Tentamentsskrivning i **Matematisk Statistik TMA321**

Tid: den 1 juni, 2020 fm
Hjälpmedel: Alla

---

Tentamen består av 8 frågor om sammanlagt 50 poäng. Preliminära betygsgränser är satta till:
betyg "3": 20 till 29 poäng
betyg "4": 30 till 39 poäng
betyg "5": 40 eller fler poäng.

---

OBS! Alla lösningar skall vara väl redovisade, motiverade och fullständiga. Talen är ej ordnade efter svårighetsgrad.

Kom också ihåg att även om alla hjälpmedel är tillåtna så är det inte samma sak som att enbart hänvisa till dessa hjälpmedel. Vad innebär då detta?

(i) Summering av siffror eller annan enklare räkning behöver ej redovisas.

(ii) Nyckelsteg i t.ex. en summation, eller lösning av integral *skall* redovisas om inget annat explicit anges. **Ex:** Man behöver inte *bevisa* formeln för partiell integration men man kan inte bara skriva "partiell integration" och sedan svaret. Genomför istället den partiella integrationen.

1. Låt $X \sim \mathrm{N}(0,1)$, $Y \sim \mathrm{N}(0,3)$ vara oberoende normalfördelade slumpvariabler.

   (a) Beräkna $\mathbb{P}(X \geq Y + 1)$. (3p)

   (b) Låt $U \sim \mathrm{U}[0,1]$ vara oberoende av $X, Y$, och låt $Z = UX + (1-U)Y$. Beräkna $\mathrm{Var}(Z)$. (3p)

2. Alexandra har en sekvens av oberoende likafördelade slumpvariabler (i.i.d.) $X_1, X_2, \ldots X_n$ med täthetsfunktion

$$f(x) = \frac{1+\alpha}{(1+x)^{2+\alpha}} \text{ för } 0 < x < \infty,$$

   där $\alpha > -1$ är en parameter vars värde är okänt.

   (a) Alexandras chef Magnus sa att $\alpha = 1$, men Alexandra är skeptisk. Hjälp Alexandra att designa ett lämpligt likelihood-ratio test för att se ifall Magnus hade fått det hela om bakfoten. (3p)

   (b) När testet ovan var genomfört så kom Alexandras medarbetare José på att han hade hört att $\alpha$ minsann var två ($\alpha$=2). Hjälp Alexandra och José att göra ett likelihood ratio test för att testa Magnus hypotes mot Josés alternativa hypotes. Bestäm även formen för förkastingsregionen (RR) för enklast möjliga härledda teststatistika. Är ditt svar rimligt? (3p)

3. Låt $X, Y$ vara två slumpvariabler med följande gemensamma sannolikhetstäthetsfunktion

$$f(x,y) = Cx^3 y e^{-\lambda(x^2+y^2)} \quad 0 \leq x, y < \infty.$$

   (a) Hitta $C$ så att detta verkligen blir en gemensam täthetsfunktion. (3p)

   (b) Är $(X, Y)$ beroende eller oberoende? Bevisa ditt påstående. (3p)

4. En transportfirma transporterar paket från Fjollträsk till Mars. Firman väger de första 9 paketen de transporterar. Resultatet blev som följer:

| paket: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| vikt i kilo: | 16.5 | 6.7 | 21 | 22.3 | 10.8 | 29.5 | 29.5 | 19.7 | 22.6 |

   Transportfirman antar att vikterna är normalfördelade med parametrar $\mu, \sigma^2$ vars värden är okända. De antar också att paketens vikter är oberoende av varandra.

   (a) Skatta $\mu, \sigma^2$. (1p)

   (b) Hitta konfidensintervall för $\mu$ och $\sigma^2$, båda med konfidensgrad 95%. (3p)

   (c) Transportfirman tar emot $M$ paket innan de stuvas in på ett rymdskepp vars maxlast är 2 ton. Bestäm största möjliga värde på $M$ (uttryckt i parametrarna $\mu$ och $\sigma^2$) så att sannolikheten att de $M$ paketens totala vikt överstiger 2 ton är mindre än 0.01. (3p)

5. Uno har en magisk låda som kan generera slumpvariabler $V_1, V_2, \ldots$ vilka är oberoende och likafördelade (i.i.d.). På Unos låda står det att

$$\mathbb{P}(V_1 = k) = \frac{2k}{N(N+1)}$$

   för $k = 1, \ldots, N$. Tyvärr så har Unos bror Sune suddat ut vad värdet på heltalet $N$ är.

   (a) Använd momentmetoden för att hitta en lämplig skattare (dvs MME) åt Uno. (3p)
   **OBS: Eventuella jobbiga uträkningar behöver ej redovisas i denna del. Enbart hänvisning till tabeller eller programvara räcker just här.**

   (b) Uno vill gärna hitta ett numeriskt värde på $N$. Tyvärr så drar Unos låda en hel del el, så han vill inte göra för många simuleringar. De sju stycken han gör ger följande resultat:

| Simulering nr: | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Resultat: | 14 | 16 | 5 | 12 | 18 | 19 | 25 |

   Vilken numerisk skattning får ni på $N$? Reflektera över ert resultat. (3p)

6. Låt $X$ ha täthetsfunktion

$$f(x) = \frac{a}{2(e^a - 1)} e^{a|x|} \text{ för } x \in [-1, 1],$$

där $a > 0$ är en parameter.

(a) Beräkna den momentgenererande funktionen (här betecknad $M_{X,a}(t)$) för $X$. (3p)

(b) Hitta en diskret slumpvariabel (dvs bestäm sannolikhetsfunktionen) $Y$ vars momentgenererande funktion $M_Y(t)$ uppfyller villkoret

$$\lim_{a \to \infty} M_{X,a}(t) = M_Y(t).$$

(3p)

7. Låt $X$ vara likformigt fördelad på mängden $\{1, 2, \ldots, M\}$ (dvs $U \sim U\{1, 2, \ldots, M\}$). Betingat på att $U = m$, låt

$$Y = Y_1 + Y_2 + \ldots + Y_m,$$

där $Y_1, Y_2, \ldots$ är en sekvens av oberoende och likafördelade (dvs i.i.d.) geometriskt fördelade slumpvariabler med parameter $p$, dvs

$$\mathbb{P}(Y_1 = k) = p(1-p)^{k-1} \text{ för } k = 1, 2 \ldots$$

(a) Bestäm sannolikhetsfunktionen för $Y$. (3p)

(b) Beräkna $\mathbb{E}[Y]$. (3p)

8. I långtbortistan förlitar man sig på kolkraft för sin energiförsörjning. Några miljöaktivister misstänker att detta leder till högre halt av smutspartiklar i luften, något som regimen helt avvisar. Miljöaktivisterna mäter därför under åtta år den genomsnittliga partikelhalten under första veckan i juni. De observerar följande dataserie:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| År nr: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Antal kraftverk: | 45 | 49 | 53 | 54 | 57 | 65 | 67 | 72 |
| Partikelhalt (ppm): | 265 | 275 | 287 | 275 | 247 | 309 | 334 | 346 |

Data kan sammanfattas med att $S_{xx} \approx 617.5, S_{yy} \approx 8325.5$ och $S_{xy} \approx 1895.5$.

Miljöaktivisterna ansätter en modell där partikelhalten anses växa linjärt med antalet kraftverk.

(a) Den av kol-lobbyn köpta politikern Täd Krus hävdar att inget samband finns mellan antalet kraftverk och partikelhalten i luften. Använd data ovan för att skatta modellens parametrar och testa Täds hypotes på signifikansnivån 0.001. (2p)

(b) Täd har valt att selektera data och visar därför ofta upp en graf på enbart år 2 till 5 ur datamängden ovan när han debatterar i parlamentet. Vad blir resultatet av testet då? (2p)

(c) Ange förklaringsgraden och beräkna residualerna (för hela datamängden). Kommentera dina resultat och framför kritik mot modellen och hur den används av miljöaktivisterna och av Täd. (3p)

Tentamentsskrivning i **Matematisk Statistik TMA321**

Tid: den 1 juni, 2020 fm
Hjälpmedel: Alla

---

Tentamen består av 8 frågor om sammanlagt 50 poäng. Preliminära betygsgränser är satta till:
betyg "3": 20 till 29 poäng
betyg "4": 30 till 39 poäng
betyg "5": 40 eller fler poäng.

---

OBS! Alla lösningar skall vara väl redovisade, motiverade och fullständiga. Talen är ej ordnade efter svårighetsgrad.

Kom också ihåg att även om alla hjälpmedel är tillåtna så är det inte samma sak som att enbart hänvisa till dessa hjälpmedel. Vad innebär då detta?

(i) Summering av siffror eller annan enklare räkning behöver ej redovisas.

(ii) Nyckelsteg i t.ex. en summation, eller lösning av integral *skall* redovisas om inget annat explicit anges. **Ex:** Man behöver inte *bevisa* formeln för partiell integration men man kan inte bara skriva "partiell integration" och sedan svaret. Genomför istället den partiella integrationen.

1. Låt $X \sim \mathrm{N}(0,1)$, $Y \sim \mathrm{N}(0,3)$ vara oberoende normalfördelade slumpvariabler.

   (a) Beräkna $\mathbb{P}(X \geq Y + 1)$. (3p)

   (b) Låt $U \sim \mathrm{U}[0,1]$ vara oberoende av $X, Y$, och låt $Z = UX + (1-U)Y$. Beräkna $\mathrm{Var}(Z)$. (3p)

   **Lösning:**

   (a) Vi använder att summan av oberoende normalfördelade slumpvariabler är normalfördelade. I detta fall har vi att $X - Y \sim \mathrm{N}(0,4)$. Vi ser då att

   $$\begin{aligned}\mathbb{P}(X \geq Y+1) &= \mathbb{P}(X - Y \geq 1) = \mathbb{P}\left(\frac{X-Y}{2} \geq \frac{1}{2}\right) \\ &= \mathbb{P}(Z \geq 1/2) = 1 - \mathbb{P}(Z \leq 1/2) \approx 1 - 0.6915 = 0.3085,\end{aligned}$$

   där $Z \sim \mathrm{N}(0,1)$.

   (b) Vi har att

   $$\begin{aligned}\mathrm{Var}(Z) &= \mathrm{Var}(UX + (1-U)Y) \\ &= \mathrm{Var}(UX) + \mathrm{Var}((1-U)Y) + \mathrm{Cov}(UX, (1-U)Y).\end{aligned}$$

Vidare är

$$\begin{aligned}\mathrm{Var}(UX) &= \mathbb{E}[(UX)^2] - \mathbb{E}[UX]^2 \\ &= \mathbb{E}[U^2]\mathbb{E}[X^2] - \mathbb{E}[U]^2\mathbb{E}[X]^2 = \mathbb{E}[U^2] = \int_0^1 u^2 du = \left[\frac{u^3}{3}\right]_0^1 = \frac{1}{3},\end{aligned}$$

där vi använder att $\mathbb{E}[X] = 0$ och att $\mathbb{E}[X^2] = 1$. Dessutom är

$$\begin{aligned}\mathrm{Var}((1-U)Y) &= \mathbb{E}[((1-U)Y)^2] - \mathbb{E}[(1-U)Y]^2 \\ &= \mathbb{E}[(1-U)^2]\mathbb{E}[Y^2] - \mathbb{E}[1-U]^2\mathbb{E}[Y]^2 \\ &= 3\mathbb{E}[(1-U)^2] = 3\int_0^1 (1-u)^2 du = 3\left[-\frac{(1-u)^3}{3}\right]_0^1 = 1,\end{aligned}$$

där vi använder att $\mathbb{E}[Y] = 0$ och att $\mathbb{E}[Y^2] = 3$. Till sist ser vi att

$$\begin{aligned}&\mathrm{Cov}(UX, (1-U)Y) \\ &\quad = \mathbb{E}[(UX - \mathbb{E}[UX])((1-U)Y - \mathbb{E}[(1-U)Y])] \\ &\quad = \mathbb{E}[UX(1-U)Y] = \mathbb{E}[U(1-U)]\mathbb{E}[X]\mathbb{E}[Y] = 0,\end{aligned}$$

där vi använder att $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, så att

$$\mathrm{Var}(Z) = \frac{1}{3} + 1 + 0 = \frac{4}{3}.$$

2. Alexandra har en sekvens av oberoende likafördelade slumpvariabler (i.i.d.) $X_1, X_2, \ldots X_n$ med täthetsfunktion

$$f(x) = \frac{1+\alpha}{(1+x)^{2+\alpha}} \text{ för } 0 < x < \infty,$$

där $\alpha > -1$ är en parameter vars värde är okänt.

   (a) Alexandras chef Magnus sa att $\alpha = 1$, men Alexandra är skeptisk. Hjälp Alexandra att designa ett lämpligt likelihood-ratio test för att se ifall Magnus hade fått det hela om bakfoten. (3p)

   (b) När testet ovan var genomfört så kom Alexandras medarbetare José på att han hade hört att $\alpha$ minsann var två ($\alpha$=2). Hjälp Alexandra och José att göra ett likelihood ratio test för att testa Magnus hypotes mot Josés alternativa hypotes. Bestäm även formen för förkastingsregionen (RR) för enklast möjliga härledda teststatistika. Är ditt svar rimligt? (3p)

**Lösning:**

   (a) Vi skall testa

$$H_0 : \alpha = 1 \text{ mot } H_1 : \alpha \neq 1,$$

då vi inte har någon alternativ information. Detta innebär att vi måste använda oss av ett generaliserat likelihood test då $H_1$ är en sammansatt hypotes.
Vi betraktar alltså (med $\Omega_0 = \{1\}$ och $\Omega = \mathbb{R}^+$)

$$\Lambda = \frac{\max_{\alpha\in\Omega_0} \text{lik}(\alpha)}{\max_{\alpha\in\Omega} \text{lik}(\alpha)} = \frac{\text{lik}(1)}{\text{lik}(\hat{\alpha})},$$

där $\hat{\alpha}$ är MLE (maximum likelihood skattaren) för $\alpha$ och $\text{lik}(\alpha)$ är likelihooden. I detta fall

$$\text{lik}(\alpha) = f(X_1, X_2, \ldots, X_n|\alpha) = \prod_{k=1}^{n} f(X_k|\alpha) = \prod_{k=1}^{n} \frac{1+\alpha}{(1+X_k)^{2+\alpha}},$$

så att

$$\begin{aligned} l(\alpha) = \log \text{lik}(\alpha) &= \sum_{k=1}^{n} \log \frac{1+\alpha}{(1+X_k)^{2+\alpha}} \\ &= n\log(1+\alpha) - (2+\alpha)\sum_{k=1}^{n} \log(1+X_k). \end{aligned}$$

Vi ser att

$$l'(\alpha) = \frac{n}{1+\alpha} - \sum_{k=1}^{n} \log(1+X_k)$$

och att

$$l''(\alpha) = -\frac{n}{(1+\alpha)^2} < 0.$$

Detta ger att $l'(\alpha) = 0$ motsvarar ett maximum och villkoret blir

$$\frac{n}{1+\alpha} = \sum_{k=1}^{n} \log(1+X_k) \text{ så att } \hat{\alpha} = \frac{n}{\sum_{k=1}^{n} \log(1+X_k)} - 1.$$

Slutligen ser vi då att

$$\Lambda = \frac{\text{lik}(1)}{\text{lik}(\hat{\alpha})} = \frac{\prod_{k=1}^{n} \frac{2}{(1+X_k)^3}}{\prod_{k=1}^{n} \frac{1+\hat{\alpha}}{(1+X_k)^{2+\alpha}}} = \left(\frac{2}{1+\hat{\alpha}}\right)^n \prod_{k=1}^{n} (1+X_k)^{1-\hat{\alpha}}$$

(b) Denna uppgift blir enklare då vi har två enkla hypoteser ($H_0 : \alpha = 1$ och $H_1 : \alpha = 2$) som ställs mot varandra. Här betraktar vi därför kvoten

$$\frac{\text{lik}(1)}{\text{lik}(2)} = \frac{\prod_{k=1}^{n} \frac{2}{(1+X_k)^3}}{\prod_{k=1}^{n} \frac{3}{(1+X_k)^4}} = \left(\frac{2}{3}\right)^n \prod_{k=1}^{n} (1+X_k).$$

Denna kvot är liten ifall $\prod_{k=1}^{n}(1+X_k)$ är liten, dvs $RR$ är på formen $\{\prod_{k=1}^{n}(1+X_k) \leq c\}$.

Informellt ser vi att vi förkastar testet ifall slumpvariablerna $X_k$ blir små. Detta är rimligt då ett större $\alpha$ ger en snabbare avtagande täthetsfunktion, vilket i sin tur innebär att slumpvariablerna tenderar att bli mindre.

3. Låt $X, Y$ vara två slumpvariabler med följande gemensamma sannolikhetstäthetsfunktion

$$f(x,y) = Cx^3 y e^{-\lambda(x^2+y^2)} \quad 0 \le x, y < \infty.$$

(a) Hitta $C$ så att detta verkligen blir en gemensam täthetsfunktion. (3p)

(b) Är $(X, Y)$ beroende eller oberoende? Bevisa ditt påstående. (3p)

**Lösning:**

(a) Vilkoret är att integralen av täthetsfunktionen över området skall vara 1. Vi har därför att

$$\begin{aligned}
1 &= \int_0^\infty \int_0^\infty Cx^3 y e^{-\lambda(x^2+y^2)} dy dx \\
&= C \int_0^\infty x^3 e^{-\lambda x^2} dx \int_0^\infty y e^{-\lambda y^2} dy \\
&= C \left( \left[ \frac{-x^2}{2\lambda} e^{-\lambda x^2} \right]_0^\infty + \int_0^\infty \frac{x}{\lambda} e^{-\lambda x^2} dx \right) \cdot \left[ \frac{1}{-2\lambda} e^{-\lambda y^2} \right]_0^\infty \\
&= C \left( 0 + \left[ -\frac{1}{2\lambda^2} e^{-\lambda x^2} dx \right]_0^\infty \right) \frac{1}{2\lambda} = C \frac{1}{4\lambda^3},
\end{aligned}$$

så att $C = 4\lambda^3$.

(b) Marginalfördelningarna blir

$$\begin{aligned}
f_X(x) &= \int_0^\infty Cx^3 e^{-\lambda x^2} y e^{-\lambda y^2} dy \\
&= Cx^3 e^{-\lambda x^2} \left[ -\frac{1}{2\lambda} e^{-\lambda y^2} \right]_0^\infty = Cx^3 e^{-\lambda x^2} \frac{1}{2\lambda} = 2\lambda^2 x^3 e^{-\lambda x^2},
\end{aligned}$$

för $0 \le x < \infty$, och vidare är

$$f_Y(y) = Cy e^{-\lambda y^2} \int_0^\infty x^3 e^{-\lambda x^2} dx = Cy e^{-\lambda y^2} \frac{1}{2\lambda^2} = 2\lambda y e^{-\lambda y^2},$$

för $0 \le y < \infty$, där vi använder uträkningen av $\int_0^\infty x^3 e^{-\lambda x^2} dx$ i del $(a)$. Vi ser att $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ och därmed är $(X, Y)$ oberoende.

4. En transportfirma transporterar paket från Fjollträsk till Mars. Firman väger de första 9 paketen de transporterar. Resultatet blev som följer:

| paket: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| vikt i kilo: | 16.5 | 6.7 | 21 | 22.3 | 10.8 | 29.5 | 29.5 | 19.7 | 22.6 |

Transportfirman antar att vikterna är normalfördelade med parametrar $\mu, \sigma^2$ vars värden är okända. De antar också att paketens vikter är oberoende av varandra.

(a) Skatta $\mu, \sigma^2$. (1p)

(b) Hitta konfidensintervall för $\mu$ och $\sigma^2$, båda med konfidensgrad 95%. (3p)

(c) Transportfirman tar emot $M$ paket innan de stuvas in på ett rymdskepp vars maxlast är 2 ton. Bestäm största möjliga värde på $M$ (uttryckt i parametrarna $\mu$ och $\sigma^2$) så att sannolikheten att de $M$ paketens totala vikt överstiger 2 ton är mindre än 0.01. (3p)

**Lösning:**

(a) Låt $X_k$ beteckna vikten av paket nr $k$. Enligt förutsättningarna så är $X_1, \ldots, X_9$ i.i.d. med $X_k \sim N(\mu, \sigma^2)$. Vi använder skattarna $\bar{X}$ och $s^2(X)$ och med våra data får vi

$$\bar{x} \approx 19.84 \text{ och } s^2(x) = \frac{1}{8}\sum_{k=1}^{n}(x_k - \bar{x})^2 \approx 58.4$$

(b) För $\mu$ använder vi referensvariabeln

$$\frac{\bar{X} - \mu}{s(X)/\sqrt{n}} \sim t_8$$

och får då med hjälp av en $t-$tabell att

$$\begin{aligned} 0.95 &= \mathbb{P}\left(-t_{0.025}(8) \leq \frac{\bar{X} - \mu}{s(X)/\sqrt{n}} \leq t_{0.025}(8)\right) \\ &= \mathbb{P}\left(\bar{X} - 2.306\frac{s(X)}{\sqrt{n}} \leq \mu \leq \bar{X} + 2.306\frac{s(X)}{\sqrt{n}}\right). \end{aligned}$$

Vårt 95% numeriska konfidensintervall blir då

$$I_\mu = \bar{x} \pm 2.306\frac{s(x)}{\sqrt{n}} = 19.84 \pm 2.306\frac{\sqrt{58.4}}{\sqrt{n}} \approx [13.97, 25.71].$$

För $\sigma^2$ använder vi istället att

$$\frac{(n-1)s^2(X)}{\sigma^2} \sim \chi^2(n-1)$$

och enligt tabell är $\chi^2_{0.025}(8) \approx 2.18$ medans $\chi^2_{0.975}(8) \approx 17.53$. Vi får

$$\begin{aligned} 0.95 &= \mathbb{P}\left(\chi^2_{0.025}(8) \leq \frac{(n-1)s^2(X)}{\sigma^2} \leq \chi^2_{0.975}(8)\right) \\ &= \mathbb{P}\left(\frac{(n-1)s^2(X)}{\chi^2_{0.975}(8)} \leq \mu \leq \frac{(n-1)s^2(X)}{\chi^2_{0.025}(8)}\right). \end{aligned}$$

Vårt 95% numeriska konfidensintervall blir då

$$I_{\sigma^2} = \left[\frac{(n-1)s^2(x)}{\chi^2_{0.975}(8)}, \frac{(n-1)s^2(x)}{\chi^2_{0.025}(8)}\right] = \left[\frac{8 \cdot 58.4}{17.53}, \frac{8 \cdot 58.4}{2.18}\right] \approx [26.7, 214.3]$$

(c) Vi skall hitta största möjliga $M$ så att $Y = X_1 + \cdots + X_M \leq 2000$ med sannolikhet 0.99. Då $X_k$ är oberoende normalfördelade slumpvariabler har vi att

$$Y \sim N\left(M\mu, M\sigma^2\right) \text{ så att } \frac{Y - M\mu}{\sqrt{M}\sigma} \sim N(0,1).$$

Vi skall välja $M$ så stort som möjligt men sådan att

$$0.99 \geq \mathbb{P}(Y \leq 2000) = \mathbb{P}\left(\frac{Y - M\mu}{\sqrt{M}\sigma} \leq \frac{2000 - M\mu}{\sqrt{M}\sigma}\right) = \mathbb{P}\left(Z \leq \frac{2000 - M\mu}{\sqrt{M}\sigma}\right),$$

där $Z \sim$N$(0,1)$. Ur tabell har att $0.99 = \mathbb{P}(Z \leq 2.33)$ och vi ser därför att $M$ måste uppfylla villkoret

$$\frac{2000 - M\mu}{\sqrt{M}\sigma} \leq 2.33.$$

och vi vill därför lösa ekvationen

$$M + \sqrt{M}\frac{2.33 \cdot \sigma}{\mu} - \frac{2000}{\mu} = 0.$$

Det enklaste är att substituera $M = K^2$ och lösa

$$K^2 + K\frac{2.33 \cdot \sigma}{\mu} - \frac{2000}{\mu} = 0$$

så att

$$K = -\frac{2.33 \cdot \sigma}{2\mu} \pm \sqrt{\left(\frac{2.33 \cdot \sigma}{2\mu}\right)^2 + \frac{2000}{\mu}}.$$

Minustecknet framför rotuttrycket måste vara en falsk rot då vi annars får ett negativt $K$. Vi ser därför att

$$M = K^2 = \left(-\frac{2.33 \cdot \sigma}{2\mu} + \sqrt{\left(\frac{2.33 \cdot \sigma}{2\mu}\right)^2 + \frac{2000}{\mu}}\right)^2.$$

Vi får så klart inte glömma att runda av svaret nedåt till närmaste heltal.

5. Uno har en magisk låda som kan generera slumpvariabler $V_1, V_2, \ldots$ vilka är oberoende och likafördelade (i.i.d.). På Unos låda står det att

$$\mathbb{P}(V_1 = k) = \frac{2k}{N(N+1)}$$

för $k = 1, \ldots, N$. Tyvärr så har Unos bror Sune suddat ut vad värdet på heltalet $N$ är.

   (a) Använd momentmetoden för att hitta en lämplig skattare (dvs MME) åt Uno. (3p)
   **OBS: Eventuella jobbiga uträkningar behöver ej redovisas i denna del. Enbart hänvisning till tabeller eller programvara räcker just här.**

   (b) Uno vill gärna hitta ett numeriskt värde på $N$. Tyvärr så drar Unos låda en hel del el, så han vill inte göra för många simuleringar. De sju stycken han gör ger följande resultat:

| Simulering nr: | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Resultat: | 14 | 16 | 5 | 12 | 18 | 19 | 25 |

   Vilken numerisk skattning får ni på $N$? Reflektera över ert resultat. (3p)

**Lösning:**

   (a) Enligt momentmetoden skall vi beräkna $E[V_1]$ och se hur denna beror på $N$. Vi får att

$$\begin{aligned}\mathbb{E}[V_1] &= \sum_{k=1}^{N} k\mathbb{P}(V_1 = k) = \sum_{k=1}^{N} k\frac{2k}{N(N+1)} \\ &= \frac{2}{N(N+1)}\sum_{k=1}^{N} k^2 = \frac{2}{N(N+1)}\frac{N(N+1)(2N+1)}{6} = \frac{2N+1}{3},\end{aligned}$$

   där vi använde t.ex. WolframAlpha för att beräkna summan. Vi löser nu ekvationen

$$\bar{V} = \frac{2\hat{N}+1}{3},$$

   så att

$$\hat{N} = \frac{3\bar{V}-1}{2}$$

   är vår momentskattare.

   (b) Med insatta värden får vi att $\bar{v} \approx 15.57$ så att

$$\hat{N}(v) \approx 22.86.$$

Vi vet ju att $N$ är ett heltal, så ur det perspektivet är gissning 22.86 rätt konstig. Men även om vi avrundar till säg 23 så är det ändå inte speciellt bra då vi har en realisation som var 25 (vilket ju innebär att $N \geq 25$). Slutsatsen är att momentmetoden kanske inte är så lyckad här och att vi istället bör använda en annan metod.

6. Låt $X$ ha täthetsfunktion

$$f(x) = \frac{a}{2(e^a - 1)} e^{a|x|} \text{ för } x \in [-1, 1],$$

där $a > 0$ är en parameter.

(a) Beräkna den momentgenererande funktionen (här betecknad $M_{X,a}(t)$) för $X$. (3p)

(b) Hitta en diskret slumpvariabel (dvs bestäm sannolikhetsfunktionen) $Y$ vars momentgenererande funktion $M_Y(t)$ uppfyller villkoret

$$\lim_{a \to \infty} M_{X,a}(t) = M_Y(t).$$

(3p)

**Lösning**:

(a) Vi har att

$$\begin{aligned} M_{X,a}(t) = \mathbb{E}[e^{tX}] &= \int_{-1}^{1} e^{tx} \frac{a}{2(e^a - 1)} e^{a|x|} dx \\ &= \frac{a}{2(e^a - 1)} \left( \int_{-1}^{0} e^{(t-a)x} dx + \int_{0}^{1} e^{(t+a)x} dx \right) \\ &= \frac{a}{2(e^a - 1)} \left( \frac{1 - e^{a-t}}{t - a} + \frac{e^{a+t} - 1}{t + a} \right). \end{aligned}$$

(b) Vi får då

$$\frac{a}{2(e^a - 1)} \left( \frac{1 - e^{a-t}}{t - a} \right) = \frac{a}{2(t - a)} \frac{e^{-a} - e^{-t}}{1 - e^{-a}} \to \frac{-1}{2}(-e^{-t}) = \frac{e^{-t}}{2} \text{ då } a \to \infty.$$

På liknande sätt ser vi att

$$\frac{a}{2(e^a - 1)} \left( \frac{e^{a+t} - 1}{t + a} \right) =\to \frac{e^t}{2} \text{ då } a \to \infty,$$

och vi ser alltså att $\lim_{a \to \infty} M_{X,a}(t) = \frac{e^t + e^{-t}}{2}$. Den momentgenererande funktionen för en diskret slumpvariabel $Y$ ges av

$$M_Y(t) = \sum_{k=-\infty}^{\infty} e^{kt} P(Y = k)$$

som blir likamed $\frac{e^t+e^{-t}}{2}$ omm $P(Y=1)=P(Y=-1)=1/2$.
Detta är högst rimligt. Då $a$ växer så blir $X$ mer och mer koncentrerad nära kanterna av området $[-1,1]$, och då vi går i gräns koncentreras all sannolikhet i dessa två punkter.

7. Låt $U$ vara likformigt fördelad på mängden $\{1,2,\ldots,M\}$ (dvs $U \sim U\{1,2,\ldots,M\}$). Betingat på att $U=m$, låt

$$Y = Y_1 + Y_2 + \ldots + Y_m,$$

där $Y_1, Y_2, \ldots$ är en sekvens av oberoende och likafördelade (dvs i.i.d.) geometriskt fördelade slumpvariabler med parameter $p$, dvs

$$\mathbb{P}(Y_1 = k) = p(1-p)^{k-1} \text{ för } k = 1, 2 \ldots$$

(a) Bestäm sannolikhetsfunktionen för $Y$. (3p)

(b) Beräkna $\mathbb{E}[Y]$. (3p)

**Lösning:**

(a) Lösningen förenklas om man vet att summan av oberoende geometriskt fördelade slumpvariabler är negativt binomialfördelad. Vi har därför att

$$\mathbb{P}(Y=k|U=m) = \binom{k-1}{m-1} p^m (1-p)^{k-m} \text{ för } k = m, m+1, \ldots$$

och får då att

$$\begin{aligned}\mathbb{P}(Y=k) &= \sum_{m=1}^{M} \mathbb{P}(Y=k|U=m)\mathbb{P}(U=m) \\ &= \sum_{m=1}^{\min(M,k)} \binom{k-1}{m-1} p^m (1-p)^{k-m} \frac{1}{M},\end{aligned}$$

för $k = 1, 2, \ldots$.

(b) Vi använder oss här av att $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|U]]$ och börjar därför med att bestämma $\mathbb{E}[Y|U=m]$. Återigen kan vi använda att betingat på $U=m$ så är $Y$ negativt binomialfördelad, eller också kan man använda enkel linearitet. På båda sätten ser vi att

$$\mathbb{E}[Y|U=m] = \mathbb{E}[Y_1 + \ldots + Y_m] = m\mathbb{E}[Y_1] = m\frac{1}{p},$$

där vi använder oss av att $Y_1$ är geometriskt fördelad i sista likheten. Därför är $\mathbb{E}[Y|U] = U\frac{1}{p}$ och vi ser då att

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|U]] = \mathbb{E}[Y] = \mathbb{E}\left[U\frac{1}{p}\right] = \frac{1}{p}\mathbb{E}[U] = \frac{1}{p}\frac{M+1}{2}.$$

8. I långtbortistan förlitar man sig på kolkraft för sin energiförsörjning. Några miljöaktivister misstänker att detta leder till högre halt av smutspartiklar i luften, något som regimen helt avvisar. Miljöaktivisterna mäter därför under åtta år den genomsnittliga partikelhalten under första veckan i juni. De observerar följande dataserie:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| År nr: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Antal kraftverk: | 45 | 49 | 53 | 54 | 57 | 65 | 67 | 72 |
| Partikelhalt (ppm): | 265 | 275 | 287 | 275 | 247 | 309 | 334 | 346 |

Data kan sammanfattas med att $S_{xx} \approx 617.5, S_{yy} \approx 8325.5$ och $S_{xy} \approx 1895.5$.

Miljöaktivisterna ansätter en modell där partikelhalten anses växa linjärt med antalet kraftverk.

(a) Den av kol-lobbyn köpta politikern Täd Krus hävdar att inget samband finns mellan antalet kraftverk och partikelhalten i luften. Använd data ovan för att skatta modellens parametrar och testa Täds hypotes på signifikansnivån 0.001. (2p)

(b) Täd har valt att selektera data och visar därför ofta upp en graf på enbart år 2 till 5 ur datamängden ovan när han debatterar i parlamentet. Vad blir resultatet av testet då? (2p)

(c) Ange förklaringsgraden och beräkna residualerna (för hela datamängden). Kommentera dina resultat och framför kritik mot modellen och hur den används av miljöaktivisterna och av Täd. (3p)

**Lösning:**

(a) Vi ansätter $y = \beta_0 + \beta_1 x$ och har att

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \approx 3.07 \text{ och } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \approx 115.$$

Vi vill testa

$$H_0 : \beta_1 = 0 \text{ mot } \beta_1 \neq 0.$$

Vi använder sedan att

$$\frac{\hat{\beta}_1 - \beta_1}{s_r/\sqrt{S_{xx}}} \sim t(6)$$

så att ett 99.9% konfidensintervall (ensidigt är också ok) ges av (Många använde $t_{0.005}(6) \approx 3.707$ fast det skulle vara $t_{0.0005}(6) \approx 5.96$. Så länge man varit konsekvens är båda ok. den andra räknas enklast ut i Matlab med komandot tinv ifall man inte hittar lämplig tabell)

$$\begin{aligned} 0.999 &= \mathbb{P}\left(-t_{0.0005}(6) \leq \frac{\hat{\beta}_1 - \beta_1}{s_r/\sqrt{S_{xx}}} \leq t_{0.0005}(6)\right) \\ &= \mathbb{P}\left(\hat{\beta}_1 - t_{0.0005}(6)\frac{s_r}{\sqrt{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{0.0005}(6)\frac{s_r}{\sqrt{S_{xx}}}\right). \end{aligned}$$

Vidare har vi att

$$s_r^2(y) = \frac{1}{6}\sum_{k=1}^{8}(y_i - \hat{y}_i)^2 \approx 418$$

Den numeriska varianten blir då ifall man använder $t_{0.005}(6) \approx 5.96$

$$I_{\beta_1} = \hat{\beta}_1 \pm 3.707\frac{s_r}{\sqrt{S_{xx}}} \approx [0.02, 6.12].$$

men ifall man använder $t_{0.0005}(6) \approx 5.96$ så får man istället

$$I_{\beta_1} = \hat{\beta}_1 \pm 5.96\frac{s_r}{\sqrt{S_{xx}}} \approx [-1.83, 7.97].$$

I variant 1 förkastas $H_0$ men i variant 2 förkastas den ej.

(b) Här är datamängden mycket mindre och vi tvingas själva räkna ut $S_{xx} \approx 32.75, S_{yy} \approx 864$ och $S_{xy} \approx -108$. Vi får nu att

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \approx -3.3 \text{ och } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \approx 446.6$$

$$s_r^2(y) = \frac{1}{2}\sum_{k=1}^{4}(y_i - \hat{y}_i)^2 \approx 254$$

och med $t_{0.005}(2) \approx 9.925$ får vi då att

$$I_{\beta_1} = \hat{\beta}_1 \pm 9.93\frac{s_r}{\sqrt{S_{xx}}} \approx [-30.9, 24.4].$$

Om vi istället använder $t_{0.0005}(2) \approx 31.6$

$$I_{\beta_1} = \hat{\beta}_1 \pm 3.707\frac{s_r}{\sqrt{S_{xx}}} \approx [-91.3, 84.7].$$

(c) Förklaringsgraden ges av

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} \approx 0.70.$$

Vi beräknar residualerna mha uttrycket $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ och får följande tabell (efter avrundningar):

| Res nr: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Värde: | 11.9 | 9.6 | 9.3 | −5.73 | −43 | −5.5 | 13.4 | 10.0 |

Vi observerar följande:

i. Förklaringsgraden är medioker/relativt låg.

ii. De försummar självklart alldeles för mycket genom att inte ta hänsyn till andra faktorer såsom andra utsläppskällor i landet och resten av världen.

iii. De tar ingen som helst hänsyn till hur stora kraftverken är eller vilken teknologi de använder.

iv. Det är mycket anmärkningsvärt att en residual avviker så mycket från de andra. Varför avvek detta värde så kraftigt? Det hade varit intressant att analysera vad som händer ifall man tar bort den avvikande punkten.

v. Signifikansnivån är anmärkningsvärt låg.

vi. Täd är antingen dum eller medvetet svekfull. Att klippa ur lämpliga delmängder ur data är djupt oetiskt och usel vetenskap. Det behöver inte i sig vara fel att ta bort en enstaka datapunkt, men om man gör det måste detta redovisas och motiveras extremt noggrannt.

vii. Att överhuvudtaget försöka dra meningsfulla statistiska slutsatser av åtta datapunkter är tveksamt, och fyra gränsar till vansinne.