TMA4240 Statistikk
Høst 2020

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Anbefalte oppgaver 10

Oppgave 1

I forkant av et stortingsvalg blir det gjennomført en meningsmåling der et representativt utvalg av velgerne blir spurt om de ønsker et regjeringsskifte eller ikke. Anta at andelen av velgerne som ønsker et skifte er p, og la X være antall personer blant n spurte som svarer JA på spørsmålet "Ønsker du et regjeringsskifte ved høstens valg?".

a) Under hvilke antagelser vil X her være binomisk fordelt? Du må relatere antagelsene til situasjonen som er beskrevet i oppgaveteksten.

Anta i resten av dette punktet at andelen av velgerne som ønsker et regjeringsskifte, er $p = 0.7$, og at $n = 20$ personer blir spurt. Bruk at X er binomisk fordelt.

Hva er sannsynligheten for at 18 eller flere av de 20 spurte svarer JA på spørsmålet om regjeringsskifte?

Hva er sannsynligheten for at flere enn 10, men færre enn 15, av de 20 sier JA?

Anta at to aviser på en bestemt dag presenterer resultater fra to meningsmålinger, gjennomført av hvert sitt meningsmålingsinstitutt, Byrå A og Byrå B. La n_1 være antall spurte og X_1 antall som svarer JA i målingen fra Byrå A, og n_2 og X_2 tilsvarende størrelser for Byrå B. Vi antar at X_1 er binomisk fordelt med parametre n_1 og p, og X_2 er binomisk fordelt med parametre n_2 og p, og at X_1 og X_2 er uavhengige.

Vi ønsker å estimere p ved å kombinere resultatene fra de to målingene. To aktuelle estimatorer er

$$\begin{aligned} \hat{P} &= \frac{1}{2}\left(\frac{X_1}{n_1} + \frac{X_2}{n_2}\right) \quad \text{og} \\ P^* &= \frac{X_1 + X_2}{n_1 + n_2}. \end{aligned}$$

b) Finn forventning og varians til hver av de to estimatorene $\hat{P}$ og P^*.

Dersom $n_1 = 500$ og $n_2 = 1000$, hvilken estimator vil du da velge? Begrunn svaret.

Anta nå at $n_1 = n_2 = n$, slik at X_1 og X_2 er uavhengige og binomisk fordelte, med *samme* parametre p og n. Dette medfører at

$$\hat{P} = P^* = \frac{X_1 + X_2}{2n}.$$

Utled et tilnærmet 95% konfidensintervall for p ved å bruke at fordelingen til

$$\frac{\hat{P} - p}{\sqrt{\frac{1}{2n}\hat{P}(1-\hat{P})}}$$

er tilnærmet standard normalfordelt.

Et tredje meningsmålingsinstitutt, Byrå C, har annonsert at de snart kommer med resultater fra en tilsvarende måling med n_3 spurte. La X_3 være antall som svarer JA på spørsmålet om regjeringsskifte i målingen fra Byrå C, og anta at X_3 er uavhengig av X_1 og X_2. Vi vil nå bruke resultatene fra Byrå A og Byrå B til å predikere hvor mange som svarer JA i den nye målingen. Vi antar i resten av oppgaven at $n_1 = n_2 = n_3 = n = 1000$, og at observerte verdier for X_1 og X_2 er $x_1 = 645$ og $x_2 = 692$.

c) La $Y = X_3 - n\hat{P}$, der $\hat{P} = \dfrac{X_1 + X_2}{2n}$.

Begrunn at det i vår situasjon er rimelig å anta at Y er tilnærmet normalfordelt, og vis at variansen til Y er $\frac{3}{2}np(1-p)$.

Bruk dette til å utlede et tilnærmet 95% prediksjonsintervall for antallet spurte som i målingen fra Byrå C svarer JA på spørsmålet om regjeringsskifte.

Bestem også intervallet numerisk basert på de observerte verdiene.

Oppgave 2

Levetiden T (målt i antall døgn) til en ny type ventiler som eventuelt skal benyttes på oljeplattformer i Nordsjøen, skal undersøkes. Det er velkjent at levetiden påvirkes av blant annet temperatur og trykk der ventilen benyttes, og av den kjemiske sammensetningen av olja som går gjennom ventilene. Anta at effekten av disse faktorene måles som en *stress*-faktor z, og at en vet av erfaring at kumulativ fordelingsfunksjon for levetiden T er gitt ved

$$F(t) = \mathrm{P}(T \leq t) = \begin{cases} 1 - e^{-\frac{zt^2}{\theta}} & \text{for } t > 0, \\ 0 & \text{ellers,} \end{cases}$$

der θ er en ukjent parameter. Parameteren θ er altså karakteristisk for en bestemt type ventiler, mens z beskriver miljøet der ventilen benyttes.

a) Når $z = 2.0$ og $\theta = 2 \cdot 10^6$, finn sannsynlighetene

$$\mathrm{P}(T > 1000) \quad \text{og} \quad \mathrm{P}(T > 2000 | T > 1000).$$

La T_1, T_2 og T_3 være levetidene til tre ventiler som alle opererer under forhold med $z = 2.0$, og anta at T_1, T_2 og T_3 er uavhengige stokastiske variabler. Når $\theta = 2 \cdot 10^6$, finn sannsynligheten for at minst to av de tre levetidene er større enn 1000 døgn.

b) Vis at sannsynlighetstettheten til T er gitt ved

$$f(t) = \begin{cases} \frac{2zt}{\theta} e^{-\frac{zt^2}{\theta}} & \text{for } t > 0, \\ 0 & \text{ellers.} \end{cases}$$

Skisser $f(t)$ for $t \in [0, 3000]$ når $z = 2.0$ og $\theta = 2 \cdot 10^6$, og skraver i denne skissen arealet som er lik sannsynligheten $\mathrm{P}(T > 1000)$.

La en stokastisk variabel V være definert ved

$$V = \frac{2zT^2}{\theta}.$$

c) Bruk transformasjonsformelen til å vise at V er χ^2-fordelt med 2 frihetsgrader.

Benytt så dette til å vise at $\mathrm{E}\left(T^2\right) = \frac{\theta}{z}$ og $\mathrm{Var}\left(T^2\right) = \left(\frac{\theta}{z}\right)^2$.

For å undersøke kvaliteten på denne type ventiler har man prøvd ut $n = 10$ ventiler. La $z_1, z_2, \ldots, z_n$ betegne stress-faktorene som disse ventilene opererer under, og la $T_1, T_2, \ldots, T_n$ betegne tilhørende levetider, og anta at $T_1, T_2, \ldots, T_n$ er uavhengige stokastiske variabler. De observerte levetidene er gitt i følgende tabell:

Ventil i	1	2	3	4	5	6	7	8	9	10
z_i	1.0	3.4	1.9	2.4	1.2	4.0	3.2	2.2	1.4	3.2
t_i	1297.2	834.2	1265.8	331.7	1937.8	727.6	869.6	746.7	1965.3	280.9

Det oppgis at $\sum_{i=1}^{n} z_i t_i^2 = 23\ 287\ 125$.

d) Utled sannsynlighetsmaksimeringsestimatoren (*maximum likelihood*-estimatoren) for θ og vis at den kan skrives på formen

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} z_i T_i^2.$$

Er $\widehat{\theta}$ forventningsrett? Finn $\mathrm{Var}(\widehat{\theta})$.

e) Begrunn at

$$U = \sum_{i=1}^{n} \frac{2z_i T_i^2}{\theta} \sim \chi^2_{2n}.$$

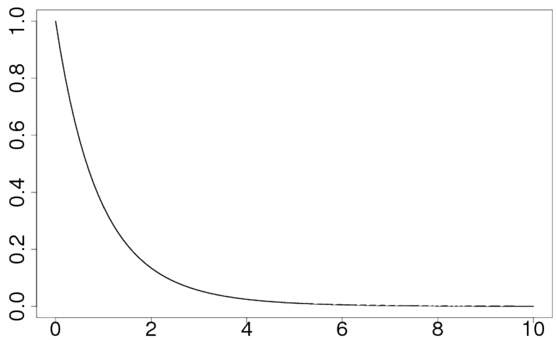
Benytt så dette til å utlede et $(1-\alpha)\cdot 100\%$-konfidensintervall for θ. Hva blir konfidensintervallet når dataene er som gitt over og $\alpha = 0.05$?

Ut fra de observerte dataene ønsker man også å lage et prediksjonsintervall for levetiden, T_0, til en ny ventil som skal operere under forhold med stress-faktor lik z_0. For å gjøre dette kan man ta utgangspunkt i den stokastiske variabelen

$$Y = n \cdot \frac{\frac{2T_0^2 z_0}{\theta}}{\sum_{i=1}^{n} \frac{2z_i T_i^2}{\theta}},$$

der teller og nevner i brøken altså er uavhengige stokastiske variabler. Teller og nevner er begge χ^2-fordelt og har henholdsvis 2 og $2n$ frihetsgrader. Når $n = 10$ kan det vises at sannsynlighetstettheten til Y blir som vist i figur 1. Merk at noen kvantiler i denne fordelinga er oppgitt til høyre i samme figur.

f) Utled et 90% prediksjonsintervall for T_0. Hva blir prediksjonsintervallet når dataene er som gitt over og $z_0 = 3.0$?



$P(Y > y_\alpha) = \alpha$

α	y_α
0.975	0.025
0.950	0.051
0.900	0.106
0.100	2.59
0.050	3.49
0.025	4.46

Figur 1: Sannsynlighetstettheten for Y og noen kvantiler i denne fordelinga.

Fasit

1. **a**) 0.035, 0.536 **b**) $\mathrm{E}[\widehat{P}] = p$, $\mathrm{Var}[\widehat{P}] = \frac{1}{4}(\frac{1}{n_1}+\frac{1}{n_2})p(1-p)$, $\mathrm{E}[P^*] = p$, $\mathrm{Var}[P^*] = \frac{1}{n_1+n_2}p(1-p)$ **c**) $[633, 704]$

2. **a**) 0.368,0.050,0.306 **d**) $\mathrm{E}[\widehat{\theta}] = \theta$, $\mathrm{Var}[\widehat{\theta}] = \theta^2/n$ **e**) $[1\,363\,016, 4\,856\,037]$ **f**) $[198.97, 1\,645.92]$