-2$

	Konfidensintervall för $\mu_1-\mu_2$	Testfunktion
Båda pop. nf σ_1 och σ_2 kända	$\bar{x}_1 - \bar{x}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$z = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$
Båda pop. nf σ_1 och σ_2 okända n_1 och n_2 små	$\bar{x}_1 - \bar{x}_2 \pm t_{1-\alpha/2;\,n_1+n_2-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$	$t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
Pop. ej nf n_1 och n_2 stora	$\bar{x}_1 - \bar{x}_2 \pm z_{1-\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	$z = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$

MASB11 – Biostatistisk grundkurs
Formelsamling

Centrala gränsvärdessatsen

Antag att X_i, $i=1, \ldots, n$ är oberoende och har samma fördelning, $E(X_i)=\mu$ och $Var(X_i)=\sigma^2$
Då gäller för stora värden på n att

$\sum_1^n X_i$ är appr. $N(n\mu, n\sigma^2)$ och att $\overline{X}$ är appr $N(\mu, \sigma^2/n)$

Procenttal (binomialfördelning)

Om $np(1-p) \geq 10$

Konfidensintervall för p	Testfunktion
$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$

Om $n_i p_i(1-p_i) \geq 10$

Konfidensintervall för p_1-p_2	Testfunktion
$\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$z = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

där $\hat{p}_0 = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2} = \frac{x_1+x_2}{n_1+n_2}$

χ^2-test

För en variabel med r kategorier (goodness-of-fit)

$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i}$ där O_i = observerat antal och E_i = förväntat antal (f.g. = $r-1$)

För en tabell med r rader och c kolumner (homogenitetstest)

$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ f.g. = $(r-1)(c-1)$

MASB11 – Biostatistisk grundkurs
Formelsamling

Ensidig variansanalys

Modell: $y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ där
$\varepsilon_{ij} \in N(0, \sigma^2)$ $\quad i = 1, 2, \ldots, a \quad , \quad j = 1, 2, \ldots, n_i \quad , \quad N = \sum n_i$

$$SS_T = SS_A + SS_e$$

$$SS_T = \sum_i \sum_j \left(y_{ij} - \bar{y}_{..}\right)^2 = \sum\sum y_{ij}^2 - \frac{T_{..}^2}{N}$$

$$SS_A = \sum_i n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 = \sum \frac{T_{i.}^2}{n_i} - \frac{T_{..}^2}{N}$$

$$SS_e = \sum_i \sum_j \left(y_{ij} - \bar{y}_{i.}\right)^2 = SS_T - SS_A = \sum_i \left(n_i - 1\right) s_i^2$$

Variansanalystabell

Källa	Kvadratsumma	f.g.	Medelkvadratsumma
Behandling	SS_A	$a-1$	$MS_A = SS_A/(a-1)$
Residual	SS_e	$N-a$	$MS_e = SS_e/(N-a)$
Total	SS_T	$N-1$	

Testkvantitet: $F = MS_A/MS_e$ f.g. = $(a-1)/(N-a)$

Ensidig variansanalys med block

Modell: $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ där
$\varepsilon_{ij} \in N(0, \sigma^2)$ $\quad i = 1, 2, \ldots, a \quad , \quad j = 1, 2, \ldots, b \quad , \quad N = a \cdot b$

$$SS_T = SS_A + SS_{Block} + SS_e$$

$$SS_T = \sum\sum y_{ij}^2 - \frac{T_{..}^2}{N}$$

$$SS_A = \sum \frac{T_{i.}^2}{b} - \frac{T_{..}^2}{N}$$

$$SS_{Block} = \sum \frac{T_{.j}^2}{a} - \frac{T_{..}^2}{N}$$

$$SS_e = SS_T - SS_A - SS_{Block}$$

Testkvantitet: $F = MS_A/MS_e$ f.g. = $(a-1)/(a-1)(b-1)$

MASB11 – Biostatistisk grundkurs
Formelsamling

Enkel linjär regression och korrelation

Beräkningssummor:

$$SS_x = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \qquad SS_y = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

$$SP_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

Modell: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ där $\varepsilon_i \in N(0, \sigma^2), i = 1, 2, \ldots, n$

Skattningar:

$$b_1 = SP_{xy} / SS_x \qquad \hat{Var}(b_1) = s_e^2 / SS_x$$

$$b_0 = \bar{y} - b_1 \bar{x} \qquad \hat{Var}(b_0) = s_e^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x} \right)$$

där

$$s_e^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{SS_y - \frac{SP_{xy}^2}{SS_x}}{n-2}$$

Konfidensintervall:

$$b_1 \pm t_{1-\alpha/2;\, n-2} \sqrt{\hat{Var}(b_1)} \qquad b_0 \pm t_{1-\alpha/2;\, n-2} \sqrt{\hat{Var}(b_0)}$$

Prognoser:

Förväntat värde för y givet $x=x_0$

$$b_0 + b_1 x_0 \pm t_{1-\alpha/2;\, n-2} \sqrt{s_e^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x} \right)}$$

För enskilt värde på y givet $x=x_0$

$$b_0 + b_1 x_0 \pm t_{1-\alpha/2;\, n-2} \sqrt{s_e^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x} \right)}$$

Korrelationskoefficient: $r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$ **Test:** $t = r \sqrt{\frac{(n-2)}{(1-r^2)}}$

Icke-parametriska metoder

Två oberoende stickprov (Wilcoxon-Mann-Whitney)

Rangordna det sammanslagna stickprovet. Ersätt värden med rangtal. Beräkna rangsummorna R_1 och R_2

Testfunktion: Det minsta av

$$U_1 = n_1 n_2 + \left[n_1(n_1+1)\right]/2 - R_1$$
$$U_2 = n_1 n_2 + \left[n_2(n_2+1)\right]/2 - R_2$$

Tabell 11

Om $n_i \leq 15$ så använd tabell för kritiska värden. För övriga stickprovsstorlekar använd att:
R är appr. $N\left(\frac{1}{2}n_1(n_1+n_2+1), \frac{1}{12}n_1 n_2(n_1+n_2+1)\right)$ där R är rangsumman i ett av stickproven

Teckentest (parvisa jämförelser)

(x_i,y_i), $i=1, 2, \ldots, n$ observationer i par
n_+= antal par där $y_i > x_i$ => under H_0 är n_+ *Bin(n, 0,5)*

Fler än två oberoende stickprov (Kruskal-Wallis)

a stycken ($a \geq 3$) oberoende stickprov med n_i observationer i det *i:te* stickprovet. Rangordna det sammanslagna stickprovet. Beräkna rangsummorna R_i i samtliga stickprov.

Testfunktion: $K = \frac{12}{N(N+1)}\sum_{i=1}^{a}\frac{R_{i\cdot}^2}{n_i} - 3(N+1)$ där $N = \sum_{i=1}^{a} n_i$

Under H_0: $K \in \chi^2(a-1)$

Spearmans rangkorrelation

Rangordna x_i:na och y_i:na var för sig. Ersätt värden med rangtal. Bilda d_i=skillnad i rangtal för observation i.

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$