TMA4245 Statistikk
Eksamen 9. desember 2013

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Oppgave 1

I kortspillet Blackjack får man den høyeste gevinsten hvis de to første kortene man får er et ess sammen med enten en tier, en knekt, en dame eller en konge. Dette kalles «å få en blackjack».

I kasinoer der man spiller Blackjack for penger, blander man gjerne 8 kortstokker som man trekker disse kortene fra. (Hensikten med dette er å gjøre det vanskeligere å «telle kort», det vil si å regne seg fram til hvor mange ess og billedkort som er igjen i stokken.) Det er da så mange kort i stokken at vi kan betrakte dette som en situasjon «med tilbakelegging», altså at man trekker ett og ett kort fra en vanlig kortstokk og legger forrige kort tilbake før man trekker et nytt. En vanlig kortstokk har 52 kort, med 4 farger som har hver sin tier, knekt, dame, konge og ess.

a) Hva er sannsynligheten for å få en blackjack?

Hva er sannsynligheten for å få en blackjack hvis det første kortet du har fått er et ess?

La oss si at en spiller som teller kort har funnet ut at sannsynligheten for å få blackjack nå er 0.06, at sannsynligheten for å få et ess som første kort er 0.1 og at sannsynligheten for at det første kortet ditt var et ess hvis du har fått en blackjack er 0.4. Bruk disse tallene til å regne ut hva sannsynligheten nå er for å få en blackjack hvis det første kortet du har fått er et ess.

Lars er på ferie i Las Vegas, og går inn på det største kasinoet han finner. Han setter seg ved Blackjack-bordet og bestemmer seg for å spille fram til han vinner og for å doble innsatsen for hvert spill. Han satser én dollar i første spill, to dollar i andre spill, og så videre fram til han vinner.

Anta (for enkelthets skyld) at han alltid får igjen det dobbelte av det han satset hvis han vinner, at sannsynligheten for at han vinner er 0.3 i hvert spill, og at Lars slutter å spille etter å ha vunnet én gang.

b) La X være antall ganger Lars spiller før han gir seg. Hva er sannsynlighetsfordelingen til X?

Anta i resten av oppgaven at Lars går tom for penger og dermed slutter å spille, hvis han ikke har vunnet etter å ha spilt fem ganger. La W være antall dollar Lars vinner i kasinoet (uavhengig av hvor mye han har satset først). Hva er forventningsverdien til W?

La Y være gevinsten Lars sitter igjen med etter at innsatsen er trukket fra. Hva er forventningsverdien til Y?

Oppgave 2

Agent John Bang går regelmessig til skytetrening. Erfaring sier at hans sannsynlighet for et treff er $p = 0.8$. I en treningssesjon skyter han 20 skudd. Anta at skuddene er uavhengige og at hvert enkeltskudd er enten en treff eller bom.

a) Hva er forventet antall treff?

Hva er sannsynligheten for at han treffer flere enn forventet?

Hva er den betingete sannsynligheten for at han treffer 20 når vi vet at han treffer flere enn forventet?

Sjefen bestemmer at John skal ha en ny pistol. De håper at denne skal gi forbedret treffsannsynlighet. De ønsker å undersøke om dette holder, og John bruker den nye pistolen i en vanlig treningssesjon med 20 skudd.

b) Formuler problemet som en hypotesetest.

Bruk den vanlige normalapproksimasjonen til å gjennomføre hypotesetesten på signifikansnivå $\alpha = 0.1$ når observert antall treff er 18.

c) Forklar hvordan testen kan gjennomføres eksakt ved bruk av binomisk fordeling.

Hva blir P-verdien til den eksakte testen når han treffer på 18 skudd?

Anta signifikansnivå $\alpha = 0.1$ for testen. Regn ut teststyrken for den eksakte testen når sann treffsannsynlighet er $p = 0.9$.

Oppgave 3

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 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

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

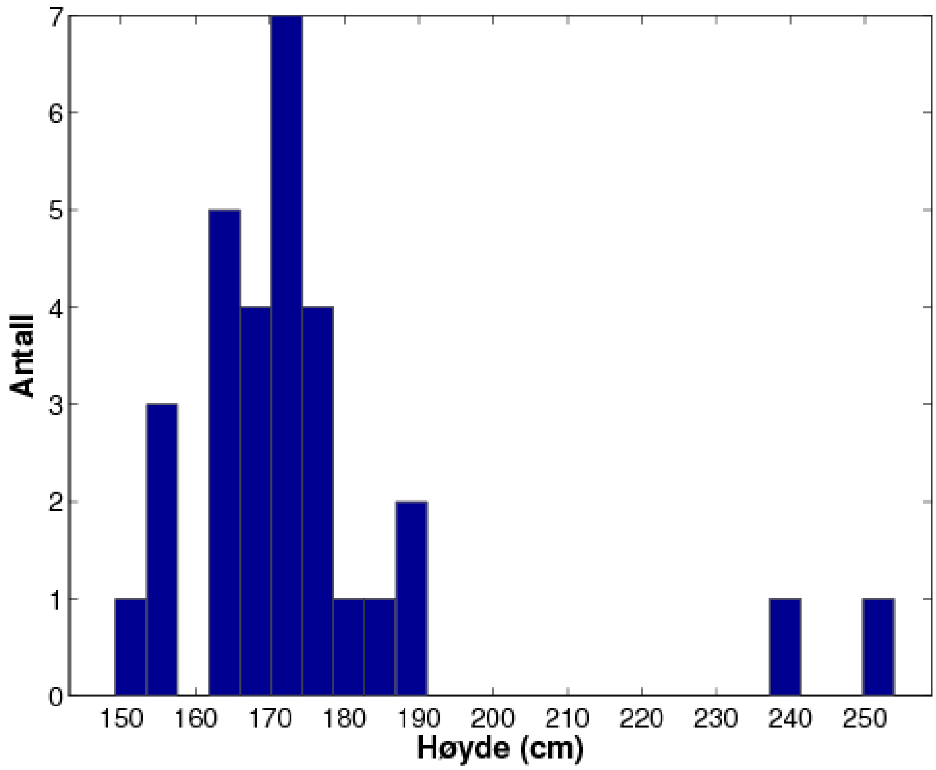
der $f(x)$ er sannsynlighetstettheten til normalfordelingen.

a) For dette tilfellet, vis at

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

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?



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

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 $\bar{X}$?

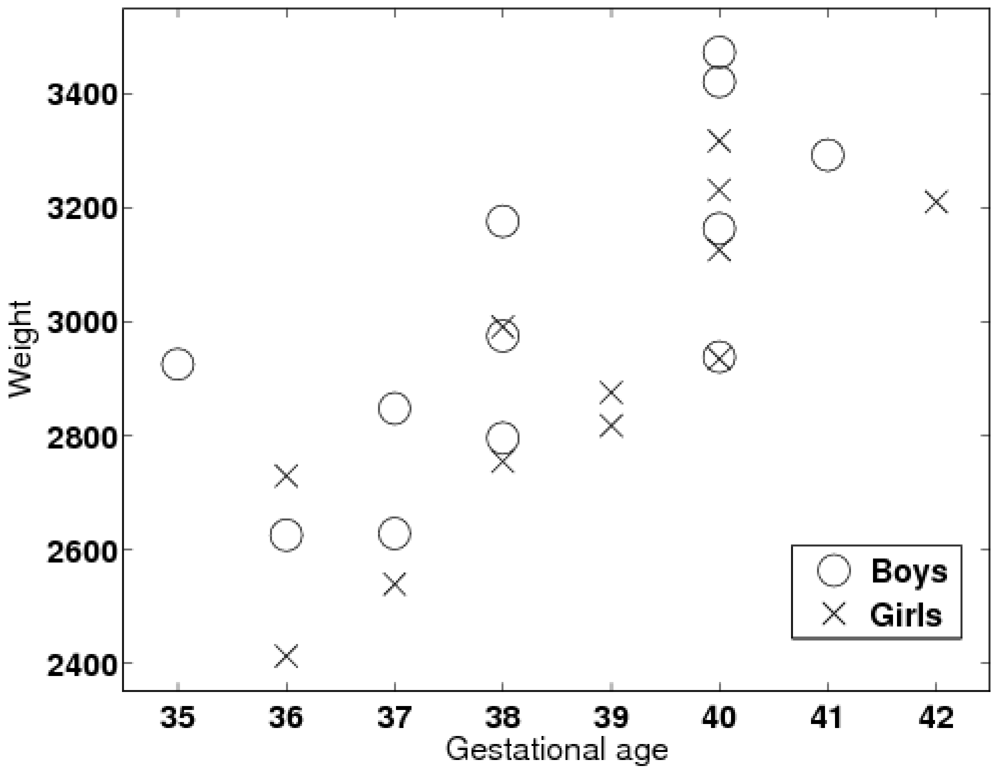
Ville du ha brukt medianen eller gjennomsnittet til å estimere forventningsverdien μ her? Begrunn svaret.

Oppgave 4

I medisin er det nyttig å studere vekten til nyfødte som funksjon av deres terminalder (*gestational age* eller tid siden unnfangelse). Data er her terminalder x_i (uker) og vekt y_i (gram) for $i = 1, \ldots, n$ babyer, og $n = 24$.

For dette datasettet har vi $\sum_{i=1}^n x_i y_i = 2\ 752\ 667$, $\sum_{i=1}^n x_i^2 = 35\ 727$, $\sum_{i=1}^n x_i = 925$ og $\sum_{i=1}^n y_i = 71\ 194$.

Anta en lineær regresjonsmodell: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$, der $\epsilon_1, \ldots, \epsilon_n$ antas



Figur 2: Kryssplott av terminuke og fødselsvekt for 12 gutter og 12 jenter.

uavhengige og normalfordelte med forventning 0 og varians σ^2.

a) Bruk oppsummeringen av tallmateriale gitt over til å regne ut estimatene for skjæringspunkt og stigningstallet for regresjonsmodellen: $\hat{\beta}_0$ og $\hat{\beta}_1$.

Vi regner ut et estimat for σ^2 ved $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 194^2$.

Regn ut et 95 prosent konfidensintervall for stigningstallet.

b) Bruk data til å finne et 90 prosent prediksjonsintervall for vekten til en nyfødt i terminuke 40.

Hvor bredt er 90 prosent prediksjonsintervallet for terminuke 42 sammenlignet med det vi fant for uke 40?

Figur 2 viser et kryssplott av terminuke og vekt. I dette plottet er data delt inn i to grupper: gutter og jenter. Det er $n_b = 12$ gutter (nummerert 1 til n_b) og 12 jenter (nummerert $i = n_b + 1$ til n).

Vi foreslår følgende modell for data

$$Y_i = \beta_b + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n_b.$$

$$Y_i = \beta_g + \beta_1 x_i + \epsilon_i, \quad i = n_b + 1, \ldots, n.$$

der vi fortsatt antar at $\epsilon_1, \ldots, \epsilon_n$ er uavhengige og normalfordelte med forventning 0 og varians σ^2.

c) Bruk plottet til å forklare hvorfor denne modellen kan være hensiktsmessig. Forklar videre hvilke elementer av modellen som kan være uønsket.

Regn ut minste kvadratsums estimater (eller maximum likelihood estimater) for parametrene i modellen, her benevnt ved $\hat{\beta}_b$, $\hat{\beta}_g$ og $\hat{\beta}_1$. I tillegg til summene gitt tidligere i oppgaven har vi at $\sum_{i=1}^{n_b} y_i = 36\ 258$, $\sum_{i=1}^{n_b} x_i = 460$, $\sum_{i=n_b+1}^{n} y_i = 34\ 936$ og $\sum_{i=n_b+1}^{n} x_i = 465$.

Fasit

1. **a**) 0.04734, 0.3077, 0.24 **b**) 6.567, -4.4

2. **a**) 16, 0.41, 0.028 **b**) ikke forkast H_0 **c**) 0.21, 0.39

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

4. **a**) -1465, 115, [69,161] **b**) [2789,3481], 690, 732 **c**) -1587, -1747, 120.22