TMA4245 Statistikk
Eksamen mai 2016

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Oppgave 1

Gustav og Margrethe er nyutdannede sivilingeniører fra NTNU og er nå på boligjakt i Trondheim. Begge ser etter en leilighet i en bestemt bydel.

Vi antar at prisen per kvadratmeter (kvadratprisen) for denne bydelen er normalfordelt. I punkt **a**) og **b**) antar vi at forventningsverdien er $\mu = 30$ kkr (30.000 kr) og standardavviket er $\sigma = 2.5$ kkr.

a) Hva er sannsynlighetene for at kvadratprisen for en tilfeldig leilighet er:

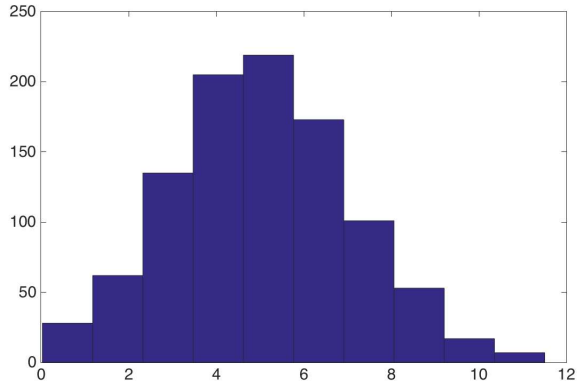
- lavere enn 30 kkr?
- høyere enn 25 kkr?
- høyere enn 25 kkr gitt at kvadratprisen er lavere enn 30 kkr.

b) Gustav vurderer en leilighet på 40 kvadratmeter, og Margrethe vurderer en leilighet på 50 kvadratmeter. La X_G være kvadratprisen for leiligheten Gustav vurderer og la X_M være kvadratprisen for leiligheten Margrethe vurderer. Bruk disse til å finne uttrykk for prisen (kjøpssummen, ikke kvadratprisen) til hver av leilighetene. Finn også et uttrykk for prisforskjellen mellom de to leilighetene når vi antar at prisene på leilighetene er uavhengige. Hva er sannsynligheten for at leiligheten Margrethe vurderer er billigere enn leiligheten Gustav vurderer?

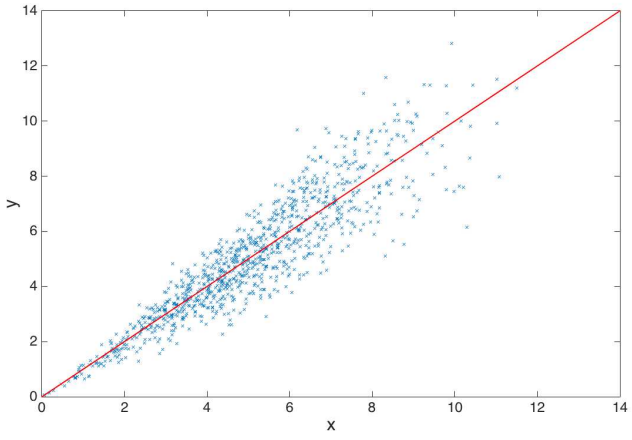
c) Gustav og Margrethe har samlet inn data $(x_1, x_2, \ldots, x_n)$ for kvadratpris (i kkr) fra de siste $n = 15$ boligsalgene i bydelen, og ønsker basert på disse å finne et 95% konfidensintervall for forventet kvadratpris. Utled et uttrykk for konfidensintervallet (du kan ta utgangspunkt i en kjent observator). Regn ut konfidensintervallet numerisk når gjennomsnittet av kvadratprisene er $\bar{x} = 32$ kkr og $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 74.1$.

Oppgave 2

Firmaet SkaffData prøver ut en ny sensor som skal gi billige data på gjennomstrømning av vann i rør. De prøver ut sensoren i en realistisk situasjon. I tillegg til målingene sensoren gir, $(y_1, y_2, \ldots, y_n)$, måler de også tilhørende sann gjennomstrømning, $(x_1, x_2, \ldots, x_n)$, for $n = 1000$ uavhengige tidsperioder. I figur 1 er histogrammet over sann gjennomstrømning, og i figur 2 er sann gjennomstrømning (x) plottet mot sensormålt gjennomstrømning (y).



Figur 1: Histogram for observasjoner av sann gjennomstrømning (x)



Figur 2: Observasjoner av sann gjennomstrømning (x) plottet mot sensormålt gjennomstrømning (y). Den heltrukne linja er $y = x$.

a) Basert på figur 1 og 2 svar på følgende spørsmål, og begrunn alle svarene kort:

- Hva er forventningsverdi og standardavvik til sann gjennomstrømning (X)? (Både forventningsverdi og standardavvik er heltall)
- Hva er forventningsverdi og standardavvik til sensormålt gjennomstrømning (Y) gitt at sann gjennomstrømning er $X = 6$.(Både forventningsverdi og standardavvik er igjen heltall)
- Er korrelasjonen (og kovariansen) mellom sann gjennomstrømning (X) og sensormålt gjennomstrømning (Y) positiv, negativ eller omtrent null?

En enkel lineær regresjonsmodell er som kjent ofte definert som $Y_i = a + bx_i + \epsilon_i$, for $i = 1, 2, \ldots, n$ der Y_i er responsen vi er interessert i, a og b er regresjonsparametre, x_i er en forklaringsvariabel som antas kjent, og støyleddene ϵ_i blir antatt uavhengige identisk normalfordelte med forventningsverdi 0 og varians σ_ϵ^2.

b) Svar på følgende spørsmål, og begrunn alle svarene kort:

- Dersom man tilpasser en enkel lineær regresjonsmodell til dataene i figur 2, hva blir omtrentlig estimatene for a og b?
- Basert på dine anslag for a og b, hva blir predikert sensormålt gjennomstrømning (y_0) for en sann gjennomstrømning på $x_0 = 4$.
- Diskuter om antakelsene i en enkel lineær regresjonsmodell passer for dataene i figur 2.

Oppgave 3

For firmaet SeMeg er antall besøk på deres webside viktig. La X_i være antall besøk i løpet av t_i timer, og la $X_1, X_2, \ldots, X_n$ være antall besøk i n ikke-overlappende tidsintervall. Vi antar at besøk på websiden er en poissonprosess med besøksintensitet λ. Dermed er $X_1, X_2, \ldots, X_n$ uavhengige poissonfordelte stokastiske variabler med sannsynlighetsfordeling

$$f(x_i) = \frac{(\lambda t_i)^{x_i}}{x_i!} e^{-\lambda t_i} \quad \text{for } x_i = 0, 1, 2, \ldots.$$

a) Anta (bare i dette punktet) at $\lambda = 10$ og $t_1 = 1$. Finn sannsynlighetene

$$P(X_1 = 8) \quad , \quad P(X_1 \geq 8) \quad \text{og} \quad P(8 \leq X_1 \leq 12).$$

Vi antar nå at besøksintensiteten λ er ukjent. SeMeg ønsker å estimere intensiteten λ fra data på antall besøk fra n ikke-overlappende tidsintervall. Det er foreslått tre estimatorer,

$$\widetilde{\lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad , \quad \widehat{\lambda} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} t_i} \quad \text{og} \quad \widehat{\widehat{\lambda}} = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{t_i},$$

og vi oppgir at $\mathrm{E}[\widehat{\lambda}] = \lambda$ og $\mathrm{Var}[\widehat{\lambda}] = \lambda / \sum_{i=1}^{n} t_i$.

b) Hvilken av de tre estimatorene vil du foretrekke når $n = 5$ og $t_1 = 1$, $t_2 = 2$, $t_3 = 5$, $t_4 = 1$, $t_5 = 5$? Begrunn svaret.

c) Utled sannsynlighetsmaksimeringsestimatoren (SME) for λ basert på X_1, $X_2, \ldots, X_n$.

For punkt **d**) og **e**) i denne oppgaven skal du, uavhengig av dine resultat i punkt **b**) og **c**), ta utgangspunk i estimatoren $\widehat{\lambda}$ definert over. Videre kan du forutsette at $\lambda \sum_{i=1}^{n} t_i$ er stor og bruke at da er $\widehat{\lambda}$ tilnærmet normalfordelt med forventningsverdi og varians som gitt over.

SeMeg får vite at besøksintensiteten til deres største konkurrent er på $\lambda_0 = 10$ besøk per time, og ønsker å benytte observerte verdier av $X_1, X_2, \ldots, X_n$ til å avgjøre om det er grunnlag for å påstå at deres webside har en høyere besøksintensitet.

d) Formuler hypotesene H_0 og H_1 for situasjonen beskrevet over.

Angi hvilken testobservator du vil benytte og hvilken (tilnærmet) sannsynlighetsfordeling testobservatoren har når H_0 er riktig.

Regn ut p-verdien til hypotesetesten når n og $t_1, t_2, \ldots, t_n$ er som i **b**), og observert antall besøk er $x_1 = 8$, $x_2 = 20$, $x_3 = 48$, $x_4 = 10$ og $x_5 = 62$. Med utgangspunkt i den utregnede p-verdien, diskuter kort om det er grunnlag for å påstå at SeMeg har høyere besøksintensitet enn konkurrenten.

e) Dersom man benytter signifikansnivå $\alpha = 0.05$ i hypotesetesten i **d**), hvor stor må besøksintensiteten til SeMeg være for at sannsynligheten for å konkludere med at den er høyere enn konkurrentens intensitet skal være minst 0.9? Bruk her de samme verdiene for n og $t_1, t_2, \ldots, t_n$ som i punkt **b**).

Fasit

1. **a**) $0.5, 0.9772, 0.9544$ **b**) $40X_G$, $50X_M$, $40X_G - 50X_M$, 0.0307 **c**) $(30.7, 33.3)$

2. **a**) $5, 2$ **b**) $a = 0, b = 1$, $\widehat{y} = 4$

3. **a**) 0.1126, 0.7798, 0.5714 **b**) Foretrekker $\widehat{\lambda}$ **d**) $H_0 : \lambda = \lambda_0$, $H_1 : \lambda > \lambda_0$, 0.2483 **e**) $\lambda \geq 12.6068$