Løsning eksamen 26. mai 2021

Oppgave 1

a) Vi har:

- Flere enkeltforsøk som hvert resulterer i "suksess" eller ikke "suksess" - flere personer med korona som enten får blodpropp eller ikke.
- Sannsynligheten for "suksess" er den samme i alle enkeltforsøk, p - samme sannsynlighet p for å få blodpropp for hver tilfeldig person med korona.
- Enkeltforsøkene er uavhengige - uavhengig fra person til person om de utvikler blodpropp eller ikke.
- Et bestemt antall, n enkeltforsøk - et bestemt antall personer med korona, n.

Dermed er X="antall personer med korona som få blodpropp" binomisk fordelt med parametre n og p.

Estimat av p : $\hat{p} = 312/513284 = \underline{\underline{0.00061}}$.

Et (tilnærmet) $(1-\alpha)100\%$ konfidensintervall for p er gitt ved:

$$[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

Innsatt $\hat{p} = 0.00061$, $n = 513284$ og $z_{\alpha/2} = z_{0.025} = 1.96$ gir dette tilnærmet 95% konfidensintervall for p :

$$[0.00061 - 1.96\sqrt{\frac{0.00061(1-0.00061)}{513284}},\ \ 0.00061 + 1.96\sqrt{\frac{0.00061(1-0.00061)}{513284}}] = \underline{\underline{[0.00054\ ,\ \ 0.00068]}}$$

b) Antall uten blodpropp finner vi ved å ta det totale antallet og trekke fra antall med. Vi får da tabellen:

	blodpropp	ikke blodpropp
korona	312	512972
vaksine	33	489838

Første skritt i testen er å regne ut forventet antall i hver celle i tabellen under antagelsen om lik fordeling i gruppene. Vi trenger da også rad- og kolonnesummene. Disse og de forventede verdiene er gitt i tabellen under (de forventede verdiene står i parantesene i hver celle). For eksempel finner vi forventet verdi for kombinasjonen "korona" og "blodpropp" som: $513284 \cdot (345/1003155) = 176.5$ og for kombinasjonen "vaksine" og "ikke blodpropp" som: $489871 \cdot (1002810/1003155) = 489702.5$.

	blodpropp	ikke blodpropp	totalt
korona	312 (176.5)	512972 (513107.5)	513284
vaksine	33 (168.5)	489838 (489702.5)	489871
totalt	345	1002810	1003155

Testobservatoren blir da:

$$\begin{aligned}\mathcal{Q} &= \sum_{\text{alle celler}} \frac{(\text{observert-forventet})^2}{\text{forventet}} \\ &= \frac{(312-176.5)^2}{176.5} + \frac{(512972-513107.5)^2}{513107.5} + \frac{(33-168.5)^2}{168.5} + \frac{(489838-489702.5)^2}{489702.5} = 213.1\end{aligned}$$

Verdien på testobservatoren skal vi sammenligne med 5% kvantilen i kjikvadratfordelingen med parameter (frihetsgrader) $(r-1)\cdot(k-1) = (2-1)\cdot(2-1) = 1$. Fra kjikvadrat-tabellen finner vi at denne kvantilen har verdi 3.84. Siden $\mathcal{Q} = 213.1 > 3.84$ blir konklusjonen at vi forkaster nullhypotesene om like andeler med blodpropp blant the koronasyke og de vaksinerte. Dvs, vi kan konkludere at der er reell forskjell mellom de koronasyke og de vaksinerte i hvor stor andel som utvikler blodpropp. Estimert andel som utvikler blodpropp blant the vaksinerte er $33/489871 = 0.000067$, dvs ca ni ganger så høy estimert andel som blant de koronasyke.

Oppgave 2

a) Vi lar X være konsentrasjonen. Vi får da:

$$
\begin{aligned}
P(X > 1.5) &= 1 - P(\frac{X-1.25}{0.15} < \frac{1.5-1.25}{0.15}) = 1 - P(Z < 1.67) = 1 - 0.9525 = \underline{\underline{0.0475}} \\
P(1.0 < X < 1.5) &= P(X < 1.5) - P(X < 1.0) \\
&= P(Z < 1.67) - P(Z < -1.67) = 0.9525 - 0.0475 = \underline{\underline{0.905}} \\
P(\text{minst en gang}) &= 1 - P(\text{ingen}) = 1 - P(X < 1.5)^{10} = 1 - (1 - 0.0475)^{10} = \underline{\underline{0.385}}
\end{aligned}
$$

b) La μ betegne forventet konsentrasjon. Vi skal da teste

$$H_0: \quad \mu \geq 1.0 \qquad \text{mot} \qquad H_1: \quad \mu < 1.0$$

Vi har at styrke=1-P(type II-feil), dvs vi ønsker $1-\beta = 0.90$ eller $\beta = 0.10$. Videre skal vi bruke nivå $\alpha = 0.05$.

Formelen for utvalgsstørrelse fra formelarket gir oss da:

$$n = \frac{(z_\beta + z_\alpha)^2\sigma^2}{(\mu_0 - \mu)^2} = \frac{(z_{0.1} + z_{0.05})^2\sigma^2}{(1.0-0.9)^2} = \frac{(1.282 + 1.645)^2 0.15^2}{(1.0-0.9)^2} = 19.3$$

Dvs de må gjøre minst $\underline{\underline{20}}$ målinger for å få styrke på minst 0.90.

c) Situasjonen her er normalfordeling med ukjent μ og kjent σ. Dersom H_0 er korrekt er da

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{X} - 1.0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Med signifikansnivå 5%, dvs $\alpha = 0.05$, forkaster vi H_0 dersom $Z \leq -z_{0.05} = -1.645$.
Observert: $z = \frac{0.932-1.0}{0.15/\sqrt{20}} = -2.03$
Siden $-2.03 < -1.645$ blir konklusjonen at vi $\underline{\underline{\text{forkaster } H_0}}$. Dvs, dataene gir grunnlag for å konkludere at forventet flouridkonsentrasjon er lavere enn 1.0. Målet om forventet fluoridkonsentrasjon under 1.0 er nådd.

p-verdien for denne ensidige testen blir p-verdi$=P(Z \leq z_{obs}) = P(Z \leq -2.03) = \underline{\underline{0.021}}$.

Oppgave 3

a)

La Y være antall forsøk på angrep som skjer per døgn. La videre Z være antall forsøk på angrep i løpet av tre dager. Da er Y Poisson-fordelt med forventning $\lambda t = 2.5 \cdot 1 = 2.5$ og Z er Poisson-fordelt med forventning $\lambda t = 2.5 \cdot 3 = 7.5$

$$\begin{aligned}
P(Y = 2) &= \frac{2.5^2}{2!}e^{-2.5} = \underline{\underline{0.257}} \\
P(Y > 2) &= 1 - P(Y \leq 2) = 1 - (\frac{2.5^0}{0!}e^{-2.5} + \frac{2.5^1}{1!}e^{-2.5}\frac{1^{2.5}}{2!}e^{-2.5}) \\
&= 1 - (0.0821 + 0.2052 + 0.2565) = \underline{\underline{0.456}} \\
P(Z = 5) + P(Z = 6) &= \frac{7.5^5}{5!}e^{-7.5} + \frac{7.5^6}{6!}e^{-7.5} = \underline{\underline{0.246}}
\end{aligned}$$

b) La X betegne størelsen på en tilfeldig utbetaling. Vi har for eksponentialfordeling at $\mathrm{E}(X) = 1/\lambda$ som her gir at $1/\lambda = 0.16$ dvs $\lambda = 1/0.16 = 6.25$

Vi kan finne sannsynligheten enten ved å integrere sannsynlighetstettheten, eller, litt enklere, ved å bruke den kumulative fordelingsfunksjonen: $P(X < x) = F(x) = 1 - e^{-\lambda x} = 1 - e^{-6.25x}$. Vi får da:

$$P(X > 0.2) = 1 - P(X \leq 0.2) = 1 - F(0.2) = 1 - (1 - e^{-6.25 \cdot 0.2}) = \underline{\underline{0.29}}$$

La $S = X_1 + \cdots + X_{50}$ være summmen av alle utbetalingene. Vi har oppgitt at utbetalingene kan antas å være uavhengige og siden vi har en sum av $50 > 30$ utbetalinger gir sentralgrenseteoremet at summer vil være tilnærmet normalfordelt. Siden vi har at $\mathrm{E}(X) = 0.16$ og $\mathrm{Var}(X) = 1/\lambda^2 = 1/6.25^2 = 0.0256$ får vi:

$$P(S > 10) = 1 - P(S \leq 10) \approx 1 - P(Z \leq \frac{10 - 50 \cdot 0.16}{\sqrt{50 \cdot 0.0256}}) = 1 - P(Z \leq 1.77) = 1 - 0.9616 = \underline{\underline{0.04}}$$

Oppgave 4

a)

Estimert regresjonsline: $\hat{y} = \hat{\alpha} + \hat{\beta}x = \underline{\underline{3.75 + 0.92 \cdot x}}$.

Her kan vi tolke $\hat{\beta} = 0.92$ som estimert økning i forventet sentraltemperatur når øretemperaturen øker en grad.

Dersom de to temperaturene er like vil dataene falle på en rett linje med stigningstall 1 og som går gjennom origo dersom den trekkes til punktet den skjærer y-aksen. Dvs vi vil få $\underline{\underline{\alpha = 0}}$ og $\underline{\underline{\beta = 1}}$.

Vi skal her finne et prediksjonsintervall (et intervall som med stor sannsynlighet inneholder en ny enkeltverdi). Et $(1 - \alpha)100\%$ prediksjonsintervall for regresjonslinja er gitt ved:

$$\hat{\alpha} + \hat{\beta}x \pm t_{\alpha/2, n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(s/SE(\hat{\beta}))^2}}$$

Ved $x = 37.6$ grader får vi $\hat{y} = 3.75 + 0.92 \cdot 37.6 = 38.34$.

Fra datautskriften finner vi videre at: $s = 0.5172$ og $SE(\hat{\beta}) = 0.04075$. Det er i oppgaven ikke oppgitt hvilken verdi på α man skal bruke. Et standard valg er $\alpha = 0.05$ som er brukt i løsningen under (men $\alpha = 0.01$ eller $\alpha = 0.1$ kan også brukes, kvantilene dette leder til er oppgitt). Ved $\alpha = 0.05$ blir kvantilen $t_{\alpha/2, n-2} = t_{0.025, 235} = 1.970$ og $\bar{x} = 37.11$. Intervallet blir da:

$$38.34 \pm 1.970 \cdot 0.5172\sqrt{1 + \frac{1}{237} + \frac{(37.6 - 37.11)^2}{(0.5172/0.04075)^2}} = 38.34 \pm 1.02 = \underline{\underline{[37.3,\ 39.4]}}$$

b)

Testen som p-verdien er angitt for er $H_0: \;\beta = 0$ mot $H_1: \;\beta \neq 0$. Vi ser fra dataustksriften at p-verdien $< 2 \cdot 10^{-16} < 0.05$ dvs $\underline{\underline{\text{vi forkaster } H_0}}$. Praktisk tolkning av dette er at det er sammenheng mellom x- og Y-variabelen, dvs mellom øretemperatur og sentraltemperatur.

Dersom $\beta = 1$ betyr det at målt øretemperatur og sentraltemperatur endrer seg likt. F.eks. vil en grad økning i øretemperatur også bety en grad økning i forventet sentraltemperatur. Det vil være naturlig å sjekke om de to måtene å måle temperatur på har en slik overenstemmelse eller ikke.

Vi har fra penusm/formelarket at

$$T = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} \sim t(n-2)$$

Under $H_0: \;\beta = 1$ får vi da at

$$T = \frac{\hat{\beta} - 1}{SE(\hat{\beta})} \sim t(n-2)$$

og vi forkaster da H_0 dersom $T \geq t_{\alpha/2,n-2} = t_{0.025,235} = 1.970$ eller $T \leq -t_{\alpha/2,n-2} = -1.970$.

Observert: $t_{obs} = \frac{0.92-1}{0.04075} = -1.96$

Dvs, $\underline{\underline{\text{vi forkaster ikke } H_0 \text{ på 5\% nivå}}}$, dataene gir ikke grunn til å konkludere at de to temperaturene ikke endrer seg likt.

c) Det er her gjort en parvis sammenligning med direkte differansemålinger på 237 personer. Ved slike parvise målinger bruker vi vanlig ett-utvalgs t-intervall/test basert på differansemålingene. Dvs, vi kan her bruke det vanlige ett-utvalgs t-intervallet. Med oppgitte tall og $\alpha = 0.1$ som gir $t_{\alpha/2,n-1} = t_{0.05,236} = 1.651$ følgende 90% konfidensintervall for μ_D:

$$[\bar{D} - t_{\alpha/2,n-1}\frac{S_D}{\sqrt{n}},\; [\bar{D} - t_{\alpha/2,n-1}\frac{S_D}{\sqrt{n}}] = [0.78 - 1.651\frac{0.52}{\sqrt{237}},\; 0.78 + 1.651\frac{0.52}{\sqrt{237}}] = \underline{\underline{[0.72,\; 0.84]}}$$

Antagelsene som ligger til grunn for intervallet er at differansemålingene er $\underline{\underline{\text{uavhengige og normalfordelte}}}$.

Vi ser fra intervallet at forventet forskjell ser ut til å ligge rundt 0.7-0.8 grader - og klart over en differanse på 0. Dvs, det ser ut for at sentraltemperaturen systematisk er høyere en øretemperaturen. Det er derfor grunn til å være forsiktig med å bruke øretemperatur, spesielt hvis man har med kritisk syke pasienter å gjøre, eller pasienter som ikke tåler å ha høy feber.