Mathematical statistics

March 4^{th}, 2018

Lecture 9: Introduction to parameter estimation

Where are we?

Week 1	Probability reviews
Week 2	Chapter 6: Statistics and Sampling Distributions
Week 4	**Chapter 7: Point Estimation**
Week 7	Chapter 8: Confidence Intervals
Week 10	Chapter 9: Test of Hypothesis
Week 14	Regression

The Central Limit Theorem

Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then, in the limit when $n \to \infty$, the standardized version of $\bar{X}$ have the standard normal distribution

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z\right) = \mathbb{P}[Z \leq z] = \Phi(z)$$

Rule of Thumb:

If $n > 30$, the Central Limit Theorem can be used for computation.

Example

Problem

When a batch of a certain chemical product is prepared, the amount of a particular impurity in the batch is a random variable with mean value 4.0 g and standard deviation 1.5 g.

If 50 batches are independently prepared, what is the (approximate) probability that the sample average amount of impurity X is between 3.5 and 3.8 g?

Hint:

- First, compute $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$
- Note that
$$\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$
is (approximately) standard normal.

Example

Problem

The tip percentage at a restaurant has a mean value of 18% and a standard deviation of 6%.

What is the approximate probability that the sample mean tip percentage for a random sample of 40 bills is between 16% and 19%?

Chapter 7: Overview

7.1 Point estimate

- unbiased estimator
- mean squared error

7.2 Methods of point estimation

- method of moments
- method of maximum likelihood.

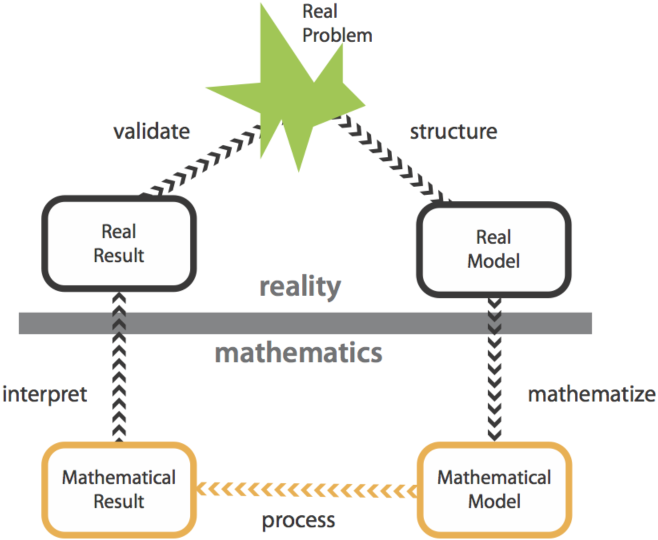
7.3 Sufficient statistic

7.4 Information and Efficiency

- Large sample properties of the maximum likelihood estimator

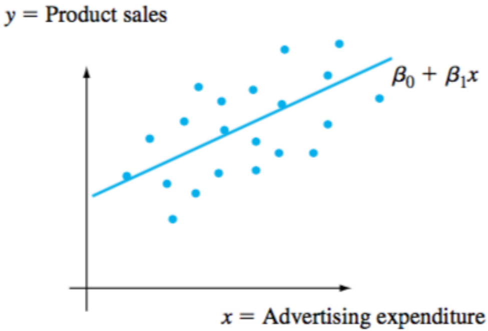
- Bootstrap

Mathematical modelling



Parameter estimation

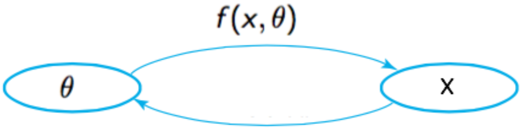
- In a mathematical model, parameters are used to define a whole family of functions that relate the inputs and the outputs
- Example:
$$y = ax + b$$
represents a family of linear functions parameterized by (a, b)
- Parameter estimation: from collected data, determine the values of the parameter



Mathematical model:

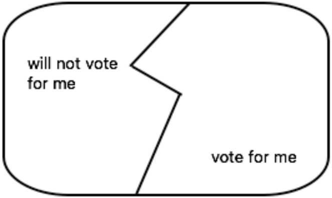
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- Given a random sample $X_1, \ldots, X_n$ from a distribution with pmf/pdf $f(x, \theta)$ parameterized by a parameter θ
- Goal: Estimate θ



Example 1

- Setting: I'm running for president of the US
- I want to estimate how many people support me



- Denote
  - A: the total number of people who will vote for me
  - B: the total number of people who will not

$$p = \frac{A}{A+B}$$

is an unknown quantity that I'm interested in

Step 1: Random sample

- Choose one random person.
- Record the response by a random variable X
  - Yes $\rightarrow X = 1$
  - No $\rightarrow X = 0$
- The pmf of X is as follows

x	0	1
p(x)	1-p	p

- Repeat 2000 times $\rightarrow$ a sample $X_1, X_2, \ldots, X_{2000}$
- Obtained data: $x_1 = 1, x_2 = 0, \ldots, x_{2000} = 1$
- Summary statistics: $n_{yes} = 1200$, $n_{no} = 800$
- Question: What is a good estimate of p?

Step 2: Analysis

- A good estimate of p is

$$\hat{p} = \frac{n_{yes}}{n} = \frac{1200}{2000} = 0.6$$

Step 2: Analysis

- A good estimate of p is

$$\hat{p} = \frac{n_{yes}}{n} = \frac{1200}{2000} = 0.6$$

- A more proper way to write $\hat{p}$

$$\hat{p} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \bar{X}$$

- The strong law of large number

$$\hat{p} = \bar{X} \approx E[X]$$

and

$$E[X] = p.1 + (1 - p).0 = p$$

Step 2: Analysis

Central Limit Theorem: $(n > 40)$

$$P\left[-1.96 \leq \frac{\hat{p} - E[X]}{\sigma_X/\sqrt{n}} \leq 1.96\right] = 95\%$$

or

$$P\left[p - 1.96\frac{1}{\sqrt{n}}\sqrt{p(1-p)} \leq \hat{p} \leq p - 1.96\frac{1}{\sqrt{n}}\sqrt{p(1-p)}\right] = 95\%$$

Step 2: Analysis

- Simplified expression:

$$P\left[\hat{p} - 1.96\frac{\hat{p}(1-\hat{p})}{\sqrt{n}} \leq p \leq \hat{p} + 1.96\frac{\hat{p}(1-\hat{p})}{\sqrt{n}}\right] = 95\%$$

- Plug $\hat{p} = 0.6$ in, we can say

$$0.579 \leq p \leq 0.621$$

with 95% confidence