Stat 5421 Notes: Likelihood Inference

Charles J. Geyer

October 04, 2023

Contents

1 License **1**

2 Likelihood **1**

3 The Method of Maximum Likelihood **2**
3.1 Local and Global . . . 3
3.2 Global . . . 3

4 One-Parameter Example **4**
4.1 Maximum Likelihood Estimation . . . 5
4.2 Asymptotic Distribution . . . 6
4.3 Plug-In . . . 6
4.4 Wald Interval . . . 7
4.5 Likelihood Interval . . . 8
4.6 Rao (Score) Interval . . . 8
4.7 Summary . . . 9

5 Two-Parameter Example **9**

6 Wald, Wilks, Rao **10**
6.1 Nested Models . . . 10
6.2 The Tests . . . 11
6.3 Specializing to the One-Parameter Case . . . 12
6.4 Applications . . . 13
6.5 Binomial Examples . . . 13

Bibliography **15**

1 License

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (http://creativecommons.org/licenses/by-sa/4.0/).

2 Likelihood

In statistics, the word "likelihood" has a technical meaning, given it by Fisher (1912). The likelihood for a statistical model is the probability mass function (PMF) or probability density function (PDF) thought of as a function of parameters rather than as a function of data

$$L_x(\theta) = f_\theta(x).$$

The right-hand side is a function of x that is nonnegative and sums (for a PMF) or integrates (for a PDF) to one. The left-hand side is a function of θ that doesn't sum or integrate to anything in particular (and need not sum or integrate to anything finite).

Except that is not the most general notion of likelihood. We are allowed to drop multiplicative terms that do not contain the parameter from the likelihood

$$L_x(\theta) = \frac{f_\theta(x)}{\text{term that does not contain } \theta}.$$

Likelihood plays a key role in frequentist statistical inference through the method of maximum likelihood (the rest of this document) and Bayesian inference (another document). And in neither does it matter if we to drop multiplicative terms that do not contain the parameter from the likelihood. We get the same point estimates, hypothesis tests, and confidence intervals (for frequentist inference) and the same posterior distributions (for Bayesian inference) whether we drop such terms or not.

In this course we are interested in statistical models for discrete data so the likelihood will always be a PMF perhaps with multiplicative terms that do not contain the parameters dropped.

For the binomial distribution the PMF is

$$f_\pi(x) = \binom{n}{x} \pi^x (1 - \pi)^{n - x}$$

and the likelihood is either the same right-hand side or

$$L_x(\pi) = \pi^x (1 - \pi)^{n - x}$$

(dropping a multiplicative term that does not contain the parameter π).

For the Poisson distribution the PMF is

$$f_\mu(x) = \frac{\mu^x}{x!} e^{-\mu}$$

and the likelihood is either the same right-hand side or

$$L_x(\mu) = \mu^x e^{-\mu}$$

(dropping a multiplicative term that does not contain the parameter μ).

3 The Method of Maximum Likelihood

Fisher (1912) proposed the method of maximum likelihood which takes the parameter value with the highest likelihood (for the observed data) to be the point estimator of the unknown parameter. Fisher (1922), the paper that really starts mathematical statistics, giving the subject its first coherent theory, gave heuristic arguments for why the maximum likelihood estimator (MLE) is consistent and asymptotically normal (CAN) and efficient, where

- consistent means the estimator converges to the true unknown parameter value as sample size goes to infinity (this is the modern definition, not Fisher's definition),
- asymptotically normal means the distribution of the estimator converges to a normal distribution as sample size goes to infinity, and
- efficient means the estimator is best possible in the sense of having the smallest possible variance of its asymptotic normal distribution.

Fisher's heuristic arguments were made completely rigorous by many later authors, although some qualifications and regularity conditions were necessary.

3.1 Local and Global

In theory, we distinguish two kinds of MLE. We can seek the global maximizer of the likelihood or be satisfied with a mere local maximizer. In one dimension, we can graph the function and just see where the global maximizer is. In higher dimensions, global optimization is hard. Neither we nor our computers know how to do it except in very special cases (notes on exponential families).

We know from calculus that, if function f has a maximum at a point x where it is twice differentiable, then $f'(x) = 0$ and $f''(x) \leq 0$. So these are necessary conditions for x to be a maximizer of f. We also know from calculus that $f'(x) = 0$ and $f''(x) < 0$ is a sufficient condition for x to be at least a local maximizer of f. A *local maximizer* x of f maximizes f over a (perhaps very small) neighborhood of x; every point y sufficiently close to x has $f(y) \leq f(x)$.

There is analogous theory from multivariable calculus. This requires that we know what multivariable calculus considers first and second derivatives. We write $\nabla f(x)$ to denote the vector of first partial derivatives, which has i-th component

$$\frac{\partial f(x)}{\partial x_i}$$

(here x is a vector having components x_i), and we write $\nabla^2 f(x)$ to denote the matrix of second partial derivatives, which has (i, j)-th component

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j}.$$

It also requires that we know what the terms positive semidefinite, positive definite, negative semidefinite, and negative definite mean as descriptions of symmetric matrices. A symmetric matrix is

- positive semidefinite if all its eigenvalues are nonnegative,
- positive definite if all its eigenvalues are positive,
- negative semidefinite if all its eigenvalues are nonpositive, and
- negative definite if all its eigenvalues are negative.

The necessary condition for a maximum of f at a point x where f is twice differentiable is $\nabla f(x) = 0$ and $\nabla^2 f(x)$ is a negative semidefinite matrix. The sufficient condition for a local maximum of f at a point x where f is twice differentiable is $\nabla f(x) = 0$ and $\nabla^2 f(x)$ is a negative definite matrix.

The R function `eigen` calculates eigenvalues and eigenvectors. Suppose we have already defined a symmetric matrix `m`. Then

```r
v <- eigen(m, symmetric = TRUE, only.values = TRUE)$values
all(v < 0)
```

```
## [1] TRUE
```

says `TRUE` if and only if `m` is negative definite (modulo inexactness of computer arithmetic — if `m` has eigenvalues very close to zero, it may give the wrong answer). If we wanted to check for negative semidefinite, we would replace `all(v < 0)` with `all(v <= 0)`.

So we and our computers know how to find local maximizers (the computer goes uphill on f until it finds a point x that satisfies the sufficient condition to within the inexactness of computer arithmetic or gives up and emits an error if it cannot find such a point).

3.2 Global

Under the "usual regularity conditions" for consistency of MLE (Wald (1949); Kiefer and Wolfowitz (1956); Wang (1985)) the global MLE is consistent. (Statisticians say "usual regularity conditions" to refer to some unspecified regularity conditions for some theorem in some paper or textbook that they may have seen in some theory class or just in their reading or only have heard about without getting into the gory details. The reason is that there are no simple necessary and sufficient conditions for either consistency or asymptotic normality of MLE, so every theorem in every paper or textbook has somewhat different conditions. The conditions that are easier to understand are not as sharp. The sharp conditions are harder to understand. So authors have produced a lot of variants. No conditions are easy to understand. That is why the "easier" above. All tend to be a long laundry list of unmotivated technicalities that are assumed just to make a particular method of proof work. So no such regularity conditions are easily motivated or explained. The cited papers are difficult to read. Ferguson (1996), Chapter 17, gives a simpler theorem, but still one appropriate for Stat 8112, which is several levels above this course, but not as sharp as the theory in the cited papers, where "not as sharp" means Ferguson's regularity conditions do not apply to as many statistical models as those of the papers cited.)

So this lends support to the idea that MLE should mean the *global* maximizer of the likelihood. It also corresponds to the naive notion: maximum means maximum (not some technicality like local maximum). It also corresponds to Fisher (1912) and Fisher (1922) who did not discuss local and global maximizers and thus lead people to believe he meant global.

Under the "usual regularity conditions" for asymptotic normality of MLE (Cramér (1946), Chapters 32 and 33; Ferguson (1996), Chapters 16–18, van der Vaart (1998), Chapters 5–7, Le Cam and Yang (2000), Chapters 6–7, Geyer (2013)) the global MLE, if consistent,[1] is asymptotically normal. Moreover it is efficient (LeCam (1953); van der Vaart (1998), Section 8.6). In short the MLE is consistent, asymptotically normal, and efficient.

But these theorems also assert that even if the global MLE is not consistent (this can happen) or does not even exist (this can also happen) that a "good" local maximizer of the likelihood (a local MLE) will also be consistent, asymptotically normal, and efficient. So the question becomes how to find such a "good" local maximizer. Theory says start at a "good enough" estimator (more on this below) and then go uphill to a local maximum[2] and this will produce an estimator (a local MLE) that is consistent, asymptotically normal, and efficient.

An estimator is a "good enough" starting point for finding a local MLE if it is what the jargon calls $\sqrt{n}$-consistent, that is, if it obeys what a textbook formerly used for Stat 1001 Freedman *et al.* (2007) calls the "square root law" (statistical precision varies inversely with sample size).[3]

4 One-Parameter Example

This example is not categorical data, because as we will eventually find out (exponential families notes) that maximum likelihood estimation is much easier for most models in categorical data analysis than for general models. This is not about continuous versus discrete. The model in the follow-on notes for these notes also needs good starting points for maximum likelihood to work. But that model has two parameters and we want to start with something simpler.

Suppose we have a vector of data `x` that is supposed to be an independent and identically distributed (IID) sample from a Cauchy distribution. In real life the location and scale parameters would both be unknown, but here we just simulate some data (so we actually know the parameter values, but we pretend we don't know them so we can do some statistics.

```r
set.seed(42)
x <- rcauchy(30, location = pi, scale = exp(1))
```

[1] That is, if the statistical model also satisfies the other set of "usual regularity conditions," the ones for consistency.

[2] Or even do not go all the way to the local maximum. The theory says from a "good enough" estimator doing one step of Newton's method to maximize the likelihood is enough to produce an estimator that is consistent, asymptotically normal, and efficient (@van-der-vaart, Section 5.7; @geyer-simple).

[3] Technically, an estimator $\tilde{\theta}_n$ is $\sqrt{n}$-consistent if $\sqrt{n}(\tilde{\theta}_n - \theta)$ is bounded in probability, where n is the sample size and θ is the true unknown parameter value, and where bounded in probability means for every $\varepsilon > 0$ there exists an $r < \infty$ such that $\Pr\{\sqrt{n}(\tilde{\theta}_n - \theta) > r)\} < \varepsilon$ for all sample sizes n. If $\sqrt{n}(\tilde{\theta}_n - \theta)$ converges in distribution to anything whatsoever (normality is not necessary), then it is bounded in probability.

For this example, we will pretend that we know the scale parameter. But we assume we know the wrong value. We use the default parameter value `scale = 1`. (Later we will show how to estimate both parameters.)

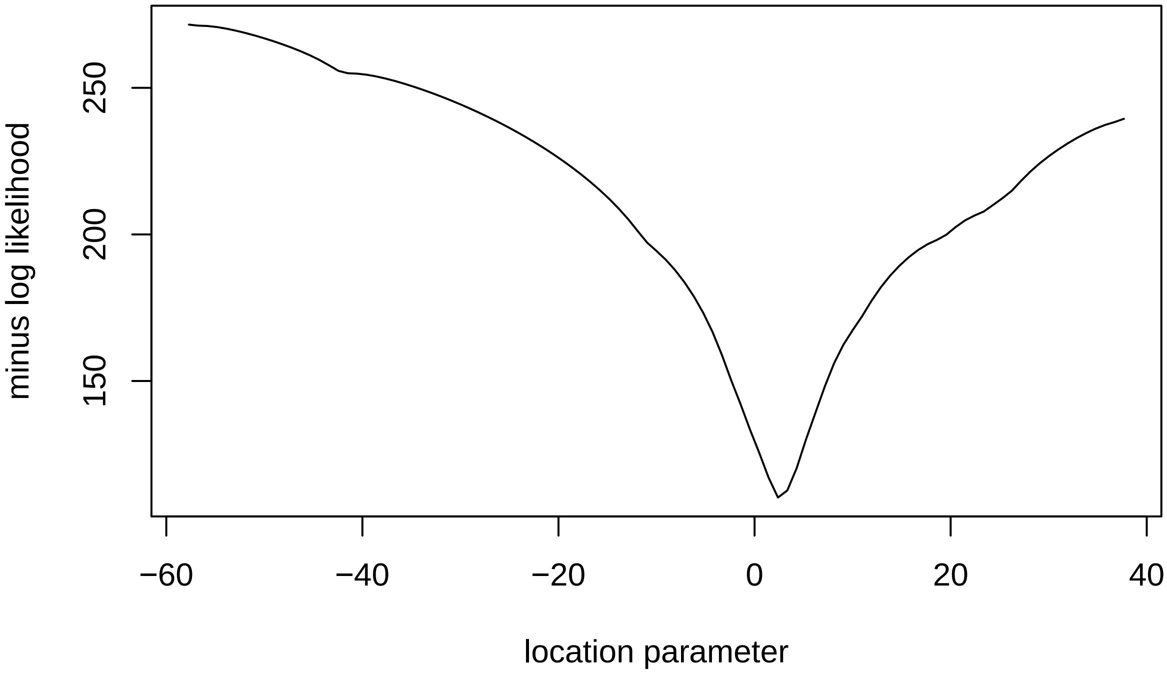
4.1 Maximum Likelihood Estimation

This log likelihood can be coded in R as follows.

```
mlogl <- function(mu)
    sum(- dcauchy(x, location = mu, log = TRUE))
```

This likelihood does not usually have a unique local maximum so we need a good starting point.

```
fran <- Vectorize(mlogl)
curve(fran, from = min(x), to = max(x), xlab = "location parameter",
    ylab = "minus log likelihood")
```



A "good enough" starting point is the sample median, which is consistent and asymptotically normal but not efficient (Ferguson (1996), Chapter 13). Starting there and going downhill on minus the log likelihood produces the efficient local MLE. Here's how to do that in R.

```
nout <- nlm(mlogl, median(x))
nout$code <= 2
```

```
## [1] TRUE
```

```
nout$estimate
```

```
## [1] 2.62256
```

```r
median(x)
```

```
## [1] 2.610152
```

If the optimizer converges (if `nout$code` ≤ 2), then `nout$estimate` is the efficient local MLE. It is crucial that we supply `median(x)` as the starting point, not some point that may be far away from the true unknown parameter value.

We wrote `mlogl` to calculate minus the log likelihood because `nlm` only minimizes functions (minimizing minus the log likelihood is the same as maximizing the log likelihood, and maximizing the log likelihood is the same as maximizing the likelihood).

Similar code to that above works if we have multiple parameters and make `mlogl` a function of a vector argument (more on this later).

We also show the median, just to show that the two estimators are different. Only theory tells us that the MLE is a better estimator than the median.

4.2 Asymptotic Distribution

Since global and local MLE have the same asymptotic distribution under "the usual regularity conditions" we can ignore the distinction when talking about this asymptotic distribution. If $\hat{\theta}_n$ is the MLE for sample size n and θ is the true unknown parameter value, then

$$\hat{\theta}_n \approx \text{Normal}(\theta, I_n(\theta)^{-1}) \tag{1}$$

where the double wiggle sign means "approximately distributed as" and

$$I_n(\theta) = -E_\theta\{\nabla^2 l_n(\theta)\}, \tag{2}$$

where

$$l_n(\theta) = \log L_n(\theta),$$

and L_n denotes the likelihood for sample size n (what was L_x before). The matrix (2) is called the *Fisher information matrix.*

In this example with only one parameter, the matrix is one-by-one, so can be treated as a single number rather than a matrix. But we will need Fisher information as a matrix by the end of these notes.

4.3 Plug-In

The fundamental equation (1), although theoretically important, is practically useless because we don't know θ (it is the true unknown parameter value we are trying to estimate). To use (1) we have to "plug in" a consistent estimator for θ in the asymptotic variance, and the obvious estimator is the MLE. This gives

$$\hat{\theta}_n \approx \text{Normal}(\theta, I_n(\hat{\theta}_n)^{-1}), \tag{3}$$

and hence, if θ is a scalar parameter,

$$\hat{\theta}_n \pm 1.96\sqrt{I_n(\hat{\theta}_n)^{-1}} \tag{4}$$

is a 95% approximate large-sample confidence for θ and

$$\frac{\hat{\theta}_n - \theta}{\sqrt{I_n(\hat{\theta}_n)^{-1}}} \tag{5}$$

is an approximately standard normal quantity that can be used as a test statistic for constructing hypothesis tests.

Sometimes calculating the expectations in the definition of Fisher information (2) is onerous. Then we use another consistent estimator of asymptotic variance. For large n, under the "usual regularity conditions" for maximum likelihood

$$I_n(\hat{\theta}_n) \approx J_n(\hat{\theta}_n)$$

where

$$J_n(\theta) = -\nabla^2 l_n(\theta) \tag{6}$$

(that is, (2) and (6) are the same except that in (6) we don't take the expectation). The matrix (6) is called the *observed information matrix*. Some people, including your humble author, call (2) the *expected Fisher information matrix* and call (6) the *observed Fisher information matrix*. Thus we can replace (3), (4), and (5) with

$$\hat{\theta}_n \approx \text{Normal}(\theta, J_n(\hat{\theta}_n)^{-1}), \tag{7}$$

$$\hat{\theta}_n \pm 1.96\sqrt{J_n(\hat{\theta}_n)^{-1}}, \tag{8}$$

and

$$\frac{\hat{\theta}_n - \theta}{\sqrt{J_n(\hat{\theta}_n)^{-1}}}. \tag{9}$$

The confidence intervals and hypothesis tests recommended in this section are Wald intervals and tests because they use Fisher information evaluated at $\hat{\theta}_n$ rather than at θ.

If we did use $I_n(\theta)$ and $J_n(\theta)$ we would get Rao confidence intervals and hypothesis tests.

4.4 Wald Interval

The log likelihood is messy. We let R calculate it for us. We can also get R to calculate $J_n(\hat{\theta}_n)$ by using optional argument `hessian` when doing maximum likelihood.

```
nout <- nlm(mlogl, median(x), hessian = TRUE)
nout$code <= 2
```

```
## [1] TRUE
```

```
nout$estimate
```

```
## [1] 2.62256
```

```
nout$hessian
```

```
##          [,1]
## [1,] 13.27041
```

R function `mlogl` calculates minus the log likelihood, the Hessian is the second derivative of this function, observed Fisher information is minus the second derivative; and the two minus signs cancel.

Thus the Wald interval is

```
conf.level <- 0.95
crit <- qnorm((1 + conf.level) / 2)
crit
```

```
## [1] 1.959964
```

```
nout$estimate + c(-1, 1) * crit / sqrt(nout$hessian)
```

```
## Warning in c(-1, 1) * crit/sqrt(nout$hessian): Recycling array of length 1 in vector-array arithmeti
##   Use c() or as.vector() instead.
```

```
## [1] 2.084531 3.160590
```

Arrrrgh! That is a truly annoying error message

```
nout$estimate + c(-1, 1) * crit / sqrt(c(nout$hessian))
```

```
## [1] 2.084531 3.160590
```

The formula we used for the Wald interval is (8) above or (18) below.

4.5 Likelihood Interval

We have only an implicit formula for the likelihood interval, (19) below. Thus we must solve equations with R function uniroot to find the endpoints of the interval.

The likelihood interval goes down from the maximum on the likelihood an amount determined by the chi-square critical value

```
crit <- qchisq(conf.level, df = 1)
crit
```

```
## [1] 3.841459
```

```
sqrt(crit)
```

```
## [1] 1.959964
```

Of course, chi-squared with one degree of freedom is the square of a standard normal, hence the relation between the critical value here and in the preceding section.

```
theta.hat <- nout$estimate
fred <- function(theta) 2 * (mlogl(theta) - mlogl(theta.hat)) - crit

uout <- uniroot(fred, interval = c(theta.hat - 1, theta.hat),
    extendInt = "downX")
lower <- uout$root
uout <- uniroot(fred, interval = c(theta.hat, theta.hat + 1),
    extendInt = "upX")
upper <- uout$root
c(lower, upper)
```

```
## [1] 2.069188 3.202467
```

4.6 Rao (Score) Interval

We have only an implicit formula for the score interval, (20) below. Thus we must solve equations with R function uniroot to find the endpoints of the interval.

We will use observed rather than expected Fisher information, so the interval is the set of θ values satisfying

$$\frac{\left(l_n'(\theta)\right)^2}{J_n(\theta)} \le \text{critical value}$$

or

$$\frac{\left(-l_n'(\theta)\right)^2}{-l_n''(\theta)} \le \text{critical value}$$

But now we have the additional problem that we don't know the first and second derivatives of the log likelihood function. (The computer knows. R function `nlm` is calculating them using finite-difference approximation. But it doesn't give us access to them.) R function `nlm` only returns the second derivative evaluated at $\hat{\theta}_n$, which is not what we need here.

Hence we turn to R package `numDeriv` to calculate these derivatives.

```
library(numDeriv)
mlogl.first <- function(theta) grad(mlogl, theta)
mlogl.second <- function(theta) hessian(mlogl, theta)

# try them out
mlogl.first(theta.hat)
```

```
## [1] -1.744026e-05
```

```
mlogl.second(theta.hat)
```

```
##          [,1]
## [1,] 13.27127
```

```
# avoid that annoying warning
mlogl.second <- function(theta) c(hessian(mlogl, theta))
mlogl.second(theta.hat)
```

```
## [1] 13.27127
```

Now we are ready to try to compute the endpoints of the interval

```
fred <- function(theta) mlogl.first(theta)^2 / mlogl.second(theta) - crit

uout <- uniroot(fred, interval = c(theta.hat - 1, theta.hat),
    extendInt = "downX")
lower <- uout$root
uout <- uniroot(fred, interval = c(theta.hat, theta.hat + 1),
    extendInt = "upX")
upper <- uout$root
c(lower, upper)
```

```
## [1] 2.129046 3.111660
```

4.7 Summary

	lower	upper
Wald	2.084531	3.160589
Wilks	2.069188	3.202467
Rao	2.129046	3.111660

Similar, as they must be for large sample size (asymptotic equivalence).

But the Wald interval is much easier to do and to explain.

5 Two-Parameter Example

We redo our Cauchy example with two parameters, that is, we assume both location and scale are unknown parameters.

```
mlogl <- function(theta) {
    stopifnot(length(theta) == 2)
    stopifnot(is.numeric(theta))
    stopifnot(is.finite(theta))
    sum(- dcauchy(x, location = theta[1],
        scale = theta[2], log = TRUE))
}
```

```
theta.twiddle <- c(median(x), IQR(x) / 2)
nout <- nlm(mlogl, theta.twiddle)
nout$code <= 2
```

```
## [1] TRUE
```

```
nout$estimate
```

```
## [1] 2.649775 2.067405
```

```
c(median(x), IQR(x) / 2)
```

```
## [1] 2.610152 1.937020
```

The MLE is now two-dimensional, first component location and second component scale. As before we use the sample median as a "good enough" estimator of the location parameter. Since the interquartile range (IQR) of the Cauchy distribution is twice the scale parameter, IQR / 2 is a consistent and "good enough" estimator of the scale.

We output both the MLE (a two-dimensional vector) and our "good enough" estimator that we used as a starting point only to show that they are different. Only theory tells us that the MLE is the better estimator.

If we ask nicely, the computer will also calculate the observed Fisher information matrix evaluated at the MLE.

```
nout <- nlm(mlogl, theta.twiddle, hessian = TRUE)
nout$code <= 2
```

```
## [1] TRUE
```

```
nout$hessian
```

```
##             [,1]        [,2]
## [1,] 4.399394073 0.002009904
## [2,] 0.002009904 2.618737940
```

As in the one-dimensional case we can use this to easily make Wald intervals.

Standard errors for the estimators are square roots of the diagonal elements of inverse Fisher information

```
se <- sqrt(diag(solve(nout$hessian)))
foo <- cbind(nout$estimate, se)
rownames(foo) <- c("location", "scale")
colnames(foo) <- c("estimate", "std. err.")
round(foo, 4)
```

```
##          estimate std. err.
## location   2.6498    0.4768
## scale      2.0674    0.6180
```

Rather than demonstrate Wilks or Rao intervals for this example, we will leave that for a categorical data analysis example in the follow-on notes. But Rao turned out to be too horrible to do, so only likelihood intervals are illustrated there. Rao would be horrible in the same way for this example. Too much calculus.

6 Wald, Wilks, Rao

6.1 Nested Models

There are three kinds of tests of model comparison associated with maximum likelihood estimation. These tests have three different test statistics but are asymptotically equivalent (a concept explained below). The idea is that we have two nested models to compare, "nested" meaning one is a submodel of the other.

The simplest situation in which we get nested models is when some of the parameters in the bigger model are set to zero (constrained to be zero) in the smaller model. When models are specified by regression formulas, as when using R functions like `lm` and `glm`, this happens when all of the terms in the formula for the smaller model are in the formula for the bigger model, but not vice versa.

More generally, for regression-like models (linear and generalized linear models, for example) models are nested if the column space of the model matrix of the bigger model contains the column space of the model matrix of the smaller model, but not vice versa (and offset vectors are the same, if there are offset vectors).

More generally, for general statistical models (not necessarily regression-like), models are nested if every probability distribution in the smaller model is also in the bigger model but not vice versa. This happens for parametric statistical models when bigger and smaller models are parameterized the same way and the parameter space of the bigger model contains the parameter space of the smaller model but not vice versa.

It is not enough for Wald, Wilks, and Rao tests that the models are nested; they must be nested smoothly, meaning the parameter space of the bigger model is a manifold (a possibly curved hypersurface in n-dimensional space and the parameter space of smaller model is a submanifold of the parameter space of the bigger model. We won't even try to explain that condition, but be content to explain the special case when some of the parameters of the bigger model are constrained to be equal to zero in the smaller model. Then the submanifold requirement is that each parameter constrained to be equal to zero in the smaller model can take any value in an open interval that includes zero in the bigger model and can do this regardless of the values of the other parameters.

That is, we are doing two-sided rather than one-sided tests of the constrained parameters (collectively). If we were doing the multiparameter analog of one-sided tests in which some parameters are equal to zero in the smaller model but greater than or equal to zero in the bigger model (variance components in random effects models, for example), then the theory for this is known but not widely understood and hard to use (Chernoff, 1954; Geyer, 1994; Le Cam, 1970; Self and Liang, 1987).

6.2 The Tests

We also need the "usual regularity conditions" for asymptotic normality of maximum likelihood estimates in the bigger and smaller models, which we denote $\hat{\theta}_n$ and $\tilde{\theta}_n$, respectively.

Then the Wilks test statistic, also called the likelihood ratio test statistic, is

$$T_n = 2[l_n(\hat{\theta}_n) - l_n(\tilde{\theta}_n)] \tag{10}$$

where l_n is the log likelihood for the bigger model, which means that $\hat{\theta}_n$ and $\tilde{\theta}_n$ have the same dimension (but some of the components of $\tilde{\theta}_n$ are equal to zero). And Wilks's theorem says (assuming the "usual regularity conditions and the submanifold nesting condition)

$$T_n \approx \text{ChiSq}(p - d),$$

where p is the dimension of the parameter space of the model and d is the dimension of the parameter space of the submodel.

It is sometimes easy to calculate one of $\hat{\theta}_n$ and $\tilde{\theta}_n$ and difficult or impossible to calculate the other. This motivates two other procedures that are asymptotically equivalent to the Wilks test.

The Rao test statistic is

$$R_n = \left(\nabla l_n(\tilde{\theta}_n)\right)^T I_n(\tilde{\theta}_n)^{-1} \nabla l_n(\tilde{\theta}_n), \tag{11}$$

where $I_n(\theta)$ is expected Fisher information for sample size n given by (2). We can also replace expected Fisher information by observed Fisher information

$$R_n' = \left(\nabla l_n(\tilde{\theta}_n)\right)^T J_n(\tilde{\theta}_n)^{-1} \nabla l_n(\tilde{\theta}_n), \tag{12}$$

where $J_n(\theta)$ is observed Fisher information for sample size n given by (6). We call either (11) or (12) the Rao test statistic.

Under the conditions for the Wilks theorem, the Rao test statistic is asymptotically equivalent to the Wilks test statistic. Both $T_n - R_n$ and $T_n - R'_n$ converge in distribution to zero as n goes to infinity, so the differences between the three test statistics T_n, R_n, and R'_n are negligible compared to their values for large sample sizes n.

The test using the statistic R_n called the *Rao test* or the *score test* or the *Lagrange multiplier test*.

An important point about the Rao test statistic is that, unlike the likelihood ratio test statistic, it only depends on the MLE for the null hypothesis $\tilde{\theta}_n$.

The Wald test statistic is

$$W_n = g(\hat{\theta}_n)^T \left[\nabla g(\hat{\theta}_n) I_n(\hat{\theta}_n)^{-1} \bigl(\nabla g(\hat{\theta}_n)\bigr)^T \right]^{-1} g(\hat{\theta}_n), \tag{13}$$

or

$$W'_n = g(\hat{\theta}_n)^T \left[\nabla g(\hat{\theta}_n) J_n(\hat{\theta}_n)^{-1} \bigl(\nabla g(\hat{\theta}_n)\bigr)^T \right]^{-1} g(\hat{\theta}_n), \tag{14}$$

where, as with the Rao statistic, $I_n(\theta)$ and $J_n(\theta)$ are expected and observed Fisher information given by (2) and (6) and where g is a vector-to-vector constraint function such that the submodel is the set of θ such that $g(\theta) = 0$. In the case of interest to us, where the smaller model sets some of the parameters of the bigger model to zero, it is usual to take $g(\theta)$ to be the vector of those constrained parameters, that is, $g(\theta)$ is the vector of values of the parameters that are constrained to be zero in the smaller model.

Under the conditions for the Wilks theorem, the Wald test statistic is asymptotically equivalent to the Wilks test statistic. Both $T_n - W_n$ and $T_n - W'_n$ converge in distribution to zero as n goes to infinity, so the differences between the five test statistics T_n, R_n, R'_n, W_n, and W'_n are negligible compared to their values for large sample sizes n.

An important point about the Wald test statistic is that, unlike the likelihood ratio test statistic, it only depends on the MLE for the alternative hypothesis $\hat{\theta}_n$.

6.3 Specializing to the One-Parameter Case

The formula (10) for the likelihood ratio test statistic does not simplify in the one-parameter case.

The formula (11) for the Rao test statistic does simplify to

$$R_n = \bigl(l'_n(\tilde{\theta}_n)\bigr)^2 \Big/ I_n(\tilde{\theta}_n) \tag{15}$$

and it simplifies even further in the case of a point null hypothesis $\theta = \theta_0$ to

$$R_n = \bigl(l'_n(\theta_0)\bigr)^2 \Big/ I_n(\theta_0) \tag{16}$$

And, of course, (12) simplifies to either of the above with I_n replaced by J_n.

How (13) simplifies depends on what the constraint function g is. If we use the simplest form $g(\theta) = \theta - \theta_0$, where θ_0 is the hypothesized value under the null hypothesis, then $g'(\theta) = 1$ and (13) simplifies to

$$W_n = (\hat{\theta}_n - \theta_0)^2 I_n(\hat{\theta}_n) \tag{17}$$

And, of course, (14) simplifies to the above with I_n replaced by J_n.

Inverting these hypothesis tests gives confidence intervals.

The Wald interval is

$$\hat{\theta}_n \pm \frac{c}{\sqrt{I_n(\hat{\theta}_n)}} \tag{18}$$

where c is the critical value derived from the standard normal distribution (the square root of the critical value derived from the chi-squared distribution with one degree of freedom).

The other two intervals have no explicit form. They are given implicitly as follows.

The likelihood interval (Wilks interval) is the set of θ values satisfying

$$2[l_n(\hat{\theta}_n) - l_n(\theta)] \leq c \tag{19}$$

where c is the critical value derived from the chi-squared distribution with one degree of freedom.

The score interval (Rao interval) is the set of θ values satisfying

$$\frac{\left(l_n'(\theta)\right)^2}{I_n(\theta)} \leq c \tag{20}$$

where c is the critical value derived from the chi-squared distribution with one degree of freedom. And, as always, we can replace I_n with J_n in the above.

6.4 Applications

All five of these are widely used. Here are some cases where T_n, R_n, and W_n are widely used.

T_n is the hardest to calculate, since either $\hat{\theta}_n$ or $\tilde{\theta}_n$ or both may require numerical optimization (like we saw with the Cauchy examples). But it is nevertheless widely used when such optimization has to be done anyway. The R generic function `anova` when applied to objects produced by the R function `glm` does likelihood ratio tests (which it calls analysis of deviance, "deviance" being another name for the likelihood ratio test statistic (10)). So everyone who does tests of model comparison using these R functions is doing likelihood ratio tests.

The Pearson chi-square test is a special case of the Rao test. So everyone who does tests of model comparison using the Pearson chi-square test is doing Rao tests (also called score tests). Since this is a very old and widely understood procedure, many users use it (in all statistical software, not just R).

The test statistics and P-values in the output of the R generic function `summary` applied to objects produced by the R function `glm` are Wald tests. This is obvious from the fact that only the MLE in the big model is computed. The relevant smaller models are those that drop the predictor for the line of the output having a particular test statistic and P-value. Those models are not fitted in order to print the output of the `summary` function. So everyone who pays attention to test statistics and P-values in the output of the R function `summary` is doing Wald tests.

An example where R_n' or W_n' would be used is when we use the computer to calculate Fisher information (the `hessian` component of the output of the R function `nlm` as in the Cauchy example). That is observed Fisher information because the computer only knows how to approximate derivatives but does not know how to do expectations. If we use such to make confidence intervals (as we did in the Cauchy example), those are Wald confidence intervals (obtained by inverting Wald tests).

6.5 Binomial Examples

Recall that the log likelihood for a binomial distribution with sample size n and parameter π is

$$l_n(\pi) = x \log(\pi) + (n - x) \log(1 - \pi),$$

the score (first derivative) is

$$\begin{aligned} l_n'(\pi) &= \frac{x}{\pi} - \frac{n - x}{1 - \pi} \\ &= \frac{x - n\pi}{\pi(1 - \pi)}, \end{aligned}$$

the MLE is

$$\hat{\pi}_n = x/n,$$

observed Fisher information is

$$J_n(\pi) = \frac{x}{\pi^2} + \frac{n - x}{(1 - \pi)^2},$$

and expected Fisher information is

$$I_n(\pi) = \frac{n}{\pi(1 - \pi)}.$$

We consider two-tailed tests with hypotheses

$$\begin{aligned} H_0 &: \pi = \pi_0 \\ H_1 &: \pi \neq \pi_0 \end{aligned}$$

The MLE in the smaller model is π_0 because there are no free parameters and nothing to estimate: the parameter is constrained to have this value.

6.5.1 Wilks

The Wilks (likelihood ratio) test statistic is

$$\begin{aligned} T_n &= 2\big[x \log(\hat{\pi}_n) + (n - x) \log(1 - \hat{\pi}_n)\big] - 2\big[x \log(\pi_0) + (n - x) \log(1 - \pi_0)\big] \\ &= 2 \left[x \log\left(\frac{\hat{\pi}_n}{\pi_0}\right) + (n - x) \log\left(\frac{1 - \hat{\pi}_n}{1 - \pi_0}\right) \right] \end{aligned}$$

6.5.2 Rao

Observed and expected Fisher information are different after plug-in of the MLE for the smaller model

$$\begin{aligned} I_n(\pi_0) &= \frac{n}{\pi_0(1 - \pi_0)} \\ J_n(\pi_0) &= \frac{x}{\pi_0^2} + \frac{n - x}{(1 - \pi_0)^2} \\ &= \frac{n\hat{\pi}_n}{\pi_0^2} + \frac{n(1 - \hat{\pi}_n)}{(1 - \pi_0)^2} \end{aligned}$$

so R_n and R_n' are different

$$\begin{aligned} R_n &= \left(\frac{x - n\pi_0}{\pi_0(1 - \pi_0)}\right)^2 \frac{\pi_0(1 - \pi_0)}{n} \\ &= \frac{(x - n\pi_0)^2}{n\pi_0(1 - \pi_0)} \\ R_n' &= \left(\frac{x - n\pi_0}{\pi_0(1 - \pi_0)}\right)^2 \left(\frac{n\hat{\pi}_n}{\pi_0^2} + \frac{n(1 - \hat{\pi}_n)}{(1 - \pi_0)^2}\right)^{-1} \\ &= \frac{(x - n\pi_0)^2}{n\hat{\pi}_n(1 - \pi_0)^2 + n(1 - \hat{\pi}_n)\pi_0^2} \end{aligned}$$

6.5.3 Wald

Observed and expected Fisher information are the same after plug-in of the MLE for the bigger model

$$\begin{aligned} I_n(\hat{\pi}_n) &= \frac{n}{\hat{\pi}_n(1 - \hat{\pi}_n)} \\ J_n(\hat{\pi}_n) &= \frac{x}{\hat{\pi}_n^2} + \frac{n - x}{(1 - \hat{\pi}_n)^2} \\ &= \frac{n\hat{\pi}_n}{\hat{\pi}_n^2} + \frac{n(1 - \hat{\pi}_n)}{(1 - \hat{\pi}_n)^2} \\ &= \frac{n}{\hat{\pi}_n(1 - \hat{\pi}_n)} \end{aligned}$$

so W_n and W_n' are the same. The function g we want to use for this application of Wald tests is

$$g(\pi) = \pi - \pi_0,$$

which has derivative

$$\frac{dg(\pi)}{d\pi} = 1,$$

so

$$\begin{aligned} W_n &= (\hat{\pi}_n - \pi_0)^2 \left(I_n(\hat{\pi}_n)^{-1} \right)^{-1} \\ &= \frac{n(\hat{\pi}_n - \pi_0)^2}{\hat{\pi}_n(1 - \hat{\pi}_n)} \end{aligned}$$

If we try these out with $\pi_0 = 1/3$, $n = 200$, and $x = 53$, we obtain

$$\begin{aligned} T_n &= 4.3708 \qquad (P = 0.0366) \\ R_n &= 4.2025 \qquad (P = 0.0404) \\ R_n' &= 4.6825 \qquad (P = 0.0305) \\ W_n &= 4.7947 \qquad (P = 0.0285) \end{aligned}$$

We see the asymptotic equivalence in the fact that all four tests give nearly the same results.

The test recommended by introductory statistics books has test statistic

$$Z_n = \frac{\hat{\pi}_n - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}$$

which has an asymptotic standard normal distribution. Since $R_n = Z_n^2$ and the square of a standard normal random variable is a chi-square distribution with one degree of freedom, which is the reference distribution for Wilks, Rao, and Wald tests in this situation with dimension zero for the smaller model and dimension one for the bigger model, the two tests are the same. So the test recommended by introductory statistics books for this situation is the Rao test.

Bibliography

Chernoff, H. (1954) On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, **25**, 573–578.

Cramér, H. (1946) *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Ferguson, T. S. (1996) *A Course in Large Sample Theory*. London: Chapman & Hall.

Fisher, R. A. (1912) On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, **41**, 155–160.

Fisher, R. A. (1922) On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, **222**, 309–368.

Freedman, D., Pisani, R. and Purves, R. (2007) *Statistics*. fourth. New York: W. W. Norton & Company.

Geyer, C. J. (1994) On the asymptotics of constrained M-estimation. *Annals of Statistics*, **22**, 1993–2010.

Geyer, C. J. (2013) Asymptotic of maximum likelihood without the LLN or CLT or sample size going to infinity. In *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton* (eds G. L. Jones and X. Shen), pp. 1–24. IMS collections. Hayward, CA: Institute of Mathematical Statistics.

Kiefer, J. and Wolfowitz, J. (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, **27**, 887–906.

Le Cam, L. and Yang, G. L. (2000) *Asymptotics in Statistics: Some Basic Concepts*. Second. New York: Springer-Verlag.

LeCam, L. (1953) On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *University of California Publications in Statistics*, **1**, 277–329.

Le Cam, L. (1970) On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Annals of Mathematical Statistics*, **41**, 802–828.

Self, S. G. and Liang, K.-Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

van der Vaart, A. W. (1998) *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20**, 595–601.

Wang, J.-L. (1985) Strong consistency of approximate maximum likelihood estimators with applications in nonparametrics. *Annals of Statistics*, **13**, 932–946.