Obligatorisk øving for STK1110, Høsten 2023

Øving 1 av 2

Innleveringsfrist

Torsdag 5. oktober 2023, klokken 14:30 i Canvas (`canvas.uio.no`).

Instruksjoner

Du velger selv om du skriver besvarelsen for hånd og scanner besvarelsen eller om du skriver løsningen direkte inn på datamaskin (for eksempel ved bruk av Latex). Besvarelsen skal leveres som **én PDF-fil**. Scannede ark må være godt lesbare. Besvarelsen skal inneholde navn, emne og oblignummer.

Det forventes at man har en klar og ryddig besvarelse med tydelige begrunnelser. Husk å inkludere alle relevante plott og figurer. Det er kun **ett forsøk**, og det er dermed ikke mulighet til å levere en revidert besvarelse dersom en ikke består. Samarbeid og alle slags hjelpemidler er tillatt og det er spesielt mulig å få hjelp på gruppene. Den innleverte besvarelsen skal imidlertid være skrevet av deg selv og reflektere din forståelse av stoffet. Er vi i tvil om du virkelig har forstått det du har levert inn, kan vi be deg om en muntlig redegjørelse. I oppgaver der du blir bedt om å programmere må du legge ved programkoden og levere den sammen med resten av besvarelsen.

Søknad om utsettelse av innleveringsfrist

Hvis du blir syk eller av andre grunner trenger å søke om utsettelse av innleveringsfristen, må du ta kontakt med studieadministrasjonen ved Matematisk institutt (e-post: studieinfo@math.uio.no) i god tid før innleveringsfristen. For å få adgang til avsluttende eksamen i dette emnet, må man bestå alle obligatoriske oppgaver i ett og samme semester.

Spesielt om dette oppgavesettet

Du **skal bruke programpakken R** til å gjøre beregninger i oppgavene, og du må angi hvilke kommandoer du har brukt for å komme fram til svarene dine. For å få godkjent besvarelsen, må du ha gjort et hederlig forsøk på besvare begge oppgavene (Oppgave 1 og 2).

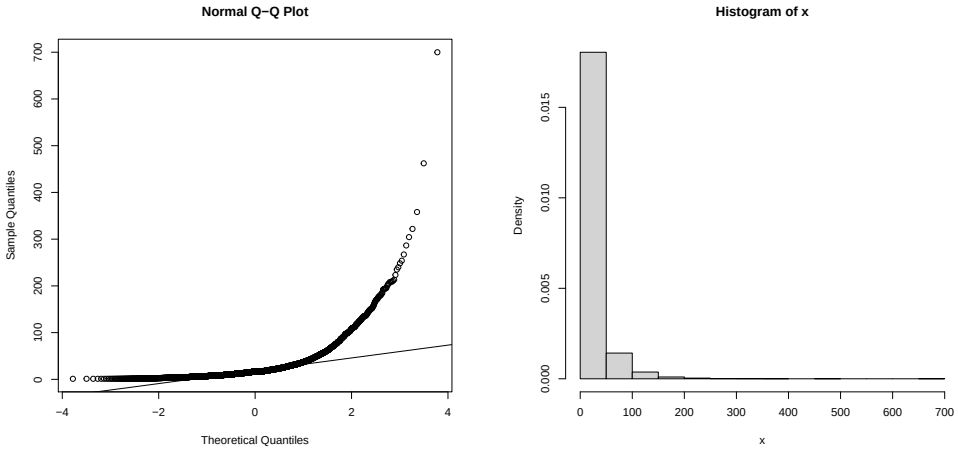
For fullstendige retningslinjer for innlevering av obligatoriske oppgaver, se her:

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

LYKKE TIL!

Oppgave 1

Fila `https://www.uio.no/studier/emner/matnat/math/STK1110/data/forsikringskrav.txt` inneholder 6377 bilforsikringskrav til et norsk forsikringsselskap et gitt år. Figur 1 viser et histogram og et kvantilplott av disse dataene. Vi ser at de åpenbart ikke er normalfordelt.



Figur 1: Kvantilplott (til venstre) og et histogram (til høyre) av forsikringskravene $x_1, \ldots, x_n$.

Siden dataene alltid er positive, er en mer rimelig modell Gamma fordelingen. Anta derfor alle observasjoner er uavhengige og identisk fordelte med sannsynlighetstetthet

$$f(x; \alpha, \gamma) = \frac{\gamma^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\gamma x}, \quad \alpha, \gamma > 0.$$

Merk at denne definisjonen er noe anderledes enn slik den er definert i boka (likning (4.6)) ved at den er representert ved rate parameteren γ istedet for skala parameteren $\beta = 1/\gamma$. For denne parametriseringen har man at

$$E[X] = \frac{\alpha}{\gamma}, \quad V[X] = \frac{\alpha}{\gamma^2}.$$

(a) Utled moment-estimatorene for α og γ. Gitt moment-estimatorene for α og β (gitt i boka), er disse resultatene rimelige?

Beregn momentestimatorene basert på forsikringsdataene.

(b) Skriv ned log-likelihood funksjonen og beregn log-likelihood verdien basert på dine estimater for α, γ.

(c) La oss nå se på maksimum likelihood estimatorene for α, γ. Her eksisterer det ikke noe analytisk uttrykk. Vis imidlertid at for gitt α, så

er det et analytisk uttrykk for maksimum likelihood estimatoren for γ (som en funksjon av α).

Utled dette uttrykket. Vis at dette svarer til moment-estimatoren for γ (for gitt α).

(d) For å finne maksimum likelihood estimate for α må en imidlertid bruke numerisk optimering. Nedenfor er en funksjon som beregner (negativ) log-likelihood verdi for en gitt verdi av α der maksimum likelihood estimatet for γ er satt inn.

```
negloglikgamma = function(logalpha,x=x)
{
 n = length(x)
 alpha=exp(logalpha)
 gamma = alpha/mean(x)
 logL = n*alpha*log(gamma)-n*lgamma(alpha)+
         (alpha-1)*sum(log(x))-gamma*sum(x)
 -logL
}
```

Selve optimeringen kan gjøres ved kommandoen

```
fit.ml=optim(log(alpha0),negloglikgamma,x=x,method="BFGS")
```

der `alpha0` er en initial verdi for α (f.eks moment-estimatoren). Opsjonen `method="BFGS"` angir hvilken optimeringsmetode som brukes (se `help(optim)`).

Forklar hvorfor en bør ha $\log(\alpha)$ som input til funksjonen og ikke α selv.

Bruk denne funksjonen til å finne maksimum likelihood estimater for (α, γ). Beregn log-likelihood verdien for dette settet av estimater. Er denne verdien rimelig i forhold til den verdi du fikk i (b)?

(e) Basert på ikke-parametrisk bootstrapping, finn anslag på standard feil for maksimum likelihood estimatene for α og γ.

Bruk dette til å lage 95% konfidens-intervaller for α og γ.

Hint: Kommandoen `quantile(z,c(0.025,0.975))` gir deg nedre og øvre empiriske 2.5% kvantiler i datavektoren `z`.

(f) Anta nå vi er mer interessert i $E[X_i] = \mu = \alpha/\gamma$. Bruk dine bootstrap simuleringer til å lage et 95% konfidensintervall for μ.

Anta vi ønsker å teste hypotesene

$$H_0 : \mu = 25 \text{ mot } H_a : \mu \neq 25$$

med et signifikansnivå $\alpha = 0.05$ Hva blir konklusjonen på en slik test?

Hvis du endrer signifikansnivået til $\alpha = 0.01$, hva blir så konklusjonen?

Basert på dette, hva kan du si om P-verdien for testen?

Oppgave 2

En av de viktigste ingrediensene i såkalt 'supermat' er blåbær. At blåbærene regnes som supermat, skyldes at fargestoffet antocyan, som gjør bærene blå, er en kraftig antioksidant.

I forbindelse med en studie av antioksidanter og antocyaner, ble innholdet av antocyan i 15 beger med blåbær målt. De målte verdiene var (i mg per 100 gram bær):

525 587 547 558 591 531 571 551 566 622 561 502 556 565 562

Vi antar at målingene kan betraktes som realisasjoner av uavhengige normalfordelte variabler med forventning μ og varians σ^2.

(a) Lag et 95% konfidensintervall for forventet antocyaninnhold μ basert på målingene over.

(b) På Wikipedia kan vi lese at forventet antocyaninnhold i blåbær er 558 mg/100g. Nå skal du bruke simuleringer til å late som om du måler antocyan i 15 prøver med blåbær veldig mange ganger. Generér 10 000 datasett, hvert av størrelse $n = 15$, bestående av observasjoner av de stokastiske variablene $X_1, \ldots, X_{15} \overset{uif}{\sim} N(\mu, \sigma)$ der $\mu = 558$ og $\sigma = 30$. Du kan bruke `rnorm()`-funksjonen i R til dette. Selv om du har simulert fra en fordeling med kjent forventning og varians, skal du late som om begge disse er ukjent i det følgende. Lag et 95% konfidensintervall for μ som i punkt a), basert på hvert av de simulerte datasettene, slik at du får 10 000 intervaller. Tell opp andelen av disse intervallene som inneholder verdien 558. Kommentér og forklar.

(c) Bruk nå i stedet det tilnærmede intervallet for store utvalg, altså

$$\left(\bar{X} - 1.96\frac{S}{\sqrt{15}}, \bar{X} + 1.96\frac{S}{\sqrt{15}}\right)$$

med

$$\bar{X} = \frac{1}{15}\sum_{i=1}^{15} X_i \text{ og } S^2 = \frac{1}{15-1}\sum_{i=1}^{15}\left(X_i - \bar{X}\right)^2,$$

og beregn dette for hvert av 10 000 datasett, generert som i b). Hvor stor andel av intervallene inneholder $\mu = 558$? Kommentér og forklar resultatet.

(d) Trekk 10 000 datasett som i (b) og lag et 95% konfidensintervall for σ for hvert av dem. Hvor stor andel av intervallene inneholder $\sigma = 30$?

(e) Under antakelsen om normalfordeling er $Z_i = \frac{X_i - \mu}{\sigma} \sim N(0,1)$, $i = 1, \ldots, n$, med $\mu = 558$ og $\sigma = 30$. Anta nå at $Z_1, \ldots, Z_{15}$ i virkeligheten er t-fordelt med 7 frihetsgrader, altså $Z_1, \ldots, Z_{15} \overset{uif}{\sim} t_7$. Trekk nå 10 000 datasett fra denne fordelingen ved å

1. trekke $z_1, \ldots, z_{15}$ fra t_7 med R-funksjonen `rt()`
2. la $x_i = \mu + \sigma z_i$, $i = 1, \ldots, n$.

Gjenta deretter oppgave b) med de nye datasettene. Hvor robust er metoden for å lage konfidensintervall for forventningsverdien for antakelsen om normalfordeling?

(f) Trekk datasett som i (e) og lag deretter 95%konfidenintervall for standardavviket til X_i slik som i (d). Merk imidlertid at $\mathrm{V}(Z_i) = \frac{7}{7-2}$ slik at variansen til X_i nå er $\tilde{\sigma}^2 = \mathrm{V}(X_i) = \mathrm{V}(\mu + \sigma Z_i) = \sigma^2 \mathrm{V}(Z_i) = 1.4\sigma^2$. Det er altså $\tilde{\sigma}$ du skal lage konfidensintervall for, og sjekke andelen intervaller som inneholder $\tilde{\sigma}$. Sammenlign med resultatene fra (d) og kommentér.