

# Obligatorisk øving for STK1110, Høsten 2022

## Øving 1 av 2

### Innleveringsfrist

Torsdag 6. oktober 2022, klokken 14:30 i Canvas ([canvas.uio.no](https://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 obliqnummer.

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, men den innleverte besvarelsen skal være skrevet av deg 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](mailto: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](https://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. Øvre rad av Figur 1 nedenfor viser et histogram og et kvantilplott av disse dataene. Vi ser at de åpenbart ikke er normalfordelt. Raden under viser samme plott av de log-transformerte dataene. For disse ser normalfordelingsantakelsen ut til å være grei.

Anta nå at  $Y_1, \dots, Y_n \stackrel{uif}{\sim} N(\mu, \sigma^2)$ , og at  $X_i = e^{Y_i}$ ,  $i = 1, \dots, n$ . Da er  $X_1, \dots, X_n$  uavhengige, identisk log-normalfordelte variabler med parametere  $(\mu, \sigma^2)$ ,  $\log - N(\mu, \sigma^2)$ . Vi antar heretter at forsikringskravene er observasjoner fra en log-normal fordeling.

- a) Finn maksimum likelihood-estimatorene  $\hat{\mu}_{mle}$  og  $\hat{\sigma}_{mle}^2$  for  $\mu$  og  $\sigma^2$  på vanlig måte, ved å
  1. spesifisere likelihood-funksjonen
  2. ta logaritmen til denne og derivere den
  3. sette den deriverte lik 0 og løse med hensyn på  $\mu$  og  $\sigma^2$ . Beregn de tilsvarende estimatene for bilforsikringskravene.
- b) Husk at  $Y_i = \log(X_i) \sim N(\mu, \sigma^2)$ ,  $i = 1, \dots, n$ . Bruk dette, samt maksimum likelihood-estimatorene for parametere i normalfordelingen (disse behøver du ikke å utlede) til å finne  $\hat{\mu}_{mle}$  og  $\hat{\sigma}_{mle}^2$  på en alternativ (og mye enklere) måte.
- c) Vis at informasjonsmatrisa for én observasjon er gitt ved

$$I(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

og bruk denne til å finne et estimat for standardfeilen til  $\hat{\mu}_{mle}$  og  $\hat{\sigma}_{mle}^2$ .

- d) Estimér nå standardfeilen til  $\hat{\mu}_{mle}$  og  $\hat{\sigma}_{mle}^2$  for forsikringsdataene ved hjelp av ikke-parametrisk bootstrapping. Sammenlign med estimatene du fikk i c) og kommentér.
- e) Lag et 95% konfidensintervall for  $E(X_i)$ . Her kan du benytte deg av at  $n$  er stor.
- f) Vi ønsker nå et 95% konfidensintervall for  $V(X_i)$ . Lag et konfidensintervall for  $V(X_i)$  basert på antakelsen om at  $X_i$  er normalfordelt (selv om vi vet at det ikke stemmer).
- g) Bruk nå i stedet persentilintervallet basert på ikke-parametrisk bootstrapping til å lage et 95% konfidensintervall for  $V(X_i)$ . Sammenlign med resultatene fra f) og kommentér.

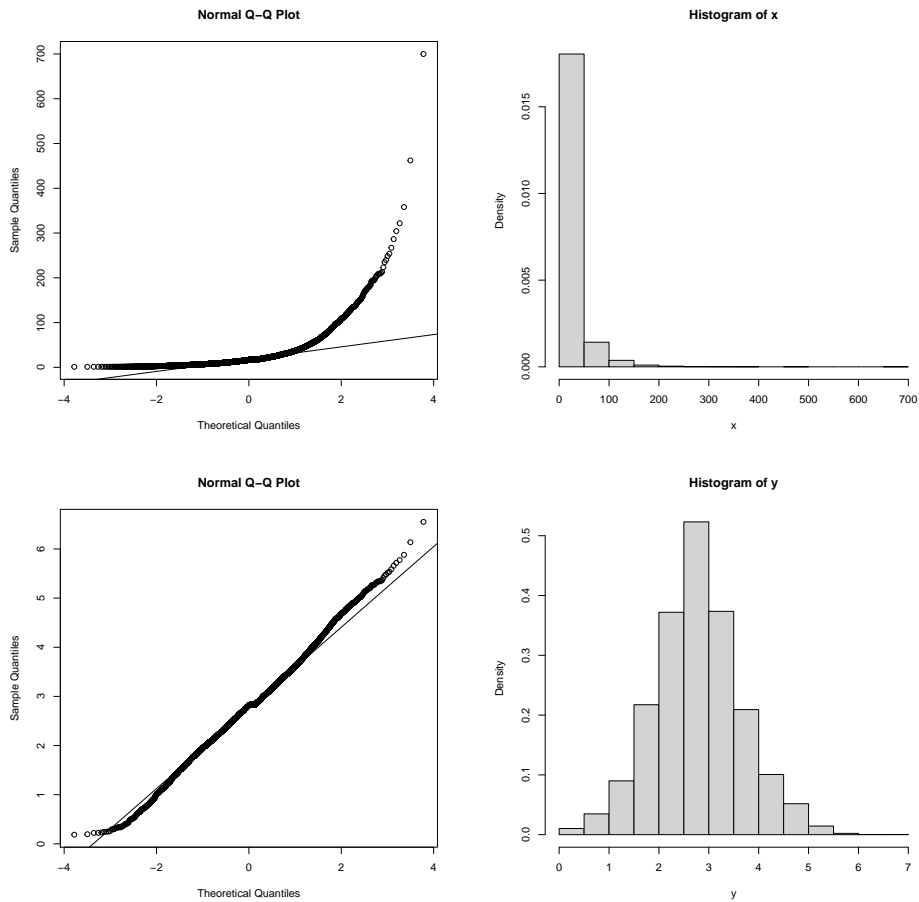


Figure 1: Øvre rad viser et kvantilplott (til venstre) og et histogram (til høyre) av forsikringskravene  $x_1, \dots, x_n$ . Nedre rad viser et kvantilplott (til venstre) og et histogram (til høyre) av de loggede forsikringskravene  $y_1, \dots, y_n$ , med  $y_i = \log(x_i)$ .

## 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  $\mu$  og varians  $\sigma^2$ .

- Lag et 95% konfidensintervall for forventet antocyaninnhold  $\mu$  basert på målingene over.
- 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 10000 datasett, hvert av størrelse  $n = 15$ , bestående av observasjoner er av de stokastiske variablene  $X_1, \dots, X_{15} \stackrel{uif}{\sim} N(558, 30^2)$ . 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  $\mu$  som i punkt a), basert på hvert av de simulerte datasettene, slik at du får 10000 intervaller. Tell opp andelen av disse intervallene som inneholder verdien 558. Kommentér og forklar.
- 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

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

og beregn dette for hvert av 10000 datasett, generert som i b). Hvor stor andel av intervallene inneholder  $\mu = 558$ ? Kommentér og forklar resultatet.

- Trekk 10000 datasett som i b) og lag et 95% konfidensintervall for  $\sigma$  for hvert av dem. Hvor stor andel av intervallene inneholder  $\sigma = 30$ ?
- Under antakelsen om normalfordeling er  $Z_i = \frac{X_i - \mu}{\sigma} \sim N(0, 1)$ ,  $i = 1, \dots, n$ , med  $\mu = 558$  og  $\sigma = 30$ . Anta nå at  $Z_1, \dots, Z_{15}$  i virkeligheten er t-fordelt med 7 frihetsgrader, altså  $Z_1, \dots, Z_{15} \stackrel{uif}{\sim} t_7$ . Trekk nå 10000 datasett fra denne fordelingen ved å
  - trekke  $z_1, \dots, z_{15}$  fra  $t_7$  med R-funksjonen `rt()`

2. la  $x_i = \mu + \sigma z_i$ ,  $i = 1, \dots, n$ .

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  $V(Z_i) = \frac{7}{7-2}$  slik at variansen til  $X_i$  nå er  $\tilde{\sigma}^2 = V(X_i) = V(\mu + \sigma Z_i) = \sigma^2 V(Z_i) = 1.4\sigma^2$ . 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.