NTNU

Kunnskap for ei betre verd

Institutt for matematiske fag

Eksamensoppgåve i **TMA4245 Statistikk**

Fagleg 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 (frå–til): 9.00–13.00

Hjelpemiddelkode/Tillatne hjelpemiddel: Stempla gult A5-ark med eigne handskrivne notat, bestemd enkel kalkulator, *Tabeller og formler i statistikk*, *Matematisk formelsamling* (K. Rottmann)

Målform/språk: nynorsk

Sidetal: 4

Sidetal vedlegg: 0

Kontrollert av:

Dato Sign

Oppgåve 1 Samleserien

Agnes samlar på kort i samleserien *Verdas dyr*. Serien består av θ forskjellige kort. På kvart kort er det bilde av ein dyreart og opplysningar om arten. I tillegg er eit av tala 1, 2, ..., θ trykt på kortet – dette talet er kortet sitt nummer i samleserien.

La X vere nummeret på eit kort som blir kjøpt 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 seie at det er same sannsynet for å få kvar type kort. Vi antar òg at når vi kjøper fleire kort, er kortnummera uavhengige.

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

Agnes kjøper 2 kort. Kva er sannsynet for at dei er forskjellige?

Kva er sannsynet for at alle korta er forskjellige dersom Agnes kjøper 8 kort?

Produsenten av korta reklamerer med at det er 200 kort i serien. Agnes har kjøpt 20 kort, men har aldri fått noko høgre kortnummer enn 170. Anta at $X_1, X_2, \ldots, X_n$ er uavhengige kortnummer, og la $\max X_i$ vere det største av desse kortnummera.

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 $x = 1, 2, \ldots, \theta$.

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

Anta at θ er ukjend. Nummera på korta 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 seie høgste kortnummer, som testobservator. Ho finn eit forkastningsområde som er gjeve ved $\max x_i \leq 172$.

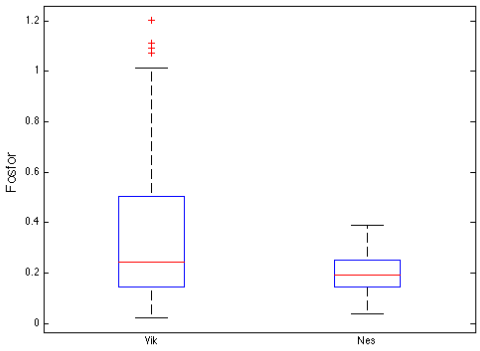
c) Kva blir konklusjonen av hypotesetesten med Agnes sine data?

Finn signifikansnivået for testen.

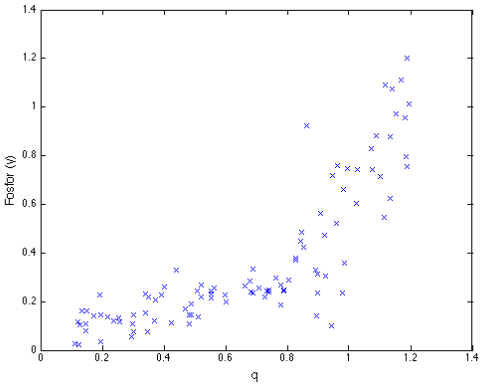
Finn teststyrken i $\theta = 180$ og i $\theta = 160$.

d) Vis at sannsynsmaksimeringsestimatet for θ er 170 med Agnes sine data.

Er sannsynsmaksimeringsestimatoren forventningsrett? Grunngje svaret (du treng ikkje å rekne ut forventningsverdien til estimatoren).



Figur 1: Boksplott frå to anlegg



Figur 2: 100 observasjonar av fosforinnhald og gjennomstrøyming

Oppgåve 2 Fosfor frå renseanlegg

Vi er interesserte i fosforinnhald (i gram pr. kubikkmeter) i ferdig rensa vatn frå reinseanlegg.

a) Figur 1 viser boksplott av målingar av fosforinnhald frå to anlegg, Vik og Nes. Vurder ut frå boksplotta om fosforinnhaldet kan kome frå normalfordelingar, og om fosforinnhald frå dei to anlegga har like medianar, like forventningsverdiar og like variansar. Grunngje kort svara dine.

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

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

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

Finn det vilkårsbunde (betinga) sannsynet for at Y er mindre enn 0,5 gjeve at Y er større enn 0,3.

Ein grunn til at fosforinnhaldet varierer, kan vere at det avheng av gjennomstrøyminga i anlegget. La q_i vere gjennomstrøyminga (i kubikkmeter pr. sekund) der prøve nr. i vart tatt og Y_i fosforinnhaldet i prøve nr. i. Vi antar ein enkel lineær regresjonsmodell

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

der α og β er regresjonsparametrar. Vidare antar vi at støyledda ϵ_i er uavhengige og normalfordelte med forventningsverdi 0 og varians σ_ϵ^2.

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

Vis at fosforinnhaldet i ein prøve ved gjennomstrøyming 0,5 er normalfordelt med forventningsverdi 0,2 og varians $0{,}05^2$, og at fosforinnhaldet i ein prøve ved gjennomstrøyming 1,0 er normalfordelt med forventningsverdi 0,35 og varians $0{,}05^2$.

Kva er sannsynet for at den største av tre uavhengige fosformålingar ved gjennomstrøyming 1,0 er større enn 0,4?

Kva er sannsynet for at ei måling ved gjennomstrøyming 0,5 er større enn ei (uavhengig) måling ved gjennomstrøyming 1,0?

Figur 2 viser $n = 100$ observasjonar av fosforinnhald og gjennomstrøyming. Vi ønsker no å estimere α og β ved minste kvadrats metode basert på desse dataa.

d) Forklar kort kva minste kvadrats metode er, og illustrer med ein figur.

Sett opp uttrykka du treng, og forklar framgangsmåten. Du treng ikkje utleie uttrykka for estimatorane.

Vi antar no at variansen σ_ϵ^2 er kjend. La $\hat{Y}_0$ vere prediksjonen av fosforinnhald gjeve av den tilpassa (estimerte) regresjonsmodellen ved gjennomstrøyming q_0. Det blir oppgjeve at $\hat{Y}_0$ er normalfordelt med forventningsverdi $\alpha + \beta q_0$ og varians

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

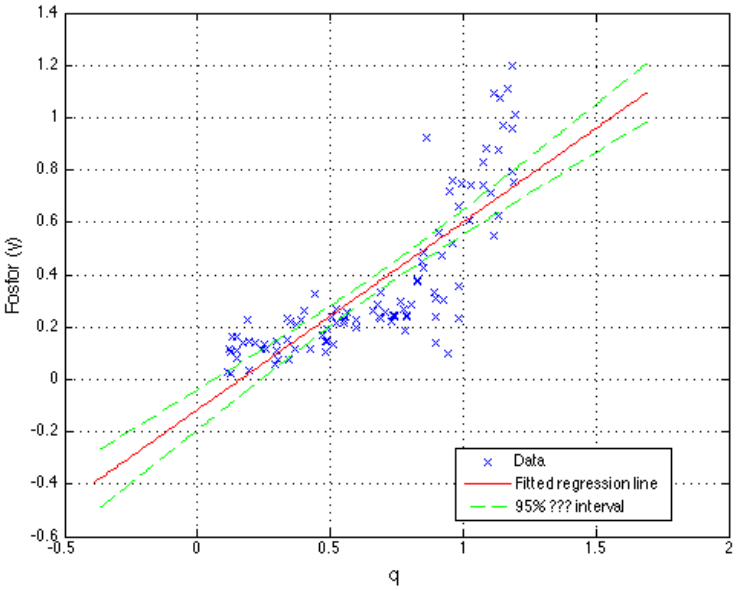
e) Utlei eit 95 %-prediksjonsintervall for ein ny (uavhengig) observasjon av fosforinnhaldet når gjennomstrøyminga er q_0.

Forklar kort forskjellen på eit konfidensintervall og eit prediksjonsintervall.

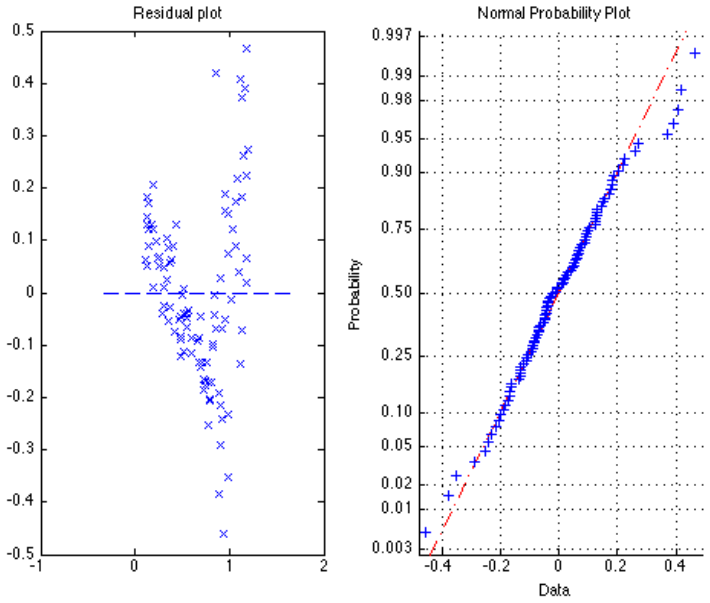
I figur 3 er data plotta saman med tilpassa (estimert) regresjonslinje og grensene for eit intervall. Er dette eit 95 %-prediksjonsintervall eller eit 95 %-konfidensintervall? Grunngje svaret.

f) Spesifiser antakingane som er gjorde i regresjonsmodellen.

Diskuter ut frå figur 2, 3 og 4 om desse antakingane er oppfylte.



Figur 3: Estimert regresjonslinje med grenser for intervall



Figur 4: Venstre: Residualplott (differanse mellom data og estimert regresjonslinje langs y-akse, gjennomstrøyming langs x-akse). Høgre: Normalsannsynsplott (normalkvantil-kvantilplott, QQ-plott) for residualar (differansar mellom data og estimert regresjonslinje)