Prøve-eksamen II	MET 1190 Statistikk
Dato	31. mai 2019 kl 1100 - 1400

Alle svar skal begrunnes. Når besvarelsen evalueres, blir det lagt vekt på at framgangsmåte og resultat presenteres så klart, presist og kortfattet som mulig.

Utregninger kan gjøres ved hjelp av godkjent kalkulator. Begrunnelse baseres på teori i kurset, og ikke hvordan man taster på kalkulator. Dersom du trenger noen antagelser for å løse oppgavene, skriv ned hvilke antagelser du gjør.

Oppgave 1.
En bedrift produserer to typer skruer, som vi kaller Type A og Type B. Maskinen som produserer skruer av Type A produserer 2% feilvare (skruer med så store feil at de ikke kan brukes). Maskinen som produserer skruer av Type B er eldre, og produserer 6% feilvare. Skruer av Type A pakkes i esker med 50 skruer, mens skruer av Type B pakkes i esker med 100 skruer.

a) **(6p)** Finn sannsynligheten for at en eske med skruer av Type A inneholder feilvare.

b) **(6p)** Bestem sannsynligheten for at en eske med skruer av Type B inneholder minst 10 skruer som er feilvare.

c) **(6p)** En kunde kjøper en eske med skruer av Type A og en eske med skruer av Type B. Finn sannsynligheten for at kunden får minst to skruer som er feilvare.

Oppgave 2.
For aksjen Norwegian har man observert følgende årlige avkastninger de siste årene:

211.7%	2.2%	-53.0%	160.5%	30.8%	46.8%	17.2%	-11.3%	-38.7%	-0.8%

a) **(6p)** Finn andelen av observasjoner i utvalget som ligger mindre enn ett utvalgsstandardavvik fra gjennomsnittet i utvalget, og framstill dataene i utvalget i et histogram.

b) **(6p)** Bestem et 67% konfidensintervall for forventet årsavkastning. Angi forutsetningene du gjør.

c) **(6p)** Undersøk om forventet årsavkastning er positiv på signifikansnivå 10%.

Oppgave 3.
Tabellen nedenfor viser sammenhørende verdier for to variabler X og Y. Vi skal bruke lineær regresjon med X som forklaringsvariabel og Y som responsvariabel.

X	10	8	13	9	11	14	6	4	12	7	5
Y	804	695	758	881	833	996	724	426	1084	482	568

a) **(6p)** Finn korrelasjonskoeffisienten, og estimer regresjonlikningen.

b) **(6p)** Finn et 96% konfidensintervall for stigningstallet β til regresjonslinjen.

c) **(6p)** Anta at vi vil bruke X som responsvariabel og Y som forklaringsvariabel i stedet. Hvilket fortegn ville stigningstallet til regresjonslinjen få da? Det er ikke nødvendig å estimere den nye regresjonslinjen.

Oppgave 4.
Et datasett består av $n = 100$ observasjoner, som framkommer ved å gjøre n uavhengige trekninger fra en normalfordelt variabel $X \sim N(\mu,\sigma)$ med ukjente parametre μ og σ. Vi skal lage et konfidensintervall for forventningen μ på signifikansnivå α.

a) **(6p)** Konfindensintervallet er gitt ved $\overline{x} \pm t^{n-1}_{\alpha/2} \cdot s/\sqrt{100}$. Forklar hva de ulike symbolene står for. Hvorfor skal man dividere med $\sqrt{100}$ i siste ledd?

b) **(6p)** Hvis et 95% konfidensintervall har bredde 12, hvilket signifikansnivå må vi velge for at intervallbredden skal bli halvparten så stor?

Oppgave 5.
La R_1 og R_2 være avkastningen til Aksje 1 og Aksje 2. Vi antar at R_1 og R_2 er normalfordelte, med $R_i \sim N(\mu_i,\sigma_i)$ for $i = 1,2$, der $(\mu_1, \sigma_1) = (0.12, 0.20)$ og $(\mu_2, \sigma_2) = (0.20, 0.30)$.

a) **(6p)** Finn $p(R_1 > 0.20)$ og $p(R_2 < 0)$.

b) **(6p)** Dersom du investerer like mye i hver av aksjene, vil avkastningen være $R = 0.5R_1 + 0.5R_2$. Hva er sannsynligheten for at denne avkastningen er minst 0.20 dersom R_1 og R_2 er uavhengige?

c) **(6p)** Vi antar nå at R_1 og R_2 ikke nødvendigvis er uavhengige, og kaller kovariansen til R_1 og R_2 for $\gamma = \text{Cov}(R_1,R_2)$. For hvilken verdi av γ er standardavviket til $R = 0.5R_1 + 0.5R_2$ minst mulig?

LØSNING

Oppgave 1.

a) La X være antall skruer i boksen som er feilvare. Da er X binomisk fordelt med $n = 50$ og $p = 0.02$, og vi har

$$p(X \geq 1) = 1 - p(X = 0) = 1 - \binom{50}{0} p^0 (1-p)^{50} = 1 - 0.98^{50} \approx 0.636 \approx 63.6\%$$

b) La Y være antall skruer i boksen som er feilvare. Da er Y binomisk fordelt med $n = 100$ og $p = 0.06$, som gir $E(Y) = np = 100 \cdot 0.06 = 6$ og $\mathrm{Var}(Y) = np(1-p) = 100 \cdot 0.06 \cdot 0.94 = 5.64$. Vi kan bruke normaltilnærming for å regne ut sannsynligheten; det regnes som en god tilnærming ettersom $\mathrm{Var}(Y) > 5$. Vi bruker heltallskorreksjon og finner at

$$p(Y \geq 10) = 1 - p(Y < 9) \approx 1 - G\left(\frac{9 + 0.5 - 6}{\sqrt{5.64}}\right) \approx 1 - 0.930 = 0.070 = 7.0\%$$

hvor G er kumulativ fordelingsfunksjon for standard normalfordeling.

c) Totalt antall skruer som er feilvare er $X + Y$, der X og Y er som tidligere i oppgaven. Vi får da

$$\begin{aligned} p(X+Y \geq 2) &= 1 - p(X+Y \leq 1) = 1 - (p(0,0) + p(1,0) + p(0,1)) \\ &= 1 - \left(0.98^{50} \cdot 0.94^{100} + (50 \cdot 0.02 \cdot 0.98^{49}) \cdot 0.94^{100} + 0.98^{50} \cdot (100 \cdot 0.06 \cdot 0.94^{99})\right) \\ &\approx 1 - (0.00075 + 0.00076 + 0.00478) \approx 1 - 0.0063 \approx 0.994 = 99.4\% \end{aligned}$$

Vi har brukt at $p(x,y) = p(X = x, Y = y) = p(X = x) \cdot p(Y = y)$ siden X og Y er uavhengige.

Oppgave 2.

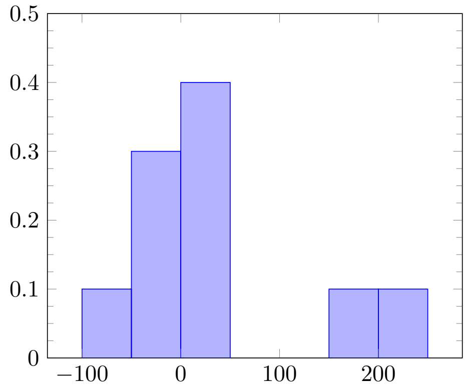
a) Vi lar $x_1, x_2, \ldots, x_n$ betegne observasjoenene i utvalget, med $n = 10$. Da er utvalgsgjennomsnitt $\overline{x}$ gitt ved

$$\overline{x} = \frac{1}{10} \sum_{i=1}^{10} x_i \approx 36.5\%$$

og utvalgsstandardavviket $s = s_x$ gitt ved

$$s^2 = \frac{1}{9} \sum_{i=1}^{10} (x_i - \overline{x})^2 \approx 7233 \quad \Rightarrow \quad s \approx \sqrt{7233} \approx 85.0\%$$

Observasjoner ligger mindre enn ett standardavvik fra gjennomsnittet om de ligger i intervallet $(\overline{x} - s, \overline{x} + s) = (-48.5\%, 121.5\%)$, og $7/10 = 0.7 = 70\%$ av observasjonene ligger i dette intervallet. Histogram med intervallbredde 50% er vist nedenfor.



b) Vi antar at årsavkastningen X til Norwegian er normalfordelt med ukjent forventning μ og varians σ^2, og at utvalget framkommer ved uavhengige trekninger fra X. For å finne et 67% konfidensintervall for μ bruker vi $\alpha = 0.33$ slik at $1 - \alpha = 0.67$. Punktestimatet for μ er $\overline{x} = 36.54\%$, og s brukes som estimat for standardavviket σ, som er ukjent. Konfidensintervallet er gitt ved

$$\overline{x} \pm t^{n-1}_{\alpha/2} \cdot s/\sqrt{n} \approx 36.54 \pm 1.0298 \cdot 85.05/\sqrt{10} \approx 36.54 \pm 27.70$$

siden $t^{n-1}_{\alpha/2} = t^{9}_{0.165} \approx 1.0298$. Dermed blir konfidensintervallet $[8.8\%, 64.2\%]$. Vi bemerker at det er tvilsomt om forutsetningene er oppfyllt.

c) Vi bruker samme forutsetninger som ovenfor. Vi skal undersøke om $\mu > 0$, og velger derfor nullhypotese $H_0 : \mu \leq 0$, og $H_1 : \mu > 0$ som alternativ hypotese. Siden det er en høyresidig T-test med $\alpha = 0.10$, har vi forkastningsområde $T > t^{n-1}_{\alpha} = t^{9}_{0.10}$, som gir $T > 1.383$. Basert på dataene i utvalget, er verdien til testobservatoren gitt ved

$$T = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \quad \Rightarrow \quad t = \frac{36.54 - 0}{85.0/\sqrt{10}} \approx 1.36$$

Vi forkaster dermed ikke H_0 og kan ikke konkludere med at $\mu > 0$. Vi bemerker at det er tvilsomt om forutsetningene er oppfyllt.

Oppgave 3.

a) Vi lar $(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$ betegne observasjoenene i utvalget, med $n = 11$. Korrelasjonskoeffisienten r er da gitt ved

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{11}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{11}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{11}(y_i - \overline{y})^2}} \approx 0.8164$$

Regresjonslinjen er gitt ved $y = \alpha + \beta x + \epsilon$, og punktestimater for α og β er gitt ved

$$\widehat{\beta} = r \cdot \frac{s_y}{s_x} = r \cdot \frac{\sqrt{\sum_{i=1}^{11}(y_i - \overline{y})^2}}{\sqrt{\sum_{i=1}^{11}(x_i - \overline{x})^2}} \approx 0.8164 \cdot \frac{203.16}{3.317} \approx 50.01$$

$$\widehat{\alpha} = \overline{y} - \widehat{\beta} \cdot \overline{x} \approx 750.1 - 50.01 \cdot 9 \approx 300.0$$

Beste estimat for regresjonslinjen er derfor $y = 300 + 50x$.

b) Konfidensintervallet er gitt ved

$$\widehat{\beta} \pm t^{n-2}_{\alpha/2} \cdot \mathrm{SE}(\widehat{\beta})$$

med $\widehat{\beta} \approx 50.01$. Vi har $n = 11$ og $\alpha = 0.04$, og dette gir kritisk verdi $t^{n-2}_{\alpha/2} = t^{9}_{0.02} \approx 2.398$. Standardfeilen er gitt ved formelen

$$\mathrm{SE}(\widehat{\beta})^2 = \frac{\sigma^2}{(n-1)s_x^2} = \frac{\sigma^2}{\sum_{i=1}^{11}(x_i - \overline{x})^2}$$

hvor σ^2 er variansen til feilleddet $\varepsilon \sim N(0,\sigma)$ i den lineære regresjonen. Vi estimerer σ^2 ved hjelp av s^2, gitt ved

$$s^2 = \frac{\sum_{i=1}^{11}(y_i - \overline{y})^2(1 - r^2)}{n - 2}$$

Dermed får vi følgende estimat for $\mathrm{SE}(\widehat{\beta})^2$:

$$\frac{s^2}{\sum_{i=1}^{11}(x_i - \overline{x})^2} = \frac{1}{n-2}\left(\frac{\sum_{i=1}^{11}(y_i - \overline{y})^2(1 - r^2)}{\sum_{i=1}^{11}(x_i - \overline{x})^2}\right) = \frac{s_y^2\,(1 - r^2)}{(n-2)\,s_x^2}$$

$$= \frac{1}{9} \cdot \frac{41273}{11} \cdot (1 - (0.8164)^2) \approx 139.0$$

Dette gir konfindensintervall

$$\widehat{\beta} \pm t^{n-2}_{\alpha/2} \cdot \mathrm{SE}(\widehat{\beta}) \approx 50.01 \pm 2.398 \cdot \sqrt{139.0} \approx 50.0 \pm 28.3$$

Konfidensintervallet for β er derfor $[21.7, 78.3]$.

c) Korrelasjonskoeffisienten ville være uforandret, og siden vi har

$$\widehat{\beta} = r \cdot \frac{s_y}{s_x}$$

der de to standardavvikene er positive, ville stigningstallet til regresjonslinjen ha samme fortegn om vi bytter om rollene for X og Y. Stigningstallet ville altså fortsatt være positivt.

Oppgave 4.

a) Symbolene $\overline{x}$ og s står for utvalgsgjennomsnitt og -standardavvik. De er gitt ved

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{og} \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Symbolet $t^{n-1}_{\alpha_2}$ er definert ved $p(T > t^{n-1}_{\alpha/2}) = \alpha/2$ og kalles kritisk verdi for T-fordelingen. Vi dividerer med $\sqrt{100} = \sqrt{n}$ i siste ledd siden standardavviket til den stokastiske variabelen

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

er gitt ved $\sqrt{n\sigma^2/n^2} = \sigma/\sqrt{n}$, og estimeres ved $s/\sqrt{n}$ når σ er ukjent.

b) Intervallbredden er gitt ved $2t^{n-1}_{\alpha/2} \cdot s/\sqrt{100}$. For å halvere intervallbredden med samme datasett, må vi halvere den kristiske verdien $t^{n-1}_{\alpha/2}$. Vi har at

$$t^{99}_{0.025} \approx 1.9842 \quad \Rightarrow \quad t^{99}_{0.025}/2 \approx 0.9921$$

når $\alpha = 0.05$. Vi må derfor velge et nytt signifikansnivå (som vi for enkelhets skyld også kaller α) slik at

$$t^{99}_{\alpha/2} = 0.9921 \quad \Rightarrow \quad p(T > 0.9921) = \alpha/2$$

Siden $p(T > 0.9921) = 1 - p(T \leq 0.9921) \approx 1 - 0.8382 = 0.1618$ når T har 99 frihetsgrader, må vi velge $\alpha = 2 \cdot 0.1618 \approx 0.324 = 32.4\%$.

Oppgave 5.

a) Siden R_1 og R_2 er normalfordelte, finner vi sannsynligheten ved standardisering. Vi får da

$$p(R_1 > 0.20) = p\left(Z > \frac{0.20 - 0.12}{0.2}\right) = p(Z > 0.4) = 1 - G(0.4) \approx 1 - 0.655 = 0.345 = 34.5\%$$

$$p(R_2 < 0) = p\left(Z < \frac{0 - 0.20}{0.30}\right) = p(Z < -2/3) = G(-2/3) \approx 0.252 = 25.2\%$$

Vi skriver G for den kumulative fordelingsfunksjonen til standard normalfordeling.

b) Avkastningen $R = 0.5R_1 + 0.5R_2$ er normalfordelt $R \sim N(\mu, \sigma)$, og $E(R) = 0.5E(R_1) + 0.5E(R_2)$ gir $\mu = 0.06 + 0.10 = 016$. Dersom R_1 og R_2 er uavhengige, er variansen til R gitt ved

$$\mathrm{Var}(R) = 0.5^2\,\mathrm{Var}(R_1) + 0.5^2\,\mathrm{Var}(R_2) = 0.25 \cdot 0.2^2 + 0.25 \cdot 0.3^2 = 0.0325$$

Dette betyr at $\sigma = \sqrt{\mathrm{Var}(R)} = \sqrt{0.0325} \approx 0.1803$. Dette gir ved standardisering

$$p(R \geq 0.20) = p\left(Z \geq \frac{0.20 - 0.16}{0.1803}\right) \approx 1 - G(0.2218) \approx 0.412 = 41.2\%$$

c) Standardavviket σ til R er minst mulig når $\mathrm{Var}(R) = \sigma^2$ er minst mulig, og

$$\mathrm{Var}(R) = 0.5^2 \cdot \mathrm{Var}(R_1) + \cdot 0.5^2\,\mathrm{Var}(R_2) + 2 \cdot 0.5 \cdot 0.5 \cdot \mathrm{Cov}(R_1,R_2) = 0.0325 + 0.5\gamma$$

Siden korrelasjonskoeffisienten ρ til R_1 og R_2 har verdier i intervallet $[-1, 1]$, og vi har

$$\rho = \frac{\mathrm{Cov}(R_1,R_2)}{\sqrt{\mathrm{Var}(R_1)} \cdot \sqrt{\mathrm{Var}(R_2)}} = \frac{\gamma}{\sigma_1 \cdot \sigma_2} = \frac{\gamma}{0.2 \cdot 0.3} = \frac{\gamma}{0.06}$$

så må kovariansen γ være i intervallet $[-0.06,0.06]$. Det er klart at $\mathrm{Var}(R) = 0.0325 + 0.5 \cdot \gamma$ er minst mulig om $\gamma = -0.06$. Dette gir $\mathrm{Var}(R) = 0.0325 - 0.03 = 0.0025$, og $\sigma = 0.05$.