**Norges teknisk-naturvitenskapelige universitet**
**Institutt for matematiske fag**



*Bokmål*

Faglig kontakt under eksamen: Professor Jarle Tufto
Telefon: 99705519

# Statistisk modellering for biologer og bioteknologer, ST2304

11. august, 2012
Kl. 9–13
Sensur: 3 uker etter eksamen

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

Hjelpesider for noen R funksjoner det kan hende du får bruk for følger på side 5.

**Oppgave 1** Anta at vi trekker et tilfeldig utvalg på 20 invidid uten tilbakelegging fra en populasjon bestående av totalt 100 indvidid. Det er 60 hunner og 40 hanner i populasjonen. La den stokastiske variabelen $X$ betegne antall hanner i utvalget.

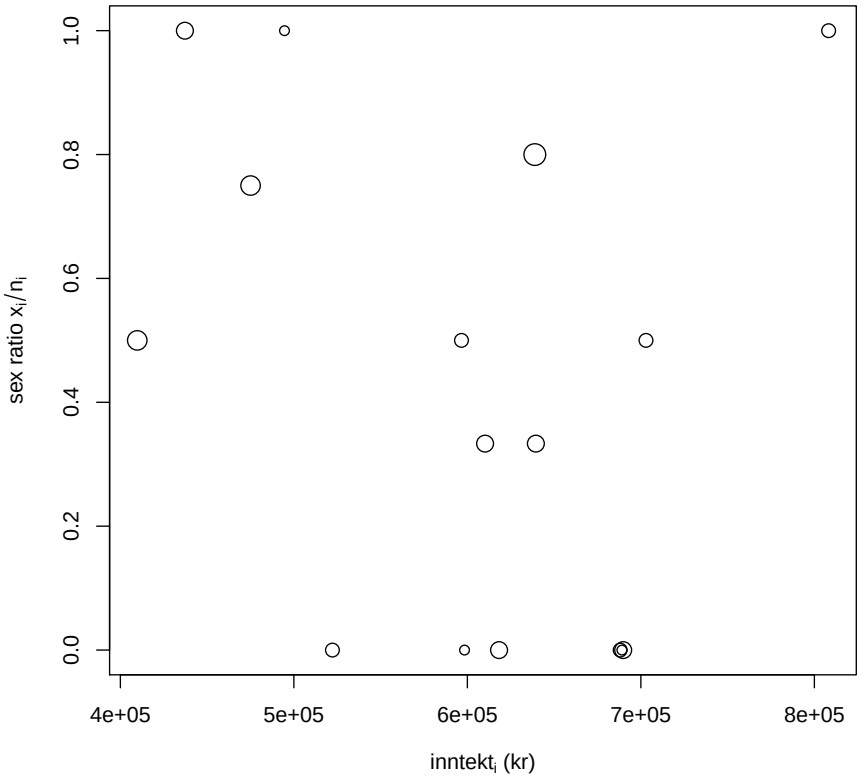
**a)** Hvilken fordeling har $X$ og hvilke verdier har parameterne i fordelingen? Skriv et R-uttrykk som simulerer én realisasjon av $X$.

**b)** Skriv et R-uttrykk som beregner sannsynligheten for at det er nøyaktig 12 hanner i utvalget.

**c)** Skriv et R-uttrykk som beregner sannsynligheten for at det er flere hanner enn hunner i utvalget.

**d)** Skriv et R-uttrykk som beregner en verdi $x$ slik at $P(X \leq x) = 1/2$. Hva kalles størrelsen $x$ når den er definert som her?

**Oppgave 2** Hos mange pattedyr, blant annet hos hjort, er det mer kostbart å oppfostre hannlige avkom. Dette kan medføre at kjønnet til avkom kan avhenge av kondisjon til moren.




Figur 1: Inntekt versus kjønnsrate i ulike familier

Vi ønsker å undersøke om noe av det samme er tilfelle hos menneske ved å se på om kjønnsraten $p_i$ (sannsynlighet for gutt) avhenger av inntekt i ulike familier $i = 1, 2, ..., N$.

Vi observerer dataene gjengitt i fig. 1 hvor $x_i$ er antall gutter i hver familie og $n_i$ er totalt antall avkom. Arealet av hver sirkel er valgt proposjonal med $n_i$. Vi tilpasser så en generalisert lineær modell på følgende måte.

```
> data.frame(n,x,inntekt)
   n x  inntekt
1  1 0 689059.3
2  1 0 598366.6
3  1 1 494596.1
4  2 0 688080.4
5  2 0 522227.6
```

```
6  2 2 808220.4
7  3 0 618272.8
8  3 1 610209.0
9  3 3 437193.1
10 3 0 689810.7
11 3 1 639486.0
12 4 2 409728.7
13 4 3 475050.0
14 5 4 638889.7
15 2 1 702950.6
16 2 1 596607.6
> mod <- glm(cbind(x,n-x) ~ inntekt,fam=binomial)
> summary(mod)

Call:
glm(formula = cbind(x, n - x) ~ inntekt, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8538  -1.1679  -0.3434   0.8548   2.2534

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.963e+00  1.801e+00   1.090    0.276
inntekt     -3.592e-06  3.022e-06  -1.188    0.235

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.387  on 15  degrees of freedom
Residual deviance: 26.921  on 14  degrees of freedom
AIC: 42.41

Number of Fisher Scoring iterations: 3
```

**a)** Skriv opp modellen vi har tilpasset i matematisk notasjon. Forklar hvorfor modellantakelsene er fornuftige.

**b)** Skriv opp nullhypotesen og alternativ hypotese som vi har fått utført en test av i linjen som begynner med `inntekt` i utskriften fra `summary( )` ovenfor. Uttrykk hypotesene ved parameterne du har definert i forrige punkt. Forklar også hva hypotesene innebærer med ord. Hva er testens konklusjon?

**c)** Hva blir estimatet av sannsynligheten for gutt i en familie med inntekt lik 600000 kr?

**d)** Test om det er overdispersjon i dataene. Diskuter ulike mekansismer som kan tenkes å generere overdispersjon i dette konkrete tilfelle.

**Oppgave 3** Anta at vi trekke et tilfeldig utvalg på 100 individ fra en tilnærmet uendelig stor populasjon. Vi ønsker å undersøke om populasjonen er i Hardy-Weinberg-likevekt.

**a)** Om vi lar de stokastiske variablene $X_1, X_2, X_3$ beregne de observerte antallene av hver genotype i utvalget, hvilken fordeling har $X_1, X_2, X_3$?

I utvalget observerer vi 10 individ av genotype $AA$, 65 av genotype $Aa$ og 25 av genotype $aa$.

**b)** Hva blir estimat av frekvensen av genvarianten (allelet) $A$? Hva blir estimat av de tre genotypefrekvensen i hele populasjonen om vi forutsetter at populasjonen er i Hardy-Weinberg likevekt?

**c)** Utfør en test av om populasjonen er i Hardy-Weinberg likevekt. Hvor mange frihetsgrader har testobservatoren (må begrunnes).

```
Hypergeometric              package:stats              R Documentation

The Hypergeometric Distribution

Description:

     Density, distribution function, quantile function and random
     generation for the hypergeometric distribution.

Usage:

     dhyper(x, m, n, k, log = FALSE)
     phyper(q, m, n, k, lower.tail = TRUE, log.p = FALSE)
     qhyper(p, m, n, k, lower.tail = TRUE, log.p = FALSE)
     rhyper(nn, m, n, k)

Arguments:

    x, q: vector of quantiles representing the number of white balls
          drawn without replacement from an urn which contains both
          black and white balls.

       m: the number of white balls in the urn.

       n: the number of black balls in the urn.

       k: the number of balls drawn from the urn.

       p: probability, it must be between 0 and 1.

      nn: number of observations. If 'length(nn) > 1', the length is
          taken to be the number required.

log, log.p: logical; if TRUE, probabilities p are given as log(p).

lower.tail: logical; if TRUE (default), probabilities are P[X <= x],
          otherwise, P[X > x].

Details:

     The hypergeometric distribution is used for sampling _without_
     replacement.  The density of this distribution with parameters
     'm', 'n' and 'k' (named Np, N-Np, and n, respectively in the
     reference below) is given by

          p(x) =      choose(m, x) choose(n, k-x) / choose(m+n, k)

     for x = 0, ..., k.

     The quantile is defined as the smallest value x such that F(x) >=
     p, where F is the distribution function.

Value:

     'dhyper' gives the density, 'phyper' gives the distribution
     function, 'qhyper' gives the quantile function, and 'rhyper'
     generates random deviates.

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

Source:

     'dhyper' computes via binomial probabilities, using code
     contributed by Catherine Loader (see 'dbinom').

     'phyper' is based on calculating 'dhyper' and
     'phyper(...)/dhyper(...)' (as a summation), based on ideas of Ian
     Smith and Morten Welinder.

     'qhyper' is based on inversion.

     'rhyper' is based on a corrected version of

     Kachitvichyanukul, V. and Schmeiser, B. (1985).  Computer
     generation of hypergeometric random variates.  _Journal of
     Statistical Computation and Simulation_, *22*, 127-145.

References:

     Johnson, N. L., Kotz, S., and Kemp, A. W. (1992) _Univariate
     Discrete Distributions_, Second Edition. New York: Wiley.

See Also:

     Distributions for other standard distributions.

Examples:

     m <- 10; n <- 7; k <- 8
     x <- 0:(k+1)
     rbind(phyper(x, m, n, k), dhyper(x, m, n, k))
     all(phyper(x, m, n, k) == cumsum(dhyper(x, m, n, k)))# FALSE
     ## but error is very small:
     signif(phyper(x, m, n, k) - cumsum(dhyper(x, m, n, k)), digits=3)
```