Kunnskap for ei betre verd

Institutt for matematiske fag

Eksamensoppgåve i **TMA4245 Statistikk**

Fagleg kontakt under eksamen: Ingelin Steinsland[a], Jo Eidsvik[b]

Tlf: [a] 92 66 30 96 , [b] 90 12 74 72

Eksamensdato: 30. mai 2016

Eksamenstid (frå–til): 09.00-13.00

Hjelpemiddelkode/Tillatne hjelpemiddel: *Tabeller og formler i statistikk*, Tapir Forlag, K. Rottmann: *Matematisk formelsamling*, Kalkulator Casio fx-82ES PLUS, CITIZEN SR-270X, CITIZEN SR-270X College eller HP30S, Gult stempla A5-ark med eigne handskrivne notat.

Annan informasjon:
Alle svar skal grunngjevast og besvarelsen skal innehalde naturlege mellomrekningar.

Målform/språk: nynorsk

Sidetal: 4

Sidetal vedlegg: 0

Kontrollert av:

Dato Sign

Oppgåve 1 Huskjøp

Gustav og Margrethe er nyutdanna sivilingeniørar frå NTNU og er no på bustadjakt i Trondheim. Begge ser etter ei leilegheit i ein 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) Kva er sannsynet for at kvadratprisen for ei tilfeldig leilegheit er:

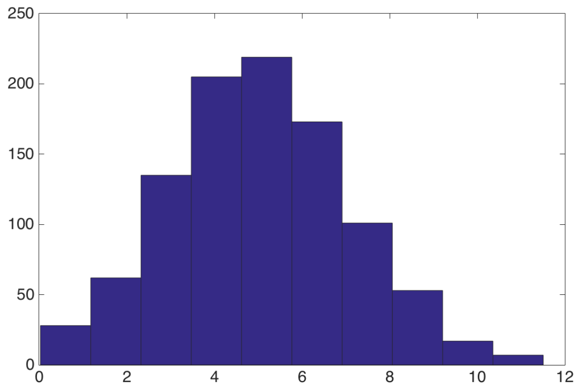
- lågare enn 30 kkr?
- høgare enn 25 kkr?
- høgare enn 25 kkr gjeve at kvadratprisen er lågare enn 30 kkr.

b) Gustav vurderer ei leilegheit på 40 kvadratmeter, og Margrethe vurderer ei leilegheit på 50 kvadratmeter. La X_G vere kvadratprisen for leilegheit Gustav vurderer og la X_M vere kvadratprisen for leilegheita Margrethe vurderer. Bruk desse til å finne uttrykk for prisen (kjøpssummen, ikkje kvadratprisen) til kvar av leiligheitene. Finn og eit uttrykk for prisforskjellen mellom dei to leiligheitene når vi antar at prisane på leiligheitene er uavhengige. Kva er sannsynet for at leiligheita Margrethe vurderer er billigare enn leilegheita Gustav vurderer?

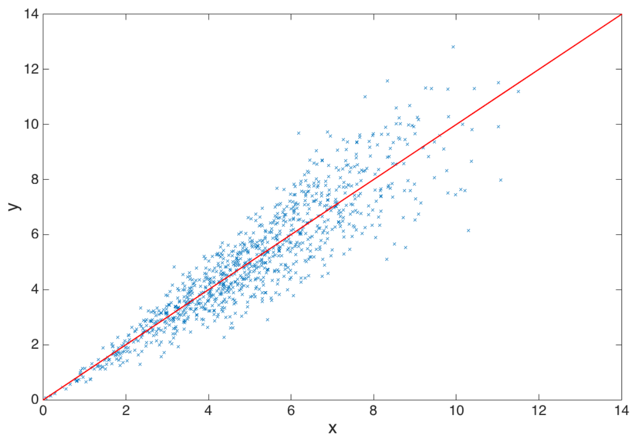
c) Gustav og Margrethe har samla inn data $(x_1, x_2, \ldots, x_n)$ for kvadratpris (i kkr) frå dei siste $n = 15$ bustadsala i bydelen, og ønsker basert på desse å finne eit 95% konfidensintervall for forventa kvadratpris. Utlei eit uttrykk for konfidensintervallet (du kan ta utgangspunkt i ein kjent observator). Rekn ut konfidensintervallet numerisk når gjennomsnittet av kvadratprisane er $\bar{x} = 32$ kkr og $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 74.1$.

Oppgåve 2 Den nye sensoren

Firmaet SkaffData prøver ut ein ny sensor som skal gje billige data på vassgjennomstrauming i røyr. Dei prøver ut sensoren i ein realistisk situasjon. I tillegg til målingane sensoren gjev, $(y_1, y_2, \ldots, y_n)$, måler dei og tilhøyrande sann vassgjennomstrauming, $(x_1, x_2, \ldots, x_n)$, for $n = 1000$ uavhengige tidsperiodar. I figur 1 er histogrammet over sann gjennomstrauming, og i figur 2 er sann gjennomstrauming (x) plotta mot sensormålt gjennomstrauming (y).



Figur 1: Histogram for observasjonar av sann gjennomstrauming (x)



Figur 2: Observasjonar av sann gjennomstrauming (x) plotta mot sensormålt gjennomstrauming (y). Den heiltrukne linja er $y = x$.

a) Basert på figur 1 og 2 svar på følgjande spørsmål, og grunngje alle svara kort:

- Kva er forventningsverdien og standardavviket til sann gjennomstrauming (X)? (Både forventningsverdien og standardavviket er heiltal)
- Kva er forventningsverdien og standardavviket til sensormålt gjennomstrauming (Y) gjeve at sann gjennomstrauming er $X = 6$.(Både forventningsverdien og standardavviket er igjen heiltal)
- Er korrelasjonen (og kovariansen) mellom sann gjennomstrauming (X) og sensormålt gjennomstrauming (Y) positiv, negativ eller omlag null?

Ein 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 regresjonsparameterar, x_i er ein forklaringsvariabel som vi antar er kjent, og støyledda ϵ_i antar vi er uavhengige identisk normalfordelte med forventningsverdi 0 og varians σ_ϵ^2.

b) Svar på følgjande spørsmål, og grunngje alle svara kort:

- Dersom ein tilpassar ein enkel lineær regresjonsmodell til dataene i figur 2, kva blir omlag estimata for a og b?
- Basert på anslaga dine for a og b, kva blir predikert sensormålt gjennomstrauming (y_0) for ei sann gjennomstrauming på $x_0 = 4$.
- Diskuter om antakingane i ein enkel lineær regresjonsmodell passar for dataene i figur 2.

Oppgåve 3 Besøk på webside

For firmaet SjåMeg er talet på besøk på websida deira viktig. La X_i vere talet på besøk i løpet av t_i timar, og la $X_1, X_2, \ldots, X_n$ vere talet på besøk i n ikkje-overlappande tidsintervall. Vi antar at besøk på websida er ein poissonprosess med besøksintensitet λ. Dermed er $X_1, X_2, \ldots, X_n$ uavhengige poissonfordelte stokastiske variabler med sannsynsfordeling

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

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

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

Vi antar no at besøksintensiteten λ er ukjent. SjåMeg ønsker å estimere intensiteten λ frå data på talet på besøk frå n ikkje-overlappande tidsintervall. Det er foreslått tre estimatorar,

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

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

b) Kven av dei tre estimatorane vil du foretrekke når $n = 5$ og $t_1 = 1$, $t_2 = 2$, $t_3 = 5$, $t_4 = 1$, $t_5 = 5$? Grunngje svaret.

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

For punkt **d**) og **e**) i denne oppgåven skal du, uavhengig av resultata dine i punkt **b**) og **c**), ta utgangspunk i estimatoren $\widehat{\lambda}$ definert over. Vidare kan du forutsette at $\lambda \sum_{i=1}^{n} t_i$ er stor og bruke at då er $\widehat{\lambda}$ tilnærma normalfordelt med forventningsverdi og varians som gjeve over.

SjåMeg får vite at besøksintensiteten til deira største konkurrent er på $\lambda_0 = 10$ besøk per time, og ønsker å bruke observerte verdiar av $X_1, X_2, \ldots, X_n$ til å avgjere om det er grunnlag for å påstå at deira webside har ein høgare besøksintensitet.

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

Oppgje kva av testobservatorane du vil bruke og kva (tilnærma) sannsynsfordeling testobservatoren har når H_0 er rett.

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

e) Dersom ein bruker signifikansnivå $\alpha = 0.05$ i hypotesetesten i **d**), kor stor må besøksintensiteten til SjåMeg vere for at sannsynet for å konkludere med at intensiteten er høgere enn konkurrenten sin intensitet skal vere minst 0.9? Bruk her dei same verdiane for n og $t_1, t_2, \ldots, t_n$ som i punkt **b**).