Løsning	MET 1190 Statistikk
Dato	24. juni 2020 kl 0900 - 1300

Oppgave 1.

a) Vi har at $E(\widehat{p}) = E(X/n) = E(X)/n = np/n = p$, derfor er estimatoren forventningsrett. Den har standardfeil gitt ved

$$\text{Var}(\widehat{p}) = \text{Var}(X/n) = \frac{\text{Var}(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \quad \Rightarrow \quad \text{SE}(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

b) Vi utfører en hypotesetest for andelen p av bilag med avvik, og bruker hypotesene $H_0 : p \leq 0.10$ og $H_1 : p > 0.10$. Siden n er forholdsvis stor, er test-observatoren

$$Z = \frac{\widehat{p} - p_0}{\text{SE}(\widehat{p})}$$

tilnærmet normalfordelt når $p = p_0 = 0.10$. Siden $\widehat{p} = 20/180 \approx 0.111$, er realisert z-verdi gitt ved

$$z = \frac{0.111 - 0.10}{\sqrt{0.10 \cdot 0.90/180}} \approx 0.497$$

Forkastningsområdet er høyresidig siden $H_1 : p > 0.10$, og vi finner p-verdien til hypotesetesten ved å regne ut sannsynligheten for en minst like ekstrem Z-verdi som den realiserte. Den er gitt ved

$$p = p(Z \geq 0.497) = 1 - p(Z \leq 0.479) = 1 - 0.690 = 0.31 = 31\%$$

Dette er en stor p-verdi. Basert på undersøkelsen, er det derfor ikke grunn til å hevde at mer enn 10% av bilagene har avvik om vi bruker vanlige signifikansnivåer som $\alpha = 5\%$.

Oppgave 2.

a) Vi bruker formelen $\overline{x} \pm t^{n-1}_{\alpha/2} \cdot s/\sqrt{n}$ for å finne konfidensintervallet, med $\overline{x} = 32.75$, $s = s_x \approx 7.605$, $n = 12$, og $t^{11}_{0.04} \approx 1.928$. Dette gir konfindensintervallet

$$\overline{x} \pm t^{n-1}_{\alpha/2} \cdot s/\sqrt{n} = 32.75 \pm 1.928 \cdot 7.605/\sqrt{12} \approx 32.75 \pm 4.23$$

Vi kan skrive konfindensintervallet $[28.52, 36.98]$.

b) Vi bruker en T-test med $H_0 : \mu \geq 36$ og $H_1 : \mu < 36$. Når $\mu = \mu_0 = 36$, så følger testobservatoren

$$T = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

en T-fordeling med $n - 1 = 11$ frihetsgrader. Forkastningsområdet blir venstresidig, og er gitt ved $T < -t^{11}_{0.10} = -1.363$. Vi finner realisert verdi av testobservatoren:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{32.75 - 36}{7.605/\sqrt{12}} \approx -1.480$$

Siden realisert t-verdi er i forkastningsområdet, forkaster vi H_0. Det er altså grunn for å hevde at $\mu < 36$ basert på undersøkelsen.

c) I så fall blir konfidensintervallet $\overline{x} \pm 2 = 32.75 \pm 2$, som vi kan skrive $[30.75, 34.75]$. Fra formelen ovenfor, ser vi at intervallbreden er gitt ved uttrykket $2t^{n-1}_{\alpha/2} \cdot s/\sqrt{n}$. Vi får derfor

$$2t^{n-1}_{\alpha/2} \cdot s/\sqrt{n} = 4 \quad \Rightarrow \quad t^{11}_{\alpha/2} = \frac{4\sqrt{12}}{2s} = \frac{2\sqrt{12}}{7.605} \approx 0.911$$

Dette betyr at $\alpha/2 = p(T > 0.911) \approx 0.191$, og dermed er $\alpha \approx 2 \cdot 0.191 \approx 38\%$. Det vil si at $[30.75, 34.75]$ er et 62% konfidensintervall.

d) La oss kalle de $n = 12$ datapunktene for $x_1, \dots, x_n$. Da har vi at $\sum x_i = n\overline{x} = 12 \cdot 32.75 = 393$. Ved å summere de ti kjente datapunktene, får vi

$$342 + a + b = 393 \quad \Rightarrow \quad a + b = 51$$

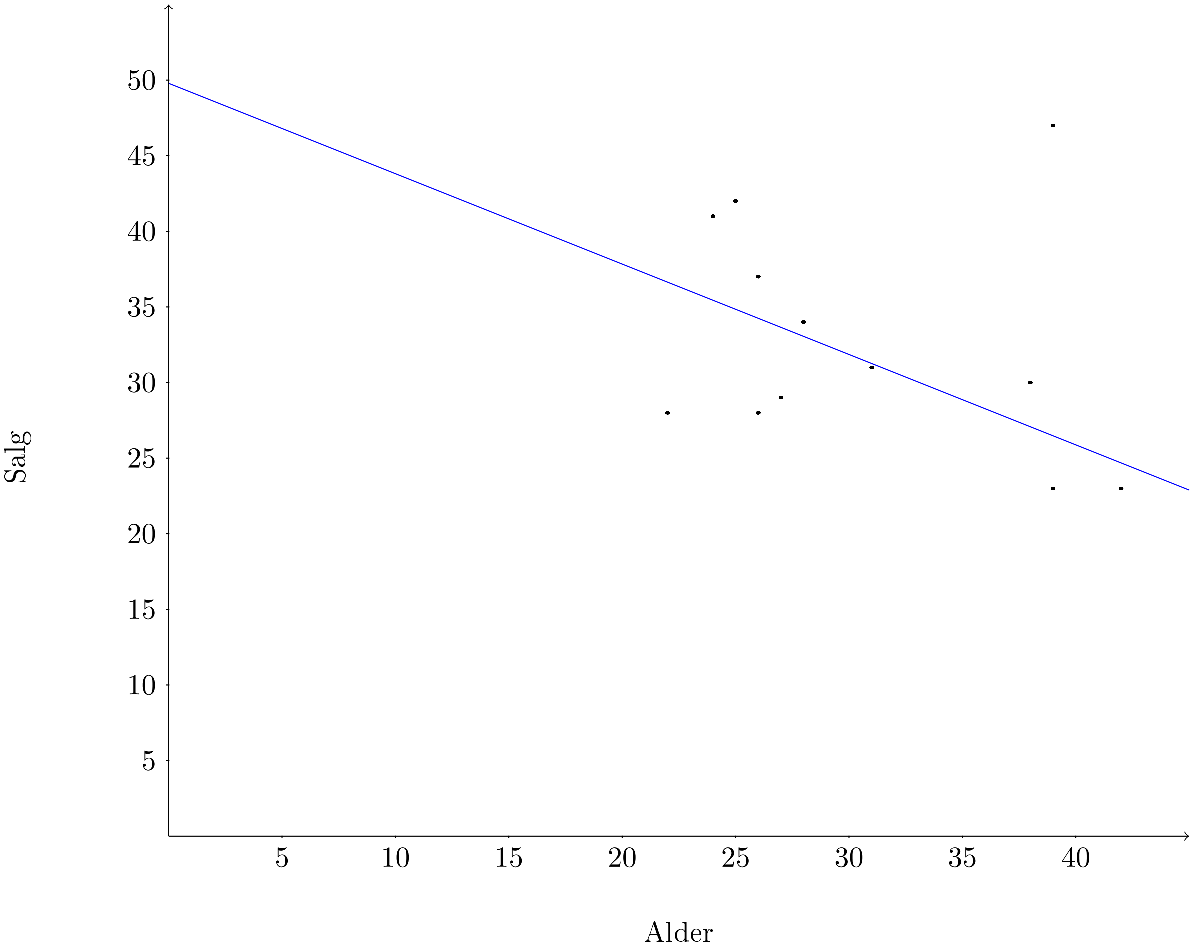
Vi har også at $(n-1)s_x^2 = \sum x_i^2 - n\overline{x}^2$, eller at $\sum x_i^2 = (n-1)s_x^2 + n\overline{x}^2$. Dette gir

$$12194 + a^2 + b^2 \approx 11 \cdot 7.605^2 + 12 \cdot 32.75^2 \approx 13507$$

Ettersom a og b er heltall, får vi $a^2 + b^2 = 13507 - 12194 = 1313$. Sammen med $a + b = 51$, gir dette andregradslikingen

$$a^2 + (51-a)^2 = 1313 \quad \Rightarrow \quad 2a^2 - 102a + 1288 = 0$$

Dermed er $a = 28$ og $b = 23$ siden $a \geq b$.



e) Spredningsdiagrammet er vist ovenfor, med alder som x (første akse) og antall salg som y (andre akse). Korrelasjonskoeffisienten r er gitt ved

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{12}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{12}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{12}(y_i - \overline{y})^2}} \approx -0.485$$

Det er en svak negativ sammenheng mellom alder og antall salg. Vi har at regresjonslinjen $y = \widehat{\alpha} + \widehat{\beta}\,x$ er gitt ved

$$\widehat{\beta} = r \cdot \frac{s_y}{s_x} \approx -0.598, \quad \widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} \approx 49.8$$

Derfor blir likningen til regresjonslinjen $y = -0.598x + 49.8$, og denne linjen er vist i spredningsplottet ovenfor.

f) 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 $n-2 = 10$ frihetsgrader om nullhypotesen $\beta = 0$ er oppfyllt. Hypotesetesten er tosidig, dermed blir forkastningsområdet

$$|T| > t_{\alpha/2}^{n-2} = t_{0.035}^{10} \approx 2.028$$

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

$$\mathrm{SE}(\widehat{\beta})^2 = \frac{\sigma^2}{(n-1)s_x^2} = \frac{\sigma^2}{\sum_{i=1}^{12}(x_i - \overline{x})^2}$$

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

$$s^2 = \frac{\sum_{i=1}^{12}(y_i - \overline{y})^2(1 - r^2)}{n-2}$$

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

$$\frac{s^2}{\sum_{i=1}^{12}(x_i - \overline{x})^2} = \frac{1}{n-2}\left(\frac{\sum_{i=1}^{12}(y_i - \overline{y})^2(1 - r^2)}{\sum_{i=1}^{12}(x_i - \overline{x})^2}\right) = \frac{s_y^2\,(1 - r^2)}{(n-2)\,s_x^2}$$

$$= 0.10 \cdot \frac{57.84}{38.09} \cdot (1 - (-0.485)^2) \approx 0.116$$

Dette gir $T = -0.598/\sqrt{0.116} \approx -1.755$. Ettersom denne verdien ikke ligger i forkastningsområdet, beholder vi nullhypotesen. Vår konklusjon er at det ikke er sammenheng mellom alder og antall salg.

Oppgave 3.

a) Siden komplementet L^C er at alle tre terningene viser et ulikt antall øyne, har vi at

$$1 - p = p(L^c) = \frac{6 \times 5 \times 4}{6 \times 6 \times 6} = \frac{20}{36} = \frac{5}{9}$$

Det følger derfor at $p = 1 - 5/9 = 4/9$.

b) La X være antall ganger vi får minst to like terninger ved $n = 108$ gjentagelser. Da er X binomisk fordelt med $n = 108$ og $p = 4/9$. Sannsynligheten for at $X > 50$ kan vi finne ved normaltilnærming siden X er tilnærmet normalfordelt med $\mu = E(X) = np = 108 \cdot 4/9 = 48$ og $\sigma^2 = \mathrm{Var}(X) = np(1-p) = 108 \cdot 4/9 \cdot 5/9 = 80/3$. Med heltallskorreksjon gir dette

$$p(X > 50) = 1 - p(X \leq 50) = 1 - p\left(Z \leq \frac{50 - 48 + 0.5}{\sqrt{80/3}}\right) \approx 1 - p(Z \leq 0.484) \approx 1 - 0.686 = 0.314$$

c) La oss skrive A, AB, ABA, $ABAB$, og så videre for de ulike forløpene spillet kan ha. Her betyr A at A vinner på første forsøk, AB betyr at B vinner på sitt første forsøk, og så videre. Sannsynligheten for at A vinner er da

$$p(A) + p(ABA) + p(ABABA) + \cdots = 4/9 + (5/9)^2 \cdot 4/9 + (5/9)^4 \cdot 4/9 + \ldots$$

Vi ser at dette er en geometrisk rekke med koeffisient $k = (5/9)^2 = 25/81 < 1$, og sannsynligheten for at A vinner blir dermed summen av den konvergente rekken:

$$\frac{a_1}{1-k} = \frac{4/9}{1 - 25/81} = \frac{4 \cdot 9}{81 - 25} = \frac{36}{56} = \frac{9}{14}$$

Oppgave 4.

a) La $Z = (U-6)/\sqrt{8}$ være standardiseringen av U, som er standard normalfordelt. Da er

$$p(5 \leq U \leq 7) = p\left(\frac{5-6}{\sqrt{8}} \leq Z \leq \frac{7-6}{\sqrt{8}}\right) = p(-1/\sqrt{8} \leq Z \leq 1/\sqrt{8})$$

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

$$p(5 \leq U \leq 7) = G(1/\sqrt{8}) - G(-1/\sqrt{8}) \approx G(0.354) - G(-0.354) \approx 0.276 \approx 0.28$$

b) Vi regner ut kovariansen:

$$\mathrm{Cov}(U+V,U-V) = \mathrm{Var}(U) - \mathrm{Cov}(U,V) + \mathrm{Cov}(V,U) - \mathrm{Var}(V) = \mathrm{Var}(U) - \mathrm{Var}(V) = 8-8 = 0$$

c) Siden $\mathrm{Cov}(U+V,U-V) = 0$ og $U+V, U-V$ er lineærkombinasjoner av U og V, er $U+V$ og $U-V$ uavhengige. Dermed har vi at

$$p(U+V<8, U-V>2) = p(U+V<8) \cdot p(U-V>2)$$

Vi har at $U+V$ er normalfordelt med $E(U+V) = E(U) + E(V) = 12$ og

$$\mathrm{Var}(U+V) = \mathrm{Var}(U) + \mathrm{Var}(V) = 8+8 = 16$$

siden U og V er uavhengige, slik at $\mathrm{Cov}(U,V) = 0$. Dermed følger $U+V$ normalfordelingen $N(\mu,\sigma^2)$ med $\mu = 12$ og $\sigma = 4$. Dette gir

$$p(U+V<8) = p(Z<(8-12)/4) = G(-1) \approx 0.159$$

Tilsvarende har vi at $U-V$ er normalfordelt med $E(U-V) = E(U) - E(V) = 0$ og

$$\mathrm{Var}(U-V) = \mathrm{Var}(U) + \mathrm{Var}(V) = 8+8 = 16$$

siden U og V er uavhengige, slik at $\mathrm{Cov}(U,V) = 0$. Dermed følger $U+V$ normalfordelingen $N(\mu,\sigma^2)$ med $\mu = 0$ og $\sigma = 4$. Dette gir

$$p(U-V>2) = 1 - p(U-V \leq 2) = 1 - p(Z<(2-0)/4) = 1 - G(0.5) \approx 1 - 0.691 = 0.309$$

Dermed får vi at

$$p(U+V<3, U-V>2) = p(U+V<3) \cdot p(U-V>2) \approx 0.159 \cdot 0.309 \approx 0.049$$