Chapter 5

Estimation

5.1 Introduction to statistical inference

Statistical inference:

What are we trying to infer? It depends on the question of interest.

Example 5.1.1 (GDP per capita)**.** *GDP per capita based on purchasing power parity (PPP). PPP GDP is gross domestic product converted to international dollars using purchasing power parity rates. An international dollar has the same purchasing power over GDP as the U.S. dollar has in the United States. GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2011 international dollars.*

In the data set, we have the data for GDP per capita for 188 countries in the year of 2010. We want to model the GDP per capita for these countries.

Example 5.1.2 (Traffic light)**.** *For a traffic engineers, they want to model the counts of cars in a given time interval or space area.*

- pdf/cdf: defining the probability distribution – lots of information;

- A subset of information given by pdf/cdf;
- Location or scale parameters (e.g., mean, median, mode).

There are three basic types of inference:

- Point estimation;
- Hypothesis testing;
- Confidence intervals;

Basic procedure

1. Data collection;
2. Formulate statistical models;
3. Calculate statistic, which is a function of the data, a random variable;
4. Draw conclusions.

5.2 Parametric Estimation

The distribution of the population can be described by a certain number of parameters, such as the population mean, population variance, and etc. What we want to infer is these quantities regarding the whole population. These quantities are called the parameters. In parametric models, we make lots of assumptions by choosing a probability density function indexed by parameters as the basis of the model.

1. Because there are only a few parameters, we can estimate them very accurately;
2. Because there are only a few parameters, the estimation can go very bad if the distribution is not right.

Either success or failure boils down to the same person

5.3 Method of moments

Principle: Assume a parametric model, match the moments of the distribution to the moments of the sample.

Suppose that the random sample $Y_1, Y_2, \cdots, Y_n$ are i.i.d. observation from a distribution with pdf $f(y)$. The r-th population moment of Y is EY^r. Let the r-th sample moment be

$$m_r = \frac{1}{n} \sum_i Y_i^r.$$

Match the r-th population moments with the r-th sample moments:

$$\begin{cases} m_1 = EY, \\ m_2 = EY^2, \\ \cdots \\ m_r = EY^r. \end{cases}$$

GDP example
Method I: Assume that the GDP per capita follows a normal distribution with mean μ and variance σ^2. Note that $EY = \mu$ and $EY^2 = \mu^2 + \sigma^2$. Then we can estimate μ and σ^2 as

$$\hat{\mu} = m_1, \hat{\sigma}^2 = m_2 - m_1^2.$$

Is this a good estimator?

Method II: Assume that $X \sim LogNormal(\mu, \sigma^2)$. Note that for the log-normal distribution, the mean is $\exp(\mu + \frac{\sigma^2}{2})$ and the second moment is $\exp(\sigma^2)\exp(2\mu + \sigma^2)$. Consequently, the estimator based on the method of moments is

$$\begin{cases} \hat{\sigma}^2 = \log \frac{m_2}{m_1^2} \\ \hat{\mu} = \log m_1 - \frac{\hat{\sigma}^2}{2} \end{cases}$$

Method III:

Assess the performance of estimators.

1. *Bias.* The bias associated with an estimated parameter is defined to be: $\text{Bias}(\hat{\beta}) = E(\hat{\beta}) - \beta$. $\hat{\beta}$ is an unbiased estimator if the mean or the expected value of $\hat{\beta}$ is equal to the true value, that is, $E(\hat{\beta}) = \beta$.

2. *Consistency.* $\hat{\beta}$ is an *consistent* estimator of β if for any $\delta > 0$, $\lim\limits_{n \to \infty} P(|\beta - \hat{\beta}| < \delta) = 1$. As sample size n approaches infinity, the probability that $\hat{\beta}$ will differ from β will get very small.

3. *Efficiency.* We say that $\hat{\beta}$ is an *efficient* unbiased estimator if for a given sample size, the variance of $\hat{\beta}$ is smaller than the variance of any other unbiased estimators.

Tradeoff between bias and variance of estimators. When the goal is to maximize the precision of the predictions, *an estimator with low variance and some bias may be more desirable than an unbiased estimator with high variance.* We may want to minimize the **mean square error**, defined as

$$\text{Mean square error}(\hat{\beta}) = E(\hat{\beta} - \beta)^2.$$

It can be shown that $\text{Mean square error}(\hat{\beta}) = [\text{Bias}(\hat{\beta})]^2 + \text{Var}(\hat{\beta})$. The criterion of minimizing mean square error take into account of both the variance and the bias of the estimator.

Theorem 5.3.1. *Assume that $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} f(x)$ with mean of μ and variance of $\sigma^2 < \infty$, then MOM estimators are unbiased.*

Theorem 5.3.2. *Let $\hat{\beta}$ be an estimator of β such that $MSE(\hat{\beta}, \beta) \to 0$ as $n \to \infty$. Then $\hat{\beta}$ is consistent.*

Consider the setting that $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} LogNormal(\mu, \sigma^2)$. We have two MOM estimators, one based on the moment of X, one based on the moment of $\log(X)$. Which estimator is better? We can use simulation to compare the mean squared error.

Steps of running simulations.

1. Set the sample size $n = 1000$, generate $X_i \overset{\text{iid}}{\sim} LogNormal(\mu, \sigma)$ with $\mu = 0, \sigma = 1$;
2. Estimate the parameters μ, σ using the two different methods;
3. For each estimate, calculate the squared error $(\hat{\mu} - \mu)^2 + (\hat{\sigma} - \sigma)^2$;
4. Replicate steps (1-3) 2000 times and calculate the mean squared error;
5. Based on this simulation, which method works better?

Example 5.3.1. Traffic lights *The Poisson distribution has been used by traffic engineers as a model for light traffic, based on the rationale that if the rate is approximately constant and the traffic is light (so the individual cars move independently of each other), the distribution of counts of cars in a given time interval or space area should be nearly Poisson (Gerlough and Schuhl 1955). The following table shows the number of right turns during 300 3-min intervals at a specific intersection. What is the MOM of the rate parameter?*

n	0	1	2	3	4	5	6
Frequency	14	30	36	68	43	43	30
n	7	8	9	10	11	12	13+
Frequency	14	10	6	4	1	1	0

The method of moments is widely used in many *empirical* Bayes estimation.

Pros of MOM estimators:

- The moment estimator $\hat{\theta}$ exists;
- The estimator is consistent: $\hat{\theta} \xrightarrow{\mathcal{P}} \theta$;

Cons of MOM estimators:

- MOM is not unique
- Sometimes MOM doesn't make sense.

5.4 Maximum likelihood estimation

Definition 5.4.1 (Likelihood)**.** *Likelihood is a joint density of the data viewed as a function of the parameter that characterize the family of distributions.*

Assume that $X \sim Exp(\theta)$. Then

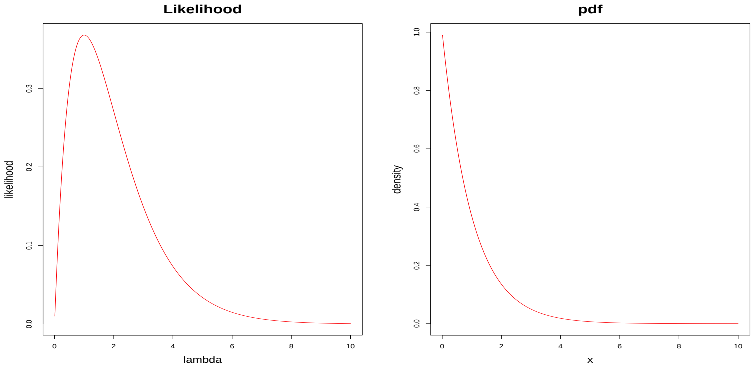

Figure 5.1: Left: Likelihood function; Right: pdf function.

θ	Y		
	-1	0	1
1	0.2	0.3	0.5
2	0.7	0.2	0.1
3	0.2	0.6	0.2

Example 5.4.1. *Consider the following situation for a density of Y indexed by θ: Now consider taking a single sample of Y's. The observed value of Y is $Y = 0$. What is the most likely value of θ?*

Definition 5.4.2 (MLE)**.** *More generally, suppose $Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} f(y, \theta)$, where $\theta \in \Theta$. If the observed values are $y_1, \cdots, y_n$ in the sample, then the likelihood is*

$$L_n(\theta) = \prod_{i=1}^{n} f(y_i, \theta).$$

The maximum likelihood estimator (MLE) is

$$\hat{\theta} = argmax_{\theta \in \Theta} L_n(\theta).$$

It is equivalently to work with the log-likelihood function

$$l_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log f(y_i, \theta),$$

and maximizes $l_n(\theta)$ instead.

How do we find MLE's?

- Graphing (simple, intuitive);

- Calculus (exponential families);
- Numerical techniques (e.g. Newton Raphson algorithm, EM algorithm).

Example 5.4.2 (GDP example continued.). *1. Assume the normal model, what is the MLE?*

2. Assume the log-normal model, what is the MLE?

Example 5.4.3. *Challenger disaster.*

On January 28, 1986, the space shuttle, Challenger, had a catastrophic failure due to burn-through of an O-ring seal at a joint in one of the solid-fuel rocket boosters. This was the 25th shuttle flight. Of the previous 24 shuttle flights, 7 has incidents of damage to joints, 16 had no incidents of damage and 1 was unknown. The data consists of the flight, number of damaged o-rings, and the temperature of that day. One question is whether the temperature is related to the probability of damaged o-rings?

Variable Description

1. Temperature (X)

2. Number of damaged o-rings, (Y).

Assume that $Y \overset{\text{iid}}{\sim} Bin(2, p)$. What is the MLE of p?

Example 5.4.4. *In the traffic lights example, what is the MLE of the rate parameter?*