Formelsamling til statistik-del af metodekursus, 4. semester, lægevidenskab

Version 3 (26/9-2011)

Kære læser

Denne formelsamling er lavet med udgangspunkt i "Medical Statistics, second edition" af Betty R. Kirkwood og A. C. Sterne og skulle meget gerne dække de formler som man kan få brug for under eksamen samt store dele af kurset i statistik. Den kan dog på ingen måde erstatte læsning af lærebogen og bør ses som et supplement og ikke et alternativ.
Selve formelsamlingen er delt op i kapitler, svarende til dem angivet i læseplanen efteråret 2011 og sideangivelser i form af fodnoter vil lede dig hen i lærebogen, hvor du kan læse uddybende om formelen.
Hvis du finder fejl, mangler, uklarheder eller synes at en formel bør uddybes, skal du være mere end velkommen til at skrive til mig på vlw432@alumni.ku.dk

Asger Mølgaard Andreasen
Hold 408, Efteråret 2011

Kapitel 3: Displaying the data

- Median[i]: $\frac{(n+1)}{2}$ = observationsnummer som er medianen (med *n* observationer)
- Nedre kvartil[ii]: $\frac{(n+1)}{4}$ = observationsnummer som er nedre kvartil
- Øvre kvartil[iii]: $\frac{3\cdot(n+1)}{4}$ = observationsnummer som er nedre kvartil

Hvis medianen/nedre/øvre kvartil giver et tal som ligger mellem to observationer (f.eks. observation nr. 8,5) tages gennemsnittes mellem de to omkringliggende observationer (dvs. snittet mellem observation 8 og 9 i eksemplet).

- Range[iv] = største værdi – laveste værdi
- Interquartile range[v] = øvre kvartil – nedre kvartil
- [vi] $k'ende\ \ percentil = \frac{k\cdot(n+1)'nde}{100}$ observations værdi (eksempel: hvis *k*=50 (vi ønsker at finde medianen) og *n*=70, fås 35,5. Dvs. medianen er mellem obs. 35 og 36.

Kap. 4: Means, standard deviations and standard errors

- Middelværdi[vii]:
  $\bar{x} = \frac{\sum x}{n}$
- Varians/spredning[viii]:
  $s^2 = \frac{\sum\left(x-\bar{x}\right)^2}{(n-1)}$
- Standard deviation/spredning[ix]:
  $s.d. = s = \sqrt{s^2} = \sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{(n-1)}} = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x^2\right)}{n}}{(n-1)}}$
- Stand error[x]:
  $s.e. = \frac{s}{\sqrt{n}}$

Kap. 5: The Normal Distribution

- Ligning for standard normal fordeling[xi] med middelværdien μ og s.d.'en σ:

---

[i] S. 22
[ii] S. 23
[iii] S. 23
[iv] S.24
[v] S. 24
[vi] S. 25
[vii] S. 33
[viii] S. 35
[ix] S. 36
[x] S. 39
[xi] S. 43

$$y = \frac{\exp\left(\frac{-z^2}{2}\right)}{\sqrt{2\pi}}$$

Hvor *z* kaldes Standard Normal Deviation og givet ved:

$$z = \frac{x - \mu}{\sigma}$$

- Ønsker man at finde hvor stor en proportion som ligger i normalfordelingens øvre ende, indsættes den *x*-værdi som ønskes undersøgt for, i formelen for *z* og *z*-værdien aflæses i Tabel A1 (bag i bogen). Eks.: Man ønsker at vide hvor mange mænd i en sample med $\mu = 171{,}5cm$, som er større end 180cm. Man indsætter 180 på *x* plads og får *z*=1,31. Denne værdi aflæses i A1 til værende 0,0951. Dvs. 9,51% af mændene var større end 180cm.
- Fremgangsmåde for at finde proportion i nedre hale af normalfordeling er analog til proportion i øvre ende (blot skal x< μ).
- For at finde proportion mellem to punkter, findes de procent som er under nedre punkt og over øvre punkt. Disse trækkes fra 1 og den samlede proportion fås. Eks.: Proportion af mænd mellem 165cm og 175cm = 1 – proportion under 165 – proportion over 175cm= 1-0,1587-0,2946=0,5467 eller 54,67%
- Reference interval[xii]:
  $\bar{x} - z' \cdot s.d. \quad til \quad \bar{x} + z' \cdot s.d.$
  hvor *z'* aflæses i Tabel A2. For et 95% CI fås z'=1,96. Formelen kan også omformuleres til: $\left[\bar{x} \pm z' \cdot s.d.\right]$. Den udtrykker, hvor vi forventer at 95% af vores værdier vil ligge.

Kap. 6: Confidence Interval for a mean

- Konfidens interval (CI) for store samples (n>25)[xiii]:
  $\bar{x} - z' \cdot s.e. \quad til \quad \bar{x} + z' \cdot s.e.$
- Konfidens interval (CI) for små samples (n<25)[xiv]:
  $\bar{x} - t' \cdot s.e. \quad til \quad \bar{x} + t' \cdot s.e.$
  Hvor *t'* slås op i tabel A3 med *(n-1)* frihedsgrader
- CI 95% giver os et interval, hvor den sande middelværdi med 95% sikkerhed ligger.

Kap. 7: Comparison of two means: confidence intervals, hypothesis test and *P*-values

- Standard error for $\left(\bar{x}_1 - \bar{x}_0\right)$[xv]:

$$s.e._{\left(\bar{x}_1 - \bar{x}_0\right)} = \sqrt{\left(s.e._1\right)^2 + \left(s.e._0\right)^2} = \sqrt{\frac{\left(\sigma_1\right)^2}{n_1} + \frac{\left(\sigma_0\right)^2}{n_0}} = \sqrt{\frac{\left(s_1\right)^2}{n_1} + \frac{\left(s_0\right)^2}{n_0}}$$

  Bemærk at σ er den sande spredning i populationen, mens *s* er vores estimat på spredningen beregnet sample. Ofte kender vi dog ikke σ og må nøjes med *s*.
- Konfidens interval på differens mellem gennemsnit:
  $$CI = \left(\bar{x}_1 - \bar{x}_0\right) - \left(z' \cdot s.e._{\left(\bar{x}_1 - \bar{x}_0\right)}\right) \quad til \quad \left(\bar{x}_1 - \bar{x}_0\right) + \left(z' \cdot s.e._{\left(\bar{x}_1 - \bar{x}_0\right)}\right)$$
- *z*-test/approksimativ t-test (kun for large samples)[xvi]:

[xii] S. 49
[xiii] S. 51
[xiv] S. 55
[xv] S. 60
[xvi] S: 62

$$z = \frac{\left(\bar{x}_1 - \bar{x}_0\right)}{s.e._{\left(\bar{x}_1 - \bar{x}_0\right)}}$$

Herefter opslag i Tabel A1 og ved signifikant *P*-værdi (husk at bruge den to-sidede!) afvises nul-hyposen (H_0: at der ingen statistisk differens er på gennemsnittene).

- *t*-test/unpaired *t*-test bruges på små samples[xvii]:

$$s = \sqrt{\frac{\left(n_1 - 1\right)\cdot\left(s_1\right)^2 + \left(n_0 - 1\right)\cdot\left(s_0\right)^2}{\left(n_1 + n_0 - 2\right)}}$$

$$t = \frac{\left(\bar{x}_1 - \bar{x}_0\right)}{s.e.} = \frac{\left(\bar{x}_1 - \bar{x}_0\right)}{s\cdot\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}}$$

Med antal frihedsgrader (degrees of freedom):

$d.f. = n_1 + n_0 - 2$

Opslag i Tabel A4 giver *P*-værdien

- For parrede observationer undersøges differensen. Dvs. vi omdanner vores data-par til en enkelt sample bestående af differenser.

CI for store samples[xviii]:

$\bar{x} - z'\cdot s.e.$ *til* $\bar{x} + z'\cdot s.e.$

CI for små samples:

$\bar{x} - t'\cdot s.e.$ *til* $\bar{x} + t'\cdot s.e.$ (husk *n-1* frihedsgrader)

- Hypotese test for parrede-data udføres med enten en *parret z-test* eller en *parret t-test*:

Parret *z*-test (store samples)[xix]:

$$z = \frac{\bar{x}}{s.e.} = \frac{\bar{x}}{\frac{s}{\sqrt{n}}}$$

Parret *t*-test (små samples):

$$t = \frac{\bar{x}}{s.e.} = \frac{\bar{x}}{\frac{s}{\sqrt{n}}},\ d.f. = n - 1$$

$\bar{x}$ er gennemsnittet af de parrede differenser og husk opslag i tabel A1 eller A4

Kap. 10: Lineær Regression and correlation (kun afsnittene 10.1 og 10.2)

- Lineær regressions forskrift[xx]:

$y = \beta_0 + \beta_1 \cdot x$

- Estimat på parametre[xxi]:

$$\beta_1 = \frac{\sum\left(\left(x - \bar{x}\right)\cdot\left(y - \bar{y}\right)\right)}{\sum\left(\left(x - \bar{x}\right)^2\right)} \quad og \quad \beta_0 = \bar{y} - \beta_1 \cdot \bar{x}$$

- Præcisionen måles med deres standard error[xxii]:

[xvii] S. 66
[xviii] S. 68
[xix] S. 69
[xx] S. 88
[xxi] S. 90

$$s.e.(\beta_0) = s \cdot \sqrt{\frac{1}{n} \cdot \frac{(\bar{x})^2}{\sum\left((x-\bar{x})^2\right)}} \qquad og \qquad s.e.(\beta_1) = \frac{s}{\sqrt{\sum\left((x-\bar{x})^2\right)}}$$

$$s = \sqrt{\frac{\sum\left((x-\bar{x})^2\right) - (\beta_1)^2 \cdot \sum\left((x-\bar{x})^2\right)}{(n-2)}}$$

- Konfidensinterval for regressionskoefficient (benævnt $\beta_{0/1}$ da formelen gælder begge)[xxiii]:
  $CI = \beta_{0/1} - t' \cdot s.e.(\beta_{0/1}) \quad til \quad \beta_{0/1} + t' \cdot s.e.(\beta_{0/1})$
- Antagelser for lineærregression: For enhver værdi af x, er y normalfordelt. Punkterne er fordelt med samme spredning omkring linien over hele plottet. Hvorfor der indsættes en spredningskoeffcient (ε) som er forskellig for hvert punkt, så forskriften bliver:
  $y = \beta_0 + \beta_1 \cdot x + \varepsilon$

Kap. 11: Multiple regression

- En variabel som kun antager værdierne 0 eller 1 kaldes en *indicator variabel* (f.eks. køn).
- Interceptet (β_0) er værdien af *y* hvis alle eksponeringsvariable er 0.
- General formel for *multipel regression* med *n* eksponeringer[xxiv]:
  $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + ... + \beta_n \cdot x_n + \varepsilon$

Kap. 14: Probability, risk and odds (of disease)

- *Prob* er forkortelse for *Probability*, ligesom *ssh* er for *sandsynlighed*.
- *Ssh* for at både A og B sker[xxv]:
  $prob(A \; and \; B) = prob(A) \cdot prob(B)$
- *Ssh* for at enten A eller B sker:
  $prob(A \; or \; B \; or \; both) = prob(A) + prob(B) - prob(both)$
- Odds defineres ved *ssh* for at A sker, divideret med *ssh* for at A ikke sker:
  $$Odds = \frac{p}{1-p} = \frac{d}{h} = \frac{diseased(d)}{healthy(h)}$$
  $$p = \frac{odds}{1+odds}$$

Kap. 15: Proportions and the binomial distribution

- Proportion/risiko/*ssh* for sygdom:
  $$p = \frac{antal_{syge}}{total} = \frac{d}{n}$$
- For en sample er risk=*p*. Den sande (og ofte ukendte) risiko i hele populationen benævnes π.
- Den generelle formel for *ssh* for at få *d* events i en sample med *n* individer[xxvi]:
  $$prob(d \; events) = \frac{n!}{d! \cdot (n-d)!} \cdot \pi^d \cdot (1-\pi)^{n-d}$$

[xxii] S. 91
[xxiii] S. 91
[xxiv] S. 105
[xxv] S. 134
[xxvi] S. 141

- Proportioner er også binomial fordelt, hvorfor konfidensinterval og standard error også kan beregnes for en proportion[xxvii]:

  $CI = p - (z' \cdot s.e.) \quad til \quad p + (z' \cdot s.e.)$

  $s.e. = \sqrt{\frac{p \cdot (1-p)}{n}}$

- Med en *z-test* kan vi teste om vores estimat *p* på proportionen er magen til en bestemt værdi (kaldet π. Gør at vi kan teste om *p* er lig med den sande π (eller den værdi vi tror, er sand)).

  $z = \frac{p - \pi}{s.e.(p)} = \frac{p - \pi}{\sqrt{\frac{\pi \cdot (1-\pi)}{n}}}$

  Efter opslag i tabel A1, får vi en *P*-værdi. En signifikant *P*-værdi betyder at vi har evidens for at $p \neq \pi$, mens at en usignifikant betyder at $p = \pi$

- $risiko = (kumulativ)\ incidens = \frac{nye\ \ sygdomstilfælde\ \ i\ \ en\ \ given\ \ periode}{antal\ \ raske\ \ ved\ \ periodens\ \ start}$

- $prævalens = \frac{antal\ \ sygdomstilfælde\ \ til\ \ specifikt\ \ tidspunkt}{total_{population}}$

Kap. 16: Comparing two proportions

- For at sammenligne binære outcome variable mellem to eksponeringsgrupper, kan det altid svare sig at bruge 2x2-tabeller (*d*=disease, *h*=healthy):

	Outcome		
Exposure	Experienced event: D (disease)	Did not experience event: H (healthy)	Total
Group 1 (exposed)	d_1	h_1	n_1
Group 0 (unexposed)	d_0	h_0	n_0
Total	d	h	n

- Vi husker at $p = \frac{d}{n}$ samt at $odds = \frac{d}{h}$
- Differens mellem to proportioner[xxviii]:

  $p_{diff} = p_1 - p_0$

  $s.e.(p_1 - p_0) = \sqrt{\frac{p_1 \cdot (1-p_1)}{n_1} + \frac{p_0 \cdot (1-p_0)}{n_0}} = \sqrt{s.e.(p_1)^2 + s.e.(p_0)^2}$

  $CI = (p_1 - p_0) - z' \cdot s.e.(p_1 - p_0) \quad til \quad (p_1 - p_0) + z' \cdot s.e.(p_1 - p_0)$

- Test for at to proportioner ens (nulhypose (H_0): $\pi_1 = \pi_0 = \pi$ [xxix]):

  $z = \frac{p_1 - p_0}{\sqrt{p \cdot (1-p) \cdot \left(\frac{1}{n_1} + \frac{1}{n_0}\right)}}$

  Hvor $p = \frac{d_0 + d_1}{n_0 + n_1} = \frac{d}{n}$

- Risk-ratio (RR), proportion mellem to risikoer[xxx]:

[xxvii] S. 143
[xxviii] S. 151
[xxix] S. 153
[xxx] S. 153

$$RR = \frac{p_1}{p_0} = \frac{d_1 / n_1}{d_0 / n_0}$$

- Konfidensinterval for RR[xxxi]:

$$CI\left(RR\right) = \frac{RR}{EF} \;\; til \;\; RR \cdot EF$$

$$EF = error \;\; factor = \exp\left(z' \cdot s.e.\left(\log\left(RR\right)\right)\right)$$

$$s.e.\left(\log\left(RR\right)\right) = \sqrt{\frac{1}{d_1} - \frac{1}{n_1} + \frac{1}{d_0} - \frac{1}{n_0}}$$

- Test af nulhyposen om at der ingen forskel i risiko er i de to grupper[xxxii]. Dvs. at RR=1 og log(RR)=0:

$$z = \frac{\log\left(RR\right)}{s.e.\left(\log\left(RR\right)\right)}$$

- Odds ratio (OR): ratioen af odds mellem den eksponerede og den ueksponerede gruppe. (næste side)[xxxiii]

$$odds = \frac{p}{1-p} = \frac{d / n}{1-\left(d / n\right)} = \frac{d / n}{h / n} = \frac{d}{h}$$

$$Oddsratio\left(OR\right) = \frac{odds_{gruppe1}}{odds_{gruppe0}} = \frac{d_1 / h_1}{d_0 / h_0} = \frac{d_1 \cdot h_0}{d_0 \cdot h_1}$$

$$OR\left(disease\right) = \frac{1}{OR\left(healthy\right)}$$

- For udredning af rationalet for, at bruge OR, se s. 163.
- Konfidens interval for odds:

$$CI = \frac{odds}{EF} \;\; til \;\; odds \cdot EF$$

$$EF = \exp\left(z' \cdot s.e.\left(\log\left(odds\right)\right)\right)$$

$$s.e.\left(\log\left(odds\right)\right) = \sqrt{\frac{1}{d} + \frac{1}{h}}$$

- Konfidens interval for odds ratio (OR):

$$CI = \frac{OR}{EF} \;\; til \;\; OR \cdot EF$$

$$EF = \exp\left(z' \cdot s.e.\left(\log\left(OR\right)\right)\right)$$

$$s.e.\left(\log\left(OR\right)\right) = \sqrt{\frac{1}{d_1} + \frac{1}{h_1} + \frac{1}{d_0} + \frac{1}{h_0}}$$

- Test af nulhyposen (*H_0: OR=1*):

$$z = \frac{\log\left(OR\right)}{s.e.\left(\log\left(OR\right)\right)}$$

[xxxi] S. 156
[xxxii] S. 158
[xxxiii] S. 159

Kap. 17: Chi-squared tests for 2x2 and larger contingency tables

- Chi-i-anden-test bruges til at undersøge om der er en association mellem række-variablen og søjle-variablen. Eller sagt på en anden måde: om fordelingen af individer blandt kategorierne af en variabel, er uafhængig af deres fordeling blandt kategorierne af en anden variabel.
- Bemærkninger om chi-i-anden-tests validitet: s. 168.
- Chi-i-anden-tests kan kun bruges på reelle tal og ikke på proportioner.
- χ^2-testens bestanddele:
  O = observed = antal observationer/individer i en celle
  E = expected = ved beregning finder man et "forventet" antal af individer i cellen.
  r = antallet af rækker (rows)
  c = antallet af søjler (colums)

  $$E = \frac{total_{søjle} \cdot total_{række}}{total_{alle}}$$

  $$\chi^2 = \sum \frac{(O-E)^2}{E}$$

  Med antal frihedsgrader: $d.f. = (r-1)\cdot(c-1)$

  χ^2 slås op i tabel A5 og *P*-værdi findes. Hvis *P*-værdien er lille (signifikant: under 0,05), da er der sammenhæng mellem række- og søjlevariablene. Dvs. data er ikke fordelt efter nul-hypotesen om at data er fordelt uafhængigt af søjle- og række-variable.

Kap. 19: Logistic regression: comparing two or more exposure groups

- I denne model multipliceres parametrene (til forskel fra den multiple regression som adderer dem (kap. 11)). Hvis der er to eksponeringer, fås følgende:
  Odds = baseline x exposure(A) x exposure(B)
  (så hvis A fordobler odds og B øger odds x3, fås at odds er 6xbaseline)
- Logistisk regression laves altid på logistisk skala (der af navnet). Vigtigt at bemærke.
- Generel formel:

  $$\log(odds_{outcome}) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_n \cdot x_n$$

  β kaldes for regressions koefficienter.
- Transformering af *ssh*/risiko/π til log(odds) kaldes logit funktionen:

  $$\log it(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

- *Ssh* vil altid ligge mellem 0 og 1. Odds vil altid ligge mellem 0 og ∞. Log(odds) vil altid ligge mellem $-\infty$ og ∞.
- Kategoriske/binære data kan også indføjes i logistiske regressioner. De vil blot blive sammenlignet med hinanden, således at f.eks. yngste alders gruppe er baseline eller at kvinde er baseline (så får *x*=0 er kvinde og *x*=1 er mand).

Kap. 20: Logistic regression: controlling for confounding and other extensions

- Ingen vigtige formler som man umiddelbart kan regne på. Til dette kræves pc-kraft som man ikke vil have til rådighed/tid til under eksamen.