GOV 2001/ 1002/ E-200 Section 3
Inference and Likelihood

Anton Strezhnev

Harvard University

February 10, 2016

Logistics

Reading Assignment- Unifying Political Methodology ch 4 and Eschewing Obfuscation

Problem Set 3- Due by 6pm, 2/24 on Canvas.

Assessment Question- Due by 6pm, 2/24 on on Canvas. You must work alone and only <u>one</u> attempt.

Replication Paper

1. Read *Publication, Publication*

2. Find a coauthor. See the Canvas discussion board to help with this.

3. Choose a paper based on the crieria in *Publication, Publication*.

4. Have a classmate sign-off on your paper choice.

Overview

- In this section you will...
  - learn how to derive a likelihood function for some data given a data-generating process.
  - learn how to calculate a Bayesian posterior distribution and generate quantities of interest from it.
  - learn about common pitfalls in hypothesis testing and think about how to interpret p-values more critically.
  - learn that Frequentists and Bayesians aren't really that different after all!

Outline

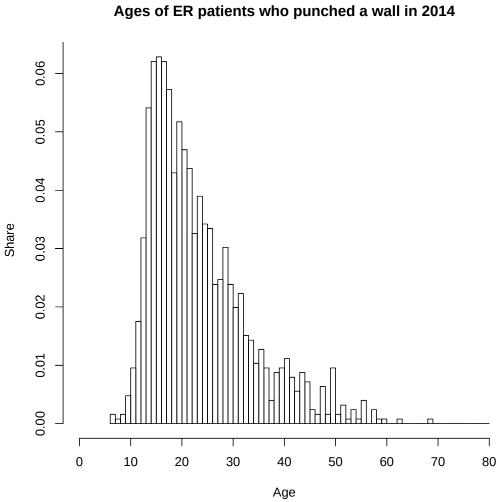
Likelihood inference

- Last week we talked about probability – Given parameters, what's the probability of the data.
- This week we're talking about inference – Given the data, what can we say about the parameters.
- Likelihood approaches to inference ask "What parameters make our data most likely?"

Example: Age distribution of ER visits due to wall punching

- We have a dataset from the U.S. Consumer Product Safety Commission's National Electronic Injury Surveillance System (NEISS) containing data on ER visits in 2014.
- Let's take a look at one injury category – wall punching. We're interested in modelling the distribution of the ages of individuals who visit the ER having punched a wall.
- To do this, we write down a probability model for the data.

Empirical distribution of wall-punching ages



A Model for the Data – Log-Normal distribution

- We observe n observations of ages, $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$.
- A normal distribution doesn't seem like a reasonable model since age is strictly positive and the distribution is somewhat right-skewed.
- But a log-normal might be reasonable!
- We assume that each $Y_i \sim \text{Log-Normal}(\mu, \sigma^2)$, and that each Y_i is independently and identically distributed. We could extend this model by adding covariates (e.g. $\mu_i = X_i\beta$).

Example: Age distribution of ER visits due to wall punching

The density of the log-normal distribution is given by

$$f(Y_i|\mu,\sigma^2) = \frac{1}{Y_i\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln(Y_i)-\mu)^2}{2\sigma^2}\right)$$

Basically the same as saying $\ln(Y_i)$ is normally distributed!

Writing a likelihood

- After writing a probability model for the data, we can write the likelihood of the parameters given the data
- By definition of likelihood

$$\mathcal{L}(\mu, \sigma^2|\mathbf{Y}) \propto f(\mathbf{Y}|\mu, \sigma^2)$$

- Unfortunately, $f(\mathbf{Y}|\mu, \sigma^2)$ is an n-dimensional density, and n is huge! How do we simplify this? The *i.i.d.* assumption lets us factor the density!

$$\mathcal{L}(\mu, \sigma^2|\mathbf{Y}) \propto \prod_{i=1}^{N} f(Y_i|\mu, \sigma^2)$$

Writing a likelihood

- Now we can plug in our assumed density for Y.

$$\mathcal{L}(\mu, \sigma^2 | \mathbf{Y}) \propto \prod_{i=1}^{N} \frac{1}{Y_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}\right)$$

- However, if we tried to calculate this in R, the value would be incredibly small! It's the product of a bunch of probabilities which are between 0 and 1. Computers have problems with numbers that small and round them to 0.
- It's also often analytically easier to work with sums over products.
- This is why we typically work with the log-likelihood (often denoted ℓ). Because taking the log is a monotonic transformation, it retains the proportionality!

$$\mathcal{L}(\mu, \sigma^2 | \mathbf{Y}) \propto \ell(\mu, \sigma^2 | \mathbf{Y})$$

Logarithm review!

- Logs turn exponentiation into multiplication and multiplication into summation.
  - $\log(A \times B) = \log(A) + \log(B)$
  - $\log(A/B) = \log(A) - \log(B)$
  - $\log(A^b) = b \times \log(A)$
  - $\log(e) = \ln(e) = 1$
  - $\log(1) = 0$
- Notational note: log in math is almost always used as short-hand for the natural log (ln) as opposed to the base-10 log.

Deriving the log-likelihood

$$
\begin{aligned}
\ell(\mu, \sigma^2 | \mathbf{Y}) &\propto \ln \left[\prod_{i=1}^{N} f(Y_i | \mu, \sigma^2) \right] \\
&\propto \ln \left[\prod_{i=1}^{N} \frac{1}{Y_i \sigma \sqrt{2\pi}} \exp \left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2} \right) \right] \\
&\propto \sum_{i=1}^{N} \ln \left[\frac{1}{Y_i \sigma \sqrt{2\pi}} \exp \left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2} \right) \right] \\
&\propto \sum_{i=1}^{N} - \ln(Y_i) - \ln(\sigma) - \ln(\sqrt{2\pi}) + \ln \left[\exp \left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2} \right) \right] \\
&\propto \sum_{i=1}^{N} - \ln(Y_i) - \ln(\sigma) - \ln(\sqrt{2\pi}) - \frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}
\end{aligned}
$$

Deriving the log-likelihood

- To simplify further, we can drop multiplicative (additive on the log scale) constants that are not functions of the the parameters since that retains proportionality.

$$\propto \sum_{i=1}^{N} -\ln(Y_i) - \ln(\sigma) - \ln(\sqrt{2\pi}) - \frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}$$

$$\propto \sum_{i=1}^{N} -\ln(\sigma) - \frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}$$

Writing the log-likelihood in R

- We can often make use of the built-in PDF functions in R for distributions to write a function that takes as input μ, σ^2 and the data. Here, we want to use `dlnorm` (the density of the log-normal).

```
### Log-Likelihood function
log.likelihood.func <- function(mu, sigma, Y){
  # Return the sum of the log of dnorm evaluated for all Y with fixed mu and
       sigma
  return(sum(dlnorm(Y, meanlog=mu, sdlog=sigma, log=T))) ## Set log=T to return
        the log-density
}
```

Plotting the log-likelihood

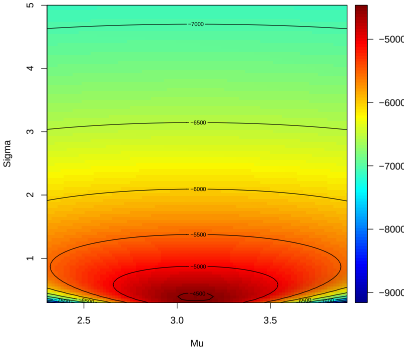


Figure : Contour plot of the log-likelihood for different values of μ and σ

Plotting the likelihood

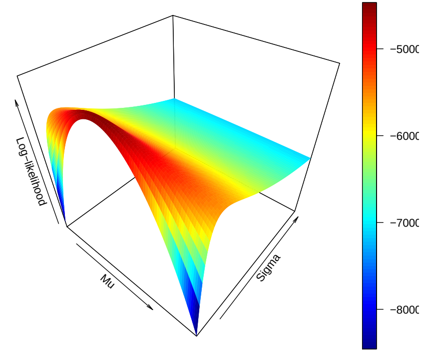


Figure : Plot of the log-likelihood for different values of μ and σ

Plotting the likelihood

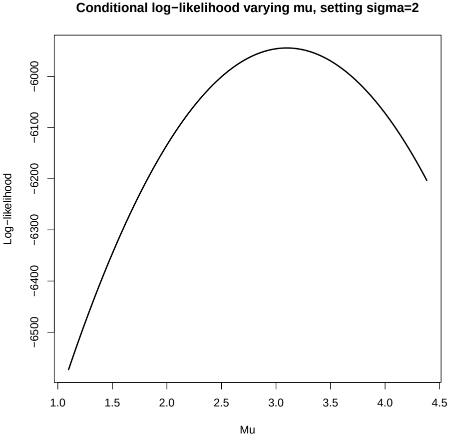


Figure : Plot of the conditional log-likelihood of μ given $\sigma = 2$

Comparing models using likelihood

- In future problem sets, you'll be directly optimizing (either analytically or using R) to find the parameters that maximize of the likelihood.
- For today, we'll eyeball it and compare the fit to the data for parameters that yield low likelihoods vs. higher likelihoods.
- Example 1: $\mu = 4$, $\sigma = .2$: Log-likelihood = -18048.79
- Example 2: $\mu = 3.099$, $\sigma = 0.379$: Log-likelihood = -4461.054 (actually the MLE)!
- Let's plot the implied distribution of Y_i for each parameter set over the empirical histogram!

COMPARING MODELS USING LIKELIHOOD

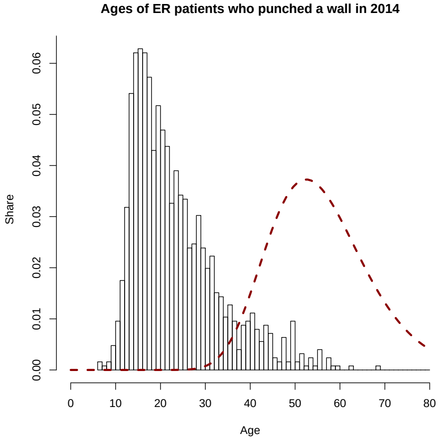


Figure : Empirical distribution of ages vs. log-normal with $\mu = 4$ and $\sigma = .2$

Comparing models using likelihood

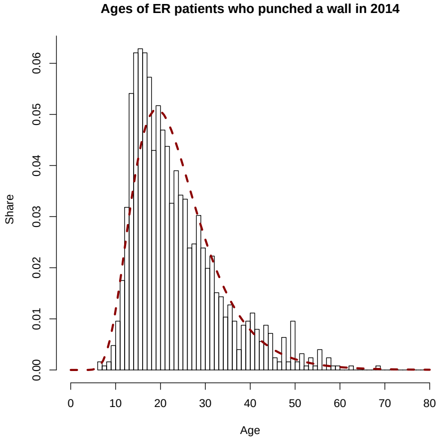


Figure : Empirical distribution of ages vs. log-normal using MLEs of parameters

OUTLINE

Likelihoods vs. Bayesian Posteriors

Likelihood:

$$
\begin{aligned}
p(\lambda|y) &= \frac{p(\lambda)p(y|\lambda)}{p(y)} \\
L(\lambda|y) &= k(y)p(y|\lambda) \\
&\propto p(y|\lambda)
\end{aligned}
$$

There is a fixed, true value of λ. We use the likelihood to estimate λ with the MLE.

Bayesian Posterior Density:

$$
\begin{aligned}
p(\lambda|y) &= \frac{p(\lambda)p(y|\lambda)}{p(y)} \\
&= \frac{p(\lambda)p(y|\lambda)}{\int_\lambda p(\lambda)p(y|\lambda)d\lambda} \\
&\propto p(\lambda)p(y|\lambda)
\end{aligned}
$$

λ is a random variable and therefore has fundamental uncertainty. We use the posterior density to make probability statements about λ.

Understanding the Posterior Density

In Bayesian inference, we have a prior *subjective* belief about λ , which we update with the data to form posterior beliefs about λ.

$$p(\lambda|y) \quad \propto \quad p(\lambda)p(y|\lambda)$$

- $p(\lambda|y)$ is the posterior density
- $p(\lambda)$ is the prior density
- $p(y|\lambda)$ is proportional to the likelihood

Bayesian Inference

The whole point of Bayesian inference is to leverage information about the data generating process along with subjective beliefs about our parameters into our inference.

Here are the basic steps:

1. Think about your subjective beliefs about the parameters you want to estimate.
2. Find a distribution that you think explains your prior beliefs of the parameter.
3. Think about your data generating process.
4. Find a distribution that you think explains the data.
5. Derive the posterior distribution.
6. Plot the posterior distribution.
7. Summarize the posterior distribution. (posterior mean, posterior standard deviation, posterior probabilities)

Example: Waiting time for a taxi on Mass Ave

If you randomly show up on Massachusetts Avenue, how long will it take you to hail a taxi?

Example: Waiting time for a taxi on Mass Ave

- Let's assume that waiting times X_i (in minutes) are distributed Exponentially with parameter λ.
- $X_i \sim \text{Expo}(\lambda)$
- The density is $f(X_i|\lambda) = \lambda e^{-\lambda X_i}$
- We observe one observation of $X_i = 7$ minutes and want to make inferences about λ. Quiz: Using what you know about the mean of the exponential, what would be a good guess for λ *without* any prior information? $\frac{1}{7}$! (since the mean of the Expo is $\frac{1}{\lambda}$)

Deriving a posterior distribution

$$
\begin{aligned}
p(\lambda|X_i) &= \frac{p(X_i|\lambda)p(\lambda)}{p(X_i} \\
&\propto p(X_i|\lambda)p(\lambda) \\
&\propto \lambda e^{-\lambda X_i} p(\lambda)
\end{aligned}
$$

- Even when deriving Bayesian posteriors, it's often easier to work without proportionality constants (e.g. $p(X_i)$). You can figure out these "normalizing" constants at the end by integration since you know that a valid probability density

Deriving a posterior distribution

- How do we choose a distribution for $p(\lambda)$? The difficulty of this question is why Bayesian methods only recently gained wider adoption. Most prior choices give posteriors that are analytically intractable (can't express them in a neat mathematical form). More advanced computational methods (like MCMC) make this less of an issue.
- However, for some distributions of the data, there are distributions called "conjugate priors." These priors retain the shape of their distribution after being multiplied by the data/likelihood.
- Example: Beta distribution is conjugate to Binomial data.

Deriving a posterior distribution

- The conjugate prior for λ in Exponential data is the Gamma distribution. So we assume a prior of the form $\lambda \sim \text{Gamma}(\alpha, \beta)$.
- α and β are "hyperparameters" – we have to assume values for them that capture our prior beliefs.
- In the case of the Expo-Gamma relationship, α and β have substantive meaning – you can think of it as denoting α previously observed taxi times that sum to a total of β.

Deriving a posterior distribution

$$
\begin{aligned}
p(\lambda|X_i) &\propto \lambda e^{-\lambda X_i} p(\lambda) \\
&\propto \lambda e^{-\lambda X_i} \times \lambda^{\alpha-1} e^{-\beta\lambda} \\
&\propto \lambda^{\alpha} e^{-(\lambda(X_i+\beta)}
\end{aligned}
$$

- By inspection, the posterior for λ is also the form of a Gamma. Here, it's Gamma$(\alpha + 1, \beta + X_i)$
- We could also integrate the above form to get the normalizing constant and get an explicit density if we didn't recognize it as a known distribution.

Plotting the posterior

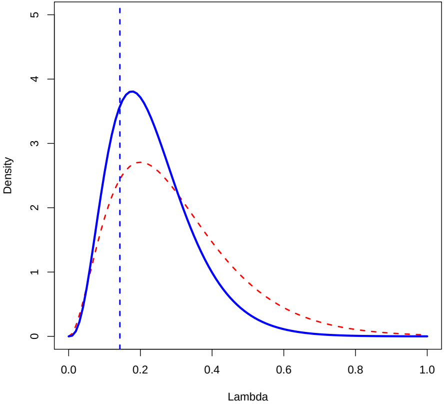


Figure : Prior and Posterior densities for λ (Red = Prior, Blue = Posterior). Vertical line denotes MLE). $\alpha = 3$, $\beta = 10$

Outline

Is ESP Real?

- Bem (2011) conducted 9 experiments purporting to show evidence of precognition.
- One experiment had 100 respondents asked to repeatedly guess which "curtain" had a picture hidden behind it.
- Under "null" hypothesis, guess rate by chance would be 50%. But Bem found that "explicit" images were significantly more likely to be predicted (53.1%) With a p-value of .01!
- Should we conclude that precognition exists? What makes Bem's p-value different from one that you calculate in your study?
- Answer: Your priors about effect size will affect how you interpret p-values.

HYPOTHESIS TESTING

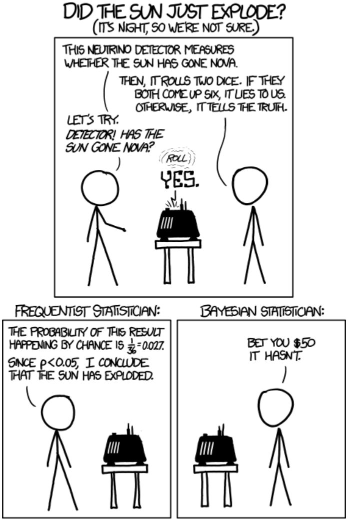


Figure : A misleading caricature - everyone uses priors

Everyone's a Little Bit Bayesian

- Frequentist inference *doesn't mean* that prior information is irrelevant! (despite popular interpretations). All inferences depend on prior beliefs about the plausibility of a hypothesis.[1]
- Where Bayesians and Frequentists differ is in *how* that information is used.
- Bayesians use a formally defined prior
  - Advantage: Explicitly incorporates prior beliefs into final inferences in a rigorous way.
  - Disadvantages: Prior needs to be elicited explicitly (in the form of a distribution). Wrong priors give misleading results. Computational issues with non-conjugate priors.
- Frequentists use prior information in the design and interpretation of studies.
  - Advantage: Not necessary to formulate prior beliefs in terms of a specific probability distribution.
  - Disadvantages: No clear rules for how prior information should be weighed relative to the data at hand.

[1] See Andy Gelman's comments at `http://andrewgelman.com/2012/11/10/16808/`

Everyone's a Little Bit Bayesian

- Don't forget what you learned in Intro to Probability!
- Classic example: A disease has a very low base rate (.1% of the population). A test for the disease has a 5% false positive rate and a 5% false negative rate. Given that you test positive, what's the probability you have the disease?
- Bayes' rule: $P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D)+P(+|\text{Not D})P(\text{Not D})}$
- $P(D|+) = \frac{.95 \times .001}{.95 \times .001 + .05 \times .999} = .01866 = 1.86\%$
- The same principles apply to hypothesis testing! Always important to ask: given my decision to reject, how likely is it that my decision is misleading?

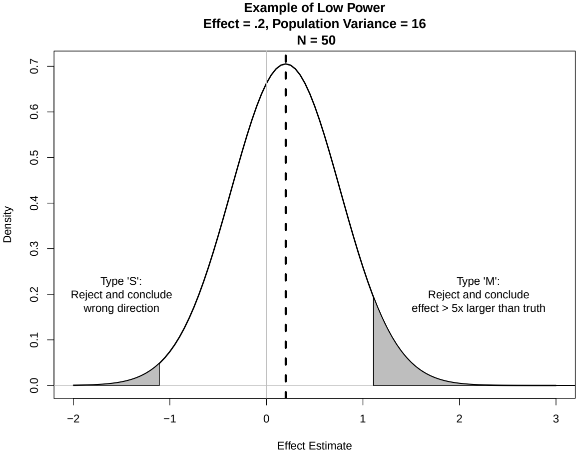
Thinking about p-values

- We typically calibrate p-values in terms of Type I error – that is, False Positive Rate.
- But false-positive rate can be misleading conditional on a positive result. Determining how "informative" our result is depends on additional design-related factors.
  - 1) The effect size
  - 2) The sample size

Type "M" and "S" Errors

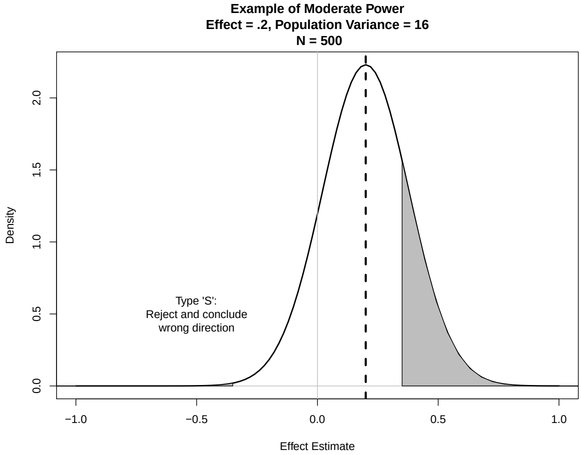
- Gelman and Carlin (2014) suggest also considering Type "S" (Sign) and Type "M" (Magnitude) error rates that are conditional on rejecting.
- Type "S" error: Given that you reject the null, what's the probability that your point estimate is the wrong sign?
- Type "M" error: Given that you reject the null, what's the probability that your estimate is too extreme?
- Both depend not only on your sampling distribution's variance, but also on the effect size.

Calculating Type "M" and "S" error rates



$Pr(\text{Reject}) = .0644$. $Pr(\text{Wrong Sign}|\text{Reject}) = .16$.
$Pr(\text{Estimate 5x Truth}|\text{Reject}) = .84$

Calculating Type "M" and "S" error rates



$Pr(\text{Reject}) = .200$. $Pr(\text{Wrong Sign}|\text{Reject}) = .005$.

Low probability of Type 'S' and our positive estimates are a lot more reasonable!

Takeaways for Hypothesis Testing

- General rule: *Smaller effects require larger samples (more data) to reliably detect.*
- A rule for tiny sample sizes and tiny effects: *You're probably getting nothing, and if you get something, it's probably wrong.*
- A rule for reading published p-values: *Just because it's peer-reviewed and published, doesn't mean its true.*

Questions

Questions?