Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

TMA4245 Statistikk

Øving nummer b5

Oppgave 1 Eksamen mai 2001, oppgave 1 av 4

Vi ser på konsentrasjonen av et giftstoff i havbunnen like utenfor en fabrikk. Miljøforskriftene sier at konsentrasjonen ikke skal overstige 12 [g/cm^3]. For å kontrollere dette tas prøver av havbunnen. Anta at en prøveverdi Y er normalfordelt med forventning μ og standardavvik σ. Sett $\mu = 13$ og $\sigma = 1{,}5$ i punkt a), og la de være ukjent i resten av oppgaven.

a) Beregn $P(Y < 12)$ og $P(11 < Y < 14)$.

b) De observerte måleverdiene er

$$11{,}7 \quad 12{,}4 \quad 12{,}8 \quad 12{,}9 \quad 13{,}3.$$

Kan vi på grunnlag av dette konkludere med at giftkonsentrasjonen på havbunnen like ved fabrikken er over 12? Formuler problemstillingen som en hypotesetest og utfør testen på signifikansnivå 0,05.

c) Det blir tatt 10 nye målinger, men denne gang i ulike avstander x fra fabrikken. Målingene er

x	10	20	30	40	50	60	70	80	90	100
y	9,9	11,1	9,3	10,6	9,2	9,3	10,0	9,2	10,3	8,4

I tillegg kommer de fem målingene i b). Her er $x = 0$. Det oppgis at $\sum x_i = 550$, $\sum (x_i - \bar{x})^2 = 18\,333{,}33$, $\sum y_i = 160{,}4$ og $\sum x_i y_i = 5245$.

Vi velger å utføre en lineær regresjonsanalyse med Y som avhengig variabel og x som uavhengig variabel. Modellen er

$$E(Y \mid x) = \alpha + \beta x.$$

Beregn estimatene for α og β. Forklar hva estimatet for α beskriver i dette eksemplet.

Regresjonsanalysen gir oss ikke grunnlag for å konkludere med at $\alpha > 12$. Hvorfor ikke? Sammenlign resultatet fra denne analysen med resultatet i b) og kommenter. Hvorfor kan det skje at to slike analyser gir forskjellig konklusjon? Bruk gjerne figur i forklaringen.

Oppgave 2 Automatisert laboratorium — Eksamen november 2002, oppgave 3 av 3

I eit laboratorium ynskjer ein å evaluere samanhengen mellom to variablar Y og x. Apparaturen er sett opp slik at ein kan fastsetje x for deretter å måle Y. Ein vel å nytte følgjande modell for samanhengen mellom variablane

$$Y = \alpha + \beta x + \varepsilon$$

der α og β er to ukjende koeffisientar og ε er ein tilfeldig variabel som er normalfordelt med forventing 0 og ukjend varians σ^2. La $x_1, x_2, \ldots, x_n$ vere n verdiar av variabelen x og $y_1, y_2, \ldots, y_n$ dei tilhøyrande verdiane som blir målt for Y. Desse skal sjåast på som realiseringar av n uavhengige variablar $Y_1, Y_2, \ldots, Y_n$. Minste kvadratsums (least squares) estimatorane, A og B, for koeffisientane α og β er då gitt ved

$$A = \frac{1}{n}\sum_{i=1}^{n} Y_i - B\bar{x} \qquad \text{og} \qquad B = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{der} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

a) Vis at estimatorane A og B er forventingsrette estimatorar for α og β.

Bruk at kovariansen mellom $\bar{Y} = (1/n)\sum_{i=1}^{n} Y_i$ og B er 0 og utlei variansen til estimatorane A og B.

Kva sannsynsfordelingar har estimatorane A og B? Grunngjev svaret.

Laboratorieforsøka er svært arbeidskrevjande, men apparaturen er automatisert slik at forsøka kan utførast automatisk for ekvidistante verdiar av x. Sjå på to måleseriar med $n = 10$

$$\begin{array}{ll} \text{Serie 1:} & x_1 = 1, x_2 = 2, \ldots, x_{10} = 10 \\ \text{Serie 2:} & x_1 = 2, x_2 = 4, \ldots, x_{20} = 20 \end{array}$$

Målet med forsøket er å prediktere Y_0 for $x_0 = 5.5$. Følgjande prediktor blir brukt: $\widehat{Y}_0 = A + Bx_0$.

b) Utlei variansen til $Y_0 - \widehat{Y}_0$.

Kva måleserie bør nyttast for å prediktere Y_0 best mogeleg for $x_0 = 5.5$? Grunngjev og kommenter svaret du har funne.

Oppgave 3 Medisinkonsentrasjon — Eksamen januar 1999, oppgave 1 av 4

Ved behandling av visse kreftformer får pasientene kurer der en bestemt type medisin blir injisert i blodet i løpet av 24 timer. Alle pasienter får tilført samme dose medisin. Ved avslutningen av kuren blir konsentrasjonen av medisin i blodet målt. Medisinkonsentrasjonen måles i milligram medisin per liter blod. For at behandlingen skal ha ønsket effekt bør medisinkonsentrasjonen ved avslutningen av kuren helst overstige 5 mg/l. På grunn av bivirkninger blir det ansett som uheldig om medisinkonsentrasjonen overstiger 12 mg/l. La Y betegne målt medisinkonsentrasjon ved avslutningen av en kur, og anta at Y er normalfordelt med forventning μ og varians σ^2. Målt medisinkonsentrasjon ved avslutningen av ulike kurer antas uavhengige. Anta i første omgang at $\mu = 8$ og $\sigma^2 = 2^2$.

a) Beregn sannsynlighetene $P(Y \leq 12)$, $P(Y > 5)$ og $P(5 < Y \leq 12)$.

Dersom en pasient går gjennom 8 kurer, hva er sannsynligheten for at målt medisinkonsentrasjon ved slutten av samtlige 8 kurer er i intervallet (5,12]?

Følgende hendelser er definert:

A_1: Målt medisinkonsentrasjon ved slutten av en kur overstiger 5 mg/l (dvs $Y > 5$).
A_2: Målt medisinkonsentrasjon ved slutten av en kur er mindre eller lik 12 mg/l (dvs $Y \leq 12$).

b) Er A_1 og A_2 disjunkte? (Begrunn svaret)

Er A_1 og A_2 uavhengige? (Begrunn svaret)

Følgende hendelse er definert:

A_3: Målt medisinkonsentrasjonen ved slutten av en kur er mellom 5 mg/l og 12 mg/l (dvs $5 < Y \leq 12$).

Uttrykk A_3 ved A_1 og A_2.

Anta nå at μ er ukjent, mens $\sigma^2 = 2^2$ fremdeles antas kjent. Fra åtte ulike kurer har man registrert dataene:

kur i	1	2	3	4	5	6	7	8
y_i	7.1	9.2	10.8	12.0	6.1	8.2	8.7	7.7

c) Skriv opp en rimelig estimator for μ, og regn ut estimatet.

Utled et 95% konfidensintervall for μ. Hva blir intervallet med de oppgitte dataene?

Legene har etterhvert funnet ut at i stedet for å gi alle pasienter samme dose medisin, vil det være gunstigere å justere dosene etter hvor syk pasienten er og hvor godt han/hun tåler bivirkningene. La x være dosen. Vi antar at x kan kontrolleres, dvs x er ikke stokastisk.

Man antar at en god lineær regresjonsmodell for sammenhengen mellom x og Y vil være

$$Y = \beta x + E,$$

der β er en ukjent konstant og E er en normalfordelt stokastisk variabel med forventningsverdi 0 og kjent varians $\sigma_E^2 = 2^2$.

d) Hvorfor er det i dette tilfellet rimelig å ikke ha med noe konstantledd i den lineære regresjonsmodellen?

Vis at sannsynlighetsmaksimeringsestimatoren (SME) for β basert på n uavhengige observasjoner blir

$$\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

der x_i og Y_i er henholdsvis dose og målt medisinkonsentrasjon for observasjon nummer i.

Regn ut forventningen og variansen til $\hat{\beta}$.

Det har i løpet av ti kurer på ulike pasienter blitt observert følgende sammenhørende verdier for x og Y:

kur i	1	2	3	4	5	6	7	8	9	10
x_i	4.5	4.0	5.5	7.0	8.0	8.5	9.0	6.5	6.0	5.0
y_i	6.2	5.2	7.3	8.7	9.0	10.5	10.3	8.2	7.4	7.0

Det oppgis at $\sum_{i=1}^{10} y_i x_i = 536.4$ og $\sum_{i=1}^{10} x_i^2 = 436$.

Før legene gir en pasient en viss dose x_0 ønsker de å vite noe om hvilken målt medisinkonsentrasjon Y_0 man kan regne med at dette vil gi. Du skal hjelpe legene ved å lage et 95% prediksjonsintervall.

e) Hva er tolkningen av et 95% prediksjonsintervall?

Utled et 95% prediksjonsintervall for Y_0 når $x_0 = 8$ ved å bruke de oppgitte dataene.

Oppgave 4 Hubble — Eksamen mai 2006, oppgave 4 av 4

En viktig vitenskapelig oppdagelse fant sted i 1929 da Edwin Hubble oppdaget at universet er ekspanderende. Hubble's tallmateriale bestod blant annet av; x_i = avstanden til galakse i (målt i millioner lysr), og y_i = hastigheten til galakse i (målt i 1000 km/s). Verdiene Hubble benyttet i en av sine analyser er som følger:

Navn	Avstand, x_i	Hastighet, y_i
Virgo	22	1.2
Pegasus	68	3.8
Perseus	108	5.1
Coma Berenices	137	7.5
Ursa Major 1	255	14.9
Leo	315	19.2
Corona Borealis	390	21.4
Gemini	405	23.0
Bootes	685	39.2
Ursa Major 2	700	41.6
Hydra	1100	60.8

Det oppgis her at $\sum_{i=1}^{11} x_i = 4185$, $\sum_{i=1}^{11} y_i = 237.7$, $\sum_{i=1}^{11} x_i^2 = 2685141$ og $\sum_{i=1}^{11} x_i y_i = 152224$.

Hubble foreslo en modell for hastighet som funksjon av avstand på formen $y = \beta x$, der β senere har blitt kalt Hubble's konstant. En statistisk versjon av ligningen kan gis ved:

$$Y_i = \beta x_i + \varepsilon_i, \qquad i = 1, \ldots, 11, \tag{4.1}$$

der ε_i, $i = 1, \ldots, 11$, er uavhengige og normalfordelte stokastiske variabler med forventning 0 og varians σ^2.

a) Vi vil i første omgang finne en estimator for β.

Bruk minste kvadraters metode (method of least squares) til å estimere β med utgangspunkt i ligning (4.1), og vis at estimatoren for β da blir gitt ved $\hat{\beta} = \frac{\sum_{i=1}^{11} x_i Y_i}{\sum_{i=1}^{11} x_i^2}$. Regn ut estimatet for β basert på dataene over.

Finn også forventning og varians til $\hat{\beta}$.

b) Anta at en annen galakse befinner seg en avstand $x_0 = 900$ millioner lysr borte.

Finn predikert hastighet, $\hat{y}_0$, til denne galaksen.

Utled et 95% prediksjonsintervall for en måling av hastigheten til denne galaksen. Det oppgis at $\sum_{i=1}^{11}(y_i - \hat{y}_i)^2 = 9.87$, der $\hat{y}_i = \hat{\beta}x_i$.

Fasit

1. **a**) 0.251, 0.657 **b**) Forkaster H_0

2. **b**) Bør benytte måleserie 2

3. **a**) $0.977, 0.933, 0.910, 0.47$ **b**) $A_3 = A_1 \cap A_2$ **c**) $\widehat{\mu} = \overline{Y}, 8.725, [7.34, 10.11]$ **d**) $\mathrm{E}(\widehat{\beta}) = \beta$, $\mathrm{Var}(\widehat{\beta}) = \sigma_E^2 / \sum_{i=1}^{n} x_i^2$ **e**) $[5.64, 14.04]$

4. **a**) 0.0567 **b**) 51.03, (48.5,53.5)