NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i
ST2304 Statistisk modellering for biologer og bioteknologer

Faglig kontakt under eksamen: Jarle Tufto

Tlf: 99 70 55 19

Eksamensdato: 10. august 2016

Eksamenstid (fra–til): 9–13

Hjelpemiddelkode/Tillatte hjelpemidler: Tabeller og formler i statistikk, Tapir Forlag, K. Rottmann: Matematisk formelsamling, Kalkulator Casio fx-82ES PLUS, CITIZEN SR-270X, CITIZEN SR-270X College eller HP30S, ett gult A4-ark med egne håndskrevne notater.

Annen informasjon:
Hjelpesider for noen R funksjoner som du kan få bruk for følger i vedlegget. Alle svar skal begrunnes og besvarelsen skal inneholde naturlig mellomregning.

Målform/språk: bokmål

Antall sider: 5

Antall sider vedlegg: 1

Kontrollert av:

Dato Sign

Oppgave 1 Anta at T er en t-fordelt variabel med 10 frihetsgrader. La $f(t)$ betegne sannsynlighetstetthetsfunksjonen til T. Skriv R-uttrykk som beregner følgende:

a) $P(T > -1)$.

b) $f(2.5)$.

c) Et tall slik at sannsynligheten for at T er mindre enn tallet er 0.25.

Oppgave 2 I denne oppgaven skal vi undersøke sammenhengen mellom lengde (cm) og vekt (gram) til ulike individ tilhørende 7 ulike fiskearter, se fig. 1.

Vi tilpasser først følgende modell hvor art er inkludert som en kategorisk faktor. Merk at både log og ln her betegner naturlige logaritmer.

```
> mod0 <- lm(log(vekt) ~ art + log(lengde))
> summary(mod0)

Call:
lm(formula = log(vekt) ~ art + log(lengde))

Residuals:
     Min       1Q   Median       3Q      Max
-0.21688 -0.07552 -0.01007  0.05846  0.36861

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.10940    0.12834 -39.810  < 2e-16 ***
art2         0.12626    0.04570   2.763  0.00645 **
art3        -0.06717    0.03308  -2.030  0.04410 *
art4         0.20461    0.04016   5.095 1.04e-06 ***
art5        -0.61635    0.04985 -12.364  < 2e-16 ***
art6        -0.72022    0.03152 -22.849  < 2e-16 ***
art7         0.05513    0.02482   2.221  0.02787 *
log(lengde)  3.15532    0.03491  90.388  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1028 on 149 degrees of freedom
Multiple R-squared:  0.9943,Adjusted R-squared:  0.994
F-statistic:  3710 on 7 and 149 DF,  p-value: < 2.2e-16

> drop1(mod0,test="F")
Single term deletions

Model:
log(vekt) ~ art + log(lengde)
```

```
            Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                    1.574 -706.61
art          6    13.299 14.873 -366.00  209.82 < 2.2e-16 ***
log(lengde)  1    86.307 87.881  -77.10 8169.91 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

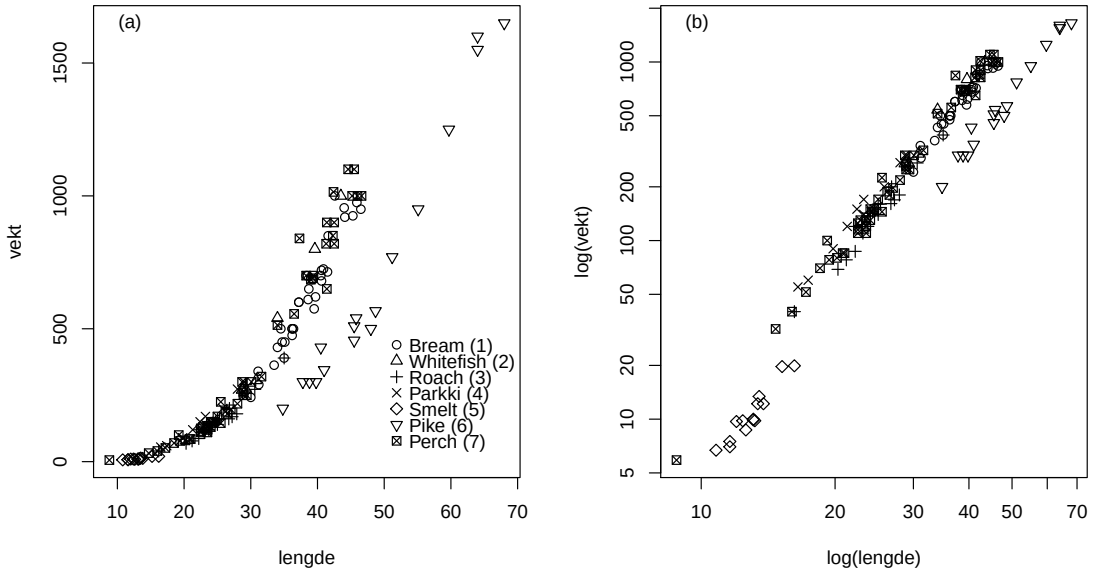
a) La $\ln y$ betegne responsvariabelen (log vekt), $\ln x$ den numeriske forklaringsvariabelen (log lengde) og i den kategoriske faktoren (art). Velg passende navn på ukjente parametere og andre variable og skriv så opp modellen i matematisk notasjon. Gjør rede for modellantakelsene.

b) Hvilke ulike alternative modeller blir sammenlignet ved bruk av `drop1` i utskriften ovenfor? Hva betyr det at testene er signifikante? Hvor mange prosent mindre er estimert forventet vekt til et individ tilhørende art nummer seks (Pike) i forhold til art nummer én (Bream) gitt at individene har samme lengde?

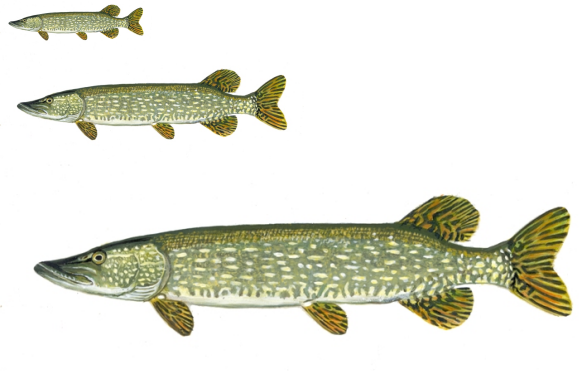
c) Hva er sammenhengen mellom vekt y og lengde x innen en gitt art i? Hvis individ av ulike størrelser er formlike (se fig. 2) og har samme tetthet (masse per volum), hva blir da sammenhengen mellom vekt y og lengde x? Utfør en test av denne nullhypotesen. Det oppgis at 2.5%-kvantilen til en t-fordelt variabel med 149 frihetsgrader er 1.976.

Vi tilpasser til sist en modell hvor vi inkluderer en interaksjon mellom $\ln x$ (log lengde) og i (art). Denne sammenlignes så med den tidligere modellen som følger:

```
> mod1 <- lm(log(vekt) ~ art + log(lengde) + art:log(lengde))
> summary(mod1)

Call:
lm(formula = log(vekt) ~ art + log(lengde) + art:log(lengde))

Residuals:
     Min       1Q   Median       3Q      Max
-0.21976 -0.07410 -0.00303  0.05732  0.36650

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4.941734   0.595909  -8.293 7.39e-14 ***
art2             -0.762713   1.136149  -0.671    0.503
art3             -0.083477   0.760774  -0.110    0.913
art4              0.413773   0.828279   0.500    0.618
art5             -0.326589   0.913790  -0.357    0.721
art6             -1.064880   0.775914  -1.372    0.172
art7             -0.137106   0.611482  -0.224    0.823
log(lengde)       3.109276   0.163559  19.010  < 2e-16 ***
art2:log(lengde)  0.250763   0.319349   0.785    0.434
art3:log(lengde) -0.001124   0.220029  -0.005    0.996
```



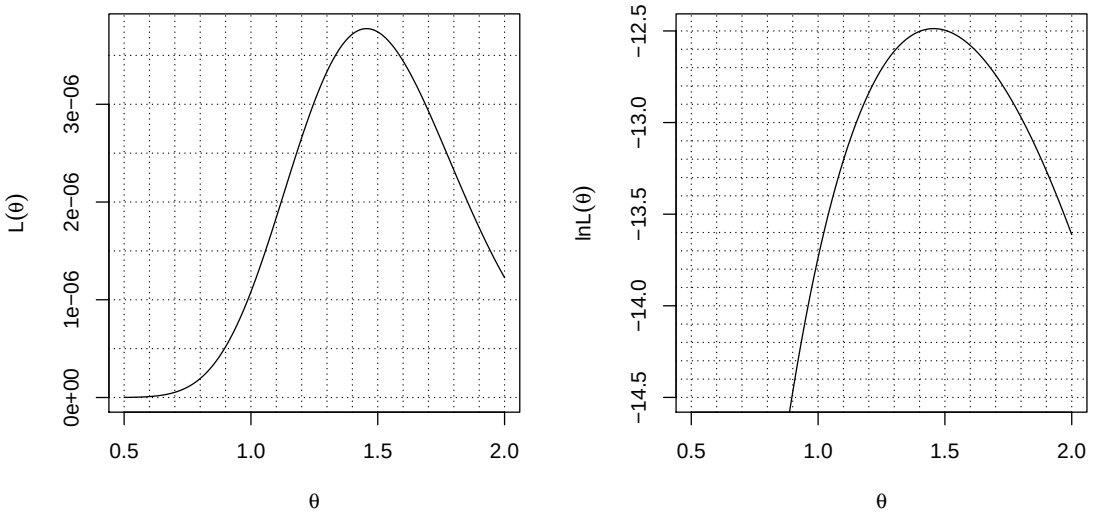
Figur 1: Lengde og vekt til 159 forskjellige individ tilhørende 7 forskjellige fiskearter plottet på orginal skala (a) og på log-log skala (b).



Figur 2: Tre formlike individ av ulike størrelse.

```
art4:log(lengde) -0.075043   0.246613  -0.304    0.761
art5:log(lengde) -0.132480   0.315800  -0.420    0.675
art6:log(lengde)  0.091820   0.207914   0.442    0.659
art7:log(lengde)  0.053412   0.168604   0.317    0.752
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1043 on 143 degrees of freedom
Multiple R-squared:  0.9944,Adjusted R-squared:  0.9939
F-statistic:  1941 on 13 and 143 DF,  p-value: < 2.2e-16

> anova(mod0,mod1,test="F")
Analysis of Variance Table

Model 1: log(vekt) ~ art + log(lengde)
Model 2: log(vekt) ~ art + log(lengde) + art:log(lengde)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    149 1.5740
2    143 1.5548  6  0.019199 0.2943 0.9388
```

d) Forklar med ord hvilken nullhypotese som testes mot hvilken alternativ hypotese i F-testen ovenfor. Hva er resultatet av testen? Virker resultatet rimelig i forhold til de observerte dataene i fig. 1b?



Figur 3: Likelihood- og log-likelihoodfunksjon for modellen og dataene i oppgave 3

.

Oppgave 3 Vi ønsker å estimere en ukjent parameter θ. Vi samler inn et datasett og lager plot av likelihood- og log-likelihood funksjonen som vist i fig. 3.

a) Hva er et fornuftig estimat $\hat{\theta}$ av den ukjente parameteren θ gitt plottet i fig. 3?

b) Utfør en tilnærmet test av nullhypotesen $H_0 : \theta = 1$ versus den alternative hypotesen $H_1 : \theta \neq 1$ ved bruk av informasjon som kan leses av samme plot.

c) Det oppgis at $\frac{d^2}{d\theta^2} \ln L = -9.44$ for θ lik estimatet i punkt a). Bruk dette til å finne et tilnærmet estimat av standardfeilen til $\hat{\theta}$.

TDist *The Student t Distribution*

Description

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

Usage

```
dt(x, df, ncp, log = FALSE)
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)
rt(n, df, ncp)
```

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 (> 0, maybe non-integer). df = Inf is allowed.
ncp	non-centrality parameter δ; currently except for rt(), only for abs(ncp) <= 37.6 If omitted, use the central t distribution.
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 t distribution with df $= \nu$ degrees of freedom has density

$$f(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}(1 + x^2/\nu)^{-(\nu+1)/2}$$

for all real x. It has mean 0 (for $\nu > 1$) and variance $\frac{\nu}{\nu-2}$ (for $\nu > 2$).

The general *non-central* t with parameters (ν, δ) = (df, ncp) is defined as the distribution of $T_\nu(\delta) := (U + \delta)/\sqrt{V/\nu}$ where U and V are independent random variables, $U \sim \mathcal{N}(0, 1)$ and $V \sim \chi^2_\nu$ (see Chisquare).

The most used applications are power calculations for t-tests:
Let $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ where $\bar{X}$ is the mean and S the sample standard deviation (sd) of $X_1, X_2, \ldots, X_n$ which are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ Then T is distributed as non-central t with df $= n - 1$ degrees of freedom and **n**on-**c**entrality **p**arameter ncp $= (\mu - \mu_0)\sqrt{n}/\sigma$.

Value

dt gives the density, pt gives the distribution function, qt gives the quantile function, and rt generates random deviates.

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

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

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

Source

The central dt is computed via an accurate formula provided by Catherine Loader (see the reference in dbinom).

For the non-central case of dt, C code contributed by Claus Ekstroem based on the relationship (for $x \neq 0$) to the cumulative distribution.

For the central case of pt, a normal approximation in the tails, otherwise via pbeta.

For the non-central case of pt based on a C translation of

Lenth, R. V. (1989). *Algorithm AS 243* — Cumulative distribution function of the non-central t distribution, *Applied Statistics* **38**, 185–189.

This computes the lower tail only, so the upper tail suffers from cancellation and a warning will be given when this is likely to be significant.

For central qt, a C translation of

Hill, G. W. (1970) Algorithm 396: Student's t-quantiles. *Communications of the ACM*, **13(10)**, 619–620.

altered to take account of

Hill, G. W. (1981) Remark on Algorithm 396, *ACM Transactions on Mathematical Software*, **7**, 250–1.

The non-central case is done by inversion.

References

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole. (Except non-central versions.)

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995) *Continuous Univariate Distributions*, volume 2, chapters 28 and 31. Wiley, New York.

See Also

Distributions for other standard distributions, including df for the F distribution.

Examples

```
require(graphics)

1 - pt(1:5, df = 1)
qt(.975, df = c(1:10,20,50,100,1000))

tt <- seq(0, 10, len = 21)
ncp <- seq(0, 6, len = 31)
ptn <- outer(tt, ncp, function(t, d) pt(t, df = 3, ncp = d))
t.tit <- "Non-central t - Probabilities"
image(tt, ncp, ptn, zlim = c(0,1), main = t.tit)
persp(tt, ncp, ptn, zlim = 0:1, r = 2, phi = 20, theta = 200, main = t.tit,
      xlab = "t", ylab = "non-centrality parameter",
      zlab = "Pr(T <= t)")

plot(function(x) dt(x, df = 3, ncp = 2), -3, 11, ylim = c(0, 0.32),
     main = "Non-central t - Density", yaxs = "i")
```