TMA4245 Statistikk
Vår 2015

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Øving nummer 5, blokk I
Løsningsskisse

Oppgave 1

X er hypergeometrisk fordelt med $N = 1000$ turer, $k = 5$ turer kjører transportfirmaet gjennom sentrum og $N - k = 995$ utenom sentrum, og vi tar en stikkprøve av størrelse $n = 5$.

Betingelser:

- Et tilfeldig utvalg av størrelse n tas *uten* tilbakelegging fra N enheter. Her: et tilfeldig utvalg av $n = 5$ turer sjekket blant N turer som totalt kjøres.
- De N enhetene deles inn i to grupper, k suksesser og $N - k$ fiaskoer. Her: $k = 5$ turer kjøres gjennom sentrum og $N - k = 995$ turer kjøres utenom sentrum.
- X er antallet suksesser blant de n. Her: X er antall turer gjennom sentrum av de $n = 5$ turene som ble sjekket.

Punktsannsynligheten i hypergeometrisk fordeling, $N = 1000, k = 5, n = 5$ er gitt som:

$$P(X = x) = \frac{\binom{5}{x}\binom{995}{5-x}}{\binom{1000}{5}}$$

og mulige verdier for x er $0, 1, 2, 3, 4, 5$.

$$P(X = 0) = \frac{\binom{5}{0}\binom{995}{5-0}}{\binom{1000}{5}} = 0.9752$$

Siden $P(X = 0) = 0.975$ må $x = 0$ være den verdien av x som gir høyest punktsannsynlighet (siden summen av alle punktsannsynligheter er 1 kan ingen annen punktsannsynlighet være større enn 1-0.975).

$$P(X = 5) = \frac{\binom{5}{5}\binom{995}{0}}{\binom{1000}{5}} = \underline{\underline{1.21 \cdot 10^{-13}}}$$

Til sammenligning er sannsynligheten for å vinne 7 rette i lotto $1.85 \cdot 10^{-7}$.

Kommentar 1: Når N er stor i forhold til n (boka nevner som tommelfingerregel at $n/N \leq 0.05$, og her er jo $5/1000 = 0.005$) så kan binomisk fordeling brukes som en tilnærming til hypergeometrisk fordeling når vi regner ut sannsynligheter. Da gjør vi $n = 5$ forsøk og i hvert forsøk sjekker vi om transporten skjer gjennom bykjernen, $p = \frac{k}{N} = \frac{5}{1000}$ er sannsynlighet for transport gjennom bykjernen, og X er antall transporter gjennom bykjernen for de $n = 5$ undersøkt. Da kan punktsansynligheten til X tilnærmes med

$$P(X = x) = \binom{n}{x}(\frac{k}{N})^x(1 - \frac{k}{N})^{n-x} = \binom{5}{x}(\frac{5}{1000})^x(1 - \frac{5}{1000})^{5-x}$$

Videre er tilnærmet:

$$\begin{aligned} P(X = 0) &= \binom{5}{0}(\frac{5}{1000})^0(1 - \frac{5}{1000})^{5-0} = (1 - \frac{5}{1000})^5 = 0.975 \\ P(X = 5) &= \binom{5}{5}(\frac{5}{1000})^5(1 - \frac{5}{1000})^{5-5} = (\frac{5}{1000})^5 = 3.125 \cdot 10^{-12} \end{aligned}$$

Kommentar 2: Denne oppgaven er basert på en henvendelse fra en tidligere bygg-student, og er basert på faktiske forhold. Dog, transportfirmaet sa først at alle 1000 turene var kjørt utenom bykjernen og kun etter at de be møtt med fakta på at stikkprøve av 5 turer viste transport gjennom bykjernen så informerte de om at det kun var akkurat disse 5 turene (av de 1000) som hadde blitt kjørt gjennom bykjernen. La oss tenke oss at vi ser på dette som en hypotesetest, der vi ønsker å finne ut om det er grunn til å tro at transportfirmaet har kjørt mer enn $k = 5$ av turene gjennom bykjernen:

$$H_0 : k = 5 \text{ vs. } H_1 : k > 5$$

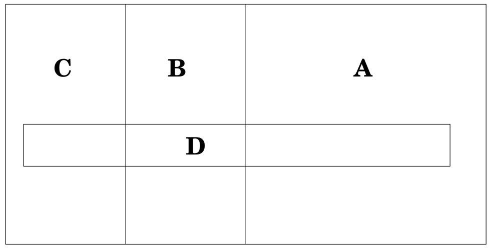
P-verdien til testen ville vært å regne ut $P(X = 5)$ som vi har gjort i oppgaven, og denne er $1.21 \cdot 10^{-13}$, som ville ført til at vi forkastet nullhypotesen og ville tro at flere enn 5 transporter var kjørt gjennom bykjernen. Men, dette var ikke med i oppgaven.

Oppgave 2

Vi ser på en tilfeldig valgt natt og definerer følgende hendelser:

A = Anne er på vakt,
B = Bernt er på vakt,
C = Cecilie er på vakt,
D = det skjer et dødsfall.
Og antar at alle dødsfall skjer naturlig.

a) Venndiagram for de fire hendelsene:



Siden det bare er Anne, Bernt og Cecilie som jobber på sykehjemmet om natten vil hendelsene A, B og C utgjøre en partisjon av utfallsrommet, og vi må ha at $P(A) +$

$P(B) + P(C) = 1$. Dette ser vi også av venndiagrammet. Siden Bernt og Cecilie jobber like ofte må $P(B) = P(C)$. Siden Anne jobber dobbelt så ofte som hver av Bernt og Cecilie må $P(A) = 2 \cdot P(B) = 2 \cdot P(C)$. Vi utrykker alt ved $P(B)$.

$$\begin{aligned} P(A) + P(B) + P(C) &= 1 \\ 2 \cdot P(B) + P(B) + P(B) &= 1 \\ P(B) &= 0.25 \end{aligned}$$

Dermed har vi at

$$\begin{aligned} P(A) &= 0.5 \\ P(B) &= 0.25 \\ P(C) &= 0.25 \end{aligned}$$

For å regne ut $P(D)$ kan vi bruke setningen om total sannsynlighet. Vi vet at A, B, C er en partisjon av utfallsrommet.

$$\begin{aligned} P(D) &= P(D \cap A) + P(D \cap B) + P(D \cap C) \\ &= P(D|A) \cdot P(A) + P(D|B) \cdot P(B) + P(D|C) \cdot P(C) \\ &= 0.06 \cdot (0.5 + 0.25 + 0.25) = \underline{\underline{0.06}} \end{aligned}$$

Definisjonen av uavhengighet sier at C og D er to uavhengige hendelser hvis og bare hvis $P(D|C) = P(D)$, dvs. at "tilleggsinformasjon ikke endrer bildet". Vi ser fra utregningene over at $P(D|C) = P(D) = 0.06$, og C og D er dermed uavhengige hendelser.

Intuitivt vil uavhengighet av C og D følge av antagelsen om naturlig død.

b) X er en stokastisk variabel som beskriver antall av $n = 10$ naturlige dødsfall som skjer på Cecilies vakter.

Betingelser for at X er binomisk fordelt:

- Vi ser på $n = 10$ dødsfall.
- For hver dødsfall sjekker vi om Cecilie var på vakt eller ikke.
- Sannsynligheten for at Cecilie er på vakt gitt at det har skjedd et dødsfall er $P(C|D) = P(C) = 0.25$, og denne sannsynligheten er det samme for alle de n dødsfallene.
- De n dødsfallene er uavhengige siden de er naturlige (og vi antar dermed at det ikke er snakk om smittsomme sykdommer eller epidemier).

Under disse 4 betingelsene er X="antall naturlig dødsfall på Cecilies vakter" binomisk fordelt med parametere $n = 10$ og $p = 0.25$. Dermed er sannsynlighetsfordelingen til X gitt ved punktsannsynligheten $f(x)$,

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, ..., n$$

Sannsynligheten for at 7 eller flere av 10 dødsfall om natten skjer på Cecilies vakter finner vi enklest ved tabelloppslag (s 13 i formelsamlingen),

$$P(X \geq 7) \quad = \quad = 1 - P(X \leq 6) = 1 - 0.996 = \underline{\underline{0.004}}$$

La Y være en stokastisk variabel som angir antall sykepleiere blant 300 sykepleiere som opplever flere enn 7 dødsfall på sine vakter av totalt 10 dødsfall. Y vil dermed være binomisk fordelt med $n = 300$ og $p = 0.004$.

Sannsynligheten for at minst en av de 300 sykepleierne opplever at 7 eller flere av 10 naturlige dødsfall skjer på sine vakter er gitt som $P(Y \geq 1)$.

$$\begin{aligned} P(Y \geq 1) &= 1 - P(Y = 0) = 1 - \binom{300}{0} 0.004^0 (1 - 0.004)^{300-0} \\ &= 1 - 0.996^{300} = 1 - 0.3 = \underline{\underline{0.7}} \end{aligned}$$

Selv om det er lite sannsynlig (bare 4 promille) at det skjer 7 av 10 naturlige dødsfall på Cecilies vakter, er det svært sannsynlig (70 prosent) at minst 7 av 10 dødsfall kan skje på vakten til en av sykepleierne i Norge som jobber i samme stillingstype som Cecilie. Disse observasjonene styrker ikke mistanken mot Cecilie.

Analogi: Hver uke er det (som regel) noen som får 7 rette i Lotto, selv om dette har en forsvinnende lav sannsynlighet for hver Lotto-spiller.

Oppgave 3

a) Antall mulige tallkombinasjoner blir $10^4 = 10000$. Sannsynligheten for å gjette riktig blir da $p = \frac{1}{10000}$.
X er geomtrisk fordelt med parameter $p = \frac{1}{10000}$ fordi det er bare to mulige utfall, gjette riktig eller feil, med konstant sannsynlighet, hvert gjett er uavhengig av de forrige gjettene, og X er antall gjett til og med først suksess (riktig gjett).

$$P(X = x) = p(1-p)^{x-1} \qquad \text{(geomtrisk fordeling)}$$

$$\Downarrow$$

$$P(X = 300) = \frac{1}{10000}\left(1 - \frac{1}{10000}\right)^{299} = \frac{1}{10303,53} = \underline{\underline{9,7 \cdot 10^{-5}}}$$

b) Antall måter å plassere to 7-tall blandt 4 posisjoner er

$$\binom{4}{2} = \frac{4 \cdot 3}{2 \cdot 1} = 6.$$

Antall mulige siffer på to gjenværende posisjoner er

$$9^2 = 81.$$

Dette gir

$$m = 6 \cdot 81 = \underline{\underline{486}}$$

Dermed blir X geometrisk fordelt med $p = 1/486$,

$$E[X] = \frac{1}{p} = 486.$$

Forventet tid til første vinner blir dermed

$$486 \cdot 1/2 \text{ min } = 243 \text{ min } = 4 \text{ timer } 3 \text{ minutter }.$$

c) Utfallsrommet til $Y = \{1, 2, 3, \ldots, m\}$. Definerer hendelsen

$$F_i : \text{Innringer nummer } i \text{ gjetter riktig kode}$$

der $i = 1, 2, \ldots, m$. Setter opp sannsynligheten

$$\begin{aligned}
P(Y = y) &= P(F_1^c \cap F_2^c \cap \ldots \cap F_{y-1}^c \cap F_y) \\
&= P(F_y | F_1^c \cap F_2^c \cap \ldots \cap F_{y-1}^c) P(F_1^c \cap F_2^c \cap \ldots \cap F_{y-1}^c) \\
&= \ldots = P(F_y | F_1^c \cap F_2^c \cap \ldots \cap F_{y-1}^c) \cdot P(F_{y-1}^c | F_1^c \cap F_2^c \cap \ldots \cap F_{y-2}^c) \\
&\qquad \ldots \cdot P(F_2^c | F_1^c) \cdot P(F_1^c) \\
&= \frac{1}{m - (y-1)} \cdot \frac{m - (y-1)}{m - (y-2)} \cdot \ldots \cdot \frac{m-2}{m-1} \cdot \frac{m-1}{m} \\
&= \frac{1}{m}.
\end{aligned}$$

Da blir

$$\underline{\underline{f(y) = P(Y = y) = \frac{1}{m}}} \text{ for } y = 1, 2, \ldots, m$$

Forventet tid til noen vinner premien:

$$\begin{aligned}
E(Y) &= \sum_{y=1}^{m} y \cdot f(y) = \sum_{y=1}^{m} y \cdot \frac{1}{m} = \frac{1}{m} \sum_{y=1}^{m} y \\
&= \frac{1}{m} \cdot \frac{m(m+1)}{2} = \frac{m+1}{2}
\end{aligned}$$

Forventet tid blir da

$$E(Y) \cdot 1/2 \text{ min } = \underline{\underline{\frac{m+1}{4} \text{ min } = 121 \text{ min } 45 \text{ sec}}}$$

Oppgave 4

a) Når rosinene er uniformt fordelt i pakken, vil en tilfeldig valgt rosin befinne seg i den første porsjonen med sannsynlighet $1/m$, siden det er andelen av den totale mengden frokostblanding i pakken som utgjør den første porsjonen. Dette gjelder uavhengig for hver av de n rosinene, slik at X_1 blir binomisk fordelt med parametre n og $1/m$,

$$X_1 \sim \text{Bin}\left(n, \frac{1}{m}\right).$$

Hvis den første prosjonen inneholder x_1 rosiner, er det $n - x_1$ rosiner igjen som fordeler seg på de $m - 1$ øvrige porsjonene. Den betingede fordelingen til X_2 gitt at $X_1 = x_1$ er dermed binomisk med parametre $n - x_1$ og $1/(m-1)$,

$$[X_2|X_1] \sim \text{Bin}\left(n - x_1, \frac{1}{m-1}\right).$$

Første prosjon inneholder minst en rosin med sannsynlighet

$$P(X_1 \geq 1) = 1 - P(X_1 = 0) = 1 - \left(1 - \frac{1}{m}\right)^n.$$

Krever vi at denne sannsynligheten skal være større enn eller lik β, får vi følgende ulikhet, som kan løses med hensyn til n,

$$\begin{aligned}
P(X_1 \geq 1) &\geq \beta \\
1 - \left(1 - \frac{1}{m}\right)^n &\geq \beta \\
\left(1 - \frac{1}{m}\right)^n &\leq 1 - \beta \\
n \log\left(1 - \frac{1}{m}\right) &\leq \log(1-\beta) \\
n &\geq \frac{\log(1-\beta)}{\log\left(1 - \frac{1}{m}\right)}.
\end{aligned}$$

Ulikheten snur fra nest siste til siste linje fordi $\log(1 - 1/m) < 0$.

b) X_1 og X_2 er ikke uavhengige. Jo flere rosiner som er inneholdt i den første porsjonen, jo færre rosiner vil en forvente å finne i den andre porsjonen. Korrelasjonskoeffisienten mellom X_1 og X_2 er negativ, siden en økning i den ene variabelen er forbundet med en (forventet) reduksjon av den andre. Den negative sammenhengen blir særs tydelig når $m = 2$, slik at $X_1 + X_2 = n$.

Den simultane punktsannsynligheten til X_1 og X_2 er

$$\begin{aligned}
&P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2|X_1 = x_1) \\
&\quad= \binom{n}{x_1}\left(\frac{1}{m}\right)^{x_1}\left(\frac{m-1}{m}\right)^{n-x_1}\binom{n-x_1}{x_2}\left(\frac{1}{m-1}\right)^{x_2}\left(\frac{m-2}{m-1}\right)^{n-x_1-x_2} \\
&\quad= \binom{n}{x_1}\binom{n-x_1}{x_2}\left(\frac{1}{m}\right)^{x_1}\left(\frac{1}{m}\right)^{x_2}\left(\frac{m}{m-1}\right)^{x_1}\left(\frac{m}{m-1}\right)^{x_2}\left(\frac{m}{m-1}\right)^{-n}\left(\frac{m-2}{m-1}\right)^{n-x_1-x_2} \\
&\quad= \binom{n}{x_1,\ x_2,\ n-x_1-x_2}\left(\frac{1}{m}\right)^{x_1}\left(\frac{1}{m}\right)^{x_2}\left(\frac{m-1}{m}\right)^{n-x_1-x_2}\left(\frac{m-2}{m-1}\right)^{n-x_1-x_2} \\
&\quad= \binom{n}{x_1,\ x_2,\ n-x_1-x_2}\left(\frac{1}{m}\right)^{x_1}\left(\frac{1}{m}\right)^{x_2}\left(\frac{m-2}{m}\right)^{n-x_1-x_2}.
\end{aligned}$$

Variablene X_1, X_2 og $\sum_{i=3}^m X_i = n - X_1 - X_2$ følger den multinomiske sannsynlighetsfordelingen med parametre n og $\mathbf{p} = (1/m,\ 1/m,\ (m-2)/m)^T$.

c) Forventingsverdien til indikatorvariabelen Y_1 er

$$\mathrm{E}(Y_1) = 0 \cdot P(Y_1 = 0) + 1 \cdot P(Y_1 = 1) = P(Y_1 = 1) = P(X_1 = 0) = \left(1 - \frac{1}{m}\right)^n.$$

Siden marginalfordelingene til Y_1, Y_2, …, Y_m er like, har vi $\mathrm{E}(Y_i) = \mathrm{E}(Y_1)$ for $i = 2, 3, \ldots, m$, som gir

$$\mathrm{E}(W) = \mathrm{E}\left(\sum_{i=1}^{m} Y_i\right) = \sum_{i=1}^{m} \mathrm{E}(Y_i) = m\mathrm{E}(Y_1) = m \cdot \left(1 - \frac{1}{m}\right)^n.$$

Variansen til Y_1 er

$$\begin{aligned}
\mathrm{Var}(Y_1) &= (0 - \mathrm{E}(Y_1))^2 P(Y_1 = 0) + (1 - \mathrm{E}(Y_1))^2 P(Y_1 = 1) \\
&= \left(-\left(1 - \frac{1}{m}\right)^n\right)^2 \left(1 - \left(1 - \frac{1}{m}\right)^n\right) + \left(1 - \left(1 - \frac{1}{m}\right)^n\right)^2 \left(1 - \frac{1}{m}\right)^n \\
&= \left(1 - \frac{1}{m}\right)^n \left(1 - \left(1 - \frac{1}{m}\right)^n\right) \underbrace{\left[\left(1 - \frac{1}{m}\right)^n + \left(1 - \left(1 - \frac{1}{m}\right)^n\right)\right]}_{1} \\
&= \left(1 - \frac{1}{m}\right)^n - \left(1 - \frac{1}{m}\right)^{2n}.
\end{aligned}$$

Kovariansen mellom Y_1 og Y_2 er

$$\begin{aligned}
\mathrm{Cov}(Y_1, Y_2) &= \mathrm{E}(Y_1 Y_2) - \mathrm{E}(Y_1)\mathrm{E}(Y_2) \\
&= \sum_{y_1=0}^{1} \sum_{y_2=0}^{1} y_1 y_2 P(Y_1 = y_1, Y_2 = y_2) - \mathrm{E}(Y_1)^2 \\
&= P(Y_1 = 1, Y_2 = 1) - \mathrm{E}(Y_1)^2 \\
&= P(Y_1 = 1) P(Y_2 = 1 | Y_1 = 1) - P(X_1 = 0)^2 \\
&= P(X_1 = 0) P(X_2 = 0 | X_1 = 0) - P(X_1 = 0)^2 \\
&= \left(\frac{m-1}{m}\right)^n \left(\frac{m-2}{m-1}\right)^n - \left[\left(\frac{m-1}{m}\right)^n\right]^2 \\
&= \left(\frac{m-2}{m}\right)^n - \left(\frac{m-1}{m}\right)^{2n}.
\end{aligned}$$

Variansen til W er

$$\begin{aligned}
\mathrm{Var}(W) = \mathrm{Var}\left(\sum_{i=1}^{m} Y_i\right) &= \sum_{i=1}^{m} \mathrm{Var}(Y_i) + \sum_{i=1}^{m} \sum_{\substack{j=1 \\ i \neq j}}^{m} \mathrm{Cov}(Y_i, Y_j) \\
&= m \cdot \mathrm{Var}(Y_1) + m(m-1) \cdot \mathrm{Cov}(Y_1, Y_2) \\
&= m \cdot \left[\left(\frac{m-1}{m}\right)^n - \left(\frac{m-1}{m}\right)^{2n}\right] \\
&\quad + m(m-1) \cdot \left[\left(\frac{m-2}{m}\right)^n - \left(\frac{m-1}{m}\right)^{2n}\right].
\end{aligned}$$