STA100 Sannsynlighetsregning og statistikk, vår 2015.

Løsning eksamen 11. september 2015

Oppgave 1:

a) $\mu = 12$ betyr i praksis at gjennomsnittslengden av veldig mange rør er 12 meter.

$$\begin{aligned}
P(X > 12.2) &= 1 - P(\frac{X-12}{0.1} < \frac{12.2-12.0}{0.1}) = 1 - P(Z < 2.00) = 1 - 0.9772 = \underline{\underline{0.023}} \\
P(11.9 < X < 12.1) &= P(X < 12.1) - P(X < 11.9) \\
&= P(\frac{X-12}{0.1} < \frac{12.1-12}{0.1}) - P(\frac{X-12}{0.1} < \frac{11.9-12}{0.1}) \\
&= P(Z < 1.00) - P(Z < -1.00) = 0.8413 - 0.1587 = \underline{\underline{0.68}}
\end{aligned}$$

b) La $T = X_1 + \cdots + X_{900}$ være totallengden.

$$\begin{aligned}
\mathrm{E}(T) &= \mathrm{E}(X_1 + \cdots + X_{900}) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_{900}) \\
&= 12 + \cdots + 12 = 900 \cdot 12 = \underline{\underline{10800}} \\
\mathrm{Var}(T) &= \mathrm{Var}(X_1 + \cdots + X_{900}) \overset{uavh.}{=} \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_{900}) \\
&= 0.1^2 + \cdots + 0.1^2 = 900 \cdot 0.1^2 = \underline{\underline{9}} \\
P(T < 10795 \cup T > 10805) &= 1 - P(10795 \leq T \leq 10805) = 1 - [P(T \leq 10805) - P(T < 10795)] \\
&= 1 - [P(\frac{T - \mathrm{E}(T)}{\sqrt{\mathrm{Var}(T)}} \leq \frac{10805-10800}{\sqrt{9}}) - P(\frac{T - \mathrm{E}(T)}{\sqrt{\mathrm{Var}(T)}} < \frac{10795-10800}{\sqrt{9}})] \\
&= 1 - [P(Z \leq 1.67) - P(Z < -1.67)] = 1 - [0.9525 - 0.0475] = \underline{\underline{0.095}}
\end{aligned}$$

En alternativ måte å regne ut sannsynligheten er å si at $P(T < 10795 \cup T > 10805) = P(T < 10795) + P(T > 10805)$ (siden $T < 10795$ og $T > 10805$ er disjunkte hendelser) og så regne ut disse to sannsynlighetene.

Siden T er normalfordelt (sum av normalfordelte variable) med forventning 10800 har vi at $P(T < 10795) = P(T > 10805)$ (pga vi ser på like store avvik fra forventningsverdien og normalfordelingen er symmetrisk om forventningsverdien) Dersom sannsynligheten for at ledningen blir for lang eller for kort skal være 0.01 må derfor $P(T < 10795) = P(T > 10805) = 0.01/2 = 0.005$. Vi kan da ta utgangspunkt i en av disse for å finne σ (der vi fra det over har at $\mathrm{Var}(T) = 900\sigma^2$), f.eks:

$$\begin{aligned}
P(T < 10795) &= P(Z < \frac{10795-10800}{\sqrt{900 \cdot \sigma^2}}) = 0.005 \\
&\Rightarrow \frac{-5}{\sqrt{900}\sigma} = -2.576 \\
&\sigma = \frac{-5}{-2.576\sqrt{900}} = \underline{\underline{0.065}}
\end{aligned}$$

c) $\bar{x} = \frac{1}{12}\sum_{i=1}^{12} x_i = \frac{1}{12}143.59 = \underline{\underline{11.966}}$.

Ved å sortere dataene i stigende rekkefølge ser man at de to observasjonene i midten blir 11.98 og 11.99, dvs medianen er (11.98+11.99)/2=$\underline{\underline{11.985}}$.

Vi har her en situasjon med normalfordelte data med kjent σ. Et $(1-\alpha)100\%$ konfidensintervall for μ er da gitt ved

$$[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}].$$

Med $\alpha = 0.05$ blir $z_{\alpha/2} = z_{0.025} = 1.96$, og med $\bar{x} = 11.966$, $n = 12$ og $\sigma = 0.1$ blir 95% konfidensintervallet:

$$[11.966 - 1.96\frac{0.1}{\sqrt{12}}, 11.966 + 1.96\frac{0.1}{\sqrt{12}}] = \underline{\underline{[11.91, 12.02]}}$$

Lengden til konfidensintervallet er:

$$\begin{aligned} L &= \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} - (\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ \text{Dvs: } L \leq 0.02 &\Rightarrow 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq 0.02 \\ &\frac{2}{0.02}z_{\alpha/2}\sigma \leq \sqrt{n} \\ &n \geq (100z_{\alpha/2}\sigma)^2 = (100 \cdot 1.96 \cdot 0.1)^2 = 19.6^2 = 384.16 \end{aligned}$$

Dvs det må gjøres minst $\underline{\underline{385}}$ målinger for å få et konfidensintervall med ønsket lengde.

Oppgave 2:

a) A og B er uavhengige hvis $P(A \cap B) = P(A) \cdot P(B)$. Har her at $P(A) \cdot P(B) = 0.35 \cdot 0.45 = 0.1575 \neq P(A \cap B) = 0.3$. Dvs A og B er $\underline{\text{ikke uavhengige}}$.

$$\begin{aligned} P(\text{minst ett av oppdragene}) &= P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.35 + 0.45 - 0.3 = \underline{\underline{0.5}} \\ P(\text{ingen av oppdragene}) &= 1 - P(\text{minst ett av oppdragene}) = 1 - 0.5 = \underline{\underline{0.5}} \end{aligned}$$

b)

$$\begin{aligned} P(O=0) &= P(\text{ingen oppdrag}) = \underline{\underline{0.5}} \\ P(O=10) &= P(A \cap \overline{B}) = P(\overline{B}|A)P(A) = (1 - P(B|A))P(A) = (1 - \frac{P(B \cap A)}{P(A)})P(A) \\ &= P(A) - P(A \cap B) = 0.35 - 0.3 = \underline{\underline{0.05}} \\ P(O=20) &= P(\overline{A} \cap B) = P(\overline{A}|B)P(B) = (1 - P(A|B))P(B) = (1 - \frac{P(A \cap B)}{P(B)})P(B) \\ &= P(B) - P(A \cap B) = 0.45 - 0.3 = \underline{\underline{0.15}} \\ P(O=30) &= P(A \cap B) = \underline{\underline{0.3}} \\ \mathrm{E}(O) &= \sum_o oP(O=o) = 0 \cdot 0.5 + 10 \cdot 0.05 + 20 \cdot 0.15 + 30 \cdot 0.3 = \underline{\underline{12.5}} \\ \mathrm{E}(O^2) &= \sum_o o^2P(O=o) = 0^2 \cdot 0.5 + 10^2 \cdot 0.05 + 20^2 \cdot 0.15 + 30^2 \cdot 0.3 = 335 \\ \mathrm{Var}(O) &= \mathrm{E}(O^2) - (\mathrm{E}(O))^2 = 335 - 12.5^2 = 178.75 \\ \mathrm{SD}(O) &= \sqrt{\mathrm{Var}(O)} = \sqrt{178.75} = \underline{\underline{13.4}} \end{aligned}$$

Oppgave 3:

a) $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{10-1} \cdot 0.901} = \underline{\underline{0.316}}$.

For å finne konfidensintervall for σ tar vi utgangspunkt i resultatet vi har i formelsamlingen som sier at $(n-1)\frac{S^2}{\sigma^2} \sim \chi(n-1)$. Fra dette får vi:

$$\begin{aligned}
P\left(\chi_{1-\alpha/2,n-1} < (n-1)\frac{S^2}{\sigma^2} < \chi_{\alpha/2,n-1}\right) &= 1-\alpha \\
P\left(\frac{\chi_{1-\alpha/2,n-1}}{(n-1)S^2} < \frac{1}{\sigma^2} < \frac{\chi_{\alpha/2,n-1}}{(n-1)S^2}\right) &= 1-\alpha \\
P\left(\frac{(n-1)S^2}{\chi_{1-\alpha/2,n-1}} > \sigma^2 > \frac{(n-1)S^2}{\chi_{\alpha/2,n-1}}\right) &= 1-\alpha \\
P\left(\sqrt{\frac{(n-1)S^2}{\chi_{\alpha/2,n-1}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi_{1-\alpha/2,n-1}}}\right) &= 1-\alpha
\end{aligned}$$

Med $\alpha = 0.05$ og $n = 10$ finner vi i tabell E.6 at $\chi_{\alpha/2,n-1} = \chi_{0.025,9} = 19.02$ og $\chi_{1-\alpha/2,n-1} = \chi_{0.975,9} = 2.70$. Med $s = 0.316$ får vi da

$$\left[\sqrt{\frac{(10-1)\cdot 0.316^2}{19.02}},\ \sqrt{\frac{(10-1)\cdot 0.316^2}{2.70}}\right] = \underline{\underline{[0.22,\ 0.58]}}$$

Standardavviket σ forteller oss hvor mye bæreevnen varierer fra bjelke til bjelke. Sammen med informasjon om forventet bæreevne er dette svært viktig informasjon for praktisk bruk av bjelkene siden det gir oss innsikt i hvor langt fra forventet bæreevne bæreevnen til enkeltbjelker kan vise seg å være. Dersom standardavviket er stort vil vi ha enkelte bjelker med bæreevne mye lavere (og enkelte med bæreevne mye høyere) enn forventet bærevne. Dersom derimot standardavviket er lite vil de fleste bjelkene ha en bæreevne ganske nær forventet bæreevne. Her ser vi at vi får et nokså bredt konfidensintervall (øvre grense er nesten tre ganger så høy som nedre) så det kan nok lønne seg å samle mer data for å få et mer presist estimat.

Oppgave 4:

a) La X være forbruk januar 2014 og Y forbruk januar 2015 og $D = X - Y$. La $\mu_D = E(D) = \mathrm{E}(X) - \mathrm{E}(Y)$. Forventet strømforbruk er redusert dersom denne er større enn 0.

Dvs vi skal teste:

$$H_0: \ \mu_D \leq 0 \qquad \text{mot} \qquad H_1: \ \mu_D > 0$$

Siden vi har direkte målinger av differanser i forbruk på de samme leilighetene kan vi basere oss på disse differansemålingene $D_1, \ldots, D_n$ og bruke en vanlig ett-utvalgs t-test. Dersom H_0 er korrekt er

$$T = \frac{\bar{D} - 0}{S_D/\sqrt{n}} \sim t(n-1)$$

Merk at vi får t-fordeling siden variansen til differansene er ukjent!

Med signifikansnivå 5%, dvs $\alpha = 0.05$, forkaster vi H_0 dersom $T \geq t_{0.05,8} = 1.860$.

Observert: $t = \frac{1076/9}{\sqrt{125720/8}/\sqrt{9}} = 2.86$ Siden $2.86 > 1.860$ blir konklusjonen at vi $\underline{\underline{\text{forkaster } H_0}}$.

Dette betyr at vi kan konkludere at forventet strømforbruk i januar 2015 var lavere enn i januar 2014. Men, dette betyr strengt tatt ikke at vi kan konkludere med at installering av varmepumpe har gitt redusert strømforbruk. Reduksjonen kan også skyldes andre faktorer som for eksempel høyere utetemperatur i januar 2015 enn i januar 2014.

Den typen forsøk som er gjort her med registrering av strømforbruk før og etter i de samme leilighetene kalles en paret sammenligning.

En alternativ måte å undersøke effekten av varmepumpe på kunne være å registrer strømforbruk i noen leiligheter med varmepumpe og noen leiligheter uten varmepumpe for samme måned og så teste om det er forskjell. Dette ville være en uparet sammenligning.

Fordelen med paret sammenligning er at vi får redusert effekten av tilfeldig variasjon i strømforbruk mellom leiligheter og man kan da klare seg med målinger fra færre leiligheten for se en eventuell effekt. Ulempen med paret sammenligning i denne situasjonen er at forskjeller kan skyldes andre faktorer enn varmepumpen (f.eks. forskjeller i utetemperatur) siden man må gjøre måingene på ulike tidspunkter.

Fordelen med uparet sammenligning er at man kan gjøre alle målinger samtidig (samme måned) og dermed slipper problemet med effekten av ulik utetemperatur. En ulempe er at man vil trenger målinger fra flere leiligheter enn ved parvis sammenligning.

Oppgave 5:

a) Den generelle lineæare regresjonsmodellen er $Y = \alpha + \beta x + e$ der vi antar at $e \sim N(0, \sigma)$ og vi antar at feilleddene $e_1, \ldots, e_n$ for ulike målinger er uavhengige.

En praktisk tolkning av α-parameteren er forventet grunnlønn for de med 0 års erfaring. En praktisk tolkning av β-parameteren er at den sier oss hvor mye forventet grunnlønn endrer seg for hvert års endring i erfaring.

Det er rimelig å tro at vi har et årsaks- virkningsforhold her som er slik at antall års erfaring påvirker lønna (og ikke motsatt), og da er det rimelig å sette erfaring som x-variabel og lønn som Y-variabel.

Residualet til en observasjon er definert som $\underline{\underline{\epsilon_i = y_i - \hat{\alpha} - \hat{\beta} x_i}}$ og for siste observasjon blir dette $\epsilon_{24} = y_{24} - 504.6 - 19.1 \cdot x_{24} = 805 - 504.6 - 19.1 \cdot 9 = \underline{\underline{128.5}}$.

Residualplottet er et plott av residualene mot x (lønn) og fra dette plottet kan vi sjekke om antagelsen om linær sammenheng mellom x og forventningen til Y holder, og om antagelsen om lik varians i Y for alle x-verdier holder. Dersom disse antagelsen holder skal vi ikke se noe bestemt mønster i dette plottet (residualene skal være symmetrisk fordelt om 0 og med lik varians for alle x-verdier). Her ser vi imidlertid et klart mønster i residualplottet. I starten ligger alle residualene under 0, så over og så til slutt under igjen. Dette er en sterk indikasjon på at antagelsen om lineær sammenheng ikke holder. Det ser vi også på plott av dataene med regresjonslinjen tegnet inn. Lønnen stiger raskere med antall års erfaring i starten enn senere. Dvs den lineære regresjonsmodellen er ikke en god modell for å beskrive disse dataene.

b) Både ut fra plottet av dataene med regresjonskurven tegnet inn og ut fra plottet av residualene så ser dette ut til å være en bedre modell. I plottet av dataene så passer andregradskurven bedre overens med dataeene enn den rette linjen i den første modellen. I residualplottet så er residualene mer jevnt fordelt rundt null uten noe særlig tendenes til mønster.

I tillegg bør det lages histogram eller normalplott av residualene for å sjekke om antagelsen om normalfordelte residualer holder. Videre bør det også lages et plott av residualene mot innsamlingsrekkefølgen for å se om der er noen avhengigheter.

Estimert forventet lønn med 10 års erfaring: $\hat{y} = 423.9 + 46.1 \cdot 10 - 1.31 \cdot 10^2 = \underline{\underline{753.9}}$.

Generelt bør estimerte regresjonsmodeller aldri brukes langt utenfor området man har data. Vi ser at den estimerte kurven her begynner å avta igjen når x blir stor (fra $x = 17.6$ der den deriverte til kurven er lik 0), og det er neppe en realistisk modell for den virkelige sammenhengen mellom erfaring og lønn. Dvs vi bør ikke bruke denne kurven utenfor området vi har data, f.eks. for erfaring mer enn 25.

c) $\underline{\underline{r^2 = 0.91}}$, og det forteller oss at ca 91% av variasjonen i grunnlønn (Y-variabelen) er forklart ved sammenhengen med erfaring beskrevet ved en tredjegradskurve.

R^2 er ikke et godt mål til å sammenligne modeller med ulike antall x-variabler fordi R^2 alltid øker når vi tar med flere x-variabler. Ved polynomisk regresjon vil da alltid R^2 øke jo høyere grad vi velger på polynomet. I stedet for bør vi bruke R^2-justert som "straffer" det å ta med for mange variabler (R^2-justert balanserer ønsket om å forklare mye av variasjonen og ønsket om å ha med så få variabler som mulig). Vi ser at R^2-justert er 0.74 for den enkle lineære modellen, og 0.89 for både modellen med andre- og tredjegradspolynom. Ved like verdier på R^2-justert velger vi modellen med færrest variabler, dvs i følge dette målet er modellen med andregradskurve å foretrekke.

$$H_0: \ \beta_1 = \beta_2 = \beta_3 = 0 \qquad \text{mot} \qquad H_1: \ \text{minst en } \beta_i \neq 0$$

Vi har fra pensum/formelarket at vi baserer denne testen på at under nullhypotesen er

$$F = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MSR}{MSE} \sim F(\alpha, k, n-k-1)$$

og vi forkaster nullhypotesen dersom F blir stor. Fra den oppgitte variansanalysetabellen ser vi at p-verdien for testen er $0.00000 < 0.05$, dvs $\underline{\underline{\text{vi forkaster } H_0 \text{ på 5\% nivå}}}$, variablene samlet sett har innvirkning på Y-variabelen.

Vi trenger tredjegradsleddet i modellen dersom $\beta_3 \neq 0$, dvs vi skal teste:

$$H_0: \ \beta_3 = 0 \qquad \text{mot} \qquad H_1: \ \beta_3 \neq 0$$

Vi leser p-verdien for testen rett ut fra datautskriften. Vi ser at for tredjegradsleddet i modellen så er p-verdien=0.214 > 0.05 dvs $\underline{\underline{\text{vi forkaster ikke } H_0}}$ og konkluderer med at tredjegradsleddet ikke trengs.

Siden tredjegradsleddet ikke er signifikant og siden R^2-justert er like god for modellen med andregradspolynom ser det her ut for at modellen med andregradspolynom er den beste modellen. Den modellen viser også god overenstemmelse mellom estimert kurve og data i plottene og ser ut for å være en god modell for å beskrive sammenhengen mellom erfaring og lønn i det området vi har data.