Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

TMA4245 Statistikk
Vår 2015

Øving nummer 10, blokk II

Oppgave 1

Når en planlegger en spørreundersøkelse har en av og til et "delikat" spørsmål som en ønsker svar på. La oss anta at vi ønsker å undersøke hvor stor andel av befolkningen som har gjort W i 1994. Dersom vi spør direkte om dette i spørreundersøkelsen, kan vi ikke forvente at alle vil svare korrekt. Et alternativ er å stille spørsmålet på en slik måte at svarpersonens sanne holdning til spørsmålet ikke er identifiserbar:

> Spørsmålet er: Har du gjort W i 1994?
> Før personen skal svare på spørsmålet, skal personen kaste en terning som ingen andre skal se. Dersom det blir 5 eller 6, skal personen svare galt på spørsmålet. Dersom det blir 1, 2, 3 eller 4, skal personen svare korrekt på spørsmålet. Dvs: Dersom en person oppnår 5 på terningen og har gjort W i 1994, skal han svare Nei.

La q være sannsynligheten for at en tilfeldig valgt person har gjort W i 1994. Vi definerer følgende hendelser,

> J = "En tilfeldig valgt person svarer ja på spørsmålet"
> K = "En tilfeldig valgt person svarer korrekt på spørsmålet"
> W = "En tilfeldig valgt person har gjort W i 1994"

La videre K^C være komplementærhendelsen til hendelsen K, og tilsvarende for J og W. Vi antar at K og W er uavhengige hendelser.

a) Finn sannsynlighetene, $P(K)$, $P(J)$ og $P(W|J)$.

Det er planlagt å dele ut spørreskjemaet til n tilfeldig valgte personer som alle vil svare. La X være antallet av de n personene som svarer ja på spørsmålet og la $p = P(J)$.

b) Hva er verdiområdet til p? (Begrunn svaret). Hva er fordelingen til X? (Begrunn svaret).

c) Finn sannsynlighetsmaksimeringsestimatoren (SME) $\hat{p}$ for p. Bruk dette resultatet til å finne sannsynlighetsmaksimeringsestimatoren (SME) $\hat{q}$ for q.

Oppgave 2

Miljøkonsulenten i en kommune ønsker å undersøke den ukjente pH-verdien i et vann. Betegn den sanne pH-verdien for μ. Konsulenten har tilgjengelig to målemetoder. Metode I er rask,

men måleresultatene er beheftet med betydelig måleusikkerhet. Metode II er mye mer tidkrevende, men gir mer nøyaktige målinger. Begge målemetodene er velbrukte og variansen i målingene er derfor kjent. Miljøkonsulenten velger å gjøre en observasjon med hver metode. La X betegne observasjonen ved bruk av metode I og Y observasjonen ved metode II. Vi antar at X og Y uavhengige og normalfordelt med

$$\mathrm{E}(X) = \mu \ \ , \ \ \mathrm{Var}(X) = \sigma_0^2 \ \ , \ \ \mathrm{E}(Y) = \mu \ \ , \ \ \mathrm{Var}(Y) = \tau_0^2$$

der σ_0^2 og τ_0^2 er kjente størrelser.

Det oppgis at en forventningsrett estimator (som forøvrig også er sannsynlighetsmaksimeringsestimator) for μ i denne situasjonen er

$$\widehat{\mu} = \frac{\tau_0^2 X + \sigma_0^2 Y}{\tau_0^2 + \sigma_0^2}.$$

Ta utgangspunkt i estimatoren $\widehat{\mu}$ og utled et $(1-\alpha)100\%$ konfidensintervall for μ.

Oppgave 3

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

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

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

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 4

I situasjoner der det er uklart hvem som er den biologiske faren til et barn kan farskapet avklares ved å sammenligne DNA-prøver fra barnet med mulige fedre. For en mulig far gjøres dette ved å sammenligne n ulike deler av DNA-strukturen til mannen med de samme n deler av DNA-strukturen hos barnet. De n undersøkte delene av DNA-strukturen antas uavhengige.

Hos et barn og en tilfeldig valgt mann (som ikke er biologisk far) er det for hver enkel del av DNA-strukturen som undersøkes en sannsynlighet $p = 0.15$ for at delen er sammenfallende hos barnet og mannen. Anta videre at en biologisk far alltid har alle de undersøkte delene av DNA-strukturen sammenfallende med barnets (dvs. vi ser bort fra mutasjoner o.l.), slik at hver undersøkte del av DNA-strukturen hos biologisk far og barn er sammenfallende med sannsynlighet $p = 1$.

La X være antall sammenfallende deler i DNA-strukturen hos et barn og en tilfeldig valgt mann (som ikke er biologisk far).

a) Begrunn at X er binomisk fordelt med parametre n og $p = 0.15$.

Dersom $n = 5$, beregn sannsynlighetene $P(X = 2)$, $P(X \geq 2)$ og $P(X = 2|X \geq 2)$.

I en farsskapssak blir en mann erklært å være biologisk far dersom alle undersøkte deler av DNA-strukturen er sammenfallende hos mannen og barnet. Dette kan vi se på som en

hypotesetest der vi tester

$$H_0: \ \ p = 0.15 \text{ (ikke far)} \qquad \text{mot} \qquad H_1: \ \ p = 1.0 \text{ (far)}$$

der H_0 forkastes (dvs. mannen erklæres som far til barnet) dersom $X = n$.

b) For $n = 5$, finn sannsynligheten for å begå type 1 feil i testen over.

For $n = 5$, finn sannsynligheten for å begå type 2 feil i testen over.

Hvor mange ulike deler, n, av DNA-strukturen må man minst sammenligne dersom man ønsker at sannsynligheten for feilaktig å erklære en mann som far skal være mindre enn 0.000001?

Fasit

1. **a**) 2/3, (1+q)/3, 2q/(1+q)

3. **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**) $[633, 704]$

4. **a**) $0.138, 0.165, 0.836$ **b**) $0.000076, 0, n = 8$