Topics in statistics III/IV (MATH3361/4071) Michaelmas term, 2020-2021

Handout 6: Pivotal values, and their asymptotics, for C.I. & H.T.

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk
References: [1, (Ch. 4)], [2, (Ch. 4)]

Notation 1. Let $X, X_1, X_2, ..., X_n$ be a sequence of IID random variables (unseen observations) generated from a distribution f_θ labeled by a d-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^d$, and admitting PDF $f(\cdot|\theta)$.

Assumption 2. *Assume that the conditions from the Cramer Theorem 19 (Handout 4) are satisfied.*

Notation 3. Consider that $\hat{\theta}_n$ is the MLE of θ.

- Although the methods below use the MLE $\hat{\theta}_n$, in fact, any asymptotic equivalent estimator $\clubsuit_n$ of the MLE can be used; e.g., the one-step-estimators with moment estimator initial guess.

  Recall that: asymptotic equivalent $\clubsuit_n - \hat{\theta}_n \xrightarrow{P} 0 \implies$ asymptotic efficient $\sqrt{n}(\clubsuit_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, \mathcal{I}(\theta_0)^{-1})$

- All the tools below are asymptotic equivalent, as you can imagine. However, for smaller samples it seems that the Likelihood ration is the most powerful.

Note 4. We present 3 pivotal values for HT and CI: the Likelihood ratio, Wald, and Score statistics. Their rational is depicted in Figure 1.

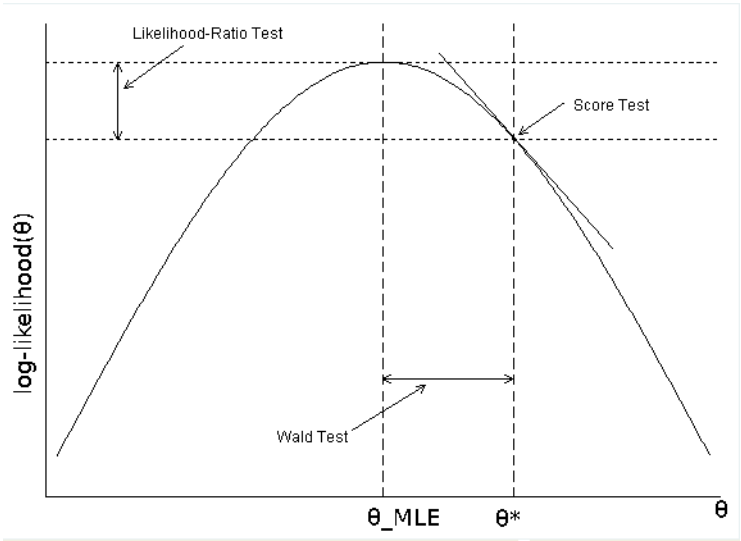

Figure 1: Assume that $\hat{\theta}$is the MLE and θ_0 is the real value of the parameter $\theta \in \mathbb{R}$. Comparison, in 1D , of:
Likelihood ratio : $W_{\mathrm{LR}}(\theta_0) = -2(\ell_n(\theta_0) - \ell_n(\hat{\theta}_n))$
Wald : $W_{\mathrm{Wald}}(\theta_0) = n(\theta_0 - \hat{\theta}_n)^2\mathcal{I}(\theta_0) = -(\theta_0 - \hat{\theta}_n)^2\mathrm{E}(\ddot{\ell}_n(\theta_0))$
Score statistic : $W_{\mathrm{Score}}(\theta_0) = n\dot{\ell}_n(\theta_0)/\mathcal{I}(\theta_0) = -\dot{\ell}_n(\theta_0)/\mathrm{E}(\ddot{\ell}_n(\theta_0))$

1 Likelihood ratio statistic as pivotal statistic

1.1 Likelihood ratio pivotal statistic

Definition 5. The log likelihood ratio statistic at $\theta \in \mathbb{R}^d$ is

$$W_{\mathrm{LR}}(\theta) = -2(\ell_n(\theta) - \ell_n(\hat{\theta}_n)) \tag{1}$$

Theorem 6. *Let $X_1, X_2, ..., X_n$ be independent random samples generated from a distribution f_θ labeled by a d-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^d$, and admitting PDF $f(\cdot|\theta)$. Assume the conditions from the Cramer Theorem 19 (Handout 4) are satisfied. Let θ_0 be the real value of θ. Then*

$$W_{LR}(\theta_0) = -2(\ell_n(\theta_0) - \ell_n(\hat{\theta}_n)) \xrightarrow{D} \chi^2_d \tag{2}$$

where $\hat{\theta}_n$ is the MLE of θ.

Proof. It is given as an Exercise 31 in Exercise sheet, as well as in the Appendix A. □

1.2 Likelihood ratio hypothesis test

Proposition 7. *Consider Hypothesis test*

$$H_0 : \theta = \theta_* \quad vs. \quad H_1 : \theta \neq \theta_*$$

The rejection area, at sig. level a, is

$$RA(X_{1:n}) = \{X_{1:n} : W_{LR}(\theta_*) \geq \chi^2_{d,1-a}\} = \{X_{1:n} : -2(\ell_n(\theta_*) - \ell_n(\hat{\theta}_n)) \geq \chi^2_{d,1-a} \tag{3}$$

1.3 Likelihood ratio confidence region

Proposition 8. *The $(1-a)$ confidence interval for θ is*

$$CI(\theta) = \{\theta \in \Theta : W_{LR}(\theta) \leq \chi^2_{d,1-a}\} = \{\theta \in \Theta : -2(\ell_n(\theta) - \ell_n(\hat{\theta}_n)) \leq \chi^2_{d,1-a}\} \tag{4}$$

as produced by inverting the $RA(x_{1:n})$

1.4 Comments

1. Regarding the HT, the comparison relies on the distance of the log-likelihood ratio $\ell_n(\theta_*)$ and $\ell_n(\hat{\theta}_n)$. The larger the distance is, the biggest doubt about the H_0 based on my data. See Figure 1.

2. It is more powerful than the other 2 tests, and hence preferable if it can by practically evaluated. The other 2 were derived possibly because ages ago people did not have computers and wanted to use something computationally cheaper.

2 Wald statistics as pivotal statistic

2.1 Wald pivotal statistic

Definition 9. The Wald statistic, is defined as

$$W_{\mathrm{W}}(\theta_0) = n(\hat{\theta}_n - \theta_0)^T \mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0) \tag{5}$$

Definition 10. Other, more tractable variations of the Wald statistic are

$$W'_{\mathrm{W}}(\theta_0) = n(\hat{\theta}_n - \theta_0)^T \mathcal{I}(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \tag{6}$$

$$W''_{\mathrm{W}}(\theta_0) = (\hat{\theta}_n - \theta_0)^T \mathcal{J}_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \tag{7}$$

Proposition 11. *Assume the conditions from the Cramer Theorem 19 (Handout 4) are satisfied. Let θ_0 be the real value of θ. Then $W_W(\theta_0)$, $W'_W(\theta_0)$, and $W''_W(\theta_0)$ are asymptotic equivalent, and*

$$W_W(\theta_0) = n(\hat{\theta}_n - \theta_0)^T \mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0) \xrightarrow{D} \chi^2_d \tag{8}$$

$$W'_W(\theta_0) = n(\hat{\theta}_n - \theta_0)^T \mathcal{I}(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \xrightarrow{D} \chi^2_d \tag{9}$$

$$W''_W(\theta_0) = (\hat{\theta}_n - \theta_0)^T \mathcal{J}_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \xrightarrow{D} \chi^2_d \tag{10}$$

Proof. Asymptotic equivalence is proved by the definition: E.g.,

$$W_{\mathrm{W}}(\theta_0) - W'_{\mathrm{W}}(\theta_0) = \sqrt{n}(\hat{\theta}_n - \theta_0)^T \left[\mathcal{I}(\theta_0) - \mathcal{I}(\hat{\theta}_n)\right] \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{p} 0$$

because $\mathcal{I}(\hat{\theta}_n) \xrightarrow{p} \mathcal{I}(\theta_0)$. Asymptotic distribution is proved as consequences of Cramer Theorem, Lemma 24 (Handout 4), and Slusky; E.g. for (8) because $\sqrt{n}\mathcal{I}(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, I)$ then

$$W_{\mathrm{W}}(\theta_0) = \left[\sqrt{n}\mathcal{I}(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0)\right]^\top \left[\sqrt{n}\mathcal{I}(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0)\right] \xrightarrow{D} Z^\top Z$$

where $Z = (z_1, ..., z_d) \sim \mathrm{N}(0, I)$ and hence $\sum_{i=1}^d z_i^2 \sim \chi^2_d$. For 9, and 10, I do the same by using

$$\sqrt{n}\mathcal{I}(\hat{\theta}_n)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, I), \quad \text{and} \quad \mathcal{J}_n(\hat{\theta}_n)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, I)$$

from Lemma 24 (Handout 4), and Slusky rules. □

Remark 12. (Wald) pivotal statistics in (8, 9, and 10) are asymptotically equivalent for large samples (this is obvious by construction). However, the order of preference is $W_{\mathrm{W}}(\theta_0)$, $W'_{\mathrm{W}}(\theta_0)$, $W''_{\mathrm{W}}(\theta_0)$ when the sample size is not that large; the proof is out of scope.

2.2 Wald hypothesis test

Proposition 13. *Consider Hypothesis test*

$$H_0 : \theta = \theta_* \quad vs. \quad H_1 : \theta \neq \theta_*$$

The rejection area, at sig. level a, is

$$RA(X_{1:n}) = \{X_{1:n} : W_{Wald}(\theta_0) \geq \chi^2_{d,1-a}\} = \{X_{1:n} : n(\hat{\theta}_n - \theta_*)^T \mathcal{I}(\theta_*)(\hat{\theta}_n - \theta_*) \geq \chi^2_{d,1-a}\} \tag{11}$$

2.3 Wald confidence sets

Proposition 14. *The* $(1-a)$ *confidence interval for* θ *is*

$$CI(\theta) = \{\theta \in \Theta : W_{Wald}(\theta) \leq \chi^2_{d,1-a}\} = \{\theta \in \Theta : n(\hat{\theta}_n - \theta)^T \mathcal{I}(\theta)(\hat{\theta}_n - \theta) \leq \chi^2_{d,1-a}\} \tag{12}$$

produced by inverting the $RA(x_{1:n})$

2.4 Comment

1. The Wald pivotal statistics are asymptotically equivalent to the LR one.
2. Regarding the HT, the comparison relies on the distance of the θ_* and $\hat{\theta}_n$, calibrated by the Information matrix (Fisher or Observed). The larger the distance is, the biggest doubt about the H_0 based on my data. See Figure 1.
3. Wald type of HT, CI are less expensive than the likelihood ratio ones because they require the computation of the expensive likelihood less number of times.

3 Score statistics as pivotal statistic

3.1 Score statistic

Definition. The Score statistic is defined as

$$U(\theta) = [\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\theta)]^T = [\sum_{i=1}^{d} \underbrace{\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_i|\theta)}_{=\Psi(X_i,\theta)}]^T \tag{13}$$

Proposition 15. *The asymptotic distribution is*

$$\frac{1}{\sqrt{n}}U(\theta) \xrightarrow{D} N(0, \mathcal{I}(\theta)) \tag{14}$$

which results as in Example/Proposition 16 (Handout 4).

Definition 16. The following score pivotal statistic is produced from the score statistic:

$$W_{\text{Score}}(\theta) = \frac{1}{n}U(\theta)^T \mathcal{I}(\theta)^{-1} U(\theta) \tag{15}$$

Definition 17. Other, more tractable variations of the score pivotal statistic are

$$W'_{\text{Score}}(\theta) = \frac{1}{n}U(\theta)^T \mathcal{I}(\hat{\theta}_n)^{-1} U(\theta) \tag{16}$$

$$W''_{\text{Score}}(\theta) = U(\theta)^T \mathcal{J}_n(\hat{\theta}_n)^{-1} U(\theta) \tag{17}$$

Proposition 18. *Assume the conditions from the Cramer Theorem 19 (Handout 4) are satisfied. Let θ_0 be the real value of θ. Then $W_{Score}(\theta_0)$, $W'_{Score}(\theta_0)$, and $W''_{Score}(\theta_0)$ are asymptotic equivalent, and*

$$W_{Score}(\theta_0) = \frac{1}{n} U(\theta_0)^T \mathcal{I}(\theta_0)^{-1} U(\theta_0) \xrightarrow{D} \chi^2_d \tag{18}$$

$$W'_{Score}(\theta_0) = \frac{1}{n} U(\theta_0)^T \mathcal{I}(\hat{\theta}_n)^{-1} U(\theta_0) \xrightarrow{D} \chi^2_d \tag{19}$$

$$W''_{Score}(\theta_0) = U(\theta_0)^T \mathcal{J}_n(\hat{\theta}_n)^{-1} U(\theta_0) \xrightarrow{D} \chi^2_d \tag{20}$$

Proof. The asymptotic equivalent is proven as in Proposition 11. Regarding the asymptotic distributions: by using (14) with Slusky rules, it can be derived

$$\frac{1}{\sqrt{n}} \mathcal{I}(\theta)^{-1/2} U(\theta) \xrightarrow{D} \mathrm{N}(0, I), \ ; \ \frac{1}{\sqrt{n}} \mathcal{I}(\hat{\theta}_n)^{-1/2} U(\theta) \xrightarrow{D} \mathrm{N}(0, I), \text{ and } \mathcal{J}_n(\hat{\theta}_n)^{-1/2} U(\theta) \xrightarrow{D} \mathrm{N}(0, I)$$

and hence 18, 19, and (20) can be derived by applied Slusky rules again. □

3.2 Score hypothesis test

Proposition 19. *Consider Hypothesis test*

$$H_0 : \theta = \theta_* \quad vs. \quad H_1 : \theta \neq \theta_*$$

Hence the rejection area, at sig. level a, is

$$RA(X_{1:n}) = \{X_{1:n} : W_{Score}(\theta_0) \geq \chi^2_{d,1-a}\} = \{X_{1:n} : \frac{1}{n} U(\theta_*)^T \mathcal{I}(\theta_*)^{-1} U(\theta_*) \geq \chi^2_{d,1-a}\} \tag{21}$$

3.3 Score confidence intervals

Proposition 20. *The $(1-a)$ confidence interval for θ is*

$$CI(\theta) = \{\theta \in \Theta : W_{Score}(\theta) \leq \chi^2_{d,1-a}\} = \{\theta \in \Theta : \frac{1}{n} U(\theta)^T \mathcal{I}(\theta)^{-1} U(\theta) \leq \chi^2_{d,1-a}\} \tag{22}$$

etc... produced by inverting the $RA(X_{1:n})$.

3.4 Comments

1. The Score pivotal statistics are asymptotically equivalent to the LR one.
2. Regarding the HT, the comparison relies on the slop of the log-likelihood at θ_* (aka the $U(\theta_*)$) calibrated by the curvature (Hessian matrix) at θ_*. The larger/steeper the slope, the bigger the distance from the peak (MLE $\hat{\theta}_n$), hence the biggest doubt about the H_0 based on my data. See Figure 1.
3. Score type of HT, CI are less expensive than the likelihood ratio ones because they require the computation of the expensive likelihood less number of times.

4. Score HT and CI are computational convenient, in situations when the practitioner wants to calculate the HT or CI for parameter $\phi \in \mathbb{R}^k$, which is a function of parameter $\theta \in \mathbb{R}^d$ whose Score type HT or CI have already been calculated.

   Let $\phi := \phi(\theta)$. The score statistic

$$U^*(\phi) = [\frac{\mathrm{d}}{\mathrm{d}\phi}\ell_n(\theta)]^T = [\sum_{i=1}^{d}\frac{\mathrm{d}}{\mathrm{d}\phi}f(X_i|\phi)]^T$$

   is such that

$$U^*(\phi) = [\frac{\mathrm{d}}{\mathrm{d}\phi}\ell_n(\phi)]^T = [\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\theta)\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]^T = [\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]^T U(\theta)$$

   It has expected value

$$\mathrm{E}(U^*(\phi)) = \mathrm{E}([\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]^T U(\theta)) = 0$$

   it has variance

$$\mathrm{var}(U^*(\phi)) = \mathrm{var}([\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]^T U(\theta)) = [\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]^T \mathcal{I}(\theta)[\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]$$

   and hence, by Delta method it has asymptotic distribution

$$\frac{1}{\sqrt{n}}U^*(\phi) \xrightarrow{D} \mathrm{N}(0, \underbrace{[\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]^T \mathcal{I}(\theta)[\frac{\mathrm{d}\theta}{\mathrm{d}\phi}]}_{=\mathcal{I}^*(\phi)})$$

   Hence, one can derive HT and CI as in (21) and (22) by substituting properly [1]. Notice that if the score HT and CI for θ are available then the score HT and CI for the transformation $\phi = \phi(\theta)$ can by computed by avoiding to recompute the expensive likelihood function.

Example 21. Let random sample $x_1, ..., x_n \overset{IID}{\sim} \mathrm{Poi}(\theta)$, $\theta > 0$ with PDF

$$f(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}1(x > 0)$$

For the hypothesis test

$$\mathrm{H}_0 : \theta = \theta_* \quad \text{vs.} \quad \mathrm{H}_1 : \theta \neq \theta_*$$

Calculate

1. the log-likelihood ratio RA at a sig. level
2. the Wald's type RA at a sig. level (the 3 of them)
3. the Score's type RA at a sig. level (the 3 of them)

[1] please write down the derivation

Solution. Ok before that, let's calculate all the quantities required.

$$\begin{aligned}
\log f(x|\theta) \propto & x\log(\theta)-\theta && ;; && 0=\dot{\ell}(\theta)|_{\theta=\hat{\theta}} \implies \hat{\theta}=\bar{x}\\
\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x|\theta) = & \frac{x}{\theta}-1 && ;; && \ddot{\ell}(\theta) = -n\bar{x}\frac{1}{\theta^2}\\
\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x|\theta) = & -\frac{x}{\theta^2} && ;; && \mathcal{J}_n(\theta) = -\ddot{\ell}(\theta) = n\bar{x}\frac{1}{\theta^2}\\
\mathcal{I}(\theta) = & -\mathrm{E}(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x|\theta)) = \frac{1}{\theta} && ;; && \mathcal{J}_n(\hat{\theta}) = \frac{n}{\bar{x}}\\
\ell(\theta) = & n\bar{x}\log(\theta)-n\theta && ;; && U(\theta)=\dot{\ell}(\theta)^T = n\bar{x}\frac{1}{\theta}-n\\
\dot{\ell}(\theta) = & n\bar{x}\frac{1}{\theta}-n && ;;
\end{aligned}$$

1. It is

$$\begin{aligned}
\mathrm{CI}(\theta) &= \{\theta\in(0,\infty) : -2(\ell_n(\theta)-\ell_n(\hat{\theta}_n))\le\chi^2_{1,1-a}\}\\
&= \{\theta\in(0,\infty) : -2((n\bar{x}\log(\frac{\theta}{\bar{x}})-n(\theta-\bar{x}))\le\chi^2_{1,1-a}\}
\end{aligned}$$

   well, here we do not have the condition $n\theta - n\bar{x}$ like in the contingency tables ...

2. It is

$$\begin{aligned}
\mathrm{RA}(x_{1:n}) &= \{x_{1:n} : n(\hat{\theta}_n-\theta_*)^T\mathcal{I}(\theta_*)(\hat{\theta}_n-\theta_*)\ge\chi^2_{1,1-a}\}\\
&= \{x_{1:n} : n(\bar{x}-\theta_*)^2\frac{1}{\theta_*}\ge\chi^2_{1,1-a}\}
\end{aligned}$$

   and

$$\begin{aligned}
\mathrm{RA}(x_{1:n}) &= \{x_{1:n} : n(\hat{\theta}_n-\theta_*)^T\mathcal{I}(\hat{\theta}_n)(\hat{\theta}_n-\theta_*)\ge\chi^2_{1,1-a}\}\\
&= \{x_{1:n} : n(\bar{x}-\theta_*)^2\frac{1}{\bar{x}}\ge\chi^2_{1,1-a}\}
\end{aligned}$$

   and

$$\begin{aligned}
\mathrm{RA}(x_{1:n}) &= \{x_{1:n} : (\hat{\theta}_n-\theta_*)^T\mathcal{J}_n(\hat{\theta}_n)(\hat{\theta}_n-\theta_*)\ge\chi^2_{1,1-a}\}\\
&= \{x_{1:n} : (\bar{x}-\theta_*)^2\frac{n}{\bar{x}}\ge\chi^2_{1,1-a}\}
\end{aligned}$$

   where the latter two CI coinside, however this is just a coinsidence...

3. It is

$$\begin{aligned}
\mathrm{RA}(x_{1:n}) &= \{x_{1:n} : \frac{1}{n}U(\theta_*)^T\mathcal{I}(\theta_*)^{-1}U(\theta_*)\ge\chi^2_{1,1-a}\}\\
&= \{x_{1:n} : \frac{1}{n}(n\bar{x}\frac{1}{\theta_*}-n)^2(\frac{1}{\theta_*})^{-1}\ge\chi^2_{1,1-a}\}\\
&= \{x_{1:n} : n(\bar{x}\frac{1}{\theta_*}-1)^2\theta_*\ge\chi^2_{1,1-a}\}
\end{aligned}$$

and

$$\begin{aligned}\mathrm{RA}(x_{1:n}) &= \{x_{1:n} : \frac{1}{n}U(\theta_*)^T\mathcal{I}(\hat{\theta}_n)^{-1}U(\theta_*) \geq \chi^2_{1,1-a}\} \\ &= \{x_{1:n} : n(\bar{x}\frac{1}{\theta_*} - 1)^2\hat{\theta}_n \geq \chi^2_{1,1-a}\}\end{aligned}$$

and

$$\begin{aligned}\mathrm{RA}(x_{1:n}) &= \{x_{1:n} : \frac{1}{n}U(\theta_*)^T\mathcal{J}_n(\hat{\theta}_n)^{-1}U(\theta_*) \geq \chi^2_{1,1-a}\} \\ &= \{x_{1:n} : n(\bar{x}\frac{1}{\theta_*} - 1)^2\hat{\theta}_n \geq \chi^2_{1,1-a}\}\end{aligned}$$

where the latter two CI coincide, however it is just a coincidence...

Confidence intervals can be computed by inverting the RA, based on the theory learnt in Concepts in Stats 2 (Term 1).

Example 22. Consider an M way contingency table $(n_{i,j})$ generated by a Poisson sampling scheme. Consider it is modeled by a log-Linear model with link function

$$\log(\mu) = X^T\beta \tag{23}$$

in the vectorized form, where vector $\beta \in \mathbb{R}^d$ contains the unknown coefficients. Consider that identifiability contains have be considered in (23). Show that

1. the asymptotic distribution of the MLEs $\hat{\beta}_n$ is such that

$$(X\mathrm{diag}(\mu_n)X^T)^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{D} \mathrm{N}(0, I) \tag{24}$$

   where $\hat{\mu}_n = \exp(X^T\hat{\beta}_n)$, and β_0 is the true value of β.

2. An $(1-a)100\%$ asymptotic confidence interval for β_0 is

$$\{\beta_0 \,:\, (\hat{\beta}_n - \beta_0)^T(X\mathrm{diag}(\hat{\mu}_n)X^T)(\hat{\beta}_n - \beta_0) \leq \chi^2_{d,1-a}\}$$

Solution.

1. Since $\hat{\beta}_n$ is an MLE then it is asymptotically Normal as

$$\mathcal{J}_n(\hat{\theta}_n)^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{D} \mathrm{N}(0, I) \tag{25}$$

   It is

$$\mathcal{J}_n = -\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\ell_n(\beta)|_{\beta=\hat{\beta}_n} = -\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\sum_{i=1}^{n}\log(\mathrm{Poi}(n|\mu(\beta)))|_{\beta=\hat{\beta}_n}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\beta_j}\ell_n(\underset{\sim}{\mu}(\beta)) &= -\sum_i n_i X_{i,j} + \sum_i \exp(\sum_j X_{i,j}\beta_j)X_{i,j} \\ \frac{\mathrm{d}^2}{\mathrm{d}\beta_j\mathrm{d}\beta_k}\ell_n(\underset{\sim}{\mu}(\beta)) &= \sum_i \exp(\sum_j X_{i,j}\beta_j)X_{i,j}X_{i,k}\end{aligned}$$

and

$$\mathcal{J}_n = X^T \text{diag}(\hat{\mu}_n) X$$

Therefore from (25)

$$\underbrace{(X^T \text{diag}(\hat{\mu}_n) X)^{1/2} (\hat{\beta}_n - \beta_0)}_{=Z_n} \xrightarrow{D} \text{N}(0, I)$$

2. I use the statistic

$$T_n = (\hat{\beta}_n - \beta_0)^T (X^T \text{diag}(\hat{\mu}_n) X)(\hat{\beta}_n - \beta_0) = Z_n^T Z_n = \sum_{i=1}^{d} Z_{n,i}$$

where $T_n \sim \chi^2_d$ as summation of d standard normal variables. Then

$$1 - a = P_{\chi^2_d}(T_n < q)$$

where $q = \chi^2_{d,1-a}$.

Exercise sheet

Exercise #31, 32, 33, 34, 36

References

[1] Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

[2] T. A. Severini. *Likelihood methods in statistics*. Oxford University Press, 2000.

Appendix

A Appendix

Sketch of the derivation of the likelihood ratio statistic asymptotic distribution in Theorem 6. For more details see the Exercise sheet.

Proof. Let's expand it,

$$\begin{aligned}\ell_n(\theta_0) &= \ell_n(\hat{\theta}_n) + \dot{\ell}_n(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n) + (\theta_0 - \hat{\theta}_n)^T \int_0^1 \int_0^1 u\, \ddot{\ell}_n(\hat{\theta}_n + uv(\theta_0 - \hat{\theta}_n))\mathrm{d}u\mathrm{d}v\ (\theta_0 - \hat{\theta}_n) \\ &= \ell_n(\hat{\theta}_n) + \dot{\ell}_n(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n) + (\theta_0 - \hat{\theta}_n)^T\, n \int_0^1 \int_0^1 u\, \frac{1}{n}\ddot{\ell}_n(\hat{\theta}_n + uv(\theta_0 - \hat{\theta}_n))\mathrm{d}u\mathrm{d}v\ (\theta_0 - \hat{\theta}_n)\end{aligned}$$

Rearranging the terms

$$-2(\ell_n(\theta_0) - \ell_n(\hat{\theta}_n)) = -\underbrace{\dot{\ell}_n(\hat{\theta}_n)}_{=0}(\theta_0 - \hat{\theta}_n) - n(\theta_0 - \hat{\theta}_n)^T \underbrace{\int_0^1 \int_0^1 u\, \frac{1}{n}\ddot{\ell}_n(\hat{\theta}_n + uv(\theta_0 - \hat{\theta}_n))\mathrm{d}u\mathrm{d}v}_{\xrightarrow{a.s} -\frac{1}{2}\mathcal{I}(\theta_0)}\ (\theta_0 - \hat{\theta}_n)$$

It is $\dot{\ell}_n(\hat{\theta}_n) = 0$ because $\hat{\theta}_n$ is an MLE.

It is

$$\int_0^1 \int_0^1 u\, \frac{1}{n}\ddot{\ell}_n(\hat{\theta}_n + uv(\theta_0 - \hat{\theta}_n))\mathrm{d}u\mathrm{d}v \xrightarrow{a.s} -\frac{1}{2}\mathcal{I}(\theta_0)$$

by using USLLN properly as in the proof of Cramer theorem for MLE asymptotic distiribution. □

So to sum up

$$-2(\ell_n(\theta_0) - \ell_n(\hat{\theta}_n)) \xrightarrow{a.s} n(\theta_0 - \hat{\theta}_n)^T \mathcal{I}(\theta_0)(\theta_0 - \hat{\theta}_n)$$

which implies that

$$\Big[-2(\ell_n(\theta_0) - \ell_n(\hat{\theta}_n))\Big] - \Big[n(\theta_0 - \hat{\theta}_n)^T \mathcal{I}(\theta_0)(\theta_0 - \hat{\theta}_n)\Big] \xrightarrow{p} 0 \tag{26}$$

From Cramer' Theorem I know that

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) &\xrightarrow{D} \mathrm{N}(0, \mathcal{I}(\theta_0)^{-1}) \\ \implies n(\hat{\theta}_n - \theta_0)^T \mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0) &\xrightarrow{D} \chi^2_d\end{aligned}$$

But $-2(\ell_n(\theta_0) - \ell_n(\hat{\theta}_n))$ is asymptotic equivalent to $n(\hat{\theta}_n - \theta_0)^T \mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0)$ from 26. So by the Slutsky's theorem

$$-2(\ell_n(\theta_0) - \ell_n(\hat{\theta}_n)) \xrightarrow{D} \chi^2_d$$

B Repeatation of results from Handout 4.

Fact 23. *Cholesky decomposition: Every symmetric, positive definite matrix Σ can be decomposed into a product of a unique lower triangular matrix L and its transpose, i.e. $\Sigma = LL^T$.*

Lemma 24. *(Which will be used as a Proposition later on) Show that given that the assumptions [C.1-C.5] of Theorem 19 are satisfied, and that $\mathcal{I}(\theta)$ and $\mathcal{J}_n(\theta)$ are continuous on θ, then*

$$\sqrt{n}\mathcal{I}(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, I) \tag{27}$$

$$\sqrt{n}\mathcal{I}(\hat{\theta}_n)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, I) \tag{28}$$

$$\mathcal{J}_n(\hat{\theta}_n)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, I) \tag{29}$$

where $A^{1/2}$ denotes the lower triangular matrix of the Cholesky decomposition of A; i.e., $A = A^{1/2}(A^{1/2})^T$.

Solution.

- Eq 27 results from Cramer Theorem, and the properties of covariance matrix.
- Eq. 28 results by using Cramer Theorem and Slutsky theorems. Precisely, because $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$, Slutsky implies $\mathcal{I}(\hat{\theta}_n) \xrightarrow{a.s.} \mathcal{I}(\theta_0)$ which implies $\mathcal{I}(\hat{\theta}_n)^{1/2}\mathcal{I}(\theta_0)^{-1/2} \xrightarrow{a.s.} I$. Therefore, by Slutsky

$$\underbrace{\mathcal{I}(\hat{\theta}_n)^{1/2}\mathcal{I}(\theta_0)^{-1/2}\sqrt{n}\mathcal{I}(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0)}_{=\sqrt{n}\mathcal{I}(\hat{\theta}_n)^{1/2}(\hat{\theta}_n-\theta_0)} \xrightarrow{D} \underbrace{I \times \mathrm{N}(0, I)}_{=\mathrm{N}(0,I)}$$

- Eq. 29 results by using the USLLN and Slutsly theorems. So I just need to show that

$$\frac{1}{n}\mathcal{J}_n(\hat{\theta}_n) \xrightarrow{a.s.} \mathcal{I}(\theta_0)$$

Set $U(x, \theta) = -\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log(f(x|\theta))$, and $\mathcal{I}(\theta) = \mathrm{E}(U(x, \theta))$. Then

$$|\frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log(f(x_i|\hat{\theta}_n))\right)}_{U(x_i,\hat{\theta}_n)} - \mathcal{I}(\theta_0)| \leq |\frac{1}{n}\sum_{i=1}^{n}U(x_i, \hat{\theta}_n) - \mathcal{I}(\hat{\theta}_n)| + |\mathcal{I}(\hat{\theta}_n) - \mathcal{I}(\theta_0)| \tag{30}$$

$$\leq \sup_{|\hat{\theta}_n - \theta_0| \leq \delta} |\frac{1}{n}\sum_{i=1}^{n}U(x_i, \theta) - \mathcal{I}(\theta)| + |\mathcal{I}(\hat{\theta}_n) - \mathcal{I}(\theta_0)| \tag{31}$$

The first term converges to zero because the assumptions of the USLLN are satisfied. The second term converges to zero because $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$ and hence $\mathcal{I}(\hat{\theta}_n) \xrightarrow{a.s.} \mathcal{I}(\theta_0)$ by using Slutsky theorem.

So by Slutsy $(\frac{1}{n}\mathcal{J}_n(\hat{\theta}_n))^{1/2}\mathcal{I}(\theta_0)^{-1/2} \xrightarrow{a.s.} I$, and by Slutsky again

$$\underbrace{(\frac{1}{n}\mathcal{J}_n(\hat{\theta}_n))^{1/2}\mathcal{I}(\theta_0)^{-1/2}I(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0)}_{=\mathcal{J}_n(\hat{\theta}_n)^{1/2}(\hat{\theta}_n-\theta_0)} \xrightarrow{D} \underbrace{I \times \mathrm{N}(0, I)}_{=\mathrm{N}(0,I)}$$