Løsningsforslag øving 10

10.1

Sannsynlighetstettheten til en standard normalfordelt variabel er

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Vi innfører notasjonen $\Phi(x) = \int_{-\infty}^{x} f(t)dt$. Da kan vi skrive

$$I = \frac{1}{\sqrt{2\pi}} \int_{1.25}^{2.50} e^{-\frac{t^2}{2}} dt = \Phi(2.50) - \Phi(1.25).$$

Vi kan finne tabellverdier for Φ i "Tabeller og formler i statistikk". Da får vi

$$I = \Phi(2.50) - \Phi(1.25) \approx 0.9938 - 0.8944 = 0.0994.$$

10.2

Sentralgrenseteoremet sier at $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ er tilnærmet normalfordelt for store verdier av n. Vi vet at hvis X er binomisk fordelt med parametre n og p så kan X skrives som summen av n Bernoullifordelte variabler med parameter p, $X = \sum_{i=1}^{100} Y_i$. Da er også $X = n\bar{Y}$, så vi kan bruke sentralgrenseteoremet på $\bar{Y}$. Vi har $\mu = EY = p$ og $\sigma^2 = Var(Y) = p(1-p)$. Dette gir

$$\begin{aligned} P(38 < X < 43) &= P(38 < n\bar{Y} < 43) \\ &= P(\frac{38/n - \mu}{\sigma/\sqrt{n}} < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < \frac{43/n - \mu}{\sigma/\sqrt{n}}) \\ &= P(\frac{38/n - p}{\sqrt{p(1-p)/n}} < Z < \frac{43/n - p}{\sqrt{p(1-p)/n}}). \end{aligned}$$

På grunn av sentralgrenseteoremet kan vi anta at Z er tilnærmet normalfordelt. Da kan vi tilnærme denne sannsynligheten. Siden X er en diskret variabel bør vi derimot tenke litt over integrasjonsgrensene våre. For å gjøre opp for overgangen fra en diskret variabel til en kontinuerlig variabel velger vi derfor å kun integrere fra 38.5 til 42.5 istedenfor fra 38 til 43. Da har vi

$$P(38 < X < 43) \approx \Phi(\frac{42.5/n - p}{\sqrt{p(1-p)/n}}) - \Phi(\frac{38.5/n - p}{\sqrt{p(1-p)/n}}) \approx \Phi(0.51) - \Phi(-0.31) \approx 0.6950 - 0.3783 = 0.3167.$$

Poisson-approksimasjonen kan brukes her, dvs. den kan alltid brukes, det er bare usikkert hvor god den er. Det ville derimot vært svært lite grunn til å bruke denne tilnærmingen her, da det ikke ville blitt noe lettere å regne ut sannsynligheten $P(38 < X < 43)$ for en Poisson-fordeling.

10.3

Vi har 400 passasjerer som velger mellom selskap A og B med sannsynlighet p. Vi kan altså beskrive antall kunder som velger selskap A ved en binomisk fordeling. Da er $N \sim \text{Bin}(n, p)$, med $n = 400$ og $p = 1/2$. Sannsynligheten for at selskap A får for mange passasjerer er lik $P(N > 230)$. Her er n så stor at vi fint kan tilnærme denne sannsynligheten ved hjelp av sentralgrenseteoremet. Vi ser på N som en sum av Bernoullifordelte variabler, akkurat som i oppgave 10.2 Da får vi

$$P(N > 230) = P(\frac{N - np}{\sqrt{np(1-p)}} > \frac{230 - np}{\sqrt{np(1-p)}}) = P(Z > \frac{230 - np}{\sqrt{np(1-p)}}).$$

Denne sannsynligheten kan tilnærmes med en normalfordeling,

$$P(Z > \frac{230 - np}{\sqrt{np(1-p)}}) \approx 1 - \Phi(\frac{230 - np}{\sqrt{np(1-p)}}) = 1 - \Phi(3) \approx 1 - 0.9987 = 0.0013.$$

Det er altså svært liten sannsynlighet for at selskap A får for mange passasjerer.

10.6

Betegn blodtrykket til 18 år gamle kvinner som X. Da ønsker vi å finne sannsynligheten for at X er større enn 150. Vi er ikke i stand til å regne ut kumulative sannsynligheter for normalfordelingen, da det ikke finnes et analytisk uttrykk for denne. Vi har tilgang til tabellverdier for normalfordelingen, men disse gjelder kun for standard normalfordeling $N(0, 1)$. Vi ønsker derfor å transformere X til en standard normalfordelt variabel.

$$P(X > 150) = P(\frac{X - \mu}{\sigma} > \frac{150 - \mu}{\sigma}) = P(Z > \frac{150 - 120}{12}) = 1 - \Phi(2.5) \approx 1 - 0.9938 = 0.0062.$$

Sannsynligheten for at X ligger mellom 110 og 130 er lik

$$\begin{aligned} P(110 < X < 130) &= P(X < 130) - P(X < 110) \\ &= P(Z < \frac{130 - 120}{12}) - P(Z < \frac{110 - 120}{12}) \\ &= \Phi(0.83) - \Phi(-0.83) \\ &\approx 0.7967 - 0.2033 \\ &= 0.5934. \end{aligned}$$

10.8

Gjennomsnittet av normalfordelte variable $X_1, \ldots, X_n \sim N(\mu, \sigma)$ er en lineær kombinasjon av normalfordelte variable, som betyr at det i seg selv er normalfordelt. Vi har $E[\bar{X}] = \mu$ og $\text{Var}(\bar{X}) =$

σ^2/n. Da får vi, for enhver positive konstant a,

$$\begin{aligned} P(|\bar{X}-\mu|<a) &= P(-a<\bar{X}-\mu<a) \\ &= P(-\frac{a}{\sigma/\sqrt{n}}<\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}<\frac{a}{\sigma/\sqrt{n}}) \\ &= P(-\frac{a}{\sigma/\sqrt{n}}<Z<\frac{a}{\sigma/\sqrt{n}}) \\ &= \Phi(\frac{a}{\sigma/\sqrt{n}})-\Phi(-\frac{a}{\sigma/\sqrt{n}}), \end{aligned}$$

siden Z er standard normalfordelt. Vi vil finne verdien av n som gjør at utvalgets gjennomsnitt med sannsynlighet større enn 90% ligger mindre enn et standardavvik unna. Da får vi

$$P(|\bar{X}-\mu|<\sigma) = \Phi(\frac{\sigma}{\sigma/\sqrt{n}})-\Phi(-\frac{\sigma}{\sigma/\sqrt{n}}) = \Phi(\sqrt{n})-\Phi(-\sqrt{n}) > 90\%.$$

Normalfordelingen er symmetrisk om 0, som betyr at $\Phi(x) = 1-\Phi(-x)$ Da har vi $\Phi(\sqrt{n}) - \Phi(-\sqrt{n}) = 2\Phi(\sqrt{n})-1$. Dette gir likningen

$$2\Phi(\sqrt{n}) > 1.9 \implies \Phi(\sqrt{n}) > 0.95.$$

Fra tabellverdiene for Φ finner vi at verdien av n som løser denne likningen er $\sqrt{n} > 1.65 \implies n > 2.7 \implies n \geq 3$. Vi kan gjøre det samme for to standardavvik. Da får vi

$$\Phi(2\sqrt{n}) > 0.95 \implies 2\sqrt{n} > 1.65 \implies n \geq 1.$$

Dette er klart siden sannsynligheten for at en enkel normalfordelt variabel ligger mindre enn to standardavvik unna μ er lik $\Phi(2)-\Phi(-2) \approx 0.95 > 0.90$.

10.10

Y_j er uniformfordelt mellom -5 og 5. Da er $\mu = E[Y_j] = 0$ og $\mathrm{Var}(Y_j) = 10^2/12 = 25/3$, som gir $\sigma = 5/\sqrt{3}$. Vi bruker sentralgrenseteoremet:

$$\begin{aligned}
P(|\sum_{j=1}^{n} Y_j| > 50) &= 1 - P(|\sum_{j=1}^{n} Y_j| < 50) \\
&= 1 - P(-50 < \sum_{j=1}^{n} Y_j < 50) \\
&= 1 - P(-\frac{50}{n} < \bar{Y} < \frac{50}{n}) \\
&= 1 - P(-\frac{50-0}{\sigma\sqrt{n}} < Z < \frac{50-0}{\sigma\sqrt{n}}) \\
&\approx 1 + \Phi(-\frac{50}{\sigma\sqrt{n}}) - \Phi(\frac{50}{\sigma\sqrt{n}}) \\
&= 1 + \Phi(-\sqrt{3}) - \Phi(\sqrt{3}) \\
&\approx 1 + \Phi(-1.73) - \Phi(1.73) \\
&\approx 1 + 0.0418 - 0.9582 \\
&= 0.0836.
\end{aligned}$$

Sannsynligheten for at avrundingsfeilen i løpet av en måned blir større enn 50 kroner er lik 0.0836.

10.12

Vi kan bruke konvolusjon eller momentgenererende funksjoner for å finne fordelingen til summen av n tilfeldige variabler. Vi velger å bruke momentgenererende funksjoner, da dette er mye enklere. For uavhengige variabler $X_1, \ldots, X_n$ har vi at $M_{\sum_i X_i}(t) = \prod_i M_{X_i}(t)$. Den momentgenererende funksjonen til en Poissonfordeling er kjent fra "Tabeller og formler i statistikk" som $M_X(t) = \exp\{\lambda(e^t - 1)\}$. Dette gir

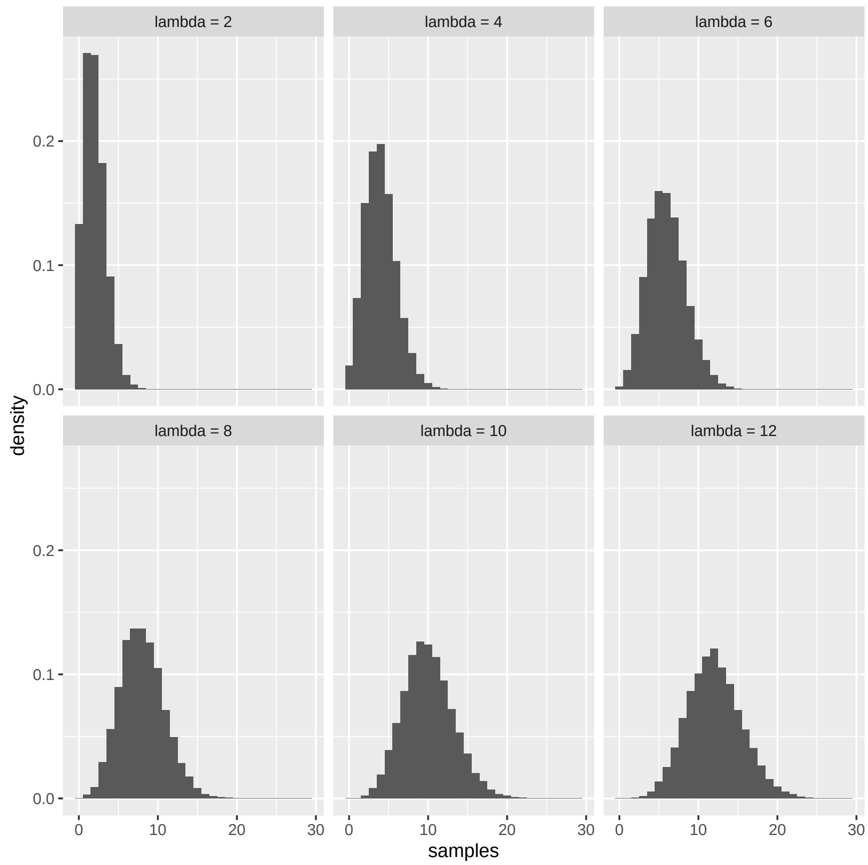
$$M_{\sum_{i=1}^{n}}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} \exp\{\lambda(e^t - 1)\} = \exp\{\sum_{i=1}^{n} \lambda(e^t - 1)\} = \exp\{n\lambda(e^t - 1)\}.$$

Vi kjenner dette igjen som den momentgenererende funksjonen til en Poissonfordeling med parameter $n\lambda$.

Sentralgrenseteoremet sier at summer av mange tilfeldige variabler er tilnærmet normalfordelt. Vi ser her at vi kan skrive en hvilken som helst Poissonfordelt variabel med parameter λ som en sum av n Poissonfordelte variabler med parameter λ/n. Derfor burde egentlig alle Poissonfordelinger være tilnærmet normalfordelte. Dette gjelder derimot ikke, fordi Poissonfordelingen ikke er definert for negative verdier. Hvis vi derfor lar λ være så stor at vi ikke trenger å vise hensyn til randeffektene ved $x = 0$ kan vi se at Poissonfordelingen kan tilnærmes godt med normalfordelingen. Eksempel kan sees i figuren under. Det er tydelig at fordelingen blir stadig mer lik normalfordelingen når vi flytter oss bort fra randen.

```
lambda_vec <- seq(from = 2, to = 12, by = 2)
samples <- lapply(
  X = lambda_vec,
  FUN = function(lambda) {
    samples <- rpois(n = 20000, lambda)
    data.frame(samples = samples, lambda = paste0("lambda = ", lambda))
  })
data <- do.call(rbind, samples)

library(ggplot2)
plot <- ggplot(data) +
  geom_histogram(aes(x = samples, y = ..density..)) +
  facet_wrap(~lambda)
print(plot)
```



10.15

Definer $U = \sum_{j=1}^{4} Z_j^2$. Siden $Z_j \sim N(0,1)$ har vi fra side 98 i notatet at U er kji-kvadratfordelt med 4 frihetsgrader. Fra Tabeller og formler i statistikk finner vi at sannsynlighetstettheten til en kji-kvadratfordeling med ν frihetsgrader er

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}.$$

Dette gir sannsynlighetstettheten til U:

$$f(u) = \frac{1}{4} x e^{-x/2}.$$

Vi kan regne ut sannsynligheten for at U ligger mellom 2 og 3.

$$
\begin{aligned}
P(2 < U < 3) &= \int_2^3 \frac{1}{4} x e^{-x/2} dx \\
&= -\frac{1}{2} x e^{-x/2} \Big|_2^3 + \int_2^3 \frac{1}{2} e^{-x/2} \\
&= -\frac{3}{2} e^{-3/2} + e^{-1} - \left(e^{-x/2} \Big|_2^3 \right) \\
&= -\frac{3}{2} e^{-3/2} + e^{-1} - e^{-3/2} + e^{-1} \\
&= 2e^{-1} - \frac{5}{2} e^{-3/2} \\
&\approx 0.1779.
\end{aligned}
$$

Sannsynligheten for at U ligger mellom 2 og 3 for ett enkelt eksperiment er altså lik $p \approx 0.1779$. Da kan antall suksesser når vi gjentar forsøket 100 ganger beskrives som et binomisk forsøk med parametre $n = 100$ og $p \approx 0.1779$. Det forventede antallet suksesser blir da $E = np \approx 17.79$.

10.16

Vi har at $X \sim F(m, n)$. Da ser vi fra side 99 i notatet at X kan uttrykkes som $X = (U/m)/(V/n)$ hvor $U \sim \chi^2_m$ og $V \sim \chi^2_n$. Det er klart at inversen av X da blir $1/X = (V/n)/(U/m)$, som vi kan se at må være Fisherfordelt, $1/X \sim F(n, m)$.

10.17

T er Student-t-fordelt med ν frihetsgrader. Da finner vi tabellen for kritiske verdier i t-fordelingen i Tabeller og formler i statistikk. Der finner vi $t_{0.05,24} = P(T > t_{0.05,24}) = 1.711$.