Fakultet for naturvitenskap og teknologi

UiT
NORGES
ARKTISKE
UNIVERSITET

EKSAMENSOPPGAVE

Eksamen i:	STA-0001 Brukerkurs i statistikk 1
Dato:	Mandag 29. mai 2017
Klokkeslett:	15:00-19:00
Sted:	Åsgårdvegen 9
Tillatte hjelpemidler:	Alle trykte og skrevne, samt kalkulator
Type innføringsark (rute/linje):	Ulinjerte ark
Antall sider inkl. forside:	4
Kontaktperson under eksamen:	Elinor Ytterstad
Telefon/mobil:	77644015

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.

Oppgavesettet består av 15 delpunkter som alle teller likt ved bedømming.

Oppgave 1

Det er utviklet en ny diagnostisk test for en sjelden sykdom som kun forekommer i 0.05 % av befolkningen, Dvs $P(S) = 0.0005$, der $S =$ 'sykdom'. Testen er ikke perfekt, men den gir positivt utslag på 99% av alle som har sykdommen, altså $P(D|S) = 0.99$ der $D =$ positiv diagnostisk test. Friske personer kan også få positivt utslag på testen, det skjer med 3% sannsynlighet.

a) Formulér denne siste opplysningen som en sannsynlighet med D og S.
Løsningsforslag: $P(D|\bar{S}) = 0.03$

b) Vis at $P(D) = 0.03048$. Alle mellomregninger må komme tydelig frem.
Løsningsforslag:
$P(D) = P(D \cap S) + P(D \cap \bar{S}) = P(D|S)P(S) + P(D|\bar{S})P(\bar{S}) = 0.99 \cdot 0.0005 + 0.03 \cdot (1 - 0.0005) = 0.03048$

c) Regn ut sannsynligheten for at en person med positiv test, har sykdommen.
Regn også ut $P(\bar{S}|D)$, og forklar med ord hva dette er sannsynligheten for.

Løsningsforslag:
"at en person med positiv test, har sykdommen" betyr "S gitt D".
$P(S|D) = \frac{P(D|S)P(S)}{P(D)} = \frac{0.99 \cdot 0.0005}{0.03048} = 0.01624$

$P(\bar{S}|D) = \frac{P(D|\bar{S})P(\bar{S})}{P(D)} = \frac{0.03 \cdot (1-0.0005)}{0.03048} = 0.98376$
Dette siste er sannsynlighet for at en person med positiv test, likevel er frisk.

Oppgave 2

a) La X være en normalfordelt variabel med $\mathrm{E}(X) = 2$ og $\mathrm{Var}(X) = 9$.
Regn ut $P(X \leq -2)$ og $P(0 \leq X \leq 4)$

Løsningsforslag:
$P(X \leq -2) = G(\frac{-2-2}{\sqrt{9}}) = G(-1.33) = 0.0918$ og
$P(0 \leq X \leq 4) = P(X \leq 4) - P(X < 0) = G(\frac{4-2}{\sqrt{9}}) - G(\frac{0-2}{\sqrt{9}}) = G(0.67) - G(-0.67) = 0.7486 - 0.2514 = 0.4972$

b) La T være en t-fordelt variabel med 8 frihetsgrader.
Bruk tabellen i boka og finn t_1 og t_2 slik at $P(T > t_1) = 0.005$ og $P(T \leq t_2) = 0.9$

Løsningsforslag:
$P(T > 3.355) = 0.005$ og $P(T \leq 1.397) = 1 - P(T < 1.397) = 1 - 0.1 = 0.9$
Altså $t_1 = 3.355$ og $t_2 = 1.397$

To målemetoder benyttes for å bestemme mengde kalsiumoksid i malm. Begge metoder er brukt på ni malmstykker. Differansene mellom de to metodene for hver malmstykke, ga følgende verdier:

-0.2 0.2 0.3 0.7 0.1 0.6 0.2 0.3 0.5

c) Regn ut gjennomsnitt og finn medianen i datamaterialet.

Løsningsforslag:
Gjennomsnitt: $\bar{x} = \frac{-0.2+0.2+0.3+0.7+0.1+0.6+0.2+0.3+0.5}{9} = \frac{2.7}{9} = 0.3$
Ordnede data: -0.2 0.1 0.2 0.2 0.3 0.3 0.5 0.6 0.7 Median er den midterste (fordi odde antall observasjoner) observerte verdien av de ordnede verdiene: Median = 0.3

d) Vis at $\sum_{i=1}^{9}(x - \bar{x})^2 = 0.60$ og regn ut varians og standardavvik til datamaterialet.

Løsningsforslag:
$\sum_{i=1}^{9}(x-\bar{x})^2 = (-0.2-0.3)^2 + (0.2-0.3)^2 + (0.3-0.3)^2 + (0.7-0.3)^2 + (0.1-0.3)^2 + (0.6-0.3)^2 + (0.2-0.3)^2 + (0.3-0.3)^2 + (0.5-0.3)^2 = (-0.5)^2 + (-0.1)^2 + (0)^2 + (0.4)^2 + (-0.2)^2 + (0.3)^2 + (-0.1)^2 + (0)^2 + (0.2)^2 = 0.60$
$S^2 = \frac{0.60}{9-1} = 0.075$ og dermed er $s = \sqrt{0.075} = 0.274$

Vi skal anta at disse kalsiumoksid-differanseverdier er normalfordelte med forventningsverdi μ.

e) Regn ut et 95% konfidensintervall (KI) for μ.
Regn også ut et 99% KI for μ.
Hvordan ville du forklart andre forskjellen på et 95% og et 99% KI, med maksimalt to setninger?

Løsningsforslag:
Populasjonsvariansen σ^2 er ikke oppgitt og må betraktes som ukjent. Vi bruker derfor t-fordelingen i konfidensintervallet (KI), der s^2 er estimert varians. Den aktuelle t-fordelingen har $n-1 = 9-1 = 8$ frihetsgrader, og $t_{0.025} = 2.306$. Til 99% KI trenger vi $t_{0.005} = 3.355$
Nedre grense=$\bar{X} - 2.306 \cdot \frac{s}{\sqrt{9}} = 0.30 - 2.306 \cdot \frac{0.274}{\sqrt{9}} = 0.3 - 0.211 = 0.089$
Øvre grense=$0.30 + 0.211 = 0.511$
95% KI for μ: (0.089, 0.511)

Tilsvarende for et 99% KI for μ: $0.30 \pm 3.355 \cdot \frac{0.274}{\sqrt{9}} = 0.30 \pm 0.306$
da får i intervallet: $(-0.006,\quad 0.606)$

Et 99% KI er alltid bredere enn et 95% KI. Et 99% KI har 99% sjanse for å dekke den sanne verdien av μ, mens et 95% KI har tilsvarende 95% sjanse.

f) Test om det er forskjell på de to metodenes bestemmelse av mengde kalsiumoksid i malm. Bruk 5% signifikansnivå.
I gjennomføringen av denne testen skal du: Formulére hypoteser, skrive opp testobservator og konkludere.
Løsningsforslag:
Fortsatt ukjent populasjonsvarians σ^2 og vi bruker t-test. Egentlig en paret t-test der X_i er kalsiumokid-differansen mellom de to målemetodene brukt på malmstykke nr i. Denne er også kalt D_i i læreboka kapittel 8.

$H_0 : \mu = 0$ mot $H_0 : \mu \neq 0$
Testobservator $T = \frac{\bar{X}-0}{\frac{S}{\sqrt{9}}}$
Forkast nullhypotesen dersom $T \geq t_{0.025}$ eller $T \leq -t_{0.025}$, der $t_{0.025} = 2.306$ fra t-fordelingen med $9 - 1 = 8$ frihetsgrader.
Vi regner ut $T = \frac{0.3}{\frac{0.274}{\sqrt{9}}} = 3.28$. $T > 2.306$ og dermed forkastes nulhypotesen på 5% nivå .
Vi kunne også brukt 95% KI fra e).
Nullhypotesen forkaste på 5% nivå fordi 95% KI: (0.09, 0.51) ikke inneholder nullhypoteseverdien $\mu = 0$

Oppgave 3

Anta X binomisk fordelt med $p = 0.7$ og $n = 73$.

a) Regn ut forventning og varians til X.
Løsningsforslag:
$\mathrm{E}(X) = np = 73 \cdot 0.7 = 51.1$ og $\mathrm{Var}(X) = np(1-p) = 73 \cdot 0.7 \cdot (1 - 0.7) = 15.33$

b) Forklar forutsetningene og vis med regning hvorfor denne binomiske fordelingen kan tilnærmes med normalfordeling.
Bruk normaltilnærming med heltallskorreksjon og regn ut: $P(X \leq 41)$.
Løsningsforslag:
To betingelser må være oppfylt for bruk av normaltilnærmint til binomisk variabel:
Variansen ≥ 5 og p må ikke være for nær null eller 1. Her er variansen lik 15.33 (altså større enn 5), og $p = 0.7$ som ikke er nær 1, $P(X \leq 41) \approx G(\frac{41+0.5-51.1}{\sqrt{15.33}}) = G(-2.45) = 0.0071$

Blant elever med foreldre med lang høyere utdanning, gjennomførte 78.3 % videregående skole (VGS) på normert tid. Dette er basert på tall fra hele landet. I Troms var det 263 elever med foreldre med lang høyere utdanning, hvorav 182 gjennomførte på normert tid. I Finnmark tilsvarende 73, hvorav 41 gjennomførte på normert tid.

c) Bruk tallene fra Troms og estimér andel p som gjennomfører på normert tid.
Løsningsforslag:
$\hat{p} = \frac{182}{263} = 0.692$

e) Undersøk med hypotesetest om andel elever i Troms som gjennomfører på normert tid er forskjellig fra tilsvarende andel i hele landet.
Bruk 5% signifikansnivå.
Løsningsforslag:
Tosidig test: Hypoteser $H_0 : p = 0.783$ mot $H_1 : p \neq 0.783$
Vi kan bruke normaltilnærnming av samme grunn som i oppgave b).
Med 5% signifikansnivå skal vi forkaste nullhypotesen dersom testobservator $Z \geq 1.96$ eller $Z \leq -1.96$
Testobservator $Z = \frac{182-263\cdot 0.783}{\sqrt{263\cdot 0.783(1-0.783)}} = -3.58$
Denne verdien er mindre enn -1.96 og nullhypotesen forkastes.
Det er grunnlag i data til å hevde at gjennomstrømning av denne gruppe elever er forskjellig i Troms enn i landet forøvrig.

f) Bruk tall fra Finnmark og lag et 95% konfidensintervall for p.
Bruk dette konfidensintervallet og sammenlign Finnmark med hele landet. (Ingen beregninger her.)
Løsningsforslag:
Nedre grense: $\hat{p} - 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{73}} = 0.5616 - 1.96 \cdot 0.0581 = 0.5616 - 0.1138 = 0.4478$
Øvre grense: $0.5616 + 0.1138 = 0.6754$
Dette intervallet (0.4478, 0.6754) er langt fra å dekke den sanne verdien av p for hele landet, som er 0.783.

Nå skal vi inkludere alle VGS elever i Troms

	Normert tid	Mer enn normert tid	Totalt
Foreldre med lang høyere utdanning	182	81	263
Alle andre elever	938	968	1906
Totalt	1120	1049	2169

g) Utfør en kjikvadrattest for å undersøke om det å gjennomføre på normert tid er avhengig av foreldrenes utdanningsnivå .

Løsningsforslag:

Signifikansnivå er ikke oppgitt og vi står fritt til å velge et selv. Her bruker vi $\alpha = 0.05$.

Hypoteser

H_0 : Gjennomføringsgrad og foreldres utdanningsnivå er uavhengig

H_1 : Gjennomføringsgrad er avhengig av foreldres utdanningsnivå.

Vi regner ut forventede tall for hver celle i tabellen:

Forventet antall: $E_1 = 1120 \cdot \frac{263}{2169} = 135.8$, osv.

	Normert tid	Mer enn normert tid	Totalt
Foreldre med lang høyere utdanning	182(135.8)	81(127.2)	263
Alle andre elever	938(984.2)	968(921.8)	1906
Totalt	1120	1049	2169

Testobservator Q er tilnærmet kjikvadratfordelt med $(r-1)(k-1) = 1$ frihetsgrad.

Med 5 % signifikansnivå forkastes nullhypotesen dersom $Q \geq 3.84$.

Finner $Q = \frac{(182-135.8)^2}{135.8} + \frac{(81-127.2)^2}{127.2} + \frac{(938-984.2)^2}{984.2} + \frac{(968-921.8)^2}{921.8} = 36.9$, som er større enn 3.84 og nullhypotesen forkastes.

Kilde: www.ssb.no og data fra perioden 2010-2015