TMA4245 Statistikk
Vår 2016

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Anbefalte oppgaver 8, blokk II

Denne anbefalte øvingen er basert på stoffet som gjennomgås i den åttende uka med forelesninger. Det vil si kap. 8 og første del av kap. 9 i læreboka (Walpole og andre). Oppgavene handler om utvalgsfordelinger, sentralmål og sentralgrenseteoremet, samt punktestimering.

Oppgave 1

I denne oppgaven skal vi simulere utfall fra en kjent sannsynlighetsfordeling. Disse utfallene skal vi bruke til å undersøke effekten av sentralgrenseteoremet på fordelingen av gjennomsnittsverdien til utfallene.

I fila SGT.m som er tilgjengelig fra hjemmesiden finner du kode for å utføre lignende oppgaver som du skal gjøre her.

a) Simuler 1000 datasett i MATLAB. Hvert datasett skal bestå av 100 utfall fra en normalfordeling med forventningsverdi 5 og standardavvik 2.

MATLAB tips: Bruk funksjonen `normrnd`.

b) Regn ut gjennomsnittsverdien av alle de 1000 datasettene. Lag et histogram basert på gjennomsnittsverdiene du har regnet ut. Minner formen på histogrammet om formen til en normalfordeling? Var dette forventet? Forklar.

MATLAB tips: Bruk funksjonene `hist` og `normplot`.

c) Gjør det samme som i **a**), men nå skal utfallene komme fra en binomisk fordeling, Bin(N,p), $N = 5$,$p = 0.2$. Prøv deg frem med $n = 2, 5, 10, 20, 30, 50, 100$ utfall for hvert datasett.

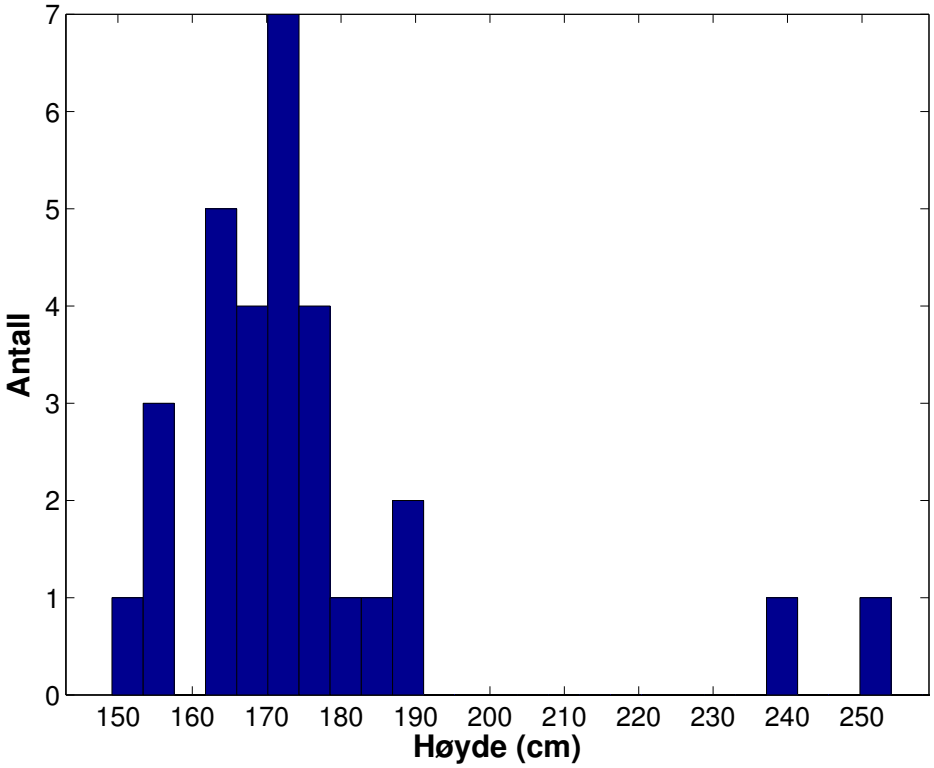
MATLAB tips: Bruk funksjonen `binornd`.

d) Hvilke av simuleringene gir et histogram som ligner en normalfordeling? Bruk sentralgrenseteoremet til å forklare resultatet du får.

MATLAB tips: Bruk funksjonene `hist` og `normplot`.

Oppgave 2

Medianen til et datasett, $\tilde{X}$, er den midterste verdien. Hvis vi har stokastiske (tilfeldige) variabler $X_1, X_2, ..., X_n$ og ordner dem etter størrelse slik at $X_{(1)} < X_{(2)} < ... < X_{(n)}$, så er



Figur 1: Høydene til 30 rekrutter, kanskje fra 1814.

medianen definert som

$$\tilde{X} = \begin{cases} X_{(\frac{n+1}{2})} & \text{hvis } n \text{ er et oddetall,} \\ \frac{1}{2}\left(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}\right) & \text{hvis } n \text{ er et partall.} \end{cases}$$

Når de stokastiske variablene våre er uavhengige og normalfordelte med forventningsverdi μ og varians σ^2, altså $X_i \sim N(\mu, \sigma^2)$, og vi har at antallet variabler, n, er stort, kan vi anta at variansen til medianen er

$$\mathrm{Var}(\tilde{X}) = \frac{1}{4n\big(f(\mu)\big)^2},$$

der $f(x)$ er sannsynlighetstettheten til normalfordelingen.

a) For dette tilfellet, vis at

$$\mathrm{Var}(\tilde{X}) = \frac{\pi}{2}\mathrm{Var}(\bar{X}),$$

der gjennomsnittet $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$\tilde{X}$ er en forventningsrett estimator for forventningsverdien μ. Hvorfor foretrekker vi vanligvis $\bar{X}$ framfor $\tilde{X}$ som estimator for μ?

Statistisk sentralbyrå har data for høydene til mannlige norske rekrutter til hæren hvert år tilbake til 1878. I denne oppgaven kan du anta at du vet med sikkerhet at høydene til rekruttene i et hvilket som helst år er normalfordelte.

På Terningmoen leir har løytnant Munthe funnet et skjema med høydene på 30 rekrutter som han mener må være fra 1814. Papiret er gulnet og blekket har falmet en del, men løytnanten får en av sine nåværende rekrutter til å skrive dataene inn i et regneark etter beste evne. Figur 1 viser et histogram av disse dataene.

b) For dette datasettet, vil medianen $\tilde{X}$ være større enn, mindre enn eller omtrent like stor som gjennomsnittet $\bar{X}$?

Ville du ha brukt medianen eller gjennomsnittet til å estimere forventningsverdien μ her? Begrunn svaret.

Oppgave 3

Utspørring av deltakere i humor- og realityprogram på TV har den siste tiden blitt svært populært. Spørsmålene som stilles kan vi dele inn i tre typer, og vi definerer følgende disjunkte hendelser:

A_1="det stilles et spørsmål av ikke-sensitiv natur, f.eks. hva heter du?",
A_2="det stilles et spørsmål av delvis sensitiv natur, f.eks. hvor gammel er du?",
A_3="det stilles et spørsmål av sensitiv natur, f.eks. har du vært utro?".

I tillegg definerer vi hendelsen:
L="deltakeren lyver"

Følgende sannsynligheter er oppgitt:
$P(A_1) = 0.1$, $P(A_2) = 0.4$, $P(A_3) = 0.5$, $P(L|A_1) = 0.05$, $P(L|A_2) = 0.2$, $P(L|A_3) = 0.6$.

a) Vis de fire hendelsene i et venndiagram.

Gitt at en deltaker blir spurt et spørsmål av type A_2, hva er sannsynligheten for at deltakeren ikke lyver, $P(L'|A_2)$?

Hva er sannsynligheten for at en tilfeldig valgt deltaker lyver, $P(L)$?

Et av spørsmålene som regnes å være av delvis sensitiv natur er "hvor gammel er du?". En gruppe på n personer ble stilt dette spørsmålet, deretter ble svarene registrert og sammenlignet med informasjon i offentlige registre. La X være en stokastisk variabel som angir antall personer som lyver blant n personer, og la p være sannsynligheten for at en person lyver.

b) Under hvilke antagelser vil X være binomisk fordelt?

Vi antar at $p = 0.2$ og at vi spør $n = 20$ personer. Hva er $P(X = 4)$?

Hva er $P[(X \leq 2) \cup (X > 5)]$?

Vi antar nå at p er ukjent. For å estimere p er det foreslått to estimatorer,

$$\hat{p} = \frac{X}{n} \quad \text{og} \quad p^* = \frac{X}{n-1}.$$

c) Finn forventningsverdi og varians til hver av estimatorene $\hat{p}$ og p^*.

Hvilke to egenskaper kjennetegner en god estimator?

Hvilken av estimatorene $\hat{p}$ og p^* vil du foretrekke? Begrunn svaret.

Oppgave 4

For å vurdere nøyaktigheten av en ny måleprosedyre gjør man n målinger av samme størrelse. La $X_1, X_2, \ldots, X_n$ betegne resultatene av disse målingene, og anta at disse er et tilfeldig utvalg fra en normalfordeling med forventning μ og standardavvik σ. Vi er her interessert i verdien til σ, men vi skal anta at verdien til μ også er ukjent.

La $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Fra pensum er det da kjent at $S^2(n-1)/\sigma^2$ er kji-kvadratfordelt med $n-1$ frihetsgrader.

a) Vis at dersom en stokastisk variabel Y er kji-kvadratfordelt med v frihetsgrader, dvs. har sannsynlighetstetthet

$$f(y) = \frac{1}{2^{\frac{v}{2}}\Gamma(\frac{v}{2})} y^{\frac{v}{2}-1} e^{-\frac{y}{2}},$$

så er

$$E(\sqrt{Y}) = \frac{\sqrt{2}\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})}.$$

b) Bruk resultatet i forrige punkt til å undersøke om

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

er forventningsrett for σ. Foreslå eventuelt en estimator for σ hvor forventningsfeilen er korrigert.

En estimator $\hat{\theta}$ som over- og underestimerer θ med like stor sannsynlighet er såkalt medianrett. For kontinuerlig fordelte medianrette estimatorer er derfor $P(\hat{\theta} \leq \theta) = 1/2$. Foreslå en medianrett estimator for σ.

Fasit

2. **b**) medianen er mindre enn gjennomsnittet

3. **a**) 0.8, 0.385 **b**) 0.218, 0.402