

# Assignment 1, ST2304

**Problem 1** Last inn datasettet `mammals.dat` ved hjelp av kommandoen

```
mammals <- read.table("http://www.math.ntnu.no/~jarlet/statmod/mammals.dat",  
                      header=T)
```

Datasettet inneholder hjernestørrelse (i gram) og kroppsvekt (i kg) til 62 landpattedyr. Inspiser datasettet ved å skrive `mammals`.

Gjør variablene i `data.frame` direkte tilgjengelige ved å skrive

```
attach(mammals)
```

Beregn hvor stor andel hjernestørrelse i gjennomsnitt utgjør av total kroppsstørrelse for dette konkrete datasettet (merk at variablene har forskjellig benevning).

Lag et spredningsplot (`?plot`) med kroppstørrelse langs førsteaksen og hjernestørrelse langs andreaksen. Ser sammenhengen mellom hjernestørrelse og kroppstørrelse ut til å være lineær?

Lag nye variabler `logbrain` og `logbody` ved å logtransformere. Bruk naturlige logaritmer (se `?log`) og lag et nytt spredningsplot av log hjernestørrelse versus log kroppstørrelse. Velg passende tekst på hver akse (inkluder benevning) ved hjelp av argumentene `xlab` og `ylab` (se `?plot`).

Ut i fra dette plottet er det rimelig å anta at log hjernestørrelse er normalfordelt og med forventning som er lineært avhengig av log kroppsvekt, altså modellen

$$\log \text{ brain} = \alpha + \beta \log \text{ body} + e \quad (1)$$

hvor  $e \sim N(0, \sigma^2)$ . Hva slags sammenheng mellom de opprinnelige ikke-transformerte variablene medfører dette?

Tilpass denne lineære regresjonsmodellen ved å skrive

```
linreg <- lm(logbrain~logbody)
```

Inspiser objektet `linreg`. Studér også den mer detaljerte informasjon om dette objektet (en tilpasset modell) du får ved å skrive `summary(linreg)`. Hva er estimatet av det vi har kalt  $\alpha$ ,  $\beta$  og  $\sigma$  ovenfor? Er effekten av kroppsvekt på hjernestørrelse statistisk signifikant? Legg den estimerte regresjonslinjen til spredningsplottet v.h.a. kommandoen `abline`.

Legg til navnet på hver art i spredningsplottet ved hjelp av kommandoen `text`. Bruk argumentet `cex` for å kontrollere størrelsen fontstørrelsen på disse artsnavnene. Hvilken art har størst hjernestørrelse i forhold til sin kroppsvekt?

Hva er forventet log hjernestørrelse til denne noe spesielle arten? Hva er sannsynligheten for så stor eller større hjernestørrelse ut i fra den estimerte modellen? Her må du bruke tabell eller eventuelt funksjonen `pnorm`.

Hva er tolkningen av stigningstallet  $\beta$  i regresjonsmodellen vi har tilpasset? Forklar spesielt hvorfor det kanskje kan forventes at dette stigningstallet er eksakt lik 1 og at dette er en rimelig nullhypotese. Utfør en test av denne null-hypotesen, altså  $H_0 : \beta = 1$  vs.  $H_1 : \beta \neq 1$  ut i fra informasjon du får ut ved hjelp av `summary(linreg)`. Hvis konklusjonen blir at  $\beta < 1$ , hva blir i tilfelle tolkningen av dette?

Gå til slutt tilbake til plottet av de originale ikke-transformerte variablene og legg til en kurve som beskriver sammenhengen mellom hjerne- og kroppsvekt (bruk funksjonen `curve` med tilleggsargumentet `add=T`).

**Problem 2** Beregn sannsynligheten for at minst to av 23 personer har fødselsdag på samme dag. Anta at fødselsdagene er uniformt fordelt på årets 365 dager. Du kan hende du får bruk for bl.a. funksjonene `factorial` og `choose`. Hint: Jobb om nødvendig med logaritmene til tallstørrelsene som inngår i uttrykket.

**Problem 3** Simuler tre tilfeldig utvalg med utvalgstørrelse  $n$  lik henholdvis 50, 1000 og 1000000 fra en normalfordeling med forventning  $\mu = 2.5$  og standardavvik  $\sigma = 1.5$  ved hjelp av funksjonen `rnorm`. Lag histogram over de tilfeldige utvalgene ved hjelp av funksjonen `hist`. Merk at grafikkvinduet kan deles i f.eks. tre med kommandoen `par(mfrow=c(3,1))`. Legg plot av tetthetsfunksjonen  $f(x)$  til i histogrammene ved hjelp av `curve`. Studer hjelpesidene for å oppnå resultatet du ønsker, spesielt `hist` sine argumenter `breaks` og `freq`.