Chapter 7. Statistical Estimation

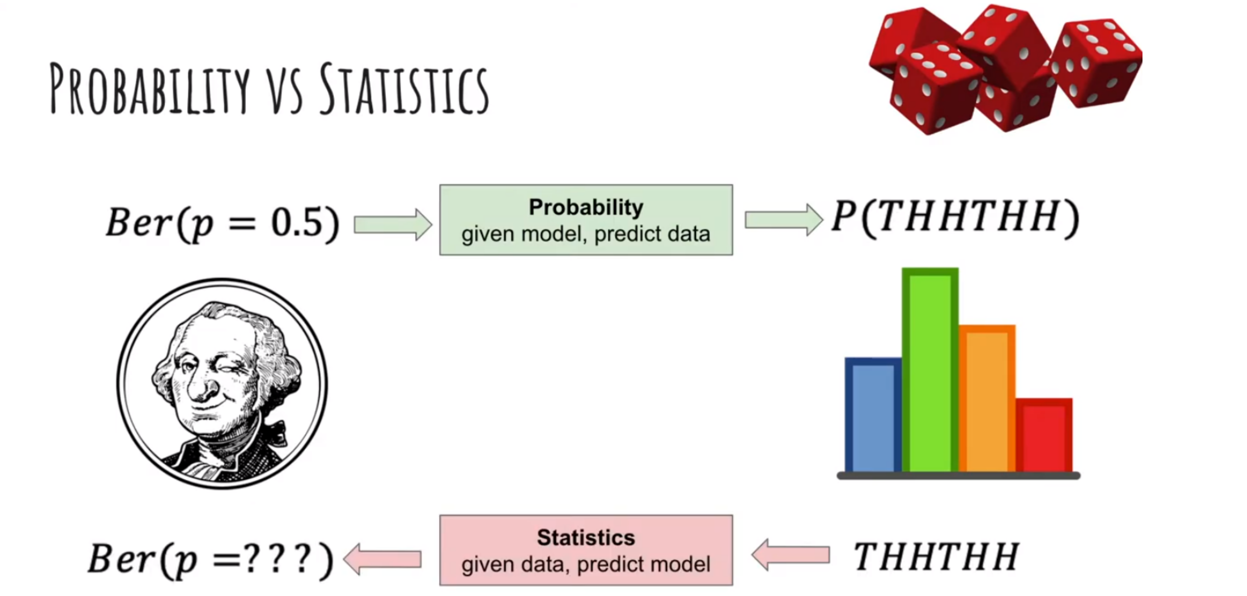
7.1: Maximum Likelihood Estimation

Slides (Google Drive) Alex Tsun Video (YouTube)

7.1.1 Probability vs Statistics

Before we start, we need to make an important distinction: what is the difference between probability and statistics? What we've been doing up until this point is probability. We're given a model, in this picture Ber($p = 0.5$) (our assumption), and we're trying to find the probability of some data. So, given this model, what is the probability of THHTHH, or $\mathbb{P}$ (THHTHH)? That's something you know how to do now!

What we're going to focus now is going the opposite way. Given a coin with unknown probability of heads is, I flip it a few times and I get THHTHH. How can I use this data to predict/estimate this value of p?



PROBABILITY VS STATISTICS

$Ber(p = 0.5)$ → **Probability** given model, predict data → $P(THHTHH)$

$Ber(p = ???)$ ← **Statistics** given data, predict model ← $THHTHH$

7.1.2 Likelihood

Let's say I give you and your classmates each 5 minutes with a coin with unknown probability of heads p. Whoever has the closest estimate will get an A+ in the class. What do you do in your precious 5 minutes, and what do you give as your estimate?

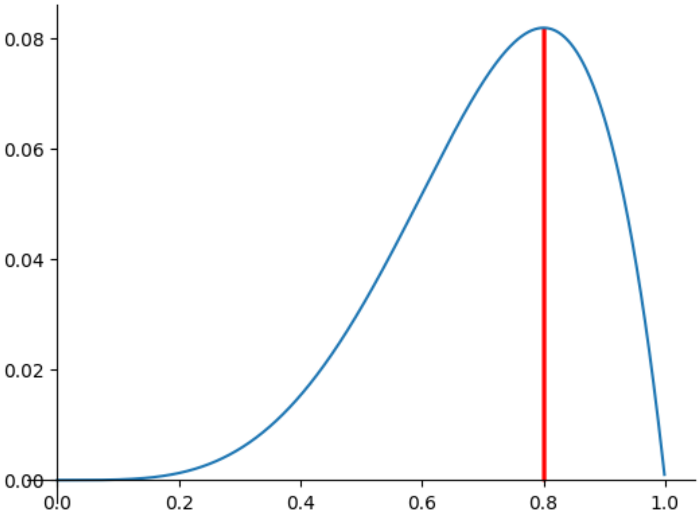
I don't know about you, but I would flip the coin as many times as I can, and return the total number of heads over the total number of flips, or

$$\frac{\text{Heads}}{\text{Heads} + \text{Tails}}$$

which actually turns out to be a really good estimate.

To make things concrete, let's say you saw 4 heads and 1 tail. You tell me that $\hat{p} = \frac{4}{5}$ (the hat above the p just means it is an estimate). How can you argue, objectively, that this is the "best" estimate?

Is there some objective function that it maximizes? It turns out yes, $\frac{4}{5}$ maximizes this blue curve, which is called the likelihood of the data. The x-axis has the different possible values of p, and they y-axis has the probability of seeing the data if the coin had probability of heads p.



You assume a model (Bernoulli in our case) with unknown parameter θ (the probability of heads), and receive iid samples $\mathbf{x} = (x_1, ..., x_n) \sim \text{Ber}(\theta)$ (in this example, each x_i is either 1 or 0). The likelihood of the data given a parameter θ is defined as the probability of seeing the data, given θ, or:

$$\begin{aligned} L(\mathbf{x} \mid \theta) &= \mathbb{P}\left(\text{seeing data} \mid \theta\right) && \text{[def of likelihood]} \\ &= \mathbb{P}\left(x_1, ..., x_n \mid \theta\right) && \text{[plug in data]} \\ &= \prod_{i=1}^{n} p_X(x_i \mid \theta) && \text{[independence]} \end{aligned}$$

(**Note**: When estimating unknown parameters, we typically use θ instead of p, λ, μ, etc.)

Definition 7.1.1: Realization / Sample

A realization/sample x of a random variable X is the value that is actually observed (will always be in Ω_X).

For example, for Bernoulli, a realization is either 0 or 1, and for Geometric, some positive integer ≥ 1.

Definition 7.1.2: Likelihood

Let $\mathbf{x} = (x_1, ..., x_n)$ be iid samples from probability mass function $p_X(t \mid \theta)$ (if X is discrete), or from density $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). We define the likelihood of $\mathbf{x}$ given θ to be the "probability" of observing $\mathbf{x}$ if the true parameter is θ.
If X is discrete,

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i \mid \theta)$$

If X is continuous,

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

In the continuous case, we have to multiply densities, because the probability of seeing a particular value with a continuous random variable is always 0. We can do this because the density preserves relative probabilities; i.e., $\dfrac{\mathbb{P}(X \approx u)}{\mathbb{P}(X \approx v)} \approx \dfrac{f_X(u)}{f_X(v)}$. For example, if $X \sim \mathcal{N}(\mu = 3, \sigma^2 = 5)$, the realization $x = -503.22$ has much lower density/likelihood than $x = 3.12$.

Example(s)

Give the likelihoods for each of the samples, and take a guess at which value of θ maximizes the likelihood!

1. Suppose $\mathbf{x} = (x_1, x_2, x_3) = (1, 0, 1)$ are iid samples from $\text{Ber}(\theta)$ (recall θ is the probability of a success).
2. Suppose $\mathbf{x} = (x_1, x_2, x_3, x_4) = (3, 0, 2, 7)$ are iid samples from $\text{Poi}(\theta)$ (recall θ is the historical average number of events in a unit of time).
3. Suppose $\mathbf{x} = (x_1, x_2, x_3) = (3.22, 1.81, 2.47)$ are iid samples from $\text{Exp}(\theta)$ (recall θ is the historical average number of events in a unit of time).

Solution

1. The samples mean we got a success, then a failure, then a success. The likelihood is the "probability" of observing the data.

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{3} p_X(x_i \mid \theta) = p_X(1 \mid \theta) \cdot p_X(0 \mid \theta) \cdot p_X(1 \mid \theta) = \theta(1-\theta)\theta = \theta^2(1-\theta)$$

Since we observed two successes out of three trials, my guess for the maximum likelihood estimate would be $\hat{\theta} = \dfrac{2}{3}$.

2. The samples mean we observed 3 events in the first unit of time, then 0 in the second, then 2 in the third, then 7 in the fourth. The likelihood is the "probability" of observing the data (just multiplying Poisson PMFs $p_X(k \mid \lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$).

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{4} p_X(x_i \mid \theta) = p_X(3 \mid \theta) \cdot p_X(0 \mid \theta) \cdot p_X(2 \mid \theta) \cdot p_X(7 \mid \theta)$$

$$= \left(e^{-\theta}\frac{\theta^3}{3!}\right)\left(e^{-\theta}\frac{\theta^0}{0!}\right)\left(e^{-\theta}\frac{\theta^2}{2!}\right)\left(e^{-\theta}\frac{\theta^7}{7!}\right)$$

Since there were a total of $3+0+2+7=12$ events over 4 units of time (samples), my guess for the maximum likelihood estimate would be $\hat{\theta}=\frac{12}{4}=3$ events per unit time.

3. The samples mean we waited until three events happened (x_1, x_2, x_3), and it took 3.22 units of time until the first event, 1.81 until the second, and 2.47 until the third. The likelihood is the "probability" of observing the data. The likelihood is the "probability" of observing the data (just multiplying Exponential PDFs $f_X(y \mid \lambda)=\lambda e^{-\lambda y}$).

$$L(\mathbf{x} \mid \theta)=\prod_{i=1}^{3} f_X(x_i \mid \theta)=f_X(x_1 \mid \theta)\cdot f_X(x_2 \mid \theta)\cdot f_X(x_3 \mid \theta)=\left(\theta e^{-3.22\theta}\right)\left(\theta e^{-1.81\theta}\right)\left(\theta e^{2.47\theta}\right)$$

Since it took an average of $\frac{3.22+1.81+2.47}{3}=2.5$ units of time to observe each events, my guess for the maximum likelihood estimate would be $\hat{\theta}=\frac{3}{7.5}=0.4$ events happen per unit of time.

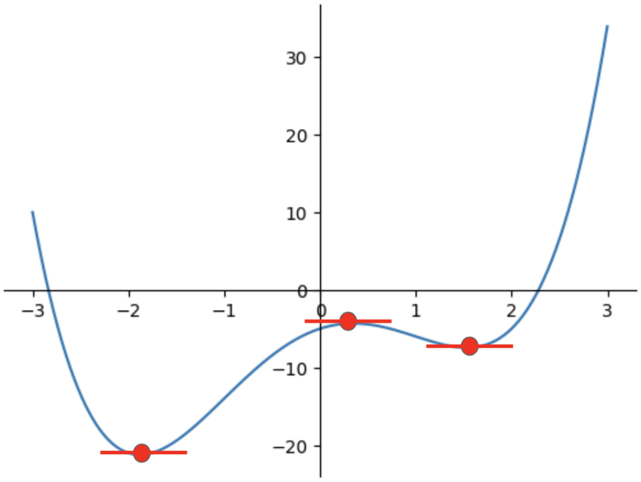
□

7.1.3 Maximum Likelihood Estimation

Now, we'll formally define what the maximum likelihood estimator of an unknown parameter is. Intuitively, it is just the value of θ which maximizes the "probability" of seeing the data $L(\mathbf{x} \mid \theta)$.

In the previous three scenarios, we set up the likelihood of the data. Now, the only thing left to do is find out which value of θ maximizes the likelihood. Everything else in this section is just explaining how to use calculus to optimize this likelihood! There is no more "probability" or "statistics" involved in the remaining pages.

Before we move on, we have to go back and review calculus really quickly. How do we optimize a function? Each of these three points is a local optima; what do they have in common? Their derivative is 0. We're going to try and set the derivative of our likelihood to 0, so we can solve for the optimum value.



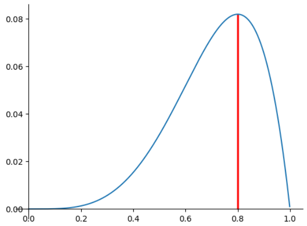
Example(s)

Suppose $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5) = (1, 1, 1, 1, 0)$ are iid samples from the $\text{Ber}(\theta)$ distribution with unknown parameter θ. Find the maximum likelihood estimator $\hat{\theta}$ of θ.

Solution The data $(1, 1, 1, 1, 0)$ can be thought of the sequence $HHHHT$, which has likelihood (assuming the probability of heads is θ)

$$\begin{aligned} L(HHHHT \mid \theta) &= \theta^4(1-\theta) \\ &= \theta^4 - \theta^5 \end{aligned}$$

The plot of the likelihood with θ on the x-axis and $L(HHHHT \mid \theta)$ on the y-axis is (copied from above):



and we can actually see the θ which maximizes the likelihood is $\hat{\theta} = 4/5$. But sometimes we can't plot the likelihood, so we will solve for this analytically now.

We want to find the θ which maximizes this likelihood, so we take the derivative with respect to θ and set it to 0:

$$\begin{aligned} \frac{\partial}{\partial \theta} L(\mathbf{x} \mid \theta) &= 4\theta^3 - 5\theta^4 \\ &= \theta^3(4 - 5\theta) \end{aligned}$$

Now, when we set the derivative to 0 (remember the optimum points occur when the derivative is 0), we replace θ with $\hat{\theta}$ because we are now estimating θ. After solving for θ, we end up with

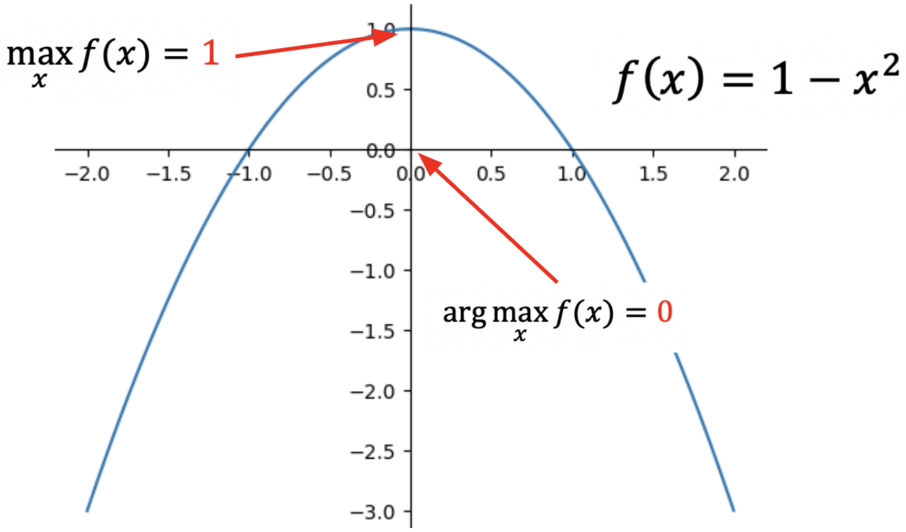
$$\hat{\theta}^3(4 - 5\hat{\theta}) = 0 \rightarrow \hat{\theta} = \frac{4}{5} \text{ or } 0$$

We switch θ to $\hat{\theta}$ when we set the derivative to 0, as that is when we start estimating. To see which is the maximizer, you can just plug in the candidates (0 and 4/5) and the endpoints (0 and 1: the min and max possible values of θ)! That is, compute the likelihood at $0, 4/5, 1$ and see which is largest. □

To summarize, we defined $\hat{\theta}_{MLE} = \arg\max_\theta L(\mathbf{x} \mid \theta)$, the **arg**ument (input) θ that **max**imizes the likelihood function. The difference between max and argmax is as follows. Here is a function,

$$f(x) = 1 - x^2$$

where the maximum value is 1, it's the highest value this function could ever achieve. The argmax, on the other hand, is 0, because argmax just means the *argument* (input) that maximizes the function. So, which x actually achieved $f(x) = 1$? Well that was $x = 0$. And so, in MLE, we're trying to find the θ that maximizes the likelihood, and we don't care what the maximum value of the likelihood is. We didn't even compute it! We just care that the argmax is $\frac{4}{5}$.



Definition 7.1.3: Maximum Likelihood Estimation (MLE)

Let $\mathbf{x} = (x_1, ..., x_n)$ be iid realizations from probability mass function $p_X(t \mid \theta)$ (if X is discrete), or from density $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). We define the **maximum likelihood estimator** $\hat{\theta}_{MLE}$ of θ to be the parameter which maximizes the likelihood (or equivalently, the log-likelihood) of the data.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\mathbf{x} \mid \theta) = \arg\max_{\theta} \ln L(\mathbf{x} \mid \theta)$$

The (usual) recipe to find the MLE goes as follows:

1. Compute the likelihood and log-likelihood of data.
2. Take the partial derivative(s) with respect to θ and set to 0. Solve the equation(s).
3. Optionally, verify $\hat{\theta}_{MLE}$ is indeed a (local) maximizer by checking that the second derivative at $\hat{\theta}_{MLE}$ is negative (if θ is a single parameter), or the Hessian (matrix of second partial derivatives) is negative semi-definite (if θ is a vector of parameters).

7.1.3.1 Optimizing Function vs Log(Function)

You may have notice we also included this “log-likelihood” that we hadn’t talked about earlier. In the next section, we’ll do several more examples of maximum likelihood estimation, and you’ll see that taking the log makes our derivatives easier. Recall that the likelihood is the product of PDFs or PMFs, and taking the derivative of a product is quite annoying, especially with more than 2 terms:

$$\frac{d}{dx}(f(x) \cdot g(x)) = f'(x)g(x) + f(x)g'(x)$$

whereas the derivative of the sum is just the sum of the derivatives:

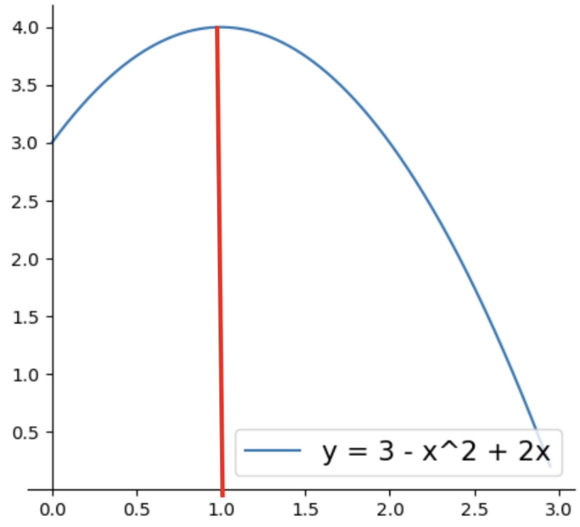
$$\frac{d}{dx}(f(x) + g(x)) = f'(x) + g'(x)$$

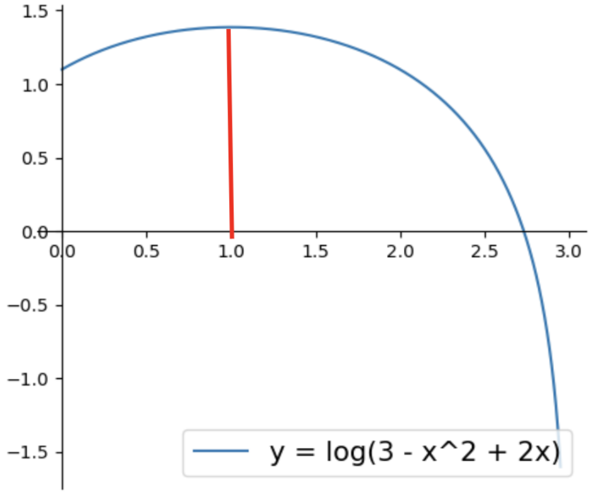
Taking the log of a product (such as the likelihood) results in the sum of logs because of log properties:

$$\log(a \cdot b \cdot c) = \log(a) + \log(b) + \log(c)$$

We see now why we might **want** to take the log of the likelihood before differentiating it, but why **can** we?

Below there are two images: the left image is a function, and the right image is the log of that function.





The values are different (see the y-axis), but if you look at the x-axis, it happens that both functions are maximized at 1 (the argmax's are the same). Log is a monotone increasing function, so it preserves order, so whatever was the maximizer (argmax) in the original function, will also be maximizer in the log function.

See below to see what happens when you apply the natural log (ln) to a product in our likelihood scenario! And see the next section 7.2 for examples of maximum likelihood estimation in action.

Definition 7.1.4: Log-Likelihood

Let $\mathbf{x} = (x_1, ..., x_n)$ be iid realizations from probability mass function $p_X(t \mid \theta)$ (if X is discrete), or from density $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). We define the likelihood of $\mathbf{x}$ given θ to be the probability of observing $\mathbf{x}$ if the true parameter is θ.
If X is discrete,

$$\ln L(\mathbf{x} \mid \theta) = \sum_{i=1}^{n} \ln p_X(x_i \mid \theta)$$

If X is continuous,

$$\ln L(\mathbf{x} \mid \theta) = \sum_{i=1}^{n} \ln f_X(x_i \mid \theta)$$