Topic 18: Parameter Estimation and Maximum Likelihood Estimation

02-680: Essentials of Mathematics and Statistics

December 10, 2024

What is a "Statistic"?

Definition: Anything that can be computed from the collected data (i.e., must be observable).

Statistics deals with data.

Goal: Make inferences based on data

This process can be divided into three overlapping phases

(1) **Collecting data** — collect data in experiments; Preceded by forming hypotheses about phenomena of interest

(2) **Describing data** — describe the results

(3) **Analyzing data** — infer from the results the strength of the evidence with respect to the hypotheses

Generally two types of statistics

Point statistic — a single value computed from data (e.g., $\overline{X_n}$)

Interval or range statistics — an interval $a \leq x \leq b$ computed from the data

Note that a statistic is itself a random variable because a new experiment will produce new data to compute it.

1 Some Basics

Consider a dataset $X_1, X_2, \cdots, X_n$ of i.i.d. (independent and identically distributed) random variables from the same unknown distribution. That is, for each X_is, the underlying distributions have the same μ and σ.

Let $\overline{X_n}$ be the average of the actual value of the observations:

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Note this is different from the expected value of the underlying distribution μ. Note also that $\overline{X_n}$ is also a random variable in and of itself.

Example. Consider a dataset $C_1, C_2, \cdots, C_{100}$ of 100 fair coin flips (chosen iid), with the corresponding random variables taking on values of 1 for Heads and 0 for Tails. Assume 48 of the 100 coins are heads, then $\overline{C_{100}} = \frac{48}{100}$.

Law of Large Numbers. The ***law of large numbers*** says that as n grows, the probability that $\overline{X_n}$ is close to μ approaches 1.

$$\lim_{n\to\infty} p\left(\overline{X_n} = \mu\right) \mapsto 1.$$

Central Limit Theorem (brief introduction). While similar to the law of large numbers the ***Central Limit Theorem*** says that n grows, the distribution of $\overline{X_n}$ converges to the Gaussian (normal) distribution $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

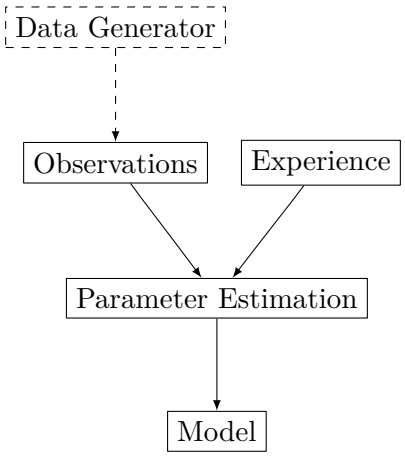
$$\lim_{n\to\infty} \overline{X_n} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

These both say something similar and speak to the fact that more data will make the empirical mean will approach the underlying (theoretical) mean. But the Central Limit Theorem is more precise, it allows us to approximate the probability that the sample size is not adequate (gives some confidence on the probability from the LLN). We will see much more about this over the next couple weeks.

2 Statistical Inference

We've been talking a lot about the models themselves and hinting at how to use those distributions. We will use ***Statistical Interference*** or ***Learning*** to try to make the models we've been learning about match the observations we've made about the world.

Thinking about the coin flip simulator from the other day: 1 out of 1000 times it outputs a nonsense number, the other outcomes are split between heads and tails. We call the simulator a *data generator* that produces things we can see which we call *observations*. On the other side, we have a set of models, one of which we think may match our intuition about the generator, remember that each of those models had a set of *parameters*. Statical inference is the task of trying to learn the parameters from the observations.



Assume you have a set

$$\mathcal{D} = X_1, X_2, \cdots, X_n$$

of i.i.d. random variables from some unknown distribution.

Let $\mathcal{F}$ be a statistical model defined as a set of probability distributions.

We will focus on parametric models defined as

$$\mathcal{F}\{f(x \mid \theta) : \theta \in \Theta\},$$

where

$$\theta = \{\theta_1, \theta_2, \cdots, \theta_k\}$$

are unknown parameters $\theta \in \Theta$.

The task is to find a distribution $\hat{f} \in \mathcal{F}$ that models the phenomenon well, which includes modeling the associated θ.

We want to do this in a way that has:

- the ability to generalize well,
- the ability to incorporate prior knowledge and assumptions, and
- the ability to scale.

2.1 Data Likelihood

Remember Bayes' Theorem (written slightly differently):

$$p(hypothesis \mid data) = \frac{p(data \mid hypothesis)p(hypothesis)}{p(data)}$$

If we can figure out all the values on the right, we can get the probability of interest on the left (how good our model parameters are given the data).

Lets start by looking at the most complicated component of the RHS: $p(data \mid hypothesis)$, or in our case $p(\mathcal{D} \mid \theta)$. Remember that the elements of $\mathcal{D}$ are i.i.d. thus we can rewrite it:

$$p(\mathcal{D} \mid \theta) = p(X_1, X_2, \cdots, X_n \mid \theta) = \prod_{i=1}^{n} p(X_i \mid \theta)$$

log likelihood

As we saw in the activity last week, when taking the product of probabilities the values can get very small. So we many times take the log likelihood of the function. This also take the product to a sum.

$$\log(p(X_1, X_2, \cdots, X_n \mid \theta)) = \sum_{i=1}^{n} \log(p(X_i \mid \theta))$$

3 Maximum Likelihood

The ***maximum likelihood estimate*** (MLE) is a way to estimate the value of a parameter of interest θ. This is the most frequently used method for parameter estimation.

Thinking about how we maximize a function (or really find the parameter that maximizes the function), we need to take its derivative and check the 0 points.

$$\frac{d}{d\theta} p(\mathcal{D} \mid \theta) = 0$$

(but we also need to ensure its a max and not a min or saddle.) If there is no critical point, then we will use the maximum allowable parameter value.

Example. Lets look at flipping a coin n times, but we don't know if its fair or not. So each of the $\mathcal{D} = X_1, X_2, \cdots, X_n$ trials are independent, thus $X_i \sim Bernouli(\alpha)$, and thus $f(X_i = x \mid \theta) = \alpha^x (1-\alpha)^{1-x}$, here $\theta = \{\alpha\}$. But we also combine the trials, so it ends up looking a little like the binomial distribution, assuming k of our n flips were heads:

$$p(\mathcal{D} \mid \theta) = \binom{n}{k} \alpha^k (1-\alpha)^{n-k}$$

and we want to estimate α.

Notice that while through Bayes' rule we need $p(\theta)$ and $p(\mathcal{D})$ the former is typically uniform and the latter does not change thus neither need to be considered when optimizing.

So lets assume that we have a sequence $\mathcal{D} = X_1, X_2, \cdots, X_{10} = 1, 1, 1, 0, 0, 0, 0, 0, 1, 0$. So we want to know

$$\underset{\theta}{\operatorname{argmax}} f(\mathcal{D} \mid \theta) = \underset{\theta}{\operatorname{argmax}} \left(\binom{10}{4} \alpha^4 (1-\alpha)^6 \right)$$

To find θ we using calculus we need the derivative:

$$\frac{d}{d\theta} \left(\binom{10}{4} \alpha^4 (1-\alpha)^6 \right) = \binom{10}{4} \left(4\alpha^3 \cdot (1-\alpha)^6 - 6\alpha^4 (1-\alpha)^5 \right)$$

Doing some algebra (notice the binomial coefficient is a constant thus does not impact the maximal θ)

$$\begin{aligned} 4\alpha^3 \cdot (1-\alpha)^6 &= 6\alpha^4 (1-\alpha)^5 \\ 4(1-\alpha) &= 6\alpha \\ 4 &= 10\alpha \end{aligned}$$

So the zero point(s) are 4/10, since thats a maximum we know $\hat{\alpha} = \frac{4}{10}$. Thus, given the sample we have, it looks like the coin is not fair.

4 Deriving MLE for generic distributions

Remember the basic steps to finding the MLE are:

- Write down the log likelihood of the data, then
- Maximize log likelihood (usually using calculus.

4.1 Bernoulli

Can we make the derivation in the example above more general.

Assume we have a set $\mathcal{D} = X_1, X_2, \cdots, X_n$ where $X_i \sim Bernouli(\alpha)$. Find the MLE of $\theta = \{\alpha\}$.

$$
\begin{aligned}
\frac{d}{d\theta}\ln p(\mathcal{D}\mid\theta) &= \frac{d}{d\theta}\ln\prod_{i=0}^{n} p(X_i\mid\theta) \\
&= \frac{d}{d\alpha}\ln\alpha^s(1-\alpha)^{n-s} && \text{assuming } s \text{ successes} \\
&= \frac{d}{d\alpha}s\ln\alpha + (n-s)\ln(1-\alpha) \\
&= \frac{s}{\alpha} - \frac{n-s}{1-\alpha} && \text{log rule, assuming } \ln\frac{d}{dx}a\ln f(x) = \frac{a}{x}\frac{d}{dx}f(x)
\end{aligned}
$$

Therefore, for a set of n Bernulli random variables with s successes, (when we solve the previous equation equal to 0)

$$\hat{\alpha} = \frac{s}{n}$$

(which aligns with the previous example).

4.2 Normal (Gaussian)

Assume we have a set $\mathcal{D} = X_1, X_2, \cdots, X_n$ where $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Find the MLE of $\theta = \{\mu, \sigma^2\}$.

$$
\begin{aligned}
p(\mathcal{D}\mid\mu,\sigma^2) &= \prod_{i=1}^{n} p(X_i\mid\mu,\sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\ e^{\frac{(x-\mu)^2}{-2\sigma^2}} \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{\frac{1}{-2\sigma^2}\sum_{i=1}^{n}(x-\mu)^2}
\end{aligned}
$$

$$
\begin{aligned}
\ln p(\mathcal{D}\mid\mu,\sigma^2) &= \ln\left(\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{\frac{1}{-2\sigma^2}\sum_{i=1}^{n}(x-\mu)^2}\right) \\
&= \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x-\mu)^2 \\
&= -n\ln\sigma - n\ln\sqrt{2\pi} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x-\mu)^2
\end{aligned}
$$

Then we need to do two derivatives:

$$\begin{aligned}
\frac{d}{d\mu} -n\ln\sigma - n\ln\sqrt{2\pi} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x-\mu)^2 &= 0 \\
\frac{d}{d\mu}\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x-\mu)^2 &= 0 \\
\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(2\cdot(x-\mu)\cdot(-1)) &= 0 \\
\sum_{i=1}^{n}(x-\mu) &= 0
\end{aligned}$$

Thus $\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

$$\begin{aligned}
\frac{d}{d\sigma} - n\ln\sigma - n\ln\sqrt{2\pi} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x-\mu)^2 &= 0 \\
\frac{d}{d\mu} - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x-\mu)^2 &= 0 \\
\frac{-n}{\sigma} + \sigma^{-}3\sum_{i=1}^{n}(x-\mu)^2 &= 0 \\
\sigma^{-}2\sum_{i=1}^{n}(x-\mu)^2 &= n
\end{aligned}$$

Thus $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x-\hat{\mu}_{MLE})^2$.

4.3 Poisson

Assume we have a set $\mathcal{D} = X_1, X_2, \cdots, X_n$ where $X_i \sim Poisson(\lambda)$. Find the MLE of $\theta = \{\lambda\}$.

$$\begin{aligned}
\ln p(\mathcal{D} \mid \mu, \sigma^2) &= \ln \left(\prod_{i=1}^{n} p(X_i \mid \mu, \sigma^2) \right) \\
&= \ln \left(\prod_{i=1}^{n} e^{-\lambda} \frac{\lambda_i^x}{x_i!} \right) \\
&= \ln \left(e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \right) \\
&= -n\lambda + \ln \lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!)
\end{aligned}$$

Then find the max (0 of the derivative)

$$\begin{aligned}
\frac{d}{d\lambda} \left(-n\lambda + \ln \lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!) \right) &= 0 \\
-n + \frac{1}{\lambda} \sum_{i=1}^{n} x_i &= 0 \\
\sum_{i=1}^{n} x_i &= n\lambda
\end{aligned}$$

And thus $\hat{\lambda}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Useful References

Wasserman. "All of Statistics: A Concise Course in Statistical Inference" §9 Degroot and Schervish. "Probability and Statistics" §§7.5