Høgskolen i Telemark

Fakultet for teknologiske fag

Sluttprøve i statistikk

FAG: BY 3212 Statistikk og Fysikk
EE 3512 Statistikk og Fysikk
FB 3112 Statistikk og Fysikk

LÆRER: Kai F. Kristensen

<table>
<tr><td colspan="2">Klasse(r):
2. klasse</td><td>Dato:
16.06.2015</td><td>Tid for sluttprøven, fra - til: 09.00-13.00</td></tr>
<tr><td>Sluttprøven består av følgende:</td><td>Antall sider: 3</td><td>Antall oppgaver: 5</td><td>Antall vedlegg:
Ingen</td></tr>
<tr><td>Tillatte hjelpemidler:</td><td colspan="3">Alle trykte og skrevne. PC og kalkulator.
Ingen kommunikasjon er tillatt. Bruk av internett er ikke tillatt.</td></tr>
<tr><td colspan="4">Kandidaten må selv kontrollere at oppgavesettet er fullstendig.</td></tr>
</table>

Oppgave 1

Anta at det fra en radioaktiv substans i gjennomsnitt registreres stråling med $\lambda = 1.3$ gamma-partikler per millisekund. Antall gamma-partikler X som sendes ut i løpet av et vilkårlig millisekund, antas å være poissonfordelt.

(a) Hvor stor er sannsynligheten for at det fra substansen sendes ut minst fire gamma-partikler i løpet av et gitt millisekund?

Tiden (i millisekunder) mellom hver gang substansen sender ut en gamma-partikkel er eksponensialfordelt.

(b) Hvor stor sannsynlighet er det for at det går høyst ett millisekund mellom to på hverandre følgende partikkel-utsendelser?

Fra en annen radioaktiv substans ble det i løpet av 100 millisekunder registrert utsendt 103 gamma-partikler.

(c) Er gammastrålingen fra denne siste substansen signifikant lavere enn fra den første dersom signifikansnivået er 1%?

Oppgave 2

En funksjon f er gitt ved $f(t) = \begin{cases} 0, & t < 0 \\ kt^4, & 0 \leq t \leq 1 \\ 0, & t > 1. \end{cases}$

(a) Forklar hvorfor vi må ha $k = 5$ for at f skal være en sannsynlighetstetthet.

La X være en stokastisk variabel med f som sannsynlighetstetthet.

(b) Bestem forventningen, medianen og variansen til X.

La $X_1, X_2, \ldots, X_{25}$ være 25 uavhengige stokastiske variable som alle har f som sannsynlighetstetthet.

(c) Bestem $P(X_1 > 0.9)$ og tilnærmet $P(\bar{X} > 0.9)$, der $\bar{X} = (X_1 + \ldots + X_{25})/25$. Kommenter forskjellen i svarene.

Oppgave 3

Anta at det trekkes fem siffer tilfeldig, med tilbakelegging, fra mengden $M = \{0, 1, 2, \ldots, 9\}$. De uttrukne sifrene noteres, etter hvert som de trekkes, fra venstre mot høyre, slik at de til slutt danner en femsifret kombinasjon.

(a) Hva er sannsynligheten for at det finnes nøyaktig én treer i en slik femsifret kombinasjon? I hvor mange femsifrede kombinasjoner finnes det nøyaktig én treer?

Anta nå at det trekkes fem sifre fra mengden M, uten tilbakelegging.

(b) Hvor stor er sannsynligheten for at det trekkes akkurat to oddetall?

Oppgave 4

For å finne ut hvor lang tid en laboratorietekniker bruker på en bestemt arbeidsoppgave, blir tidsforbruket registrert ved 16 anledninger. Vi antar at observasjonene er uavhengige realiseringer av en normalfordelt variabel med forventning μ og standardavvik $\sigma = 1.2$. Det viser seg at $\bar{x} = 7.9$ minutter for de 16 målingene.

(a) Vil det, med et signifikansnivå $\alpha = 0.05$, finnes statistisk bevis for at laboratorieteknikeren i gjennomsnitt bruker mer tid enn 7.5 minutter på arbeidsoppgaven?

(b) Bestem styrken til hypotesetesten i (a), gitt at $\mu = 8.2$.

Oppgave 5

Kapasiteten Y (i prosent) til en bestemt type bilbatteri avtar (øker) når temperaturen x (i $°C$) avtar (øker). Fem tilfeldig valgte, nye batterier av denne typen ble utsatt for fem ulike temperaturer og kapasiteten ble målt. Resultatene er gitt i Tabell 1.

Batteri nr i	**Temperatur** x_i **i** $°C$	**Kapasitet** y_i **i %**
1	-30	45
2	-20	54
3	-10	65
4	0	80
5	10	88

Tabell 1: Data for temperatur og kapasitet av bilbatterier

(a) Vis, ved å benytte formler for konstantledd og stigningstall, at den rette linja som er best tilpasset observasjonene i Tabell 1 etter minste kvadraters metode, kan uttrykkes ved $y = 77.6 + 1.12x$. Gi et estimat på den temperaturen som gir 100% kapasitet på batteriet.

Anta at observasjonene i Tabell 1 er realiseringer av en modell $Y_i = \alpha + \beta x_i + e_i$, der e_i-ene er uavhengige og normalfordelte residualer med forventning $\mu = 0$ og ukjent standardavvik σ.

(b) Gi en fortolkning av parametrene α og β.

(c) Lag et 95% konfidensintervall for β basert på observasjonene i Tabell 1 og forklar hvilken informasjon konfidensintervallet gir.

SLUTT

Løsningsforslag på sluttprøve i statistikk 16.06.2015

Oppgave 1

(a) Hvis X er antall gamma-partikler som sendes ut i løpet av et millisekund, søker vi altså $P(x \geq 4)$. Vi får

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - e^{1.3}\left(\frac{1.3^0}{0!} + \frac{1.3^1}{1!} + \frac{1.3^2}{2!} + \frac{1.3^3}{3!}\right) = 0.043.$$

Maple:

```
with(Statistics):
X:=RandomVariable(Poisson(1.3)):
Probability(X>=4);
```

(b) La T betegne tiden (i millisekunder) mellom to partikkelutsendelser. Da søker vi $P(T \leq 1)$. Parameteren i eksponensialfordelingen blir den samme som i poissonfordelingen, altså $\lambda = 1.3$. Vi får

$$P(T \leq 1) = 1 - e^{-1.3 \cdot 1} = 0.727.$$

Maple:

```
with(Statistics):
T:=RandomVariable(Exponential(1/1.3)):
Probability(T<=4);
```

(c) La Y betegne antall partikler som sendes fra den siste substansen i løpet av en ny periode på 100 millisekunder. Da kan vi anta at Y er poissonfordelt med parameter μ. Siden det spørres om den observerte strålingen er signifikant lavere enn for den første substansen, svarer dette til ensidig testing:
$H_0: \ \mu = 130$ mot $H_1: \ \mu < 130$ på 1% signifikansnivå.
P-verdien for den observerte strålingen på 103 partikler blir derfor

$$\text{P-verdi} = P(Y \leq 103) = 0.008.$$

Maple:

```
with(Statistics):
Y:=RandomVariable(Poisson(130)):
Probability(Y<=103);
```

Siden p-verdien er lavere enn signifikansnivået, er avviket signifikant og vi konkluderer med at strålingen er lavere fra den siste substansen.

En *alternativ test* vil være en z-test fordi poisson-parameteren i løpet av 100 millisekunder er tilstrekkelig høy til at Y blir tilnærmet normalfordelt.

Under nullhypotesen vil Y være tilnærmet $N(130, \sqrt{130})$-fordelt, slik at z-observatoren i dette tilfellet blir

$$z = \frac{103 - 130}{\sqrt{130}} = -2.368.$$

Vi har $z_{0.01} = 2.326$, så siden $-2.368 < -2.326$, kan nullhypotesen forkastes og konklusjonen blir den samme som i testen over. P-verdien blir for øvrig i dette tilfellet $P(Z < -2.368) = 0.009$.

Oppgave 2

(a) Vi må kreve at arealet under sannsynlighetstettheten blir 1. Siden $f(t) = 0$ utenfor intervallet $[0, 1]$, krever vi $\int_0^1 kt^4 dt = 1$. Dette gir $\left[\frac{k}{5}t^5\right]_0^1 = 1$, slik at $k/5 = 1$ og $k = 5$.

Maple:

```
k=solve(int(k*x∧4, x=0..1)=1,k);
```

(b) Hvis vi lar μ, m og σ^2 betegne henholdsvis forventning, median og varians, får vi

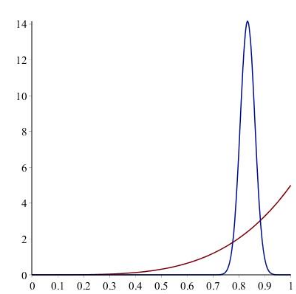
$$\mu = \int_0^1 t \cdot 5t^4 dt = \left[\frac{5}{6}t^6\right]_0^1 = \frac{5}{6} \approx 0.83$$

Videre vil likningen $\int_0^m 5t^4 dt = 0.5$ gi m. Vi får $[t^5]_0^m = 0.5$, slik at $m^5 = 0.5$. Dette gir $m = \sqrt[5]{0.5} = 0.87$.

$$\sigma^2 = E(X^2) - \mu^2 = \int_0^1 t^2 \cdot 5t^4 dx - (5/6)^2 = 5/7 - 25/36 = 5/252 \approx 0.02.$$

(c) $P(X_1 > 0.9) = \int_{0.9}^1 5t^4 dt = [t^5]_{0.9}^1 = 1 - 0.9^5 = 0.41.$
På grunn av sentralgrenseteoremet vil $\bar{X}$ være tilnærmet normalfordelt med forventning 5/6 og standardavvik $\sqrt{5/252}/\sqrt{25} = 1/\sqrt{1260}$. Vi får derfor

$$P(\bar{X} > 0.9) = P\left(\frac{\bar{X} - 5/6}{1/\sqrt{1260}} > \frac{0.9 - 5/6}{1/\sqrt{1260}}\right) \approx P(Z > 2.366) = 0.009$$



For å forklare at den siste sannsynligheten blir lavere enn den første, kan vi se på figuren over. Tettheten til $\bar{X}$ ligger mye mer konsentrert omkring forventningsverdien 5/6 og har dermed en tynnere hale til høyre for 0.9.

Oppgave 3

(a) Hvis X betegner antall treere i en slik femsifret kombinasjon, vil X være binomisk fordelt med $n = 5$ og $p = 1/10$. Vi søker derfor $P(X = 1)$:

$$P(X = 1) = \binom{5}{1}(0.1)^1 \cdot (0.9)^4 = 0.328.$$

Maple:

```
with(Statistics):
X:=RandomVariable(Binomial(5,0.1)):
Probability(X=1);
```

Vi kan også skrive dette svaret som $5 \cdot 9^4/10^5$, slik at antall "gunstige" kombinasjoner, med nøyaktig én treer, blir $5 \cdot 9^4 = 32805$.

(b) Det er fem oddetall og fem partall i mengden det trekkes fra, så hvis Y er antall oddetall blant de fem uttrukne, vil Y være hypergeometrisk fordelt med parametre $N = 10$, $M = 5$ og $n = 5$. Vi får

$$P(Y = 2) = \frac{\binom{5}{2} \cdot \binom{5}{3}}{\binom{10}{5}} = \frac{25}{63} \approx 0.397.$$

Maple:

```
with(Statistics):
Y:=RandomVariable(Hypergeometric(10,5,5)):
Probability(Y=2);
```

Oppgave 4

(a) Vi skal altså teste $H_0 : \ \mu <= 7.5$ mot $H_1 : \ \mu > 7.5$ på 5% signifikansnivå. Vi beregner z-observatoren $z = \frac{7.9-7.5}{1.2/\sqrt{16}} = 1.33$. Siden $z_{0.05} = 1.645$, og altså $1.33 < 1.645$, følger at avviket ikke er signifikant. Testen påviser altså ikke at laboratorieteknikeren bruker signifikant lengre tid enn 7.5 minutter.

(b) Forkastingsgrensen c er $c = 7.5 + 1.645 \cdot 1.2/\sqrt{16} = 7.99$. Teststyrken $\gamma(8.2)$ finner vi derfor ved

$$\gamma(8.2) = P(\bar{X} > 7.99|\mu = 8.2) = P\left(\frac{\bar{X} - 8.2}{1.2/\sqrt{16}} > \frac{7.99 - 8.2}{1.2/\sqrt{16}}\right) = P(Z > -0.7) = 0.76.$$

Oppgave 5

(a) Stigningstallet b og konstantleddet a kan vi finne fra formlene

$$b = \frac{\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{5}(x_i - \bar{x})^2} \quad \text{og} \quad a = \bar{y} - b \cdot \bar{x}.$$

Her er det greit å bruke for eksempel *Maple*:

```
with(Statistics):
X:=[seq(-30+10*i, i=0..4)]:
Y:=[45,54,65,80,88]:
SXY:=add((X[i]-Mean(X))*(Y[i]-Mean(Y)), i=1..5):
SSX:=add((X[i]-Mean(X))∧2, i=1..5):
b:=SXY/SSX;
a=Mean(Y)-b*Mean(X);
```

Dette gir $b = 1.12$ og $a = 77.6$. For å finne temperaturestimatet som svarer til 100% kapasitet, løser vi likningen $77.6 + 1.12x = 100$, og får $x = 20$.

(b) α er den forventede kapasiteten for batteriet ved $0°C$, mens β er den forventede kapasitetsøkningen i prosentpoeng som svarer til en temperaturøkning på $1°C$.

(c) Et $100(1 - \delta)\%$ konfidensintervall for β er gitt på formen

$$\hat{\beta} \pm t_{\delta/2} \cdot \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Siden vi har $\hat{\beta} = b = 1.12$, $\sum_{i=1}^{5}(x_i - \bar{x})^2 = 1000$, $\sum_{i=1}^{5}(y_i - \bar{y})^2 = 11265.2$, $t_{0.025} = 3.182$ (3 frihetsgrader) og

$$s^2 = \frac{1}{5-2}\left(\sum_{i=1}^{5}(y_i - \bar{y})^2 - \hat{\beta}^2 \sum_{i=1}^{5}(x_i - \bar{x})^2\right) = \frac{1}{3}(1265.2 - 1.12^2 \cdot 1000) = 3.6,$$

følger at konfidensintervallets grenser blir $1.12 \pm 3.182 \cdot \sqrt{3.6}/\sqrt{1000}$, slik at konfidensintervallet blir [0.93,1.31].

Dette konfidensintervallet sier oss at vi kan være 95% sikre på at den forventede økningen i batteriets kapasitet ved å øke temperaturen med $1°C$ ligger mellom 0.93 og 1.31 prosentpoeng.

Høgskolen i Telemark

Fakultet for teknologiske fag

Sluttprøve i statistikk

FAG: BY 3212 Statistikk og Fysikk
EE 3512 Statistikk og Fysikk
FB 3112 Statistikk og Fysikk

LÆRER: Kai F. Kristensen

Klasse(r): 2. klasse		**Dato:** 16.12.2015	**Tid for sluttprøven, fra - til:** 09.00-13.00
Sluttprøven består av følgende:	**Antall sider:** 2	**Antall oppgaver:** 5	**Antall vedlegg:** Ingen
Tillatte hjelpemidler:	Alle trykte og skrevne. PC og kalkulator. Ingen kommunikasjon er tillatt. Bruk av internett er ikke tillatt.		
Kandidaten må selv kontrollere at oppgavesettet er fullstendig. Besvarelsen leveres på papir.			

Oppgave 1
I en test av 400 lyspærer av en viss type ble levetiden (i timer) registrert og presentert i følgende tabell:

Levetid	**Frekvens**
$[0, 5500\rangle$	319
$[5500, 11000\rangle$	70
$[11000, 16500\rangle$	10
$[16500, 22000\rangle$	1

(a) Beregn gjennomsnitts- og medianlevetiden til lyspærene, basert på tabelldataene. Kommentér forskjellen på de to svarene.

Tabelldataene er simulert fra en fordeling der levetiden T (som ikke er eksponensialfordelt) har fordelingsfunksjon F, gitt ved

$$F(t) = P(T \leq t) = 1 - e^{-(t/4000)^{1.2}}.$$

(b) Vis at $P(T > 5500) = 0.231$. Hvordan stemmer dette resultatet overens med de simulerte dataene i frekvenstabellen gitt innledningsvis?

(c) Hvor stor sannsynlighet er det for at akkurat to av ti tilfeldig valgte lyspærer av denne typen lever lenger enn 5500 timer?

(d) Hvor stor sannsynlighet er det for at en lyspære lever ytterligere minst 5500 timer hvis den allerede har levd 5500 timer?

Oppgave 2

I en stor bedrift med 2000 ansatte registreres det gjennomsnittlig 13 arbeidsulykker per år, som medfører langvarig fravær. Hvis X er antall slike fremtidige arbeidsulykker per år i bedriften, regner vi X som poissonfordelt med parameter $\lambda = 13$.

(a) Finn sannsynligheten for at det neste år vil bli registrert maksimalt (høyst) ti arbeidsulykker med langvarig fravær.

(b) Løs punkt (a) på nytt, men benytt en passende binomisk variabel.

(c) Løs punkt (a) på nytt, men benytt en passende normalfordelt variabel.

Oppgave 3

(a) La $T = 3X - Y$, der $X \sim N(1, 2)$ og $Y \sim N(2, 3)$ er uavhengige. Regn ut $E(T)$ og $Var(T)$ og bestem sannsynligheten $P(T > 0)$.

(b) La $X \sim N(\mu, 2)$ og $Y \sim N(2\mu, 3)$ være uavhengige, og la μ være ukjent. Finn en forventningsrett estimator for μ på formen $\hat{\mu} = \alpha X + \beta Y$, der α og β velges slik at variansen til $\hat{\mu}$ blir minst mulig?

Oppgave 4

Graden av opplevd postoperativ smerte, målt på en visuell, analog skala fra 0 til 10, antas for en viss type operasjon å være $N(\mu, 2)$-fordelt. Man vet av erfaring at $\mu = 4.3$ for pasientgruppe A, men mistenker at $\mu > 4.3$ for pasientgruppe B. Man fant et gjennomsnitt på 5.2 for åtte tilfeldig valgte personer fra pasientgruppe B.

(a) Finn ut, ved først å beregne p-verdi, om personer fra pasientgruppe B opplever den postoperative smerten sterkere enn de fra pasientgruppe A. Benytt 5% signifikansnivå.

(b) Bestem teststyrken til testen i (a) dersom $\mu = 5.0$. Hva uttrykker denne teststyrken?

Oppgave 5

I tabellen under finnes tall over utviklingen av antall drepte i trafikkulykker i Norge fra 2009 (år 1) til 2014 (år 6).

År (x)	1	2	3	4	5	6
Antall drepte (y)	214	208	168	145	180	147

(a) (i) Finn (den empiriske) korrelasjonskoeffisienten mellom tid (x) og antall drepte i trafikken (y), basert på dataene over. Hvilken informasjon gir fortegnet i denne sammenhengen?

(ii) Finn den rette linja $y = a + bx$ som er best tilpasset dataene etter minste kvadraters metode. Lag en prognose for antall drepte i trafikken i 2015.

La oss anta at tabellverdiene over antall drepte i trafikken er realiseringer fra en lineærnormal modell $Y_i = \alpha + \beta x_i + e_i, \quad (i = 1, \ldots, 6)$, der e_i-ene er $N(0, \sigma)$-fordelte med ukjent σ.

(b) Gir observasjonene støtte for at det har vært en signifikant nedadgående trend når det gjelder antall drepte i trafikken over disse årene? Besvar spørsmålet ved å gjennomføre en passende ensidig T-test på 5% signifikansnivå.

SLUTT

Løsningsforslag på sluttprøve i statistikk, 16. desember 2015

Oppgave 1

(a) Gjennomsnittslevetiden blir

$$\bar{x} = \frac{319 \cdot 2750 + 70 \cdot 8250 + 10 \cdot 13750 + 1 \cdot 19250}{400} = 4028.75.$$

Medianlevetiden blir

$$\tilde{x} = 0 + \frac{200 - 0}{319} \cdot 5500 = 3448.28.$$

Høyrehale-strukturen vil påvirke og trekke opp gjennomsnittslevetiden men ikke medianlevetiden. Derfor blir $\bar{x} > \tilde{x}$.

(b) Vi har $P(T > 5500) = 1 - P(T \leq 5500) = 1 - (1 - e^{-(5500/4000)^{1.2}}) = 0.231$. Fra tabellen ser vi at 81 av de 400 levetidene, altså en andel på ca 20%, lever lenger enn 5500 timer. Dette stemmer greit overens med den beregnede sannsynligheten på ca 23%.

(c) Antall lyspærer X av ti som lever lenger enn 5500 timer, blir da $Bin(10, 0.231)$-fordelt, slik at vi får

$$P(X = 2) = \binom{10}{2} 0.231^2 \cdot (1 - 0.231)^{10-2} = 0.29.$$

Sannsynligheten for at akkurat to av ti tilfeldig valgte lyspærer lever lenger enn 5500 timer, er altså 0.29.

(d) Her søker vi sannsynligheten $P(T \geq 11000 | T \geq 5500)$. Vi benytter definisjonen av betinget sannsynlighet og får

$$P(T \geq 11000 | T \geq 5500) = \frac{P(T \geq 11000 \cap T \geq 5500)}{P(T \geq 5500)} = \frac{P(T \geq 11000)}{P(T \geq 5500)}.$$

Fra fordelingsfunksjonen F finner vi $P(T \geq 11000) = e^{-(11000/4000)^{1.2}} = 0.0345$, slik at

$$P(T \geq 11000 | T \geq 5500) = \frac{0.0345}{0.231} = 0.15.$$

Det er altså ca 15% sjanse for at en lyspære, som allerede har levd i 5500 timer, skal leve i minst 5500 timer til.

Oppgave 2

(a) Sannsynligheten for høyst ti slike ulykker neste år kan vi beregne ved

$$P(X \leq 10) = e^{-13}(\frac{13^0}{0!} + \cdots + \frac{13^10}{10!}) = 0.252.$$

Dette svaret kan vi også lese av direkte i tabell E2 i læreboka. Alternativt kan vi bruke *Maple*:

```
with(Statistics):
X:=RandomVariable(Poisson(13)):
sannsynlighet=evalf(CDF(X,10));
```

(b) Vi kan tenke binomisk ved å anta at hver av de $n = 2000$ ansatte har sannsynligheten $p = 13/2000$ for å bli utsatt for arbeidsulykke. Hvis X nå betegner en binomisk variabel, som teller opp antall ulykker, får vi

$$\begin{aligned} P(X \leq 10) &= \binom{2000}{0}\left(\frac{13}{2000}\right)^0\left(\frac{1987}{2000}\right)^{2000} + \cdots + \binom{2000}{10}\left(\frac{13}{2000}\right)^{10}\left(\frac{1987}{2000}\right)^{2000-10} \\ &= 0.251. \end{aligned}$$

Maple:

```
with(Statistics):
X:=RandomVariable(Binomial(2000,13/2000)):
sannsynlighet=evalf(CDF(X,10));
```

(c) Siden $\lambda = 13 \geq 10$, kan vi tilnærme med normalfordelingen. X vil da være tilnærmet $N(13, \sqrt{13})$-fordelt. Hvis vi benytter heltallkorreksjon, får vi:

$$P(X \leq 10) = P(X \leq 10.5) = P\left(\frac{X-13}{\sqrt{13}} \leq \frac{10.5-13}{\sqrt{13}}\right) \approx P(Z < -0.69) = 0.245.$$

Vi ser at overensstemmelsen er god, selv om vi bruker normaltilnærming.

Oppgave 3

(a) Vi har $E(T) = E(3X - Y) = 3 \cdot 1 - 2 = 1$ og $Var(T) = Var(3X - Y) = 3^2 \cdot 2^2 + (-1)^2 \cdot 3^2 = 45$. Vi har altså at $T = 3X - Y$ er $N(1, \sqrt{45})$-fordelt. Dette gjør at vi finner

$$P(T > 0) = P\left(\frac{T-1}{\sqrt{45}} > \frac{-1}{\sqrt{45}}\right) = P(Z > -0.15) = P(Z < 0.15) = 0.56.$$

(b) Vi har

$$E(\hat{\mu}) = E(\alpha X + \beta Y) = \alpha \cdot \mu + \beta \cdot 2\mu = (\alpha + 2\beta)\mu.$$

For at $\hat{\mu}$ skal bli forventningsrett, må vi kreve $E(\hat{\mu}) = \mu$, noe som fører til at

$$\alpha + 2\beta = 1.$$

Vi har videre

$$Var(\alpha X + \beta Y) = \alpha^2 \cdot 2^2 + \beta^2 \cdot 3^2 = 4\alpha^2 + 9\beta^2.$$

Vi setter så $\alpha = 1 - 2\beta$ inn i uttrykket for variansen og får:

$$Var(\alpha X + \beta Y) = f(\beta) = 4(1 - 2\beta)^2 + 9\beta^2 = 4 - 16\beta + 25\beta^2.$$

Dette blir en parabel med et minimum (siden koeffisienten foran andregradsleddet er positivt), som vi finner ved derivasjon:

$$f'(\beta) = -16 + 50\beta = 0,$$

som gir $\beta = 0.32$, slik at $\alpha = 1 - 2 \cdot 0.32 = 0.36$. Dette viser at

$$\hat{\mu} = 0.36X + 0.32Y$$

er den forventningsrette estimatoren på den gitte formen, som har minst varians.

Oppgave 4

(a) Her er det naturlig å teste $H_0 : \ \mu = 4.3$ mot $H_1 : \ \mu > 4.3$. P-verdien er sannsynligheten for å observere noe som er minst like avvikende fra H_0 som det vi har observert, gitt at H_0 er sann. Vi får altså

$$\text{P-verdi} = P(\bar{X} > 5.2 | \mu = 4.3) = P\left(\frac{\bar{X} - 4.3}{2/\sqrt{8}} > \frac{5.2 - 4.3}{2/\sqrt{8}}\right) = P(Z > 1.27) = 0.10.$$

Siden p-verdien er større enn signifikansnivået på 0.05, beholder vi H_0. Vi har altså ikke tilstrekkelig grunnlag for å kunne si at personer fra pasientgruppe B opplever de postoperative smertene som sterkere enn personer fra pasientgruppe A.

(b) Vi forkaster H_0 og aksepterer H_1 dersom $\bar{x} > \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$. I dette tilfellet betyr det at H_0 forkastes dersom $\bar{x} > 4.3 + 1.645 \cdot \frac{2}{\sqrt{8}}$, altså når $\bar{x} > 5.46$. Teststyrken for $\mu = 5.0$ beregnes ved

$$\gamma(5.0) = P(\bar{X} > 5.46 | \mu = 5.0) = P\left(\frac{\bar{X} - 5.0}{2/\sqrt{8}} > \frac{5.46 - 5.0}{2/\sqrt{8}}\right) = P(Z > 0.65) = 0.26.$$

Dette forteller oss at det er 26% sjanse for at denne testen oppdager at nullhypotesen er feil (og forkaster den) dersom den sanne μ-verdien skulle være 5.0, altså dersom forventet opplevd smerte i pasientgruppe B i virkeligheten er 5.0.

Oppgave 5

(a) (i) Vi finner $\bar{x} = (1 + \cdots + 6)/6 = 3.5$ og $\bar{y} = (214 + \cdots + 147)/6 = 177$. Videre finner vi (for eksempel ved hjelp av Maple)

$$\begin{aligned} SSX &= \sum_i (x_i - 3.5)^2 = 17.5 \\ SSY &= \sum_i (y_i - 177)^2 = 4344 \\ SXY &= \sum_i (x_i - 3.5)(y_i - 177) = -221 \end{aligned}$$

Korrelasjonskoeffisienten r blir dermed

$$r = \frac{SXY}{\sqrt{SSX \cdot SSY}} = -\frac{221}{\sqrt{17.5 \cdot 4344}} = -0.80.$$

Det negative fortegner indikerer en avtagende tendens i datamaterialet, mot et lavere antall dødsulykker ettersom tida går.

(ii) Vi får at stigningstallet b og skjæringskoeffisienten a blir

$$b = \frac{SXY}{SSX} = -12.63 \quad \text{og} \quad a = \bar{y} - b \cdot \bar{x} = 221.2.$$

Den best tilpassede rette linja etter minste kvadraters metode er derfor

$$y = 221.2 - 12.63x.$$

Prognosen for 2015 blir dermed $y(7) = 221.2 - 12.63 \cdot 7 = 132.8 \approx 133$.

(b) Utfra problemstillingen bør vi teste $H_0: \ \beta = 0$ mot $H_1: \ \beta < 0$. Vi har $\hat{\beta} = b = -12.63$. Videre har vi $s = \sqrt{(SSY - b^2 SSX)/4} = 19.704$, slik at t-observatoren blir

$$t = \frac{-12.63 - 0}{19.704/\sqrt{17.5}} = -2.68.$$

Siden vi har en venstresidig T-test på signifikansnivå 5%, må vi sjekke om $t < -t_{0.05} = -2.132$ (6-2=4 frihetsgrader). Vi har at $-2.68 < -2.132$ er sant, og vi forkaster derfor H_0. Dette betyr at tallene for drepte i trafikken gjennom disse årene viser en signifikant avtakende trend på 5% signifikansnivå.

BOKMÅL

Høgskolen i Sørøst-Norge

SLUTTPRØVE I STATISTIKK

<table>
<tr><td>Emnekode:
BY 3215
EE 3515
FB 3115</td><td colspan="2">Emnenavn:
Statistikk og fysikk for bygg
Statistikk og fysikk for elektro
Statistikk og fysikk</td></tr>
<tr><td>Dato:
14.12.2016</td><td>Tid fra/til:
09.00-12.00</td><td>Ant. timer:
3</td></tr>
<tr><td colspan="3">Ansv. faglærer:
Kai F. Kristensen</td></tr>
<tr><td>Campus:
Porsgrunn</td><td colspan="2">Fakultet:
Fakultet for teknologiske fag</td></tr>
<tr><td>Antall oppgaver: 5</td><td>Antall vedlegg: 0</td><td>Ant. sider inkl. forside og vedlegg: 3</td></tr>
<tr><td colspan="3">Tillatte hjelpemidler (jfr. emnebeskrivelse):
Alle trykte og skrevne, kalkulator og bærbar datamaskin</td></tr>
<tr><td colspan="3">Opplysninger om vedlegg:</td></tr>
<tr><td colspan="3">Merknader:
Alle delpunkter teller like mye. En oppgave uten delpunkter teller som ett delpunkt. Manuelle utregninger må vises. Svar fremkommet ved hjelp av datakraft skal dokumenteres ved hjelp av passende kommandoer/prosedyrer. Ingen kommunikasjon er tillatt. Bruk av internett er ikke tillatt.</td></tr>
</table>

Kryss av for type eksamenspapir

Ruter ☐ Linjer ☐

KANDIDATEN MÅ SELV KONTROLLERE AT OPPGAVESETTET ER FULLSTENDIG

Oppgave 1

For et bestemt flyselskap er sannsynligheten for forsinkelse 0.49 på en tilfeldig avgang.

(a) Bestem sannsynligheten for at minst 8 av 12 tilfeldig valgte avganger er forsinket.

Etter å ha gjort tiltak, hevder flyselskapet at sannsynligheten for forsinkelse er blitt lavere. En uavhengig instans tester dette og finner at 20 av 50 tilfeldig valgte avganger er forsinket.

(b) Sett opp en nullhypotese og en mothypotese (alternativ hypotese) for å teste om flyselskapet har rett i at sannsynligheten for forsinkelse er blitt lavere. Sett signifikansnivået til 5%. Hva blir konklusjonen?

Oppgave 2

En mynthandler har 15 gullmynter, 10 norske og 5 svenske, liggende i en monter. La oss tenke oss en situasjon, der innbruddstyver tilfeldig rasker med seg 11 av de 15 myntene. Hva er sannsynligheten for at det finnes akkurat to norske og to svenske mynter igjen?

Oppgave 3

Levetiden T (i år) til en maskindel antas for $t \geq 0$ å ha sannsynlighetstettheten

$$f(t) = 0.01te^{-0.1t}.$$

(a) Finn sannsynligheten for at en maskindel av denne typen lever minst 15 år.

(b) Finn sannsynligheten for at en maskindel, som allerede har fungert i 5 år, skal leve ytterligere minst 15 år.

(c) Finn tilnærmet, ved hjelp av sentralgrenseteoremet, sannsynligheten for at gjennomsnittslevetiden til 25 uavhengige maskindeler, er minst 15 år. Hvorfor er denne sannsynligheten høyere enn den du fant i (a)?

Oppgave 4

La vekten X til en levende, fullvoksen elgokse være normalfordelt med forventning μ og standardavvik $\sigma = 75$ kg (som gjelder for hele oppgaven).

(a) Finn sannsynligheten for at en elgokse veier mellom 400 kg og 520 kg dersom $\mu = 480$.

Oppgave 4 fortsetter på neste side

Oppgave 4 fortsetter

I et bestemt område er normalvekten til en levende, fullvoksen elgokse 480 kg om høsten. Mistanke om tettere bestand har også gitt mistanke om at elgoksene i dette området har fått lavere vekt enn det normale. Registreringer i forbindelse med jakt viste at gjennomsnittsvekten til 12 fullvoksne elgokser var 430 kg.

(b) Sett opp en nullhypotese og en passende mothypotese i en situasjon der man ønsker å finne ut om de voksne elgoksene i området har lavere vekt enn normalt. Avgjør ved hjelp av p-verdi om nullhypotesen kan forkastes på 5% signifikansnivå.

Oppgave 5

Tabell 1 gir en oversikt over gjennomsnittsproduksjon y per turbin (målt i GWh) i 2015 ved fem ulike vindkraftverk, samt årstallet x da vindkraftverket ble igangsatt.

Vindkraftverk	**Igangsettelsesår** (x)	**Produksjon per turbin** (y)
Fjeldskår	1998	1.7
Havøygavlen	2002	5.7
Bessakerfjellet	2008	7.3
Lista	2012	8.1
Raggovidda	2014	13.1

Tabell 1: Gjennomsnittsproduksjon (GWh) per turbin i 2015 og år for igangsetting.

(a) Tegn dataene i Tabell 1 inn i et spredningsdiagram og beregn den empiriske korrelasjonskoeffisienten r. Hvilken informasjon gir r-verdien i dette tilfellet?

Anta at observasjonene i Tabell 1 tilpasses en lineærnormal modell $Y_i = \alpha + \beta x_i + e_i$, der e_i-ene ($i = 1, \ldots, 5$) er uavhengige og $N(0, \sigma)$-fordelte med ukjent σ.

(b) Lag et 95% konfidensintervall for β. Hvorfor er det rimelig å tro at $\beta > 0$?

LØSNINGSFORSLAG TIL SLUTTPRØVE I STATISTIKK
HSN-PORSGRUNN 14. DESEMBER 2016

Oppgave 1

(a) X=antall avganger som er forsinket, må antas å være binomisk fordelt med $n = 12$ og $p = 0.49$. Vi får derfor

$$P(X \geq 8) = \binom{12}{8} 0.49^8 \cdot 0.51^4 + \cdots + \binom{12}{12} 0.49^{12} \cdot 0.51^0 = 0.175.$$

Sannsynligheten for at minst 8 av 12 avganger skal være forsinket, er derfor 0.175.

Maple:

```
with(Statistics):
X:=RandomVariable(Binomial(12,0.49)):
Probability(X>=8);
```

(b) En naturlig nullhypotese, dersom p betegner sannsynligheten for forsinkelse etter tiltakene, bør uttrykke at det ikke er bedring, så vi bruker enten $H_0: p = 0.49$ eller $H_0: p \geq 0.49$. Siden flyselskapet påstår lavere sannsynlighet for forsinkelse, vil $H_1: p < 0.49$. La X være antall forsinkede avganger av 50, gitt at $p = 0.49$ (nullhypotesen stemmer). Vi gjennomfører testingen ved å beregne p-verdien når vi vet at det er observert 20 forsinkelser og alternativet er venstresidig:

$$\text{p-verdi} = P(X \leq 20) = 0.129.$$

Siden signifikansnivået er $\alpha = 0.05$ og siden $0.129 > 0.05$, følger at vi beholder H_0. Vi har derfor ikke tilstrekkelig bevis til å kunne si at sannsynligheten for forsinkelse er blitt lavere, på dette signifikansnivået.

Maple:

```
with(Statistics):
X:=RandomVariable(Binomial(50,0.49)):
p-verdi=Probability(X<=20);
```

Oppgave 2

Vi kan se på de fire gjenværende myntene som et tilfeldig utvalg uten tilbakelegging fra de 15 opprinnelige. Hvis vi lar Y betegne antall norske mynter av de fire, så vil Y være hypergeometrisk fordelt med $N = 15$, $M = 10$ og $n = 4$, og vi søker $P(Y = 2)$. Vi får

$$P(Y = 2) = \frac{\binom{10}{2}\binom{5}{2}}{\binom{15}{4}} = \frac{30}{91} = 0.33$$

Sannsynligheten for at det er to svenske og to norske gullmynter igjen, er altså 0.33. Alternativt kan man komme til samme svaret ved å se på myntene som blir stjålet, som et utvalg.

Maple:

```
with(Statistics):
Y:=RandomVariable(Hypergeometric(15,10,4)):
Probability(Y=2);
```

Oppgave 3

(a) Sannsynligheten $P(T \geq 15)$ for at maskindelen lever i minst 15 år, finnes ved å integrere sannsynlighetstettheten:

$$P(T \geq 15) = \int_{15}^{\infty} 0.01te^{-0.1t}dt = 0.558.$$

Vi kan regne manuelt (delvis integrasjon), eller benytte Maple (eller kalkulator) ved å skrive integralet inn slik det står.

(b) Vi søker her den betingede sannsynligheten $P(T \geq 20|T \geq 5)$ og bruker derfor definisjonsformelen for betinget sannsynlighet:

$$P(T \geq 20|T \geq 5) = \frac{P(T \geq 20 \cap T \geq 5)}{P(T \geq 5)} = \frac{P(T \geq 20)}{P(T \geq 5)} = \frac{\int_{20}^{\infty} 0.01te^{-0.1t}dt}{\int_{5}^{\infty} 0.01te^{-0.1t}dt} = 0.446.$$

Sannsynligheten for at en maskindel som allerede har fungert i 5 år, skal leve i minst 15 nye år, er altså 0.446, som er lavere enn for en ny maskindel (jfr (a)). Vi har altså slitasje.

(c) Vi kan bruke Maple for å finne at $\mu = E(T) = \int_0^{\infty} t \cdot 0.01te^{-0.1t}dt = 20$ (eller vi kan regne manuelt ved delvis integrasjon). Videre finner vi $\sigma = SD(T) = \sqrt{\int_0^{\infty}(t-20)^2 \cdot 0.01te^{-0.1t}dt} = 10\sqrt{2}$. Gjennomsnittslevetiden $\bar{T}$ for 25 maskindeler blir ved sentralgrenseteoremet tilnærmet $N(20, 10\sqrt{2}/\sqrt{25})$-fordelt, altså tilnærmet $N(20, 2\sqrt{2})$-fordelt. Vi søker en tilnærmingsverdi for $P(\bar{T} \geq 15)$ og får ved standardisering

$$P(\bar{T} \geq 15) = P\left(\frac{\bar{T}-20}{2\sqrt{2}} \geq \frac{15-20}{2\sqrt{2}}\right) \approx P(Z \geq -1.77) = P(Z \leq 1.77) = 0.96.$$

Sannsynligheten er derfor tilnærmet 0.96 for at gjennomsnittslevetiden til 25 maskindeler blir minst 15 år.

Maple:

```
with(Statistics):
Snittlevetid:=RandomVariable(Normal(20,2*sqrt(2))):
evalf(Probability(Snittlevetid>=15));
```

Sannsynligheten i (c) er større enn sannsynligheten i (a) fordi sannsynlighetstettheten til $\bar{T}$ ligger mer konsentrert om forventningsverdien 20 enn det sannsynlighetstettheten til T gjør. Det betyr at en større andel av sannsynlighetsmassen vil ligge over 15.

Oppgave 4

(a) Hvis X betegner elgoksevekten, har vi $X \sim N(480, 75)$, og vi skal finne $P(400 \leq X \leq 520)$. Ved hjelp av standardisering får vi:

$$\begin{aligned} P(400 \leq X \leq 520) &= P(\tfrac{400-480}{75} \leq \tfrac{X-480}{75} \leq \tfrac{520-480}{75}) \\ &= P(-1.07 \leq Z \leq 0.53) \\ &= G(0.53) - G(-1.07) = 0.70 - 0.14 = 0.56. \end{aligned}$$

Maple:
```
with(Statistics):
X:=RandomVariable(Normal(480,75)):
evalf(CDF(X,520)-CDF(X,400));
```

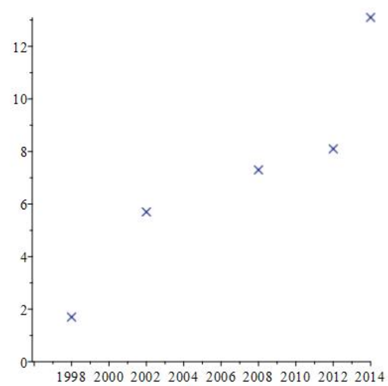
(b) Hvis vi antar at elgoksevekten X er $N(\mu, 75)$-fordelt, vil det være naturlig at nullhypotesen er enten $H_0: \mu = 480$ eller $H_0: \mu \geq 480$. Mothypotesen blir, siden man har mistanke om lavere vekt, $H_1: \mu < 480$. Siden $\bar{X} \sim N(480, 75/\sqrt{12})$ under nullhypotesen, følger at

$$\text{p-verdi} = P(\bar{X} \leq 430) = P\left(\frac{\bar{X} - 480}{75/\sqrt{12}} \leq \frac{430 - 480}{75/\sqrt{12}}\right) = P(Z \leq -2.31) = 0.01.$$

Siden p-verdien er lavere enn signifikansnivået på 0.05, forkastes H_0, og vi konkluderer med at elgoksene i området har lavere vekt enn normalt.

Oppgave 5

(a) Et spredningsplott vil kunne se slik ut:



For å beregne korrelasjonskoeffisienten, trenger vi noen nøkkeltall:
$\bar{x} = (1998 + \cdots + 2014)/5 = 2006.8$, $\bar{y} = (1.7 + \cdots + 13.1)/5 = 7.18$,
$SSX = \sum_{i=1}^{5}(x_i - \bar{x})^2 = 180.8$, $SSY = \sum_{i=1}^{5}(y_i - \bar{y})^2 = 68.128$ og
$SXY = \sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})^2 = 102.88$.
Vi får $r = SXY/\sqrt{SSX \cdot SSY} = 0.93$, som viser at det er en relativt sterk lineær, positiv sammenheng mellom igangsettelsesår og gjennomsnittsproduksjon per vindkraftturbin.

(b) Et 95% konfidensintervall for β er gitt ved

$$\hat{\beta} \pm t_{0.025} \cdot \frac{s}{\sqrt{SSX}},$$

der $t_{0.025}$ er 0.025-kvantilet til T-fordelingen med $5 - 2 = 3$ frihetsgrader. Fra Tabell E5 i boka får får vi $t_{0.025} = 3.182$. Videre har vi formlene

$$\hat{\beta} = \frac{SXY}{SSX} = \frac{102.88}{180.8} = 0.569$$

og

$$s = \sqrt{\frac{1}{5-2}(SSY - \hat{\beta}^2 SSX)} = \sqrt{\frac{1}{3}(68.128 - 0.569^2 \cdot 180.8)} = 1.788.$$

Vi får av dette at konfidensintervallets grenser blir $0.569 \pm 3.182 \cdot \frac{1.788}{\sqrt{180.8}}$, som gir $[0.146, 0.992]$. Siden vi kan være 95% sikre på at β ligger i dette intervallet, kan vi også være rimelig sikre på at $\beta > 0$.

Nøkkeltallene som trenges i beregningene kan finnes ved å definere de aktuelle summene etter å ha lest inn lister med x- og y-observasjoner:

```
with(Statistics):
X:=[1998,2002,2008,2012,2014]:
Y:=[1.7,5.7,7.3,8.1,13.1]:
```