Aalto-universitetet

MS-A0509 Grundkurs i sannolikhetskalkyl och statistik Gripenberg
Övning 4
Vecka 5, 26–30.1.2015

**I1.** Vattenhöjden i Saramojoki var under åren 2008–2010 följande:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 76 | 54 | 52 | 98 | 170 | 59 | 42 | 80 | 69 |
| 99 | 131 | 86 | 58 | 48 | 42 | 53 | 137 | 50 |
| 45 | 35 | 43 | 65 | 64 | 73 | 46 | 39 | 38 |
| 108 | 129 | 111 | 45 | 24 | 37 | 51 | 70 | 43 |

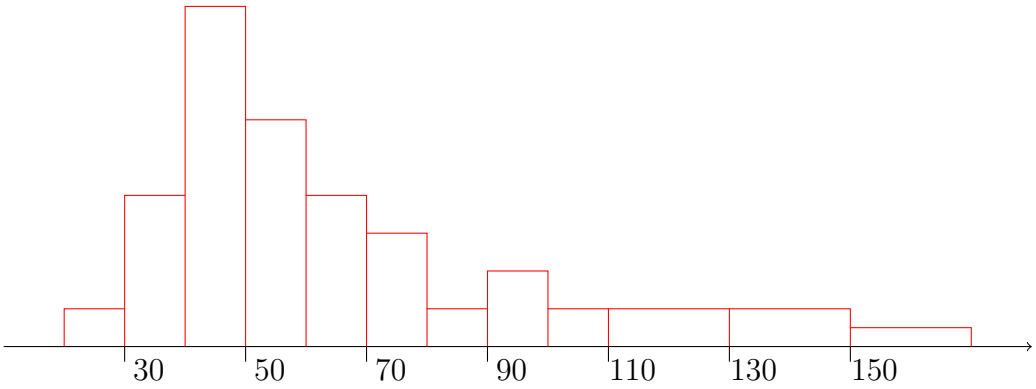
Bilda en frekvenstabell med klasserna $21-30$, $31-40$, $41-50$, $51-60$, $61-70$, $71-80$, $81-90$, $91-100$, $101-110$, $111-130$, $131-150$ och $151-170$ och rita ett histogram över materialet (och kom ihåg att areorna av rektanglarna skall vara proportionella mot frekvenserna).

*Ledning: I storleksordning är siffrorna* 24, 35, 37, 38, 39, 42, 42, 43, 43, 45, 45, 46, 48, 50, 51, 52, 53, 54, 58, 59, 64, 65, 69, 70, 73, 76, 80, 86, 98, 99, 108, 111, 129, 131, 137, 170.

*Lösning:* Klassfrekvenserna är följande:

| Klass | 21-30 | 31-40 | 41-50 | 51-60 | 61-70 | 71-80 |
|---|---|---|---|---|---|---|
| Frekvens | 1 | 4 | 9 | 6 | 4 | 3 |
| Klass | 81-90 | 91-100 | 101-110 | 111-130 | 131-150 | 151-170 |
| Frekvens | 1 | 2 | 1 | 2 | 2 | 1 |

Histogrammet ser ut på följande sätt:




---

**I2.** Använd samma data som i uppgift I1. Beräkna medianen, variationsbredden (största minus minsta), och avståndet mellan kvantilerna $x_{0.75}$ och $x_{0.25}$. Vilken av siffrorna 33.533, 68.611 och 1124.5 är medelvärdet, vilken varians och vilken standardavvikelse. (Du behöver inte siffrorna i uppgift I1 för att svara på den senare frågan.)

Om du skulle vara tvungen att beskriva siffermaterialet med två tal, vilka skulle du välja? Blir det en bra beskrivning med de två tal du valt? (Det finns inte ett entydigt rätt svar på dehär frågorna!)

*Lösning:* Medianen är vilket tal som helst i intervallet $[54, 58]$, tex. 56.

Variationsbredden är $170 - 24 = 146$.

Kvantilerna $x_{0.25}$ och $x_{0.75}$ är 44 och 83 (om vi igen väljer mittpunkten av de intervall i vilka kvantilerna ligger) så att $x_{0.75} - x_{0.25} = 39$.

Eftersom $\sqrt{1124.5} = 33.354$ så ser vi att 1124.5 måste vara variansen, 33.533 måste vara standardavvikelsen så att 68.611 blir medelvärdet.

Det kan vara förnuftigt att använda medelvärdet och variansen, medianen och variationsbredden eller medianen och avståndet mellan kvantilerna $x_{0.75}$ och $x_{0.25}$. Att här byta ut medelvärdet mot medianen är inte helt fel men kanske inte riktigt motiverat. Däremot är det inte förnuftigt att ta med variansen eller att varken ta med medelvärdet eller medianen.

Det som är svårt att beskriva med två av de givna talen är att fördelningen av mätvärdena är så osymmetrisk.

---

**I3.** Antag att du har ett observerat stickprov $x_j$, $j = 1, \ldots, 8$ av en slumpvariabel som du antar har fördelningen $\mathrm{N}(\mu, \sigma^2)$. Du har räknat ut medelvärdet $\overline{x}$ och stickprovsvariansen $s^2 = 2.6753$, och sedan på vanligt sätt bestämt ett (symmetriskt) konfidensintervall som blev $[2.6543, 6.7017]$. Vad är konfidensgraden?

*Lösning:* Konfindensintervallet har i detta fall formen

$$\left[\overline{x} - t_{\frac{\alpha}{2},n-1}\sqrt{\frac{s^2}{n}}, \overline{x} + t_{\frac{\alpha}{2},n-1}\sqrt{\frac{s^2}{n}}\right],$$

där $t_{\frac{\alpha}{2},n-1} = -F^{-1}_{\mathrm{t}(n-1)}(\frac{\alpha}{2}) = F^{-1}_{\mathrm{t}(n-1)}(1 - \frac{\alpha}{2})$. Längden av konfidensintervallet är alltså

$$2t_{\frac{\alpha}{2},n-1}\sqrt{\frac{s^2}{n}} = 6.7017 - 2.6543 = 4.0474.$$

Eftersom $s^2 = 2.341$ och $n = 8$ så blir

$$t_{\frac{\alpha}{2},7} = \frac{4.0474}{2\sqrt{\frac{2.6753}{8}}} = 3.4995$$

Eftersom

$$F_{\mathrm{t}(7)}(-3.4995) = 0.0049999$$

så drar vi slutsatsen att $\alpha = 0.01$.

---

**I4.** Antag att $X_i, i = 1, 2, \ldots, n$ är oberoende slumpvariabler så att $\mathrm{E}(X_i) = 0$ och $\mathrm{Var}(X_i) = \sigma^2$ då $i = 1, 2, \ldots, n$.
Visa att $\mathrm{E}(S^2) = \sigma^2$ då

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

och $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ dvs. stickprovsvariansen är en **väntevärdesriktig** estimator av variansene.

*Ledning: Räkna* $\mathrm{E}\big((n-1)S^2\big)$ *och observera att eftersom* $\mathrm{E}(X_i) = 0$ *så är* $\mathrm{E}(X_i^2) = \mathrm{Var}(X_i)$ *och* $\mathrm{E}(\overline{X}) = 0$ *så att* $\mathrm{E}(\overline{X}^2) = \mathrm{Var}(\overline{X})$ *och kom ihåg vad* $\mathrm{Var}(\overline{X})$ *blir då slumpvariablerna* $X_i$ *är oberoende.*

Obs! *Om* $\mathrm{E}(X_i) = \mu \neq 0$ *så kunde vi byta ut slumpvariablerna* $X_i$ *mot slumpvariablerna* $Y_i = X_i - \mu$ *utan att* $S^2$ *skulle ändras och detta betyder att antagandet att* $\mathrm{E}(X_i) = 0$ *inte är en begränsning, endast en förenkling.*

*Lösning:*

$$\begin{aligned}\mathrm{E}\big((n-1)S^2\big) &= \mathrm{E}\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right) = \mathrm{E}\left(\sum_{i=1}^{n}\Big(X_i^2-2X_i\overline{X}+(\overline{X})^2\Big)\right)\\ &= \mathrm{E}\left(\sum_{i=1}^{n}X_i^2\right)+\mathrm{E}\left((-2)\sum_{i=1}^{n}X_i\overline{X}\right)+\mathrm{E}\left(\sum_{i=1}^{n}\overline{X}^2\right)\\ &= \sum_{i=1}^{n}\mathrm{E}\big(X_i^2\big)+\mathrm{E}\left((-2n)\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)\overline{X}\right)+\mathrm{E}\left(n\overline{X}^2\right)\\ &= \left(\sum_{i=1}^{n}\sigma^2\right)+(-2n+n)\mathrm{E}(\overline{X}^2) = n\sigma^2-n\frac{1}{n}\sigma^2=(n-1)\sigma^2,\end{aligned}$$

och av detta följer påståendet när man dividerar med $n-1$.

---

**I5.** Slumpvariablen $X$ har täthetsfunktionen

$$f(t,\theta)=\begin{cases}2^{\theta-1}(\theta-1)t^{-\theta}, & t\geq 2,\\ 0, & t<2,\end{cases}$$

där $\theta>1$. Ett stickprov gav värdena 4, 6, 8 och 14. Bestäm ett estimat för $\theta$ med momentmetoden och ett med "maximum likelihood"-metoden.
*Ledning: Kom ihåg att* $\int_2^\infty t2^{\theta-1}(\theta-1)t^{-\theta}\,\mathrm{d}t = 2\frac{\theta-1}{\theta-2}$ *då* $\theta>2$ *och att* $\frac{\mathrm{d}}{\mathrm{d}\theta}a^\theta = a^\theta\ln(a)$.

*Svar:* 2.3333 och 1.7806

*Lösning:* (a) Slumpvariabelns $X$ väntevärde är

$$\mathrm{E}(X)=\int_2^\infty t2^{\theta-1}(\theta-1)t^{-\theta}\,\mathrm{d}t = 2\frac{\theta-1}{\theta-2},$$

förutsatt att $\theta>2$ för om $\theta\leq 2$ så har slumpvariabeln inte något väntevärde.

Estimatet av $\theta$ fås ur ekvationen $\mathrm{E}(X)=\overline{x}$ dvs.

$$2\frac{\theta-1}{\theta-2}=\frac{1}{4}(4+6+8+14)=8,$$

och genom att dividera med 2 och multiplicera med $\theta-2$ får vi $\theta-1=4(\theta-2)$ vilket ger estimatet

$$\hat{\theta}_{MM}=\frac{7}{3}\approx 2.3333.$$

(b) "Likelihood"-funktionen för ett observerat stickprov $x_i, i=1,2,\ldots,n$ är

$$\begin{aligned}L(\theta,x_1,\ldots,x_n)=f(x_1,\theta)\cdot f(x_2,\theta)\cdot\ldots\cdot f(x_n,\theta) &= 2^{n(\theta-1)}(\theta-1)^n(x_1x_2\ldots x_n)^{-\theta}\\ &= 2^{-n}(\theta-1)^n(2^{-n}x_1x_2\ldots x_n)^{-\theta}.\end{aligned}$$

Maximipunkten fås i något av derivatans nollställen:

$$\begin{aligned}0=L'(\theta,x_1,\ldots,x_n) &= n2^n(\theta-1)^{n-1}(2^{-n}x_1x_2\ldots x_n)^{-\theta}\\ &\quad -2^n(\theta-1)^n(2^{-n}x_1x_2\ldots x_n)^{-\theta}\ln(2^{-n}x_1x_2\ldots x_n)\\ &= L(\theta,x_1,\ldots,x_n)\left(\frac{n}{\theta-1}-\ln(2^{-n}x_1x_2\ldots x_n)\right),\end{aligned}$$

av vilket vi får

$$\hat{\theta} = 1 + \frac{n}{\ln(2^{-n} x_1 x_2 \dots x_n)}.$$

(Eftersom funktionen $\frac{n}{\theta - 1}$ är avtagande är detta verkligen en maximipunkt.) I detta fall blir

$$\hat{\theta} = 1 + \frac{4}{\ln(2^{-4} \cdot 4 \cdot 6 \cdot 8 \cdot 14)} = 1 + \frac{4}{\ln(168)} \approx 1.7806.$$

---