UiT NORGES ARKTISKE UNIVERSITET

EKSAMENSOPPGAVE

Eksamen i:	STA-0001 Brukerkurs i statistikk 1
Dato:	4. juni 2019
Klokkeslett:	15:00-19:00
Sted:	Hall del 2 Kraft sportssenter
Tillatte hjelpemidler:	Alle trykte og skrevne samt kalkulator
Type innføringsark (rute/linje):	Rute
Antall sider inkl. forside:	**4**
Kontaktperson under eksamen: Telefon/mobil:	**Elinor Ytterstad** **775 44015**
Vil det bli gått oppklaringsrunde i eksamenslokalet? Svar: JA Hvis JA: ca. kl.___1600_______	

NB! Det er ikke tillatt å levere inn kladdepapir som del av eksamensbesvarelsen. Hvis det likevel leveres inn, vil kladdepapiret bli holdt tilbake og ikke bli sendt til sensur.

Postboks 6050 Langnes, N-9037 Tromsø / 77 64 40 00 / postmottak@uit.no / uit.no

Oppgavesettet består av 10 delpunkter som alle teller likt ved bedømming. Det brukes desimalpunktum i dette oppgavesettet.

Oppgave 1

Per produserer bjørkeved i 60 liters sekker.

De fire første sekkene veier (i kg): 36.0, 38.5, 34.0, 35.5.

a)
- Regn ut medianen og gjennomsnittsvekten til vedsekkene.
- Regn ut empirisk standardavvik s og finn et 95% konfidensintervall for standardavviket σ. Avrund til to desimaler.
  Hvilken fordelingsantagelse må gjøres?

Løsningsforslag:
Median av 34, 35.5, 36, 38.5: $\frac{35.5+36}{2} = 35.75$
Gjennomsnitt: $\frac{144}{4} = 36$
Datamateriale:

	x	$x - \bar{x}$	$(x - \bar{x})^2$
	36.0	0	0
	38.5	2.5	6.25
	34.0	-2	4
	35.5	-0.5	0.25
sum	144		10.50

Empirisk varians: $s^2 = \frac{10.50}{4-1} = 3.5$
og da blir standardavviket: $s = \sqrt{3.5} = 1.87$

Må først finne et 95% KI for σ^2 der vi bruker kjikvadratfordeling med $n - 1 = 3$ frihetsgrader. Finner $\chi_{0.025}$ og $\chi_{0.975}$ i tabell E.6:
$N = \frac{(4-1)s^2}{\chi_{0.025}} = \frac{10.5}{9.35} = 1.12$ og $Ø = \frac{(4-1)s^2}{\chi_{0.975}} = \frac{10.5}{0.22} = 47.73.$
Et 95% KI for σ blir da $(\sqrt{1.12}, \quad \sqrt{47.73})$, altså $(1.06, \quad 6.91)$.
Fordelingsantagelse: Y er normalfordelt eller tilnærmet normalfordelt.

Nyhogd ved inneholder mye vann og må tørke før det er egnet som brensel. Per har lagret vedsekkene utendørs i ett år og måler Y = fuktigheten i 9 tilfeldig valgte vedskier. (Det oppgis ikke her hvordan fuktighet måles og heller ikke hva måleenheten for fuktighet er.)

Vi skal anta at $Y \sim N(\mu, \sigma)$ med ukjent σ.
Resultatet av de $n = 9$ målingene ble: $\bar{y} = 20.8$ og $s = 0.90$

b)
- Finn et estimat for forventet fuktighet μ.
- Regn ut et 90 % konfidensintervall for μ.

Løsningsforslag:
Estimert forventet fuktighet μ: $\hat{\mu} = \bar{y} = 20.8$
Vi har normalfordelte data med ukjent standardavvik σ, da bruker vi t-fordelingen med $n - 1 = 8$ frihetsgrader. $N = \bar{y} - t_{0.05} \frac{s}{\sqrt{9}} = 20.8 - 1.860 \cdot \frac{0.90}{3} = 20.24$
Øvre grense: $\text{Ø} = 20.8 + 1.860 \cdot 0.30 = 21.36$
Et 90% KI for μ: $(20.24, \quad 21.36)$

I følge Norsk standard NS4414 skal fuktigheten i salgbar ved ikke overstige 20.

c) Undersøk med hypotesetest om Pers ved har for høy fuktighet.
Bruk 5 % signifikansnivå.
Formulèr passende hypoteser, regn ut testobservator og konkludèr.

Løsningsforslag:
Hypoteser: $H_0 : \mu \leq 20$ mot $H_1 : \mu > 20$
Bruker t-test med 8 frihetsgrader.
Da skal en forkaste H_0 hvis $T > 1.860$
Testobservator, T-test: $T = \frac{20.8-20}{\frac{s}{\sqrt{9}}} = \frac{0.8}{0.30} = 2.67$
Obersvert verdi av testobservator $T > 2.67$ er større enn 1.860 og H_0 forkastes på 5% nivå.
Det er grunnlag i data til å hevde at Pers ved har for høy fuktighet.

d) Forklar hvordan konfidensintervallet i b) kan brukes til å gjennomføre testen i c).
Finn også en omtrentlig p-verdi for testresultatet i c).

Løsningsforslag:
Et 90% konfidensintervall for μ kan brukes direkte i test av tosidig alternativ hypotese på 10% nivå (100-90).
Her har vi imidlertid en ensidig alternativ hypotese: $\mu > 20$, og da forkastes nullhypotesen på 5% nivå (merk endringen fra 10% nivå test til 5% test) hvis hele konfidensintervallet ligger i H_1 området $\mu > 20$. Det er tilfelle her og vi kan konkludere fra 90% konfidensintervallet at nullhypotesen forkastes på 5% nivå.

Eksakt P-verdi er $P(T > 2.67)$. Bruker t-fordelingstabellen for 8 frihetsgrader og finner de to nærmeste verdiene. Det er $P(T > 2.896) = 0.01$ og $P(T > 2.306) = 0.025$.
Dermed er: $0.01 <$ P-verdi < 0.025.

Noe av veden Per produserte ble ikke pakket i sekker, men stablet og lagret samme sted som sekkene.
Fuktigheten (X) i $n = 10$ tilfeldig valgte vedskier av denne veden ble også målt: $\bar{x} = 20.3$, $s_x = 0.93$.

e) Undersøk med hypotesetest om det er grunnlag for å hevde at løs ved (ikke lagret i sekker) tørker bedre enn den i sekker.
Hvilke fordelingsantagelser må du gjøre?
Bruk 5% signifikansnivå.

Løsningsforslag: Forutsetter uavhengige og normalfordelte X og Y med identisk varians σ^2.
Variansen σ^2 er ukjent og estimeres med: $s^2_{pooled} = \frac{(9-1)0.90^2+(10-1)0.93^2}{9+10-2} = 0.839$

Vi gjennomfører en uparet T-test.

$H_0 : \mu_y \leq \mu_x$ mot $H_1 : \mu_y > \mu_x$

Testboservatoren $T = \frac{\bar{Y}-\bar{X}}{s_{pooled}\sqrt{\frac{1}{9}+\frac{1}{10}}}$
er t-fordelt med $9 + 10 - 2 = 17$ frihetsgrader.

Forkaster nullhypotesen hvis $T > 1.740$ (fra tabell E.5).

Regner ut verdien av testobservator: $T = \frac{20.8-20.3}{s_{pooled}\sqrt{\frac{1}{9}+\frac{1}{10}}} = 1.19$.

Konklusjon: Nullhypotesen forkastes ikke på 5% nivå. Det er ikke grunnlag i data for å hevde at løs ved tørker bedre enn den som lagres i sekker.

Alle tall i denne oppgaven er oppdiktet.

Oppgave 2

Lyskilder med LED-teknologi har lange levetider sammenlignet med lyspærer med glødetråder. La T være levetid (i år) for ei LED-lyspære fra en produsent som oppgir at forventet levetiden er 5.7 år (ca 50000 timer).

Vi skal anta at T er eksponentialfordelt med forventningsverdi 5.7 år. Det vil si at

$$f(t) = \frac{1}{5.7}e^{-\frac{t}{5.7}} = 0.1754 \cdot e^{-0.1754 \cdot t} \quad \text{for } t > 0$$

og

$$P(T \leq t) = F(t) = 1 - e^{-0.1754 \cdot t} \quad \text{for } t > 0$$

.

a)
- Bruk kjente egenskaper ved eksponentialfordelingen og finn standardavviket til T.

Regn ut og vis at (avrundet til to desimaler):

- Sannsynligheten for at ei tilfeldig valg LED-pære skal slutte å virke (dø) før det har gått ett år, er 0.16
- $P(T \leq 2) = 0.30$

Løsningsforslag:
Standardavviket til T er identisk med forventningsverdien i denne eksponensialfordelingen, altså 5.7 år.

$P(T \leq 1) = F(1) = 1 - e^{-0.1754 \cdot 1} = 0.16$
$P(T \leq 2) = F(2) = 1 - e^{-0.1754 \cdot 2} = 0.30$

I Annes nye leilighet er det 10 slike LED-lys i stuetaket.
Vi forusetter gjennom hele oppgaven at lysene er tent 24 timer i døgnet.

La X være antall taklys som har sluttet å virket før det har gått 2 år.

b)
- Hvilken fordeling har X, begrunn svaret. Hvilke antagelser må du gjøre?
- Regn ut forventet antall taklys som har sluttet å virke før to år.
- Vis at $P(X \geq 6) = 0.047$.

Løsningsforslag:
Antar lysene virker uavhengig av hverandre, f.eks at det ikke er noe med strømforsyningen som påvirker alle lys. Antar også at alle lys har samme sannsynlighet for å slutte å virke, $p = 0.30$ (hentet fra a)).
Da er X binomisk fordelt med $n = 10$ og $p = 0.30$.
Forventning i binomisk fordeling er: $E(X) = np = 10 \cdot 0.30 = 3$.

Bruker tabell E.1: $P(X \geq 6) = 1 - P(X \leq 5) = 1 - 0.953 = 0.047$.

Anne er skeptisk til den oppgitte levetiden, og vil undersøke med hypotesetest om den er mindre enn 5.7 år. Anne trenger mer data, og avtaler derfor med tre naboer i identiske leiligheter om å la taklysene være tent 24 timer i døgnet. Til sammen har de $n = 40$ LED-lys, og etter ett år viser det seg at $Y = 12$ av taklysene har sluttet å virke.
Her kan du bruke at Y er en binomisk fordelt variabel.

c) Formulér passende hypoteser og gjennomfør testen på 5 % signifikansnivå. Hva blir konklusjonen av testen?

Løsningsforslag:
$Y \sim Bino(40, 0.16)$ der $p = 0.16$ er hentet fra a).
Variansen er da $np(1-p) = 40 \cdot 0.16 \cdot 0.84 = 5.376$, altså større enn 5 og vi kan bruke normaltilnærming.

$H_0 : p \leq 0.16 \quad H_1 : p > 0.16$ (dvs forventet levetid < 5.7 år).

Forkast nullhypotesen dersom $Z \geq 1.645$

Vi gjennomfører testen:

Testobservator: $Z = \frac{12 - 40 \cdot 0.16}{\sqrt{5.376}} = 2.42$

Konklusjon: $Z = 2.42$ er større enn 1.645 og H_0 forkastes på 5 % signifikansnivå. Det er grunnlag for å tro at den oppgitte levetiden er for høy.

Oppgave 3

I forbindelse med en helseundersøkelse måles blodtrykket (Y) til 6 tilfeldig valgte kvinner. Kvinnens alder er også registrert.

Alder x	25	35	45	55	65	75
Blodtrykk y	108	114	124	143	158	166

Det oppgis at: $\sum_{i=1}^{6}(x_i - \bar{x})(y_i - \bar{y}) = 2205$, $\sum_{i=1}^{6}(y_i - \bar{y})^2 = 2843.5$, $\sum_{i=1}^{6} y_i = 813$ og $\sum_{i=1}^{6}(x_i - \bar{x})^2 = 1750$.

a)
- Lag et spredningsplott av datamaterialet.
- Regn ut korrelasjonskoeffisienten r.
- Er det rimelig å anta en lineær sammenheng mellom alder og blodtrykk? Begrunn.

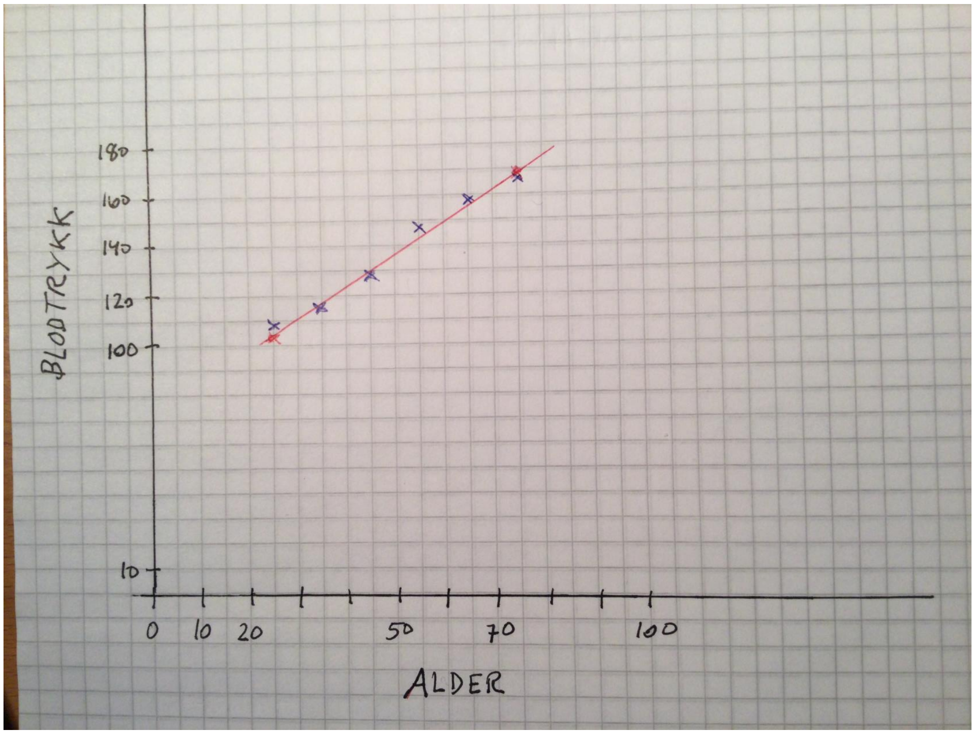
Løsningsforslag:

- Spredningsplott: Se etter b)
- Korrelasjonskoeffisienten $r = \frac{2205}{\sqrt{1750}\sqrt{2843.5}} = 0.9885$.
- Ja. Korrelasjonskoefisienten er veldig nær 1. Det er bare mulig hvis sammenhengen er lineær (rettlinjet)

b)
- Regn ut regresjonslinja og tegn den inn i spredningsplottet i a).
- Hvordan tolker du stigningstallet til linja?

Løsningsforslag:

- $\hat{\beta} = \frac{2205}{1750} = 1.26$
  $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{813}{6} - 1.26 \cdot 50 = 72.5$
  Regresjonslinja er: $\hat{y} = 72.5 + 1.26x$
  Regner ut to punkt på linja og tegner den inn i spredningsplottet.
  $x = 25$ gir $\hat{y} = 104$
  $x = 75$ gir $\hat{y} = 167$
- Blodtrykket hos kvinner generelt øker i gjennomsnitt med 1.26 måleenheter (mmHg) hver år kvinnen eldes.



De to punktene som er beregnet for å plassere regresjonslinja, er markert med rødt.