

7

What is likelihood?

Likelihood theory is the single most important inference framework. Invented by Ronald A. Fisher in 1922 it provides a systematic way for automatically constructing estimators with good properties. In this chapter we derive the likelihood function as approximation to the KL information and describe the many advantages but also the limits of likelihood inference.

7-1 Likelihood function and Kullback-Leibler information

The center stage of likelihood theory is taken by the **likelihood function**, which is defined

$$L(\boldsymbol{\theta}|x_1, \dots, x_n) = \prod_{i=1}^n f(x_i|\boldsymbol{\theta}).$$

Roughly speaking, the likelihood is the probability $\Pr(x_1, \dots, x_n|F_\theta)$ to observe independent data x_1, \dots, x_n under the hypothesis F_θ . This interpretation is not strictly correct because for continuous variables f is not a probability mass function but a density that can assume values larger than one. The likelihood must not be confused with and is not the same as $\Pr(F_\theta|x_1, \dots, x_n)$ (see Bayes chapter). Instead of the likelihood function one usually considers the log-likelihood

$$\log L(\boldsymbol{\theta}|x_1, \dots, x_n) = \sum_{i=1}^n \log f(x_i|\boldsymbol{\theta})$$

because it is numerically more stable, and as we see shortly this is actually the more fundamental quantity. By construction, the log-likelihood is additive for independent observations x_i .

For a better understanding of the log-likelihood function we have to go back to the Kullback-Leibler information (**Equation 4-1**). We denote by F_0 the unknown true data generating model and by F_θ a family of distributions that we employ to approximate F_0 (note it is not necessary to assume that F_θ contains F_0). The KL information is given by

$$\begin{aligned} I^{\text{KL}}(F_0; F_\theta) &= E_{F_0} \log f_0(X) - E_{F_0} \log f(X|\boldsymbol{\theta}) \\ &= C - E_{F_0} \log f(X|\boldsymbol{\theta}). \end{aligned}$$

The first term, the constant C , doesn't contain the parameter θ and is therefore irrelevant for relative comparison of models (however, its presence ensures that I^{KL} is transformation-invariant, see Chapter 4). A simple way to estimate the second term is to replace the expectation with its empirical counterpart (as in $E(X) \approx \frac{1}{n} \sum_{i=1}^n x_i$). This is valid if the sample size is large relative to the dimension of θ and leads to

$$-E_{F_0} \log f(X|\theta) \approx -\frac{1}{n} \sum_{i=1}^n \log f(x_i|\theta) = -\frac{1}{n} \log L(\theta|x_1, \dots, x_n).$$

Thus, the negative average log-likelihood function has a very simple interpretation — it acts as large sample proxy to the Kullback-Leibler information via

$$I^{\text{KL}}(F_0; F_\theta) \approx C - \frac{1}{n} \log L(\theta|x_1, \dots, x_n).$$

7–2 Maximum likelihood estimation

The close link between the likelihood function and the KL divergence provides the reasoning for the method of **maximum likelihood**.

Ideally, we would minimize the KL divergence between a set of candidate models F_θ and the unknown true model. However, as we don't know the true model we cannot actually compute the KL information. However, if the sample size is sufficiently large, we can use the above approximation and maximize the (log-)likelihood instead. This leads to the **maximum likelihood estimate** (MLE),

$$\hat{\theta}_{\text{ML}} = \arg \max l_n(\theta).$$

Here and in the following we simplify notation and write $l_n(\theta)$ instead of the more lengthy $\log L(\theta|x_1, \dots, x_n)$. The subscript n serves as a reminder that we have used n iid samples for computing the likelihood.

The **score function** is the gradient of the log-likelihood

$$S_n(\theta) = \nabla l_n(\theta).$$

In most cases the MLE $\hat{\theta}_{\text{ML}}$ is obtained as a solution of the equation $S(\theta) = 0$. However, that is not always possible, e.g., when the MLE lies at a boundary (see **Example 7–5**). The **observed Fisher information matrix** is defined as the *negative* Hessian of the log-likelihood

$$J_n(\theta) = -\nabla^T \nabla l_n(\theta).$$

The quantity $J_n(\theta)$ is related but not identical to the *expected* Fisher information $I^{\text{Fisher}}(\theta)$ (see **Equation 4–5**). The precise relationship is

$$I^{\text{Fisher}}(\theta) = E\left(J_1(\theta)\right)$$

so that for large sample size $J_n(\theta)/n \rightarrow I^{\text{Fisher}}(\theta)$. Of particular interest is the observed Fisher information $J_n(\hat{\theta}_{\text{ML}})$ assumed at the maximum of the likelihood function. As this corresponds to the curvature of the likelihood function at its peak it is intuitively clear that it provides a measure of precision of the maximum likelihood estimate.

Example 7–1: Maximum likelihood estimate of a proportion

We conduct a Bernoulli experiment r times and observe n_1 successes and n_2 failures, with $r = n_1 + n_2$. What is the MLE of the underlying success probability μ ?

The probability to observe n_1 successes is given by the Binomial distribution. The corresponding log-likelihood for μ is

$$l_r(\mu) = n_1 \log \mu + (r - n_1) \log(1 - \mu)$$

and the score function is

$$S_r(\mu) = \frac{n_1}{\mu} - \frac{r - n_1}{1 - \mu}.$$

Setting $S_r(\mu)$ equal to zero gives

$$\hat{\mu}_{\text{ML}} = \frac{n_1}{r},$$

i.e. the standard empirical proportion. The observed Fisher information at the MLE is

$$J_r(\hat{\mu}_{\text{ML}}) = \frac{r}{\hat{\mu}_{\text{ML}}(1 - \hat{\mu}_{\text{ML}})}$$

which we recognize as the *inverse* empirical variance of the parameter of the Binomial distribution.

Example 7–2: MLE of the parameters of the normal distribution

Suppose we have n independent observations x_1, \dots, x_n from a univariate normal distribution $N(\mu, \sigma^2)$. What is the MLE of the parameters μ and σ^2 ?

The normal density is $f(x) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. With parameter vector $\theta = (\mu, \sigma^2)^T$ and dropping all constant terms not depending on θ we have as log-likelihood

$$l_n(\theta) = \frac{n}{2} \log\left(\frac{1}{\sigma^2}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

The corresponding score function is

$$S_n(\theta) = \left(\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu), -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 \right),$$

which equated to zero leads to maximum likelihood estimates

$$\hat{\theta}_{\text{ML}} = (\hat{\mu}_{\text{ML}}, \widehat{\sigma^2}_{\text{ML}}) = \left(\frac{1}{n} \sum_{i=1}^n x_i, \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_{\text{ML}})^2 \right)^T.$$

This shows that the standard average is the MLE of the mean and the empirical variance with a factor of $\frac{1}{n}$ the MLE of the variance. The latter implies that the MLE of the variance is biased. Some further calculation gives the observed Fisher information at the maximum likelihood point

$$J_n(\hat{\theta}_{\text{ML}}) = \begin{pmatrix} \frac{n}{\widehat{\sigma^2}_{\text{ML}}} & 0 \\ 0 & \frac{n}{2(\widehat{\sigma^2}_{\text{ML}})^2} \end{pmatrix}.$$

The expected Fisher information for this model is very similar (cf. **Example 4–4**), except for a factor n and the fact that $I^{\text{Fisher}}(\theta)$ is a constant and not a random matrix as $J_n(\hat{\theta}_{\text{ML}})$.

From the previous chapter we know the exact variance of the empirical mean $\hat{\mu}$ and of $\widehat{\sigma^2}_{\text{ML}}$. Specifically, if we observe n iid samples drawn from a distribution (not necessarily normal!) with mean μ and variance σ^2 then $\text{Var}(\hat{\mu}) = \sigma^2/n$ and $\text{Var}(\widehat{\sigma^2}_{\text{ML}}) = \frac{2}{n-1}\sigma^4$. The inverse of the observed Fisher information matrix provides estimates of these variances.

Example 7–3: Log-likelihood of the multivariate normal distribution

In generalization of **Example 7–2** we study the multivariate normal model with density

$$f(x|\mu, \Sigma) = (2\pi)^{-d/2} \det(\Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right\}.$$

The log-likelihood function based on independently sampled data X_1, \dots, X_n is

$$\begin{aligned} l_n(\mu, \Sigma) &= \frac{n}{2} \log \det(\Sigma^{-1}) - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \\ &= \frac{n}{2} \log \det(\Sigma^{-1}) - \frac{1}{2} \text{tr} \left(\Sigma^{-1} \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T \right). \end{aligned}$$

The score function becomes (with the help of some clever matrix calculus)

$$S_n(\mu, \Sigma) = \left(\sum_{i=1}^n (x_i - \mu)^T \Sigma^{-1}, -\frac{n}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-2} \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T \right)$$

which results in the maximum likelihood estimates

$$(\hat{\mu}_{\text{ML}}, \hat{\Sigma}_{\text{ML}}) = \left(\frac{1}{n} \sum_{i=1}^n x_i, \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_{\text{ML}})(x_i - \hat{\mu}_{\text{ML}})^T \right)$$

with observed Fisher information at the MLE $J_n(\hat{\mu}_{\text{ML}}) = n\hat{\Sigma}_{\text{ML}}^{-1}$ and $J_n(\hat{\Sigma}_{\text{ML}}) = \frac{n}{2}\hat{\Sigma}_{\text{ML}}^{-2}$.

Example 7–4: Relationship of maximum likelihood and the method of least squares

In the univariate normal model the log-likelihood is a function of the sum of the squared residuals $\varepsilon^2 = \sum_{i=1}^n (x_i - \mu)^2$. Maximizing the log-likelihood for the parameter μ is identical to minimizing ε^2 . Hence, the method of least squares is simply a special case of maximum-likelihood assuming a normal distribution. Consequently, least squares inherits the many favorable properties of maximum likelihood (but also its defects).

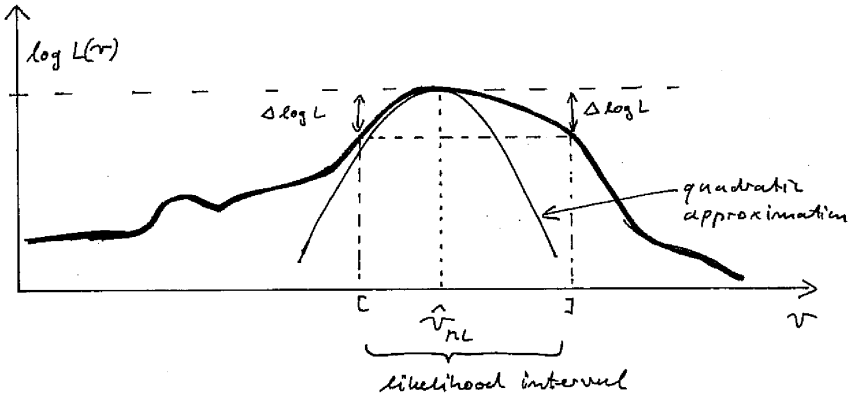


Figure 7-1: Likelihood function and its quadratic approximation.

7-3 Quadratic approximation and normal asymptotics

If the underlying model is regular, i.e. smooth and sufficiently often differentiable, we can quadratically approximate the log-likelihood by the first terms of a Taylor series around the MLE $\hat{\theta}^{ML}$,

$$l_n(\theta) \approx l_n(\hat{\theta}_{ML}) - \frac{1}{2}(\hat{\theta}_{ML} - \theta)^T J_n(\hat{\theta}_{ML}) (\hat{\theta}_{ML} - \theta). \quad \text{Equation 7-1}$$

There is no linear term because in regular situations the score vector vanishes at the MLE (however, see **Example 7-5** for a counterexample). In case of a univariate parameter θ the approximation takes the shape of a parabola as illustrated in **Figure 7-1**.

Intriguingly, **Equation 7-1** has the form of a multivariate normal log-likelihood with $\hat{\theta}_{ML}$ as single data point and model $N_d(\theta, J_n(\hat{\theta}_{ML})^{-1})$. This directly implies that regular maximum likelihood estimates are approximately normally distributed with covariance given by the inverse observed Fisher information at the MLE. Thus we have

$$\hat{\theta}_{ML} \overset{a}{\sim} N_d(\theta, J_n(\hat{\theta}_{ML})^{-1}).$$

The “a” in the symbol $\overset{a}{\sim}$ reminds us that the distribution is assumed asymptotically for large n (or large J_n) and for finite sample size holds only approximately. It is often useful to standardize and Mahalanobis-decorrelate the maximum likelihood estimates. This results in the **Wald pivot** or **Wald statistic**

$$t(\theta) = J_n(\hat{\theta}_{ML})^{1/2}(\hat{\theta}_{ML} - \theta) \quad \text{Equation 7-2}$$

that is approximately standard normal distributed

$$\mathbf{t}(\boldsymbol{\theta}) \stackrel{a}{\sim} N_d(0, \mathbf{I}_d).$$

Closely related to the Wald statistic is the **standardized score statistic**

$$\mathbf{s}(\boldsymbol{\theta}) = (n\mathbf{I}^{\text{Fisher}}(\boldsymbol{\theta}))^{-1/2} \mathbf{S}_n(\boldsymbol{\theta})^T.$$

From Chapter 4 (Problems) we recall that the score $S_1(\boldsymbol{\theta})$ for a single random sample X has vanishing expectation $E(S_1(\boldsymbol{\theta})) = 0$ and that $\text{Var}(S_1(\boldsymbol{\theta})) = \mathbf{I}^{\text{Fisher}}(\boldsymbol{\theta})$. Hence, for n iid samples we have $E(\mathbf{S}_n(\boldsymbol{\theta})) = 0$ and $\text{Var}(\mathbf{S}_n(\boldsymbol{\theta})) = n\mathbf{I}^{\text{Fisher}}(\boldsymbol{\theta})$. Thus, for large n we find with the central limit theorem

$$\mathbf{s}(\boldsymbol{\theta}) \stackrel{a}{\sim} N_d(0, \mathbf{I}_d).$$

The Wald and standardized score statistic have in common that they employ Fisher information for standardization and decorrelation. As we know their asymptotic distribution both can be used for constructing approximate confidence intervals for the MLE.

Example 7-5: Uniform model

Likelihood inference in uniform model is an example where the quadratic normal approximation fails. Suppose that we observe data $x_1, \dots, x_n \geq 0$ from the uniform distribution on the interval $[0, \theta]$. The corresponding density is

$$f(x|\theta) = \begin{cases} \theta^{-1} & \text{for } x \in [0, \theta], \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Correspondingly, with $x_{[n]} = \max(x_1, \dots, x_n)$ the log-likelihood function is

$$l_n(\theta) = \begin{cases} -n \log \theta & \text{if } \theta \geq x_{[n]}, \text{ and} \\ -\infty & \text{otherwise.} \end{cases}$$

The likelihood is maximized at $\hat{\theta}_{\text{ML}} = x_{[n]}$. However, the score function is *not* zero at $\hat{\theta}_{\text{ML}}$, as $\log L(\theta)$ is not even differentiable at the MLE. As a result, the Taylor series needed in the quadratic approximation of the log-likelihood cannot be applied, and the Fisher information cannot be computed.

However, the distribution of the MLE is still amenable. $X_{[n]}$ is the n -th order statistic on the interval $[0, \theta]$ which is distributed $X_{[n]} \sim \theta \text{Beta}(n, 1)$ with mean $E(X_{[n]}) = \frac{n}{n+1} \theta$ and variance $\text{Var}(X_{[n]}) = \frac{n}{(n+1)^2(n+2)} \theta^2$. As the variance (and also MSE) declines with $1/n^2$ rather than with the usual $1/n$ the MLE is **superefficient**.

Example 7-6: Wald and standardized score statistic for the normal mean

In the normal model with fixed variance σ^2 the log-likelihood for the mean is $l_n(\mu) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$, the score function is $S_n(\mu) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)$ and the MLE is $\hat{\mu}_{\text{ML}} =$

$\frac{1}{n} \sum_{i=1}^n x_i$. Furthermore, the observed Fisher information at the MLE is $J_n(\hat{\mu}_{\text{ML}}) = \frac{n}{\sigma^2}$ and the expected Fisher information equals $I^{\text{Fisher}}(\mu) = \frac{1}{\sigma^2}$.

With this in hand it can be easily verified that the Wald statistic and the standardized score statistic are identical and equal

$$t(\mu) = s(\mu) = \frac{\hat{\mu}_{\text{ML}} - \mu}{\sigma / \sqrt{n}}.$$

In this case the distribution of the MLE is known exactly with $\hat{\mu}_{\text{ML}} \sim N(\mu, \frac{\sigma^2}{n})$, so we see that the distribution $t(\mu) = s(\mu) \sim N(0, 1)$ is also exact.

Example 7–7: Normal confidence interval

The asymptotic normality of regular MLEs makes it straightforward to obtain approximate confidence intervals. For example, for a univariate parameter θ with MLE $\hat{\theta}_{\text{ML}}$ and asymptotic standard deviation $\hat{\sigma} = J_n(\hat{\theta}_{\text{ML}})^{-1/2}$ a symmetric β -confidence interval covering β probability is given by

$$[\hat{\theta}_{\text{ML}} - c_1 \hat{\sigma}; \hat{\theta}_{\text{ML}} + c_1 \hat{\sigma}].$$

The constant $c_1 = \Phi^{-1}((1 + \beta)/2)$ is the $(1 + \beta)/2$ quantile of the unit normal density. Typically, one uses one of the following three choices: $c_1 = 2.58$ corresponding to $\beta = 99\%$, $c_1 = 1.96$ for a $\beta = 95\%$ confidence interval and $c_1 = 1.64$ for $\beta = 90\%$.

7–4 Likelihood ratio statistic

In non-regular models we may not be able evaluate the score at the MLE or compute the Fisher information needed for the Wald statistic. These problems are circumvented by directly looking at log-likelihood differences. The quantity

$$W(\theta) = 2 \log \left(\frac{L(\hat{\theta}_{\text{ML}})}{L(\theta)} \right) = 2(l_n(\hat{\theta}_{\text{ML}}) - l_n(\theta))$$

is called the **Wilks likelihood ratio statistic**. By construction, $W(\theta) \geq 0$ for any choice of θ . From **Equation 7–1** and **Equation 7–2** we see that in regular situations $W(\theta) \approx \mathbf{t}(\theta)^T \mathbf{z}(\theta)$ and therefore

$$W(\theta) \overset{a}{\sim} \chi_d^2,$$

i.e. it follows approximately a chi-square distribution with the dimension of the parameter vector θ as the degrees of freedom. In non-regular models the likelihood ratio can still be computed, however, in this case the distribution of $W(\theta)$ is *not* necessarily χ_d^2 and must be obtained, e.g., by simulation.

Instead of the likelihood ratio statistic $W(\theta)$ one can also consider the **signed likelihood root**

$$r(\theta) = \text{sign}(\hat{\theta}_{\text{ML}} - \theta) \sqrt{W(\theta)} \overset{a}{\sim} N_d(0, I_d)$$

with $W(\boldsymbol{\theta}) = \mathbf{r}(\boldsymbol{\theta})^T \mathbf{r}(\boldsymbol{\theta})$ and

$$\mathbf{r}(\boldsymbol{\theta}) \stackrel{a}{\sim} N_d(0, \mathbf{I}_d).$$

Example 7–8: Signed likelihood root for the normal mean

For the normal mean with fixed variance σ^2 Wilks likelihood ratio is

$$W(\mu) = 2l_n(\hat{\mu}_{\text{ML}}) - 2l_n(\mu) = \frac{n}{\sigma^2} (\hat{\mu}_{\text{ML}} - \mu)^2$$

so the signed likelihood root equals

$$r(\mu) = \frac{\hat{\mu}_{\text{ML}} - \mu}{\sigma/\sqrt{n}}$$

which is in this case identical to the Wald statistic $t(\mu)$ and the standardized score statistic $s(\mu)$. The distribution of $r(\mu) \sim N(0, 1)$ is exact as well.

Example 7–9: Likelihood interval

Alternatively, we may define a univariate **likelihood interval**

$$\left\{ \theta, \frac{L(\theta)}{L(\hat{\theta}_{\text{ML}})} > c_2 \right\}$$

in terms of likelihood ratios around $\hat{\theta}_{\text{ML}}$. Using the Wilks statistic the same interval is given by

$$\{ \theta, W(\theta) < -2 \log(c_2) \}$$

By applying the quadratic log-likelihood approximation (**Equation 7–1**) and by comparison with the normal intervals we see that $\log L(\theta) - \log L(\hat{\theta}_{\text{ML}}) > -\frac{1}{2}c_2^2$ and hence $c_2 = e^{-c_1^2/2}$. Therefore, the threshold in the likelihood interval equals $c_2 = 0.04$ for $\beta = 0.99$, $c_2 = 0.15$ for $\beta = 0.95$ and $c_2 = 0.26$ for $\beta = 0.90$. The corresponding log-likelihood differences $\Delta \log L = \log L(\hat{\theta}_{\text{ML}}) - \log L(\theta)$ are 3.32, 1.92 and 1.35, respectively.

By construction, likelihood intervals need not be symmetric (cf. **Figure 7–1**). However, if the log-likelihood around the MLE is not well approximated by a quadratic function then the normal calibration will be incorrect.

Example 7–10: Uniform model (II)

likelihood interval for uniform model?

Example 7–11: Exponential model

The exponential distribution with mean μ implies a log-likelihood function $l_n(\mu) = -n \log(\mu) - \frac{1}{\mu} \sum_{i=1}^n x_i$, a score function $S_n(\mu) = -\frac{n}{\mu} + \frac{1}{\mu^2} \sum_{i=1}^n x_i$, and MLE $\hat{\mu}_{\text{ML}} = \frac{1}{n} \sum_{i=1}^n x_i$. The observed Fisher information at the MLE equals $J_n(\hat{\mu}_{\text{ML}}) = \frac{n}{\mu^2}$ and the expected Fisher information is $I^{\text{KL}}(\mu) = \frac{1}{\mu^2}$.

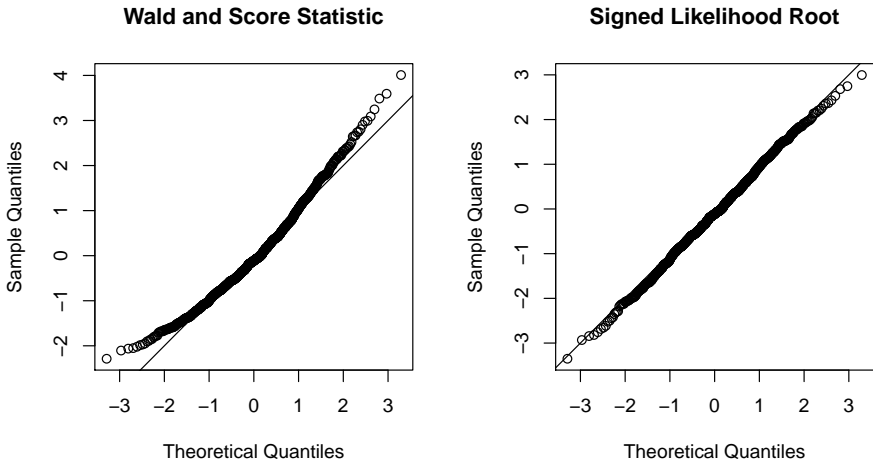


Figure 7–2: Normal QQ-plots of the Wald and standardized score statistics and the signed likelihood root for an exponential model (**Example 7–11**).

The Wald statistic and the standardized score statistic are identical and equal

$$t(\mu) = s(\mu) = \frac{\hat{\mu}_{\text{ML}} - \mu}{\mu / \sqrt{n}}.$$

In contrast, the likelihood ratio statistic is

$$W(\mu) = 2n \left(\frac{\hat{\mu}_{\text{ML}}}{\mu} - \log \left(\frac{\hat{\mu}_{\text{ML}}}{\mu} \right) - 1 \right).$$

In **Figure 7–2** we evaluate by simulation the quality of the normal asymptotics for $t(\mu) = s(\mu)$ as well as for the signed likelihood root $r(\mu) = \text{sign}(\hat{\mu}_{\text{ML}} - \mu) \sqrt{W(\mu)}$. The true mean was set to $\mu = 4$, the sample size was $n = 10$, and 1000 repetitions were performed to obtain 1000 t and r values. Clearly, the signed likelihood root is the most accurate and hence it and $W(\mu)$ are preferable to construct a confidence interval.

Example 7–12: p^* formula

[TO DO: examples for non-normal confidence intervals; comparison of wald type and likelihood interval; variance stabilization to improve wald interval; notation: score is a ROW vector!]

first order approximation

higher order approximations are available (e.g. to get t-distribution) but generally — as likelihood point estimate only works well for large samples — it is doubtful whether it makes sense computer higher order confidence intervals for small samples.