Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

TMA4240 Statistikk
Høst 2019

Anbefalt øving 11
Løsningsskisse

Oppgave 1

a) Målet er å finne forventningsverdi og varians til den tilfeldige variabelen X, som er poissonfordelt med parameter $\lambda\nu_0$. Her kan bruke at poissonfordelingen med parameter μ har både forventningsverdi og varians lik μ. I dette tilfellet har vi derfor

$$\mathrm{E}(X) = \underline{\underline{\lambda\nu_0}}, \quad \text{og} \quad \mathrm{Var}(X) = \underline{\underline{\lambda\nu_0}}.$$

Intensiteten λ er lik forventet antall kolibakterier per liter drikkevann,

$$\lambda = \frac{\mathrm{E}(X)}{\nu_0}$$

og er derfor et mål på forurensningsgrad.

b) Sannsynligheten for en tilfeldig valgt liter av drikkevannet er fri for kolibakterier, tilsvarer sannsynligheten for at $X = 0$ når $\nu_0 = 1$. Det er også oppgitt at $\lambda = 3$. Vi regner ut sannsynligheten ved å sette disse tallverdiene inn i uttrykket for punktsannsynligheten til X,

$$P(X = 0) = \frac{(\lambda\nu_0)^0}{0!}\exp(-\lambda\nu_0) = \frac{1}{1}\exp(-3 \cdot 1) = \underline{\underline{0.0498}}.$$

Det andre spørsmålet er hvor stor prøven må være for at den med sannsynlighet minst 0.9975 skal inneholde én eller flere kolibakterier. For å svare på dette trenger vi sannsynligheten for at $X \geq 1$. Siden dette er komplementærhendelsen til $X = 0$ har vi

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{(\lambda\nu_0)^0}{0!}\exp(-\lambda\nu_0) = 1 - \exp(-\lambda\nu_0).$$

Sannsynligheten $P(X \geq 1)$ er en voksende funksjon av ν_0, og vi ønsker å finne verdien ν_c slik at $P(X \geq 1) \geq 0.9975$ for enhver $\nu_0 \geq \nu_c$. Vi får dermed ulikheten

$$\begin{aligned} 1 - \exp(-\lambda\nu_0) &\geq 0.9975 \\ 1 - 0.9975 &\geq \exp(-\lambda\nu_0) \\ \ln 0.0025 &\geq -\lambda\nu_0 \\ \nu_0 &\geq -\frac{\ln 0.0025}{\lambda}, \end{aligned}$$

og ved å sette inn $\lambda = 3$ får vi

$$\nu_0 \geq -\frac{\ln 0.0025}{3} = \underline{\underline{1.9972}}.$$

For at prøven skal inneholde én eller flere bakterier med en sannsynlighet på minst 0.9975, trengs det altså knappe to liter vann.

c) Vi blir her bedt om å sammenligne egenskapene til de to estimatorene, og deretter velge den ene over den andre basert på sammenligningen. Vi begynner med å finne forventningsverdiene til λ^* og $\hat{\lambda}$,

$$\begin{aligned}
\mathrm{E}(\lambda^*) &= \mathrm{E}\left(\frac{\nu_1 X_1 + \nu_2 X_2}{\nu_1^2 + \nu_2^2}\right) \\
&= \frac{\nu_1 \mathrm{E}(X_1) + \nu_2 \mathrm{E}(X_2)}{\nu_1^2 + \nu_2^2} \\
&= \frac{\nu_1 \cdot \lambda\nu_1 + \nu_2 \cdot \lambda\nu_2}{\nu_1^2 + \nu_2^2} \\
&= \frac{\lambda(\nu_1^2 + \nu_2^2)}{\nu_1^2 + \nu_2^2} = \underline{\lambda},
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(\hat{\lambda}) &= \mathrm{E}\left(\frac{X_1 + X_2}{\nu_1 + \nu_2}\right) \\
&= \frac{\mathrm{E}(X_1) + \mathrm{E}(X_2)}{\nu_1 + \nu_2} \\
&= \frac{\lambda\nu_1 + \lambda\nu_2}{\nu_1 + \nu_2} \\
&= \frac{\lambda(\nu_1 + \nu_2)}{\nu_1 + \nu_2} = \underline{\lambda}.
\end{aligned}$$

Begge estimatorene er forventningsrette, siden $\mathrm{E}(\lambda^*) = \mathrm{E}(\hat{\lambda}) = \lambda$. Vi fortsetter derfor med å regne ut variansene til λ^* og $\hat{\lambda}$,

$$\begin{aligned}
\mathrm{Var}(\lambda^*) &= \mathrm{Var}\left(\frac{\nu_1 X_1 + \nu_2 X_2}{\nu_1^2 + \nu_2^2}\right) \\
&= \left(\frac{\nu_1}{\nu_1^2 + \nu_2^2}\right)^2 \mathrm{Var}(X_1) + \left(\frac{\nu_2}{\nu_1^2 + \nu_2^2}\right)^2 \mathrm{Var}(X_2) \\
&= \left(\frac{\nu_1}{\nu_1^2 + \nu_2^2}\right)^2 \lambda\nu_1 + \left(\frac{\nu_2}{\nu_1^2 + \nu_2^2}\right)^2 \lambda\nu_2 \\
&= \frac{\nu_1^3}{(\nu_1^2 + \nu_2^2)^2}\lambda + \frac{\nu_2^3}{(\nu_1^2 + \nu_2^2)^2}\lambda \\
&= \underline{\underline{\frac{\nu_1^3 + \nu_2^3}{(\nu_1^2 + \nu_2^2)^2}\lambda}},
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(\hat{\lambda}) &= \mathrm{Var}\left(\frac{X_1 + X_2}{\nu_1 + \nu_2}\right) \\
&= \left(\frac{1}{\nu_1 + \nu_2}\right)^2 \mathrm{Var}(X_1) + \left(\frac{1}{\nu_1 + \nu_2}\right)^2 \mathrm{Var}(X_2) \\
&= \left(\frac{1}{\nu_1 + \nu_2}\right)^2 \lambda\nu_1 + \left(\frac{1}{\nu_1 + \nu_2}\right)^2 \lambda\nu_2 \\
&= \frac{\lambda\nu_1}{(\nu_1 + \nu_2)^2} + \frac{\lambda\nu_2}{(\nu_1 + \nu_2)^2} \\
&= \frac{\lambda(\nu_1 + \nu_2)}{(\nu_1 + \nu_2)^2} \\
&= \underline{\underline{\frac{1}{\nu_1 + \nu_2}\lambda.}}
\end{aligned}$$

Forholdet mellom variansene til λ^* og $\hat{\lambda}$ er

$$\begin{aligned}
\frac{\mathrm{Var}(\lambda^*)}{\mathrm{Var}(\hat{\lambda})} &= \frac{\lambda(\nu_1^3 + \nu_2^3)/(\nu_1^2 + \nu_2^2)^2}{\lambda/(\nu_1 + \nu_2)} \\
&= \frac{(\nu_1^3 + \nu_2^3)(\nu_1 + \nu_2)}{(\nu_1^2 + \nu_2^2)^2} \\
&= \frac{\nu_1^4 + \nu_2^4 + \nu_1^3\nu_2 + \nu_1\nu_2^3}{\nu_1^4 + 2\nu_1^2\nu_2^2 + \nu_2^4} \\
&= \frac{(\nu_1^4 + \nu_2^4) + (\nu_1^3\nu_2 + \nu_1\nu_2^3)}{(\nu_1^4 + \nu_2^4) + 2\nu_1^2\nu_2^2}.
\end{aligned}$$

I dette uttrykket er telleren større enn nevneren, siden

$$\begin{aligned}
\nu_1^3\nu_2 + \nu_1\nu_2^3 - 2\nu_1^2\nu_2^2 &= \nu_1\nu_2(\nu_1^2 + \nu_2^2) - 2\nu_1^2\nu_2^2 \\
&= \nu_1\nu_2(\nu_1^2 - 2\nu_1\nu_2 + \nu_2^2) \\
&= \nu_1\nu_2(\nu_1 - \nu_2)^2 > 0
\end{aligned}$$

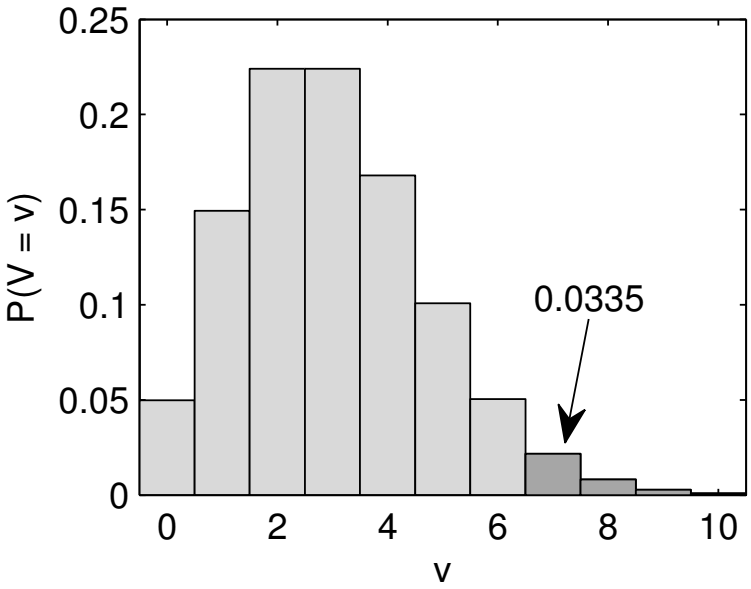
hvor ulikheten holder for ikkenegative volumer $\nu_1, \nu_2 \geq 0$. Dette betyr at $\mathrm{Var}(\lambda^*) \geq \mathrm{Var}(\hat{\lambda})$, og vi foretrekker derfor $\hat{\lambda}$ som estimator for λ.

d) Vi skal sette opp nullhypotese H_0 og alternativ hypotese H_1 for å undersøke om det er grunnlag for å påstå at $\lambda > 1$. Vi må da velge $H_1 :\ \lambda > 1$, og H_0 blir dermed $\lambda = 1$. Det vil si at vi skal teste

$$H_0 :\ \lambda = 1 \quad \text{og} \quad H_1 :\ \lambda > 1.$$

For å kunne teste hypotesene, trenger vi en testobservator med kjent sannsynlighetsfordeling. Vi kan bruke at en sum av uavhengige, poissonfordelte tilfeldige variable også er poissonfordelt, og parameteren vil være lik summen av parameterne til variablene som inngår i summen. Konkret har vi at siden $X_1 \sim \mathrm{Po}(\lambda\nu_1)$ og $X_2 \sim \mathrm{Po}(\lambda\nu_2)$, og X_1 og X_2 er uavhengige, så vil

$$V = (\nu_1 + \nu_2)\hat{\lambda} = X_1 + X_2 \sim \mathrm{Po}(\lambda\nu_1 + \lambda\nu_2) = \mathrm{Po}(\lambda(\nu_1 + \nu_2)).$$



Figur 1: Punktsannsynligheten $P(V = v)$ for $0 \leq v \leq 10$ til poissonfordelingen med parameter 3. Hver søyle er 1 enhet bred, slik at det totale arealet for $v = 0, 1, 2, \ldots$ er lik 1. En del av arealet som tilsvarer $P(V > 6)$ er markert.

Hvis nullhypotesen er sann vil vi ha $\lambda = 1$, slik at $V \sim \text{Po}(\nu_1 + \nu_2)$. Det er rimelig å forkaste H_0 dersom $V > k$ der k velges minst mulig slik at

$$P(\text{Forkast } H_0 | H_0 \text{ er riktig}) \leq \alpha$$
$$P(V > k | \lambda = 1) \leq \alpha.$$

Fra tabell over poissonfordelingen med parameter $\nu_1 + \nu_2$ får vi at

$$P(V > v | \lambda = 1) = 1 - P(V \leq v | \lambda = 1)$$

for ulike verdier av v, og vi kan dermed finne verdien til k. Hvilken verdi vi her får vil avhenge av prøvevolumene ν_1 og ν_2.

e) Her får vi oppgitt volumene $\nu_1 = 1$ og $\nu_2 = 2$, samt observasjonene $x_1 = 1$ og $x_2 = 5$ av de tilfeldige variablene X_1 og X_2. Med disse tallverdiene vil testobservatoren være poissonfordelt med parameter $\nu_1 + \nu_2 = 1 + 2 = 3$. For denne fordelingen, som er illustrert i figur 1, har vi

$$P(V > 5) = 1 - P(V \leq 5) = 1 - 0.9161 = 0.0839$$
$$P(V > 6) = 1 - P(V \leq 6) = 1 - 0.9665 = 0.0335$$

slik at den kritiske verdien på signifikansnivå $\alpha = 0.05$ er $k = 6$, og vi forkaster nullhypotesen på dette signifikansnivået hvis vi observerer at $V > k$. Den observerte verdien av testobservatoren er $v_{\text{obs}} = x_1 + x_2 = 1 + 5 = 6$. Dette er ikke større enn k, så nullhypotesen forkastes ikke.

Til slutt skal vi finne p-verdien til resultatet, det vil si sannsynligheten for å observere et like ekstremt eller mer ekstremt utfall, altså

$$p = P(V \geq v_{\text{obs}}) = P(V > v_{\text{obs}} - 1) = P(V > 5) = \underline{\underline{0.0839}}.$$

f) Hvis $\lambda = 3$ og prøvene er like store som i d), blir parameteren i fordelingen til V lik

$$\lambda(\nu_1 + \nu_2) = 3 \cdot (1 + 2) = 9.$$

For at testen ikke skal gi forkastning, må vi observere $V \leq k = 6$. Tabelloppslag gir sannsynligheten

$$P(V \leq 6) = \underline{\underline{0.2068}},$$

når $V \sim \text{Po}(9)$.

Oppgave 2

a) Expected value in the binomial distribution is $E(X) = np = 20 \cdot 0.8 = \underline{\underline{16}}$.

$P(X > 16) = 1 - P(X \leq 16) = 1 - 0.59 = \underline{\underline{0.41}}$.

$P(X = 20|X > 16) = \frac{P(X=20)}{P(X>16)} = \frac{0.8^{20}}{0.41} = \frac{0.0115}{0.41} = \underline{\underline{0.028}}$.

b) $H_0 : p = 0.8$, $H_1 : p > 0.8$.

Normal approximation means $Z = \frac{X - E(X)}{\sqrt{\text{Var}(X)}} = \frac{X - np}{\sqrt{np(1-p)}} = \frac{X - 20 \cdot 0.8}{\sqrt{20 \cdot 0.8 \cdot 0.2}}$ is assumed Gaussian.

We reject H_0 when the observed fraction of success is significantly large.

Setting $\alpha = 0.1$ the rejection value is $z_{0.1} = \underline{1.28}$. We get $X = 18$ and $Z = \frac{18-16}{\sqrt{3.2}} = \underline{1.12}$. Since $Z < z_{0.1}$, the observed number is not significantly large. $\underline{\underline{\text{We do not reject } H_0.}}$

Half correction could also be used in this exercise, and gives more accurate results. In that case we get $Z = \frac{17.5-16}{\sqrt{3.2}} = 0.8385$. Full score is given to both methods.

c) The hypothesis test is rejected when the observation is extreme. In particular, we reject if the observation is larger than c, and the critical level c is determined as the smallest c that satisfies $P(X > c|p = 0.8) \leq \alpha$. For various c we get: $P(X > 19|p = 0.8) = 0.011$, $P(X > 18|p = 0.8) = 0.07$, $P(X > 17|p = 0.8) = 0.21$. At significance level $\alpha = 0.1$ we reject when $X > 18$.

Based on the observed value $X = 18$, the P value is

$$P(X \geq 18|p = 0.8) = 1 - P(X \leq 17|p = 0.8) = \underline{\underline{0.21}}.$$

The power is the probability of rejection, given that the parameter is $H_1 : p = 0.9$:

$P(\text{reject}|p = 0.9) = P(X > 18|p = 0.9) = 1 - P(X \leq 18|p = 0.9) = 1 - 0.61 = \underline{\underline{0.39}}$.

Oppgave 3

a) Sannsynligheten for å få 5 kron er

$$P(5 \text{ kron}) = \frac{1}{2^5} = 1/32 = \underline{\underline{0.031}}.$$

Sannsynligheten for å få 3 kron er lik punktsannsynligheten $P(X = 3)$ der X er binomisk fordelt med parametre $n = 5$ og $p = 0.5$, altså

$$P(X = 3) = \binom{5}{3} 0.5^3 \cdot (1 - 0.5)^{5-3} = 10 \cdot 0.5^3 \cdot 0.5^2 = \underline{\underline{0.3125}}.$$

Fire kron på rad kan inntreffe på 3 forskjellige måter: Kron på alle 5 kastene, kron på de første 4 kastene, og mynt på siste, eller mynt på første kast og kron på de 4 siste. Antall mulige utfall av de fem kastene er $2^5 = 32$, og alle er like sannsynlige, så sannsynligheten for å få fire kron på rad er

$$P(4 \text{ kron på rad}) = \frac{3}{32} = \underline{\underline{0.0938}}.$$

b) Sannsynligheten for at lengste sekvens har lengde 5 eller 6 kan anslås ved å regne ut andelen utfall hvor lengste sekvens var på 5 eller 6 kast, av de 10000 simulasjonene. Fra figuren leser vi av at lengste sekvens hadde lengde 5 i omtrent 2700 tilfeller, og lengde 6 i omtrent 1700 tilfeller, og vi får estimatet

$$\widehat{P(5 \text{ eller } 6)} = \frac{2700 + 1700}{10000} = \underline{\underline{0.44}}.$$

I Miriams myntkastsekvens har den lengste uavbrutte sekvensen av kron lengde 2. For en tilfeldig generert myntkastsekvens av lengde 30, vil lengden av lengste uavbrutte sekvens av kron ha en sannsynlighetsfordeling som er svært lik den i figuren. At denne lengden er så lav som 2 er ganske usannsynlig, og Miriams myntkastsekvens er dermed mistenkelig.

Vi vil teste nullhypotesen

$$H_0: \text{ Sekvensen er tilfeldig generert}$$

mot den alternative hypotesen

$$H_1: \text{ Sekvensen er ikke tilfeldig generert.}$$

Vi antar at under nullhypotesen er lengden av lengste sammenhengende sekvens av kron fordelt som i figuren. For å avgjøre om nullhypotesen skal forkastes eller ikke, regner vi ut p-verdien, altså sannsynligheten for å observere et like ekstremt eller mer ekstremt utfall. Her er dette lik sannsynligheten for at lengste uavbrutte sekvens av kron er 0, 1 eller 2. Utfra figuren ser det ut som om antall utfall i søylene for 0, 1 og 2 er henholdsvis 0, 0 og 25. Vi får dermed følgende estimat for p-verdien:

$$\widehat{P(0, 1 \text{ eller } 2)} = \frac{25}{10000} = 0.0025.$$

Dette er en lav p-verdi som tilsier at nullhypotesen forkastes f.eks. på signifikansnivå 0.05. Det er altså grunn til å hevde at Miriam har funnet på tallene.