t-test for en normalfordelt stikprøve

Lad $\mathbf{x} = (x_1, \ldots, x_n)$ være en normalfordelt stikprøve med ukendt middelværdi $\theta \in \mathbb{R}$ og ukendt varians $\phi > 0$.

1. Angiv likelihoodfunktionen og log-likelihoodfunktionen.

2. Vis, at maximum likelihood estimatet er givet ved

$$\hat{\theta} = \bar{x}, \qquad \hat{\phi} = \sum_{i=1}^{n} (x_i - \bar{x})^2 / n,$$

forudsat x_i-erne ikke alle er ens (hvor $\bar{x} = \sum_{i=1}^{n} x_i / n$). Er $(\hat{\theta}, \hat{\phi})$ sufficient for (θ, ϕ)?

3. Betragt punkthypotesen $H_0 : \theta = \theta_0$, hvor θ_0 er et givet tal. Vis, at maximum likelihood estimatet under H_0 er givet ved

$$\hat{\theta} = \theta_0, \qquad \hat{\phi}_0 = \sum_{i=1}^{n} (x_i - \theta_0)^2 / n,$$

forudsat x_i-erne ikke alle er lig med θ_0.

4. Hvad er sandsynligheden for, at alle x_i-erne er ens?

5. Vis, at likelihood ratio testet for H_0 er givet ved $Q(\mathbf{x}) = (\hat{\phi}/\hat{\phi}_0)^{n/2}$. Mere præcist er likelihood ratio testet veldefineret med sandsynlighed 1 - hvorfor?

6. Vis, at

$$Q(\mathbf{x}) = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \theta_0)^2} \right)^{n/2}.$$

7. Slut heraf, at små værdier af $Q(\mathbf{x})$ svarer til numerisk store værdier af t-testoren givet ved

$$t(\mathbf{x}) = \frac{\bar{x} - \theta_0}{s/\sqrt{n}}$$

(hvor $s = \sqrt{s^2}$ og $s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1) > 0$ med sandsynlighed 1).

NB: Det kan vises, at $t = t(\mathbf{x})$ under antagelse af H_0 er t_{n-1}-fordelt (dette ønskes ikke bevist her men må gerne benyttes nedenfor).

8. Antag $\mathbf{x} = \mathbf{x}_0$ er observeret (således at $\mathbf{x}_0$ ikke har alle koordinater ens). Vis at p-værdien for likelihood ratio testet for H_0 er givet ved

$$p(\mathbf{x}_0) = P(|t| > |t(\mathbf{x}_0)|).$$

9. I klassisk statistik konstrueres et 95% konfidensinterval for θ som følger. Hvis q betegner 97.5% fraktilen for t_{n-1}-fordelingen, så er $P(|t| > q) = 5\%$. Altså accepteres H_0 på niveau 5% netop når data $\mathbf{x}$ er således at

$$\left|\frac{\bar{x}-\theta_0}{s/\sqrt{n}}\right| \leq q, \text{ dvs. } \quad \bar{x} - qs/\sqrt{n} \leq \theta_0 \leq \bar{x} + qs/\sqrt{n}.$$

Følgelig er et 95% konfidensinterval for θ givet ved

$$[\bar{x} - qs/\sqrt{n}, \bar{x} + qs/\sqrt{n}]. \tag{1}$$

Diskuter om dette giver mening.

10. Hvilken priorfordeling for (θ, ϕ) leder os til at slutte, at (1) er et 95% centralt posterior interval for θ?

11. Hvilken fraktil skal q være, hvis vi i stedet for ønsker, at (1) er et $(1-p)$%-konfidensinterval (eller $(1-p)$% centralt posterior interval) for θ?