TMA4240 Statistikk
Høst 2017

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Anbefalt øving 5
Løsningsskisse

Oppgave 1

a) Vi har at en gjennomlesing av teksten tilsvarer n repeterte forsøk, ett forsøk for hver skrivefeil i teksten. Hvert forsøk resulteterer i suksess (feilen oppdages) eller ikke-suksess (feilen oppdages ikke). Sannsynligheten for suksess er p, og denne er konstant for alle forsøkene. Vi må i tillegg anta at hvert ord leses uavhengig av alle andre ord i teksten, slik at forsøkene er uavhengige.

Vi har $\lambda = 2$ og $s = 8$ og ønsker å finne sannsynligheten for at antall trykkfeil, N, er større enn 10. Vi har $\mu = \lambda s = 2 \cdot 8 = 16$

$$
\begin{aligned}
P(N > 10) &= 1 - P(N \leq 10) \\
&= 1 - \sum_{i=1}^{10} P(N = i) \\
&= 1 - 0.0774 \\
&= 0.923.
\end{aligned}
$$

Vi har nå gitt $N = 12$ og $p = 0.6$ og œnsker å finne sannsynligheten for at korrekturleseren oppdager alle trykkfeilene.

$$
\begin{aligned}
P(X = 12 | N = 12) &= \binom{12}{12} \cdot 0.6^{12} \cdot 0.4^{0} \\
&= 0.6^{12} \\
&= 0.0022.
\end{aligned}
$$

b) Y_k = antall trykkfeil som gjenstår etter k uavhengige gjennomlesninger. Vi finner først simultanfordelingen til Y_1 og N.
Vi har

$$
P(X = x | N = n) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, ..., n
$$

Simultanfordelingen til Y_1 og N er da gitt ved

$$
\begin{aligned}
P(Y_1 = u, N = n) &= P(Y_1 = u|N = n) \cdot P(N = n) \\
&= P(N - X = u|N = n) \cdot P(N = n) \\
&= P(X = n - u|N = n) \cdot P(N = n) \\
&= \binom{n}{n-u} p^{n-u}(1-p)^u \cdot P(N = n)
\end{aligned}
$$

for $u = 0, 1, \ldots$ og $n = u, u + 1, \ldots$.

Vi finner deretter marginalfordelingen til Y_1.

$$
\begin{aligned}
P(Y_1 = u) &= \sum_{n=u}^{\infty} P(Y_1 = u, N = n) \\
&= \sum_{n=u}^{\infty} \binom{n}{n-u} p^{n-u}(1-p)^u e^{-\lambda s} \cdot \frac{(\lambda s)^n}{n!} \\
&= \sum_{n=0}^{\infty} \binom{n+u}{n} p^{n}(1-p)^u e^{-\lambda s} \cdot \frac{(\lambda s)^{n+u}}{(n+u)!} \\
&= \sum_{n=0}^{\infty} \frac{1}{n!u!} p^{n}(1-p)^u e^{-\lambda s} (\lambda s)^{n+u} \\
&= \frac{(\lambda s)^u e^{-\lambda s}}{u!}(1-p)^u \sum_{n=0}^{\infty} \frac{(\lambda p s)^n}{n!} \\
&= \frac{(\lambda s)^u e^{-\lambda s}}{u!}(1-p)^u e^{\lambda p s} \\
&= \frac{(\lambda s(1-p))^u}{u!} e^{-\lambda s(1-p)}.
\end{aligned}
$$

Vi ser at marginalfordelingen til Y_1 er $Y_1 \sim \text{Poisson}(\lambda s(1-p))$.

Oppgave 2

a) For hver deltaker har vi følgende situasjon:

- Deltakeren får en serie oppgaver.
- Hver runde har to mulige utfall: Deltakeren klarer ikke oppgaven og går ut av konkurransen (hendelse A), eller han/hun klarer oppgaven og går videre til neste runde (hendelse A').
- Sannsynligheten for ikke å klare oppgaven, $p = P(A)$, er lik i hver runde.
- Resultatene fra hver runde er uavhengige.

Denne situasjonen svarer til en Bernoulli-forsøksrekke, der vi ikke bestemmer antall forsøk på forhånd, men repeterer forsøket (gir nye oppgaver) inntil første gang hendelsen A (klarer ikke oppgaven) inntreffer. Siden X er antall forsøk inntil A inntreffer *første* gang (deltakeren første gang ikke klarer oppgaven), er det rimelig å anta at X er geometrisk fordelt.

Sannsynligheten for at deltakeren går ut i første runde:

$$P(X=1) = f(1) = p(1-p)^{1-1} = p = \underline{\underline{0.10}}$$

Sannsynligheten for at deltakeren fortsatt er med etter fem runder:

$$P(X>5) = 1 - P(X \le 5) = 1 - F(5) = 1 - (1-(1-p)^5) = (1-p)^5 = 0.90^5 = \underline{\underline{0.59}}.$$

Sannsynligheten for at deltakeren ikke klarer oppgaven i niende runde ($X=9$), dersom deltakeren klarer oppgavene til og med femte runde ($X>5$): Her bruker vi betinget sannsynlighet, og resultatet fra forrige spørsmål.

$$\begin{aligned} P(X=9 \mid X>5) &= \frac{P(X=9 \cap X>5)}{P(X>5)} = \frac{P(X=9)}{P(X>5)} = \frac{f(9)}{1-F(5)} \\ &= \frac{p(1-p)^{9-1}}{(1-p)^5} = p(1-p)^3 = 0.10 \cdot 0.90^3 = \underline{\underline{0.073}} \end{aligned}$$

b) Vi har følgende situasjon for hver oppgavelager:

- Resultater for et visst antall (n_1 eller n_2) deltakere blir registrert
- To mulig utfall: Deltakeren klarer færre enn fem oppgaver (hendelse C), eller ikke (dvs. klarer fem eller flere, hendelse C').
- Sannsynligheten for C er lik i for hver deltaker.
- Resultatene for hver deltaker er uavhengige.

Dette svarer til et binomisk forsøk, og Z_1 og Z_2 er dermed binomisk fordelte, med parametre som gitt i oppgaven.

Oppgave 3

a) X er en stokastisk variabel som beskriver antall korrekte svar på $n=20$ spørsmål på midtveiseksamen (flervalgsoppgave).

Betingelser for at X er binomisk fordelt:

- Vi ser på $n=20$ svar.
- For hvert svar sjekker vi om svaret er korrekt eller ikke.
- Sannsynligheten for at svaret er korrekt er $\frac{1}{m}$ fordi Ole tipper, og det er bare ett svar som er korrekt blant m mulig svar. Denne sannsynligheten er den samme for alle n svarene.
- De n svarene er uavhengige av hverandre, siden Ole tipper svaret på hvert spørsmål.

Under disse 4 betingelsene er X binomisk fordelt med parametere $n = 20$ og $p = \frac{1}{m}$. 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 å svare korrekt på minst 8 spørsmål finner vi enklest ved tabelloppslag (s 17 i formelsamlingen). Vi gjør dette for de tre verdiene av m som er oppgitt, $m = 2, 4, 5$. (Grunnen til at disse tre verdiene er valgt er kun pga at de finnes i tabellen...)

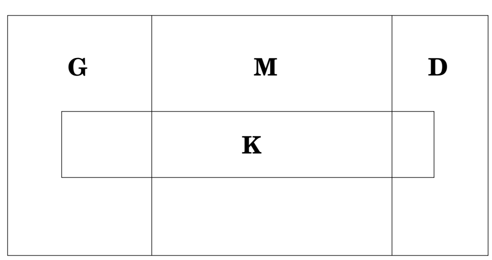
$$\begin{aligned}
m = 2 : P(X \geq 8) &= 1 - P(X \leq 7 | p = 0.5, n = 20) = 1 - 0.132 = \underline{\underline{0.87}} \\
m = 4 : P(X \geq 8) &= 1 - P(X \leq 7 | p = 0.25, n = 20) = 1 - 0.898 = \underline{\underline{0.10}} \\
m = 5 : P(X \geq 8) &= 1 - P(X \leq 7 | p = 0.2, n = 20) = 1 - 0.968 = \underline{\underline{0.03}}
\end{aligned}$$

Forventningen til X er $\mathrm{E}(X) = np$. Forventet antall korrekte svar for $m = 2, 4, 5$ blir da:

$$\begin{aligned}
m = 2 : \mathrm{E}(X) &= 20 \cdot \frac{1}{2} = 10 \\
m = 4 : \mathrm{E}(X) &= 20 \cdot \frac{1}{4} = 5 \\
m = 5 : \mathrm{E}(X) &= 20 \cdot \frac{1}{5} = 4
\end{aligned}$$

b) G = spørsmålet er fra den delen av pensum som Ole kan godt,
M = spørsmålet er fra den delen av pensum som Ole kan middels godt,
D = spørsmålet er fra den delen av pensum som Ole kan dårlig,
K = Ole svarer korrekt på spørsmålet.

Venndiagram for de fire hendelsene:



Sannsynligheten for at Ole svarer korrekt på et tilfeldig valgt spørsmål, $P(K)$, finner vi vet å bruke setningen om total sannsynlighet. Vi vet at G, M, D er en partisjon av utfallsrommet (det ser vi lett av venndiagrammet og at summen av sannsynlighetene er 1).

$$\begin{aligned}
P(K) &= P(K \cap G) + P(K \cap M) + P(K \cap D) \\
&= P(K|G) \cdot P(G) + P(K|M) \cdot P(M) + P(K|D) \cdot P(D) \\
&= 0.8 \cdot 0.3 + 0.4 \cdot 0.5 + 0.2 \cdot 0.2 = \underline{\underline{0.48}}
\end{aligned}$$

Bayes regel kan benyttes til å finne sannsynligheten for at spørsmålet var fra den delen av pensum som Ole kan dårlig, gitt at Ole svarte korrekt på spørsmålet.

$$\begin{aligned} P(D|K) &= \frac{P(K \cap D)}{P(K)} \\ &= \frac{P(K|D) \cdot P(D)}{P(K)} = \frac{0.2 \cdot 0.2}{0.48} = \underline{\underline{0.08}} \end{aligned}$$

Oppgave 4

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.