Løsningsforslag øving 13

11.51

Fordelingen Q kan skrives som

$$Q(A) = \sum_{x \in A} f(x) + \int_A g(x)dx = \frac{1}{2} I(\theta \in A) + \int_{(\theta, \theta+1/2) \cap A} 1 dx,$$

hvor $I(\cdot)$ er indikatorfunksjonen. Dette er altså en fordeling som består av både en kontinuerlig og en diskret sannsynlighetstetthet. Vi har sannsynlighetstettheten $q(x) = f(x) + g(x)$. Vi finner forventningsverdien til en til en tilfeldig variabel med fordelingen Q:

$$\begin{aligned}
\mathrm{E}X &= \sum_{x \in \mathbb{R}} x f(x) + \int_{-\infty}^{\infty} x g(x) dx \\
&= \theta \cdot f(\theta) + \sum_{x \in \mathbb{R} \setminus \{\theta\}} x \cdot 0 + \int_{\theta}^{\theta+1/2} x dx \\
&= \frac{1}{2}\theta + \frac{1}{2}\left((\theta + \frac{1}{2})^2 - \theta^2 \right) \\
&= \frac{1}{2}\theta + \frac{1}{2}(\theta^2 + \theta + \frac{1}{4} - \theta^2) \\
&= \theta + \frac{1}{8}
\end{aligned}$$

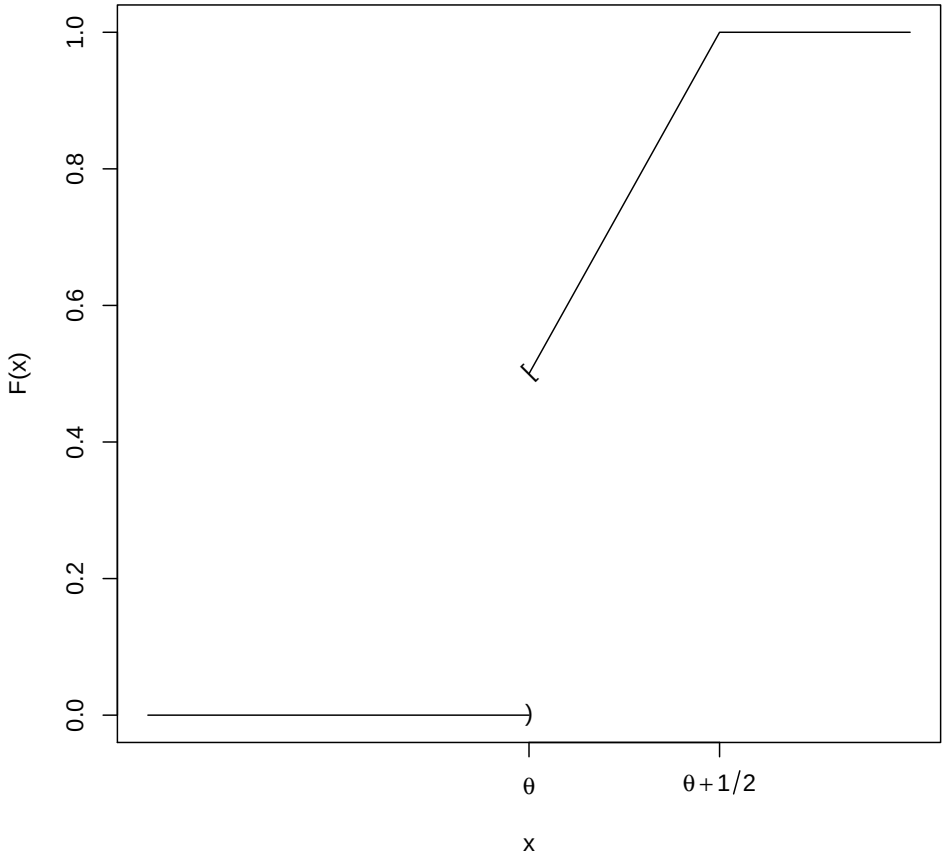
Ut ifra fordelingen Q kan vi finne den kumulative fordelingsfunksjonen

$$\begin{aligned}
F(x) = Q((-\infty, x]) &= \frac{1}{2} I(\theta \leq x) + \int_{(\theta, \theta+1/2) \cap (-\infty, x]} dy \\
&= \frac{1}{2} I(\theta \leq x) + \min\{\theta + 1/2, x\} - \min\{x, \theta\}.
\end{aligned}$$

Vi kan tegne grafen til F:

```
theta = 1
first_part = seq(theta - 1, theta, by = .01)
plot(first_part, rep(0, length(first_part)), "l",
xlim = c(theta - 1, theta + 1),
ylim = c(0, 1),
xlab = "x", ylab = "F(x)", xaxt = "n")
```

```
text(x = theta, y = 0, ")")
F = function(x, theta) {
  ifelse(theta <= x, .5, 0) + pmin(x, theta + .5) - pmin(x,
 theta)
}
x = seq(theta, theta + 1, by = .01)
y = F(x, theta)
lines(x, y)
text(x = theta, y = F(theta, theta), "[", srt = 45)
xticks = c(theta, theta + .5)
xlabs = expression(theta, theta + 1 / 2)
axis(side = 1, at = xticks, labels = xlabs)
```



Vi ønsker nå å estimere θ ved momentmetoden, basert på verdier fra en tilfeldig trekning fra fordelingen. Vi trekker verdiene $X_1, X_2, \ldots, X_n$ og finner gjennomsnittet $\overline{X}$. Momentmetoden gir $\hat{\mu} = \hat{\theta} + 1/8 = \overline{X}$ som fører til moment-estimatoren $\hat{\theta} = \overline{X} - 1/8$.

Vi approksimerer den kontinuerlige delen av Q med funksjonen

$$g_n(x) = 1/(2n), \quad x = \theta + i/(2n), \quad i = 1, \ldots, n.$$

Da kan vi skrive $Q_n(A) = \sum_{x \in A}(f(x) + g_n(x))$, og

$$q_n(x) = f(x) + g_n(x) = (1/2)I(x = \theta) + (1/2n)\sum_{i=1}^{n} I(x = \theta + i/(2n)).$$

For N observasjoner får vi rimelighetsfunksjonen

$$L_n(\theta) = \prod_{j=1}^{N}(1/2)I(x_j = \theta) + (1/2n)\sum_{i=1}^{n} I(x_j = \theta + i/(2n)), \theta \leq \min_j\{x_j\},\ \theta \geq \max_j\{x_j - 1/2\}.$$

Vi ønsker å finne verdien av θ som maksimerer rimelighetsfunksjonen.

Vi kan starte med å se på tettheten $q_1(x) = f(x) + g_1(x)$. Denne tettheten er uniform, og tar verdien $1/2$ i punktene $x = \theta$ og $x = \theta + 1/2$. For denne uniformfordelingen vil rimelighetsestimatoren rett og slett være den verdien av θ som gjør at alle observasjonene fra et tilfeldig utvalg ligger innenfor det godkjente intervallet. Da får vi enten $\hat{\theta} = \min_j\{x_j\}$ eller $\hat{\theta} = \max_j\{x_j\}$, hvor $x_j,\ j = 1, \ldots, N$ er de N observasjonene våre. Nå øker vi verdien av n for fordelingen q_n. Vi ser da at det største enkeltbidraget til rimelighetsfunksjonen kommer fra funksjonen f. Bortsett fra dette leddet så har alle andre ledd identisk verdi. Vi kunne derfor tenkt at for et tilfeldig utvalg av observasjoner så ville rimelighetsestimatoren for θ vært lik moden av fordelingen, dvs. den observasjonen vi fant flest ganger. Dessverre så er rimelighetsfunksjonen null med mindre $\theta \leq \min_j\{x_j\}$. Vi bruker det vi lærte fra $n = 1$ og finner at rimelighetsestimatoren blir lik $\hat{\theta} = \min_j\{x_j\}$, da dette er den enste verdien som garanterer en rimelighetsfunksjon som ikke er lik null og som har et bidrag fra f-funksjonen. Om vi lar $n \to \infty$ kan vi se at funksjonen $g_n(x)$ konvergerer til $g(x)$. Rimelighetsestimatoren er alltid lik $\min_j\{x_j\}$ for alle verdier av n er, så dette må også holde i grensen $n \to \infty$. Da har vi funnet rimelighetsestimatoren for θ for fordelingen Q.

11.57

I eksempel 11.16 finner vi at intervallestimatet for μ når σ^2 er ukjent, for en normalfordeling, er lik $\left[\overline{X} - \frac{t_{\frac{\alpha}{2},n-1}S}{\sqrt{n}}, \overline{X} + \frac{t_{\frac{\alpha}{2},n-1}S}{\sqrt{n}}\right]$. I eksempel 11.15 finner vi at intervallestimatet med kjent varians er lik $\left[\overline{X} - \frac{u_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \overline{X} + \frac{u_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right]$. Vi lar $\alpha = 5\%$ og $n = 20$. Da har vi $u_{\frac{\alpha}{2}} \approx 1.96$ og $t_{\frac{\alpha}{2},n-1} \approx 2.093$. Dette betyr at intervallestimatet for μ med ukjent varians er smalere enn intervallestimatet med kjent varians dersom

$$t_{\frac{\alpha}{2},n-1}S < u_{\frac{\alpha}{2}}\sigma \implies \frac{S}{\sigma} < \frac{u_{\frac{\alpha}{2}}}{t_{\frac{\alpha}{2},n-1}} \approx \frac{1.96}{2.093} \approx 0.9365.$$

Dersom vi f.eks har $\sigma^2 = 1^2$ og $S^2 = 0.9^2$ er dette oppfylt.

11.67

Vi kan skrive $P(-0.883 < T < 1.383) = P(T > -0.833) - P(T > 1.383)$. Vi slår opp i en tabell for student-t-fordelingen. Dessverre er ikke tabellen i "Tabeller og formler i statistikk" stor nok, så vi må velge en annen en, som f.eks tabeller i Larsen-Marx, eller en tabell fra wikipedia: https://en.wikipedia.org/wiki/Student%27s_t-distribution#Table_of_selected_values. Fra en slik tabell kan vi finne at $P(T > -0.8333) - P(T > 1.383) \approx 0.8 - 0.1 = 0.7$. Fordelingen til T er definert ved sannsynlighetstettheten

$$f_T(x) = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)\sqrt{\pi\nu}} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}.$$

11.68

Hvert sykehus rapporterer andelen av pasientene sine som fikk bieffekter. Dette betyr at hvert sykehus rapporterer et gjennomsnitt $Y_i = \overline{X}^{(i)}$, hvor $X_1^{(i)}, \ldots, X_{n_i}^{(i)}$ er resultatene som viser om de n_i pasientene ved sykehus nummer i fikk bieffekter eller ikke. Fra sentralgrenseteoremet vet vi at gjennomsnittet av tilfeldige variabler er tilnærmet normalfordelt når n er stor nok. Vi kan derfor anta at resultatet fra hvert av de 15 sykehusene kan modelleres som en tilfeldig variabel Y_i med normalfordeling. Vi antar at fordelingen ikke varierer fra sykehus til sykehus. Da er vi i situasjonen at vi har 15 uavhengige normalfordelte variabler $Y_1, \ldots, Y_{15} \sim N(\mu, \sigma^2)$, hvor både forventningsverdien og variansen er ukjent. Vi kan da finne, fra eksempel 11.16, at et 95%-konfidensintervall for μ er intervallet

$$\left[\overline{Y} - \frac{t_{2.5\%,14}S}{\sqrt{15}}, \overline{Y} + \frac{t_{2.5\%,14}S}{\sqrt{15}},\right]$$

hvor

$$S^2 = \frac{1}{14}\sum_{i=1}^{15}\left(Y_i - \overline{Y}\right)^2 = \frac{n}{n-1}\left(\overline{y^2} - \overline{y}^2\right).$$

Vi finner $t_{2.5\%,14} \approx 2.145$ og $\overline{y}^2 = (5.96\%)^2 = 35.5216\%^2$, $\overline{y^2} = 37.54133\%^2$ som gir $S \approx 1.4711\%$. Da får vi konfidensintervallet $[5.1453\%, 6.7747\%]$

Vi kan nå beregne et 95% konfidensintervall ved hjelp av Tsjebysjevs ulikhet. Vi antar, som sagt i oppgaven, at variansen til Y er lik $\sigma^2 \leq 2s^2$, som betyr at $\mathrm{Var}(\overline{Y}) = \sigma^2/n \leq 2s^2/n$. Da gir Tsjebysjevs ulikhet oss

$$P(|\overline{Y} - \mu| \geq \frac{ks}{\sqrt{n}}) = P(|\overline{Y} - \mu| \geq \frac{\sigma}{\sqrt{n}} \cdot \frac{ks}{\sigma}) \leq \frac{\sigma^2}{k^2 s^2} \leq \frac{2s^2}{k^2 s^2} = \frac{2}{k^2}.$$

Vi ønsker å finne et 95% konfidensintervall, som betyr at vi må ha $2/k^2 = 0.05$, som gir $k \approx 6.32$. Dette gir et mye bredere konfidensintervall. Vi får intervallet

$$\left[\overline{Y} - \frac{kS}{\sqrt{15}}, \overline{Y} + \frac{kS}{\sqrt{15}},\right] \approx [3.5578\%, 8.3622\%].$$

11.74

Vi antar at sannsynligheten for at en tunfisksalat er dårlig er lik for alle tunfisksalater. Da kan vi beskrive tunfiskundersøkelsen ved et binomisk forsøk, hvor kvaliteten til hver enkelt salat er Bernoullifordelt, $X \sim Ber(p)$. Sentralgrenseteoremet forteller oss da at antall uegnede tunfisksalater, $n\overline{X}$ er tilnærmet normalfordelt med forventning np og varians $np(1-p)$, siden $n = 220$ er relativt stor. For å finne et 95% konfidensintervall for p setter vi opp likningen

$$P(-z_{\alpha/2} \leq \frac{\overline{X} - p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}) = 1 - \alpha.$$

Dette kan skrives om til

$$\begin{aligned} P(|\overline{X} - p| \leq z_{\alpha/2}\sqrt{p(1-p)/n}) &= P(|\overline{X} - p|^2 \leq z_{\alpha/2}^2 p(1-p)/n) \\ &= P(\overline{X}^2 - 2\overline{X}p + p^2 \leq \frac{z_{\alpha/2}^2}{n} p(1-p)) \\ &= P(p^2(1 + z_{\alpha/2}^2/n) - p(2\overline{X} + z_{\alpha/2}^2/n) + \overline{X}^2 \leq 0). \end{aligned}$$

Vi finner nullpunktene til ulikheten,

$$p = \frac{2\overline{X} + z_{\alpha/2}^2/n \pm \sqrt{(2\overline{X} + z_{\alpha/2}^2/n)^2 - 4\overline{X}^2(1 + z_{\alpha/2}^2/n)}}{2(1 + z_{\alpha/2}^2/n)}.$$

Vi skriver disse nullpunktene som $\{p_1, p_2\}$ slik at $p_1 < p_2$. Dette gir konfidensintervallet $[p_1, p_2] \approx [0.7570, 0.8596]$, hvor $n = 220$, $\overline{X} = 179/220$ og $z_{\alpha/2} = 1.96$.

12.1

Siden definisjonen av et kritisk område K_W for W er at det er en hendelse i verdiområdet til observatoren er det mulig å finne sannsynligheten for at W havner i det kritiske området. Da er det altså mulig å finne sannsynligheten for at en nullhypotese forkastes, som er svært nyttig for å beregne teststyrker.

12.2

Dersom en hypotesetest har nivå α betyr det at $P(W \in K_W | W \in \Omega_0) \leq \alpha$, altså at sannsynligheten for at H_0 forkastes når den er sann er mindre enn α. Da er det klart at nullhypotesen i det lange løp ikke vil bli feilaktig forkastet mer enn $\alpha \cdot 100\%$ av gangene. Dersom koeffisienten til testen er lik α betyr det at α er den minste mulige verdien slik at ulikheten $P(H_0 \text{ forkastes} | H_0 \text{ er sann}) \leq \alpha$ er sann.

12.3

Her er parameterrommet $\Omega_\theta = \{p : 1 \geq p \geq 1/2\}$, mens nullhypotesen er $\Omega_0 = \{1/2\}$. Vi ønsker å finne det kritiske området på formen $K_Y(y^*) = \{y^*, \ldots, 18\}$ slik at $P(Y \in K_Y(y^*)|p = 1/2) \leq 10\%$. Dette kan skrives som

$$P(Y \geq y^*|p = 1/2) \leq 10\% \implies 1-P(Y \leq y^*-1|p = 1/2) \leq 10\% \implies P(Y \leq y^*-1|p = 1/2) \geq 90\%.$$

Vi slår opp i tabellsamlingen og finner at denne ulikheten holder for $y^* \geq 13$. Den minste mulige verdien gir derfor et kritisk område $K_Y = \{13, 14, \ldots, 18\}$. Signifikanskoeffisienten er lik $\alpha = P(Y \in K_Y(13)|p = 1/2) = 1 - P(Y \leq 12|p = 1/2) \approx 0.048$. Dersom vi observerer $y = 12$ er konklusjonen at nullhypotesen ikke forkastes ved et signifikansnivå på 10%.

12.4

- Her virker det rimelig å benytte samme type kritisk område som i oppgave 12.3, hvor testobservatoren vår er Y. Vi ønsker å finne det kritiske området på formen $K_Y(y^*) = \{y^*, \ldots, 12\}$ slik at $P(Y \in K_Y(y^*)|p = 1/2) \leq 5\%$. Dette kan skrives som

  $$P(Y \geq y^*|p = 1/2) \leq 5\% \implies 1-P(Y \leq y^*-1|p = 1/2) \leq 5\% \implies P(Y \leq y^*-1|p = 1/2) \geq 95\%.$$

  Vi slår opp i tabellsamlingen og finner at denne ulikheten holder for $y^* \geq 10$. Da velger vi det kritiske området $K_Y = \{10, 11\}$.

- Her har vi testobservatoren $W = n - Y$, og nullhypotesen $p = 3/5$. Hvis Y er liten blir W stor, og motsatt. Vi ser derfor etter et kritisk område på formen $K_W(w^*) = \{w^*, \ldots, n\}$ slik at $P(W \in K_W(w^*)|p = 3/5) \leq 5\%$. Da får vi

  $$\begin{aligned} P(W \in K_W(w^*)|p = 3/5) &= P(W \geq w^*|p = 3/5) \\ &= P(n - Y \geq w^*|p = 3/5) \\ &= P(Y \leq n - w^*|p = 3/5) \leq 5\%. \end{aligned}$$

  Fra en tabell får vi at denne ulikheten er sann for $n - w^* \leq 4$, som gir $w^* \geq 10$. Da velger vi det kritiske området $K_W = \{10, \ldots, 14\}$.

- Her velger vi samme testobservator som i oppgave 12.3, dvs. Y. Vi ser etter et kritisk område på formen $K_Y(y^*) = \{y^*, \ldots, n\}$ slik at $P(Y \in K_Y(y^*)|p = 0.3) \leq 1\%$. Fra en tabell finner vi at denne ulikheten holder for $y^* \geq 12$, så vi velger det kritiske området $K_Y = \{12, 13, \ldots, 20\}$.

12.13

Nullhypotesen er enkel og det kritiske området er et lukket intervall. Da vil vi finne intervallet $[a, b]$ slik at $P(W \in [a, b]|H_0) = \alpha$. Det finnes uendelig mange slike intervaller. Vi

kan finne dem ved hjelp av de øvre kvantilene til $\tilde{W} = [W|H_0 \text{ sann}]$ når W er kontinuerlig. Vi kan f.eks sette $a = \tilde{W}_{c+\alpha/2}$ og $b = \tilde{W}_{c-\alpha/2}$ for en konstant $c \in [\alpha/2, 1-\alpha/2]$. Når W er diskret fordelt kan vi ikke nødvendigvis finne en test som har konfidenskoeffisient α, men vi kan finne intervaller med konfidensnivå α, på liknende måte som for den kontinuerlige fordelingen. Den alternative hypotesen kan ha mye å si i valget av kritisk område. Vi vil gjerne at det kritiske området skal representere den alternative hypotesen så godt som mulig. Hvis den alternative hypotesen for eksempel er ensidig, og henger sammen med lave verdier av W, kan vi velge intervallet $[a, b] = (-\infty, b]$.