

# Likelihood-based inference: Single parameter

Patrick Breheny

September 21

## Introduction

- Previously, we constructed and plotted likelihoods and used them informally to comment on likely values of parameters
- Our goals for today:
  - Connect likelihood with probability, in order to quantify coverage and type I error rates for various likelihood-based approaches
  - See how likelihood is also a fundamental component of Bayesian inference
- With the exception of simple cases such as the two-sample exponential model, exact derivations of these quantities is typically unattainable, and we must rely on asymptotic arguments (frequentist) or numerical integration (Bayesian)

## The score statistic

- Likelihoods are typically easier to work with on the log scale (where products become sums); furthermore, since it is only relative comparisons that matter with likelihoods, it is more meaningful to work with derivatives than the likelihood itself
- Thus, we often work with the derivative of the log-likelihood, which is known as the *score*, and often denoted  $U$ :

$$U(\theta) = \frac{d}{d\theta} \ell(\theta|X)$$

## The score statistic (cont'd)

- Note that
  - $U$  is a random variable, as it depends on  $X$
  - $U$  is a function of  $\theta$
  - For independent observations, the score of the entire sample is the sum of the scores for the individual observations:

$$U = \sum_i U_i$$

- In the derivations that follow, I will use  $U$  as shorthand for the score statistic evaluated at the true value of the parameter,  $\theta^*$ , and  $U(\theta)$  when we evaluate the score at other values of  $\theta$

# Mean

- We now consider some theoretical properties of the score
- It is worth noting that there are some regularity conditions that  $f(x|\theta)$  must meet in order for these theorems to work; we'll discuss these in greater detail a little later
- **Theorem:**  $\mathbb{E}(U) = 0$
- Note that maximum likelihood can therefore be viewed as a method of moments estimator with respect to the score statistic

# Variance

- **Theorem:**

$$\mathbb{V}(U) = -\mathbb{E}(U')$$

- The variance of  $U$  is given a special name in statistics: it is called the *Fisher information*, the *expected information*, or simply the *information*
- For notation, I will use  $\mathcal{I}$  to represent the Fisher information, and  $\mathcal{I}_i$  to represent the contribution to the Fisher information coming from the  $i$ th subject; note that under independence,  $\mathcal{I} = \sum_i \mathcal{I}_i$
- Like the score, the Fisher information is a function of  $\theta$ , although unlike the score, it is not random, as the random variable  $X$  has been integrated out

## Some examples

- **Example #1:** For the normal mean model (treating  $\sigma^2$  as known),

$$\mathcal{I}_i = \frac{1}{\sigma^2};$$

this makes sense: as the data becomes noisier, less information is contained in each observation

- In the above example,  $U'$  is free of both  $X$  and  $\mu$ ; in general both can appear in the information, which gives rise to a few different ways of working with the information in practice
- **Example #2:** For the Poisson distribution,

$$U'_i = -X_i \lambda^{-2}$$

## Observed information

- The Fisher information is therefore

$$\mathcal{I}_i(\lambda) = \lambda^{-1}$$

- Here, taking the expectation was straightforward; in general, it can be complicated, and for survival data analysis in particular, typically involves the censoring mechanism
- A simpler alternative is to use the observed values of  $\{X_i\}$  rather than their expectation; this is known as the *observed information* and will be denoted  $I$
- In the Poisson example,

$$I(\lambda) = \lambda^{-2} \sum_i x_i$$



## Asymptotic distribution

We have a sum of independent terms for which we know the mean and variance; we can therefore apply the central limit theorem:

$$\sqrt{n}\{\bar{U} - \mathbf{E}(U)\} \xrightarrow{d} N(0, \mathcal{I}_i),$$

or equivalently,

$$\frac{1}{\sqrt{n}}U \xrightarrow{d} N(0, \mathcal{I}_i),$$

## Consistency and information

- **Proposition:** Any consistent estimator of the information can be used in place of  $\mathcal{I}_i$  from the previous slide, and the result still holds
- Thus, all of the following results hold:

$$\mathcal{I}(\theta^*)^{-1/2}U \xrightarrow{d} N(0, 1)$$

$$\mathcal{I}(\hat{\theta})^{-1/2}U \xrightarrow{d} N(0, 1)$$

$$I(\theta^*)^{-1/2}U \xrightarrow{d} N(0, 1)$$

$$I(\hat{\theta})^{-1/2}U \xrightarrow{d} N(0, 1)$$

provided that  $\hat{\theta}$  is a consistent estimator of the true value  $\theta^*$

## Inference: Introduction

- How can we use these results to carry out likelihood-based inference?
- It turns out that there are three widely used frequentist techniques for doing so: the *score*, *Wald*, and *likelihood ratio* methods, as well as the Bayesian approach
- For the remainder of this lecture, we will motivate these approaches and then apply them to exponentially distributed survival data as an illustration of how they work

## Score test

- The score test follows most directly from our earlier derivations
- Here, to test  $H_0 : \theta = \theta_0$ , we simply calculate

$$\frac{U(\theta_0)}{\sqrt{I(\theta_0)}}$$

and then compare it to a standard normal distribution

- As always, by inverting this test at  $\alpha = 0.05$ , we can obtain 95% confidence intervals for  $\theta$
- Note that the score test, unlike the next two approaches we will consider, does not even require estimating  $\theta$

## Wald approximation

- The score test was first proposed by C. R. Rao; an alternative approach, first proposed by Abraham Wald, relies on a Taylor series approximation to the score function about the MLE
- **Proposition:**

$$U(\theta) \approx I(\hat{\theta})(\hat{\theta} - \theta)$$

## Wald result

- Thus,

$$I(\hat{\theta})^{1/2}(\hat{\theta} - \theta^*) \sim N(0, 1), \text{ or}$$
$$\hat{\theta} \sim N(\theta^*, I(\hat{\theta})^{-1})$$

- The MLE is therefore
  - Approximately normal...
  - ...with mean equal to the true value of the parameter...
  - ...and variance equal to the inverse of the information
- Based on this result, we can easily construct tests and confidence intervals for  $\theta$

## LRT approximation

- Finally, we could also consider the asymptotic distribution of the likelihood ratio, originally derived by Samuel Wilks
- This approach also involves a Taylor series expansion, but here we approximate the log-likelihood itself about the MLE, as opposed to the score
- **Proposition:**

$$\ell(\theta) \approx \ell(\hat{\theta}) - \frac{1}{2}I(\hat{\theta})(\theta - \hat{\theta})^2$$

## LRT result

- Thus,

$$2\{\ell(\hat{\theta}) - \ell(\theta^*)\} \sim \chi_1^2$$

- Note that

$$\exp\{-\chi_{1,(1-\alpha)}^2/2\} = 0.15;$$

this was the basis for choosing 15% as a cutoff for  $L(\theta)/L(\hat{\theta})$  in our likelihood intervals

- It is worth pointing out, however, that a 15% cutoff for  $L(\theta)/L(\hat{\theta})$  is only appropriate for the single parameter case; as we will see next time, the cutoff needs to change when multiple unknown parameters are present



## Regularity conditions

The score, Wald, and LRT approaches derived here are all asymptotically equivalent to each other, and all hold provided that certain regularity conditions are met:

- $\theta$  is not a boundary parameter (otherwise we can't take an approximation about it)
- The information matrix  $I(\theta^*)$  is finite and positive
- We can take up to third derivatives of  $\int f(x|\theta)$  inside the integral, at least in the neighborhood of  $\theta^*$
- The distributions  $\{f(x|\theta)\}$  have common support and are identifiable

## Reparameterization

- It is worth noting that the score and Wald approaches will be affected by reparameterization
- For example, if we decide to carry out inference for the log-hazard  $\gamma = \log(\lambda)$  of an exponentially distributed time-to-event, we will obtain different score and Wald confidence intervals than if we constructed intervals for  $\lambda$  and then transformed them
- The likelihood ratio approach, however, since it doesn't involve any derivatives, will be unaffected by such transformations

## Bayesian inference: Main idea

The central idea of the Bayesian framework is that if we treat  $\theta$  as a random variable, then

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{f(x)},$$

where

- $f(x|\theta)$  is the likelihood
- $f(\theta)$  is the *prior*: Our beliefs about the plausible values of our parameter before seeing any data
- $f(\theta|x)$  is the *posterior*: Our updated beliefs about the plausible values for our parameter after seeing the data
- $f(x)$  is a normalizing constant typically not of interest

# Priors

- To carry out Bayesian inference, therefore, we need to specify both a prior as well as a likelihood
- Broadly speaking, there are two main ways of specifying priors:
  - *Informative priors* attempt to incorporate knowledge from other sources such as past studies in order to realistically capture one's state of knowledge about  $\theta$
  - *Reference priors* attempt to represent a vague, uninformed baseline, so that all conclusions will be based on the data alone, not from any external sources

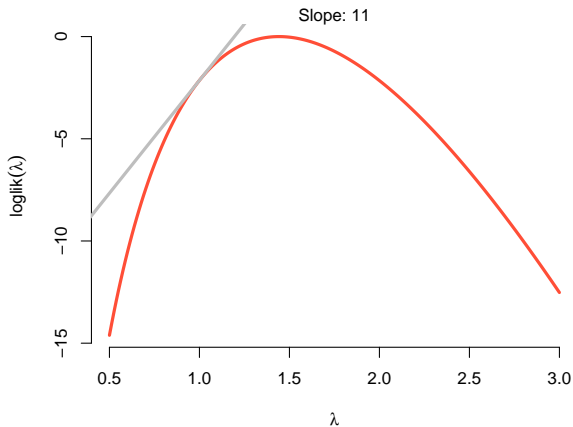
# Inference

- Once the model has been specified, all inference is based on the posterior  $f(\theta|x)$
- For example, we can obtain point estimates via the posterior mean  $\int \theta f(\theta|x) d\theta$  or posterior mode  $\max_{\theta} f(\theta|x)$
- We can obtain 95% posterior intervals  $[a, b]$  such that  $\int_a^b f(\theta|x) d\theta = 0.95$
- We can calculate tail probabilities:  $\mathbb{P}(\theta < 0) = \int_{-\infty}^0 f(\theta|x) d\theta$
- Note that with the Bayesian approach, no asymptotic arguments are required, although the integrals involved may be complicated, and thus, numerical integration methods are typically crucial to Bayesian methodology

## Pike rat example

- To illustrate these approaches and the geometry behind them, we'll apply them to the Pike rat data
- For the purposes of this illustration, we'll assume the data follow an exponential distribution (actually a pretty bad assumption here) under independent censoring
- Also, we'll just look at overall survival without respect to pretreatment regimen

# Score approach: $H_0 : \lambda = 1$



## Score approach: $H_0 : \lambda = 1$ (cont'd)

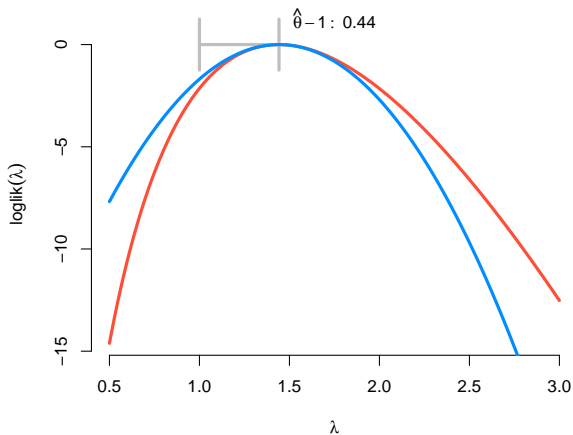
- So, we observe a score of  $d - v = 11$
- We would expect the score to be zero (i.e, if  $\lambda = 1$ , we'd expect to be near the top of the curve, where it's flat)
- Still, the standard error of the slope is  $\sqrt{d} = 6$ , so our observed score is only

$$Z = 11/6 = 1.84$$

standard deviations away from the mean, implying that we have insufficient evidence to rule out  $\lambda = 1$  ( $p = 0.07$ )



# Wald approach: $H_0 : \lambda = 1$



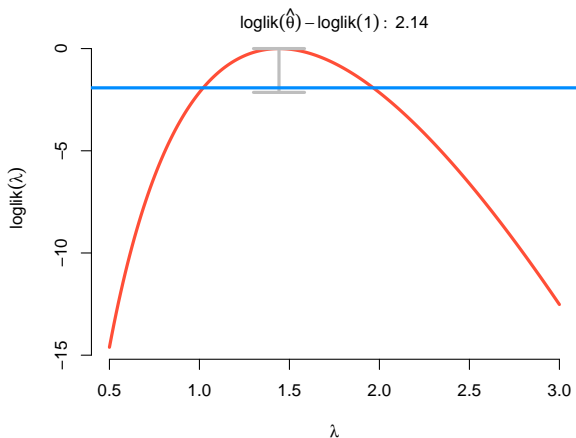
## Wald approach: $H_0 : \lambda = 1$ (cont'd)

- So, we observe a difference of  $\hat{\lambda} - \lambda_0 = d/v - 1 = 0.44$
- We would expect this difference to be near zero if  $\lambda$  was truly equal to 1
- However, the standard error  $\hat{\theta}$  is  $\sqrt{d}/v = 0.24$ , so our observed difference is only

$$Z = 0.44/0.24 = 1.84;$$

in this particular case, the score and Wald approaches coincide, but this is not true in general

# Likelihood ratio approach: $H_0 : \lambda = 1$



## Likelihood ratio approach: $H_0 : \lambda = 1$ (cont'd)

- So, we observe a difference of  $\ell(\hat{\lambda}) - \ell(\lambda_0) = 2.14$
- Our  $p$ -value is therefore the area to the right of  $2(2.14) = 4.29$  for a  $\chi_1^2$  distribution
- This turns out to be  $p = 0.04$ ; thus,  $\lambda = 1$  would be excluded from our likelihood ratio confidence interval despite being included in both the score and Wald intervals

## “Exact” result

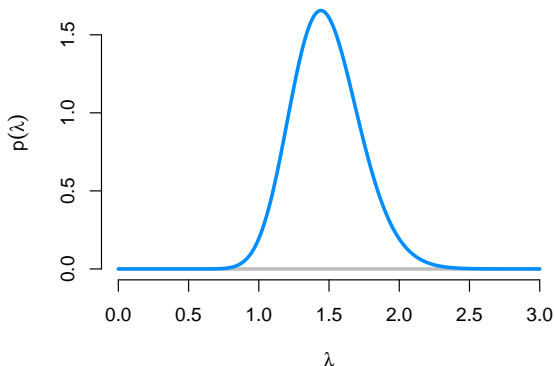
- For the exponential distribution, we could carry out something of an “exact” test based on the gamma distribution
- Here, our (one-sided)  $p$ -value would be the area to the left of  $V$  for a gamma distribution with shape parameter  $d$  and rate parameter  $\lambda_0$ , although it would only be exact in the case of type II censoring
- Nevertheless, the resulting one-sided  $p$ -value is 0.02; this is in good agreement with the two-sided  $p$ -value of 0.04 we got from the likelihood ratio test

## Accuracy

- This small anecdote doesn't necessarily prove anything; nevertheless, it is the case the the likelihood ratio approach is typically the most accurate of the three
- To see why, consider analyzing a transformation,  $g(\theta)$
- Some transformations will make the normal approximations for the score and Wald approaches more accurate (and some will make them less accurate)
- Suppose there exists a "best" transformation  $g^*$ ; you could improve your score/Wald accuracy by finding and then applying  $g^*$ , but with the likelihood ratio test, you've already achieved that accuracy without even finding  $g^*$

## Bayesian approach: Reference prior

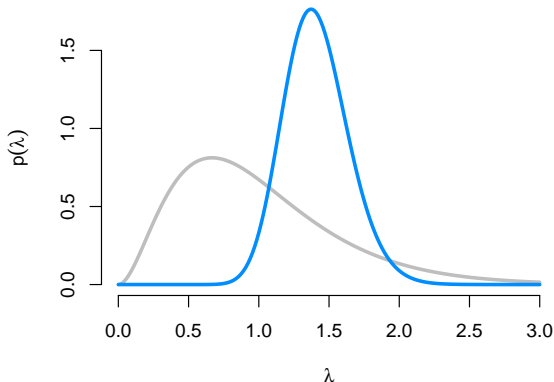
Finally, let's look at the Bayesian approach, first using an uninformative flat prior:



$$\mathbb{P}(\lambda < 1|d, v) = 0.014$$

## Bayesian approach: Gamma(3,3) prior

Suppose prior studies suggested that  $\lambda$  was likely between 0 and 2, and could reasonably be represented by a Gamma(3,3) distribution:



$$\mathbb{P}(\lambda < 1 | d, v) = 0.028$$