Eksamen i: STK1000 – Innføring i anvendt statistikk

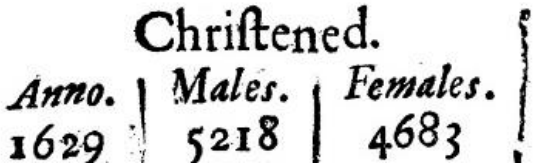
Eksamensdag: Tirsdag 12. desember 2017

Alle deloppgaver teller likt i vurderingen av besvarelsen. Lykke til!

Dette er et løsningsforslag. Studenter som har kommet frem til svaret på andre måter og begrunnet det fornuftig har også fått uttelling. I oppgave 2 vil også t-test gi god uttelling dersom det begrunnes bra og det er konsistens i besvarelsen.

Oppgave 1

John Arbuthnot blir av mange tildelt æren for å ha beregnet verdens første p-verdi. Nullhypotesen hans om at det blir født like mange gutter som jenter ble undersøkt ved å studere dåpsstatistikken i London mellom 1629 og 1710. Han observerte at i 82 år på rad ble det født (og døpt) flere gutter enn jenter. I denne oppgaven skal vi se nærmere på om det var nødvendig med observasjoner over 82 år, eller om John Arbuthnot kunne klart seg med ett år.



Christened.		
Anno.	Males.	Females.
1629	5218	4683

a) Beregn sannsynligheten for å observere flere guttefødsler enn jentefødsler i 82 år på rad, dersom sannsynligheten for å føde en gutt er lik sannsynligheten for å føde en jente.

Under antakelsen om uavhengighet:

$0.5^{82} = 2.07 \cdot 10^{-25} \approx 0.00000000000000000000000002$

La den stokastiske variabelen X være antall guttefødsler i 1629. Hvilken sannsynlighetsfordeling har X?

X er binomisk fordelt, X~bin(n, p), her X~bin(9901, 0.5)

Hvilke tre forutsetninger må være oppfylt for at X skal ha denne fordelingen?

Vi har

1) n = 9901 uavhengige forsøk (observasjoner),

2) hvert med 2 mulige utfall; guttefødsel («Suksess») eller ikke guttefødsel («Fiasko»),

3) og P(«Suksess») = P(guttefødsel)=p er lik for hver av de 9901 fødslene.

Når dette er oppfylt, vil X = Antall suksesser = Antall gutter være binomisk fordelt med parametere n=9901 og p=0.5

b) Ta utgangspunkt i X og parameteren p i fordelingen fra oppgave a). Formuler nullhypotesen og den alternative hypotesen for hvor stor andel gutter som blir født i 1629.

p = andel guttefødsler i 1629

H_0: p = 0.5 (Lik sannsynlighet for gutt som jente)

H_1: $p \neq 0.5$ (Ikke lik sannsynlighet for gutt og jente)

Gi et punktestimat for p, og beregn et 95% konfidensintervall for p.

Fra tabellen: n=5218+4683=9901 fødsler, hvorav 5218 var gutter.

Punktestimat for p:

$$\hat{p} = \frac{5218}{9901} = 0.527 \quad (= 52.7\%)$$

95% konfidensintervall for p er gitt ved:

$$\hat{p} \pm 1.96 \cdot SE(\hat{p}) = \hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

[0.517, 0.537]

Gi en tolkning av konfidensintervallet, og forklar hvilken konklusjon du kommer frem til på hypotesetesten hvis du bruker et signifikansnivå på 0.05.

Intervallet inneholder den sanne p med sikkerhet 95%.

Siden H_0-verdien p=0.5 ikke er i intervallet, forkaster vi H_0 på nivå 0.05.

c) Anta at den sanne andelen guttefødsler er 0.52. Hvor mange fødsler (n) måtte du observere for å få forkastet en tosidig H_0 på nivå 0.05? (Hint: Ta utgangspunkt i nedre grense i et 95% konfidensintervall, som i oppgave b).)

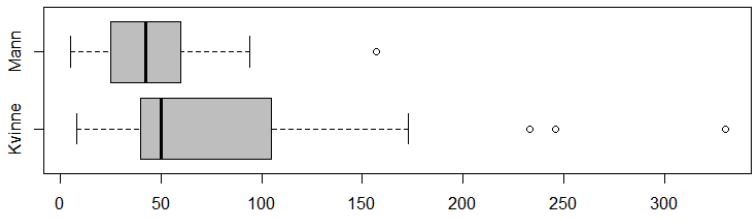
Hvis vi skal forkaste H0 for et tosidig alternativ på nivå 0.05, må den nedre grensen i et 95% konfidensintervall være akkurat på/over 0.05. Vi løser altså ligningen

$$0.50 = 0.52 - 1.96 \cdot \sqrt{\frac{0.52 \cdot (1 - 0.52)}{n}}$$

n=2398

Oppgave 2

Da STK1000-studenter høsten 2016 ble spurt om hvor mange hverdagsklær de har i skapet sitt, fordelte antallet klær seg på 20 menn og 31 kvinner på følgende måte:



Deskriptiv statistikk for de 20 mennene:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    sd
    5.0    27.5    42.5    49.0    59.0   157.0    34.3
```

Deskriptiv statistikk for de 31 kvinnene:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    sd
    8.0    40.0    50.0    85.0   105.0   330.0    75.2
```

Utskriftene viser resultatene fra en to-utvalgs t-test og en Wilcoxon Rank Sum test:

```
> t.test(hverdag~kjonn)

        Welch Two Sample t-test

data:  hverdag by kjonn
t = 2.3216, df = 45.082, p-value = 0.02483
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.780768 67.383748
sample estimates:
mean in group Kvinne   mean in group Mann
            85.03226             48.95000

> wilcox.test(hverdag~kjonn)

        Wilcoxon rank sum test with continuity correction

data:  hverdag by kjonn
W = 399.5, p-value = 0.08544
alternative hypothesis: true location shift is not equal to 0
```

a) Hvilke oppsummeringstall fra den deskriptive statistikken beskriver gruppene best? Begrunn svaret.

Boksplottene (og den deskriptive statistikken) viser at dette er skjevfordelte data. Da oppsummeres gruppene best med median og kvartiler (evt 5-talls-oppsummering):
Menn: median = 43 klesplagg, og 50% av mennene har mellom 28 og 59 klesplagg.
Kvinner: median = 50 klesplagg, og 50% av kvinnene har mellom 40 og 105 klesplagg.

b) Sett opp hypoteser for å teste om det er forskjell på antall klær mellom kjønnene.

H_0: Kvinner og menn har like mange klær
H_1: Kvinner og menn har ikke like mange klær

Forklar hvilken test du velger og hvorfor.

Fordi vi her har skjevfordelte data med outliere i hver gruppe, vil ikke t-test være aktuelt, og når vi har så små antall i gruppene, vil heller ikke CLT (z-test) slå inn, så her bør vi bruke Wilcoxon rank sum test. Hypotesene er derfor ikke formulert med parametere, siden det velges en ikke-parametrisk test.

Velg signifikansnivå 0.05. Hvilken konklusjon trekker du og hvorfor?

Utskriften viser at p-verdien fra WRS test er 0.085, altså større enn 0.05, og H_0 beholdes.

Oppgave 3

Blant ikke-gravide voksne er det funnet en sammenheng mellom høy body mass index (BMI), altså (vekt i kg)/(høyde i m)^2, og høy insulinresistens. Insulinresistensen reflekteres i høye blodsukkerverdier, spesielt etter at man har spist. I en studie av 130 gravide kvinner ønsket man å undersøke sammenhengen mellom kvinnenes BMI før graviditeten, og deres blodsukkernivåer (målt i mmol/l) to timer etter matinntak, i siste halvdel av svangerskapet. Følgende regresjonsanalyse ble gjort:

```
> summary(lm(blodsukker ~ BMI ))

Call:
lm(formula = blodsukker  ~ BMI )

Residuals:
    Min      1Q  Median      3Q     Max
-2.8228 -0.7290 -0.2273  0.7749  3.5811

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.77942    0.88482   3.141  0.00209 **
BMI            0.08225    0.03897   2.111  0.03676 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.169 on 128 degrees of freedom
Multiple R-squared:  0.03363,   Adjusted R-squared:  0.02608
F-statistic: 4.454 on 1 and 128 DF,  p-value: 0.0367
```

a) Hva er effektmålet her?

Analysen som er gjort er en regresjonsanalyse der responsvariabelen y_i (blodsukker i utskriften) relateres til forklaringsvariabelen x_i (BMI i utskriften) og uttrykkes ved regresjonsmodellen

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\quad \varepsilon_i \sim N(0, \sigma)$

Da er regresjonsparameteren β_1 effektmålet, altså tallet som oppsummerer sammenhengen mellom forklaringsvariabelen x_i og responsvariabelen y_i

Hvordan tolkes det?

β_1 viser hvor mange enheters økning i y_i som forventes når x_i øker med en enhet.

Gi et estimat for sammenhengen mellom mors BMI før svangerskapet og hennes blodsukkernivå to timer etter matinntak, i siste halvdel av svangerskapet.

Beregn et 95% konfidensintervall for denne sammenhengen.

Estimat for β_1: $\hat{\beta}_1 = 0.082$

95% konfidensintervall for β_1 er gitt ved:

$\hat{\beta}_1 \pm t_{0.025,n-2} \cdot SE(\hat{\beta}_1) = 0.08225 \pm t_{0.025,128} \cdot 0.03897$

$t_{0.025,128}$ finnes ikke i tabellen i boka, så vi må velge enten $t_{0.025,100}$ eller $t_{0.025,1000}$

qt(0.025,100) # -1.9840 qt(0.025,1000) # -1.9623

[0.005, 0.160] med df=100

[0.006, 0.159] med df=1000

Forskerne mistenkte at vektøkningen i svangerskapet kunne være en viktig faktor i dette. Det ble lagt til grunn forskning fra ikke-gravide, som viste at selv en moderat økning i BMI førte til økt insulinresistens hos personer som ble fulgt opp over tid.
Helseråd som gis til gravide om vektøkning i svangerskapet har som mål å redusere uheldige konsekvenser av overvekt, både for den gravide og barnet hun bærer. Anbefalt vektøkning vil derfor være en konsekvens av kvinnenes BMI før svangerskapet. Overvektige kvinner anbefales en mindre vektoppgang enn normalvektige kvinner.

b) Kan vektøkningen i svangerskapet antas å være en en konfunderende variabel (confounder eller lurking variable) for sammenhengen mellom mors BMI før svangerskapet og hennes blodsukkernivå to timer etter matinntak, sent i svangerskapet?
Nei.
Begrunn svaret.
Her er det grunn til å tro at vektøkning er en konsekvens (respons) av BMI før svangerskapet. Selv om det er grunn til å anta at vektøkningen påvirker blodsukkeret, vil vektøkningen i svangerskapet være en mellomliggende faktor mellom BMI før svangerskapet, og blodsukkerverdier i siste del av svangerskapet. Da er vektøkning en mediator, ikke en confounder.

Kommenter følgende utskrift i lys av det du nettopp svarte.
Det avhenger av problemstillingen om vi skal korrigere for en mediator eller ikke (ta den med i analysen eller ikke).
Hvis vi ikke korrigerer, vil det estimerte effektmålet være den totale effekten av BMI før svangerskapet, på blodsukkerverdier (og noe av denne effekten kan for eksempel skyldes vektøkning). Hvis vi er interessert i den totale effekten av BMI, beholder vi estimatet fra oppgave a).
Hvis vi derimot er interessert i den delen av effekten av BMI som skyldes andre ting enn vektøkning i svangerskapet, velger vi det justerte effektestimatet fra analysen under. Vi ser da at i dette utvalget har det svært lite å si for effektestimatet om vi velger det ujusterte eller justerte estimatet.

```
> summary(lm(blodsukker  ~ BMI + vektokning))

Call:
lm(formula = blodsukker  ~ BMI  + vektokning)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8334 -0.6902 -0.2317  0.8012  3.5537

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.63692    0.97384   2.708  0.00771 **
BMI          0.08268    0.03912   2.113  0.03653 *
vektokning   0.01789    0.05023   0.356  0.72231
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.173 on 127 degrees of freedom
Multiple R-squared:  0.03459,   Adjusted R-squared:  0.01939
F-statistic: 2.275 on 2 and 127 DF,  p-value: 0.1069
```

Oppgave 4

Et av datasettene man finner i R er et datasett med høyde, omkrets og tømmervolumet i 31 felte kirsebærtrær. Omkretsen til trærne (i cm) er målt i brysthøyde, 137 cm over bakken. I denne oppgaven er tømmervolumet kovertert til liter. Man kan gå ut i fra at en liter tilsvarer en vedkubbe. Dersom man plotter tømmervolumet y_i (kalt `vedkubbe` i utskriften) og omkretsen på treet x_i (kalt `omkrets` i utskriften) i et scatterplot, ser man at den kan uttrykkes ved regresjonsmodellen

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma).$$

Gjennomsnitt og standardavvik er $\bar{x} = 105.7$, $sd_x = 25.0$, og $\bar{y} = 854.3$, $sd_y = 465.5$.
Utskriften viser en regresjonsanalyse som ble gjort av disse 31 trærne:

```
> summary(lm(vedkubber~omkrets))

Call:
lm(formula = vedkubber ~ omkrets)

Residuals:
     Min       1Q   Median       3Q      Max
-228.386  -87.972    4.303   98.961  271.468

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1046.1223    95.2903  -10.98 7.62e-12 ***
omkrets        17.9769     0.8779   20.48  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 120.4 on 29 degrees of freedom
Multiple R-squared:  0.9353,    Adjusted R-squared:  0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

a) Hvordan må sammenhengen mellom x og y i scatterplottet se ut for at det skal være fornuftig å bruke denne regresjonsmodellen?

Det må være en lineær sammenheng, og variasjonen rundt linja må være ganske lik i hele observasjonsområdet.

Gi estimater og tolkning av estimatene for parameterne β_0 og β_1.

Estimat for β_0: $\hat{\beta}_0 = -1046$

Estimat for β_1: $\hat{\beta}_1 = 18.0$

β_0 viser hvor regresjonslinja krysser y-aksen. Dette er viktig å ha med for å få en best mulig tilpassing av linja til observasjonene. OBS: Siden observasjonene gjelder trær med en gjennomsnittlig omkrets på 106 cm (sd=25), gir det ikke mening å oppgi dette som «det forventede antall vedkubber for et tre med 0 cm omkrets er -1046».

β_1 viser hvor mange ekstra vedkubber vi kan forvente (18 vedkubber) når omkretsen til treet øker med 1 cm.

b) Beregn et 95% prediksjonsintervall for tømmervolum målt i vedkubber for et tre med en omkrets på 94.2 cm (diameter 30 cm).

95% prediksjonsintervall for y når x=94.2 cm er gitt ved:

$\hat{y} \pm t_{0.025,n-2} \cdot SE(\hat{y})$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 94.2 = -1046.1223 + 17.9769 \cdot 94.2 = 647.3$

$t_{0.025,29} = 2.0452$

$$SE(\hat{y}) = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1) \cdot sd_x^2}} = 120.4 \sqrt{1 + \frac{1}{31} + \frac{(94.2 - 105.7)^2}{30 \cdot 25^2}} = 122.7$$

647.3±2.0452·122.7

[396,898]

c) Utskriften under viser en tilsvarende regresjonsanalyse for trærnes omkrets mot trærnes høyde (målt i meter). Et tilhørende 95% prediksjonsintervall for høyden til et tre med en omkrets på 94.2 cm (diameter 30 cm) er [19.2, 26.2]. Forklar hva et prediksjonsintervall viser.

Det viser i hvilket intervall vi kan forvente å finne høyden (eller tømmervolumet, i oppgave b)) til 95% av trærne med en omkrets på 94.2 cm.

Gi en tolkning av de to prediksjonsintervallene fra b) og c).

Oppgave b): 95 % av trærne med en omkrets på 94.2 cm har et tømmervolum mellom 396 og 898 liter (vedkubber).

Oppgave c): 95 % av trærne med en omkrets på 94.2 cm har en høyde mellom 19.2 og 26.2 meter.

For de 31 kirsebærtrærne: Hva gir omkretsen til treet en mest presis prediksjon av: tømmervolumet eller høyden på treet? Begrunn svaret.

Bredden på prediksjonsintervallet for tømmervolum er (898-396)liter = 502 liter. Det er stor variasjon, men sammenlignet med den store variasjonen i tømmervolum i alle disse 31 trærne ($\bar{y} = 854.3$, $sd_y = 465.5$), er bredden på prediksjonsintervallet ikke stort mer enn et standardavvik.

Andel forklart varians, «r-squared» i modellen fra a) og b) er 0.93.

Det ser tilsynelatende ut til å være et mer presist prediksjonsintervall enn i b), for bredden er 7 meter. Men sammenligner vi med deskriptiv statistikk for høyden,

```
> round(c(mean(trehoydeimeter), sd(trehoydeimeter)),1)
[1] 23.2  1.9
```

Ser vi at prediksjonsintervallet for høyden er mer enn 3 ganger standardavviket til høyden, og det er derfor et intervall med lite presisjon.

Dette ser vi også på at andel forklart varians, «r-squared» i denne modellen er 0.24

Enten vi bruker andel forklart varians eller prediksjonsintervallenes bredder, viser begge at tømmervolumet er det som blir mest presist predikert av omkretsen.

```
> round(c(mean(trehoydeimeter), sd(trehoydeimeter)),1)
[1] 23.2  1.9
> summary(lm(trehoydeimeter~omkrets))

Call:
lm(formula = trehoydeimeter ~ omkrets)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8349 -0.8439  0.0964  0.7537  3.0314

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.90714    1.33603  14.152 1.49e-14 ***
omkrets      0.04027    0.01231   3.272  0.00276 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.688 on 29 degrees of freedom
Multiple R-squared:  0.2697,    Adjusted R-squared:  0.2445
F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```

SLUTT