STK 1110 høsten 2016
Oblig I

Dette er oppgavesettet for første obligatoriske innleveringsprosjekt for kurset STK 1110. Det legges ut på kurssiden **torsdag 22/ix/16**, og leveringsfristen er **torsdag 6/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; spesielt kreves det at hver student har samlet sitt eget datamateriale for Oppgave 1. 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. Man får også bruk for **R** i forbindelse med Oblig II (med leveringsfrist torsdag 10/xi/16).

Nils Lid Hjort

Oppgave 1

Hvor forskjellige er engelsk og norsk, med hensyn til ordlengde, preposisjonshyppighet, tegnsettingsiver, setningskonstruksjoner, leddsetningsmønstre, osv.? Kan vi se noe som ligner systematiske forskjeller mellom Solstad, Kjærstad, Fløgstad, på den ene side, og Morrison, Lessing, Munro på den annen side, ut fra slike rent kvantitative mål?

Vi skal her bry oss om ett enkelt av disse aspektene, nemlig ordlengdene. Skrid til din bokhylle, og velg én bok på norsk og én på engelsk. Dette kan være romaner, novellesamlinger, prosa-artikler eller noe helt annet, men skal ikke være teknisk fagstoff, med for eksempel matematiske formler. For hver av disse to bøkene skal du så gå gjennom følgende øvelse.

Slå opp på side 51, eventuelt på første «normale tekstside» etter side 51 i det tilfelle at denne siden ikke er en vanlig tekstside. Gå så gjennom hvert av de første hundre ord på siden, og noter antall bokstaver i ordet. (Dersom boken du har valgt har færre enn hundre ord på side 51, gå da videre til side 52, inntil du altså har hundre ordlengder.)

(a) Oppgi hvilke to bøker du har valgt (gi forfatter, tittel, forlag, årstall).

(b) Sammenfatt ordlengderesultatene i to tabeller (en for hver av de to bøkene), og i to histogrammer (en for hver bok).

(c) Kall disse ordlengdene $x_1, x_2, \ldots, x_{100}$, der x_1 er antall bokstaver i ord nr. 1, osv. Beregn gjennomsnittet $\bar{x}$ og medianen M for dine to datasett.

(d) Beregn dessuten standardavviket $\widehat{\sigma}$ for de to datasettene.

(e) Kommenter eventuelle likheter og forskjeller mellom de to datasettene.

(f) Det er forelesers intensjon at de to gjennomsnitt $\bar{x}$ og de to standardavvik $\widehat{\sigma}$ fra hver enkelt student skal tas vare på i en liten database, som så kan analyseres videre, for eksempel mandag 28. november. Er det problemstillinger du mener kan være av spesiell interesse, som kunne forfølges og belyses, gjennom et slikt datamateriale? Med andre ord, er det gode spørsmål du mener kursets foreleser bør stille på eksamen i dette kurset (dersom han altså velger å lage en oppgave basert på dette datamaterialet)?

– De tilstrekkelig interesserte kan få lov til å lese min artikkel *And Quiet Does Not Flow the Don: Statistical Analysis of a Quarrel between Nobel Laureates* fra Centre of Advanced Studies, 2006:
`www.cas.oslo.no/getfile.php/138668/CAS_publications_events/CAS_publications/Seminar_booklets/PDF/Consilience_LidHjort.pdf`

Oppgave 2

«Bra påvist! Påvis mer!», skrev *Natt og Dag* begeistret, forleden uke, om min PhD-student Céline Cunen, som ifølge artikler i *Titan* og *Universitas* har påvist at det er visse statistiske likheter mellom levetidsfordelingene i *Game of Thrones* (GoT, det litterære og filmatiske universet skapt av George R.R. Martin) og Rosekrigene (*The War of Roses*, 1455–1487, som sies å ha vært Martins inspirasjon for GoT-serien).

En del av arbeidet har bestått i å få tak i data – hvor lenge lever de ulike personene; er de menn eller kvinner; er de adelige eller «commoners»? Cunen har saumfart historiebøker og leksikonartikler (hvilket i vår tid delvis kan gjøres elektronisk, ved at visse scripts skraper seg gjennom biter av wikipedia), for å fremskaffe slike opplysninger for Rosekrigene, mens GoT-fans har laget tabeller og oversikter over de ulike skjebner i det universet. Les gjerne Cunens blogg-post *Mortality and Nobility in the Wars of the Roses and Game of Thrones* om dette på prosjektsiden for FocuStat (et NFR-støttet femårsprosjekt jeg leder), og kom på hennes foredrag under Faglig-pedagogisk dag, 3. november.

Noe av poenget med disse undersøkelsene, i tillegg til å lære mer om livet og samfunnsforhold i England på sent 1400-tall, er å sammenligne de to universene, og dette har jeg tenkt at vi skal komme tilbake til i Oblig II (leveringsfrist 10. november). Men i denne oppgaven skal vi kun diskutere aspekter forbundet med GoT-universet. Levetidene for 26 kvinner og 127 menn er gitt her, ferdig sortert (der `.5` betyr 0.5); de kan også leses inn i **R** fra kurssiden:

```
 1   3  12  14  16  16  17  19  19  19  21  24
27  31  32  33  35  36  40  45  50  54  56  64
75  85
```

```
 .5  .5  .5   1   8  10  11  12  13  13  13  13
14  14  14  15  16  16  17  17  17  17  18  19
19  21  21  22  22  23  23  23  24  24  25  25
26  26  27  27  28  29  30  30  31  32  32  33
33  33  34  34  35  35  35  35  35  36  36  36
37  37  37  37  38  38  38  39  39  40  40  40
41  41  41  42  43  43  43  44  45  46  47  47
47  47  48  48  48  48  50  51  52  54  54  55
55  55  55  57  58  59  59  59  60  60  60  60
60  62  64  65  65  65  66  67  67  67  67  67
69  78  79  79  82  84 102
```

(a) La $x_1, \ldots, x_n$ være de $n = 153$ levetidene. Finn median, gjennomsnitt $\bar{x}$ og standardavvik $\widehat{\sigma}$ for dataene.

(b) Bruk metoder fra bokens Chapter 8 til å lage et konfidensintervall for μ, den forventede levetid for populasjonen som GoT representerer, med dekningsgrad (konfidensgrad) tilnærmet 90%. Forklar (og diskuter kort) hvilke antagelser du benytter deg av.

(c) Beregn så gjennomsnitt og standardavvik for de to gruppene hver for seg, altså de 26 kvinner og 127 menn. Idet du *i dette punkt* antar at de to levetidsfordelingene er normalfordelte, lag en test for hypotesen om at det ikke er noen systematisk forskjell mellom forventet levetid μ_m for menn og forventet levetid μ_w for kvinner. Beregn p-verdien, og kommentar det du finner ut.

– Det viser seg at normalfordelingen ikke gir noen særlig god tilpasning til levetidsfordelingene her. Man får bedre tilpasning ved å anvende den såkalte Weibull-fordelingen [mrk. sv. utt.], der den kumulative fordelingsfunksjonen er på formen

$$F(x) = \Pr(X \le x) = 1 - \exp\{-(x/a)^b\} \quad \text{for } x > 0.$$

Her er (a, b) ukjente positive parametre, som må estimeres fra data.

(d) Vis at medianen i denne fordelingen kan uttrykkes som

$$x_0(\tfrac{1}{2}) = a(\log 2)^{1/b}.$$

Finn dessuten en formel for sannsynligheten for at en person fra denne populasjonen skal rekke å bli minst seksti år.

(e) For å estimere parametrene (a, b) i Weibull-fordelingen, for de to populasjonene av menn og kvinner i GoT-universet, skal vi bruke kvantiler. Vis at p-kvantilen $F^{-1}(p)$ kan skrives

$$x_0(p) = a\, c(p)^{1/b}, \quad \text{der } c(p) = -\log(1-p).$$

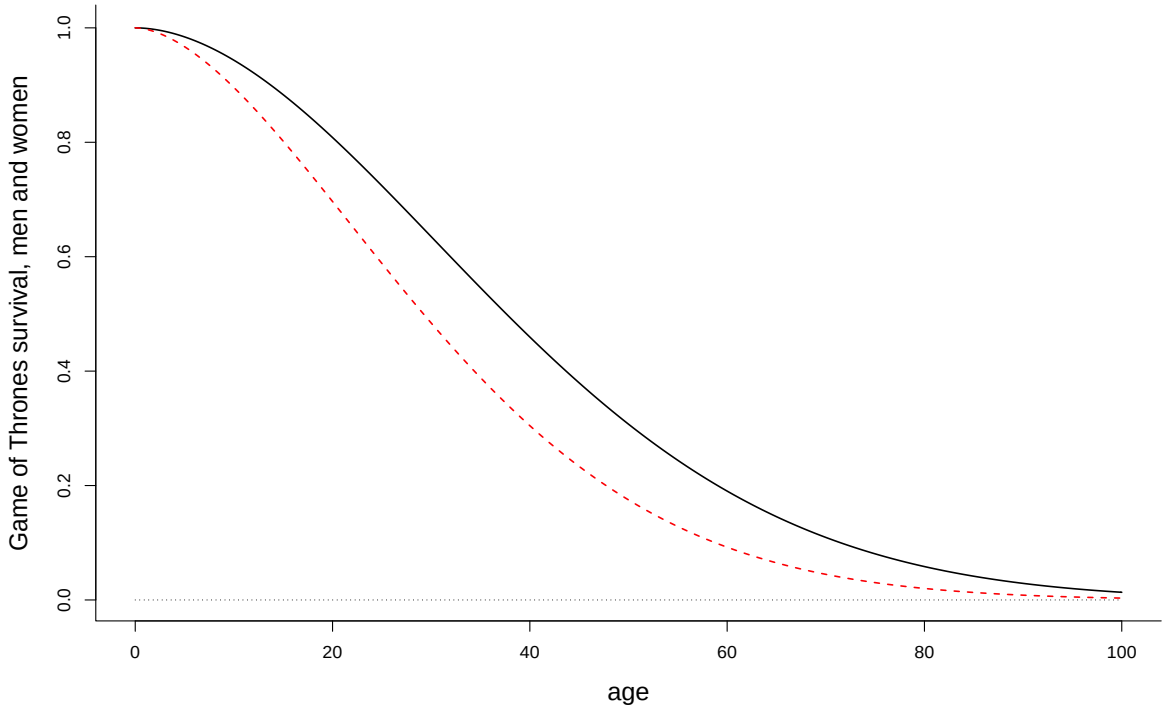
For de to observerte kvartiler z_1 og z_2, altså 0.25- og 0.75-kvantilene i data, lager vi to ligninger med to ukjente:

$$z_1 = a\, c(0.25)^{1/b} \quad \text{og} \quad z_2 = a\, c(0.75)^{1/b}.$$

Finn formler for parameterestimatene $\widehat{a}$ og $\widehat{b}$ som kommer ut av dette.

(f) Finn Weibull-parameterestimater $(\widehat{a}, \widehat{b})$ på denne måten, for menn og for kvinner. Du finner kvartilene fra data ved å skrive
`z1 <- quantile(xx,0.25)` og `z2 <- quantile(xx,0.75)`
der xx er dataene du arbeider med.

(g) Estimer medianen i levetidsfordelingen for menn og for kvinner, i GoT-universet, samt sannsynligheten for at henholdsvis en mann og en kvinne skal rekke å bli seksti år, ved hjelp av Weibull-fordelingene. Prøv også å lage en figur av denne typen, med de såkalte overlevelseskurver $\Pr(X \geq x)$.



(h) For datamateriale $x_1, \ldots, x_n$ fra en Weibull-fordeling som over, skriv ned et uttrykk for log-likelihood-funksjonen $\ell(a, b)$. Denne kan maksimeres via generelle optimiseringsalgoritmer, som `nlm` i **R**; se Appendiks. Om du har tid kan du gjerne supplere parameterestimatene du fant via kvantilmetoden i (f), med maximum-likelihood-estimater, for kvinner og for menn.

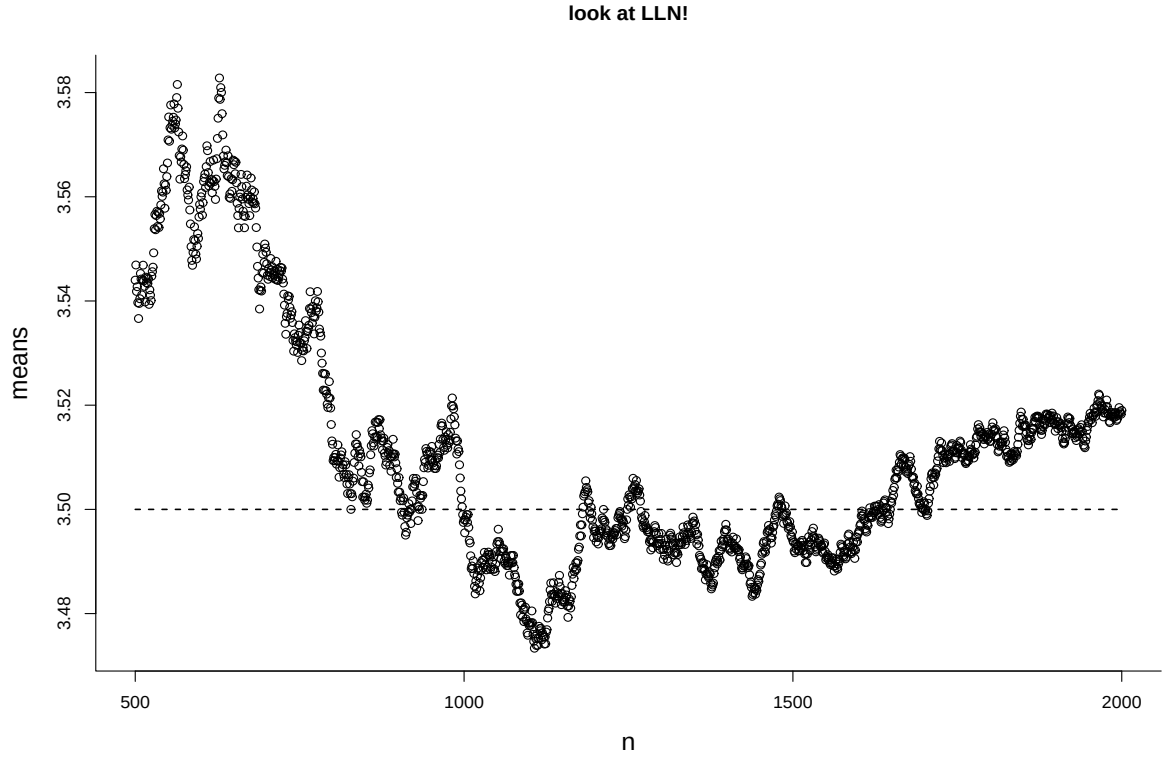
Oppgave 3

"If that die has a 'one' face up, I thought, I'm going downstairs and rape Arlene." Vel, enten man har lest Luke Rhineharts *The Dice Man*, eller for den saks skyld Carl Barks' *Flipism* fra 1953, skal denne oppgaven handle om og illustrere LLN (the Law of Large Numbers, de store talls lov) og CLT (the Central Limit Theorem, sentralgrenseteoremet).

(a) La X være resultatet av ett enkelt terningkast, med en perfekt terning, altså en der hver av utfallene 1, 2, 3, 4, 5, 6 har lik sannsynlighet. Finn forventning μ og standardavvik σ for X.

(b) Jeg kaster min terning en rekke ganger, og får resultatene $X_1, X_2, \ldots$, med gjennomsnitt $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$ etter de første n kast. Hva sier de store talls lov, om denne sekvensen av gjennomsnitt?

(c) Prøv å lage en illustrasjon av de store talls lov, ved å simulere en lang rekke slike terningkast og lage en graf over gjennomsnittene, omtrent som i min figur under. Bruk gjerne **R**-grep fra denne obligens Appendiks, og lag gjerne ditt eget spenn av n-verdier (der jeg har brukt n fra 500 til 2000).



(d) Hva er sannsynligheten for at $\bar{X}_n \in [3.40, 3.60]$, for $n = 500$? Hvor mange ganger må jeg kaste min terning, for at denne sannsynligheten skal være minst 0.99?

(e) Jeg er ikke tindrende sikker på at terningen jeg kaster med er helt perfekt. Etter $n = 500$ kast får jeg gjennomsnitt 3.66. Hva er p-verdien, for den nullhypotese at terningen faktisk er perfekt?

Appendiks: Noen R-grep

Her er noen grep man kan ty til i **R**.

◇ Lese inn data fra fil:

```
xmen <- scan("got_men", skip=3)
```

Her gir `skip=3` signal om antallet linjer i toppen av datafilen som skal ignoreres.

◇ Simulere en sekvens gjennomsnitt (her av terningkast, men det er lett å generalisere):

```
probs <- c(1,1,1,1,1,1)/6
nn <- 1000
xx <- sample(1:6,nn,replace=TRUE,prob=probs)
xbars <- 0*(1:nn)
for (j in 1:nn)
{
xbars[j] <- mean(xx[1:j])
}
plot(200:400,xbars[200:400],xlab="kasta kasta",ylab="gjennomsnitt")
```

◇ Plotting av kurver, f.eks. $f(x) = \exp(\sin(\sqrt{x}))$ over et passende intervall:

```
xval <- seq(0,5,by=0.01)
fval <- exp(sin(sqrt(xval)))
matplot(xval,fval,type="l",xlab="x",ylab="se paa f")
```

Skal man plotte to kurver i samme diagram, f.eks. $f(x)$ og $g(x)$, kan man først lage `fval` og `gval`, og så skrive

```
matplot(xval,cbind(fval,gval),type="l",xlab="x",ylab="f og g")
```

◇ Maksimering av en log-likelihood-funksjon, her for Weibull-fordelingen, basert på data lagret i `xx` (men oppsettet lar seg lett generalisere til andre typer fordelinger): Her er `nlm` den herlige non-linear-minimisation-algoritmen, som finner minimum av en gitt funksjon, selv av temmelig mange variable. Den trenger bare et startpunkt (se under).

```
logL <- function(ab)
{
a <- ab[1]
b <- ab[2]
#
heisan <- -(xx/a)^b + log(b) + (b-1)*log(xx) - b*log(a)
sum(heisan)
}
minuslogL <- function(ab)
-logL(ab)
nils <- nlm(minuslogL,c(40,2),hessian=T)
ML <- nils$estimate
Jhat <- nils$hessian
se <- sqrt(diag(solve(Jhat)))
showme <- cbind(ML,se)
print(round(showme,4))
```