Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

TMA4245 Statistikk
Vår 2015

Øving nummer 10, blokk II
Løsningsskisse

Oppgave 1

a) Sannsynligheten for at en tilfeldig valgt person svarer korrekt på spørsmålet er lik sannsynligheten for at tilfeldig valgt person kaster 4 eller lavere på terningen, det vil si

$$P(K) = P(\{1,2,3,4\}) = \frac{4}{6} = \frac{2}{3}.$$

Tilsvarende vil sannsynligheten for komplementærhendelsen, altså at en tilfeldig valgt person svarer feil på spørsmålet, være lik sannsynligheten for å kaste 5 eller høyere,

$$P(K^c) = P(\{5,6\}) = \frac{2}{6} = \frac{1}{3}.$$

Siden enhver person enten har gjort eller ikke har gjort W i 1994, vil hendelsene W, W^c utgjøre en partisjon av utfallsrommet. Vi har dermed $J = (J \cap W) \cup (J \cap W^c)$, slik at sannsynligheten for at en tilfeldig valgt person svarer ja på spørsmålet kan uttrykkes som

$$\begin{aligned} P(J) &= P(J|W)P(W) + P(J|W^c)P(W^c) \\ &= P(K) \cdot q + P(K^c) \cdot (1-q) \\ &= \frac{2q}{3} + \frac{1-q}{3} \\ &= \frac{1+q}{3}. \end{aligned}$$

Sannsynligheten $P(W|J)$ for at en tilfeldig valgt person har gjort W i 1994, gitt at samme person svarer ja på spørsmålet, finner vi fra Bayes' regel,

$$\begin{aligned} P(W|J) &= \frac{P(J|W)P(W)}{P(J)} = \frac{P(K) \cdot q}{(1+q)/3} \\ &= \frac{2q/3}{(1+q)/3} = \frac{2q}{1+q}. \end{aligned}$$

b) Sannsynligheten q har verdiområde $[0,1]$. Fra sammenhengen

$$p = P(J) = \frac{1+q}{3}$$

funnet i forrige deloppgave, ser vi at

$$0 \leq q \leq 1 \implies \frac{1}{3} \leq p \leq \frac{2}{3}$$

slik at verdiområdet til p er $[1/3, 2/3]$.

Tellevariabelen X er binomisk fordelt med parametre n og p,

$$X \sim \text{Bin}(n, p).$$

Dette kan begrunnes som følger:

- Hver av de n respondentene vil enten svare ja, eller ikke svare ja.
- De n svarene er uavhengige av hverandre.
- Hver respondent svarer ja med samme sannsynlighet p.

c) Punktsannsynligheten til X er gitt ved

$$P(X = x; p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Likelihood-funksjonen $L(p)$ defineres ved å fiksere x og betrakte uttrykket som en funksjon av p,

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

Fra likelihood-funksjonen defineres log-likelihood-funksjonen,

$$\ell(p) = \ln L(p) = \ln \binom{n}{x} + x \ln p + (n-x) \ln(1-p).$$

Derivasjon med hensyn på p gir

$$\frac{\partial \ell(p)}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p}.$$

Maximum likelihood-estimatoren $\hat{p}$ finnes ved å sette den deriverte lik null, og løse for p,

$$\frac{\partial \ell(p)}{\partial p} = 0 \implies p = \frac{x}{n}.$$

Dette betyr at estimatoren er $\hat{p} = X/n$.

Siden q kan uttrykkes som en funksjon av p, kan vi finne maximum likelihood-estimatoren $\hat{q}$ ved å sette inn $\hat{p}$ i funksjonen,

$$p = \frac{1+q}{3} \implies q = 3p - 1 \implies \hat{q} = 3\hat{p} - 1 = \frac{3X}{n} - 1.$$

Oppgave 2 $E(\hat{\mu}) = \mu$

$Var(\hat{\mu}) = \frac{\tau_0^4 Var(X) + \sigma_0^4 Var(Y)}{(\tau_0^2+\sigma_0^2)^2} = \frac{\tau_0^4\sigma_0^2 + \sigma_0^4\tau_0^2}{(\tau_0^2+\sigma_0^2)^2} = \frac{\sigma_0^2\tau_0^2}{\tau_0^2+\sigma_0^2}$

$\hat{\mu}$ er en lineærkombinasjon av normalfordelte variable, og er dermed normalfordelt.
$Z = \frac{\hat{\mu}-\mu}{\sqrt{\frac{\sigma_0^2\tau_0^2}{\tau_0^2+\sigma_0^2}}} \sim N(0,1)$

$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$
$P(\hat{\mu} - z_{\alpha/2}\sqrt{\frac{\sigma_0^2\tau_0^2}{\tau_0^2+\sigma_0^2}} < \mu < \hat{\mu} + z_{\alpha/2}\sqrt{\frac{\sigma_0^2\tau_0^2}{\tau_0^2+\sigma_0^2}}) = 1-\alpha$

Ett $1-\alpha$ konfidensintervall for μ er da :
$(\hat{\mu} - z_{\alpha/2}\sqrt{\frac{\sigma_0^2\tau_0^2}{\tau_0^2+\sigma_0^2}}, \hat{\mu} + z_{\alpha/2}\sqrt{\frac{\sigma_0^2\tau_0^2}{\tau_0^2+\sigma_0^2}})$

Oppgave 3

a) Antagelser for at X er binomisk fordelt:

- Gjør n forsøk: Spør n personer.
- Registrerer suksess eller fiasko i hvert forsøk: Får svaret JA eller ikke JA (nei eller vet ikke) i hvert forsøk.
- P(suksess) lik i alle forsøk: Sannsynlighet for JA er p for alle som blir spurt.
- Forsøka er uavhengige: Rimelig å anta at de som blir spurt svarer uavhengig av hverandre.

$P(X \geq 18) = 1 - P(X < 18) = 1 - P(X \leq 17) \overset{\text{tabell}}{=} 1 - 0.965 = 0.035.$

$P(10 < X < 15) = P(X \leq 14) - P(X \leq 10) \overset{\text{tabell}}{=} 0.584 - 0.048 = 0.536$

b)
- $E(\hat{P}) = p$ og $\text{Var}(\hat{P}) = \frac{1}{4}(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)$.
- $E(P^*) = p$ og $\text{Var}(P^*) = \frac{1}{n_1+n_2}p(1-p)$.

Egenskaper for god estimator: forventningsrett og liten varians. Begge estimatorene er forventningsrette, men P^* har minst varians, vi velger derfor P^*.

La $\alpha = 0.05$. Siden $\frac{\hat{P}-p}{\sqrt{\frac{1}{2n}\hat{P}(1-\hat{P})}}$ er tilnærmet standardnormalfordelt får vi:

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\hat{P}-p}{\sqrt{\frac{1}{2n}\hat{P}(1-\hat{P})}} < z_{\frac{\alpha}{2}}\right) \approx 1-\alpha$$
$$P\left(\hat{P} - z_{\frac{\alpha}{2}}\sqrt{\frac{1}{2n}\hat{P}(1-\hat{P})} < p < \hat{P} + z_{\frac{\alpha}{2}}\sqrt{\frac{1}{2n}\hat{P}(1-\hat{P})}\right) \approx 1-\alpha$$

Et tilnærmet 95% konfidensintervall for p blir da:

$$\left[\hat{p} - z_{0.025}\sqrt{\frac{1}{2n}\hat{p}(1-\hat{p})}, \hat{p} + z_{0.025}\sqrt{\frac{1}{2n}\hat{p}(1-\hat{p})}\right].$$

c) Vi har at

$$Y = X_3 - n\hat{P} = X_3 - n\frac{X_1 + X_2}{2n} = X_3 - \frac{1}{2}X_1 - \frac{1}{2}X_2.$$

Siden n er stor og p ikke nær 0 og 1, vil vi ha at $np > 5$ og $n(1-p) > 5$, slik at vi kan bruke normaltilnærming til binomisk fordeling. Vi kan dermed anta at X_1, X_2 og X_3 alle er tilnærmet normalfordelt, de er uavhengige, og lineærkombinasjonen Y er dermed også tilnærmet normalfordelt.

$\text{Var}(Y) = \text{Var}(X_3 - n\hat{P}) \overset{\text{uavh.}}{=} \text{Var}(X_3) + n^2\text{Var}(\hat{P}) \overset{b)}{=} np(1-p) + n^2\frac{1}{2n}p(1-p) = \frac{3}{2}np(1-p)$.

Har da at

- $X_3 - n\hat{P}$ er tilnærmet normalfordelt
- $\text{Var}(X_3 - n\hat{P}) = \frac{3}{2}np(1-p)$
- $\text{E}(X_3 - n\hat{P}) = E(X_3) - nE(\hat{P}) = np - np = 0$

Vi får da et prediksjonsintervall ved:

$$P\left(-z_{\frac{\alpha}{2}} < \frac{X_3 - n\hat{P}}{\sqrt{\frac{3}{2}np(1-p)}} < z_{\frac{\alpha}{2}}\right) \approx 1-\alpha$$

$$P\left(n\hat{P} - z_{\frac{\alpha}{2}}\sqrt{\frac{3}{2}np(1-p)} < X_3 < n\hat{P} + z_{\frac{\alpha}{2}}\sqrt{\frac{3}{2}np(1-p)}\right) \approx 1-\alpha$$

Siden n er stor, vil variansen til $\hat{P}$ være liten, og $\hat{P}$ være en god estimator for p. Vi kan derfor erstatte p med estimatet $\hat{p}$ i uttrykket for intervallgrensene.

Intervallet blir: $[n\hat{p} - z_{0.025}\sqrt{\frac{3}{2}n\hat{p}(1-\hat{p})}, n\hat{p} + z_{0.025}\sqrt{\frac{3}{2}n\hat{p}(1-\hat{p})}]$

Innsatt verdier blir intervallet $[633, 704]$.

Oppgave 4

a) $X \sim bin(n,p)$ fordi:

- Undersøker n uavhengige delar av DNA-strukturen.
- Finn for kvar del ut om denne delen av DNA-stukturen er samanfallande eller ikkje.
- Sannsynet for samanfallande er det same for alle delane ($P(\text{samanfell}) = p = 0.15$).

$$\begin{aligned}
P(X=2) &= \binom{5}{2}0.15^2(1-0.15)^{5-2} = \underline{\underline{0.138}} \\
P(X \geq 2) &= 1 - P(X<2) = 1 - P(X \leq 1) = 1 - 0.835 = \underline{\underline{0.165}} \\
P(X=2|X \geq 2) &= \frac{P(X=2 \cap X \geq 2)}{P(X \geq 2)} = \frac{P(X=2)}{P(X \geq 2)} \\
&= \frac{0.138}{0.165} = \underline{\underline{0.836}}
\end{aligned}$$

b)

$$
\begin{aligned}
P(\text{Type-I-feil}) &= P(\text{forkaste } H_0|H_0) \\
&= P(X=5|p=0.15) = 0.15^5 = \underline{\underline{0.000076}} \\
P(\text{Type-II-feil}) &= P(\text{ikkje forkaste } H_0|H_1) \\
&= P(X<5|p=1) = \underline{\underline{0}}
\end{aligned}
$$

Ser så på det generelle uttrykket for sannsynet for type-I-feil når vi har n forsøk. Finn derfrå kor stor n må vere for å oppnå ønska sannsyn for type-I-feil.

$$
\begin{aligned}
P(\text{Type-I-feil}) = P(X=n|p=0.15) = 0.15^n &< 0.000001 \\
n\ln(0.15) &< \ln(0.000001) \\
n &> \frac{\ln(0.000001)}{\ln(0.15)} = 7.28
\end{aligned}
$$

Minst $\underline{\underline{\text{8 delar}}}$ frå DNA-strukturen må undersøkast dersom sannsynet for type-I-feil skal vere mindre enn 0.000001.