EKSAMEN I: STA100 SANNSYNLIGHETSREGNING OG STATISTIKK

VARIGHET: 4 TIMER DATO: 13 MAI, 2016

HJELPEMIDLER: Godkjent enkel kalkulator (HP30S, Casio FX82, TI-30, Citizen SR-270X , Texas BA II Plus eller HP17bII+ ).

EKSAMEN BESTÅR AV 4 OPPGAVER PÅ 5 SIDER OG 5 SIDER VEDLEGG, TOTALT 10 SIDER.

EMNEANSVARLIG: Jan Terje Kvaløy TELEFON: 51 83 22 55

## Oppgave 1

I denne oppgaven skal vi se på kartlegging av hvor mange studenter som ønsker at forelesningene ved UiS legges om til rene videoforelesninger. La $p$ betegne andelen av studentene ved UiS som støtter en omlegging til rene videoforelesninger. I en spørreundersøkelse skal $n$ tilfeldige studenter spørres om de støtter en omlegging til rene videoforelesninger. La $X$ være antallet av disse som vil svare at de støtter en slik omlegging.

a) Forklar hvorfor $X$ kan antas å være binomisk fordelt.

Dersom $n = 6$ og $p = 0.4$ hva er sannsynligheten for at undersøkelsen (feilaktig) vil vise flertall for omlegging til videoforelesninger? Dvs, hva er $P(X > 3)$?

Dersom $n = 60$ og $p = 0.4$ hva er sannsynligheten for at undersøkelsen (feilaktig) vil vise flertall for omlegging til videoforelesninger? Dvs, hva er $P(X > 30)$?

Forklar kort hvorfor det er logisk at sannsynligheten er lavest i den siste av de to situasjonene over.

I evalueringen av STA100 i vår ble studentene spurt om de støtter en omlegging av forelesningene til rene videoforelesninger. Av de 214 studentene som besvarte undersøkelsen svarte 69 at de støtter en slik omlegging.

b) Anta i første omgang at tallene fra evalueringen i STA100 er representative for hva alle studenter ved UiS mener. Regn ut et estimat og et 90% konfidensintervall for $p$.

Er det rimelig å anta at tallene fra evalueringen i STA100 er representative for hva alle studenter ved UiS mener? Forklar kort hvorfor/hvorfor ikke, og hva dette har å si for praktisk bruk av estimatet og konfidensintervallet.

Oppgave 2

Antall pumper av en bestemt type på en oljeplattform som feiler i løpet av en måned er Poisson-fordelt med forventningsverdi 0.4.

a) Regn ut sannsynligheten for ingen pumper feiler i løpet av en måned.

Regn ut sannsynligheten for at færre enn tre pumper feiler i løpet av en måned.

Regn ut sannsynligheten for at flere enn 6 pumper feiler i løpet av ett år (dvs 12 måneder).

På plattformen har man et reservelager av pumper slik at pumper som feiler umiddelbart kan byttes ut. I starten av hver måned fylles lageret opp slik at man har tre pumper på lager. La $A$ være hendelsen at minst tre pumper feiler i løpet av en måned (dvs reservelageret av pumper går tomt). La $B$ være hendelsen at minst en pumpe feiler i løpet av en måned.

b) Tegn et Venn-diagram som illustrerer sammenhengen mellom hendelsene $A$ og $B$.

Forklar hvorfor $P(A) = 0.008$ og $P(B) = 0.33$.

I de månedene der minst en pumpe feiler, hva er sannsynligheten for at reservelageret av pumper går tomt? Dvs regn ut $P(A|B)$.

Er hendelsene $A$ og $B$ uavhengige? Begrunn svaret.

La $X$ være antallet av de tre pumpene i reservelageret man har i starten av en måned som blir brukt i løpet av måneden. Fordelingen til $X$ blir da som gitt i tabellen under.

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $P(X = x)$ | 0.670 | 0.268 | 0.054 | 0.008 |

Kostnaden for hver pumpe fra reservelageret som må tas i bruk er (i tusen kroner) 15. I de månedene der alle pumpene i reservelageret må tas i bruk før måneden er omme (dvs når $X = 3$) ordnes straks en ekstraordinær transport av nye pumper til plattformen før den ordinære månedlige oppfyllingen av lageret. Dette medfører en ekstrakostnad på 55 som dekker både transport og eventuelle ytterligere pumper som svikter før måneden er omme. Dvs den månedlige kostnaden ved å drifte reservelageret av pumper er 15 for hver pumpe som må tas i bruk i løpet av måneden blant de tre pumpene man har i starten av måneden, pluss en ekstrakostnad på 55 de månedene hvor alle pumpene i lageret tas i bruk. La $W$ betegne den månedlige kostnaden.

c) Regn ut $\mathrm{E}(X)$ og $\mathrm{Var}(X)$ fra tabellen over.

Regn ut $\mathrm{E}(W)$.

## Oppgave 3

I forbindelse med testing og utvikling av batterier måler man funksjonstiden ved ulike standardiserte belastningstester. Fra lang tids testing av en bestemt type batterier under betingelser som tilsvarer bruk i en vanlig type lommelykt vet man at funksjonstiden er normalfordelt med forventning $\mu = 27$ timer og standardavvik $\sigma = 2$ timer. Vi antar at funksjonstidene til ulike batterier er uavhengige av hverandre.

a) Regn ut sannsynligheten for at et batteri av den omtalte typen i den omtalte testen vil ha en funksjonstid på over 30 timer.

Regn ut sannsynligheten for at funksjonstiden til et batteri er mellom 25 og 30 timer.

Regn ut sannsynligheten for at summen av funksjonstidene til fire batterier er større enn 100 timer.

Man har utviklet en ny versjon av batteriene omtalt i starten av oppgavene. Et mål i utviklingen av batteriene var at forventet funksjonstid i testen omtalt innledningsvis skulle være over 30 timer. Man antar at standardavviket fremdeles er 2 timer. I en test av 35 batterier av den nye typen finner man en gjennomsnittlig funksjonstid på 31.1 timer. Basert på denne undersøkelsen ønsker man å fastslå om målet om en forventet funksjonstid på over 30 timer er nådd.

b) Formuler problemstillingen som en hypotesetest.

Utfør testen og forklar hva resultatet betyr i praksis. Bruk 5% signifikansnivå.

Regn ut $p$-verdien for testen.

En helt ny type batterier er utviklet, og for disse er man blant annet interessert i å kartlegge differansen i forventet funksjonstid i to ulike belastningstester. Vi antar som før at funksjonstidene er normalfordelte, men siden dette er helt nye batterier er både forventningsverdi og standardavvik ukjente. La $\mu_X$ begne forventet funksjonstid i den første testen og $\mu_Y$ betegne forventet funksjonstid i den andre testen. I den første testen blir 16 batterier testet og man får en gjennomsnittlig funksjonstid på 187.3 og et utvalgsstandardavvik på 6.4. I den andre testen blir 18 batterier testet og man får der en gjennomsnittlig funksjonstid på 196.0 og et utvalgsstandardavvik på 8.1.

c) Finn et 95% konfidensintervall for $\mu_X - \mu_Y$. Spesifiser eventuelle antagelser du må gjøre i tillegg til antagelsene gjort over.

Bruk konfidensintervallet til å på 5% signifikansnivå utføre testen:

$$H_0: \ \mu_X = \mu_Y \qquad \text{mot} \qquad H_1: \ \mu_X \neq \mu_Y$$

Oppgave 4

Når sement størkner blir det avgitt varme til omgivelsene. For å undersøke hvordan mengden av tre kjemiske forbindelser $x_1$, $x_2$ og $x_3$ som inngår i sementen påvirker avgitt varme, $Y$, (målt i kalorier per gram) er det blitt gjort 20 målinger som er viste i tabellen under.

| Y | x1 | x2 | x3 |
|---|---|---|---|
| 88,5 | 6 | 38 | 14 |
| 74,3 | 1 | 29 | 15 |
| 104,3 | 11 | 56 | 8 |
| 87,6 | 9 | 26 | 12 |
| 92,9 | 7 | 52 | 6 |
| 99,2 | 12 | 35 | 9 |
| 102,7 | 3 | 61 | 17 |
| 77,5 | 1 | 31 | 22 |
| 93,1 | 2 | 54 | 9 |
| 105,9 | 18 | 47 | 14 |
| 83,8 | 1 | 40 | 16 |
| 113,3 | 11 | 66 | 9 |
| 109,4 | 10 | 58 | 8 |
| 108,1 | 8 | 55 | 13 |
| 80,3 | 5 | 31 | 5 |
| 93,3 | 3 | 40 | 15 |
| 104,9 | 15 | 35 | 7 |
| 89,1 | 2 | 50 | 18 |
| 103,1 | 9 | 51 | 10 |
| 88,4 | 10 | 32 | 8 |

Man har først tilpasset regresjonsmodellen

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

der $e \sim N(0, \sigma)$ og vi antar at for ulike målinger så er $e_1, \ldots, e_n$ uavhengige. Når vi bruker Excel til å estimere denne modellen fra dataene over får vi utskriften under.

SAMMENDRAG (UTDATA)

| *Regresjonsstatistikk* | |
|---|---|
| Multippel R | 0,949 |
| R-kvadrat | 0,900 |
| Justert R-kvadrat | 0,882 |
| Standardfeil | 3,89 |
| Observasjoner | 20 |

Variansanalyse

| | *fg* | *SK* | *GK* | *F* | *Signifkans-F* |
|---|---|---|---|---|---|
| Regresjon | 3 | 2181 | 727 | 48 | 0,00000 |
| Residualer | 16 | 242 | 15 | | |
| Totalt | 19 | 2422 | | | |

| | *Koeffisienter* | *Standardfeil* | *t-Stat* | *P-verdi* | *Nederste 95%* | *Øverste 95%* |
|---|---|---|---|---|---|---|
| Skjæringspunkt | 55 | 5,1 | 10,7 | 0,00000 | 44 | 66 |
| x1 | 1,43 | 0,21 | 6,71 | 0,00000 | 0,98 | 1,88 |
| x2 | 0,63 | 0,07 | 8,44 | 0,00000 | 0,47 | 0,79 |
| x3 | 0,15 | 0,23 | 0,63 | 0,54 | -0,34 | 0,64 |

a) Skriv ned den estimerte regresjonsligningen.

I følge den estimerte regresjonsligningen, hvor mye endres forventet avgitt varme dersom $x_1$ øker med 1 og mengden av de øvrige kjemiske forbindelsene forblir uendret?

Har mengden av de tre kjemiske forbindelsene samlet sett innflytelse på avgitt varme? Formuler dette som en hypotesetest, gi resultatet av testen på 5% signifikansnivå og forklar hva resultatet av testen i praksis betyr.

Trenger vi å ha med variabelen $x_3$ i modellen? Formuler dette som en hypotesetest, gi resultatet av testen på 5% signifikansnivå og forklar hva resultatet av testen i praksis betyr.

Man har videre tilpasset en regresjonsmodell uten $x_3$, dvs modellen

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

og en datauskrift fra Excel for denne modellen er gitt under.

SAMMENDRAG (UTDATA)

| *Regresjonsstatistikk* | |
|---|---|
| Multippel R | 0,948 |
| R-kvadrat | 0,898 |
| Justert R-kvadrat | 0,886 |
| Standardfeil | 3,82 |
| Observasjoner | 20 |

Variansanalyse

| | *fg* | *SK* | *GK* | *F* | *Signifkans-F* |
|---|---|---|---|---|---|
| Regresjon | 2 | 2175 | 1087 | 75 | 0,00000 |
| Residualer | 17 | 248 | 15 | | |
| Totalt | 19 | 2422 | | | |

| | *Koeffisienter* | *Standardfeil* | *t-Stat* | *P-verdi* | *Nederste 95%* | *Øverste 95%* |
|---|---|---|---|---|---|---|
| Skjæringspunkt | 57 | 3,45 | 16,6 | 0,00000 | 50 | 65 |
| x1 | 1,359 | 0,179 | 7,61 | 0,00000 | 0,98 | 1,74 |
| x2 | 0,628 | 0,073 | 8,57 | 0,00000 | 0,47 | 0,78 |

b) Sammenlign $R^2$ og $R^2$-justert for de to modellene og kommenter kort hvilken modell som i følge dette ser ut for å være den beste modellen.

Det har vært vanlig å regne at effekten av variabelen $x_1$ er slik at når $x_1$ øker med 1 så øker også forventet avgitt varme, $\mathrm{E}(Y)$, med 1. Noen påstår imidlertid at effekten er større (dvs at forventet avgitt varme øker mer enn 1 når $x_1$ øker med 1). Formuler denne påstanden som en hypotesetest og utfør testen. Bruk 5% signifikansnivå og regresjonsmodellen uten $x_3$ når du utfører testen.