Oppgave 1

En lottorekke kan oppfattes som et ikke-ordnet utvalg på 7 elementer blant tallene 1 til 34, der utvelgingen skjer uten tilbakelegging.

a) Det finnes

$$\binom{34}{7} = 5379616$$

forskjellige lotto-rekker.

Skal finne sannsynligheten for at lottorekka inneholder tallet 34. Denne kan beregnes som

$$P(\text{ lottorekka inneholder tallet 34 }) = 1 - P(\text{ lottorekka inneholder IKKE tallet 34 })$$

der antallet mulige lottorekker er $\binom{34}{7}$ og antallet mulige lottorekker uten tallet 34 er $\binom{33}{7}$ slik at

$$P(\text{ lottorekka inneholder tallet 34 }) = 1 - \frac{\binom{33}{7}}{\binom{34}{7}} = 1 - \frac{27}{34} = \frac{7}{34}.$$

Eventuelt kan vi se på antall gunstige som antall rekker som inneholder tallet 34 som vi kan telle ved å telle antall mulige rekker av seks tall mellom 1 og 33, slik at

$$P(\text{ lottorekka inneholder tallet 34 }) = \frac{\binom{33}{6}}{\binom{34}{7}} = \frac{7}{34}.$$

Sannsynligheten for at en lottorekke vil oppnå 7 rette er gitt ved antall gunstige (1 riktig rekke) over antall mulige, som gir

$$P(\text{ 7 rette }) = \frac{1}{\binom{34}{7}} = 1.859 \cdot 10^{-7}.$$

Eventuelt kan vi løse oppgaven ved å definere

$$X: \text{ antall rette i lottorekken,}$$

der X følger hypergeometrisk fordeling

$$h(x; N, n, k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

med $N = 34$, $n = 7$ og $k = 7$.

Sannsynligheten for at en lottorekke vil oppnå 7 rette er gitt ved

$$p = P(\text{ 7 rette }) = h(7; 34, 7, 7) = \frac{\binom{7}{7}\binom{27}{0}}{\binom{34}{7}} = \frac{1}{\binom{34}{7}} = 1.859 \cdot 10^{-7}.$$

b) Det leveres hver uke inn $n = 19\,000\,000$ rekker i lotto. Definer

$$X: \text{ antall av de } n \text{ innleverte rekkene som oppnår 7 rette i ukens trekning}$$

X er binomisk fordelt, men kan med god tilnærmelse regnes å være poissonfordelt da poissonfordelingen er grensefordelingen til binomisk fordeling når $n \to \infty$ og $p \to 0$. Ifølge teorem 5.6 har vi at $n\,p \to \mu$ for $n \to \infty$ slik at parameteren i poissonfordelingen er gitt ved

$$\mu = n\,p = 19\,000\,000 \times 1.859 \cdot 10^{-7} = 3.53.$$

Vi ser i figurene at fordelingene er tilnærmet like.

c) Sannsynlighetstettheten til X er nå gitt ved

$$f_X(x) = \frac{\mu^x}{x!} e^{-\mu}$$

og vi har

$$P(\text{ ingen av de innleverte rekkene har 7 rette })$$
$$= P(X = 0) = \frac{3.53^0}{0!} e^{-3.53} = e^{-3.53} = 0.029.$$

Legg merke til at vi oppnår samme svar uten poissontilnærmingen, dvs dersom X er binomisk fordelt med parametre $n = 19\,000\,000$ og $p = 1.859 \cdot 10^{-7}$ gitt ved

$$f_X(x) = \binom{n}{x} p^x \, (1-p)^{n-x}.$$

Vi har da

$$\begin{aligned}
&P(\text{ ingen av de innleverte rekkene har 7 rette })\\
&= P(X=0) = \binom{n}{0} p^0 \, (1-p)^{n-0} = (1-p)^n = (1 - 1.859 \cdot 10^{-7})^{19\,000\,000} = 0.029.
\end{aligned}$$

Definer

$$Y: \text{ antall Gull-lotto omganger pr. år}$$

der vi antar at det er 52 uker pr. år. Y følger her en binomisk fordeling med $n = 52$ og $p = 0.029$. Vi har dermed at

$$E(Y) = n\,p = 52 \times 0.029 = 1.5$$

og

$$P(Y=0) = f_Y(0) = \binom{52}{0} p^0 \, (1-p)^{52-0} = (1-0.029)^{52} = 0.22.$$

d) Vi skal nå øke antall valgbare tall fra 34 til m, der $m > 34$ slik at det med sannsynlighet minst 0.1 ikke finnes rekker med 7 riktige i en uke der det innleveres $n = 19\,000\,000$ rekker. Vi må i denne oppgaven bruke de tidligere definerte ligningene. Vi har at

$$\begin{aligned}
&P(\text{ ingen av de innleverte rekkene har 7 rette })\\
&= P(X=0) = (1-p)^{19\,000\,000} \geq 0.1
\end{aligned}$$

Vi har videre at p er sannsynligheten for at en lottorekke vil oppnå 7 rette, og denne er gitt ved (se forslag for utregning fra oppgave a))

$$p = P(7 \text{ rette }) = h(7; m, 7, 7) = \frac{\binom{7}{7}\binom{m-k}{0}}{\binom{m}{7}} = \frac{1}{\binom{m}{7}}$$

slik at

$$\left(1 - \frac{1}{\binom{m}{7}}\right)^{19\,000\,000} \geq 0.1$$

Prøver med innsatte verdier av m:

m	35	36
$\left(1 - \frac{1}{\binom{m}{7}}\right)^{19\,000\,000}$	0.059	0.103

Vi ser av tabellen at ulikheten er oppfylt for $m = 36$, og har dermed at $m \geq 36$ dersom det med sannsynlighet minst 0.1 ikke skal finnes rekker med 7 riktige i en uke der det innleveres $n = 19\,000\,000$ rekker.

Oppgave 2

a) Variansen til utvalgsgjennomsnittet er

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

Sannsynlighetstetthetsfunksjonen til normalfordelingen er gitt på s. 25 i *Tabeller og formler i statistikk* som

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right),$$

slik at vi har

$$f(\mu) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\cdot\frac{0}{\sigma^2}\right) = \frac{1}{\sqrt{2\pi}\sigma}e^0 = \frac{1}{\sqrt{2\pi}\sigma}.$$

Dette gir at

$$\mathrm{Var}(\tilde{X}) = \frac{1}{4n\big(f(\mu)\big)^2} = \frac{1}{4n\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^2} = \frac{\pi\sigma^2}{2n} = \frac{\pi}{2}\mathrm{Var}(\bar{X}),$$

hvilket skulle vises.

Når vi skal velge mellom to estimatorer som begge er forventningsrette, velger vi alltid den med minst varians. Siden $\frac{\pi}{2} \approx 1.57 > 1$ har vi $\mathrm{Var}(\tilde{X}) > \mathrm{Var}(\bar{X})$, som betyr at vi foretrekker å bruke $\bar{X}$ som estimator for μ.

b) På grunn av de to tydelige outlierne på oppsiden, kommer medianen $\tilde{X}$ til å være mindre enn utvalgsgjennomsnittet $\bar{X}$ (for disse dataene er $\tilde{X} = 171.0$ mens $\bar{X} = 175.3$).

Vi har antatt at rekruttenes høyder er normalfordelte. Utfra histogrammet ser det ut til at gjennomsnittet ligger rundt 170 cm. I så fall er sannsynligheten for at to av de tretti datapunktene er større enn 235 cm neglisjerbar, så de ekstreme verdiene til disse to datapunktene skyldes antakelig en feil hos rekrutten som fylte inn dataene i regnearket – ikke spesielt usannsynlig, gitt det gulnede papiret og falmede blekket. Siden utvalgsgjennomsnittet er følsomt for outliere, mens utvalgsmedianen ikke er det, gir medianen et bedre estimat enn gjennomsnittet i dette tilfellet.

Anmerkning vedrørende dataene
Datasettet i denne oppgaven er naturligvis fiktivt. Histogrammet er laget for 28 datapunkt trukket tilfeldig fra en normalfordeling med forventningsverdi 166 cm (litt lavere enn gjennomsnittshøyden for 1878, som er 169.5 cm) og standardavvik 7 cm, og med to outliere på 239 cm og 251 cm (høyden til verdens høyeste mann). Når $X \sim N(166, 7^2)$ *så er* $P(X \geq 239) = 9 \cdot 10^{-26}$.

Oppgave 3

a) I dette punktet lar vi X være en bestemt pH-måling, og antar at X er normalfordelt med forventningsverdi $\mu = 6.8$ og varians $\sigma^2 = 0.060^2$. Sannsynligheten for at resultatet av målingen er under 6.74 er da

$$\begin{aligned} P(X < 6.74) &= P(\frac{X - 6.8}{0.06} < \frac{6.74 - 6.8}{0.06}) \\ &= \Phi(-1) = 1 - \Phi(1) \\ &= 1 - 0.841 = \underline{\underline{0.159}}. \end{aligned}$$

Videre er sannsynligheten for at resultatet av målingen ligger mellom 6.74 og 6.86 lik

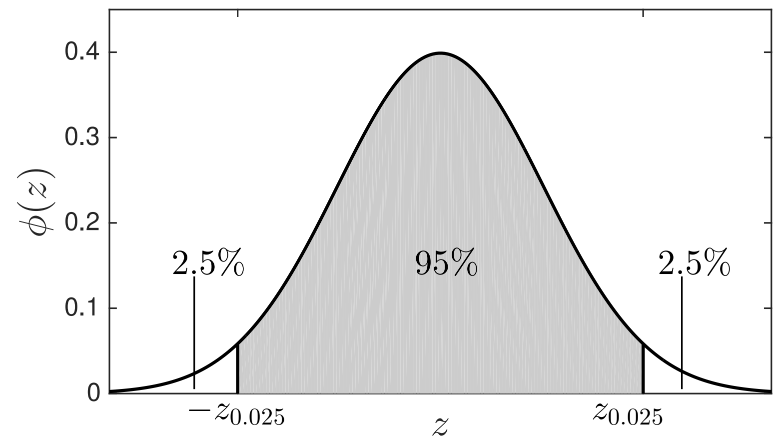
$$\begin{aligned} P(6.74 < X < 6.86) &= P(X < 6.86) - P(X < 6.74) \\ &= P(\frac{X - 6.8}{0.06} < \frac{6.86 - 6.8}{0.06}) - 0.159 \\ &= \Phi(1) - 0.159 = 0.841 - 0.159 = \underline{\underline{0.682}}. \end{aligned}$$

Sannsynligheten for at avviket $|X - \mu|$ overstiger 0.06 er

$$\begin{aligned} P(|X - \mu|) > 0.06) &= P(X - \mu < -0.06) + P(X - \mu > 0.06) \\ &= P(\frac{X - \mu}{0.06} < -1) + P(\frac{X - \mu}{0.06} > 1) \\ &= \Phi(-1) + 1 - \Phi(1) = 2(1 - \Phi(1)) = \underline{\underline{0.318}}. \end{aligned}$$

Den samme sannsynligheten kan også regnes ut som følger:

$$\begin{aligned} P(|X - \mu|) > 0.06) &= 1 - P(6.74 < X < 6.86) \\ &= 1 - 0.682 = \underline{\underline{0.318}}. \end{aligned}$$



Figur 1: Sannsynlighetstettheten $\phi(z)$ til den standard normalfordelte stokastiske variabelen Z. Kvantilene $\pm z_{0.025} = \pm 1.96$ er markert.

b) Her er Y gjennomsnittet av de fem uavhengige målingene $X_1,\ X_2, \dots,\ X_5$, som alle er normalfordelte med forventningsverdi μ og varians σ^2. Siden Y er en lineærkombinasjon av normalfordelte tilfeldige variable, vet vi at Y selv er normalfordelt. Forventningsverdien og variansen til Y er henholdsvis

$$E(Y) = E\left(\frac{1}{5}\sum_{i=1}^{5} X_i\right) = \frac{1}{5}\sum_{i=1}^{5} E(X_i) = \frac{1}{5}\cdot 5\mu = \mu$$

og

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\frac{1}{5}\sum_{i=1}^{5} X_i\right) = \frac{1}{5^2}\sum_{i=1}^{5} \mathrm{Var}(X_i) = \frac{1}{5^2}\cdot 5\sigma^2 = \frac{\sigma^2}{5},$$

slik at $Y \sim N(\mu, \sigma^2/5)$. Sannsynligheten for at Y avviker mer enn 0.06 fra μ er

$$\begin{aligned} P(|Y-\mu| > 0.06) &= P(\{Y-\mu < -0.06\} \cup \{Y-\mu > 0.06\}) \\ &= 2P(Y-\mu > 0.06) \\ &= 2\left(1 - P\left(\frac{Y-\mu}{\frac{0.06}{\sqrt{5}}} \leq \sqrt(5)\right)\right) \\ &= \underline{\underline{0.026}}. \end{aligned}$$

Den tilfeldige variabelen

$$Z = \frac{Y - E(Y)}{\sqrt{\mathrm{Var}(Y)}} = \frac{Y-\mu}{\sqrt{\sigma^2/5}}$$

er standard normalfordelt, $Z \sim N(0,1)$. Et 95% konfidensintervall for μ kan konstrueres ved å ta utgangspunkt i et tilsvarende konfidensintervall for Z (se figur 1), og så manipulere dette slik at μ isoleres,

$$\begin{aligned}
0.95 &= 1 - 2 \cdot 0.025 \\
&= P\left(-z_{0.025} \leq Z \leq z_{0.025}\right) \\
&= P\left(-z_{0.025} \leq \frac{Y - \mu}{\sqrt{\sigma^2/5}} \leq z_{0.025}\right) \\
&= P\left(Y - z_{0.025}\sqrt{\frac{\sigma^2}{5}} \leq \mu \leq Y + z_{0.025}\sqrt{\frac{\sigma^2}{5}}\right).
\end{aligned}$$

Setter vi inn den observerte verdien $y = 6.76$, variansen $\sigma^2 = 0.060^2$, og 0.025-kvantilen til standard normalfordelingen $z_{0.025} = 1.96$, får vi intervallet

$$\left[y - z_{0.025}\sqrt{\frac{\sigma^2}{5}}, y + z_{0.025}\sqrt{\frac{\sigma^2}{5}}\right] = \left[6.76 - 1.96\sqrt{\frac{0.060^2}{5}}, 6.76 + 1.96\sqrt{\frac{0.060^2}{5}}\right]$$
$$= \underline{\underline{[6.707, 6.813]}}.$$