# Lecture 7: Estimation

## STATS 101: Foundations of Statistics

Linh Tran
linh@thetahat.ai

January 15, 2019

# Announcements

- Next assignment is posted (due 1/22 @ 9:00am)
  - Partner pairing happens now.
- A *Colab* script is available for today's class.

# Outline

Estimation

- Random sample
- Statistical inference
- Point estimation
- Maximum likelihood

# Recall

For rv $X$, we have

$$F_X : \mathbb{R} \to [0,1], \text{ i.e. the cdf}$$
$$f_X : \mathbb{R} \to [0,1], \text{ i.e. the pdf}$$

For multiple rv $X_1, \ldots, X_n$, we have $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ and $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$. Note that:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)) \tag{1}$$
$$\tag{2}$$

where $Pa(X_i) \triangleq (X_1, ..., X_j) : j < i$ and $Pa(X_i) = \emptyset$

# Random samples

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an n-dimensional vector. We tend to think of $X_1, \ldots, X_n$ as a *random sample*, such that

$$X_i \overset{iid}{\sim} P_0 : i = 1, \ldots, n \tag{3}$$

Consequently, we have that

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \underbrace{=}_{\text{indep}} \prod_{i=1}^{n} F_{X_i}(x_i) \underbrace{=}_{\text{identical dist}} \prod_{i=1}^{n} F(x_i) \tag{4}$$

*Colab link*

# Statistical inference

A *parameter* is a mapping, $\theta : \mathbb{R}^p \to P_\theta \ni$

$$\mathcal{M} = \{P_\theta : \theta \in \Theta\} \tag{5}$$

# Statistical inference

A *parameter* is a mapping, $\theta : \mathbb{R}^p \to P_\theta \ni$

$$\mathcal{M} = \{P_\theta : \theta \in \Theta\} \tag{5}$$

More generally, we can define a *parameter* as the mapping

$$\Psi : \mathcal{M} \to \mathbb{R} \tag{6}$$

**This allows us to define a target parameter** $\psi_0 = \Psi(P_0)$.

# Statistical inference

A *statistic* is a mapping, $T : \mathbb{R}^n \to \mathbb{R}^p : p \geq 1$.

The random variable

$$Y = T(X_1, \ldots, X_n) \tag{7}$$

is a statistic.

Its distribution is called the *sampling distribution* of $Y$.

*Colab link*

**Examples:**

- The sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{8}$$

- The sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{9}$$

- The sample standard deviation:

$$s = \sqrt{s^2} \tag{10}$$

# Statistical inference

Statistics can be the an observed value, e.g. the max:

$$Y = \max_{i \leq n} X_i \tag{11}$$

Statistics can be the entire data, e.g. *order statistics*:

$$X_{(1)} = \min_{i \leq n} X_i \leq X_{(2)} \leq \ldots \leq X_{(n)} = \max_{i \leq n} X_i \tag{12}$$

# Statistical inference

All statistics converge to some fixed value, e.g. the sample mean

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{n\mu}{n} = \mu \tag{13}$$

# Statistical inference

All statistics converge to some fixed value, e.g. the sample mean

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{n\mu}{n} = \mu \tag{13}$$

To get the variance of the $\bar{X}$

$$\begin{aligned} var(\bar{X}) &= var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} var(X_i) \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned} \tag{14}$$

# Statistical inference

Statistics (normally) converge to some fixed value, e.g. the sample variance

$$
\begin{aligned}
\mathbb{E}[s^2] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} X_i^2 + \bar{X}^2 - 2X_i\bar{X}\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[n\bar{X}^2 + \sum_{i=1}^{n}(X_i^2) - 2\bar{X}\sum_{i=1}^{n}\left(\frac{n}{n}X_i\right)\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[n\bar{X}^2 + \sum_{i=1}^{n}(X_i^2) - 2n\bar{X}^2\right] \\
&= \frac{1}{n-1}\left(n\mathbb{E}[X_i^2] - n\mathbb{E}[\bar{X}^2]\right) \\
&= \frac{1}{n-1}\left(n\left(\mu^2 + \sigma^2\right) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right) = \frac{n\sigma^2 - \sigma^2}{n-1} \\
&= \sigma^2
\end{aligned}
\tag{15}
$$

# Point estimation

Given $X_1, \ldots, X_n$, we typically assume $X_i \overset{iid}{\sim} P_{\theta_0}$ where $\theta_0 \in \Theta$ is unknown.

A *point estimator* is a function

$$\hat{\theta}_n = \omega(X_1, \ldots, X_n) \tag{16}$$

We refer to the realized value of $\hat{\theta}_n$ as

$$\hat{\theta} = \omega(X_1 = x_1, \ldots, X_n = x_n) \tag{17}$$

**Good $\hat{\theta}_n$ will give $\hat{\theta}$ close to $\theta_0$.**

# Point estimation

Examples of point estimators:

- The mean

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{18}$$

- The median

$$\hat{\theta}_n = X_{(n/2)} \tag{19}$$

- A constant value

$$\hat{\theta}_n = 0 \tag{20}$$

**We want $\hat{\theta}_n$ so that $\hat{\theta}$ is close to $\theta_0$.**

# The likelihood

**Recall**: For $X_1, \ldots, X_n$

$$P(X_1, \ldots, X_n) = P(X_1)P(X_2|X_1) \cdots P(X_n|X_1, \ldots, X_n) \qquad (21)$$

If we assume $X_i \stackrel{iid}{\sim} P_{\theta_0}$, then

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i) \qquad (22)$$

and

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} P_\theta(x_i) \qquad (23)$$

is our *likelihood*.

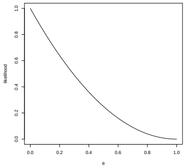
# The likelihood

**Example**: Assume $X_i \overset{iid}{\sim} Ber(\theta) : i = 1, 2$.

$$L(\theta|x_1, x_2) = P_\theta(x_1)P_\theta(x_2) \tag{24}$$

For $X_1 = X_2 = 0$, we can consider the likelihood under different $\theta$, e.g.

$$\begin{aligned} L(0.5|x_1, x_2) &= 0.5 * 0.5 \\ L(0.75|x_1, x_2) &= 0.25 * 0.25 \\ L(0.01|x_1, x_2) &= 0.99 * 0.99 \end{aligned}$$



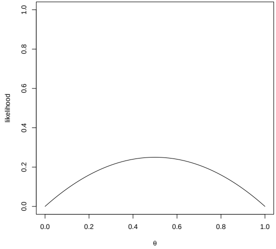

Likelihood for different values of $\theta$.

# The likelihood

**Another example**: $X_1 = 1, X_2 = 0$

$$L(\theta|x_1 = 1, x_2 = 0) = P_\theta(x_1)P_\theta(x_2) \tag{25}$$

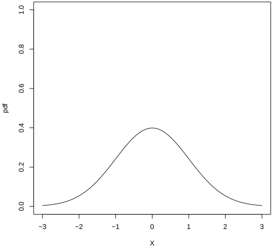
$$= \theta(1-\theta) \tag{26}$$




Likelihood for different values of $\theta$.

# The likelihood

**Example with continuous outcome**: $X_i \sim N(\theta, 1)$




*Colab link*

# Maximum likelihood estimation (MLE)

Our likelihood varies with $\theta$.

**Idea**: why not pick the $\theta$ with the highest value as our estimate?

In other words: We pick the $\theta$ that *maximizes the likelihood* (probability of the observed sample having occurred)

# Maximum likelihood estimation (MLE)

Our likelihood varies with $\theta$.

**Idea**: why not pick the $\theta$ with the highest value as our estimate?

In other words: We pick the $\theta$ that *maximizes the likelihood* (probability of the observed sample having occurred)

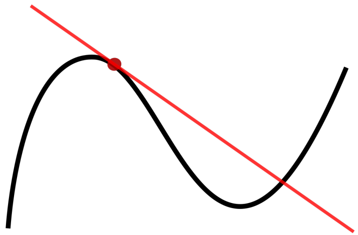
Equivalently, we can maximize the log-likelihood, i.e.

$$\ell(\theta|x_1, \ldots, x_n) = \log L(\theta|x_1, \ldots, x_n) \tag{27}$$

# Maximum likelihood estimation (MLE)

How to find the maximum value:

- Take the derivative of $\ell(\theta|x_1, \ldots, x_n)$.
- Set the derivative equal to 0.
- Solve for $\theta$.



Example function and derivative.

# Maximum likelihood estimation (MLE)

**Example**: $X_1, \ldots, X_n \overset{iid}{\sim} Ber(\theta)$

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \tag{28}$$

$$\ell(\theta|x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i \log(\theta) + (1-x_i)\log(1-\theta) \tag{29}$$

Taking the derivative:

$$\frac{\partial \ell(\theta|x_1, \ldots, x_n)}{\partial \theta} = \sum_{i=1}^{n} \frac{x_i}{\theta} - \frac{1-x_i}{1-\theta} = \sum_{i=1}^{n} \frac{x_i - \theta}{\theta(1-\theta)} \tag{30}$$

Set to 0 and solving for $\theta$:

$$0 = \sum_{i=1}^{n} \frac{x_i - \hat{\theta}}{\hat{\theta}(1-\hat{\theta})} = \sum_{i=1}^{n} x_i - \hat{\theta} \iff \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{31}$$

# MLE properties

1. The MLE might not be found via setting the derivative to zero (e.g. uniform distribution).
2. The MLE might not be analytically found (e.g. Expectation Maximization).
3. The MLE might not exist (e.g. strickly uniform distribution).
4. Equivariance: If $\hat{\theta}$ is the MLE of $\theta_0$, then $g(\hat{\theta})$ is the MLE of $g(\theta_0)$.
5. Consistency: $\hat{\theta} \xrightarrow{p} \theta_0$.
6. Asymptotic normality
$$\frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} \xrightarrow{d} N(0, 1) \tag{32}$$
7. Asymptotic efficiency: MLE has the smallest asymptotic variance among asymptotically normal estimators.

# MLE in practice

**Colab**

# References

- DeGroot & Schervish Chapters 6.5, 6.6