21. april 2022

STK1100

Obligatorisk oppgavesett 2 av 2.

Innleveringsfrist

Torsdag 5. mai 2022, 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). NB! 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, med forklaringer. Besvarelser som kun består av kode, godkjennes ikke. Besvarelser som ikke tar for seg oppgavene som krever programmering blir heller ikke godkjent. 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) i god tid før innleveringsfristen.

For å få adgang til avsluttende eksamen i STK1100, må man bestå begge de obligatoriske oppgavesettene i ett og samme semester.

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

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

Spesielt om det obligatoriske oppgavesettet i STK1100

Det anbefales på det sterkeste at du bruker Python til å gjøre beregningene i oppgave 2 og 3. Hvis du bruker et annet programmeringsspråk, kan vi ikke hjelpe deg hvis du får problemer. Uansett hvilket programmeringsspråk du bruker, må du angi hvilke kommandoer du har brukt for å komme fram til svarene dine. Hvis du trenger hjelp til å løse oppgavene, kan du få det på en av de åpne gruppene i STK1100.

LYKKE TIL!

Oppgave 1. De stokastiske variablene X og Y har simultan sannsynlighetstetthet

$$f(x,y) = \begin{cases} k(x+2y) & \text{for } 0 \le x \le 1,\ 0 \le y \le 1,\ x+y \le 1 \\ 0 & \text{ellers,} \end{cases}$$

der k er en konstant.

a) Vis at $k = 2$.

b) Vis at den marginale sannsynlighetstettheten til Y er

$$f_Y(y) = \begin{cases} 1 + 2y - 3y^2 & \text{for } 0 \le y \le 1 \\ 0 & \text{ellers} \end{cases}$$

c) Bestem den betingede sannsynlighetstettheten til $X|Y = y$.

d) Er X og Y uavhengige? Svaret skal begrunnes!

Oppgave 2. I denne oppgåva skal du studere levetider for kvinner og menn i det gamle Egypt. Dataane finn du i fila `egypt.txt` under fana Data på heimesida til kurset, og består av levetida til 82 menn og 59 kvinner for omtrent 2000 år sida. Datasettet stammar frå artikkelen *On the change in expectation of life in ein during a period of circa 2000 years*, skrive av Karl Pearson, og utgitt i Biometrika i 1902, vol. 1, s. 261–264.

PYTHON hjelp: Du kan lesa inn dataane og dele inn i menn og kvinner ved kommandoane:

```
import pandas as pd
url = "https://www.uio.no/studier/emner/matnat/math/STK1100/data/egypt_data.txt"
levetidfelles=pd.read_csv(url, header=None)
t_menn=levetidfelles[0:82]
t_kvinner=levetidfelles[82:141]
```

For eit utval $t_1, t_2, \ldots, t_n$, er den *empiriske kumulative fordelingsfunksjonen* gitt ved

$$\widehat{F(t)} = \frac{1}{n} \sum_{i=1}^{n} I(t_i \le t),$$

altså at $\widehat{F(t)}$ er gitt ved *andelen* av alle observasjonane i utvalet som er mindre enn eller lik t.

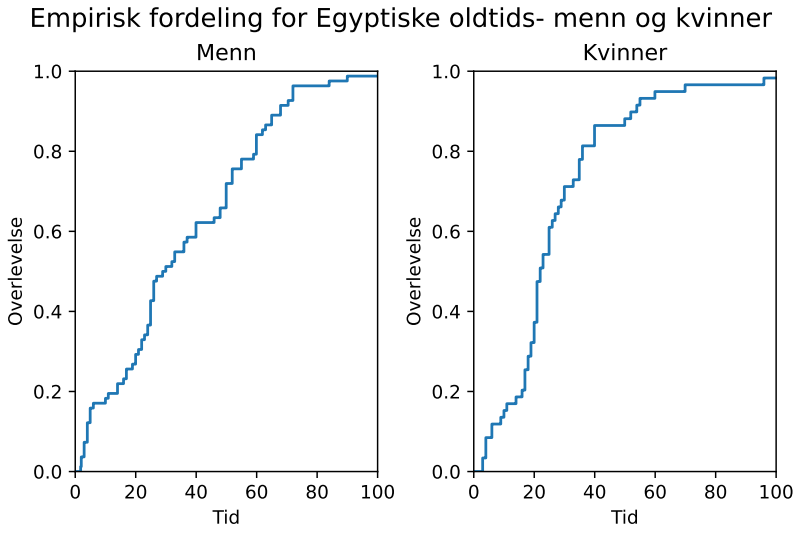
a) Plott den empiriske kumulative fordelingsfunksjonen for kvinnene, og for mennene i datasettet, kvar for seg.

PYTHON hjelp: Du kan plotte den kumulative fordelingen til t ved kommandoane:

```
import numpy as np
from scipy import interpolate as inter

def lag_empirisk_fordelingsfunksjon(x):
    empirisk_fordeling = inter.interp1d(np.sort(x),
                          np.arange(len(x))/float(len(x)),
                          kind = "zero",
                          fill_value = "extrapolate")
    return empirisk_fordeling

fordeling_kjonn = lag_empirisk_fordelingsfunksjon(t_kjonn)
z = np.linspace(0, 110, 1000)
plt.step(z, fordeling_kjonn(z))}
```



Figur 1: Empirisk kumulativ fordelingsfunksjon for menn og kvinner i oldtids-egypt.

Ein mogleg parametrisk modell for levetider er gammafordelinga, altså ein modell med tettleik.

$$f(t) = \frac{t^{\alpha-1}e^{-\frac{t}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}$$

I gammafordelinga er $\mathbb{E}(T) = \alpha\beta$, og $\text{Var}(T) = \alpha\beta^2$.

b) Finn estimatorar for α og β ved å løyse ligningsettetet under for α og β

$$\bar{t} = \alpha\beta,$$

$$s^2 = \alpha\beta^2,$$

der $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$, og $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(t_i - \bar{t})^2$ (desse kallast forresten momentestimatorane). Finn dei tilsvarande estimata for α og β ved å sette inn dei egyptiske overlevelsesdataane. Du skal her finne eitt sett med estimat for kvinner og eitt for menn. Ser levetidsfordelinga for menn og kvinner ut til å være forskjellige?

Ein annan mogleg parametrisk modell for levetider er ei såkalla *log-normal* fordeling. Denne fordelinga kan ein konstruere som $T = e^X$, der $X \sim N(\mu, \sigma)$.

c) Finn tettleiken til den log-normale fordelinga, ved å regne ut transformasjonen.

For den log-normale fordelinga har ein at

$$\mathbb{E}(T) = \exp\left(\mu + \frac{\sigma^2}{2}\right),$$

og at

$$\text{Var}(T) = \left[\exp(\sigma^2) - 1\right]\exp\left(2\mu + \sigma^2\right).$$

d) Finn momentestimatorane for μ og σ ved å løyse ligningsettetet under for μ og σ

$$\bar{t} = \exp\left(\mu + \frac{\sigma^2}{2}\right),$$

og

$$S^2 = \left[\exp(\sigma^2) - 1\right] \exp\left(2\mu + \sigma^2\right).$$

Finn dei tilsvarande estimata for μ og σ ved å sette inn dei egyptiske overlevelsesdataane slik som i punkt b). Kva kan du nå seie om forskjellen mellom kvinner og menn?

e) Lag to figurar, der du i figurane plottar den empiriske overlevelseskurven mot log normal- og gamma-overlevelseskurvene for *eitt av kjønna* kvar gong. Kva fordeling synest du passar best til levetidene til høvesvis kvinner og menn? Grunngi svaret ditt.

Oppgave 3. I denne oppgaven skal vi se hvordan du kan bruke en datamaskin til å generere observasjoner fra Lomax-fordelingen, som er oppgitt under. Vi tar som utgangspunkt at datamaskiner kan generere tilfeldige tall i intervallet $[0, 1]$, dvs. observasjoner av en stokastisk variabel U som er uniformt fordelt på $[0, 1]$.

a) La $F(x)$ være en kumulativ fordeling, og anta at $U \sim \text{uniform}(0, 1)$. Vis at da har $X = F^{-1}(U)$ kumulativ fordeling $F(x)$.

La X være tida fram til en bedrift går konkurs. Da er det er vanlig å anta at X er Lomax-fordelt, det vil si at X har sannsynlighetstettheten

$$f_X(x) = \begin{cases} \frac{\alpha}{\lambda}\left(1 + \frac{x}{\lambda}\right)^{-(\alpha+1)} & \text{for} \quad x > 0, \\ 0 & \text{ellers.} \end{cases}$$

Her er $\alpha, \lambda > 0$ parametere i fordelingen.

b) Vis at den kumulative sannsynlighetsfordelingen til X er gitt ved

$$F_X(x) = \begin{cases} 1 - \left(1 + \frac{x}{\lambda}\right)^{-\alpha} & \text{for} \quad x > 0, \\ 0 & \text{ellers.} \end{cases}$$

Bestem median tid fram til en tilfeldig bedrift går konkurs.

c) Bruk resultatene i punkt a) og b) til å angi en framgangsmåte for å generere observasjoner fra Lomax-fordelingen.

d) Bruk framgangsmåten i forrige punkt til å generere 10 000 observasjoner fra Lomax-fordelingen med $\lambda = 48$ måneder og $\alpha = 3$. Beregn medianen av de genererte observasjonene, og sammenlign med resultatene i punkt b).

PYTHON hjelp: Du kan generere n observasjoner av $U \sim \text{uniform}(0, 1)$ ved kommandoene

```
from numpy import *
u=random.uniform(0,1,n)
```

e) Lag et normert histogram av de Lomax-fordelte observasjonene fra punkt d) (dvs. et histogram der arealet av alle stolpene til sammen er lik én.)

PYTHON hjelp: Du kan tegne et normert histogram av observasjoner i en vektor `x` ved kommandoene

```
import matplotlib.pyplot as plt
plt.hist(x,density=True,edgecolor="black")
```

Siden noen av observasjonene fra Lomax-fordelingen vil være veldig store, må du avgrense de `x`-verdiene du plotter histogrammet for. For eksempel får du tegnet histogrammet over intervallet fra 0 til 360 måneder ved kommandoene

```
plt.xlim(0,360)
plt.hist(x,density=True,edgecolor="black",nbins=50).
```

Noen ganger kan du få så mange store verdier at du må jobbe litt med antall bins i histogrammet for å få det pent.

f) Tegn tettheten til Lomax-fordelingen med $\lambda = 48$ måneder og $\alpha = 3$ i samme figur som histogrammet og kommentér resultatet.

g) Lomax-fordelingen er en fordeling med tung hale. Du skal se på hvordan dette slår ut for sentralgrenseteoremet i praksis. Illustrér vha. simuleringer hvordan den empiriske fordelingen til gjennomsnittet av n Lomax-fordelte observasjoner nærmer seg normalfordelingen for $n = 10, 100$ og 1000. Dette gjør du ved å simulere 10 000 utvalg av størrelse n, ta gjennomsnittet i hvert utvalg, og se på hvordan de 10 000 gjennomsnittene fordeler seg, enten ved å lage histogrammer, eller ved å se på den empiriske kumulative fordelingen, og sammenligne med standard normalfordeling. La deg inspirere av forelesningen fra kapittel 6.

h) Gjenta punkt g), men denne gangen med den uniforme fordelingen på $(0, 1)$. Du skal med andre ord gjøre akkurat det samme som i forrige punkt, bortsett fra at du genererer og tar gjennomsnittet av uniformt fordelte observasjoner i stedet for Lomax-fordelte. Sammenligne den empiriske fordelingen for gjennomsnittet av de uniformt fordelte observasjonene med den for de Lomax-fordelte i punkt g). Kommentér og forklar den forskjellen du ser i resultatene for Lomax-fordelingen og den uniforme fordelingen.