Introduction to the Likelihood Theory of Inference

David A. Hughes, Ph.D.

Auburn University at Montgomery

david.hughes@aum.edu

January 31, 2020

Overview

Introduction

By the time students finish this unit, they should be able to explain:

- The difference between probability density and mass functions,
- The types of distributions we oftentimes deal with with categorical dependent variables, and
- The logic underpinning maximum likelihood inference.

What are density functions?

- Probability mass functions (PMFs) are useful with discrete data and describe the probability an outcome is equal to some precise value.
  - Suppose that $x_i \in \{1, 2, \ldots, k\}$.
  - Then $\sum_{i=1}^{k} Pr(x_i) = 1$.
- Probability density functions (PDFs) are useful with continuous data and describe the probability of some volume of outcomes.
  - Suppose you have a function, $f(\cdot)$ defined across $x \in (a, b)$.
  - Then: $Pr\{x \in (a, b)\} = \int_a^b f(x)dx = 1$.
  - For a normal distribution, then, $a = -\infty$ and $b = \infty$.
  - Note: $Pr(x_i = k) = 0$.

Why do I care about these distributions?

- For one thing, they can tell us something about the likelihood of an eventuality.
- When we engage in hypothesis testing, it is critical that we select appropriate distributions that can help us to measure the relative likelihood of having observed a given dataset.
- Our choice of distribution will largely come down to our level of measurement.

Bernoulli Distribution (discrete)

- This is the simplest statistical distribution
- Represents the situation where a random variable x_i has only two possible event outcomes, each with a non-zero probability of occurrence
- Example: flipping a coin
- $\Pr(x_i = 1) = \pi$ and $\Pr(x_i = 0) = 1 - \pi$.
- Formally, we represent the distribution:

$$x_i \sim f_{Bern}(x_i|\pi) = \pi^{x(1-\pi)(1-x)}$$

Binomial Distribution (discrete)

- This is a series of N Bernoulli random variables, where we only observe the sum of the observations
- The distribution is nonnegative and discrete (no fractions), with an upper bound of n
- Examples: the number of bills in a legislature, number of cases on a court's docket
- Mathematical specification:

$$f_{k,n,p} = \binom{n}{k} p^k (1-p)^{(n-k)}$$

- where:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Normal Distribution

- Most intuitively familiar distribution (pdf froms the familiar "bell-shaped" curve)
- Used in OLS regression models
- Somewhat difficult to employ in MLE, because it does not possess an analytic solution
  - Analytic solution requires computing integrals
  - Computationally, the mathematics underlying this distribution were too complex for early computers
- Mathematical specification:

$$y_i \sim \mathcal{N}(y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left[\frac{(y-\mu)^2}{\sigma^2}\right]}$$

Logistic Distribution

- Better adept at modeling probabilities for dichotomous outcomes than the Normal distribution
- Contains an analytic solution (e.g. is mathematical tractable)
- Low computational costs (can even be done by hand)
- Mathematical specification:

$$y_i \sim f_{Logistic}(y_i|\mathbf{X}\beta) = \frac{e^{\mathbf{X}\beta}}{1+e^{\mathbf{X}\beta}}$$

Poisson Distribution

- Used when dependent variable is a count with no upper bound
- Key assumption: Occurrence of one event has no influence on the expected number of subsequent events (λ)
- Mathematical specification:

$$y_i \sim f_{Poisson}(y_i|\lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

- where $\lambda > 0$ and $y_i = 0, 1, 2, \ldots$

Negative Binomial Distribution

- Two key assumptions about the Poisson distribution are often problematic:
  - That events accumulating during observation period i are independent
  - Events have a constant rate of occurrence
- If either assumption is violated, then a new distribution is required because λ is no longer constant for all observations
  - Instead, must assume that λ itself varies across observations according to a particular probability distribution
  - The most popular distribution for λ is the gamma distribution
  - This involves calculating another parameter in the equation — the variance of the distribution

Negative Binomial Distribution

- Mathematical specification:

$$y_i \sim f_{nb}(y_i|\lambda, \sigma^2) = \frac{\Gamma\left(\frac{\lambda}{\sigma^2-1} + y_i\right)}{y_i!\Gamma\left(\frac{\lambda}{\sigma^2-1}\right)} \left(\frac{\sigma^2-1}{\sigma^2}\right) y_i (\sigma^2)^{\frac{-\lambda}{\sigma^2-1}}$$

- where $\lambda > 0$ and $\sigma^2 > 0$.
- Note: the more events within observation i that are positively related, the larger σ^2 becomes. Also, as σ^2 approaches 0, the negative binomial distribution collapses into the Poisson distribution

Uncertainty and inference

- We use probability all the time to better understand our own uncertainty over events.
- We like to use probability to summarize the relative likelihood of an occurrence. For example, what's the probability I flip 10 heads in a row using a fair coin?
- We can think about probability as being either relative or subjective. Phenomena might be infinitely repeatable. But do we have the same concept in mind when we say Donald Trump has a 0.5 probability of winning reelection in 2020? This gets tricky.

Inverse probability

- Uncertainty is not the same as inference. We might want to know $Pr(y \mid \mathcal{M})$ such that y is our data and $\mathcal{M}$ represents the statistical model that describes the relationship among our data.
- Ideally, we might like to reverse this conditional probability and estimate the probability of a certain model, taking our data as a given.
- It turns out, this is actually not possible. We therefore turn to concepts such as likelihood or Bayes' Theorem to calculate *relative* uncertainty.

Likelihood as a model of inference

- To try and get at $Pr(\mathcal{M} \mid y)$, we will make use of the concept of likelihood.
- Suppose that $\tilde{\theta}$ represents the hypothetical parameter value for the data.
- Then:

$$L(\tilde{\theta} \mid y) = k(y)Pr(y \mid \tilde{\theta}) \propto Pr(y \mid \tilde{\theta}), \tag{1}$$

  where $k(y)$ is treated as an unknown, positive constant.
- This scaling parameter allows us to think about relative uncertainty and to calculate summary estimates of θ.

Intuition of maximum likelihood

- Consider a model such that $Y_i \sim N(\mu, \sigma^2)$ where $E(Y) = \mu$ and $Var(Y) = \sigma^2$.
- Suppose you have some data on Y, and you want to estimate μ and σ^2 from these data.
- The whole idea behind likelihood is to find the estimates of the parameters that maximize the probability of having observed these data.

An example

- Suppose Y is a sample of cars and their estimated fuel efficiency (in MPG): $Y = \{30, 25, 35, 15, 45\}$.
- Intuitively, how likely is it that these five data points were drawn from a normal distributin with $\mu = 50$?
- What about $\mu = 30$, which happens to be the empirical mean of Y?
- Maximium likelihood is a way of doing this.

Example continued

- We can think of the MPG observations as having been draws from a normal distribution's PDF:

$$Pr(Y_i = y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \text{exp}\left[\frac{-(Y_i - \mu)^2)}{2\sigma^2}\right],$$

which is the probability that, for any one observation, i, Y will take on the particular value y.

- We can think about the probability of a single realization being what it is, e.g.:

$$Pr(Y_1 = 30) = \frac{1}{\sqrt{2\pi\sigma^2}} \text{exp}\left[\frac{-(30 - \mu)^2}{2\sigma^2}\right].$$

Example continued

- If we assume that the observations in Y are independent, then we can consider the joint probability of the observations as simply the product of marginals.
- Recall that $Pr(a, b) = Pr(a) \times Pr(b)$.
- Therefore:

$$Pr(Y_1 = 30, Y_2 = 25) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathsf{exp}\left[\frac{-(30-\mu)^2}{2\sigma^2}\right] \times \frac{1}{\sqrt{2\pi\sigma^2}} \mathsf{exp}\left[\frac{-(25-\mu)^2}{2\sigma^2}\right].$$

The likelihood function

- We can generalize the previous example to include N marginal probabilities:

$$Pr(Y_i = y_i \forall i) \equiv \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \text{exp}\left[\frac{-(y_i - \mu)^2)}{2\sigma^2}\right]. \qquad (2)$$

- This product is generally known as the *Likelihood* $[L(Y)]$ and is the probability each observation is what it is, given the parameters.

Estimation of the likelihood function

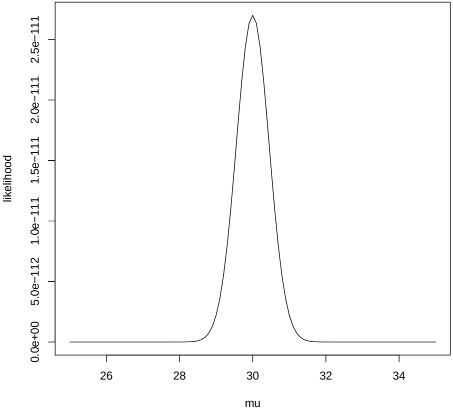
- Obviously, we don't know the paramaters as they are the very things we're kind of interested in.
- What we'd like to figure out is which values of μ and σ^2 were most likely to have generated Y in the first place.
- It turns out that $L(\hat{\mu}, \hat{\sigma}^2 \mid Y) \propto Pr(Y \mid \hat{\mu}, \hat{\sigma}^2)$.
- Essentially, we're looking for parameters that maximize the likelihood of generating the function.

The mechanics of maximum likelihood

- We could take the brute force approach and just start plugging in values for μ and σ^2 and see what gives us the biggest likelihood.
- For example, suppose I chose $\mu = 40$ and $\sigma^2 = 1$.
- Then: $L = (7.7 \times 10^{-23}) \times (5.5 \times 10^{-50}) \times \ldots$, which is a crazy small number.
- Holding $\sigma^2 = 1$, we could find out what value of μ maximizes the likelihood function.

Maximizing the likelihood function

- Turns out, the empirical mean is the answer to our problem.



Problems with the likelihood function

- Dealing with products can get tricky, especially when we're getting such teensy-tiny joint probabilities.
- Early computers really couldn't handle this (though that's not so much of a problem anymore).
- Therefore, we often take the natural log of the likelihood function, producing what we call the log-likelihood.

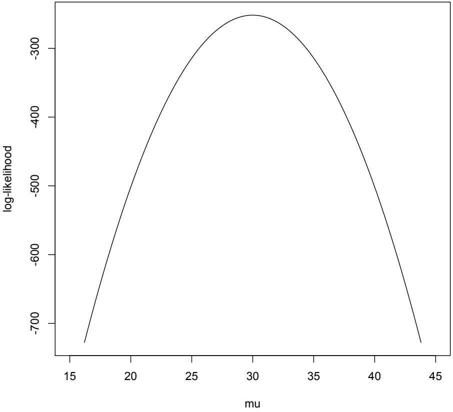
The log-likelihood function

- To calculate log-likelihood, we simply take the natural log of both sides of Equation 2.

$$
\begin{aligned}
lnL(\hat{\mu}, \hat{\sigma}^2 \mid Y) &= ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \text{exp}\left[\frac{-(y_i - \mu)^2)}{2\sigma^2}\right] \\
&= \sum_{i=1}^{N} ln \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \text{exp}\left[\frac{-(y_i - \mu)^2)}{2\sigma^2}\right] \right\} \\
&= \frac{-N}{2} ln(2\pi) - \left[\sum_{i=1}^{N} \frac{1}{2} ln\sigma^2 - \frac{1}{2\sigma^2}(Y_i - \mu)^2 \right].
\end{aligned}
$$

Visualizing the log-likelihood function

- Assuming once again that $\sigma^2 = 1$, we get:



Maximizing the log-likelihood function

- Generally, eye-balling graphs won't be sufficient to find maxima.
- We turn to differential calculus to find these points.
- Taking the derivative of the log-likelihood equation above with respect to μ and σ^2 gives:

$$\begin{aligned} \frac{\partial lnL}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{N}(Y_i - \mu), \\ \frac{\partial lnL}{\partial \sigma^2} &= \frac{-N}{2\sigma^2} + \frac{1}{2}\sigma^4\sum_{i=1}^{N}(Y_i - \mu)^2. \end{aligned}$$

Maximizing the log-likelihood function (continued)

- Setting the previous two expressions equal to zero and solving for the unknowns gives:

$$\begin{aligned}\hat{\mu} &= \frac{1}{N}\sum_{i=1}^{N} Y_i, \\ \hat{\sigma^2} &= \frac{1}{N}\sigma_{i=1}^{N}(Y_i - \bar{Y})^2,\end{aligned}$$

which are simply the formulas for mean and variance.

- That is, our estimates of $\hat{\mu}$ and $\hat{\sigma^2}$ are the maximum likelihood estimates for μ and σ^2.

Conclusion

- We'd like to be able to estimate the likelihood of having observed some parameters of interest given our data.
- This problem of inverse probability turns out to be intractable, however.
- The concept of likelihood helps us to address the problem of relative uncertainty.
- And maximum likelihood inference tells us which parameters are most likely to summarize the relationships in our data, given that data.
- Next time, we'll discuss MLE with respect to regression analysis and some of the properties of these estimates.