Mathematical Statistics

MAS 713

Chapter 5

Previous lectures

1. Interval estimation
2. Confidence interval
3. Student's t-distribution
4. Central limit theorem
5. Prediction intervals

Any questions?

This lecture

1. 5.1 Point estimators
   - 5.1.1 Introduction
   - 5.1.2 Estimation and sampling distribution
   - 5.1.3 Point estimation
   - 5.1.4 Properties of estimators

2. 5.2 Sufficient Statistics
   - 5.2.1 Introduction
   - 5.2.2 Factorization Theorem

Additional reading : Chapter 6 in the textbook

Introduction

The purpose of most statistical inference procedures is to generalise from information contained in an observed random sample about the population from which the samples were obtained

This can be divided into two major areas :

- **estimation**, including point estimation and interval estimation
- **tests of hypotheses**

In this chapter we will present point estimation.

Statistical Inference : Introduction

We introduce the main topic for this course : **statistical inference**

Recall the general problem that is addressed :

- statistical methods are used to draw conclusions and make decisions about a population of interest
- however, for some reasons, we have no access to the whole population and we must do with observations on a subset of the population only. That subset is called the sample
- if the sample is effectively representative of the population, what we observe on the sample can be generalised to the population as a whole, at least to some extent ...
- ... taking chance factors properly into account

$\rightsquigarrow$ from what we have learned about descriptive statistics, probabilities and random variables in the previous chapters, we should now be able to set all that to music

Statistical Inference : Introduction

Populations are often described by the distribution of their values

$\rightsquigarrow$ for instance, it is quite common practice to refer to a 'normal population', when the variable of interest is thought to be normally distributed

In statistical inference, we focus on drawing conclusions about one parameter describing the population

Statistical Inference : Introduction

Often, the parameters we are mainly interested in are

- the **mean** μ of the population
- the **variance** σ^2 (or standard deviation σ) of the population
- the **proportion** π of individuals in the population that belong to a class of interest
- the **difference in means of two sub-populations,** $\mu_1 - \mu_2$
- the **difference in two sub-populations proportions,** $\pi_1 - \pi_2$

Statistical Inference : Introduction

Obviously, those parameters are unknown (otherwise, no need to make inferences about them) $\rightsquigarrow$ the first part of the process is thus to estimate the unknown parameters

Random sampling

The importance of **random sampling** has been emphasized in Chapter 1 :

- to assure that a sample is representative of the population from which it is obtained, and
- to provide a framework for the application of probability theory to problems of sampling

As we said, the assumption of random sampling is very important : if the sample is not random and is based on judgment or flawed in some other way, then statistical methods will not work properly and will lead to incorrect decisions

Definition:

The set of observations $X_1, X_2, \ldots, X_n$ constitutes a random sample if

1. the X_i's are independent random variables, and
2. every X_i has the same probability distribution

Statistic, estimator and sampling distribution

Any numerical measure calculated from the sample is called a **statistic**

Denote the unknown parameter of interest θ (so this can be μ, σ^2, or any other parameter of interest to us)

The only information we have to estimate that parameter θ is the information contained in the sample

An **estimator** of θ is thus a statistic, i.e. a function of the sample

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

Statistic, estimator and sampling distribution

Note that an **estimator is a random variable**, as it is a function of random variables $\rightsquigarrow$ it must have a **probability distribution**

That probability distribution is called a **sampling distribution**, and it generally depends on the population distribution and the sample size

After the sample has been selected, $\hat{\Theta}$ takes on a particular value $\hat{\theta} = h(x_1, x_2, \ldots, x_n)$, called the **estimate** of θ

Note: as usual,
the lower case distinguishes the realization of a random sample from the upper case, which represents the random variables before they are observed

Estimation : some remarks

Remark : the hat notation conventionally distinguishes the sample-based quantities (estimator $\hat{\Theta}$ or estimate $\hat{\theta}$) from the population parameter (θ). Besides, as usual, capital letters denote the random variables, like $\hat{\Theta}$, whereas lower-case letters are for particular numerical values, like $\hat{\theta}$

$\rightsquigarrow$ two notable exceptions are :

- the sample mean, usually denoted $\bar{X}$; its observed value, calculated once we have observed a sample $x_1, x_2, \ldots, x_n$, is denoted $\bar{x}$
- the sample standard deviation (variance), usually denoted S (S^2); its observed value, calculated once we have observed a sample $x_1, x_2, \ldots, x_n$, is denoted s (s^2)

sampling distribution

Definition

The probability distribution of a statistic is called *sampling distribution*.

An example : estimating μ in a normal population

Suppose that the random variable X of interest is normally distributed, with unknown mean μ and *known* standard deviation σ

From a random sample, a natural estimator for μ is the sample mean

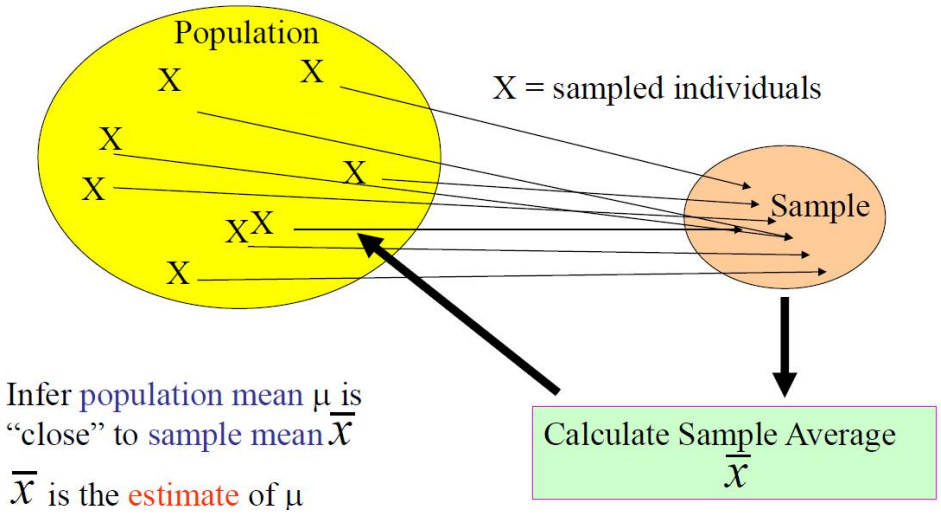
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

whose observed value is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{(= estimate)}$$

As the sample is random, it should be representative of the whole population, and the population mean μ should be "close" to the observed sample mean $\bar{x}$

An example : estimating μ in a normal population



An example : estimating μ in a normal population

What does that mean, μ should be "close" to $\bar{x}$?

$\rightsquigarrow$ the sampling distribution will answer this question

An example : estimating μ in a normal population

By the previous **assumption**, each X_i in the sample follows the $\mathcal{N}(\mu, \sigma)$ distribution, and they are independent (random sample)

Then, because linear combinations of **independent** normal r.v. remain normally distributed, we conclude that

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

has a normal distribution

$$\bar{X} \sim \mathcal{N}\left(\mu_{\bar{X}}, \sigma^2_{\bar{X}}\right)$$

An example : estimating μ in a normal population

with mean

$$\mu_{\bar{X}} = \mathbb{E}\left(\bar{X}\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \overset{\text{i.d.}}{=} \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

and variance

$$\sigma^2_{\bar{X}} = \mathbb{V}\text{ar}\left(\bar{X}\right) = \mathbb{V}\text{ar}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{\text{i.}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\text{ar}(X_i) \overset{\text{i.d.}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

An example : estimating μ in a normal population

$\rightsquigarrow$ the sampling distribution of $\bar{X}$, as an estimator of μ, is

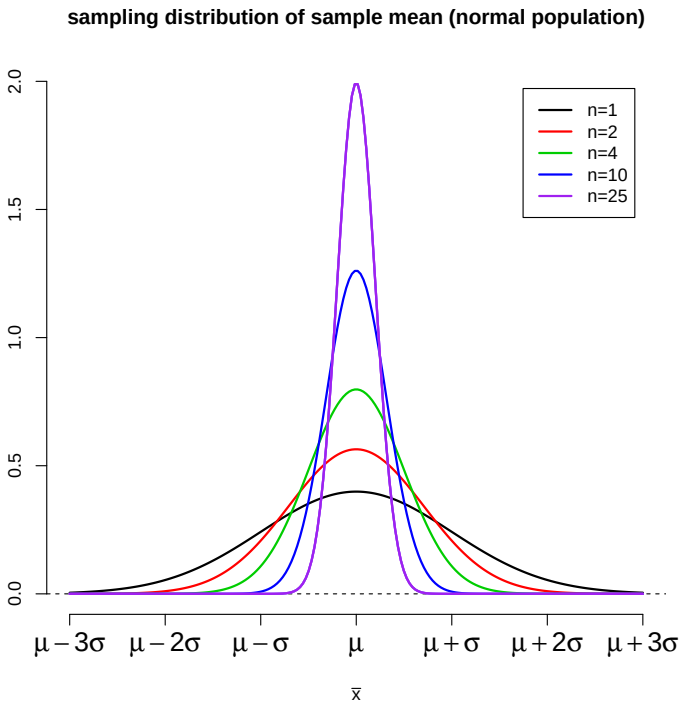
$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

if the population is normal

$\bar{X}$ is a normal random variable, centered about the population mean μ, but with spread becoming more and more reduced as the sample size increases

$\rightsquigarrow$ **the larger the sample, the more accurate the estimation is**

Note : see that $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ do not depend on the normality assumption



Point estimation

Remind that we wish to estimate an unknown parameter θ of a population, for instance the population mean μ, from a random sample of size n, say $X_1, X_2, \ldots, X_n$

To do so, we select an estimator, which must be a statistic, say $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$

$\rightsquigarrow$ **an estimator is a random variable**, which has its mean, its variance and its probability distribution, known as the sampling distribution

Point estimation

Summary: to estimate the population mean μ, we suggested to use the sample mean

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

We derived that

$$\mu_{\bar{X}} = \mu \qquad \text{and } \sigma^2_{\bar{X}} = \frac{\sigma^2}{n},$$

where σ^2 is the population variance

Point estimation

Besides, if $X_i \sim \mathcal{N}(\mu, \sigma)$ for all i, we showed

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Properties of estimators

Properties of estimators

The choice of the sample mean to estimate the population mean seems quite natural. However, there are many other estimators that can be used to calculate an estimate.
Our random sample is: $X_1, X_2, \ldots, X_n$
Why not :

- $\hat{\Theta}_0 = \frac{1}{n}\sum_{i=1}^{n} X_i$;
- $\hat{\Theta}_1 = X_1$, the first observed value;
- $\hat{\Theta}_2 = (X_1 + X_n)/2$;
- $\hat{\Theta}_3 = (aX_1 + bX_n)/(a+b)$, for two constants a, b $(a + b \neq 0)$
- Many other choices

How do we choose which estimator to use?

Properties of estimators

$\rightsquigarrow$ criteria for selecting the 'best' estimator are needed

What do we expect from an estimator for θ ?

$\rightsquigarrow$ certainly that it should give estimates reasonably close to θ, the parameter it is supposed to estimate

However, this 'closeness' is not easy to comprehend : first, θ is unknown, and second, the estimator is a random variable

$\rightsquigarrow$ we have to properly define what are the desirable properties of an estimator

Unbiasedness

Properties of estimators : unbiasedness

A first desirable property that a good estimator should possess is that it is **unbiased**

Definition

An estimator $\hat{\Theta}$ for θ is said to be unbiased **if and only if** the mean of its sampling distribution is equal θ, whatever the value of θ, i.e.

$$\mathbb{E}(\hat{\Theta}) = \theta$$

$\rightsquigarrow$ an estimator is unbiased if "on the average" its values will equal the parameter it is supposed to estimate

Properties of estimators : unbiasedness

If an estimator is not unbiased, then the difference

$$\mathbb{E}(\hat{\Theta}) - \theta$$

is called the **bias** of the estimator $\rightsquigarrow$ systematic error

For instance, we showed that $\mathbb{E}(\bar{X}) = \mu_{\bar{X}} = \mu$

$\rightsquigarrow$ the sample mean $\bar{X}$ is an unbiased estimator for μ

Properties of estimators : unbiasedness

The property of unbiasedness is one of the most desirable properties of an estimator, although it is sometimes out weighted by other factors

One shortcoming is that it will generally not provide a unique estimator for a given problem of estimation

Properties of estimators : unbiasedness

For instance, for the above defined estimators for μ,

Estimator 1:

$$\hat{\Theta}_1 = X_1$$

$$\mathbb{E}(\hat{\Theta}_1) = \mathbb{E}(X_1) = \mu$$

Properties of estimators : unbiasedness

Estimator 2:

$$\hat{\Theta}_2 = (X_1 + X_n)/2$$

$$\mathbb{E}(\hat{\Theta}_2) = \mathbb{E}\left(\frac{X_1 + X_n}{2}\right) = \frac{1}{2}(\mathbb{E}(X_1) + \mathbb{E}(X_n)) = \frac{1}{2}(\mu + \mu) = \mu$$

Properties of estimators : unbiasedness

Estimator 3:

$$\hat{\Theta}_3 = (aX_1 + bX_n)/(a+b)$$

$$\mathbb{E}(\hat{\Theta}_3) = \mathbb{E}\left(\frac{aX_1 + bX_n}{a+b}\right) = \frac{1}{a+b}(a\mathbb{E}(X_1) + b\mathbb{E}(X_n)) = \mu$$

Properties of estimators : unbiasedness

Conclusion:

$\rightsquigarrow$ $\hat{\Theta}_1$, $\hat{\Theta}_2$ and $\hat{\Theta}_3$ are also unbiased estimators for μ

$\rightsquigarrow$ we need a further criterion for deciding which of several **unbiased** estimators is best for estimating a given parameter

Efficiency

Properties of estimators : efficiency

That further criterion becomes evident when we compare the variances of $\bar{X}$ and $\hat{\Theta}_1$

We have shown that $\mathbb{V}\text{ar}(\bar{X}) = \frac{\sigma^2}{n}$, while we have directly $\mathbb{V}\text{ar}(\hat{\Theta}_1) = \mathbb{V}\text{ar}(X_1) = \sigma^2$

$\rightsquigarrow$ the variance of $\bar{X}$ is n times smaller than the variance of $\hat{\Theta}_1$!

$\rightsquigarrow$ it is far more likely that $\bar{X}$ will be closer to its mean, μ, than $\hat{\Theta}_1$ is to μ

Properties of estimators : efficiency

Fact

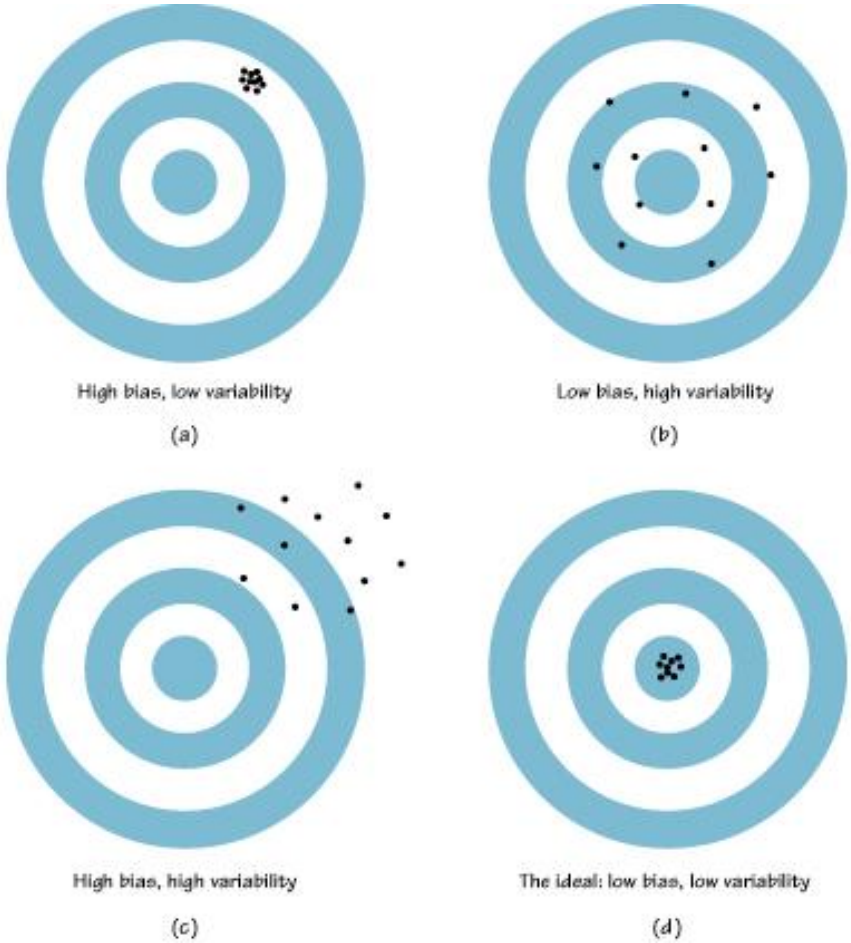
Estimators with smaller variances are more likely to produce estimates close to the true value θ

$\rightsquigarrow$ a logical principle of estimation is to choose the unbiased estimator that has minimum variance

Such an estimator is said to be **efficient** among the unbiased estimators

Properties of estimators

A useful analogy is to think at each value taken by an estimator as a shot at a target, the target being the population parameter of interest



Consistency

Properties of estimators : consistency

Put to the limit, the 'minimum variance' argument becomes :

we desire that the probability that the estimator lies 'close' to θ increases to 1 as the sample size increases

$\rightsquigarrow$ we desire an estimator $\hat{\Theta} \equiv \hat{\Theta}^{(n)}$ that is more and more precise as the number of observations increases :
as we increase n, it becomes more and more likely that the estimator will take a value very close to θ

Such estimators are called **consistent**

Definition of consistent estimator

A (sequence of) estimator $\hat{\Theta}^{(n)}$ is called consistent for the parameter θ if

$$\lim_{n\to\infty} \mathbb{P}\big(|\hat{\Theta}^{(n)} - \theta| \geq \varepsilon\big) = 0 \qquad \forall \varepsilon > 0$$

- Pointwise convergence $\hat{\Theta}^{(n)} \to \theta$ as $n \to \infty$ implies consistency

Theorem: Sufficient condition for being consistent

Let $\hat{\Theta}^{(n)}$ be a (sequence of) estimator. If :

- The variance goes to zero, meaning that: $\lim_{n\to\infty} \mathbb{V}\text{ar}(\hat{\Theta}^{(n)}) = 0$
- The bias goes to zero, meaning that: $\lim_{n\to\infty} \mathbb{E}(\hat{\Theta}^{(n)}) - \theta = 0.$

Then, $\hat{\Theta}^{(n)}$ is a consistent estimator.

Note: Conversely, under some integrability condition, we see that the above conditions are also necessary:

Theorem: Necessary condition for being consistent

Let $\hat{\Theta}^{(n)}$ be a (sequence of) estimator with $\sup_n \mathbb{E}(|\hat{\Theta}^{(n)}|^{2+\delta}) < \infty$ for some $\delta > 0$.

If $\hat{\Theta}^{(n)}$ is consistent, then

- The variance goes to zero, meaning that: $\lim_{n\to\infty} \mathbb{V}\text{ar}(\hat{\Theta}^{(n)}) = 0$
- The bias goes to zero, meaning that: $\lim_{n\to\infty} \mathbb{E}(\hat{\Theta}^{(n)}) - \theta = 0.$

Properties of estimators : consistency

Consequence of Sufficiency Theorem

An easy way to check that an **unbiased** estimator is consistent is to show that its variance decreases to 0 as n increases to ∞

For instance, $\mathbb{V}\text{ar}(\bar{X}) = \frac{\sigma^2}{n} \to 0$ as $n \to \infty \rightsquigarrow$ **$\bar{X}$ is consistent for μ**

On the other hand, it can be verified that

$$\mathbb{V}\text{ar}(\hat{\Theta}_1) = \sigma^2 \nrightarrow 0,$$

$$\mathbb{V}\text{ar}(\hat{\Theta}_2) = \frac{\sigma^2}{2} \nrightarrow 0,$$

$$\mathbb{V}\text{ar}(\hat{\Theta}_3) = \frac{a^2 + b^2}{(a+b)^2} \nrightarrow 0$$

$\rightsquigarrow$ under some intgrability coniditon, e.g., $\mathbb{E}(X_1^{2+\delta}) < \infty$ for some $\delta > 0$

$\rightsquigarrow$ none of them are consistent,

Sample mean

We have thus seen that the sample mean $\bar{X}$ is unbiased and consistent as an estimator of the population mean μ

Besides, it can be shown that in most practical situations where we estimate the population mean μ, the variance of no other estimator is less than that of the sample mean

Sample mean

Fact

In most practical situations, the sample mean is a very good estimator for the population mean μ

Note : there exist several other criteria for assessing the goodness of point estimation methods, but we shall not discuss them in this course

Here, we will always use the sample mean $\bar{X}$ when we will have to estimate the population mean μ

Standard error of a point estimate

Although we estimate the population parameter θ with an estimator that we know to have certain desirable properties (unbiasedness, consistency), the chances are slim, virtually nonexistent, that the estimate will actually equal θ

$\rightsquigarrow$ an estimate remains an approximation of the true value!

$\rightsquigarrow$ it is unappealing to report your estimate only, as there is nothing inherent in $\hat{\theta}$ that provides any information about how close it is to θ

Hence, it is usually desirable to give some idea of the precision of the estimation $\rightsquigarrow$ the measure of precision usually employed is the standard error of the estimator that has been used

Standard error of a point estimate

Definition

The **standard error** of an estimator $\hat{\Theta}$ is its standard deviation $\sigma_{\hat{\Theta}}$

Note : If the standard error involves some unknown parameters that can be estimated, substitution of those values into $\sigma_{\hat{\Theta}}$ produces an estimated standard error, denoted $\hat{\sigma}_{\hat{\Theta}}$

Standard error of the sample mean

Suppose again that we estimate the mean μ of a population with the sample mean $\bar{X}$ calculated from a random sample of size n

We know that $\mathbb{E}(\bar{X}) = \mu$ and $\mathbb{V}\text{ar}(\bar{X}) = \frac{\sigma^2}{n}$, so the standard error of $\bar{X}$ as an estimator of μ is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}},$$

Standard error of the sample mean

However, the standard deviation σ is usually unknown.

$\rightsquigarrow$ we have a natural estimate of the population standard deviation given by the observed sample standard deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$\rightsquigarrow$ estimate the standard error $\sigma_{\bar{X}}$ with $\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}}$

Note : we will study the sample standard deviation S as an estimator of σ in details later

Concluding remarks

The sample mean is a good estimator of the population mean.
In more complicated models (which is usually the case), we need a more methodical way of estimating parameters.

Example

One observation, X, is taken from a $N(0, \sigma)$ population.
Is $\widehat{\sigma}^2 = X^2$ an unbiased estimator of σ^2?

To check unbiasness we calculate:

$$\begin{aligned}\mathbb{E}\left[\widehat{\sigma}^2\right] &= \mathbb{E}\left[X^2\right] \\ &= \mathbb{V}\text{ar}[X] + \mathbb{E}[X]^2 \\ &= \sigma^2 + \mu^2 = \sigma^2\end{aligned}$$

We conclude that $\widehat{\sigma}^2 = X^2$ is an unbiased estimator of σ^2.

Example

One observation, X, is taken from a $N(0, \sigma)$ population.
Is $\widehat{\sigma}^2 = X^2$ an unbiased estimator of σ^2?

To check unbiasness we calculate:

$$\begin{aligned}\mathbb{E}\left[\widehat{\sigma}^2\right] &= \mathbb{E}\left[X^2\right] \\ &= \mathbb{V}\text{ar}[X] + \mathbb{E}[X]^2 \\ &= \sigma^2 + \mu^2 = \sigma^2\end{aligned}$$

We conclude that $\widehat{\sigma}^2 = X^2$ is an unbiased estimator of σ^2.

Example

Let $X_1, \ldots, X_n$ be i.i.d. samples with PDF

$$p(x|\theta) = \frac{1}{\theta}, \ 0 < x < \theta, \ \theta > 0$$

1. Estimator 1: $\widehat{\theta}_1 = 2\bar{X}$
2. Estimator 2: $\widehat{\theta}_2 = X_{(n)}$.

   Hint: the pdf of $\widehat{\theta}_2$ is $p(x) = nx^{n-1}/\theta^n$.

1. Are these estimators unbiased?
2. What is their variance?
3. Are these estimators consistent?

Estimator 1: $\widehat{\theta}_1 = 2\bar{X}$

Is this estimator unbiased?

$$\begin{aligned}
\mathbb{E}\left[\widehat{\theta}_1\right] &= \mathbb{E}\left[2\bar{X}\right] \\
&= \mathbb{E}\left[2\frac{1}{n}\sum_{j=1}^{n} X_j\right] \\
&= 2\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left[X_j\right] \\
&= 2\frac{1}{n}\sum_{j=1}^{n} \int_0^{\theta} x\,\tfrac{1}{\theta}\,dx \\
&= 2\frac{1}{n}\sum_{j=1}^{n} \frac{\theta}{2} \\
&= \theta
\end{aligned}$$

The estimator is unbiased.

Estimator 1: $\widehat{\theta}_1 = 2\bar{X}$

What is the variance?

$$\mathbb{V}\text{ar}(\widehat{\theta}_1) = \mathbb{E}\left[\widehat{\theta}_1^2\right] - \mathbb{E}\left[\widehat{\theta}_1\right]^2$$

$$\mathbb{E}\left[\widehat{\theta}_1^2\right] = \mathbb{E}\left[\left(2\bar{X}\right)^2\right] = 4\mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)^2\right]$$

$$= 4\frac{1}{n^2}\left(\sum_{j=1}^{n}\mathbb{E}\left[X_j^2\right] + 2\sum_{i=1}^{n}\sum_{k=1}^{i-1}\mathbb{E}\left[X_i X_k\right]\right)$$

$$= 4\frac{1}{n^2}\left(\sum_{j=1}^{n}\int_0^{\theta} x^2 \tfrac{1}{\theta}\, dx + 2\sum_{i=1}^{n}\sum_{k=1}^{i-1}\mathbb{E}[X_i]\mathbb{E}[X_k]\right)$$

$$= 4\frac{1}{n^2}\left(\tfrac{n\theta^2}{3} + 2\tfrac{n(n-1)}{2}\tfrac{\theta}{2}\tfrac{\theta}{2}\right)$$

$$= \theta^2 + \tfrac{\theta^2}{3n}$$

$$\Longrightarrow \mathbb{V}\text{ar}(\widehat{\theta}_1) = \mathbb{E}\left[\widehat{\theta}_1^2\right] - \mathbb{E}\left[\widehat{\theta}_1\right]^2 = \theta^2 + \frac{\theta^2}{3n} - \theta^2 = \frac{\theta^2}{3n}.$$

Estimator 1: $\widehat{\theta}_1 = 2\bar{X}$

Is it consistent?

$$\lim_{n\to\infty} \mathbb{V}\text{ar}(\widehat{\theta}_1) = \lim_{n\to\infty} \frac{\theta^2}{3n} = 0$$

This, and as it is **unbiased**, we get that it is a consistent estimator

Estimator 1: $\widehat{\theta}_1 = 2\bar{X}$

Is it consistent?

$$\lim_{n\to\infty} \mathbb{V}\text{ar}(\widehat{\theta}_1) = \lim_{n\to\infty} \frac{\theta^2}{3n} = 0$$

This, and as it is **unbiased**, we get that it is a consistent estimator

Estimator 2: $\widehat{\theta}_2 = X_{(n)}$

Is this estimator unbiased?

$$\begin{aligned}
\mathbb{E}\left[\widehat{\theta}_2\right] &= \mathbb{E}\left[X_{(n)}\right] \\
&= \int_0^\theta x n x^{n-1}/\theta^n \, dx \\
&= \frac{n}{\theta^n} \int_0^\theta x^n \, dx \\
&= \frac{n}{\theta^n} \frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\theta
\end{aligned}$$

The estimator is biased.

Estimator 2: $\widehat{\theta}_2 = X_{(n)}$

What is the variance?

$$\begin{aligned}
\mathbb{E}\left[\widehat{\theta}_2^2\right] &= \mathbb{E}\left[X_{(n)}^2\right] \\
&= \int_0^\theta x^2 n x^{n-1}/\theta^n \, dx \\
&= \frac{n}{\theta^n} \int_0^\theta x^{n+1} \, dx \\
&= \frac{n}{\theta^n} \frac{\theta^{n+2}}{n+2} = \frac{n}{n+2}\theta^2
\end{aligned}$$

$$\implies \mathbb{V}\mathrm{ar}(\widehat{\theta}_2) = \mathbb{E}\left[\widehat{\theta}_2^2\right] - \mathbb{E}\left[\widehat{\theta}_2\right]^2 = \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2 = \frac{n\theta^2}{(n+2)(n+1)^2}.$$

Estimator 2: $\widehat{\theta}_2 = X_{(n)}$

Is it consistent?

$$\lim_{n\to\infty} \mathbb{V}\text{ar}(\widehat{\theta}_2) = \lim_{n\to\infty} \frac{n\theta^2}{(n+2)(n+1)^2} = 0$$

$$\lim_{n\to\infty} \mathbb{E}(\widehat{\theta}_2) - \theta = \lim_{n\to\infty} \theta\left(\tfrac{n}{n+1} - 1\right) = 0$$

This is a consistent estimator

Estimator 2: $\widehat{\theta}_2 = X_{(n)}$

Is it consistent?

$$\lim_{n\to\infty} \mathbb{V}\text{ar}(\widehat{\theta}_2) = \lim_{n\to\infty} \frac{n\theta^2}{(n+2)(n+1)^2} = 0$$

$$\lim_{n\to\infty} \mathbb{E}(\widehat{\theta}_2) - \theta = \lim_{n\to\infty} \theta\left(\tfrac{n}{n+1} - 1\right) = 0$$

This is a consistent estimator

Comparison Estimator 1 with Estimator 2:

Estimator 1: $\widehat{\theta}_1 = 2\bar{X}$

- $\mathbb{E}\left[\widehat{\theta}_1\right] = \theta$ and $\mathbb{V}\text{ar}(\widehat{\theta}_1) = \frac{\theta^2}{3n}$.

Estimator 2: $\widehat{\theta}_2 = X_{(n)}$

- $\mathbb{E}\left[\widehat{\theta}_2\right] = \frac{n}{n+1}\theta$ and $\mathbb{V}\text{ar}(\widehat{\theta}_2) = \frac{n\theta^2}{(n+2)(n+1)^2}$.

- If n is large, the bias is not large because $n/(n+1)$ is close to one. But if n is small, the bias is quite large.
- On the other hand, $\mathbb{V}\text{ar}(\widehat{\theta}_2) < \mathbb{V}\text{ar}(\widehat{\theta}_1)$ for all θ.

Conclusion: So, if n is large, $\widehat{\theta}_2$ is probably preferable to $\widehat{\theta}_1$.

Mean Squared Error

Properties of estimators : Mean Squared Error

The **Mean Squared Error** of an estimator $\widehat{\theta}$ is defined as follows

Definition

The **Mean Squared Error** of an estimator $\widehat{\theta}$ is given by

$$\text{MSE}\left(\widehat{\theta}\right) = \mathbb{E}\left[\left(\widehat{\theta} - \theta\right)^2\right]$$

Properties of estimators : Mean Squared Error

The **Mean Squared Error** of an estimator $\widehat{\theta}$ is given by

$$\begin{aligned}
\text{MSE}\left(\widehat{\theta}\right) &= \mathbb{E}\left[\left(\widehat{\theta}-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right]\right)^2\right] + \left(\theta-\mathbb{E}\left[\widehat{\theta}\right]\right)^2 \\
&= \mathbb{V}\text{ar}\left(\widehat{\theta}\right) + \left(\theta-\mathbb{E}\left[\widehat{\theta}\right]\right)^2 \\
&= \text{Variance} + (\text{bias})^2
\end{aligned}$$

The bias variance trade-off is one of the most fundamental results in estimation theory.
This trade-off appears in any estimation problem (point estimation, regression, classification).

Example

Example

Let $X_1, \ldots, X_n$ be i.i.d random variables with PDF

$$p(x|\sigma) = \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right).$$

For the estimator

$$\widehat{\sigma} = \frac{\sum_{i=1}^{n} |X_i|}{n}$$

1. Is the estimator unbiased?
2. What is its MSE?
3. Is the estimator consistent?

Recall : the Gamma function is given by

$$\Gamma(y) = \int_0^{+\infty} x^{y-1} e^{-x}\, dx, \qquad \text{for } y > 0$$

It can be shown that $\Gamma(y) = (y-1) \times \Gamma(y-1)$, so that, if y is a positive integer n,

$$\Gamma(n) = (n-1)!$$

Is the estimator unbiased?

$$\mathbb{E}\left[\widehat{\sigma}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{n}|X_i|}{n}\right] = \frac{\sum_{i=1}^{n}\mathbb{E}\left[|X_i|\right]}{n}$$

$$\begin{aligned}
\mathbb{E}\left[|X_i|\right] &= \int_{-\infty}^{\infty} |x| \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right) dx \\
&= \sigma \int_{0}^{\infty} \frac{x}{\sigma} \exp\left(-\frac{x}{\sigma}\right) d\frac{x}{\sigma} \\
&= \sigma \int_{0}^{\infty} y \exp\left(-y\right) dy \\
&= \sigma\Gamma\left(2\right) = \sigma.
\end{aligned}$$

The estimator is unbiased

What is its MSE?

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\widehat{\sigma}\right] &= \mathbb{V}\mathrm{ar}\left[\frac{\sum_{i=1}^{n}|X_i|}{n}\right] \\
&= \frac{\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}\left[|X_i|\right]}{n^2} \\
&= \frac{\mathbb{V}\mathrm{ar}\left[|X_i|\right]}{n} \\
&= \frac{\mathbb{E}\left[|X_i|^2\right] - \mathbb{E}\left[|X_i|\right]^2}{n} \\
&= \frac{\mathbb{E}\left[|X_i|^2\right] - \sigma^2}{n}
\end{aligned}$$

- Now

$$\begin{aligned}
\mathbb{E}\left[|X_i|^2\right] &= \int_{-\infty}^{\infty} |x|^2 \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right) dx \\
&= \sigma^2 \int_0^{\infty} \frac{x^2}{\sigma^2} \exp\left(-\frac{x}{\sigma}\right) d\frac{x}{\sigma} \\
&= \sigma^2 \int_0^{\infty} y^2 \exp\left(-y\right) dy \\
&= \sigma\Gamma\left(3\right) = 2\sigma^2.
\end{aligned}$$

- Therefore, we can conclude that

$$\begin{aligned}
\mathbb{V}\text{ar}\left[\widehat{\sigma}\right] &= \mathbb{V}\text{ar}\left[\frac{\sum_{i=1}^{n}|X_i|}{n}\right] \\
&= \frac{\sum_{i=1}^{n} Var\left[|X_i|\right]}{n^2} \\
&= \frac{\mathbb{V}\text{ar}\left[|X_i|\right]}{n} \\
&= \frac{\mathbb{E}\left[|X_i|^2\right] - E\left[|X_i|\right]^2}{n} \\
&= \frac{\mathbb{E}\left[|X_i|^2\right] - \sigma^2}{n} \\
&= \frac{2\sigma^2 - \sigma^2}{n} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

The MSE is given by

$$\begin{aligned} MSE_{\widehat{\sigma}} &= \mathbb{V}\text{ar}\left[\widehat{\sigma}\right] + \left(E\left[\widehat{\sigma}\right] - \sigma\right)^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Is the estimator consistent?

$$\lim_{n\to\infty} \mathbb{V}\text{ar}\left[\widehat{\sigma}\right] = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

As it is also **unbiased**, the estimator is consistent

Is the estimator consistent?

$$\lim_{n\to\infty} \mathbb{V}\text{ar}\left[\widehat{\sigma}\right] = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

As it is also **unbiased**, the estimator is consistent

Sufficient Statistics

Sufficient Statistics

Definition

A statistic $T(\mathbf{X})$ is a *sufficient statistic* for θ if the conditional distribution of the sample $\mathbf{X} = (X_1, \ldots, X_n)$ given the value of $T(\mathbf{X})$ does not depend on θ.

Remark (for later:)

A sufficient statistic captures all the information about θ to compute the **likelihood function**. We will define the likelihood function later.

Sufficient Statistics

Sufficient statistic helps us compress the information, a.k.a “data reduction".
The intuition behind the sufficient statistic concept is that it contains all the information necessary for estimating θ. Therefore if one is interested in estimating θ, it is perfectly fine to “get rid" of the original data while keeping only the value of the sufficient statistic.

It is difficult to use the definition to check if a statistic is sufficient or to find a sufficient statistic.
Luckily, there is a theorem that makes it easy to find sufficient statistics.

Sufficient Statistics

Theorem

If $p(\mathbf{x}|\theta)$ is the joint PDF or PMF of $\mathbf{X}$ and $q(T(\mathbf{X})|\theta)$ is the joint PDF or PMF of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for θ if, for every $\mathbf{x}$ in the samples space, the ratio $\frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{X})|\theta)}$ is a constant as a function of θ.

- Useful to check if statistic is sufficient.

Example

Let $X_1, \ldots, X_n$ be i.i.d. from $\mathcal{N}(\mu, \sigma)$, where σ is known.
Is the sample mean

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

a sufficient statistic for μ?

Sufficient Statistics

Our statistics is

$$T(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The joint PDF of the sample X is

$$\begin{aligned}
p(\mathbf{X}|\mu) &= \prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}\exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-n/2}\exp\left(-\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-n/2}\exp\left(-\sum_{i=1}^{n}\frac{(X_i-\bar{X}+\bar{X}-\mu)^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-n/2}\exp\left(-\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2+n(\bar{X}-\mu)^2}{2\sigma^2}\right)
\end{aligned}$$

Sufficient Statistics

Therefore,

$$\frac{p(\mathbf{X}|\mu)}{q(T(\mathbf{X})|\mu)} = \frac{(2\pi\sigma^2)^{-n/2}\exp\left(-\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2+n(\bar{X}-\mu)^2}{2\sigma^2}\right)}{(2\pi\sigma^2/n)^{-1/2}\exp\left(-\frac{n(\bar{X}-\mu)^2}{2\sigma^2}\right)}$$

$$= n^{-1/2}(2\pi\sigma^2/n)^{-(n-1)/2}\exp\left(-\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{2\sigma^2}\right)$$

Since $\frac{p(\mathbf{X}|\mu)}{q(T(\mathbf{X})|\mu)}$ does not depend on μ, therefore, the sample mean is sufficient statistic for μ.

Sufficient Statistics

Procedure presented in the last Theorem may be difficult to implement, since we need to:

- first choose the statistic $T(\mathbf{X})$, and then
- test if it is indeed sufficient.

This requires a good deal of experience and intuition, and a tedious analysis.
Luckily, there is a way to find sufficient statistic via a simple inspection of the PDF or PMF.

Factorization Theorem of Fisher-Neyman

Factorization Theorem

Let $p(\mathbf{x}|\theta)$ denote the joint PDF or PMF of $\mathbf{X}$. A statistic $T(\mathbf{X})$ is a sufficient statistic for θ **if and only if** there exist nonnegative functions $t \mapsto g(t|\theta)$ and $\mathbf{x} \mapsto h(\mathbf{x})$ such that, for all sample points x and all parameters points θ,

$$p(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$$

Factorization Theorem

To use the Factorization Theorem we proceed as follows:

Application procedure:

- Factor the joint PDF of the sample into two parts, with one part not depending on θ.
- The part that does not depend on θ constitutes the $h(x)$ function.
- The other part, the one that depends on θ, usually depends on the sample $\mathbf{X}$ only through some function $T(\mathbf{X})$

  $\Longrightarrow$ this function is a sufficient statistic for θ.

Example revisited

- For the last example, we can write the joint PDF as

$$p(\mathbf{X}|\mu) = (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{2\sigma^2}\right) \times \exp\left(-\frac{n(\bar{X} - \mu)^2}{2\sigma^2}\right).$$

- We define

$$h(x) = (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{2\sigma^2}\right),$$

which does not depend on the unknown parameter μ.

- Moreover, we define the function $t \mapsto g(t|\mu)$ by

$$g(t|\mu) = \exp\left(-\frac{n(t - \mu)^2}{2\sigma^2}\right),$$

- Note that

$$p(\mathbf{x}|\mu) = h(\mathbf{x})g(\bar{x}|\mu) \quad \Longrightarrow \quad \text{choose } T(\mathbf{X}) := \bar{X}$$

Sufficient statistic: example

Example

Let X be one observation from $N(0, \sigma)$ population.
Is $|X|$ a sufficient statistic for σ?

$$\begin{aligned} p(x|\sigma) &= (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{x^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{|x|^2}{2\sigma^2}\right) \\ &= g(|x| \ |\sigma) \times 1 \end{aligned}$$

where

$$g(T(x)\,|\sigma) = g(|x| \ |\sigma) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{|x|^2}{2\sigma^2}\right)$$

$$h(x) = 1$$

Sufficient statistic: example

Example

Let $X_1, \ldots, X_n$ be be a random sample from a population with a PDF

$$p(x|\theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \ \theta > 0.$$

Is $\sum X_i$ a sufficient statistic for θ?

$$\begin{aligned} p(x_1, \ldots, x_n|\theta) &= \prod_{i=1}^{n} \theta x_i^{\theta-1} \\ &= \theta^n \left(\prod_{i=1}^{n} x_i \right)^{\theta-1} \end{aligned} \tag{1}$$

- We see that $\sum_{i=1}^{n} X_i$ is not sufficient statistics for θ
- But $\prod_{i=1}^{n} X_i$ is.

Objectives

Now you should be able to :

- Explain important properties of point estimators, including bias, variance, efficiency and consistency □
- Know how to compute and explain the precision with which a parameter is estimated □
- Understand what sufficient statistics is and how to check if a statistic is sufficient. □

Put yourself to the test !

$\rightsquigarrow$ Q6.2 p.300, Q6.5 p.300, Q6.6 p.300, Q6.20 p.302,