STA 4508: Topics in Likelihood Inference Fall 2018

SS 2010 Tuesdays 10-1. First class October 26.

Topics

1. Inference based on the likelihood function: derived quantities, limiting distributions, approximations to posterior distributions;
2. Likelihood for semi-parametric and non-parametric models: proportional hazards regression, partially linear models, penalized likelihood;
3. Composite likelihood: definition, summary statistics, asymptotic theory; applications
4. Likelihood inference for p > n;
5. Simulated likelihoods, indirect inference and approximate Bayesian computation

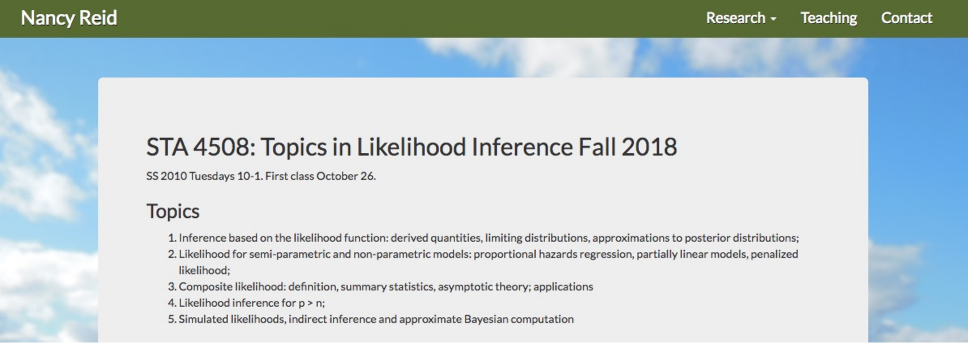
Running list of references and background reading

Review Papers

- Reid, N. (2013) Aspects of likelihood inference *Bernoulli* 19, 1404-1418.
- Reid, N. (2010) Likelihood Inference *Wiley Interdisciplinary Reviews in Computational Statistics* 5, 517-525.
  (I need to use Preview to view this, rather than Adobe.)
- Reid, N. (2011) Likelihood *International Encyclopedia of Statistical Science*, Part 5, 455-459.
- Reid, N. (2000) Likelihood. *J. Am. Stat. Assoc.*, 95, 1335-1340.

Likelihood Basics

- Davison, A.C. (2003) *Statistical Models* (SM) Cambridge University Press. -- Ch 4
- Barndorff-Nielsen, O.E. and Cox, D.R. (1994) *Inference and Asymptotics* (BNC) Chapman and Hall. -- Ch 2.2
- Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics* (CH) Chapman and Hall. -- Ch 2.1 (i), (ii)
- Cox, D.R. (2006) *Principles of Statistical Inference* (Cox) -- Ch.2.1


Nancy Reid   Research   Teaching   Contact

STA 4508: Topics in Likelihood Inference Fall 2018

SS 2010 Tuesdays 10-1. First class October 26.

Topics

1. Inference based on the likelihood function: derived quantities, limiting distributions, approximations to posterior distributions;
2. Likelihood for semi-parametric and non-parametric models: proportional hazards regression, partially linear models, penalized likelihood;
3. Composite likelihood: definition, summary statistics, asymptotic theory; applications
4. Likelihood inference for p > n;
5. Simulated likelihoods, indirect inference and approximate Bayesian computation

Nancy Reid SS 6002A
`reid@utstat.utoronto.ca`
Office Hours by appointment

Problems assigned weekly, due the following week
`http://www.utstat.utoronto.ca/reid/html/sta/sta4508.html`

Various ‘types’ of likelihood

1. likelihood, marginal likelihood, conditional likelihood, profile likelihood, adjusted profile likelihood
2. semi-parametric likelihood, partial likelihood
3. empirical likelihood, penalized likelihood
4. quasi-likelihood, composite likelihood
5. simulated likelihood, indirect inference
6. bootstrap likelihood, *h*-likelihood, weighted likelihood, pseudo-likelihood, local likelihood, sieve likelihood

Why so many?

- Principle: "The probability model and the choice of [parameter] serve to translate a subject-matter question into a mathematical and statistical one" Cox, 2006, p.3

- likelihood function is proportional to the probability model

- inference based on the likelihood function is widely accepted
- provides more than point estimate or test of point hypothesis

- models needed for applications are more and more complex

- need some analogues to the likelihood function for these complex settings

The likelihood function

- Parametric model: $f(y;\theta), \quad y \in \mathcal{Y}, \theta \in \Theta \subset \mathbb{R}^p$
- Likelihood function

$$L(\theta; y) = f(y;\theta), \quad \text{or} \quad L(\theta; y) = c(y) f(y;\theta), \quad \text{or} \quad L(\theta; y) \propto f(y;\theta)$$

- typically, $y = (y_1, \ldots, y_n)$ $\quad x_1, \ldots, x_n \quad i = 1, \ldots, n$
- $f(y;\theta)$ or $f(y \mid x;\theta)$ is joint density
- under independence $L(\theta; y) \propto \prod f(y_i \mid x_i; \theta)$
- log-likelihood $\ell(\theta; y) = \log L(\theta; y) = \sum \log f(y_i \mid x_i; \theta)$
- θ could have dimension $p > n$ (e.g. genetics), or $d \uparrow n$, or
- θ could have infinite dimension e.g.
- regular model $p < n$ and p fixed as n increases

Examples

- $y_i \sim N(\mu, \sigma^2)$:

$$L(\theta; y) = \prod_{i=1}^{n} \sigma^{-n} \exp\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\}$$

- $E(y_i) = x_i^T \beta$:

$$L(\theta; y) = \prod_{i=1}^{n} \sigma^{-n} \exp\{-\frac{1}{2\sigma^2}(y_i - x_i^T \beta)^2\}$$

- $E(y_i) = m(x_i), \quad m(x) = \Sigma_{j=1}^{J} \phi_j B_j(x)$:

$$L(\theta; y) = \prod_{i=1}^{n} \sigma^{-n} \exp\{-\frac{1}{2\sigma^2}(y_i - \Sigma_{j=1}^{J} \phi_j B_j(x_i))^2\}$$

... examples

- $y_i = \mu + \rho(y_{i-1} - \mu) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$:

$$L(\theta; y) = \prod_{i=1}^{n} f(y_i \mid y_{i-1}; \theta) f_0(y_0; \theta)$$

- $y_1, \dots, y_n$ i.i.d. observations from a $U(0, \theta)$ distribution:

$$L(\theta; y) = \prod_{i=1}^{n} \theta^{-n}, \quad 0 < y_{(1)} < \cdots < y_{(n)} < \theta$$

… examples

- $y_1, \ldots, y_n$ are the times of jumps of a non-homogeneous Poisson process with rate function $\lambda(\cdot)$:

$$\ell\{\lambda(\cdot); y\} = \sum_{i=1}^{n} \log\{\lambda(y_i)\} - \int_0^{\tau} \lambda(u) du, \quad 0 < y_1 < \cdots < y_n < \tau$$

Davison, §6.5

- multinomial: $y_i = (y_{i1}, \ldots, y_{ik}), \quad y_{ic} = 1, y_{ic'} = 0, c' \neq c$

$$\ell(\theta; y) = \sum_{i=1}^{n} \sum_{c=1}^{k} y_{ic} \log(p_{ic})$$

negative cross-entropy

Hastie et al., Ch. 7

$p_{ic} = p(x_{ic}; \theta)$, as above

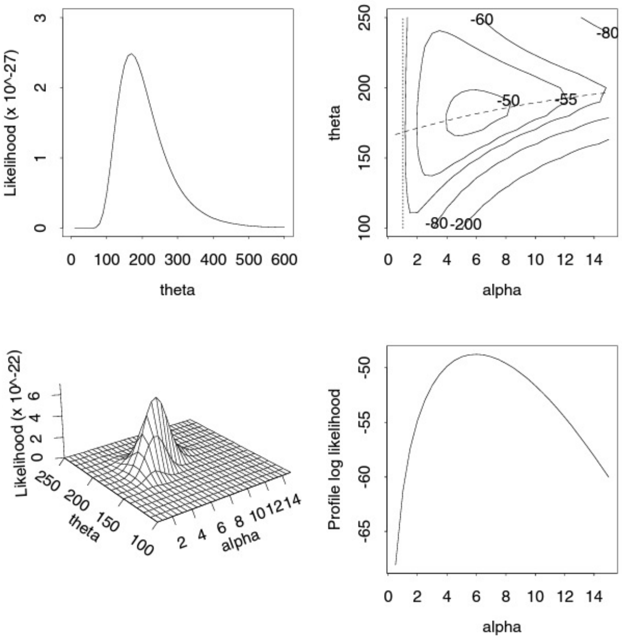


Figure 4.1 Likelihoods for the spring failure data at stress 950 N/mm^2. The upper left panel is the likelihood for the exponential model, and below it is a perspective plot of the likelihood for the Weibull model. The upper right panel shows contours of the log likelihood for the Weibull model; the exponential likelihood is obtained by setting $\alpha = 1$. that is, slicing L along the vertical dotted line. The lower right panel shows the profile log likelihood for α, which corresponds to the log likelihood values along the dashed line in the panel above, plotted against α.

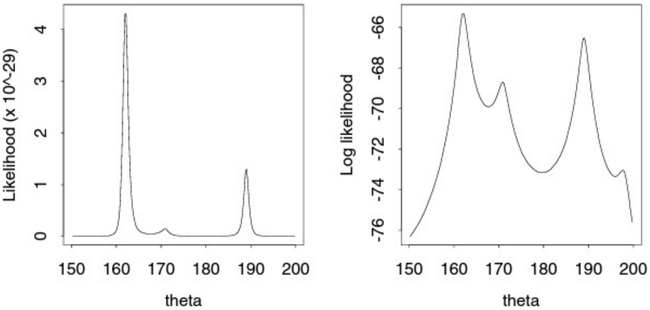


Figure 4.2 Cauchy likelihood and log likelihood for the spring failure data at stress 950N/mm^2.

Data: times of failure of a spring under stress

225, 171, 198, 189, 189, 135, 162, 135, 117, 162

Complicated likelihoods

- example: clustered binary data — Renard et al. (2004)
- latent variable: $z_{ir} = x_{ir}'\beta + b_i + \epsilon_{ir}, \quad b_i \sim N(0, \sigma_b^2), \quad \epsilon_{ir} \sim N(0,1)$
- $r = 1, \dots, n_i$: observations in a cluster/family/school...
  $i = 1, \dots, n$ clusters
- random effect b_i introduces correlation between observations in a cluster
- observations: $y_{ir} = 1$ if $z_{ir} > 0$, else 0

- $Pr(y_{ir} = 1 \mid b_i) = \Phi(x_{ir}'\beta + b_i) = p_i$ — $\Phi(z) = \int^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$

$$L(\theta; y) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{r=1}^{n_i} p_i^{y_{ir}} (1 - p_i)^{1-y_{ir}} \phi(b_i, \sigma_b^2) db_i$$

- more general: $z_{ir} = x_{ir}'\beta + w_{ir}'b_i + \epsilon_{ir}$

... complicated likelihoods

- generalized linear geostatistical models

$$E\{Y(s) \mid u(s)\} = g\{x(s)^T\beta + u(s)\}, \quad s \in \mathcal{S} \subset \mathbb{R}^d, d \geq 2$$

Diggle & Ribeiro, 2007

- random intercept u is a realization of a stationary GRF, mean 0, covariance

$$\text{cov}\{u(s), u(s')\} = \sigma^2\rho(s - s'; \alpha)$$

- n observed locations $y = (y_1, \ldots, y_n)$ with $y_i = y(s_i)$
- likelihood function

$$L(\theta; y) = \int_{\mathbb{R}^n} \prod_{i=1}^{n} f(y_i \mid u_i; \theta) \underbrace{f(u; \theta)}_{MVN(0,\Sigma)} du_1 \ldots du_n$$

- no factorization into lower dimensional integrals, as with previous example

Non-computable likelihoods

- Ising model:
$$f(y;\theta) = \exp\Big(\sum_{(i,j)\in E} \theta_{ij} y_i y_j\Big)\frac{1}{Z(\theta)}$$
- $y_i = \pm 1$; binary property of a node i in a graph with n nodes
- θ_{ij} measures strength of interaction between nodes i and j
- E is the set of edges between nodes

- partition function $Z(\theta) = \sum_y \exp(\sum_{(i,j)\in E} \theta_{ij} y_i y_j)$

Davison §6.2

Ravikumar et al. (2010).

High-dimensional Ising model selection... *Ann. Statist.* p.1287

History

IX. *On the Mathematical Foundations of Theoretical Statistics.*

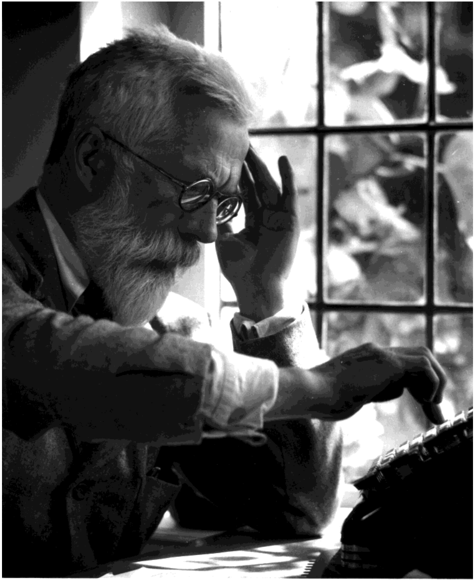
By R. A. FISHER, M.A., *Fellow of Gonville and Caius College, Cambridge, Chief Statistician, Rothamsted Experimental Station, Harpenden.*

Communicated by DR. E. J. RUSSELL, F.R.S.

Received June 25,—Read November 17, 1921.

CONTENTS.

History



History

know nothing whatever. We must return to the actual fact that one value of p, of the frequency of which we know nothing, would yield the observed result three times as frequently as would another value of p. If we need a word to characterise this relative property of different values of p, I suggest that we may speak without confusion of the *likelihood* of one value of p being thrice the likelihood of another, bearing always in mind that likelihood is not here used loosely as a synonym of probability, but simply to express the relative frequencies with which such values of the hypothetical quantity p would in fact yield the observed sample.

Why likelihood?

- makes probability modelling central
- emphasizes the inverse problem of reasoning from y^o to θ or $f(\cdot)$
- suggested by Fisher as a measure of plausibility

Royall, 1994

$L(\hat\theta)/L(\theta) \in (1,3)$	very plausible;
$L(\hat\theta)/L(\theta) \in (3,10)$	implausible;
$L(\hat\theta)/L(\theta) \in (10,\infty)$	very implausible

- converts a 'prior' probability $\pi(\theta)$ to a posterior $\pi(\theta \mid y)$ via Bayes' formula
- provides a conventional set of summary quantities for inference based on properties of the postulated model

… why likelihood?

- likelihood function depends on data only through sufficient statistics
- "likelihood map is sufficient" Fraser & Naderi, 2006

- gives exact inference in transformation models
- "likelihood function as pivotal" Hinkley, 1980

- provides summary statistics with known limiting distribution
- leading to approximate pivotal functions,
  based on normal distribution

- likelihood function + sample space derivative gives better approximate inference

Likelihood inference

- direct use of likelihood function
- note that only relative values are well-defined
- define relative likelihood $RL(\theta) = \frac{L(\theta)}{\sup_{\theta'} L(\theta')} = \frac{L(\theta)}{L(\hat{\theta})}$

$$
\begin{aligned}
1 \geq RL(\theta) &> \tfrac{1}{3}, && \theta \text{ strongly supported,} \\
\tfrac{1}{3} \geq RL(\theta) &> \tfrac{1}{10}, && \theta \text{ supported,} \\
\tfrac{1}{10} \geq RL(\theta) &> \tfrac{1}{100}, && \theta \text{ weakly supported,} \\
\tfrac{1}{100} \geq RL(\theta) &> \tfrac{1}{1000}, && \theta \text{ poorly supported,} \\
\tfrac{1}{1000} \geq RL(\theta) &> 0, && \theta \text{ very poorly supported.}
\end{aligned}
$$

SM (4.11)

… likelihood inference

- combine with a probability density for θ
- $$\pi(\theta \mid y) = \frac{f(y;\theta)\pi(\theta)}{\int f(y;\theta)\pi(\theta)d\theta}$$
- inference for θ via probability statements from $\pi(\theta \mid y)$
- e.g., "Probability $(\theta > 0 \mid y) = 0.23$", etc.
- any other use of likelihood function for inference relies on derived quantities and their distribution under the model
- the Likelihood Principle states two experiments with proportional likelihood functions lead to the same inference about the same parameter C& H, 1974, p.39 (strong likelihood)

Derived quantities, single observation

observed likelihood	$L(\theta; y) = c(y) f(y; \theta)$
log-likelihood	$\ell(\theta; y) = \log L(\theta; y) = \log f(y; \theta) + a(y)$
score	$U(\theta) = \partial \ell(\theta; y)/\partial\theta$
observed information	$j(\theta) = -\partial^2 \ell(\theta; y)/\partial\theta\partial\theta^T$
expected information	$i(\theta) = \mathrm{E}_\theta U(\theta) U(\theta)^T$ called $i_1(\theta)$ in CH

... derived quantities, i.i.d. sample

observed likelihood	$L(\theta; y) \propto \prod_{i=1}^n f(y_i; \theta)$
log-likelihood	$\ell(\theta; y) = \sum_{i=1}^n \log f(y; \theta) + a(y)$
score	$U(\theta) = \partial \ell(\theta; y)/\partial\theta = O_p(\sqrt{n})$
maximum likelihood estimate	$\hat\theta = \hat\theta(y) = \arg\sup_\theta \ell(\theta; y)$
Fisher information	$j(\hat\theta) = -\partial^2 \ell(\hat\theta; y)/\partial\theta\partial\theta^T = O_p(n)$
expected information	$i(\theta) = \mathrm{E}_\theta U(\theta)U(\theta)^T = O(n)$

Bartlett identities

$$
\begin{aligned}
1 &= \int f(y;\theta)dy \text{ endpoints not specified} \\
0 &= \frac{\partial}{\partial\theta}\int f(y;\theta)dy \\
&= \int \frac{\partial}{\partial\theta} f(y;\theta)dy \text{ but can't involve } \theta \\
&= \int \frac{\partial}{\partial\theta}\ell(\theta;y)f(y;\theta)dy = \mathrm{E}_\theta\{U(\theta;Y)\}
\end{aligned}
$$

$$
\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\int \frac{\partial}{\partial\theta}\ell(\theta;y)f(y;\theta)dy \\
&= \int [\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\ell(\theta;y) + \{\frac{\partial}{\partial\theta}\ell(\theta;y)\}\{\frac{\partial}{\partial\theta}\ell(\theta;y)\}^{\mathrm{T}}]f(y;\theta)dy \\
\Rightarrow \mathrm{E}_\theta\{U(\theta)U^{\mathrm{T}}(\theta)\} &= \mathrm{E}_\theta\{-\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\ell(\theta;y)\} \\
i(\theta) &= \mathrm{E}_\theta\{j(\theta)\}
\end{aligned}
$$

... Bartlett identities

You can keep going, as long as the endpoints don't depend on θ, the log-density is differentiable, and the required moments exist.

From the book *Tensor Methods* by McCullagh:

sample space does not depend on θ.

In the univariate case, power notation is often employed in the form

$$i_{rst} = E\left\{\left(\frac{\partial l}{\partial\theta}\right)^r\left(\frac{\partial^2 l}{\partial\theta^2}\right)^s\left(\frac{\partial^3 l}{\partial\theta^3}\right)^t;\theta\right\}.$$

The moment identities then become $i_{10} = 0$,

$$\begin{aligned}
&i_{01} + i_{20} = 0,\\
&i_{001} + 3i_{11} + i_{30} = 0,\\
&i_{0001} + 4i_{101} + 3i_{02} + 6i_{21} + i_{40} = 0.
\end{aligned}$$

Similar identities apply to the cumulants, but we refrain from writing these down, in order to avoid further conflict of notation.

Or when θ is a vector:

202 LIKELIHOOD FUNCTIONS

Differentiation with respect to θ and reversing the order of differentiation and integration gives

$$\mu_r = \kappa_r = \int u_r(\theta; y) f_Y(y;\theta)dy = 0.$$

Further differentiation gives

$$\begin{aligned}
\mu_{[rs]} &= \mu_{rs} + \mu_{r,s} = 0\\
\mu_{[rst]} &= \mu_{rst} + \mu_{r,st}[3] + \mu_{r,s,t} = 0\\
\mu_{[rstu]} &= \mu_{rstu} + \mu_{r,stu}[4] + \mu_{rs,tu}[3] + \mu_{r,s,tu}[6] + \mu_{r,s,t,u} = 0.
\end{aligned}$$

Limiting distributions

- $U(\theta) = \sum_{i=1}^{n} U_i(\theta)$
- $E\{U(\theta)\} = 0$
- $var\{U(\theta)\} = ni_1(\theta)$
- $U(\theta)/\sqrt{n} \xrightarrow{d} N\{0, i_1(\theta)\}$ need $0 < i_1(\theta) < \infty$
- Note that could have not i.d., or not independent, if we can still prove the limiting normality of the sum. E.g. Lindeberg-Feller type conditions, or weak dependence

… limiting distributions

- $U(\theta)/\sqrt{n} \xrightarrow{d} N\{0, i_1(\theta)\}$

- $U(\hat{\theta}) = 0 = U(\theta) + (\hat{\theta} - \theta)U'(\theta) + R_n$

- $(\hat{\theta} - \theta) = \{U(\theta)/i(\theta)\}\{1 + o_p(1)\}$

- $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\{0, i_1^{-1}(\theta)\}$

... limiting distributions

- $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\{0, i_1^{-1}(\theta)\}$

- $\ell(\theta) = \ell(\hat{\theta}) + (\theta - \hat{\theta})\ell'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2\ell''(\hat{\theta}) + R_n$

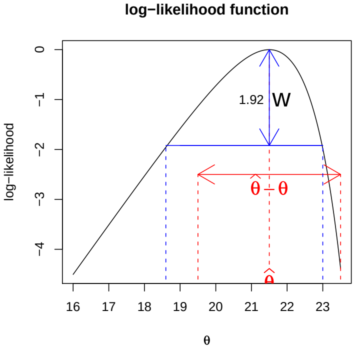
- $2\{\ell(\hat{\theta}) - \ell(\theta)\} = (\hat{\theta} - \theta)^2 i(\theta)\{1 + o_p(1)\}$

- $2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi^2_d$

Inference from limiting distributions

- $\hat{\theta} \dot{\sim} N_d\{\theta, j^{-1}(\hat{\theta})\}$ $\quad\quad j(\hat{\theta}) = -\ell''(\hat{\theta}; y)$
- "θ is estimated to be 21.5 (95% CI 19.5 – 23.5)"
- $\hat{\theta} \pm 2\hat{\sigma}$

- $w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \dot{\sim} \chi^2_d$
- "likelihood based CI for θ with confidence level 95% is (18.6, 23.0)"



p-value functions of θ

- $$\begin{aligned} r_u(\theta) &= U(\theta)j^{-1/2}(\hat{\theta}) \dot{\sim} N(0,1) \\ r_e(\theta) &= (\hat{\theta}-\theta)j^{1/2}(\hat{\theta}), \\ r(\theta) &= \text{sign}(\hat{\theta}-\theta)[2\{\ell(\hat{\theta})-\ell(\theta)\}]^{1/2} \end{aligned}$$
- approximate pivotal quantities

$$\Pr\{r_u(\theta) \le r_u^0(\theta)\} \doteq \Phi\{r_u^0(\theta)\}$$

under sampling from the model $f(y;\theta) = f(y_1,\dots,y_n;\theta)$

- p-value function (of θ, for fixed data)

$$p_u(\theta) = \Phi\{r_u^0(\theta)\}$$

- similarly $p_e(\theta) = \Phi\{r_e(\theta)\}, \quad p_r(\theta) = \Phi\{r(\theta)\}$ are also p-value functions for θ, based on limiting dist'ns

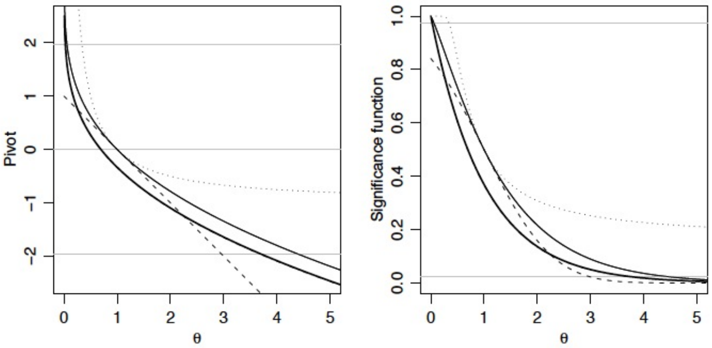
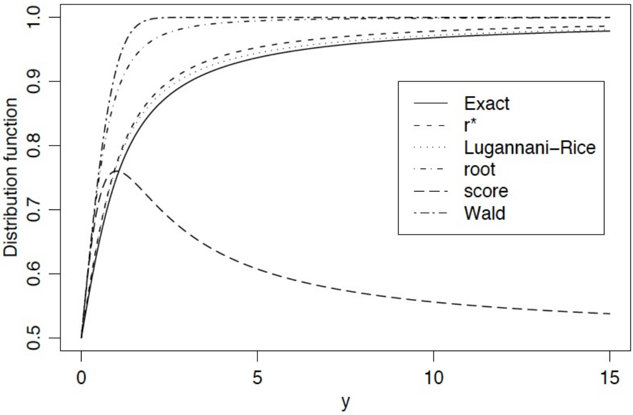


Figure 2.2: Approximate pivots and P-values based on an exponential sample of size $n = 1$. Left: likelihood root $r(\theta)$ (solid), score pivot $s(\theta)$ (dots), Wald pivot $t(\theta)$ (dashes), modified likelihood root $r^*(\theta)$ (heavy), and exact pivot $\theta \sum y_j$ (dot-dash). The modified likelihood root is indistinguishable from the exact pivot. The horizontal lines are at $0, \pm 1.96$. Right: corresponding significance functions, with horizontal lines at 0.025 and 0.975.



BDR, Ch.3.2, Cauchy, distribution functions (y) at $\theta = 0$, $n = 1$

Example

- $f(y_i;\theta) = \theta e^{-y_i\theta}, \quad i = 1, \ldots, n$
- $\ell(\theta) = n \log\theta - n\theta\bar{y}$
- $\ell'(\theta) = \frac{n}{\theta} - n\bar{y}$ $\qquad \hat{\theta} = \bar{y}^{-1}$
- $\ell''(\theta) = -\frac{n}{\theta^2}$
- $r_u(\theta) = \frac{1}{\sqrt{n}}\ell'(\theta)j^{-1/2}(\hat{\theta}) = \sqrt{n}(\frac{1}{\theta\bar{y}} - 1)$
- $r_e(\theta) = (\hat{\theta} - \theta)j^{1/2}(\hat{\theta}) = \sqrt{n}(1 - \bar{y}\theta)$
- $r(\theta) = \surd(2n)\{\theta\bar{y} - 1 - \log(\theta\bar{y})\}^{1/2}$

expand $\log(\theta\bar{y})$ around 1 to get asymptotic equivalence to r_e, r_u

Example

- $f(y_i;\theta) = \theta^{y_i} e^{-\theta}/y_i!$
- $\ell(\theta) =$
- $\ell'(\theta) =$
- $\ell''(\theta) =$
- $r_e(\theta) = (s - n\theta)/\sqrt{s}$
- $\Pr(S \le s) \ne 1 - \Pr(S \ge s)$
- upper and lower p-value functions: $\Pr(S < s), \quad \Pr(S \le s)$
- mid p-value function: $\Pr(S < sr) + 0.5\Pr(S = s)$

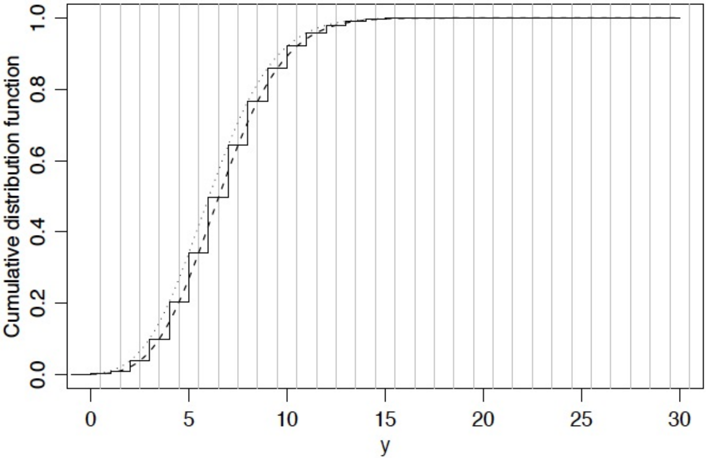


Figure 3.2: Cumulative distribution function for Poisson distribution with parameter 6.7 (solid), with approximations $\Phi\{r^*(y)\}$ (dashes) and $\Phi\{r^*(y + 1/2)\}$ (dots). The vertical lines are at $0.5, 1.5, 2.5, \ldots$

Aside

- for inference re θ, given y, plot $p(\theta)$ vs θ
- for p-value for $H_0 : \theta = \theta_0$, compute $p(\theta_0)$
- for checking whether, e.g. $\Phi\{r_e(\theta)\}$ is a good approximation,
  - compare $p(\theta) = \Phi\{r_e(\theta)\}$ to $p_{\text{exact}}(\theta)$, as a function of θ, fixed y
  - or compare $p(\theta_0)$ to $p_{\text{exact}}(\theta_0)$ as a function of y
- if $p_{\text{exact}}(\theta)$ not available, simulate
- if θ is a vector, choose one component at a time

Nuisance parameters

- $\theta = (\psi, \lambda) = (\psi_1, \dots, \psi_q, \lambda_1, \dots, \lambda_{d-q})$
- $U(\theta) = \begin{pmatrix} U_\psi(\theta) \\ U_\lambda(\theta) \end{pmatrix}, \qquad U_\lambda(\psi, \hat\lambda_\psi) = 0$
- $i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} \quad j(\theta) = \begin{pmatrix} j_{\psi\psi} & j_{\psi\lambda} \\ j_{\lambda\psi} & j_{\lambda\lambda} \end{pmatrix}$
- $i^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix} \quad j^{-1}(\theta) = \begin{pmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{pmatrix}.$
- $i^{\psi\psi}(\theta) = \{i_{\psi\psi}(\theta) - i_{\psi\lambda}(\theta) i^{-1}_{\lambda\lambda}(\theta) i_{\lambda\psi}(\theta)\}^{-1},$
- $\ell_{\mathrm{P}}(\psi) = \ell(\psi, \hat\lambda_\psi), \qquad j_{\mathrm{P}}(\psi) = -\ell''_{\mathrm{P}}(\psi)$

Inference from limiting distributions, nuisance parameters

$$\begin{aligned} w_u(\psi) = U_\psi(\psi,\hat\lambda_\psi)^T\{i^{\psi\psi}(\psi,\hat\lambda_\psi)\}U_\psi(\psi,\hat\lambda_\psi) &\ \dot\sim\ \chi^2_q \\ w_e(\psi) = (\hat\psi-\psi)\{i^{\psi\psi}(\hat\psi,\hat\lambda)\}^{-1}(\hat\psi-\psi) &\ \dot\sim\ \chi^2_q \\ w(\psi) = 2\{\ell(\hat\psi,\hat\lambda)-\ell(\psi,\hat\lambda_\psi)\} = 2\{\ell_{\mathrm{P}}(\hat\psi)-\ell_{\mathrm{P}}(\psi)\} &\ \dot\sim\ \chi^2_q; \end{aligned}$$

Approximate Pivots, $q = 1$

$$\begin{aligned} r_u(\psi) &= \ell_{\mathrm{P}}'(\psi)j_{\mathrm{P}}(\hat\psi)^{-1/2} \ \dot\sim\ N(0,1), \\ r_e(\psi) &= (\hat\psi-\psi)j_{\mathrm{P}}(\hat\psi)^{1/2} \ \dot\sim\ N(0,1), \\ r(\psi) &= \mathrm{sign}(\hat\psi-\psi)[2\{\ell_{\mathrm{P}}(\hat\psi)-\ell_{\mathrm{P}}(\psi)\}]^{1/2} \ \dot\sim\ N(0,1) \end{aligned}$$

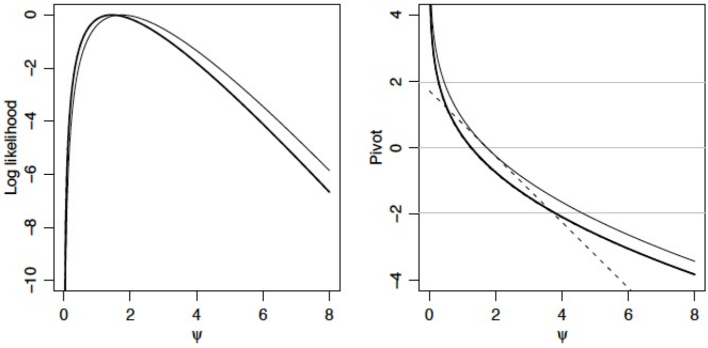


Figure 2.3: Inference for shape parameter ψ of gamma sample of size $n = 5$. Left: profile log likelihood ℓ_{p} (solid) and the log likelihood from the conditional density of u given v (heavy). Right: likelihood root $r(\psi)$ (solid), Wald pivot $t(\psi)$ (dashes), modified likelihood root $r^*(\psi)$ (heavy), and exact pivot overlying $r^*(\psi)$. The horizontal lines are at $0, \pm 1.96$.