SENSORVEILEDNING - Skriftlig eksamen

MET 11901
Statistikk

Institutt for Samfunnsøkonomi

Utlevering:	05.06.2019	Kl. 09:00
Innlevering:	05.06.2019	Kl. 12:00

For mer informasjon om formalia, se eksamensoppgaven.

Løsning	MET 1190 Statistikk
Dato	5. juni 2019 kl 0900 - 1200

Oppgave 1.

a) La X være antall riktigs svar for Student A. Da er X binomisk fordelt med $n = 90$ og $p = 1/3$, og $E(X) = np = 90 \cdot 1/3 = 30$.

b) Vi har at $\text{Var}(X) = np(1-p) = 90 \cdot 1/3 \cdot 2/3 = 20$. Siden $\text{Var}(X) > 5$, er X tilnærmet normalfordelt, med fordeling $N(\mu,\sigma)$ der $\mu = E(X) = 30$ og $\sigma = \sqrt{\text{Var}(X)} = \sqrt{20}$. Vi bruker normaltilnærming med heltallskorreksjon for å regne ut sannsynligheten:

$$p(X \geq 30) = 1 - p(X \leq 29) \approx 1 - G\left(\frac{29 + 0.5 - 30}{\sqrt{20}}\right) \approx 0.54 = 54\%$$

Dersom man bruker normaltilnærming uten heltallskorreksjon, får man en $p(X \geq 30) \approx 59\%$. Det er en dårligere tilnærming, siden den eksakte sannsynligheten er $p(X \geq 30) \approx 0.540 = 54.0\%$ med tre gjeldende siffer.

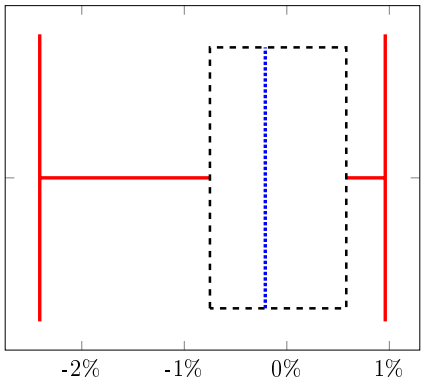
c) La R være hendelsen at Student B velger riktig svar på et tilfeldig spørsmål, og la B være hendelsen at han bommer når han velger hvilke alternativer han skal utelukke. Da er $p(B) = 0.40$, og vi får at

$$p(R) = p(R|B) \cdot p(B) + p(R|B^C) \cdot p(B^C) = 0 \cdot 0.40 + 0.50 \cdot 0.60 = 0.30$$

Siden $p(R) = 0.30 < 1/3$, har Student B mindre sannsynlighet for å velge riktig svar på hvert spørsmål enn Student A, og vi forventer derfor at Student A vil gjøre det bedre enn Student B.

Oppgave 2.

a) Når datapunktene er ordnet i stigende rekkefølge $x_1 < x_2 < \cdots < x_{11}$ er median gitt som x_6 og øvre og nedre kvartil som x_9 og x_3. Dermed er median -0.21%, øvre kvartil er 0.58%, nedre kvartil er -0.75%, og kvartilsbredden er $0.58\% - (-0.75\%) = 1.33\%$. Boksplott er vist nedenfor.



b) For å finne et 87% konfidensintervall for forventet dagsavkastning μ bruker vi $\alpha = 0.13$ slik at $1 - \alpha = 0.87$. Punktestimatet for μ er

$$\overline{x} = \frac{1}{11}\sum_{i=1}^{11} x_i \approx -0.2927$$

Standardavviket σ er ukjent, og vi estimerer det ved å bruke utvalgets standardavvik s, gitt ved

$$s^2 = \frac{1}{10}\sum_{i=1}^{11}(x_i - \overline{x})^2 \approx 1.0548 \quad \Rightarrow \quad s \approx \sqrt{1.0548} \approx 1.027$$

Vi antar at dagsavkastningen for S&P 500 er normalfordelt, og at vi kan regne utvalget som tilfeldig slik at $x_1, \ldots, x_{11}$ kan regnes som uavhengige trekninger. Da er konfidensintervallet gitt ved

$$\overline{x} \pm t_{\alpha/2}^{n-1} \cdot s/\sqrt{n} \approx -0.2927 \pm 1.6498 \cdot 1.027/\sqrt{11} \approx -0.2927 \pm 0.5109$$

siden $t_{\alpha/2}^{n-1} = t_{0.065}^{10} \approx 1.6498$. Dermed blir konfidensintervallet $[-0.80\%, 0.22\%]$.

c) Aksjeindeksen S&P 500 er satt sammen av 500 store selskaper, og man kan derfor tenke seg at endringer i indeken er drevet av summen av svært mange på kort sikt tilfeldige og uavhengige faktorer. Dette støtter antagelsen om at dagsavkastningen er tilnærmet normalfordelt. Utvalget er langt fra tilfeldig, siden det er gjort for noen få dager i løpet av en kort periode, og dette taler imot antagelsen om uavhengige trekninger.

Oppgave 3.

a) Korrelasjonskoeffisienten r er gitt ved

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{6}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{6}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{6}(y_i - \overline{y})^2}} \approx -0.9808$$

Det er en sterk negativ sammenheng mellom X og Y siden $r < 0$ og r er nær -1 ($r = -1$ svarer til en eksakt lineær negativ sammenheng $Y = \alpha + \beta X$ med $\beta < 0$).

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

$$\widehat{\beta} = r \cdot \frac{s_y}{s_x} = r \cdot \frac{\sqrt{\sum_{i=1}^{6}(y_i - \overline{y})^2}}{\sqrt{\sum_{i=1}^{6}(x_i - \overline{x})^2}} \approx -0.9808 \cdot \frac{18.4147}{12.6399} \approx -1.4290$$

$$\widehat{\alpha} = \overline{y} - \widehat{\beta} \cdot \overline{x} \approx 50.50 - 1.4290 \cdot 37.83 \approx 104.56$$

Beste estimat for regresjonslinjen er derfor $y = 105 - 1.43x$.

c) En sammenheng mellom X og Y svarer til at stigningstallet β til regresjonslinjen oppfyller $\beta \neq 0$. Vi gjør derfor en hypotesetest med nullhypotese $\mathrm{H}_0 : \beta = 0$ og alternativ hypotese $\mathrm{H}_1 : \beta \neq 0$. Vi bruker testobservatoren

$$T = \frac{\widehat{\beta} - \beta_0}{\mathrm{SE}(\widehat{\beta})} = \frac{\widehat{\beta}}{\mathrm{SE}(\widehat{\beta})}$$

som er T-fordelt med $n-2 = 4$ frihetsgrader om nullhypotesen $\beta = 0$ er oppfyllt. Hypotesetesten er tosidig, dermed blir forkastningsområdet

$$|T| > t_{\alpha/2}^{n-2} = t_{0.005}^{4} \approx 4.604$$

Standardfeilen til $\widehat{\beta}$ er gitt ved formelen

$$\mathrm{SE}(\widehat{\beta})^2 = \frac{\sigma^2}{(n-1)s_x^2} = \frac{\sigma^2}{\sum_{i=1}^{6}(x_i - \overline{x})^2}$$

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

$$s^2 = \frac{\sum_{i=1}^{6}(y_i - \overline{y})^2(1 - r^2)}{n-2}$$

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

$$\frac{s^2}{\sum_{i=1}^{6}(x_i - \overline{x})^2} = \frac{1}{n-2}\left(\frac{\sum_{i=1}^{6}(y_i - \overline{y})^2(1 - r^2)}{\sum_{i=1}^{6}(x_i - \overline{x})^2}\right) = \frac{s_y^2\,(1 - r^2)}{(n-2)\,s_x^2}$$

$$= 0.25 \cdot \frac{339.1}{159.77} \cdot (1 - (-0.9808)^2) \approx 0.0201$$

Dette gir $T = -1.43/\sqrt{0.0201} \approx -10.1$. Ettersom denne verdien ligger i forkastningsområdet, forkaster vi nullhypotesen. Vår konklusjon er at det er sammenheng mellom X og Y.

Oppgave 4.

a) En god estimator for μ er $\overline{X}$, og vi kan benytte $\overline{X}$ som testobservator. Vi kan også bruke funksjoner av $\overline{X}$, og det vanligste valget er å bruke

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

som testobservator når σ er ukjent. Dersom $\mu = \mu_0$ er den t-fordelt med $n-1$ frihetsgrader. Forkastningsområdet er $T < -t_\alpha^{n-1}$.

b) Vår beslutningsregel basert på p-verdien er:

$$\begin{cases} p < \alpha: & \text{Vi forkaster nullhypotesen } \mathrm{H}_0 \\ p \geq \alpha: & \text{Vi beholder nullhypotesen } \mathrm{H}_0 \end{cases}$$

Dette er fordi en p-verdi med $p < \alpha$ svarer til at realisert verdi t av testobservatoren T er mer ekstrem enn grenseverdien $-t_\alpha^{n-1}$, altså at $t < -t_\alpha^{n-1}$.

Oppgave 5.

a) La $Z_1 = (X_1 - 0.10)/0.12$ være standardiseringen av X_1, som er standard normalfordelt. Da er

$$p(0.05 \leq X_1 \leq 0.15) = p\left(\frac{0.05 - 0.10}{0.12} \leq Z_1 \leq \frac{0.15 - 0.10}{0.12}\right) = p(-5/12 \leq Z_1 \leq 5/12)$$

Vi skriver $G(z)$ for den kumulative fordelingsfunksjonen til standard normalfordelingen. Da blir sannsynlighet ovenfor gitt ved

$$p(0.05 \leq X_1 \leq 0.15) = G(5/12) - G(-5/12) \approx G(0.4167) - G(-0.4167) \approx 0.3231 \approx 0.32$$

b) Vi har at $\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X,Y)$ for alle konstanter a,b og alle variabler X,Y. Det gir

$$\mathrm{Var}(U) = \mathrm{Var}(0.4X_1 + 0.6X_2) = 0.4^2 \cdot 0.12^2 + 0.6^2 \cdot 0.16^2 + 2 \cdot 0.4 \cdot 0.6 \cdot 0.015 = 0.01872$$

Det betyr at standardavviket til U er gitt ved

$$\sigma_U = \sqrt{\mathrm{Var}(U)} = \sqrt{0.01872} \approx 0.1368 \approx 0.14$$

c) Vi har at V er normalfordelt med $E(V) = 0.4 \cdot 0.10 + 0.2 \cdot 0.18 + 0.4 \cdot 0.16 = 0.108$ og varians

$$\begin{aligned}\mathrm{Var}(V) = \mathrm{Var}(0.4X_1 + 0.2X_2 + 0.4X_3) &= 0.4^2 \cdot 0.12^2 + 0.2^2 \cdot 0.16^2 + 0.4^2 \cdot 0.07^2 \\ &+ 2 \cdot 0.4 \cdot 0.2 \cdot 0.015 + 2 \cdot 0.2 \cdot 0.4 \cdot (-0.01) = 0.004912\end{aligned}$$

Dermed er standardavviket til V gitt ved $\sigma_V = \sqrt{0.004912} \approx 0.0701$, og $V \sim N(0.108, 0.0701)$. Dette gir

$$p(V < 0) = p\left(Z < \frac{0 - 0.108}{0.0701}\right) = p(Z < -1.5410) = G(-1.5410) \approx 0.0617 \approx 0.062$$

der $Z = (V - 0.108)/0.0701$ er standariseringen av V, som er standard normalfordelt.