Stat100, vår 2018

**Løsning uke 15**

**Matematikkoppgave**

 = 2 og = 4 (dette ser du uten å bruke lommeregner)

 = 1 + 4 + 9 = 14 = 1 + 16 + 49 = 56

 = (1 + 2 + 3)2 = 62 = 36  = (1 + 4 + 7 )2 = 122 = 144

 = (1-2 +2-2 +3-2)2 = 02 = 0

 = (1-4 + 4-4 + 7-4 +8-4)2 = 02 = 0

Merk at det er alltid slik at  = 0

 = (1-2)2 + (2-2)2 + (3-2)2 = 12 + 02 + 12 = 2

 = (1-4)2 + (4-4)2 + (7-4)2 = 32 + 02 + 32 = 18

.

. Denne finner du ved å summere de to foregående resultatene

(2 + 18) = 20

Merk at totalsnittet = 3, dermed blir = (2 - 3)2 + (4 - 3)2 = 2

SSGruppe =  = 3∙2= 6

SSE = 20 (funnet tidligere)

SSTotal = 6 + 20 = 26

**Oppgave 1**

Modell:

La Xi være poeng for jente nummer i, der i er 1 eller 2.

La Yi være poeng for gutt nummer i der i er 1 eller 2.

Vi antar at alle poengsummer er uavhengige av hverandre.

Vi antar videre at Xi ~ N(1, ) og Yi ~ N(2, ).

Alle 3 parametere er ukjente

1 er gjennomsnittlig poengsum for alle jenter i videregående, mens 2 er tilsvarende for alle gutter.

 er et mål for spredning (standardavvik) i poeng **for alle jenter** i videregående, tilsvarende også for gutter (Merk: Spredning innenfor kjønn)



Siden vi antar felles standardavvik:

. Dermed er .

Estimert kjønnseffekt: 

*Er jenter smartere enn gutter?*

H0: Ingen effekt kjønn: 1 =  2 eller  1 -  2 = 0.

H1: Jenter er smartere enn gutter:  1 >  2 eller  1 -  2 > 0.

Vi finner T =  = 2,23

Sjekker med tabellverdi for eksempel å på 5 % nivå (2 frihetsgrader) 2,92.

Vi kan ikke forkaste H0. Vi har ikke nok data til å påvise at jenter er flinkere enn gutter.

(P-verdien er på 0,077, men dette må du bruke pc for å regne ut.)

Konstruer et 95 % konfidensintervall for kjønnseffekt.

Gitt ved  = 25 ± 4.3\*11,2 = (-23,6; 73,16).

Siden 0 er i intervallet kan vi ikke påstå at det er kjønnseffekt med hensyn på matematiske ferdigheter.

Dersom en tar lærer med i modellen:

|  |  |  |  |
| --- | --- | --- | --- |
| Lærer | Jente (X) | Gutt(Y) | Differanse (D) |
| Åge | 80 | 50 | 30 |
| Finn | 90 | 70 | 20 |

La Di = Xi, - Yi. Da er Di normalfordelt (\mu_d, \sigma_d), der  \mu_d = \mu_1 - \mu_2.

* estimeres med 25,

\sigma_d  estimeres med 7,07 (prøv selv, dette er kvadratrota av 50).

H0: Jenter er ikke flinkere:    \mu_d = 0 H1: Jenter er flinkere:  \mu_d > 0

T=\frac{25}{\frac{7.07}{\sqrt{2}}}=5

Denne må nå sammenlignes med tabellverdi 6,314 (1 frihetsgrad).

Vi kan ikke forkaste H0. Vi har ikke nok data til å påvise at jenter er flinkere enn gutter.

(P-verdien er på 0,063, dette må du bruke PC for å regne ut.)

Selv om T ble større for parvis, var tapet av frihetsgrader ikke nok til å veie opp dette.

**Oppgave 2**

Parvis sammenligning fordi hver eiendom er et par.

Anta X1, , . . .,X6 er Olsens takster på de 6 eiendommene og at Y1, , . . .,Y6 er Hansens takster på de samme 6 eiendommene.

La X-ene være normalfordelt (x, x),

La Y-ene være normalfordelt (y, y)

Standardavviken kan godt være forskjellige, og vi må (selvsagt) anta at Xi og Yi er avhengige. Hvorfor?

Di = Xi - Yi, da vil D1, . , . . ., D6 være normalfordelt med (d, d),

der d = x - y. Vi antar at d er ukjent.

Finner differanser mellom takstmenn for hver leilighet.

0,09 0,02 -0,07 0,07 0,1 0,09

Dette gir en gjennomsnittsdifferanse på 0,06 (60 000 kroner) og et standardavvik på 0,0732

(73 200 kroner).

H0: Ingen forskjell : d = 0 H1: Olsen gir høyere takst. d > 0

T = . Denne er t-fordelt med 5 frihetsgrader under H0.

Dette må sammenlignes med for eksempel t0,05, 5 = 2,015.

Her får vi et problem fordi den er signifikant på sånn cirka 5 % nivå.

Altså påstår vi at Olsens takster ligger over Hansens i snitt.

**b)** La X være antall eiendommer der Olsen gir høyere takster enn Hansen.

Da er X binomisk med n = 30 og p er sannsynligheten for at Olsen gir høyere takst enn Hansen, men der p er ukjent.

H0: Olsen gir ikke høyere takster p ≤ 0,5 H1: Olsen gir høyere takster p > 0,5

Vi estimerer p til X/n = 30/40 = 0,75

Under H0 vil  være nesten standard normalfordelt.

Enten gå løs på p-verdi som er P(>) = P(Z > 3,16) = 0,001.

Dermed er vi helt sikre på at Olsen generelt gir høyere takster enn Hansen.

Kan også løses slik (for de som ikke klarer å forholde seg til P-verdier):

Signifikansnivå 5 % forkast H= hvis  > 1,645.

Siden  = 3.16, forkaster vi H0.

**Oppgave 3 (forbudt å bruke lommeregner eller PC)**

**A**: La Yij være høyde på person j, kjønn i. der i= 1,2. j = 1,2.

Vi antar modell: Yij = i + ij, der ij ~ N(0, ) . Alle observasjoner er uavhengige.

1 er gjennomsnittlig høyde for alle gutter, 2 er gjennomsnittlig høyde for alle jenter.

 er standardavvik i høyde for alle person av samme kjønn.

**B**: Estimater: 

**C**:

DF SS MS F P

kjønn 1 324 324 162 (0,006)

Error 2 4 2

P-verdi kan du bare finne ved datamaskin.

**D**: Kan påvise forskjell fordi F er mye større enn tabellverdier på 5 % Signifikansnivå.

**E:** Det er stor forskjell mellom kjønn og liten forskjell innenfor kjønn.

-------------------------------------------------------

Situasjon 2: Estimater: 

DF SS MS F P

kjønn 1 4 4 0,02 (0,900)

Error 2 400 200

P-verdi kan du bare finne ved datamaskin.

Kan ikke påvise forskjell fordi F er mye mindre enn tabellverdier på 5 % Signifikansnivå.

Her var det liten forskjell mellom kjønn, mens det var stor forskjell innenfor kjønn.

**Oppgave 4.**

Df Sum Sq Mean Sq F value

Gruppe 4 10 2,5 1,875

Residuals 15 20 1,33

Må sammenlignes med tabellverdi, på 5 % nivå er denne 3,055. Kan ikke forkaste en hypotese om manglende gruppeeffekt.

**Oppgave 5**

SSGruppe =  = 3(10 - 11.1)2 + 4(12 - 11.1)2 + 3(11 - 11.1)2 = 6,9

SSE = = (2\*22 + 3\*32 + 2\*22) = 43

Df Sum Sq Mean Sq F value

Gruppe 2 6,9 3,45 0,56

Residuals 7 43 6,14

Må sammenlignes med tabellverdi, på 5 % nivå er denne 4,737. Kan ikke forkaste en hypotese om manglende gruppeeffekt.

**Oppgave 6 (Bruk R-commander)**

Finn gjennomsnitt og standardavvik i hver gruppe.

mean sd

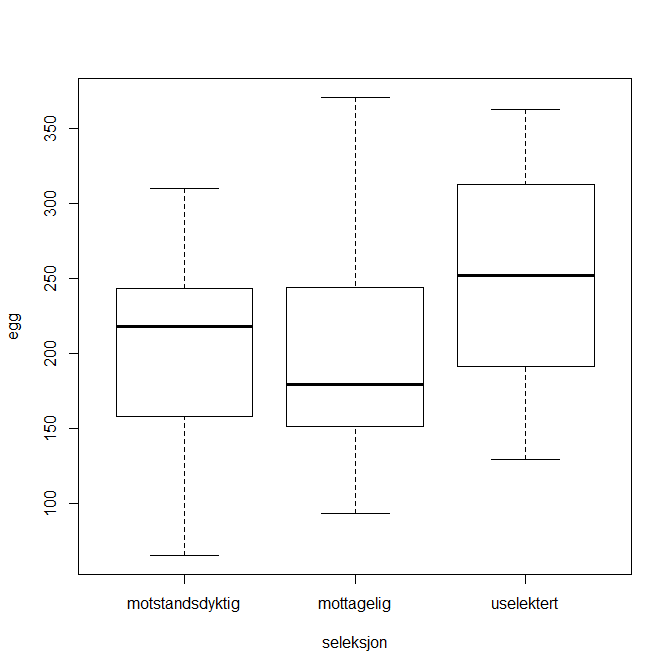
motstandsdyktig 205.48 65.11

mottagelig 194.84 65.69

uselektert 253.08 70.64

Ser det ut til at antagelsen om likt (populasjons) standardavvik er akseptabel?

Estimatene i hver gruppe er ganske like, det taler for at antagelsen ikke skulle være så gal.



Sett opp modellen for datanalysen du vil utføre og estimer alle parametere i modellen.

Yij er antall egg for flue nummer j innen gruppe nummer i.

Yij = i + ij der ij ~N(0, ). Alle observasjoner er uavhengige.

i =1, 2, 3. j =1, 2 . . . . , 25.

Dersom vi nummererer gruppen alfabetisk: 

i: Er gjennomsnittlig fruktbarhet for alle fluer som kunne tenkes å få behandling nr. i

 Er et mål for spredning i fruktbarhet for fluer med identisk behandling.

Kan du påvise signifikant forskjell mellom gruppene med hensyn på fruktbarhet?

Tester: H0: 1 = 2 = 3.

Mot alternativ hypotese: Minst to forventninger er forskjellig.

Df Sum Sq Mean Sq F value Pr(>F)

seleksjon 2 48091 24045 5.326 **0.00697**

Residuals 72 325079 4515

Sikker forskjell. Forkaster nullhypotesen (se P-verdien).

Det er kun 0,7 % sjanse for å registre så store forskjeller med hensyn på seleksjon i data dersom seleksjon ikke skulle ha noen betydning.

**Oppgave 7 (for noen detaljer se hvordan det er svart på oppgave 6)**

Mean SD n

O1 6.625 2.263 8

O2 5.625 1.84 8

O3 7.875 1.552 8

O4 4.625 1.407 8

O5 4.750 2.375 8

O6 6.125 2.167 8

Ser det ut til at antagelsen om likt (populasjons) standardavvik er akseptabel? Ja

Vi antar modellen der alle observasjoner er uavhengige.

Yij er poeng på brød j bakt med oppskrift i. i = 1, 2, 3, 4, 5, 6 j =1, 2, . . . 8.

Vi antar at poengene er uavhengige og at Yij ~ N(i, ).

**Eventuelt: Yij = i + ij der ij ~N(0, ).**

H0: ingen forskjell mellom oppskrifter. D.v.s.

H0: 1 = 2 =3 = 4 =5 = 6 H1. minst to -er er forskjellige.

R Commander gir:

Df Sum Sq Mean Sq F value Pr(>F)

oppskrift 5 59.94 11.987 3.091 0.0183

Residuals 42 162.88 3.878

Denne testen har p-verdi 0,018 (hva betyr det?), og dermed kan H0 forkastes.

Vi er altså sikre på at oppskriftene er forskjellige, men hvor ligger denne forskjellen?