LØSNING, EKSAMEN I STATISTIKK, TMA4240, DESEMBER 2006

OPPGAVE 1

Anta at sann porøsitet er r. Måling med utstyret gir da $X \sim n(x; r, 0{,}03)$.

a)

$$P(X > r) = P\left(\frac{X-r}{0{,}03} > 0\right) = P(Z > 0) = 0{,}5.$$
$$P(X - r > 0{,}05) = P\left(\frac{X-r}{0{,}03} > \frac{0{,}05}{0{,}03}\right) = P(Z > 1{,}67) = 0{,}0475.$$

For de som tolket feilen som absoluttverdi (som også er rimelig)

$$P(|X - r| > 0{,}05) = P\left(\frac{|X-r|}{0{,}03} > \frac{0{,}05}{0{,}03}\right) = P(|Z| > 1{,}67) = 0{,}0950.$$

Anta at X_1 og X_2 er uavhengig identisk fordelt (uif). Da er $\bar{X} = \frac{1}{2}(X_1+X_2) \sim n(\bar{x}; r, 0{,}03/\sqrt{2})$.

$$P(\bar{X} - r > 0{,}05) = P\left(\frac{\bar{X}-r}{0{,}03/\sqrt{2}} > \frac{0{,}05}{0{,}03/\sqrt{2}}\right) = P(Z > 2{,}357) = 0{,}0092.$$

For de som tolket avviket som absoluttverdi (som også er rimelig)

$$P(|\bar{X} - r| > 0{,}05) = P\left(\frac{|\bar{X}-r|}{0{,}03/\sqrt{2}} > \frac{0{,}05}{0{,}03/\sqrt{2}}\right) = P(|Z| > 2{,}357) = 0{,}0184.$$

b) $X_1, X_2, \ldots, X_5$ er uif $n(x; 0{,}15, \sigma)$.

Den beste (mest effektive) forventningsrette estimatoren er $\widehat{\sigma}^2 = \frac{1}{5}\sum_{i=1}^{5}(X_i - 0{,}15)^2$. Dette er sannsynlighetsmaksimeringsestimatoren. (NB: $\mu = 0{,}15$ er kjent!)

Vi har at

$$\frac{n\widehat{\sigma}^2}{\sigma^2} = \frac{5\widehat{\sigma}^2}{\sigma^2} = \sum_{i=1}^{5}\left(\frac{X_i - 0{,}15}{\sigma}\right)^2 = \sum_{i=1}^{5} Z_i^2 = V_5 \sim \chi_5^2.$$

Uttrykket er kji-kvadratfordelt ("chi-square") med 5 frihetsgrader.

$$E\left(\frac{5\widehat{\sigma}^2}{\sigma^2}\right) = E(V_5) = 5 \quad \text{dvs} \quad \frac{5}{\sigma^2}E(\widehat{\sigma}^2) = 5, \quad \text{og dermed} \quad E(\widehat{\sigma}^2) = \sigma^2.$$
$$\text{Var}\left(\frac{5\widehat{\sigma}^2}{\sigma^2}\right) = \text{Var}(V_5) = 2 \cdot 5 = 10 \quad \text{dvs} \quad \frac{25}{\sigma^4}\text{Var}(\widehat{\sigma}^2) = 10, \quad \text{og dermed} \quad \text{Var}(\widehat{\sigma}^2) = 0{,}4\sigma^4.$$

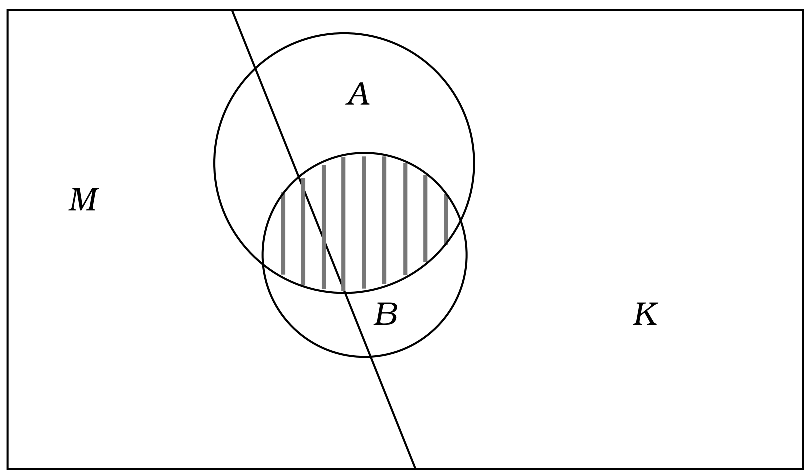
95%-konfidensintervall for σ^2: $P(\chi^2_{5,0,975} < V_5 < \chi^2_{5,0,025}) = 0{,}95$.

$$P\left(\chi^2_{5,0,975} < \frac{5\widehat{\sigma}^2}{\sigma^2} < \chi^2_{5,0,025}\right) = 0{,}95,$$
$$P\left(\frac{5\widehat{\sigma}^2}{\chi^2_{5,0,275}} < \sigma^2 < \frac{5\widehat{\sigma}^2}{\chi^2_{5,0,975}}\right) = 0{,}95.$$

Med $\widehat{\sigma} = 0{,}0177$, $\chi^2_{5,0,025} = 12{,}833$ og $\chi^2_{5,0,975} = 0{,}831$ blir konfidensintervallet for σ^2 $[0{,}000122, 0{,}00188]$. Det tilsvarer konfidensintervall for σ: $[0{,}0110, 0{,}0434]$.

OPPGAVE 2

a)



Figur 1: Venn-diagram for hendelsene M, K, A og B. Merk at M og K deler utfallsrommet i to disjunkte deler, mens A og B delvis overlapper M, K og hverandre. $A \cap B$ er vertikalt skravert.

Vi benytter at vi summerer sannsynlighet for disjunkte hendelser, og at snittene kan skrives ut med betinget sannsynlighet.

$P(A \cap K) = P(A|K)P(K) = 0{,}440 \cdot \frac{418}{632} = 0{,}2910$,

30% av responentene var kvinner som hadde utøvd psykologisk aggresjon mot partneren.

$P(A) = P(A \cap M) + P(A \cap K) = P(A|M)P(M) + P(A|K)P(K) = 0{,}272 \cdot \frac{214}{632} + 0{,}440 \cdot \frac{418}{632} = 0{,}0921 + 0{,}2910 = 0{,}3831$,

38% av respondentene hadde utøvd psykologisk aggresjon mot partneren, dvs når vi også inkluderer mennene.

b) For $P(K|A)$ bruker vi egentlig Bayes' formel, men beregningen over gir oss tallene vi trenger direkte.

$$P(K|A) = \frac{P(K \cap A)}{P(A)} \left(= \frac{P(A|K)P(K)}{P(A|M)P(M) + P(A|K)P(K)} \right) = \frac{0{,}2910}{0{,}3831} = 0{,}7596.$$

Sannsynligheten for at skjemaet er fylt ut av en kvinne er hele 76%. Dette skyldes dels at flere kvinner enn menn svarte på undersøkelsen, dels at større andel kvinner enn menn rapporterte A.

For $P(A|B)$ bruker vi igjen Bayes' formel, og bruker at $B = (A \cap B) \cup (A^* \cap B)$.

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B \cap K) + P(A \cap B \cap M)}{P(B|K)P(K) + P(B|M)P(M)} \\ &= \frac{P(A \cap B|K)P(K) + P(A \cap B|M)P(M)}{[P(A \cap B|K) + P(A^* \cap B|K)]P(K) + [P(A \cap B|M) + P(A^* \cap B|M)]P(M)} \\ &= \frac{0{,}292 \cdot 0{,}6614 + 0{,}216 \cdot 0{,}3386}{[0{,}292 + 0{,}148] \cdot 0{,}6614 + [0{,}216 + 0{,}056] \cdot 0{,}3386} = 0{,}6950. \end{aligned}$$

Eventuelt kan en selvfølgelig regne ut $P(B|K) = P(A \cap B|K) + P(A^* \cap B|K)$ osv. først, eller til og med regne ut antall personer i hver kategori og løse oppgaven derfra.

70% av respondentene som har opplevd psykologisk aggresjon fra partneren, har selv også utøvd psykologisk aggresjon, de resterende 30% er kanskje mer "uskyldige" offer?

c) Hendelsen X er binomisk fordelt hvis:

- hvert skjema kan karakteriseres som suksess (S) eller fiasko (S^*) - ingen tvilstilfeller,
- det er et fiksert antall $n = 418$ skjema i undersøkelsen, hvor suksessene telles i X,
- sannsynligheten for S, $P(S|K)$ er den samme for hvert skjema fylt ut av en kvinne; vi kan altså ikke på forhånd dele bunken i skjema hvor vi tror det er mange S og skjema med få S,
- besvarelsene er uavhengige; respondentene har ikke påvirket hverandre.

Dette er rimelige antakelser, så vi bruker binomisk fordeling for X (og Y).

Hvis andelen x/n fra undersøkelsen gjelder hele populasjonen (av kvinnelige studenter i heterogene parforhold), er $p_X = x/n = 38/418$. Dine fem venninner antas trukket tilfeldig med binomisk fordeling som over, med $p = P(S|K) = 38/418 = 0{,}0909$. (Altså, ca. 1 av 10 har opplevd seksuell aggresjon fra partner.) $n = 5$ gir punktsannsynligheten

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{5}{0} p^0 (1-p)^{5-0} = 1 - (1 - 0{,}0909)^5 = 0{,}3790 = 38\%.$$

d) Vi kan tilnærme binomisk til normal hvis $np_X > 5$ og $n(1 - p_X) > 5$ og tilsvarende $mp_Y > 5$ og $m(1 - p_Y) > 5$. Den minste av disse ser ut til å være mp_Y, men $m\widehat{p}_Y = 10$ er stor nok.

$\text{Var}(\widehat{p}_X) = \text{Var}(X/n) = \text{Var}(X)/n^2 = np_X(1-p_X)/n^2 = p_X(1-p_X)/n.$

$\text{Var}(\widehat{p}_Y) = \text{Var}(Y/m) = \text{Var}(Y)/m^2 = mp_Y(1-p_Y)/m^2 = p_Y(1-p_Y)/m.$

Eventuelt kan en legge til at vi har et gjennomsnitt av indikatorvariabler, og at sentralgrenseteoremet gir at $\frac{\widehat{p}_X - p_X}{\sqrt{p_X(1-p_X)/n}} \sim N(0,1)$ når $n \to \infty$ og tilsvarende for p_Y og m.

En rimelig estimator for andelen S hvis H_0 er sann er $\widehat{p} = \frac{X+Y}{n+m} = \frac{38+10}{632} = 0{,}0759$. Dermed setter vi inn $p = p_X = p_Y \approx \widehat{p} = 0{,}0759$ i variansuttrykkene.

Med $H_0 : p_X = p_Y$ og $H_1 : p_X \neq p_Y$ har vi en tosidig hypotesetest. Testobservatoren er $\widehat{p}_X - \widehat{p}_Y$, som under H_0 har forventningsverdi $p_X - p_Y = 0$ og varians $\sigma^2 = p_X(1-p_X)/n + p_Y(1-p_Y)/m = p(1-p)(1/n+1/m) \approx 0{,}0223^2$. Vi aksepterer H_0 hvis observatoren havner mellom $-z_{\alpha/2}\sigma$ og $-z_{\alpha/2}\sigma$. Med $\alpha = 0{,}05$ blir akseptområdet $|\widehat{p}_X - \widehat{p}_Y| < 1{,}645 \cdot 0{,}0223 = 0{,}0366$.

Vi har $\widehat{p}_X - \widehat{p}_Y = 38/418 - 10/214 = 0{,}0442$, som er utenfor akseptområdet. Dermed forkastes H_0 på signifikansnivå 0.05, og vi er enige med rapporten. (Det er grunn til å tro at flere kvinner opplever seksuell aggresjon fra sin mannlige partner enn motsatt.)

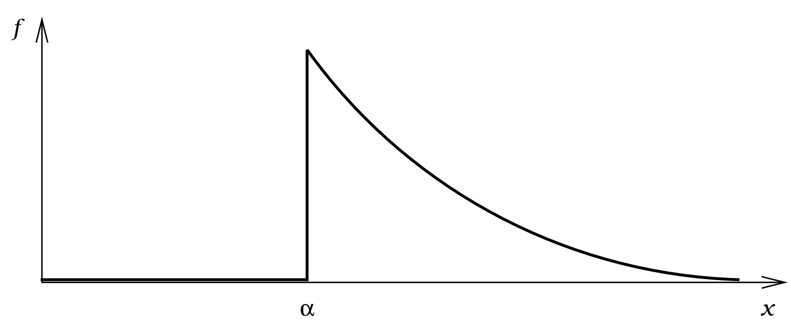
Vi har altså fått $z_{\text{obs}} = \frac{\widehat{p}_X - \widehat{p}_Y}{\sigma} = \frac{0{,}0442}{0{,}0223} = 1{,}9841$. Arealet over denne verdien i standard normalfordelingen er halvparten av p-verdien, siden vi har en tosidig test.

P-verdi $= 2P(Z > 1{,}9841) = 2P(Z < -1{,}9841) \approx 2 \cdot \frac{0{,}0239+0{,}0233}{2} = 0{,}0472 = 4{,}72\%$.

Bruk av $\sigma \approx \sqrt{\widehat{p}_X(1-\widehat{p}_X)/n + \widehat{p}_Y(1-\widehat{p}_Y)/m}$ gir omtrent samme resultat. Merk at dette er en dårligere approksimasjon av σ under H_0.

OPPGAVE 3

a)



Figur 2: Skisse av sannsynlighetstettheten $f(x;\alpha,\beta)$.

Sannsynlighetsmaksimering for β følger normal prosedyre, men en bør huske at $x_i \geq \alpha$ for alle x_i.

$$L(\alpha, \beta; x_1, \ldots, x_n) = \begin{cases} 0, & \text{hvis minst én } x_i < \alpha, \\ \prod_{i=1}^n \frac{1}{\beta} e^{-(x_i-\alpha)/\beta}, & \text{ellers.} \end{cases}$$

$$\Lambda(\alpha, \beta; x_1, \ldots, x_n) = -n \ln(\beta) - \sum_{i=1}^n \frac{x_i - \alpha}{\beta}, \quad \text{når alle } x_i \geq \alpha.$$

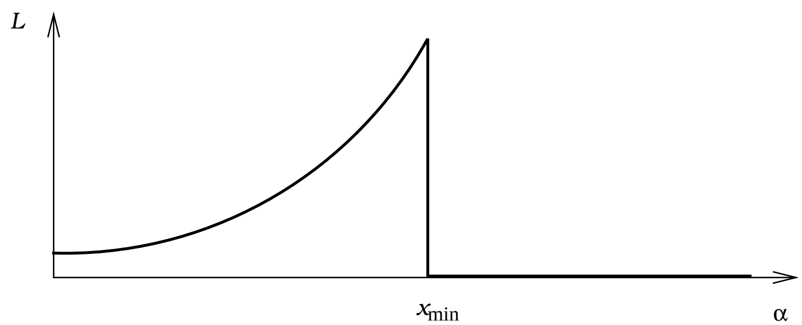
$$\frac{\partial \Lambda}{\partial \beta} = -\frac{n}{\beta} + \sum_{i=1}^n \frac{x_i - \alpha}{\beta^2}.$$

$\frac{\partial \Lambda}{\partial \beta} = 0$ for $\beta = \widehat{\beta}$ og stokastiske variabler for målingene x_i gir

$$\widehat{\beta} = \frac{1}{n} \sum_{i=1}^n (X_i - \alpha) = \bar{X} - \alpha,$$

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

b)



Figur 3: Skisse av likelihoodfunksjonen som funksjon av α. Vi vet at $\alpha > 0$ fra oppgaven, funksjonen vokser eksponensielt i starten, og at den blir null hvis α er større enn grensen definert av minste X-verdi. Poenget er at funksjonen ikke er deriverbar mhp α i toppunktet, så vi kan ikke derivere og sette lik null som før. Oppgaven krever ikke at denne figuren er med.

Vi skal altså finne en α som maksimerer

$$L(\alpha, \beta; x_1, \ldots, x_n) = \begin{cases} 0, & \text{hvis minst én } x_i < \alpha, \\ \left(\frac{1}{\beta}\right)^n \exp\left[-\frac{\sum_{i=1}^n x_i - n\alpha}{\beta}\right], & \text{ellers.} \end{cases}$$

Det er klart at vi må velge $\alpha \leq \min(x_i)$, ellers er $L = 0$ fra den første linja i den delte forskriften. Den andre linja blir størst hvis α er så stor som mulig. Derfor maksimeres L av $\widehat{\alpha} = \min(X_1, \ldots, X_n)$.

Sannsynlighetsmaksimeringsestimatoren for β har vi fra forrige oppgave. Siden α er ukjent må vi prøve $\widehat{\beta} = \bar{X} - \widehat{\alpha}$.

For å finne forventningsverdien til $\widehat{\alpha}$, brukes transformasjonen $Z_i = X_i - \alpha$. Dette skyver fordelingen inntil andreaksen, slik at vi får en standard eksponensialfordeling med parameter β. Hintet sier at $\min(Z_1, \ldots, Z_n)$ har forventningsverdi β/n.

Den minste X-verdien er α større enn den minste Z-verdien, siden $X_i = Z_i + \alpha$ også for den i-en som gir minst X_i. Dermed er $E(\widehat{\alpha}) = \alpha + \beta/n$, og ikke forventningsrett.

Forventningsverdien til X er $\alpha + \beta$, det er jo bare en vanlig eksponensialfordeling med forventning β som er flyttet α til høyre på x-aksen. Forventningsverdien til $\bar{X}$ er $E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = E(X_i) = \alpha + \beta$. Hvis vi bytter ut α med $\widehat{\alpha}$ i uttrykket for estimatoren, blir forventningsverdien

$$E(\widehat{\beta}) = E(\bar{X}) - E(\widehat{\alpha}) = \alpha + \beta - E(\widehat{\alpha}) = \beta(1 + \frac{1}{n}) = \frac{n+1}{n}\beta.$$

Denne er heller ikke forventningsrett.

Her kan vi fikse estimatoren for β ved å gange med $n/(n+1)$, og deretter justere estimatoren for α. Vi vil demonstrere en annen framgangsmåte (som gir samme resultat). Trekk fra $\widehat{\beta}/n$ i estimatoren for α. Da blir $\widehat{\alpha}$ forventningsrett hvis $\widehat{\beta}$ er forventningsrett. Til sammen har vi da to likninger som gjelder for forventningsrette estimatorer:

$$\begin{aligned}
\widehat{\alpha} &= \min(X_1, \ldots, X_n) - \frac{\widehat{\beta}}{n}, \\
\widehat{\beta} &= \bar{X} - \widehat{\alpha}.
\end{aligned}$$

Dette likningssystemet kan løses, og vi ender opp med

$$\begin{aligned}
\widehat{\alpha} &= \frac{n}{n-1}\min(X_1, \ldots, X_n) - \frac{1}{n-1}\bar{X}, \\
\widehat{\beta} &= \frac{n}{n-1}(\bar{X} - \min(X_1, \ldots, X_n)) = \frac{1}{n-1}\sum_{i=1}^{n} X_i - \frac{n}{n-1}\min(X_1, \ldots, X_n).
\end{aligned}$$

Til slutt bør vi forsikre oss om at metoden over faktisk gir forventningsrette estimatorer:

$$\begin{aligned}
E(\widehat{\alpha}) &= \frac{n}{n-1}E(\min(X_1, \ldots, X_n)) - \frac{1}{n-1}E(\bar{X}) \\
&= \frac{n}{n-1}(\alpha + \frac{\beta}{n}) - \frac{1}{n-1}(\alpha + \beta) = \alpha, \\
E(\widehat{\beta}) &= \frac{1}{n-1}\sum_{i=1}^{n} E(X_i) - \frac{n}{n-1}E(\min(X_1, \ldots, X_n)) \\
&= \frac{1}{n-1}n(\alpha + \beta) - \frac{n}{n-1}(\alpha + \frac{\beta}{n}) = \beta.
\end{aligned}$$