NTNU
Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i ST2304
Statistisk modellering for biologer og bioteknologer

Faglig kontakt under eksamen: Ola H. Diserud

Tlf.: 93218823

Eksamensdato: Onsdag 21. mai 2014

Eksamenstid: 15-19

Hjelpemiddelkode/Tillatte hjelpemidler: Et håndskrevet gult A4-ark, godkjent kalkulator, *Tabeller og formler i statistikk* (Tapir forlag), *Matematisk formelsamling* (K. Rottmann)

Annen informasjon: Hjelpesider for noen R funksjoner det kan hende du får bruk for følger på side 8

Alle svar skal begrunnes.

Målform/språk: Bokmål

Antall sider: 8

Kontrollert av:

_____

Dato Sign

Merk! Studenter finner sensur i Studentweb. Har du spørsmål om din sensur må du kontakte instituttet ditt. Eksamenskontoret vil ikke kunne svare på slike spørsmål.

Oppgave 1

Et 100(1-α) % konfidensintervall for variansen σ^2 er gitt ved intervallet

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

a) Skriv et R uttrykk som finner kvantilene i kjikvadratfordelingen, hvis vi har $\alpha = 0.05$ og åtte frihetsgrader.
Vis også, ved et R uttrykk, hvordan du kan finne medianen i kjikvadrat-fordelingen.

Kji-kvadratfordelingen kan også brukes i en kji-kvadrattest. Anta at testobservatoren i en slik test får verdien 14,3 med 8 frihetsgrader.

b) Skriv et uttrykk i R som beregner kji-kvadrattestens signifikanssannsynlighet. Hva blir sannsynligheten for at testobservatoren skal få verdi eksakt lik 8?

c) Vis hvordan du kan lage et plot som viser kji-kvadratfordelingen med 8 frihetsgrader.

Oppgave 2

En studie har undersøkt hvor effektiv måling av *creatine kinase* (CK) nivåer i blodet er for å avdekke om en pasient med symptomer på hjerteattakk faktisk har hatt et attakk. Datasettet under viser hvor mange av pasientene med et gitt CK-nivå som etter hvert viste seg å ha hatt et hjerteattakk.

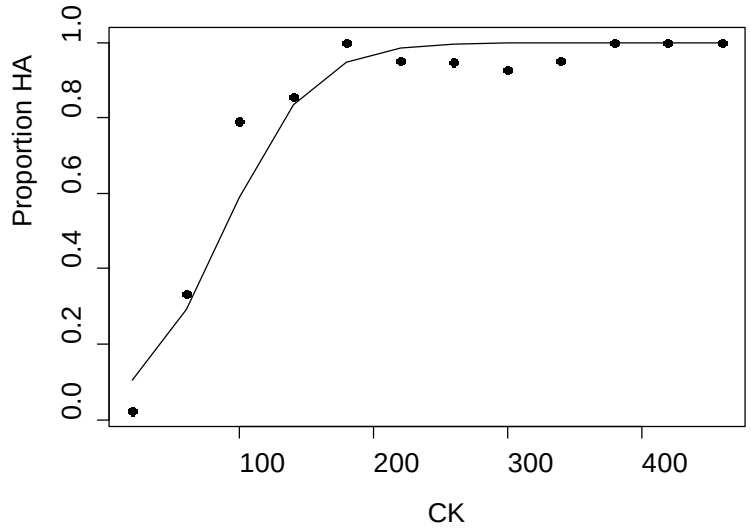
```
> Heart.CK
    CK Heart.Attack Not.Heart.Attack
1   20            2               88
2   60           13               26
3  100           30                8
4  140           30                5
5  180           21                0
6  220           19                1
7  260           18                1
8  300           13                1
9  340           19                1
10 380           15                0
11 420            7                0
12 460            8                0
```

Vi tilpasser først følgende modell (`mod.1`) hvor sannsynligheten for å ha hatt et hjerteattakk forklares ved CK-nivå.

```
> mod.1<-glm(cbind(Heart.Attack,Not.Heart.Attack)~CK,family=binomial)
> summary(mod.1)
```

```
Call:
glm(formula = cbind(Heart.Attack, Not.Heart.Attack) ~ CK, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.08184  -1.93008   0.01652   0.41772   2.60362

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.758358   0.336696  -8.192 2.56e-16 ***
CK           0.031244   0.003619   8.633  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 271.712  on 11  degrees of freedom
Residual deviance:  36.929  on 10  degrees of freedom
AIC: 62.334

Number of Fisher Scoring iterations: 6
```

Figur 1 viser de observerte andelene hjerteattakk for de forskjellige `CK` verdiene, med `mod.1` illustrert ved den heltrukne linja.

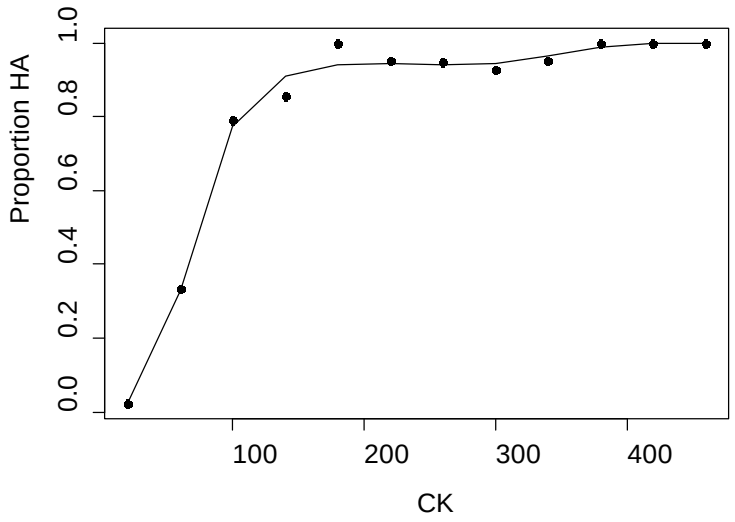


Figur 1: Andel med hjerteattakk mot CK nivå i blodet.

- **a)** Skriv opp modell `mod.1` med matematisk (algebraisk) notasjon og diskuter kort forutsetningene for modellen.
  Hvor stor andel av pasienter med `CK`-nivå på 100 forventes å ha hatt et hjerteattakk?
- **b)** Er det noen tegn til overdispersjon i dataene? Diskuter kort noen mekanismer som kan gi overdispersjon i dette tilfellet.

Basert på tidligere studier, og en residualanalyse for `mod.1`, foreslås en alternativ modell (`mod.2`). Vi introduserer her variablene `CK2` (`CK2<-CK^2`) og `CK3` (`CK3<-CK^3`).

```
> mod.2<-
glm(cbind(Heart.Attack,Not.Heart.Attack)~CK+CK2+CK3,family=binomial)
> summary(mod.2)

Call:
glm(formula = cbind(Heart.Attack, Not.Heart.Attack) ~ CK + CK2 +
    CK3, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.99572  -0.08966   0.07468   0.17815   1.61096

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.786e+00  9.268e-01  -6.243 4.30e-10 ***
CK           1.102e-01  2.139e-02   5.153 2.57e-07 ***
CK2         -4.649e-04  1.381e-04  -3.367  0.00076 ***
CK3          6.448e-07  2.544e-07   2.535  0.01125 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 271.7124  on 11  degrees of freedom
Residual deviance:   4.2525  on  8  degrees of freedom
AIC: 33.658

Number of Fisher Scoring iterations: 6
```



Figur 2: Andel med hjerteattakk mot CK nivå i blodet. Heltrukken linje illustrerer `mod.2`.

c) Hvilken modell foretrekker du? Argumenter for valget.
Forklar kort hvordan du, som en ytterligere vurdering av den valgte modellen, ville ha foretatt en analyse av residualene. Suppler gjerne forklaringen med R-kode.

Oppgave 3

Vi vil her undersøke effekten av vitamin C på tannvekst for marsvin. Dataene er lagt inn som en dataramme i `R`, hvor `len` er tannlengde, `supp` angir hva slags måte vitamin C ble gitt på (`OJ` er appelsinjuice og `VC` er som askorbinsyre), og `dose` er dosestørrelse (`low` = 0,5 mg, `med`=1,0 mg, `high`=2,0 mg). NB: Kun noen av observasjonene er vist under.

```
    len supp dose
1   4.2   VC  low
6  10.0   VC  low
11 16.5   VC  med
16 17.3   VC  med
21 23.6   VC high
26 32.5   VC high
31 15.2   OJ  low
36 10.0   OJ  low
41 19.7   OJ  med
46 25.2   OJ  med
51 25.5   OJ high
56 30.9   OJ high
…
```

En toveis variansanalyse gir følgende ANOVA tabell

```
> anova(lm(len~supp+dose))
Analysis of Variance Table

Response: len
          Df  Sum Sq Mean Sq F value    Pr(>F)
supp       1  205.35  205.35  14.017 0.0004293 ***
dose       2 2426.43 1213.22  82.811 < 2.2e-16 ***
Residuals 56  820.43   14.65
```

a) Tyder utskriften fra variansanalysen på at noen av faktorene har signifikant effekt på tannlengde?
Er det mulig å lese ut fra ANOVA tabellen hvor stor effekt `supp` har på tannlengde?
Hvor mange observasjoner har vi totalt i datasettet?

For å få en liste over parameterestimatene kan vi bruke `summary` funksjonen

```
> summary(lm(len~supp+dose))

Call:
lm(formula = len ~ supp + dose)

Residuals:
   Min     1Q Median     3Q    Max
-7.085 -2.751 -0.800  2.446  9.650
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.4550     0.9883  12.603  < 2e-16 ***
suppVC       -3.7000     0.9883  -3.744 0.000429 ***
dosemed       9.1300     1.2104   7.543 4.38e-10 ***
dosehigh     15.4950     1.2104  12.802  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.828 on 56 degrees of freedom
Multiple R-squared:  0.7623,    Adjusted R-squared:  0.7496
F-statistic: 59.88 on 3 and 56 DF,  p-value: < 2.2e-16
```

b) Hva er forventet tannlengde for et marsvin som får en lav dose med vitamin C, gitt som appelsinjuice? Hvor mye lenger forventes tennene å være for et marsvin som får appelsinjuice enn for et som får askorbinsyre?

c) Til slutt i utskriften over beskrives resultatet fra en F-test. Skriv opp hypoteser og testobservator for denne testen. Hva blir konklusjonen for denne testen?

Chisquare {stats} R Documentation

The (non-central) Chi-Squared Distribution

Description

Density, distribution function, quantile function and random generation for the chi-squared (*chi^2*) distribution with df degrees of freedom and optional non-centrality parameter ncp.

Usage

```
dchisq(x, df, ncp = 0, log = FALSE)
pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qchisq(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
rchisq(n, df, ncp = 0)
```

Arguments

x, q	vector of quantiles.
p	vector of probabilities.
n	number of observations. If length(n) > 1, the length is taken to be the number required.
df	degrees of freedom (non-negative, but can be non-integer).
ncp	non-centrality parameter (non-negative).
log, log.p	logical; if TRUE, probabilities p are given as log(p).
lower.tail	logical; if TRUE (default), probabilities are $P[X \le x]$, otherwise, $P[X > x]$.

Details

The chi-squared distribution with df= $n \ge 0$ degrees of freedom has density

$$f_n(x) = 1 / (2^{n/2} \Gamma(n/2)) \; x^{n/2-1} \; e^{-x/2}$$

for $x > 0$. The mean and variance are n and $2n$.

The non-central chi-squared distribution with df= n degrees of freedom and non-centrality parameterncp = λ has density

$$f(x) = \exp(-\lambda/2) \; SUM_{r=0}^{\infty} \; ((\lambda/2)^r / r!) \; dchisq(x, df + 2r)$$

for $x \ge 0$. For integer n, this is the distribution of the sum of squares of n normals each with variance one, λ being the sum of squares of the normal means; further,
$E(X) = n + \lambda$, $Var(X) = 2(n + 2*\lambda)$, and $E((X - E(X))^3) = 8(n + 3*\lambda)$.

Note that the degrees of freedom df= n, can be non-integer, and also $n = 0$ which is relevant for non-centrality $\lambda > 0$, see Johnson et al. (1995, chapter 29).

Note that ncp values larger than about 1e5 may give inaccurate results with many warnings forpchisq and qchisq.

Value

dchisq gives the density, pchisq gives the distribution function, qchisq gives the quantile function, and rchisq generates random deviates.

Invalid arguments will result in return value NaN, with a warning.

The length of the result is determined by n for rchisq, and is the maximum of the lengths of the numerical parameters for the other functions.

The numerical parameters other than n are recycled to the length of the result. Only the first elements of the logical parameters are used.

Note

Supplying ncp = 0 uses the algorithm for the non-central distribution, which is not the same algorithm used if ncp is omitted. This is to give consistent behaviour in extreme cases with values of ncp very near zero.

The code for non-zero ncp is principally intended to be used for moderate values of ncp: it will not be highly accurate, especially in the tails, for large values.