STK 1110 høsten 2016
Oblig II

Dette er oppgavesettet for andre obligatoriske innleveringsprosjekt for kurset STK 1110. Det legges ut på kurssiden **torsdag 27/x/16**, og leveringsfristen er **torsdag 10/xi/16** Besvarelsen skal leveres til instituttkontoret ved Matematisk institutt, senest kl. 14:30 den dagen. Sjekk de praktiske detaljene samlet i

`www.uio.no/studier/admin/obligatoriske-aktiviteter/mn-math-oblig.html`,

der det også forklares at du skal bruke en bestemt standardisert «Forside for obligatoriske innleveringer»:

`www.uio.no/studier/admin/obligatoriske-aktiviteter/mn-math-obligforside.pdf`

Hvis flere samarbeider om å løse oppgavene, må likevel hver student levere sin egen besvarelse. Det må gå frem av besvarelsen hvem du eventuelt har samarbeidet med.

Du kan levere håndskrevet eller maskinskrevet (tekstbehandlet) besvarelse, and you may write in bokmål, nynorsk, English or Latin. Der du bruker **R** må utskrifter og plott legges ved (eventuelt opereres inn i tekstbehandlingsdokumentet). Ulike introduksjonshefter for **R** er å finne på verdensveven; jfr. også det som gjennomgås i undervisningen.

Nils Lid Hjort

Oppgave 1

Falling in love is like algebra; you have to go step by step & follow the steps for it to turn out right*. Resultater fra de noe algebraiske operasjonene vi skal gjennom her, skal om ikke annet hjelpe oss i oppgave 2. Jeg minner om at for en stokastisk variabel A med forventning ξ, så gjelder $\mathrm{Var}\, A = \mathrm{E}\,(A-\xi)^2 = \mathrm{E}\, A^2 - \xi^2$. Spesielt er altså $\mathrm{E}\, A^2 = \xi^2 + \mathrm{Var}\, A$. Dessuten er det lov å vite, eller kjapt vise, uten at det er avkrevet her, at for vilkårlige tall $x_1, \ldots, x_n$ med gjennomsnitt $\bar{x} = (1/n)\sum_{i=1}^n x_i$, gjelder

$$\sum_{i=1}^n (x_i - \bar{x}) = 0 \quad \text{og} \quad \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2.$$

(a) La nå $X_1, \ldots, X_n$ være uavhengige stokastiske variable, med samme forventning μ og samme varians σ^2. Vis at $Z = \sum_{i=1}^n (X_i - \bar{X})^2$ har forventning $(n-1)\sigma^2$. Dette er grunnen at til man foretrekker $n-1$ i nevneren for den vanlig definerte empiriske varians

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2,$$

for da blir altså $\mathrm{E}\, S^2 = \sigma^2$.

* or you just end up with a big headache, and wishing you hadn't done it in the start.

(b) Anta nå at X_i-ene fortsatt har samme varians σ^2, men at de kan ha forskjellige forventninger, si $\mu_1, \ldots, \mu_n$, med gjennomsnitt $\bar{\mu} = (1/n)\sum_{i=1}^n \mu_i$. Vis at da blir

$$\mathrm{E}\, S^2 = \sigma^2 + \frac{1}{n-1}\sum_{i=1}^n (\mu_i - \bar{\mu})^2.$$

Den empiriske variansen griper altså fatt i ikke bare variablenes varians, per se, men i hvilken grad forventningene ikke er like.

(c) Av og til må man arbeide med «nokså like» variable, men der disse ikke har helt like forventninger, og heller ikke helt like varianser. Anta derfor at X_i-ene er uavhengige, med forventninger $\mathrm{E}\, X_i = \mu_i$ og varianser $\mathrm{Var}\, X_i = \sigma_i^2$. Vi skal arbeide med

$$Q = \sum_{i=1}^n \frac{(X_i - \widehat{\mu})^2}{\sigma_i^2}, \quad \text{der} \quad \widehat{\mu} = \frac{\sum_{i=1}^n X_i/\sigma_i^2}{\sum_{i=1}^n 1/\sigma_i^2} = \frac{1}{A}\sum_{i=1}^n \frac{X_i}{\sigma_i^2},$$

der altså $A = \sum_{i=1}^n 1/\sigma_i^2$. Vis at

$$\mathrm{E}\, Q = n - 1 + \sum_{i=1}^n \frac{(\mu_i - \widetilde{\mu})^2}{\sigma_i^2},$$

der

$$\widetilde{\mu} = \frac{\sum_{i=1}^n \mu_i/\sigma_i^2}{\sum_{i=1}^n 1/\sigma_i^2} = \frac{1}{A}\sum_{i=1}^n \frac{\mu_i}{\sigma_i^2}.$$

(d) Skriv ut hva resultatet fra (c) gir, for de to tilfellene (i) alle σ_i-ene er like; (ii) alle μ_i-ene er like.

Oppgave 2

ET BILDE SIER MER ENN CA. FØRTI ORD, er noen villige til å mene, men hvor lange er disse ordene? Det vet studentene i dette kurs noe om, takket være oppgavesettet til Oblig I. Hver student skred i den forbindelse til sin bokhylle, fant 100 ord og deres lengder i en norsk bok, og 100 ord og deres lengder i en engelsk bok. Vi har ikke tatt vare på absolutt all informasjonen i disse $2 \times 64 \times (100 + 100)$ ord, men på essensen, nemlig gjennomsnitt $\bar{x}$ og empirisk standardavvik $\widehat{\sigma}$, for hver av $n = 64$ studenter, for de to språk. Datafilen `wordlengths-data` på kurssiden inneholder

$$(\bar{x}_{i,N}, \widehat{\sigma}_{i,N}, \bar{x}_{i,E}, \widehat{\sigma}_{i,E}) \quad \text{for } i = 1, \ldots, n.$$

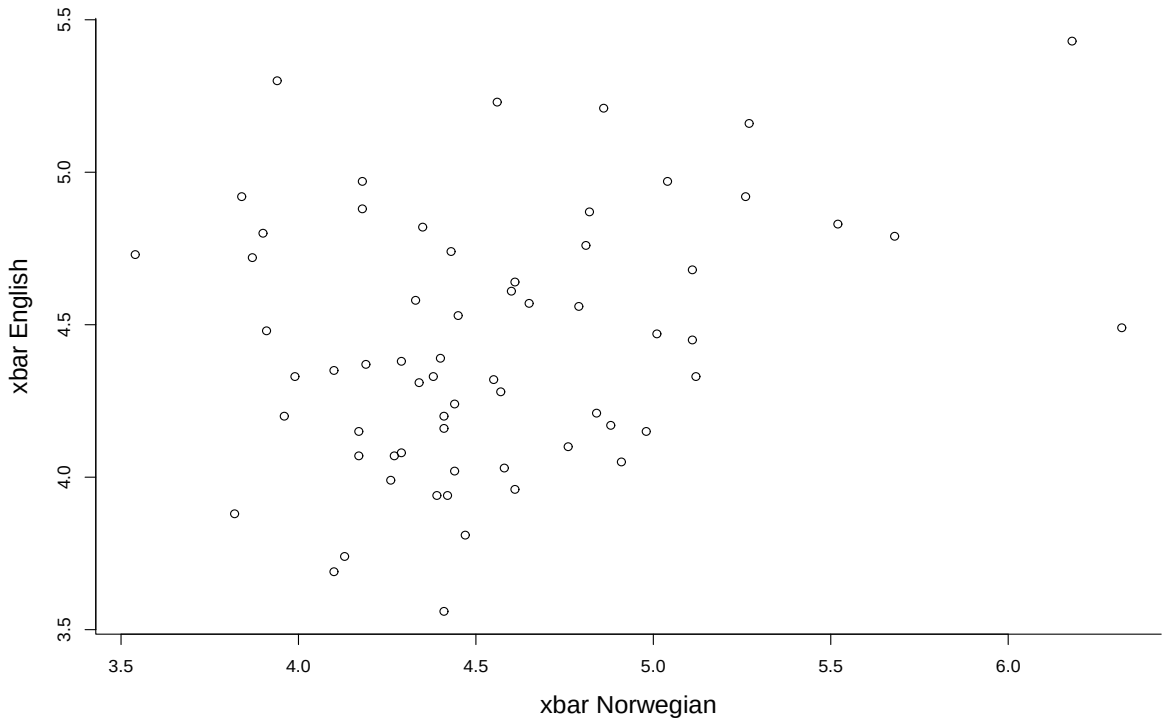
Om man lagrer denne datafilen på eget område, skal man kunne aksessere disse dataene ved å skrive

```
wordlengths <- matrix(scan("wordlengths-data",skip=10),byrow=T,ncol=5)
```

Etter dette kan man f.eks. skrive

```
xbarN <- wordlengths[ ,2]
sigmaN <- wordlengths[ ,3]
xbarE <- wordlengths[ ,4]
sigmaE <- wordlengths[ ,5]
```

fulgt av `plot(xbarN,xbarN)` og diverse statistiske operasjoner; se figur 1, der du til og med skal kunne se ditt eget bidrag til statistisklingvistisk erkjennelse.



Figur 1: $(\bar{x}_{i,N}, \bar{x}_{i,E})$*, gjennomsnittsordlengder for norske og engelske bøker, for* $n = 64$ *lesende studenter.*

(a) Stoltenberg og Hjort har altså tillatt seg å ikke ta vare på mer enn nettopp $(\bar{x}, \widehat{\sigma})$, for hver av studentene, for hvert av de to språk. I hvilken forstand kan de unne seg å ha god statistisk samvittighet, i denne anledning?

(b) Lag et slikt plott, over $(\bar{x}_{i,N}, \bar{x}_{i,E})$. Gi argumenter som gjør det rimelig å anta, ihvertfall som en approksimasjon, at disse parene kommer fra en binormal fordeling. Denne har en viss korrelasjonskoeffisient, ρ. Estimer denne, med den vanlige empiriske korrelasjonskoeffisienten, $\widehat{\rho}$. Du skal også lage et tilnærmet 95% konfidensintervall for ρ, ved å anvende «transformasjonsknepet», at

$$A(\widehat{\rho}) \approx_d \mathrm{N}(A(\rho), 1/(n-3)), \quad \text{der} \quad A(\rho) = \tfrac{1}{2} \log \frac{1+\rho}{1-\rho}.$$

Gi en tolkning av det du kommer frem til.

(c) Er gjennomsnittlig ordlengde omtrent lik, for norsk og engelsk? Undersøk dette, og formuler en konklusjon.

(d) Lag også et plott over standardavvikene, altså $(\widehat{\sigma}_{i,N}, \widehat{\sigma}_{i,E})$. For å belyse om gjennomsnittlig variansparameter for norske og engelske ordlengder er like, kan du her anvende at hvert estimert standardavvik, $\widehat{\sigma}_{i,N}$ og $\widehat{\sigma}_{i,E}$, har en tilnærmet normalfordeling. (Du kan altså anvende metoder for tester og konfidensintervaller som tar utganspunkt i dette, uten at du trenger å komme inn på de eksakte fordelinger for de $2 \times n$ empiriske standardavvikene.) Som for (c), formuler en konklusjon basert på det du finner ut her.

(e) Men er studentene, det vil si deres bokhyller, omtrent like, eller er det systematiske forskjeller mellom dem? Foretrekker noen studenter knappordet poesi, der andre best liker langordsfrekvente middelalderromaner? – En måte å nærme seg dette spørsmålet på er som følger. Vi starter med

$$\bar{x}_i \sim \mathrm{N}(\mu_i, \kappa_i^2) \quad \text{for } i = 1, \ldots, n$$

(for eksempel for det norske utvalget, med helt tilsvarende for det engelske), der vi også vet at $\kappa_i^2 = \sigma_i^2/100$, med gode estimater for σ_i-ene. Hvilken tolkning har μ_i her? Så tenker vi oss at μ_i-ene stammer fra sin egen normalfordeling, $\mu_i \sim \mathrm{N}(\mu_0, \tau^2)$. Spørsmålet er om τ er null eller om ikke annet temmelig liten, eller om den er merkbar. Hvilken tolkning vil du gi τ?

(i) Anta i første omgang at σ_i-ene er tilnærmet like, og beregn

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{x}_i - \bar{x})^2,$$

der $\bar{x}$ er det store gjennomsnittet av gjennomsnittene. Bruk så resultater fra oppgave 1 til å estimere τ, for det norske og det engelske ordutvalget. Gi også en test for hypotesen om full likhet mellom μ_i-ene (som svarer til at $\tau = 0$).

(ii) Prøv så å gjennomføre en analog, men litt mer kompleks analyse, igjen ved å bruke resultater fra oppgave 1, eller noe analogt, men der du ikke antar at σ_i-ene er like.

Oppgave 3

«Bra påvist! Påvis mer!», skrev altså *Natt og Dag* begeistret, for noen uker siden, om de temaene min PhD-student Céline Cunen har arbeidet med, der hun sammenligner aspekter ved rosekrigene (1455–1487) og Game of Thrones (det litterære univers skapt av George R.R. Martin). Dette har vi arbeidet med i Oblig I, og i denne oppgaven skal skal vi fortsette, i to nye retninger. Den første er å anvende gamma-fordelingen, med momentbaserte estimatorer (der Oblig I kjørte Weibull, med kvantilbasert estimering); den andre er bootstrapping. Se også Cunens bloggpost i FocuStat-prosjektet, og kom på hennes faglig-pedagogiske foredrag 3/11.

Gamma-fordelingen har tetthet

$$g(x, a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) \quad \text{for } x > 0. \qquad (\spadesuit)$$

Her er (a, b) positive parametre, og $\Gamma(a)$ er den såkalte gamma-funksjonen. Læreboken anvender en noe annerledes og mindre vanlig parametrisering, men jeg velger altså den i (♠), og som også er den som er benyttet i **R** via `dgamma, pgamma, qgamma, rgamma`.

(a) Vis at forventning og varians for denne fordelingen blir

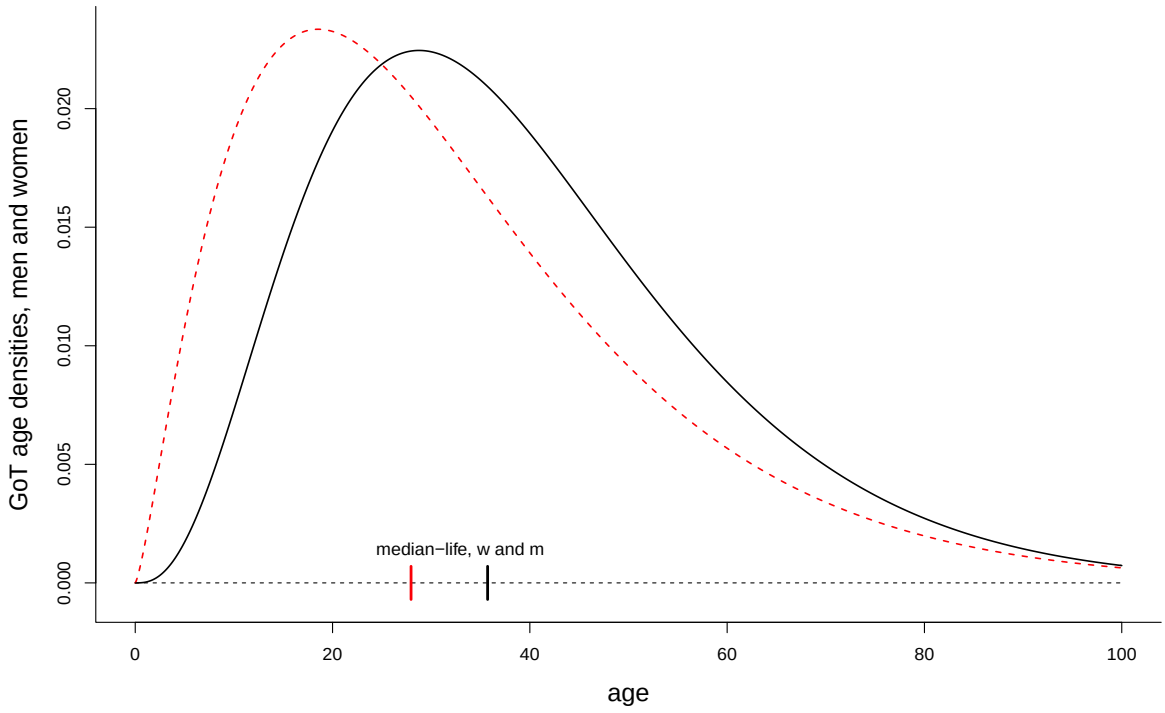
$$\mathrm{E}\, X = \frac{a}{b} \quad \text{og} \quad \mathrm{Var}\, X = \frac{a}{b^2}.$$

Her kan du få bruk at $\Gamma(u+1) = u\Gamma(u)$, for hver u, slik at også $\Gamma(u+2) = (u+1)u\Gamma(u)$, osv.

(b) Anta $X_1, \ldots, X_n$ er uavhengige observasjoner fra en slik gamma-fordeling. Forklar at momentmetoden for å estimere (a, b) består i å løse de to ligningene

$$\bar{X} = \frac{\widehat{a}}{\widehat{b}} \quad \text{og} \quad S^2 = \frac{\widehat{a}}{\widehat{b}^2},$$

der som vanlig $\bar{X}$ er gjennomsnittet og S^2 den empiriske varians. Løs disse, og sett opp eksplisitte formler for estimatorene $(\widehat{a}, \widehat{b})$.



Figur 2: Estimerte tettheter $\widehat{f}_m(x) = g(x, \widehat{a}_m, \widehat{b}_m)$ for menn og $\widehat{f}_w(x) = g(x, \widehat{a}_w, \widehat{b}_w)$ for kvinner, i GoT-universet, via gamma-fordelinger. De assosierte median-levetidene er også tegnet inn.

(c) Få igjen tak i levetidsdataene for de 26 kvinner og de 127 menn, fra GoT-universet, via oppskriften fra Oblig I. Tilpass gamma-fordelingen til hver av disse datasettene, ved å finne momentestimatene $(\widehat{a}_m, \widehat{b}_m)$ for menn og $(\widehat{a}_w, \widehat{b}_w)$ for kvinner. Lag en versjon av figur 2, og kommenter kort hva du finner ut.

(d) Beregn de estimater for medianlevetiden, q_w for kvinner og q_m for menn, som følger av de estimerte gamma-fordelinger. Det finnes ingen eksplisitt formel for

$$\widehat{q}_m = G^{-1}(\tfrac{1}{2}, \widehat{a}_m, \widehat{b}_m),$$

der $G(x, a, b)$ er den kumulative gamma-fordelingen, men det er ingen sak for **R**, der vi kan bruke

```
qhatw = qgamma(0.50,ahatw,bhatw)
qhatm = qgamma(0.50,ahatm,bhatm)
```

Disse er tegnet inn i figur 2.

(e) Så skal vi bruke bootstrapping for å lage konfidensintervaller for q_w og q_m. Dette kan foregå omtrent slik, i **R**, der maskinen først har fått vite om datasettet `xmen` for mennene, med sample size `nmen` (lik 127):

```
boot <- 1000
qhatmstar <- 0*(1:boot)
for (b in 1:boot)
{
xmenstar <- sample(xmen,nmen,replace=T)
# this is a bootstrap sample of the xxm dataset
[then a few lines and formulas here
to compute ahatmstar and bhatmstar based on dataset xmenstar
using the same formulas for parameter estimates as in (b)]
#
qhatmstar[b] <- qgamma(0.50,ahatmstar,bhatmstar)
}
```

Da har du fått laget deg `boot` bootstrapverdier $\widehat{q}^*_m$, og da gjenstår det bare å be om

```
quantile(qhatmstar,c(0.05,0.95))
```

Bruk dette oppsettet til å finne tilnærmede 90% konfidensintervaller for medianlevetidene q_w og q_m. Kommenter det du finner ut.

(f) Det er verd å registrere at bootstrapmetoden her er temmelig generell, og temmelig uavhengig av matematiske formler. Altså kan man anvende metoden også for mer komplekse problemstillinger enn dette. Velg *en statistisk parameter til*, etter eget ønske, og bruk oppsettet over til å finne estimater og konfidensintervaller, for din parameter, for GoT-populasjonene kvinner og menn. Kommenter igjen det du finner frem til.