Live Session 7

The probability function returns probabilities of the data, given the sample size and the parameters, while the likelihood function gives the relative likelihoods for different values of the parameter, given the sample size and the data.

In other words, probability is used before data are available to describe possible future outcomes given a fixed value for the parameter (or parameter vector). Likelihood is used after data are available to describe a function of a parameter (or parameter vector) for a given outcome.

Be careful not to confuse the probability function and the likelihood function; the right hand sides are the same, but there are differences in the conditioning of the left hand sides.

Let X1, . . . , Xn be an iid sample with probability density function (pdf) f(xi; θ), where θ is a parameter.

- The joint density of the sample is, by independence, equal to the product of the marginal densities

$$f(x_1, \ldots, x_n; \theta) = f(x_1; \theta) \cdots f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

- If X1,..., Xn are discrete random variables, then f(x1,...,xn;θ) = Pr(X1 = x1,...,Xn = xn) for a fixed value of θ, then the joint density satisfies

$$\begin{aligned} f(x_1, \ldots, x_n; \theta) &\geq 0 \\ \int \cdots \int f(x_1, \ldots, x_n; \theta) dx_1 \cdots dx_n &= 1. \end{aligned}$$

- The likelihood function is defined as the joint density treated as a function of the parameters θ:

$$L(\theta | x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

  Remember: the likelihood function, being a function of θ and not the data, is not a proper pdf. It is always positive but

$$\int \cdots \int L(\theta | x_1, \ldots, x_n) d\theta_1 \cdots d\theta_k \neq 1.$$

Example: Bernoulli

Let Xi˜ Bernoulli(θ). That is, Xi = 1 with probability θ and Xi = 0 with probability1−θ where 0≤θ≤1. The pdf for Xi is

$$f(x_i;\theta) = \theta^{x_i}(1-\theta)^{1-x_i},\ x_i = 0,1$$

Let X1, . . . , Xn be an iid sample with Xi˜ Bernoulli(θ). The joint density/likelihood function is given by

$$f(\mathbf{x};\theta) = L(\theta|\mathbf{x}) = \prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n}x_i}(1-\theta)^{n-\sum_{i=1}^{n}x_i}$$

For a given value of θ and observed sample x, f (x; θ) gives the probability of observing the sample.

Likelihood of Bernoulli Random Variables

Recall that the likelihood is the probability of data given the model and specific parameter values.

Suppose that Y1, . . . , Yn are independent and identically distributed Bernoulli random variables with common pmf $p^y(1-p)^{(1-y)}$ where $p \in [0, 1]$ is the probability of "success."

An intuitively appealing estimate of p is the value of p that maximizes L(p), the likelihood. In other words, the value of p that maximizes the probability of the sample $y_1, \ldots, y_n$.

Example 1:
Suppose we have 6 tosses of a coin.

Model:

1. Possible outcomes of each coin toss are heads (H) and tails (T)
2. Results of coin tosses are independent of one another
3. Probability of heads is denoted by p and this probability is the same for each toss.

Let's assume that our data for X is given by: H T T T H T.
The likelihood is given by $Pr(X|p) = p \bullet (13\text{-}p) \bullet (1\text{-}3p) \bullet (13\text{-}p) \bullet p \bullet (1\text{-}3p) = p^2 \bullet (1-p)^4$
The value of p that maximizes the likelihood is 2/(2+4).

Note: the value of p that maximizes the likelihood is the same value that maximizes the log likelihood.

Example 2:
Suppose we have a random sample $X_1, X_2, \ldots, X_n$ where:

- $X_i = 0$ if a randomly selected student does not own a sports car, and
- $X_i = 1$ if a randomly selected student does own a sports car.

Assuming that the X_i are independent Bernoulli random variables with unknown parameter p, find the maximum likelihood estimator of p, the proportion of students who own a sports car.

If the X_i are independent Bernoulli random variables with unknown parameter p, then the probability mass function of each X_i is:

$f(x_i;p) = p^{x_i}(1-p)^{1-x_i}$ for $x_i = 0$ or 1 and $0 < p < 1$.

Therefore, the likelihood function $L(p)$ is, by definition:

$$L(p) = \prod_{i=1}^{n} f(x_i, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Based on the sample given, a reasonable estimate of the long run proportion of students who own cars, p, is L(p).

Question from Class:
Does increasing n imply that the histogram of a Poisson random variable will be approximately normal?

Here are two theorems about the Poisson distribution that show why this is true.

Poisson as a limiting case of binomial

Theorem 1
Let λ be a fixed positive constant. Then for each integer x = 0,1,2,…, the following is true:

$$\lim_{n\to\infty}\binom{n}{x} = p^x(1-p)^{n-x} = \lim_{n\to\infty}\frac{n!}{x!\,(n-x)!}p^x(1-p)^{n-x} = \frac{(\lambda)^x e^{-\lambda}}{x!}$$

where $p = \frac{\lambda}{n}$.

According to the theorem the Poisson distribution is the limiting case of the binomial distribution. Given a binomial distribution where the number of trials n is large and the probability of success p is small, np is moderate in size and can be approximated using the Poisson distribution with mean λ=np.

In other words, if the probability of a success of a single trial, p, approaches 0 while the number of trials increases, i.e. approaches infinity, and the value μ = np stays fixed, then the binomial distribution B(n;p) approaches the Poisson distribution with mean *μ.*

As n increases, the binomial distribution approaches a normal distribution. Hence, Theorem 1 implies that a Poisson distribution will approach a normal distribution.

Theorem 2:

If X has a Poisson distribution with mean μ, then X ~ N(μ, $\sqrt{\mu}$), then for a sufficiently large n (i.e. n ≥ 20), X ~ N(μ, $\sqrt{\mu}$).