

Statistiek I (G0B19a): Examen

24 Januari 2020

Naam:

Vraag 1 (4pt) Stel je beschikt over een steekproef van onafhankelijke en identiek verdeelde toevalsvariabelen $X_1, \dots, X_n$ uit de Parcto verdeling met dichtheid

$$f(x) = \theta x_0^\theta x^{-\theta-1}, \quad x > x_0$$

waarbij $\theta > 0$ ongekend en x_0 gekend is.

- a. Toon aan dat $T_n = \frac{1}{n} \sum_{i=1}^n \log \frac{X_i}{x_0}$ een onvertekende schatter is van $\frac{1}{\theta}$.
- b. Bereken de variantie van deze schatter T_n .
- c. Als n voldoende groot is zodat dankzij de centrale limietstelling T_n bij benadering normaal verdeeld is, bepaal dan een 95% betrouwbaarheidsinterval voor θ .

$$I a) \quad E\left(\log \frac{X}{x_0}\right)$$

$$= \int_{x_0}^{\infty} \left(\log \frac{y}{x_0}\right) \theta x_0^{\theta} y^{-\theta-1} dy$$

$$= \int_0^{\infty} v \cdot \theta x_0^{\theta} (x_0 e^v)^{-\theta-1} x_0 e^v dv$$

$$\log \frac{y}{x_0} = v$$

$$y = x_0 e^v$$

$$= \int_0^{\infty} \theta v \cdot e^{-\theta v} dv$$

$$(w=\theta v) \int_0^{\infty} w e^{-w} \frac{dw}{\theta} = \frac{1}{\theta}$$

$$b) \quad \text{Var}\left(\log \frac{X}{x_0}\right) = E\left(\log^2 \frac{X}{x_0}\right) - \left(\frac{1}{\theta}\right)^2$$

$$= \int_0^{\infty} v^2 (\theta e^{-\theta v}) dv - \frac{1}{\theta^2}$$

$$\boxed{\text{Var}(T_n) = \frac{1}{n\theta^2}}$$

$$= \frac{2}{\theta^2} - \frac{1}{\theta^2} = \frac{1}{\theta^2}$$

Merksop: je kan a) en b) ook terug vinden door te beschouwen

$Y_i = \log \frac{X_i}{x_0}$, $i=1, \dots, n$
 en na te gaan dat Y_i exponentieel verdeeld is met dichtheid $f_Y(y) = \theta e^{-\theta y}$, $y > 0$.

Ic) De centrale limitstelling stelt dat

$$\frac{T_n - \frac{1}{\theta}}{\frac{1}{\theta\sqrt{n}}} \underset{n \rightarrow \infty}{\approx} \mathcal{N}(0,1)$$

of

$$\sqrt{n}(\theta T_n - 1) \approx \mathcal{N}(0,1)$$

Dus met 95% betrouwbaarheid hebben we dat

$$-1,96 \leq \sqrt{n}(\theta T_n - 1) \leq 1,96$$

$$1 - \frac{1,96}{\sqrt{n}} \leq \theta T_n \leq 1 + \frac{1,96}{\sqrt{n}}$$

$$\boxed{\frac{1}{T_n} \left(1 - \frac{1,96}{\sqrt{n}}\right) \leq \theta \leq \frac{1}{T_n} \left(1 + \frac{1,96}{\sqrt{n}}\right)}$$

Vraag 2 (3pt) S^2 stelt de steekproefvariantie voor van een lukrake steekproef van omvang n uit een normale verdeling met μ en σ^2 beide onbekend. De verdeling van $(n-1)S^2/\sigma^2$ is dus χ_{n-1}^2 verdeeld. We beschouwen een betrouwbaarheidsinterval voor σ^2 van niveau 0.95 van het type

$$((n-1)S^2/b; (n-1)S^2/a)$$

met a en b twee positieve getallen met $a < b$.

- Aan welke eis moeten a en b voldoen opdat het interval inderdaad het gewenste betrouwbaarheidsniveau heeft?
- Toon aan dat de lengte van het interval minimaal is als a en b voldoen aan

$$a^2 g(a) = b^2 g(b)$$

waarbij g de dichtheid van de χ_{n-1}^2 verdeling voorstelt.

$$\text{II a)} \quad P\left((n-1) \frac{S^2}{b} \leq \sigma^2 \leq (n-1) \frac{S^2}{a}\right) = 0,95$$

of

$$P\left(a \leq (n-1) \frac{S^2}{\sigma^2} \leq b\right) = 0,95$$

Noten met G de verdelingsfunctie van de χ^2_{n-1} verdeling, dan is de gezochte eis dus dat

$$\boxed{G(b) - G(a) = 0,95}$$

Merk op dat

$b = Q(0,95 + G(a))$
met Q de kwantiefunctie van de χ^2_{n-1} verdeling

b) De lengte is

$$(n-1)S^2 \left(\frac{1}{a} - \frac{1}{b} \right) = (n-1)S^2 \left(\frac{1}{a} - \frac{1}{Q(0,95 + G(a))} \right)$$

Afleiden naar a levert

$$(*) \quad -\frac{1}{a^2} + \frac{1}{Q^2(0,95 + G(a))} \times \frac{g(a)}{g(Q(0,95 + G(a)))} = 0$$

omdat

$$\frac{d}{da} Q(0,95 + G(a)) = \frac{g(a)}{g(Q(0,95 + G(a)))},$$

(*) is equivalent met

$$\frac{1}{a^2} = \frac{1}{b^2} \cdot \frac{g(a)}{g(b)} \quad \text{of} \quad a^2 g(a) = b^2 g(b)$$

Vraag 3 (4pt) Twee verschillende types gereedschap (A en B genoemd) werden gebruikt bij een belangrijk stuk in een proefopstelling. De sterktemeting (in ksi) van 40 stukken gereedschap van A en 42 van B werden bepaald, en als volgt samengevat:

Sterktemeting	A	B
26 - < 30	6	4
30 - < 34	12	9
34 - < 38	15	19
38 - < 42	7	10
	40	42

- Bepaal een 95% betrouwbaarheidsinterval voor het verschil in proporties van A en B specimens die een sterktemeting boven de 34 ksi hebben.
- Voer een hypothesetest uit om na te gaan (op significantieniveau 0.05) of de sterkteverdelingen afhankelijk zijn van het gekozen type, en dus of de twee types een andere verdeling hebben. Bepaal de P-waarde.

III a) De geschatte proportie van A met sterkte meting boven 34 ksi wordt gegeven door

$$\hat{p}_A = \frac{22}{40}$$

$$\hat{p}_B = \frac{29}{42}$$

Het betrouwbaarheidsinterval wordt dan gegeven door

$$\hat{p}_A - \hat{p}_B \pm 1,96 \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{40} + \frac{\hat{p}_B(1-\hat{p}_B)}{42}}$$

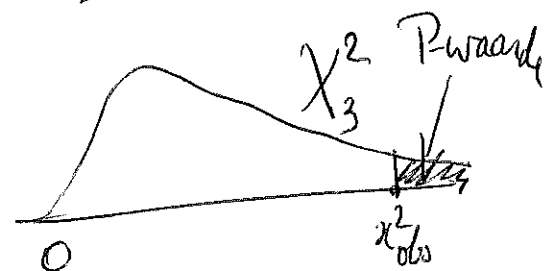
b) Hier gebruiken we de χ^2 test

Sterkte			
	A	B	
1	6	4	10
2	12	9	21
3	15	19	34
4	7	10	17
	40	42	82

$$\chi^2_{obs} = \frac{\left(6 - \frac{10 \times 40}{82}\right)^2}{\frac{10 \times 40}{82}} + \dots + \frac{\left(10 - \frac{17 \times 42}{82}\right)^2}{\frac{17 \times 42}{82}}$$

We plaatsen dit getal onder de χ^2_3 verdeling (8 termen)

$$P(\chi^2_3 > x^2_{obs})$$



Vraag 4 (3pt) Om metaalresten te detecteren in een rivier werden op 6 locaties langs die rivier de zinkconcentraties (mg/l) bepaald in het oppervlaktewater én in het slib. Kan je uit deze data afleiden dat de gemiddelde zinkconcentratie in het slib hoger is dan aan de oppervlakte? Neem significantieniveau 0.01 en beslis op basis van de P-waarde.

Locatie	1	2	3	4	5	6
Bodem (slib) concentratie	0.430	0.266	0.567	0.531	0.707	0.716
Oppervlakte concentratie	0.415	0.238	0.390	0.410	0.605	0.609

Het gaat hier om gepaarde waarnemingen
 Per locatie beschouwen we de verschillen van
 de 2 intkomsten: $V = \text{Bodem conc.} - \text{Opp. Conc.}$

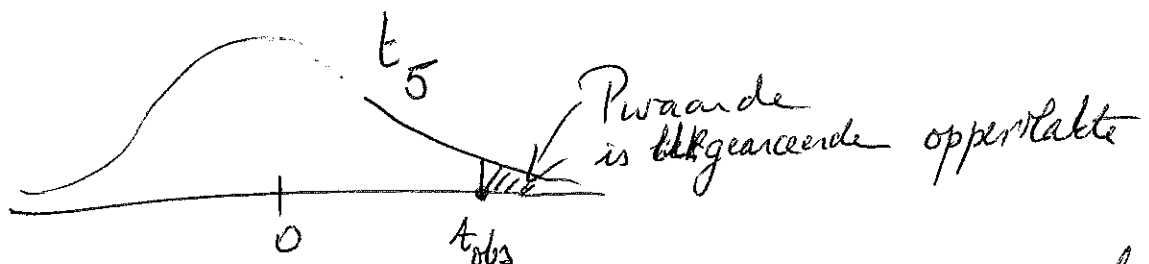
$$V = 0,015 \quad 0,028 \quad 0,177 \quad 0,121 \quad 0,102 \quad 0,107$$

$$H_0: \mu_V \leq 0 \quad \text{versus} \quad H_1: \mu_V > 0$$

met teststatistiek $T = \frac{\bar{V}}{s_V / \sqrt{n}} \stackrel{H_0}{\sim} t_5$

$$P\text{-waarde} = P(T > t_{\text{obs}})$$

$$\text{met } t_{\text{obs}} = \frac{\bar{v}}{s_v} \sqrt{6} = \dots$$



Verwerp H_0 als $P\text{-waarde} < 0,01$, wat equivalent
 is met $t_{\text{obs}} > t_{5; 0,01} = 3,365$