

Why Likelihood Intervals Work: A Graphical Guide

NRES 746 reference notes on the score, Fisher information, and the sampling distribution of the MLE

Two recipes turn a fitted likelihood into a 95% confidence interval for a parameter. The first takes the maximum likelihood estimate (MLE) plus or minus 1.96 standard errors, with the standard error computed as $1/\sqrt{\text{curvature}}$ of the negative log-likelihood at its minimum. The second keeps every parameter value whose negative log-likelihood is within 1.92 units of the minimum. These notes explain why both recipes work: why the standard error is $1/\sqrt{\text{curvature}}$, why the MLE is approximately normal, and where the 1.92 comes from. The approach is graphical; the only math needed is a derivative and the central limit theorem. Bolker's *Ecological Models and Data in R* shows how to compute these intervals (Sections 6.4 and 6.5); these notes focus on why they work.

The running example is a survey of American pikas, small mammals that live in rock fields (talus) and store vegetation in haypiles for the winter. A survey counts active haypiles in each of n talus plots, and each count is Poisson with mean λ . To see how estimates behave across repeated surveys, we simulate surveys from a known truth, $\lambda_0 = 3$.

1. Two views of the same curve

For one survey, the negative log-likelihood (NLL) is a valley whose bottom is the MLE, $\hat{\lambda}$. Its slope, with the sign flipped so that it points uphill on the log-likelihood, is called the **score**:

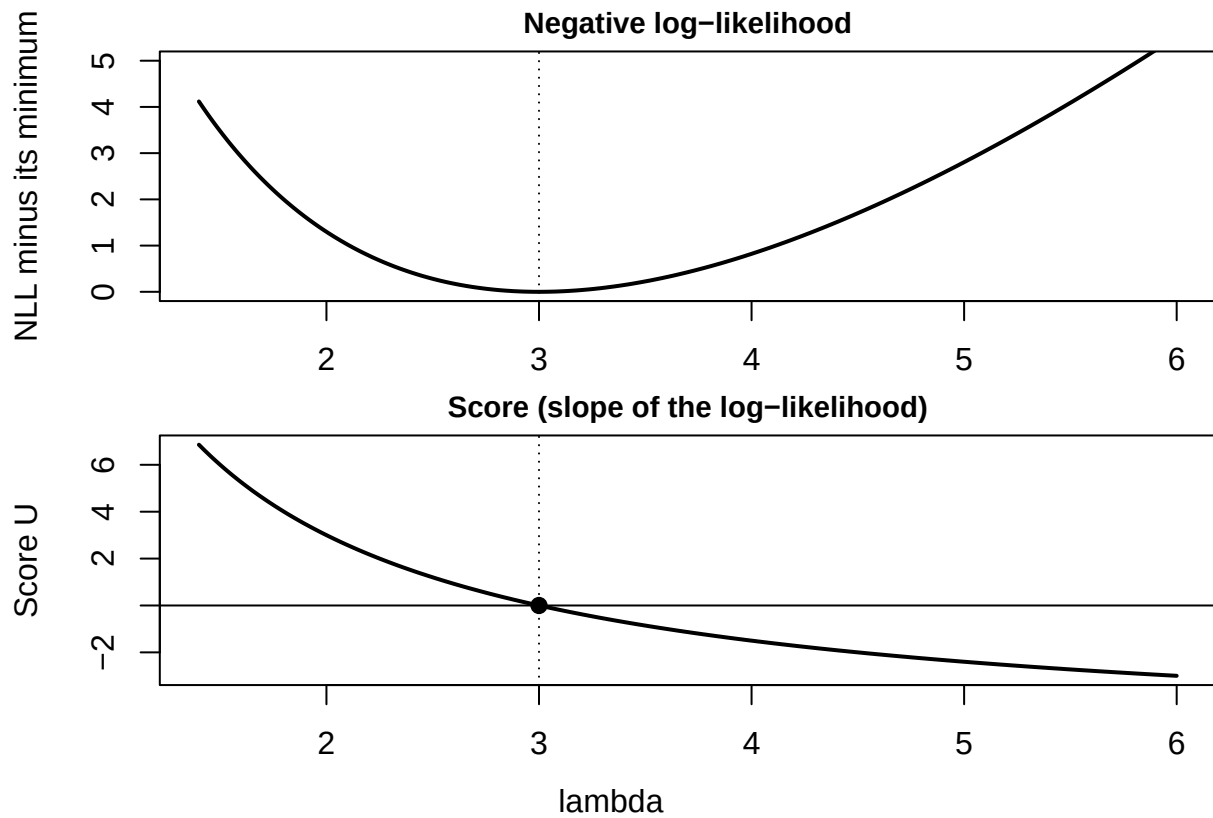
$$U(\lambda) = \frac{d}{d\lambda} \log \mathcal{L}(\lambda) = -\frac{d}{d\lambda} \text{NLL}(\lambda).$$

For the Poisson, $U(\lambda) = \sum y_i/\lambda - n$.

For the data 2, 5, 3, 0, 4, 4 (figure below):

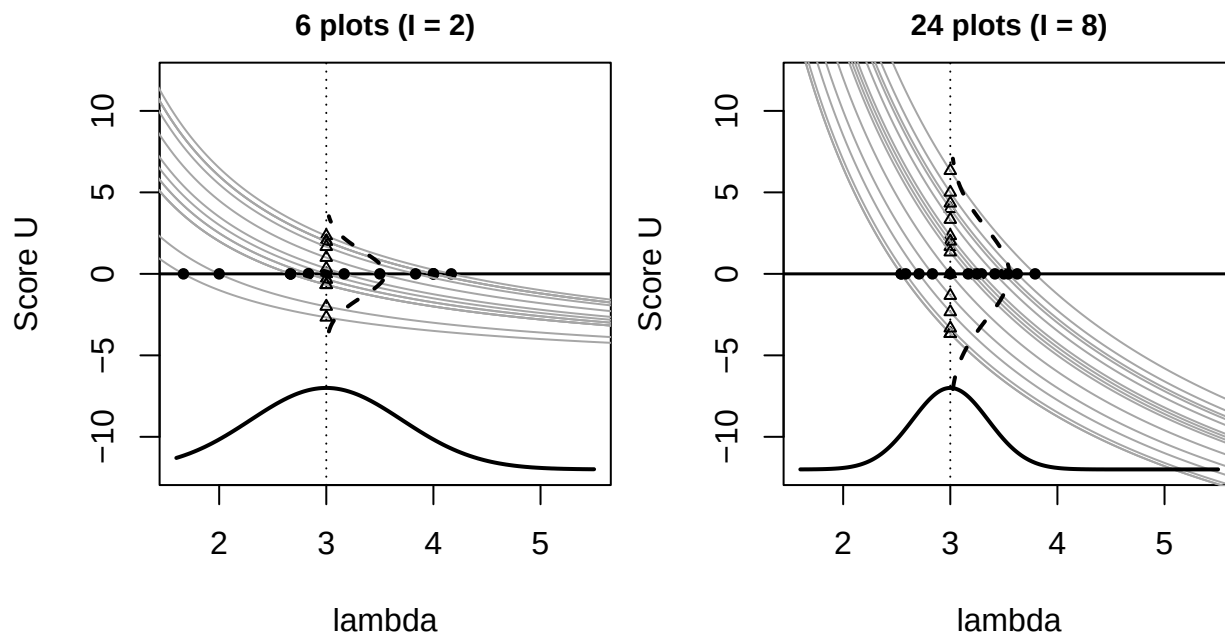
- The bottom of the valley (top panel) is where the score crosses zero (bottom panel). Both mark the MLE, $\hat{\lambda} = 3$.
- The **curvature** of the valley is the **steepness of the score**. A sharp valley means a steep score line; a broad valley means a shallow one.
- Near the MLE, a **parabola-shaped valley** and a **straight-line score** are the same statement. Everything below depends on it.

The expected steepness of the score (the expected curvature of the NLL) is the **Fisher information**, I . For the Poisson, $I = n/\lambda$.



2. Many surveys, many score curves

A new survey gives new counts, so a new score curve. The figure shows the score from 15 simulated surveys at two sample sizes (grey curves).

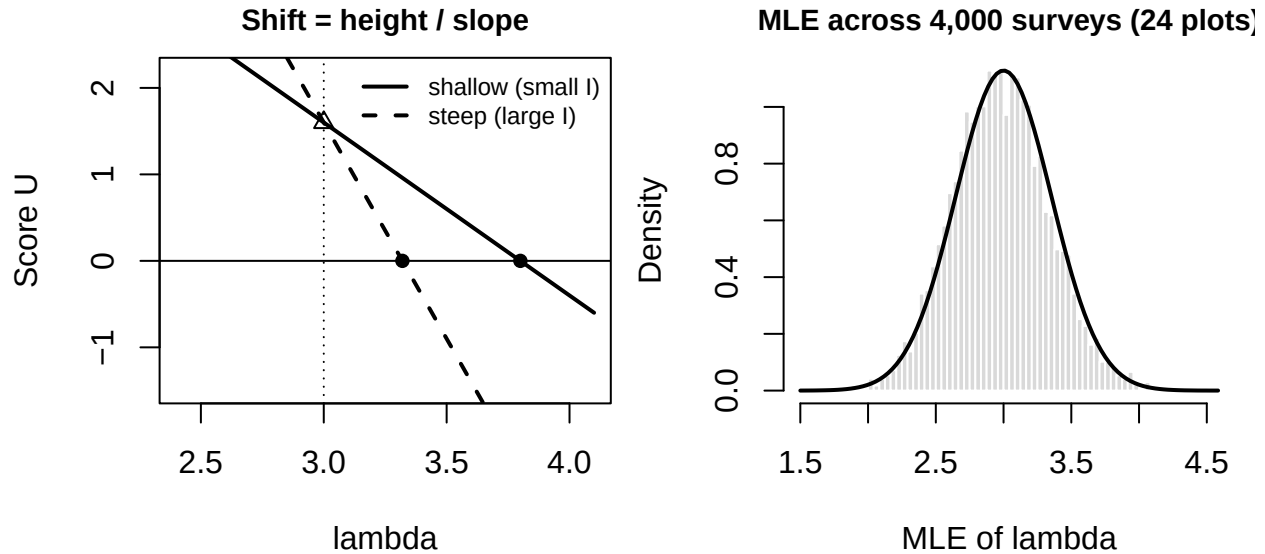


Read it in four steps.

1. **Each curve crosses zero at that survey's MLE** (filled circles). The crossings spread along the axis: that spread is the sampling distribution of $\hat{\lambda}$ (the solid bell along the bottom).
2. **At the true value, look up and down.** The open triangles on the dotted line are each curve's height at $\lambda_0 = 3$. They are centered on 0, because at the truth the log-likelihood is flat on average.
3. **Those heights are normal.** The score is a sum over plots, $U(\lambda_0) = \sum_i (y_i/\lambda_0 - 1)$, so the central limit theorem applies. Their variance turns out to be I (Section 4 explains why), drawn as the dashed bell lying on its side against the dotted line.
4. **Each curve slopes down at about $-I$.** Section 3 turns this into the spread of the filled circles.

3. Turning vertical spread into horizontal spread

Zoom in near the truth, where each score curve is close to a straight line with slope $-I$.



- **Left:** a curve that sits a height h above zero at the truth (open triangle) has to travel h/I to the right before it reaches zero (filled circles). A steeper line (dashed, larger I) gets there sooner, so its MLE lands closer to the truth.
- The heights have variance I . Dividing by the constant slope I gives the MLEs a variance of

$$\frac{\text{Var}(U)}{I^2} = \frac{I}{I^2} = \frac{1}{I}.$$

- **Right:** across 4,000 simulated surveys of 24 plots, the MLEs follow the predicted $\text{Normal}(\lambda_0, 1/I) = \text{Normal}(3, 3/24)$ (curve).

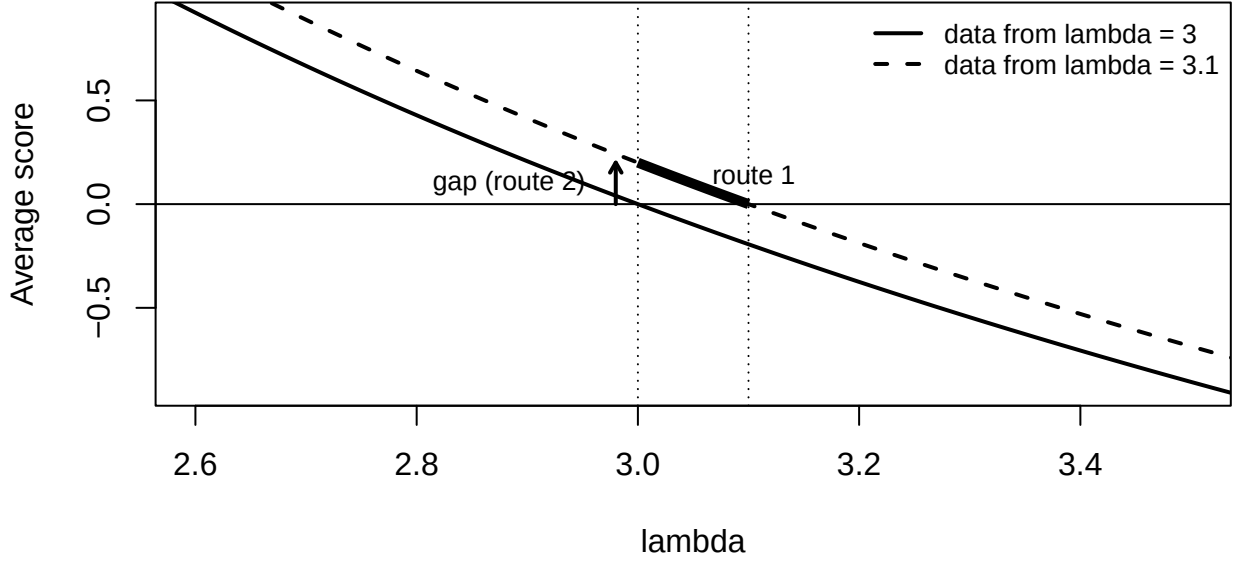
More data. With 24 plots instead of 6, the heights at the truth spread *more* (their variance I grows with n), but the lines get steeper just as fast (their slope is also I). Spread grows like $\sqrt{n}$, slope like n , so the MLEs tighten like $1/\sqrt{n}$. This also settles a common puzzle: more data lowers the maximum likelihood itself, but what matters for the estimate is the steepness, not the height.

Check with the pikas. Here the straight-line step is exact:

$$\hat{\lambda} - \lambda_0 = \frac{U(\lambda_0)}{n/\lambda_0} = \frac{\sum y_i/\lambda_0 - n}{n/\lambda_0} = \bar{y} - \lambda_0, \quad \text{Var}(\hat{\lambda}) = \frac{\lambda_0}{n} = \frac{1}{I}.$$

4. Why the variance of the score equals its slope

Section 3 needed the variance of the score and its slope to be the same number, I . This is the **information identity**, and the score curves show why it has to hold: **the curves must move when the truth moves.**



The solid curve is the average score when the data come from $\lambda_0 = 3$; it crosses zero at 3. Now nudge the truth by $\delta = 0.1$, to 3.1 (dashed curve). The new curve must cross zero at 3.1, or the MLE wouldn't track the truth. The **gap** is how far the average score at $\lambda = 3$ rises. There are two ways to get it.

Route 1: what the slope requires. The curves fall at slope $-I$, so to cross zero δ further right, the new curve must sit about $\delta \cdot I$ above zero at 3. For six plots, $I = 6/3 = 2$, so the gap should be $0.1 \times 2 = 0.2$.

Route 2: what reweighting the surveys delivers. Keep reading the score at $\lambda = 3$. For six plots with total count S , that's $U(3) = S/3 - 6$, a fixed formula applied to each survey. Changing the truth doesn't change the formula; it changes **how often each survey turns up**, and the average is a probability-weighted sum over surveys.

How much more likely does a survey get? The score is, by definition, how fast a survey's log-probability rises as λ increases. So a small step δ multiplies its probability by about $e^{\delta U} \approx 1 + \delta U$:

Total count S	Score at 3, U	Exact ratio	$1 + 0.1 U$
12	-2	0.813	0.8
18	0	0.990	1.0
24	2	1.206	1.2

Surveys whose score points toward a larger λ become more common, and surveys pointing lower become rarer. The new average score at 3 is then

$$\sum U \cdot P_3 \cdot (1 + \delta U) = \underbrace{E[U]}_0 + \delta E[U^2] = \delta \cdot \text{Var}(U).$$

The variance appears because reweighting only moves the average when surveys disagree. If every survey had a score of 0, tilting the weights would change nothing; the more the scores spread, the more the average moves. For six plots, $\text{Var}(U) = \text{Var}(S)/9 = 18/9 = 2$, so the gap is again $0.1 \times 2 = 0.2$.

Both routes describe the same gap, so $\delta \cdot I = \delta \cdot \text{Var}(U)$, which means $\text{Var}(U) = I$. In the figure, the gap is 0.2.

The path in the figure. Start where the solid curve crosses zero, at 3, and get to where the dashed curve crosses zero, at 3.1, in two moves: up the gap (switching the truth from 3 to 3.1 while still reading the score at 3; route 2), then down along the dashed curve from 3 to 3.1 (moving the reading point with the truth fixed; route 1, the thick segment). The path starts and ends on the zero line, so the two moves cancel: $\delta \cdot \text{Var}(U) - \delta \cdot I = 0$. The calculus version below is this path, made exact.

The calculus version

Write $f(y; \lambda)$ for the probability of a whole survey y (all six counts) when the true mean is λ . Because the counts are discrete, “averaging over surveys” means summing over every possible survey, weighted by f . The score is

$$U(y; \lambda) = \frac{d}{d\lambda} \log f(y; \lambda) = \frac{f'}{f}, \quad \text{so} \quad f' = Uf,$$

where a prime means a derivative with respect to λ .

Step 1: the score averages to zero at the truth. The claim is that the expected score, a sum over every possible outcome weighted by its probability, is zero when the score is read at the true λ :

$$E[U] = \sum_y U(y; \lambda) f(y; \lambda) = 0.$$

One count first. For a single plot, $f(y; \lambda) = \lambda^y e^{-\lambda} / y!$ for $y = 0, 1, 2, \dots$, and the log is $y \log \lambda - \lambda - \log y!$, so the score for one count is $U = y/\lambda - 1$. A count of 3 says nothing ($U = 0$ at $\lambda = 3$), a count of 6 says “ λ is higher” ($U = 1$), and a count of 0 says “ λ is lower” ($U = -1$). Weighting each possible count by its probability at $\lambda = 3$:

Count y	Probability $f(y; 3)$	Score $U = y/3 - 1$	U times f
0	0.050	-1.000	-0.050
1	0.149	-0.667	-0.100
2	0.224	-0.333	-0.075
3	0.224	0.000	0.000
4	0.168	0.333	0.056
5	0.101	0.667	0.067
6	0.050	1.000	0.050
7+			0.050

The negative terms add to -0.224 and the positive terms to +0.224. Low counts point down, high counts point up, and weighted by how often each turns up, they cancel. (Directly: $E[U] = E[y]/3 - 1 = 3/3 - 1 = 0$, since a Poisson's mean is λ .)

Why it has to be zero, for any model. Whatever λ is, the probabilities of all possible outcomes add to 1: $\sum_y f(y; \lambda) = 1$. Something that stays at 1 has derivative 0, so

$$\sum_y f' = \sum_y U f = 0.$$

Each f' is how fast one outcome's probability changes as λ increases. Some go up (high counts) and some go down (low counts), but the total can't change, so the changes cancel. Since $f' = U f$, the last column of the table is f' : the change in each count's probability per unit of λ . So "the score averages to zero" is another way of saying "probabilities always add to 1." This is why the open triangles in Section 2 are centered on 0.

The condition that matters: "at the truth." Nothing in the algebra above fixes λ : the substitution $f' = U f$ is just the definition of U , so $\sum_y U(y; \lambda) f(y; \lambda) = 0$ holds for every λ . What it requires is that the **same** λ appears in the score and in the weights. The truth enters when we read the sum as an average over real surveys: the weights $f(y; \lambda)$ are how often each survey turns up only if λ is the value that generated the data. So the result says "if the data come from λ , the score read at that same λ averages to zero," and for real data that means reading the score at the true value.

If the two don't match, the argument doesn't apply: $U(y; a) f(y; b)$ with $a \neq b$ isn't the derivative of anything that adds to 1, and the average isn't zero. Read the score at 3 when the data come from 3.1 and the average is $3.1/3 - 1 = 0.033$. That is the gap from route 2 for one count, $\delta \cdot \text{Var}(U) = 0.1 \times 1/3$ (for one Poisson count, $\text{Var}(U) = I = 1/\lambda$). For the six-plot survey every quantity is six times larger, which gives the gap of 0.2 and $I = 2$ in the figure. From here on, y is the whole survey again.

Step 2: separate the two jobs λ is doing. In $\sum_y U(y; \lambda) f(y; \lambda)$, λ appears twice: once as the point where the score is read (the horizontal axis of the figure), and once as the truth that generates the surveys (which curve we're on). Give them separate names:

$$g(a, b) = \sum_y U(y; a) f(y; b) = \text{average score read at } a, \text{ when the data come from } b.$$

The solid curve is $g(a, 3)$ and the dashed curve is $g(a, 3.1)$, each plotted against a . Step 1 says $g(\lambda, \lambda) = 0$ for every λ : each curve crosses zero at its own truth.

Step 3: move both together. As λ increases, $g(\lambda, \lambda)$ stays at 0, so its rate of change is 0. A change in $g(\lambda, \lambda)$ comes from two sources, the change in a and the change in b , and these add (the chain rule for a function of two inputs, each changed with the other held fixed):

$$0 = \underbrace{\sum_y U'(y; \lambda) f(y; \lambda)}_{\text{change } a, \text{ hold } b} + \underbrace{\sum_y U(y; \lambda) f'(y; \lambda)}_{\text{change } b, \text{ hold } a}.$$

- **The first term is route 1.** With the truth held fixed, it's the rate at which the average score falls as the reading point moves: the slope of a curve in the figure. It equals $E[U']$, the average slope of the score, which is $-I$.
- **The second term is route 2.** With the reading point held fixed, it's how fast the average score rises as the truth moves: the gap per unit δ . Substituting $f' = Uf$ turns it into $\sum_y U \cdot Uf = E[U^2] = \text{Var}(U)$ (the last step because the score averages to 0). The substitution is the reweighting from route 2: $f' = Uf$ says a step δ multiplies each survey's probability by about $1 + \delta U$.

So $0 = -I + \text{Var}(U)$, which is the identity. Many texts write Step 3 in one line, differentiating $\sum Uf = 0$ with the product rule: $\sum(U'f + Uf') = 0$. The product rule's two terms are the two routes.

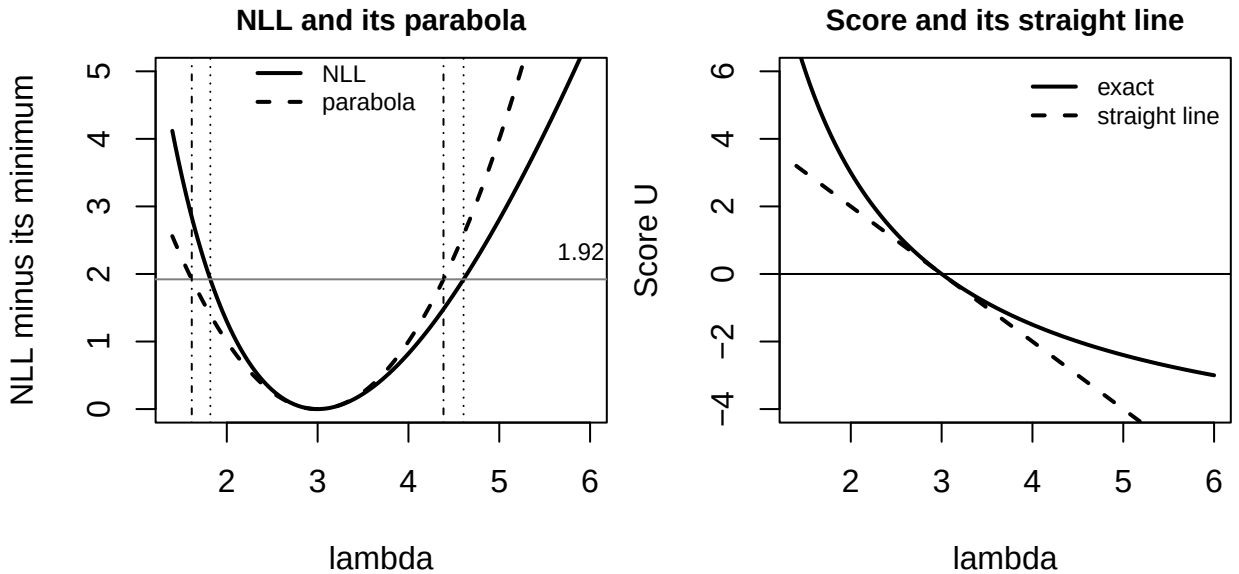
Check with the pikas. For six plots, the average score read at a when the truth is b is $g(a, b) = 6b/a - 6$ (the curves in the figure). Holding $b = 3$, the slope in a at $a = 3$ is $-6b/a^2 = -2$ (route 1, $-I$). Holding $a = 3$, the rate in b is $6/a = 2$ (route 2, $\text{Var}(U)$). They cancel.

5. The parabola, the interval, and the 1.92 cutoff

Sections 2 and 3 described many surveys. In practice we have one, and we use the shape of its NLL near the MLE. The second-order Taylor expansion around $\hat{\lambda}$ (the first derivative is zero there) is

$$\text{NLL}(\lambda) \approx \text{NLL}(\hat{\lambda}) + \frac{1}{2} I (\lambda - \hat{\lambda})^2,$$

a parabola whose curvature is the slope of the score.



- **The interval from the parabola.** Setting the rise equal to the cutoff, $\frac{1}{2} I d^2 = 1.92$, gives $d = 1.96/\sqrt{I}$: the Hessian (Wald) interval, $\hat{\lambda} \pm 1.96 \text{ SE}$, here $[1.61, 4.39]$ (dot-dash

vertical lines, where the dashed parabola crosses the grey cutoff line).

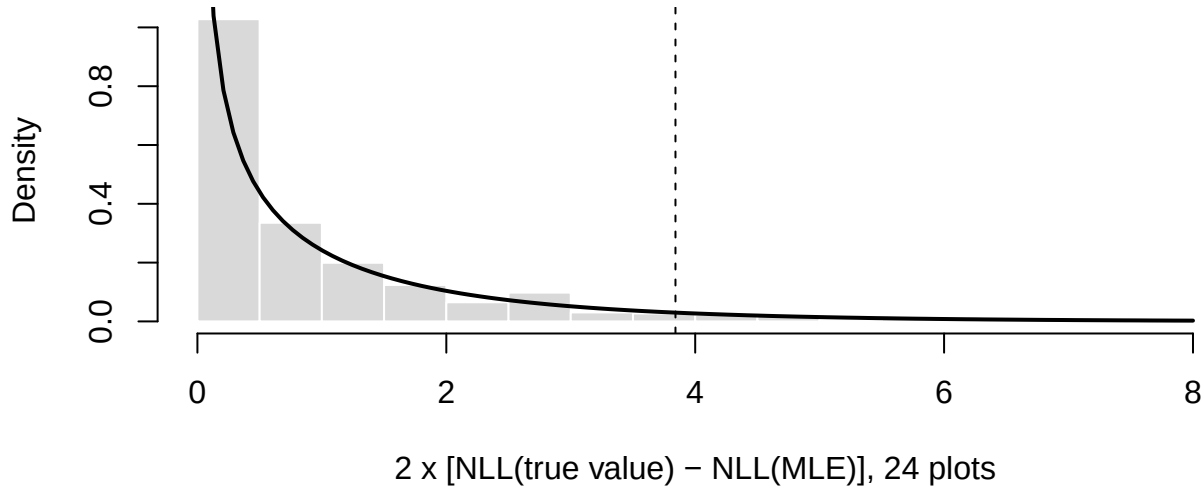
- **Where 1.92 comes from.** Put the true value into the parabola:

$$2 [\text{NLL}(\lambda_0) - \text{NLL}(\hat{\lambda})] \approx I (\hat{\lambda} - \lambda_0)^2 = \left(\frac{\hat{\lambda} - \lambda_0}{1/\sqrt{I}} \right)^2 = Z^2.$$

Section 3 showed that the quantity in parentheses is a standard normal, Z , so twice the NLL difference is χ_1^2 (Wilks' theorem). Its 95th percentile is 3.84, and half of that is 1.92.

- **The likelihood ratio interval** uses the 1.92 cutoff on the true curve instead of the parabola: [1.82, 4.61] (dotted vertical lines, where the solid NLL crosses the cutoff). It follows the curve's lopsidedness, which the parabola can't.

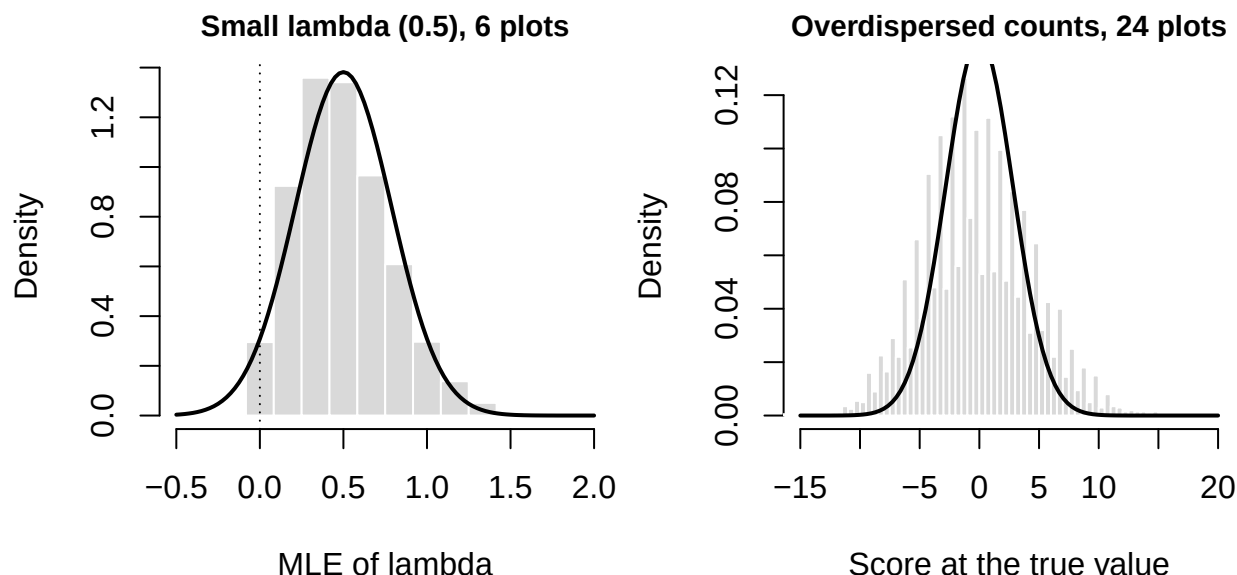
So the Hessian interval and the likelihood ratio interval come from the same idea. They agree when the score is close to a straight line, and they part ways when it bends, as it does here with only 6 plots.



Across 4,000 simulated surveys, twice the NLL difference at the true value follows a χ_1^2 distribution (curve); about 95% of surveys fall below 3.84 (dashed line). In this simulation the share is 0.948.

6. When the picture breaks

Every step above assumed the score curves are close to straight lines, their heights are close to normal, and the model is right. Each can fail.



- **Near a boundary or with little data (left).** With $\lambda_0 = 0.5$ and 6 plots, the MLE is skewed, and the normal approximation (curve) puts probability below zero, where λ can't go. This is why a Hessian interval for a prevalence of 2 in 20 can have a negative lower limit. The likelihood ratio interval or a Bayesian interval handles this better.
- **Wrong model (right).** Here the counts are more variable than a Poisson allows (negative binomial), but we still fit a Poisson. The score heights spread far more than the predicted bell (curve). Variance and slope no longer match, so the Hessian standard errors are too small. “Sandwich” (robust) standard errors estimate the two pieces separately for this reason.
- **Several parameters.** The same argument runs with vectors and matrices: $\hat{\theta} - \theta_0 \approx I^{-1}U$, so the variance-covariance matrix of the estimates is I^{-1} , the inverse of the Hessian. Its off-diagonal entries are the correlations between estimates, the tilt of the likelihood surface.

Summary

Idea	Graphical meaning	Formula
Score	slope of the log-likelihood	$U(\theta) = \frac{d}{d\theta} \log \mathcal{L}(\theta)$
MLE	where the score crosses zero	$U(\hat{\theta}) = 0$
Fisher information	steepness of the score; curvature of the NLL	$I = E[-U'(\theta_0)]$
Score at the truth	heights of the curves at θ_0 ; normal by the CLT	$U(\theta_0) \approx \text{Normal}(0, I)$
Information identity	curves must move when the truth moves	$\text{Var}(U) = I$

Idea	Graphical meaning	Formula
Sampling distribution of the MLE	crossing points = height / slope	$\hat{\theta} \approx \text{Normal}(\theta_0, 1/I)$
Hessian (Wald) interval	width of the parabola at 1.92	$\hat{\theta} \pm 1.96/\sqrt{I}$
Likelihood ratio cutoff	squared standard normal	$2 \Delta\text{NLL} \sim \chi_1^2$, cutoff 1.92

One sentence holds it together: **a parabola-shaped log-likelihood and a straight-line score are the same assumption, and the normal sampling distribution, the Hessian interval, and the 1.92 cutoff all rest on it.**

A companion handout, *Where AIC's 2k Penalty Comes From*, uses the same ideas (the parabola, $\text{Var}(\hat{\theta}) \approx 1/I$, and Wilks' theorem) to derive the penalty term in AIC.

References

- Bolker, B. M. 2008. *Ecological Models and Data in R*. Princeton University Press, Princeton, NJ. (Sections 6.4 and 6.5.)
- Pawitan, Y. 2001. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, Oxford. (A fuller treatment of the score, Fisher information, and likelihood intervals.)