Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

TMA4245 Statistikk
Vår 2017

Anbefalt øving 3
Løsningsskisse

Oppgave 1

a) Den tilfeldige variabelen X er kontinuerlig fordelt med sannsynlighetstetthet

$$f_X(x) = \begin{cases} nx^{n-1} & \text{dersom } 0 < x \leq 1 \\ 0 & \text{ellers} \end{cases}.$$

Den kumulative fordelingsfunksjonen F er lik det bestemte integralet av sannsynlighetstettheten f fra $-\infty$ til x. Siden $f(x) = 0$ for $x \leq 0$ setter vi 0 som nedre grense for integralet,

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{0}^{x} nt^{n-1}dt = [t^n]_0^x = \underline{\underline{x^n}}.$$

Fordelingsfunksjonen $F(x)$ gir sannsynligheten for at $X \leq x$, og kan brukes til å finne sannsynligheten for at $1/4 < X \leq 3/4$,

$$P\left(\frac{1}{4} < X \leq \frac{3}{4}\right) = P\left(X \leq \frac{3}{4}\right) - P\left(X \leq \frac{1}{4}\right) = F\left(\frac{3}{4}\right) - F\left(\frac{1}{4}\right) = \left(\frac{3}{4}\right)^n - \left(\frac{1}{4}\right)^n.$$

Setter vi inn $n = 1$ får vi

$$P\left(\frac{1}{4} < X \leq \frac{3}{4}; n = 1\right) = \frac{3}{4} - \frac{1}{4} = \underline{\underline{\frac{1}{2}}},$$

og med $n = 2$ får vi

$$P\left(\frac{1}{4} < X \leq \frac{3}{4}; n = 2\right) = \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = \frac{9}{16} - \frac{1}{16} = \underline{\underline{\frac{1}{2}}}.$$

For å bestemme medianen til X setter vi $P(X \leq a) = F(a) = 1/2$, slik at

$$F(a) = a^n = \frac{1}{2}.$$

Medianen er altså

$$a = \begin{cases} 1/2 & \text{dersom } n = 1 \\ \sqrt{2}/2 & \text{dersom } n = 2 \end{cases}.$$

Forventningsverdien til X er lik integralet av $xf(x)$ fra $-\infty$ til ∞,

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 xnx^{n-1}dx$$
$$= \int_0^1 nx^n dx = \frac{n}{n+1} = \begin{cases} 1/2 & \text{dersom } n = 1 \\ 2/3 & \text{dersom } n = 2 \end{cases}.$$

For $n = 1$ er medianen og forventningsverdien identiske, mens for $n = 2$ er medianen størst ($\sqrt{2}/2 \approx 0.71 > 2/3$).

b) Histogrammet $f_X(x)$ kan plottes på følgende måte:

```
data30=load('data30.txt');
figure
histogram(data30,'Normalization','pdf')
```

Vi bruker her innstillingene `'Normalization'` og `'pdf'` for å få et normalisert histogram slik at vi kan sammenligne historgrammet med sannsynlighetstettheten $f_X(x)$. Sannsynlighetstettheten $f_X(x)$ kan for eksempel plottes slik:

```
hold on
n=2;
x=linspace(0,1,300);
f=n*x.^(n-1);
plot(x,f)
```

Resultatet er vist i Figur 1. Vi ser at sannsynlighetstettheten stemmer godt overens med histogrammet, selv om vi har få observasjoner som har verdi lavere enn 0.5.

Den empiriske kumulative fordelingsfunksjonen og fordelingsfunksjonen kan plottes på følgende måte:

```
figure
ecdf(data30)
hold on
F=x.^n
plot(x,F)
```

Resultatet er vist i Figur 2. Den empiriske fordelingen $\hat{F}(x)$ har samme fasong som den teoretiske fordelingen $F(x)$, men den empiriske fordelingen tar noe lavere verdier. Andel av observasjonene som er mellom $\frac{1}{4}$ og $\frac{3}{4}$ er:

```
mean(data30>(1/4) & data30<(3/4))
ans =   0.467
```

Kommandoen data30>(1/4) returner en vektor av lengde 30 med verdi 1 for alle observasjoner i data30 som er større enn $\frac{1}{4}$ og 0 ellers. Tilsvarende gir data30<(3/4) 1 for alle observasjoner mindre enn $\frac{3}{4}$. Merk at & er og operatoren i Matlab, dvs. den sammenligner de to vektorene og returner en ny vektor av lengde 30 der hvert element har verdi

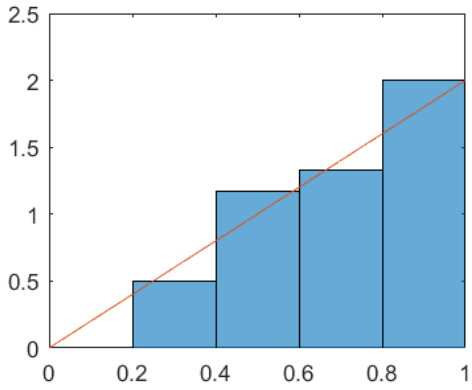
1 dersom både data30>(1/4) og data30<(3/4) har verdi 1. Den søkte sannsynligheten finnes ved å ta gjennomsnittet. Medianen til observasjonene er:

```
median(data30)
ans=0.775
```

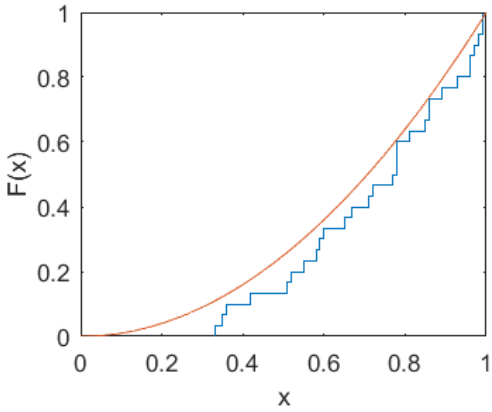
Forventningsverdien til observasjonene er:

```
mean(data30)
ans=0.724
```

Sammenligner vi dette med de teoretiske resultatene i **a)** ser vi at estimatene våre er i nærheten av de teoretiske verdiene.



Figur 1: Histogram over de 30 observasjonene og sannsynlighetstettheten $f_X(x)$.



Figur 2: Empirisk kumulativ fordelingsfunksjon $\hat{F}(x)$ og fordelingsfunksjonen $F(x)$ basert på 30 observasjoner

c) Vi studerer nå det store datasettet og gjentar deloppgave **b)**. Vi plotter historgrammet, sannsynlighetstettheten, den empiriske og den teoretiske kumulative fordelingsfunksjonen som tidligere:

```
data300=load('data300.txt');

%Histogram og sannsynlighetstetthet:
figure
histogram(data300,'Normalization','pdf')
hold on
n=2;
x=linspace(0,1,300);
f=n*x.^(n-1)
plot(x,f)

%Kumulativ fordeling:
figure
ecdf(data300)
hold on
F=x.^n
plot(x,F)
```

De resulterende plottene er gitt i Figur 3 og 4. Vi ser at de teoretiske fordelingene stemmer godt overens med observasjonene. Merk at siden vi har flere observasjoner vil Matlab øke antallet søyler/bins.

Andel av observasjonene som er mellom $\frac{1}{4}$ og $\frac{3}{4}$ er nå:

```
mean(data300>(1/4) & data300<(3/4))
ans =0.497
```

Medianen til observasjonene er:

```
median(data300)
ans=0.705
```

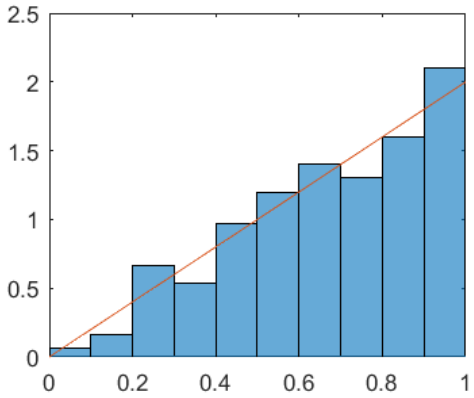
Forventningsverdien til observasjonene er:

```
mean(data300)
ans=0.671
```

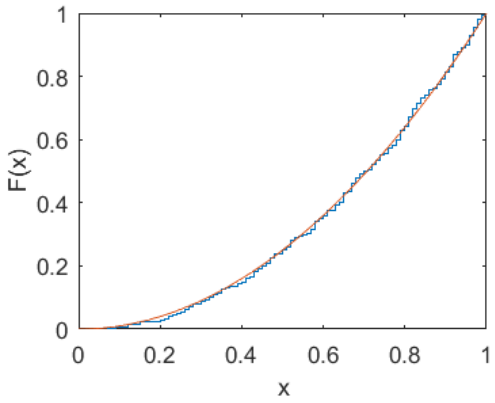
Vi ser at vi i dette tilfellet oppnår estimater nærmere den teoretiske verdien enn i oppgave **b)**. Estimatene er altså nærmere de teoretiske, utregnede verdiene fra oppgave **a)**. Dersom man har et sett med observasjoner som er trukket fra en bestemt fordelingsfunksjon $f_X(x)$, kan man forvente at medianen og gjennomsnittet til observasjonene vil komme nærmere og nærmere de teoretiske verdiene når størrelsen på datasettet øker. I grensen, dvs. med uendelig mange observasjoner kan en forvente å få eksakt den teoretisk verdien.

Oppgave 2 Togforsinkelsen — Eksamen desember 2003, oppgave 1 av 3

Oppgitt: $\int_0^\infty x^r e^{-ax} \mathrm{d}x = \frac{r!}{a^{r+1}}$ for $a > 0$, $r \geq 0$ heltall.



Figur 3: Histogram over de 300 observasjonene og sannsynlighetstettheten $f_X(x)$.



Figur 4: Empirisk kumulativ fordelingsfunksjon $\hat{F}(x)$ og fordelingsfunksjonen $F(x)$ basert på 300 observasjoner.

a) For at $g(x)$ skal være en sannsynlighetstetthet, må vi ha $\int_{-\infty}^{\infty} g(x)\mathrm{d}x = 1$, dvs at total sannsynlighet er 1. (Bruker formelen med $r = 1$ og $a = 2$.)

$$\int_{-\infty}^{\infty} g(x)\mathrm{d}x = \int_{0}^{\infty} kxe^{-2x}\mathrm{d}x = k \cdot \frac{1}{2^2} = \frac{k}{4}.$$

$k/4 = 1$ gir $k = 4$.

For forventet forsinkelse brukes formelen igjen, med $r = 2$ og $a = 2$:

$$\begin{aligned}\mathrm{E}(X) &= \int_{-\infty}^{\infty} xg(x)\mathrm{d}x = \int_{0}^{\infty} x \cdot 4xe^{-2x}\mathrm{d}x \\ &= 4\int_{0}^{\infty} x^2e^{-2x}\mathrm{d}x = 4 \cdot \frac{2}{2^3} = 1.\end{aligned}$$

For å vise at det er 0.09 i sannsynlighet for mer enn to minutters forsinkelse, bruker vi delvis integrasjon.

$$\begin{aligned}\text{Prob}(X > 2) &= \int_2^\infty 4xe^{-2x}\text{dx} \\ &= 4 \cdot \left[-\frac{1}{2}xe^{-2x}\right]_{x=2}^\infty + 4 \cdot \int_2^\infty \frac{1}{2}e^{-2x}\text{d}x \\ &= 4 \cdot \frac{1}{2} \cdot 2e^{-4} + \int_2^\infty 2e^{-2x}\text{d}x \\ &= 4e^{-4} + e^{-4} = 5e^{-4} \approx 0.09.\end{aligned}$$

b) V er binomisk fordelt med $n = 22$ og $p = 0.09$ under forutsetning av at

- Hendelsene "mer enn 2 minutter forsinket" for to forskjellige dager er uavhengige.
- Sannsynligheten for "mer enn 2 minutter forsinket" er lik 0.09 hver dag.

Antall forsøk (dager) er bestemt på forhånd, det er to utfall og vi teller antall "suksesser". (Togselskapet ville neppe kalle en dag med mer enn to minutters forsinkelse for en suksess.)

$$\begin{aligned}\text{Prob}(V \geq 2) &= 1 - \text{Prob}(V \leq 1) = 1 - \text{Prob}(V = 0) - \text{Prob}(V = 1) \\ &= 1 - 0.91^{22} - 22 \cdot 0.91^{21} \cdot 0.09^1 = 0.6012.\end{aligned}$$

Ser vi på et år med $n = 220$ virkedager, er $V \sim$ binomisk$(220, 0.09)$. Da kan vi bruke tilnærmingen til normalfordelingen, dvs

$$\begin{aligned}\text{Prob}(V > 30) &= 1 - \text{Prob}(V \leq 30) \approx 1 - \Phi(\frac{30 + 1/2 - 220 \cdot 0.09}{\sqrt{220 \cdot 0.09 \cdot (1 - 0.09)}} \\ &= 1 - \Phi(2.52) = 0.0059.\end{aligned}$$

c) Setter $x = 2$ i den betingede fordelingen. Da har oppholdstiden Y fordeling $f(y|2) = e^{-y}$ for $y > 0$. Med andre ord er $Y|X = 2$ eksponensialfordelt med parameter (og forventningsverdi) 1.

Simultantetthet finner vi ved å multiplisere;

$$f(x, y) = f(y|x)g(x) = \frac{x}{2}e^{-\frac{xy}{2}} \cdot 4xe^{-2x} = 2x^2e^{-x(2+\frac{y}{2})} \qquad \text{for } x > 0, y > 0.$$

Marginaltettheten for Y finnes ved å integrere ut x:

$$\begin{aligned}h(y) &= \int_0^\infty f(x, y)\text{d}x = \int_0^\infty 2x^2e^{-x(2+\frac{y}{2})}\text{d}x \\ &= 2 \cdot \frac{2}{(2 + \frac{y}{2})^3} = \frac{32}{(4 + y)^3}, \qquad \text{for } y > 0.\end{aligned}$$

Her brukte vi enda en gang formelen som var oppgitt, denne gangen med $a = 2 + y/2$ og $r = 2$.

Oppgave 3

Vi ønsker å finne sannsynligheten for at summen av X og Y er mindre enn 60 minutter, altså $P(X+Y \leq 60)$.

Vi vet at X og Y er uavhengige, slik at $f(x,y) = g(x)h(y)$. Dermed er simultanfordelingen gitt ved

$$f(x,y) = \begin{cases} \frac{1}{30} \cdot \frac{1}{13} & \text{for } 0 < x < 30, 39 < y < 52 \\ 0, & \text{ellers} \end{cases}$$

Vi observerer at hendelsen $X + Y \leq 60$ vil oppstå hvis X tar verdier $0 < x < 60 - y$, for enhver Y, slik at $39 < y < 52$. Dermed kan vi integrere simultantettheten over dette området for å regne ut sannsynligheten,

$$\begin{aligned} P(X+Y \leq 60) &= \int_{39}^{52} \int_{0}^{60-y} f(x,y) dx dy \\ &= \int_{39}^{52} \int_{0}^{60-y} \frac{1}{30} \cdot \frac{1}{13} dx dy \\ &= \int_{39}^{52} \left[\frac{1}{30} \cdot \frac{1}{13} x \right]_{0}^{60-y} dy \\ &= \int_{39}^{52} \frac{1}{30} \cdot \frac{1}{13} (60 - y) dy \\ &= \left[\frac{1}{30} \cdot \frac{1}{13} \left(60y - \frac{y^2}{2} \right) \right]_{39}^{52} \\ &= 0.4833 \end{aligned}$$