UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1100 — Sannsynlighetsregning og statistisk modellering.
Eksamensdag:	Tirsdag 9. juni 2015.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 3 sider.	
Vedlegg:	Ingen.
Tillatte hjelpemidler:	Godkjent kalkulator. Formelsamling for STK1100og STK1110.

Kontroller at oppgavesettet er komplett før
du begynner å besvare spørsmålene.

Oppgave 1

Vi minner om at hvis en stokastisk variabel X har sannsynlighetstetthet

$$f_X(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & \text{for } x > 0, \\ 0 & \text{ellers,} \end{cases}$$

så er X gamma-fordelt med formparameter α og skalaparameter β. Kort skriver vi $X \sim \text{Gamma}(\alpha, \beta)$.

Anta at $X \sim \text{Gamma}(\alpha, \beta)$.

a) Vis at den moment genererende funksjonen til X er

$$M_X(t) = E\left(e^{tX}\right) = \frac{1}{(1-\beta t)^\alpha} \quad \text{for } t < 1/\beta.$$

b) Bruk den momentgenererende funksjonen til å bestemme $E(X)$ og $V(X)$.

Anta så at X_1 og X_2 er uavhengige stokastiske variabler og at $X_1 \sim \text{Gamma}(\alpha_1, \beta)$ og $X_2 \sim \text{Gamma}(\alpha_2, \beta)$.

c) Bruk momentgenererende funksjoner til å vise at $X_1 + X_2 \sim \text{Gamma}(\alpha_1 + \alpha_2, \beta)$. Generaliser dette resultatet til n uavhengige gamma-fordelte stokastiske variabler som alle har samme skalaparameter.

(Fortsettes på side 2.)

Oppgave 2

Kreftpasienter som gjennomgår en behandling med stråling og kjemoterapi kan mange ganger bli symptomfrie uten at sykdommen er helbredet. De kan da seinere få tilbakefall slik at symptomene igjen kommer tilbake.

Vi vil anta at tiden T til tilbakefall for en tilfeldig valgt kreftpasient er eksponentialfordelt, slik at T har sannsynlighetstetthet

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t} & \text{for } t > 0, \\ 0 & \text{ellers.} \end{cases} \tag{1}$$

a) Bestem den kumulative fordelingsfunksjonen til T.

b) Vis at median tid til tilbakefall er gitt ved $m = \ln(2)/\lambda$ og forklar hva medianen gir uttrykk for.

La nå $T_1, T_2, \ldots, T_n$ være tid til tilbakefall for n kreftpasienter. Vi antar at T_i-ene er uavhengige og identisk fordelte med tettheten (1).

c) En estimator for median tid til tilbakefall er

$$\widehat{m} = \frac{\ln(2)}{n} \sum_{i=1}^{n} T_i.$$

Vis at $\widehat{m}$ er en forventningsrett estimator og bestem standardfeilen $\sigma_{\widehat{m}} = \sqrt{V(\widehat{m})}$. Er $\widehat{m}$ konsistent?

d) Vis at $2\lambda \sum_{i=1}^{n} T_i$ er gamma-fordelt med formparameter n og skalaparameter 2 (jf. oppgave 1).

e) Utled et $100(1-\alpha)\%$ konfidensinterval for medianen m. Intervallet skal uttrykkes ved hjelp av n, $\widehat{m}$, a og b. Her er a og b henholdsvis $100(\alpha/2)\%$ og $100(1-\alpha/2)\%$ persentilene for gamma-fordelingen med formparameter n og skalaparameter 2.

Nedenfor er det gitt tid fra behandling til tilbakefall (i uker) for 21 pasienter med akutt leukemi:

1	1	2	2	3	4	4	5	5	8	8
8	8	11	11	12	12	15	17	22	23	

(Tallene er fra 1960-tallet og er ikke representative for dagens kreftbehandling.)

f) Estimer median tid til tilbakefall for leukemipasientene. Bestem også et 95% konfidensintervall for medianen. Det opplyses at 2.5% og 97.5% persentilene i gamma-fordelingen med formparameter 21 og skalaparameter 2 er henholdsvis 26.0 og 61.8.

(Fortsettes på side 3.)

Anta til slutt at vi ikke antar at tilbakefallstidene for leukemipasientene er eksponentialfordelte, bare at de er observerte verdier av stokastiske variabler som er uavhengige og identisk fordelte med kumulative fordelingsfunksjon F.

g) Gi et estimat for median tid til tilbakefall og forklar hvordan du kan bruke ikke-parametrisk bootstrap til å bestemme standardfeilen til estimatet.

Oppgave 3

I mange sammenhenger ønsker en å anslå verdien til, eller predikere, utfallet til en framtidig (uobservert) stokastisk variabel X, for eksempel hva temperaturen blir i morgen eller global gjennomsnittstemperatur i 2050.

Vi skal gjøre det enkelt her og anta at X er en stokastisk variabel med forventning $E(X) = \mu$ og varians $V(X) = \sigma^2$.

a) Vis at $\theta = \mu$ er den verdien av θ som minimerer forventet kvadratisk avvik

$$\mathrm{MSE}_X(\theta) = E\{(\theta - X)^2\}.$$

Vi kan tolke dette som at forventningen μ er vårt beste gjett på verdien til X før den er observert (med hensyn på å minimere forventet kvadratisk avvik).

For mange naturlige fenomener er det avhengighet fra måling til måling. For eksempel vil temperaturen i morgen avhenge av temperaturen i dag. Vi skal her se på hvordan en observert måling av X påvirker vårt beste gjett på utfallet til en framtidig variabel Y når X og Y er positivt korrelert.

Vi får en enkel modell for en slik situasjon hvis vi lar $\mu = E(X) = 0$ og $\sigma^2 = V(X) = 1$ og antar at $Y = \rho X + \sqrt{1-\rho^2}\,Z$, for $0 < \rho < 1$, hvor Z er uavhengig av X, med $E(Z) = 0$ og $V(Z) = 1$.

b) Finn $E(Y)$ og $V(Y)$.

c) Vis at $\mathrm{Corr}(X, Y) = \rho$.

Hvis vi observerer $X = x_0$ og skal gjette på verdien til Y er det fornuftig, siden X og Y er korrelerte, å bruke informasjonen om X og korrelasjonen til å gi et bedre anslag på verdien til Y enn å bare bruke forventningen til Y.

d) Vis at $E(Y \mid X = x_0) = \rho x_0$ og bestem $V(Y \mid X = x_0)$.

e) Finn den verdien av θ som gjør at betinget forventet kvadratisk avvik

$$\mathrm{MSE}_{Y \mid X = x_0}(\theta) = E\{(\theta - Y)^2 \mid X = x_0\}$$

blir minst mulig. Hva er ditt beste gjett på verdien til Y hvis du har observert $X = x_0$? Diskuter fordelen med å bruke dette anslaget i stedet for forventningsverdien til Y.

SLUTT