SENSORVEILEDNING - Skriftlig eksamen

MET 11901
Statistikk

Institutt for Samfunnsøkonomi

Utlevering:	18.12.2019	Kl. 09:00
Innlevering:	18.12.2019	Kl. 12:00

For mer informasjon om formalia, se eksamensoppgaven.

Løsning	MET 1190 Statistikk
Dato	18. desember 2019 kl 0900 - 1200

Oppgave 1.

a) Når X er uniformt fordelt på intervallet $[-3,3]$, har vi at

$$p(-1 \le X \le 1) = \int_{-1}^{1} f(x)\,\mathrm{d}x = \int_{-1}^{1} \frac{1}{6}\,\mathrm{d}x = \left[\frac{1}{6}x\right]_{-1}^{1} = \frac{1}{6} - \left(-\frac{1}{6}\right) = \frac{2}{6} = \frac{1}{3}$$

Det er også mulig å argumentere for dette geometrisk, siden integralet ovenfor kan tolkes som arealet av et rektangel med høyde $1/6$ og bredde $1-(-1) = 2$, slik at arealet blir $2/6 = 1/3$.

b) Vi har at $\mathrm{Var}(X) = E(X^2) - E(X)^2$, og vi regner ut forventingsverdiene ved integrasjon. Når X er uniformt fordelt på intervallet $[-3,3]$, gir dette

$$E(X) = \int_{-3}^{3} x\,f(x)\,\mathrm{d}x = \int_{-3}^{3} \frac{1}{6}x\,\mathrm{d}x = \left[\frac{1}{6}\cdot\frac{x^2}{2}\right]_{-3}^{3} = \frac{3^2}{12} - \frac{(-3)^2}{12} = \frac{9}{12} - \frac{9}{12} = 0$$

og

$$E(X^2) = \int_{-3}^{3} x^2\,f(x)\,\mathrm{d}x = \int_{-3}^{3} \frac{1}{6}x^2\,\mathrm{d}x = \left[\frac{1}{6}\cdot\frac{x^3}{3}\right]_{-3}^{3} = \frac{3^3}{18} - \frac{(-3)^3}{18} = \frac{3}{2} - \left(-\frac{3}{2}\right) = 3$$

Dermed er $\mathrm{Var}(X) = E(X^2) - E(X)^2 = 3 - 0^2 = 3$.

c) Vi har at X er en sum av $n = 50$ uavhengige og identisk fordelte stokastiske variabler, og X er derfor tilnærmet normalfordelt. Vi har at $E(X) = 50 \cdot E(X_i) = 50 \cdot 0 = 0$ og at

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_{50}) = 50 \cdot 3 = 150$$

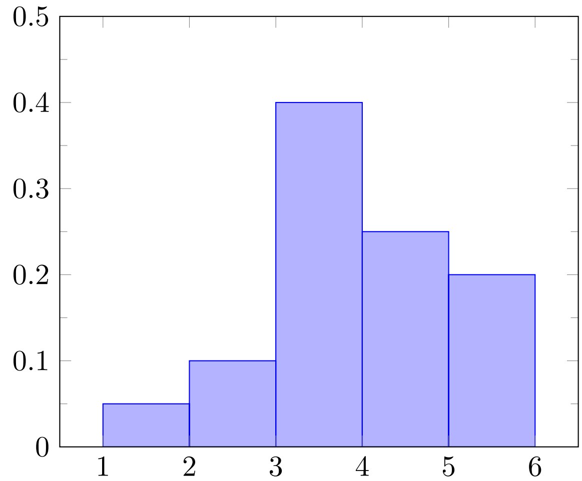
siden variablene er uavhengige. Vi kan derfor regne X som tilnærmet normalfordelt med $\mu = 0$ og $\sigma = \sqrt{150} \approx 12.25$, og at $Z = X/\sigma$ er tilnærmet standard normalfordelt. Det følger at

$$p(-1 \le X \le 1) \approx p(-1/\sqrt{150} \le Z \le 1/\sqrt{150}) \approx G(0.0816) - G(-0.0816) \approx 0.065$$

hvor G er den kumulative fordelingsfunksjonen til standard normalfordelingen.

Oppgave 2.

a) Vi bruker intervallene $[1,2)$, $[2,3)$, $[3,4)$, $[4,5)$, og $[5,6)$, og finner da følgende histogram:



b) Vi skriver $x_1, x_2, \ldots, x_n$ med $n = 20$ for dataene i utvalget. Da er gjennomsnitt $\overline{x}$ og standardavvik $s = s_x$ i utvalget gitt ved

$$\overline{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = 3.92 \quad \text{og} \quad s^2 = \frac{1}{19}\sum_{i=1}^{20}(x_i - \overline{x})^2 \approx 1.1396 \quad \Rightarrow \quad s \approx \sqrt{1.1396} \approx 1.068$$

En gjennomsnittskarakter er mindre enn ett standardavvik fra gjennomsnittet om den er i intervallet $(\overline{x} - s, \overline{x} + s) \approx (2.85, 4.99)$. Dette gjelder $13/20 = 65\%$ av studentene i utvalget.

c) Vi antar at gjennomsnittskarakteren er normalfordelt, og at studentene i utvalget er trukket tilfeldig. Da er et 88% konfidensintervall gitt ved

$$\overline{x} \pm t_{\alpha/2}^{n-1} \cdot s/\sqrt{n} \approx 3.92 \pm 1.6280 \cdot 1.068/\sqrt{20} \approx 3.92 \pm 0.389$$

med $\alpha = 1 - 0.88 = 0.12$ og $t_{\alpha/2}^{n-1} = t_{0.06}^{19} \approx 1.6280$. Dermed blir konfidensintervallet $[3.53, 4.31]$.

Oppgave 3.

a) Korrelasjonskoeffisienten r er gitt ved

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{9}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{9}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{9}(y_i - \overline{y})^2}} \approx 0.99608 \approx 0.996$$

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

$$\widehat{\beta} = r \cdot \frac{s_y}{s_x} = r \cdot \frac{\sqrt{\sum_{i=1}^{9}(y_i - \overline{y})^2}}{\sqrt{\sum_{i=1}^{9}(x_i - \overline{x})^2}} \approx -0.99608 \cdot \frac{1303.2}{55.371} \approx 23.444$$

$$\widehat{\alpha} = \overline{y} - \widehat{\beta} \cdot \overline{x} \approx 2184.7 - 23.444 \cdot 103.33 \approx -237.85$$

Beste estimat for regresjonslinjen er derfor $y = -237.8 + 23.43\,x$.

b) En positiv sammenheng mellom X og Y svarer til at stigningstallet β til regresjonslinjen oppfyller $\beta > 0$. Vi gjør derfor en hypotesetest med nullhypotese $\mathrm{H}_0 : \beta \leq 0$ og alternativ hypotese $\mathrm{H}_1 : \beta > 0$. 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 $n - 2 = 7$ frihetsgrader om $\beta = 0$ er oppfyllt. Hypotesetesten er ensidig, dermed blir forkastningsområdet

$$T > t_{\alpha}^{n-2} = t_{0.02}^{7} \approx 2.517$$

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

$$\mathrm{SE}(\widehat{\beta})^2 = \frac{\sigma^2}{(n-1)s_x^2} = \frac{\sigma^2}{\sum_{i=1}^{9}(x_i - \overline{x})^2}$$

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

$$s^2 = \frac{\sum_{i=1}^{9}(y_i - \overline{y})^2(1 - r^2)}{n-2}$$

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

$$\frac{s^2}{\sum_{i=1}^{9}(x_i - \overline{x})^2} = \frac{1}{n-2}\left(\frac{\sum_{i=1}^{9}(y_i - \overline{y})^2(1 - r^2)}{\sum_{i=1}^{9}(x_i - \overline{x})^2}\right) = \frac{s_y^2\,(1 - r^2)}{(n-2)\,s_x^2}$$

$$= \frac{1}{7}\frac{1303.2^2}{55.371^2} \cdot (1 - 0.99608^2) \approx 0.61881$$

Dette gir $T \approx 23.444/\sqrt{0.61881} \approx 29.80$. Ettersom denne verdien ligger i forkastningsområdet, forkaster vi nullhypotesen. Vår konklusjon er at det er positiv sammenheng mellom X og Y.

c) Standard forutsetninger for en lineær regresjonsmodell er at Y er en stokastisk variabel slik at for en gitt verdi av x, så er $Y = \alpha + \beta\,x + \epsilon$, der feilleddet epsilon er normalfordelt $\epsilon \sim N(0,\sigma)$ med konstant standardavvik σ. Hvis vi kaller feilleddet når $x = x_i$ for ϵ_i, så antar vi også at $\epsilon_1, \epsilon_2, \ldots$ er uavhengige stokastiske variabler.

Oppgave 4.

a) Vi har at $E(R) = (E(R_1) + E(R_2) + E(R_3))/3 = (0.05 + 0.12 + 0.22)/3 = 0.13$.

b) Vi har at $\mathrm{Var}(R) = (1/3)^2 \,\mathrm{Var}(R_1 + R_2 + R_3)$, og dermed får vi at

$$\begin{aligned}\mathrm{Var}(R) &= \frac{1}{9}\left(\mathrm{Var}(R_1) + \mathrm{Var}(R_2) + \mathrm{Var}(R_3) + 2\,\mathrm{Cov}(R_1,R_2) + 2\,\mathrm{Cov}(R_1,R_3) + 2\,\mathrm{Cov}(R_2,R_3)\right)\\ &= \frac{1}{9}\left(0.02^2 + 0.14^2 + 0.15^2 + 2(-0.0025)\right) \approx 0.00417\end{aligned}$$

Standardavviket til R blir dermed $\sigma_R = \sqrt{\mathrm{Var}(R)} \approx \sqrt{0.00417} \approx 0.065$.

c) Dersom R er normalfordelt, er $Z = (R - \mu_R)/\sigma_R \approx (R - 0.13)/0.065$ standard normalfordelt. Vi har da at

$$p(R < 0) = p(Z < \frac{0 - \mu_R}{\sigma_R}) \approx p(Z < \frac{-0.13}{0.065}) \approx G(-2.014) \approx 0.022$$

Oppgave 5.

a) Formen til forkastningsområdet er $|T| > k$ siden nullhypotesen er $H_0 : \mu = 0$ og alternativhypotesen er $H_1 : \mu \neq 0$. Forkastningsområdet er definert av $p(|T| > k|\mu = 0) = 0.03$. Dette gir forkastningsområde

$$|T| > t^{42}_{0.03/2} = 2.246$$

b) Når p-verdien er $p = 0.0245$, betyr det at observert T-verdi ville vært kritisk verdi om testen hadde blitt gjennomført med signifikansnivå $\alpha = 0.0245$. Vi har at $t^{42}_{0.0245/2} = 2.333$. Det betyr at $|T| = 2.333$, og altså at

$$T = -2.333 \quad \text{eller} \quad T = 2.333$$