NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i **TMA4245 Statistikk**

Faglig kontakt under eksamen: Ingelin Steinsland[a], Øyvind Bakke[b]

Tlf: [a] 73 59 02 39, 926 63 096 , [b] 73 59 81 26, 990 41 673

Eksamensdato: 19. mai 2014

Eksamenstid (fra–til): 9.00–13.00

Hjelpemiddelkode/Tillatte hjelpemidler: Stemplet gult A5-ark med egne håndskrevne notater, bestemt enkel kalkulator, *Tabeller og formler i statistikk*, *Matematisk formelsamling* (K. Rottmann)

Målform/språk: bokmål

Antall sider: 4

Antall sider vedlegg: 0

Kontrollert av:

Dato Sign

Oppgave 1 Samleserien

Agnes samler på kort i samleserien *Verdens dyr*. Serien består av θ forskjellige kort. På hvert kort er det bilde av en dyreart og opplysninger om arten. I tillegg er et av tallene 1, 2, ..., θ trykt på kortet – dette tallet er kortets nummer i samleserien.

La X være nummeret på et kort som kjøpes i butikken. Vi antar at $P(X = x) = 1/\theta$ for $x = 1, 2, \ldots, \theta$ og $P(X = x) = 0$ for alle andre x. Det vil si at det er samme sannsynlighet for å få hver type kort. Vi antar også at når vi kjøper flere kort, er kortnumrene uavhengige.

a) Anta (i dette punktet) at det er 50 forskjellige kort, altså at $\theta = 50$.

Agnes kjøper 2 kort. Hva er sannsynligheten for at de er forskjellige?

Hva er sannsynligheten for at alle kortene er forskjellige hvis Agnes kjøper 8 kort?

Produsenten av kortene reklamerer med at det er 200 kort i serien. Agnes har kjøpt 20 kort, men har aldri fått noe høyere kortnummer enn 170. Anta at $X_1, X_2, \ldots, X_n$ er uavhengige kortnumre, og la $\max X_i$ være det største av disse kortnumrene.

b) Finn kumulativ fordelingsfunksjon $P(X_i \leq x)$ for $x = 1, 2, \ldots, \theta$.

Vis at $P(\max X_i \leq x) = (x/\theta)^n$ for $x = 1, 2, \ldots, \theta$.

Hva er $P(\max X_i \leq 170)$ hvis $n = 20$ og det er $\theta = 200$ forskjellige kort i samleserien?

Anta at θ er ukjent. Numrene på kortene som Agnes har kjøpt, er 7, 8, 25, 32, 55, 72, 74, 74, 89, 100, 102, 114, 121, 124, 126, 129, 131, 151, 165 og 170.

Agnes vil teste nullhypotesen $\theta = 200$ mot alternativet $\theta < 200$, og bruker $\max X_i$, det vil si høyeste kortnummer, som testobservator. Hun finner et forkastningsområde som er gitt ved $\max x_i \leq 172$.

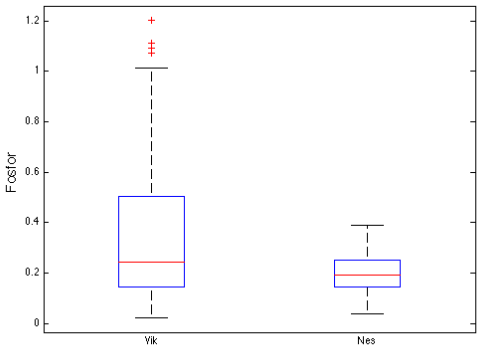
c) Hva blir konklusjonen av hypotesetesten med Agnes' data?

Finn signifikansnivået for testen.

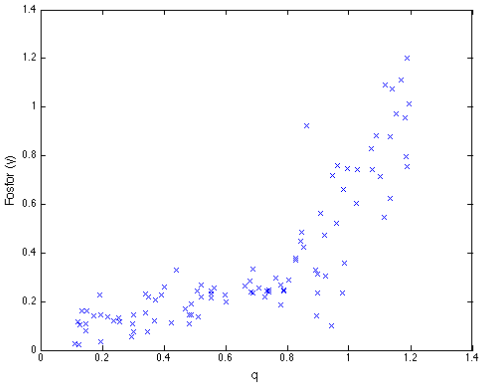
Finn teststyrken i $\theta = 180$ og i $\theta = 160$.

d) Vis at sannsynlighetsmaksimeringsestimatet for θ er 170 med Agnes' data.

Er sannsynlighetsmaksimeringsestimatoren forventningsrett? Begrunn svaret (du trenger ikke å regne ut estimatorens forventningsverdi).



Figur 1: Boksplott fra to anlegg



Figur 2: 100 observasjoner av fosforinnhold og gjennomstrømning

Oppgave 2 Fosfor fra renseanlegg

Vi er interessert i fosforinnhold (i gram pr. kubikkmeter) i ferdig renset vann fra renseanlegg.

a) Figur 1 viser boksplott av målinger av fosforinnhold fra to anlegg, Vik og Nes. Vurder ut fra boksplottene om fosforinnholdet kan komme fra normalfordelinger, og om fosforinnhold fra de to anleggene har like medianer, like forventningsverdier og like varianser. Begrunn kort svarene dine.

b) Vi antar i dette punktet at fosforinnholdet Y av en prøve er normalfordelt med forventningsverdi $\mu = 0{,}3$ og varians $\sigma^2 = 0{,}1^2$.

Finn sannsynligheten for at Y er mindre enn 0,5.

Finn sannsynligheten for at Y er større enn 0,3.

Finn den betingede sannsynligheten for at Y er mindre enn 0,5 gitt at Y er større enn 0,3.

En grunn til at fosforinnholdet varierer, kan være at det avhenger av gjennomstrømningen i anlegget. La q_i være gjennomstrømningen (i kubikkmeter pr. sekund) der prøve nr. i ble tatt og Y_i fosforinnholdet i prøve nr. i. Vi antar en enkel lineær regresjonsmodell

$$Y_i = \alpha + \beta q_i + \epsilon_i,$$

der α og β er regresjonsparametre. Videre antar vi at støyleddene ϵ_i er uavhengige og normalfordelte med forventningsverdi 0 og varians σ_ϵ^2.

c) Vi antar (bare i dette punktet) at regresjonsparametrene er kjente: $\alpha = 0{,}05$, $\beta = 0{,}3$ og $\sigma_\epsilon^2 = 0{,}05^2$.

Vis at fosforinnholdet i en prøve ved gjennomstrømning 0,5 er normalfordelt med forventningsverdi 0,2 og varians $0{,}05^2$, og at fosforinnholdet i en prøve ved gjennomstrømning 1,0 er normalfordelt med forventningsverdi 0,35 og varians $0{,}05^2$.

Hva er sannsynligheten for at den største av tre uavhengige fosformålinger ved gjennomstrømning 1,0 er større enn 0,4?

Hva er sannsynligheten for at en måling ved gjennomstrømning 0,5 er større enn en (uavhengig) måling ved gjennomstrømning 1,0?

Figur 2 viser $n = 100$ observasjoner av fosforinnhold og gjennomstrømning. Vi ønsker nå å estimere α og β ved minste kvadraters metode basert på disse dataene.

d) Forklar kort hva minste kvadraters metode er, og illustrer med en figur.

Sett opp uttrykkene du trenger, og forklar framgangsmåten. Du trenger ikke utlede uttrykkene for estimatorene.

Vi antar nå at variansen σ_ϵ^2 er kjent. La $\hat{Y}_0$ være prediksjonen av fosforinnhold gitt av den tilpassede (estimerte) regresjonsmodellen ved gjennomstrømning q_0. Det oppgis at $\hat{Y}_0$ er normalfordelt med forventningsverdi $\alpha + \beta q_0$ og varians

$$\sigma_\epsilon^2 \left(\frac{1}{n} + \frac{(q_0 - \bar{q})^2}{\sum_{i=1}^n (q_i - \bar{q})^2} \right).$$

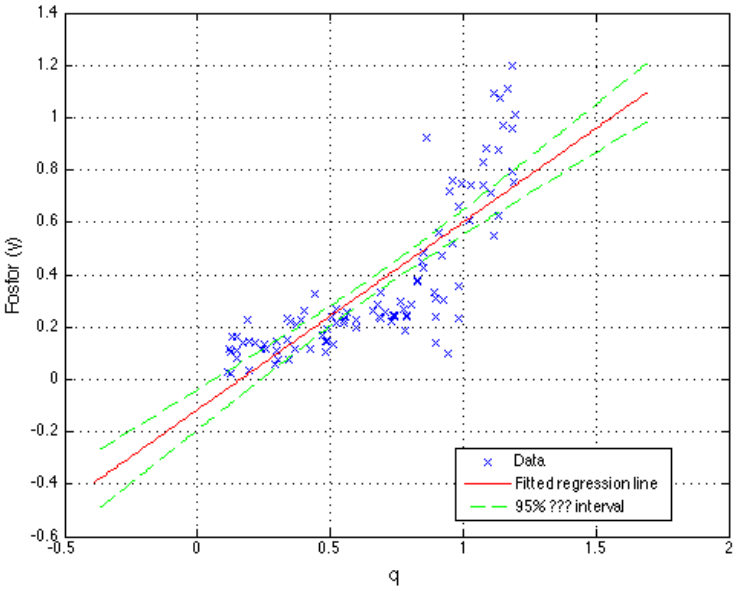
e) Utled et 95 %-prediksjonsintervall for en ny (uavhengig) observasjon av fosforinnholdet når gjennomstrømningen er q_0.

Forklar kort forskjellen på et konfidensintervall og et prediksjonsintervall.

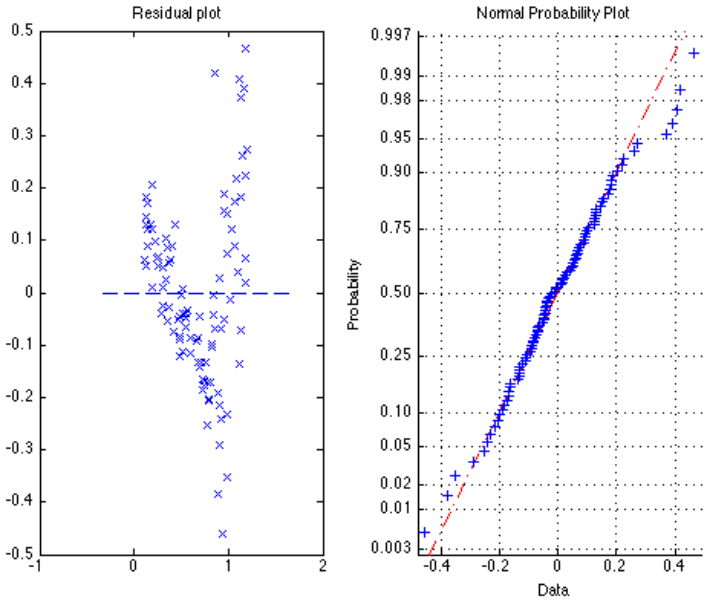
I figur 3 er data plottet sammen med tilpasset (estimert) regresjonslinje og grensene for et intervall. Er dette et 95 %-prediksjonsintervall eller et 95 %-konfidensintervall? Begrunn svaret.

f) Spesifiser antakelsene som er gjort i regresjonsmodellen.

Diskuter ut fra figur 2, 3 og 4 om disse antakelsene er oppfylt.



Figur 3: Estimert regresjonslinje med grenser for intervall



Figur 4: Venstre: Residualplott (differanse mellom data og estimert regresjonslinje langs y-akse, gjennomstrømning langs x-akse). Høyre: Normalsannsynlighetsplott (normalkvantil-kvantilplott, QQ-plott) for residualer (differanser mellom data og estimert regresjonsllinje)