

IMPERIAL COLLEGE LONDON

MATHEMATICS: YEAR 2

# Statistics and Sampling distributions

Xin Wang

September 29, 2020

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Random samples</b>	<b>3</b>
<b>3</b>	<b>Distribution of sample mean</b>	<b>4</b>
<b>4</b>	<b>Central Limit Theorem</b>	<b>6</b>
<b>5</b>	<b>Statistical analysis</b>	<b>6</b>
5.1	Point estimation . . . . .	7
5.2	Properties of estimators . . . . .	8
5.2.1	Estimator bias . . . . .	8
5.2.2	Minimum variance estimators . . . . .	9
5.2.3	Mean square error . . . . .	9
<b>6</b>	<b>Standard errors</b>	<b>9</b>
<b>7</b>	<b>Point estimation methods</b>	<b>10</b>
7.1	Method of moments . . . . .	10
7.2	Maximum likelihood (ML) . . . . .	11

# 1 Introduction

This section considers samples drawn from a larger population, and using that sample to make statements about unknown parameters e.g. the mean of the population.

There are certain levels of uncertainties when drawing a sample from a population that can be expressed as a random variable for example:

- Before the data is obtained, there is uncertainty about what values will be observed.
- There is uncertainty about the value that denotes a method of summary about the population i.e. the mean  $\bar{x}$ .

The presence of these uncertainties can be seen with the following example:

**Example 1:** 5 samples are taken. Each sample has 10 independent observations  $n = 10$  are drawn from a population with a standard normal distribution i.e. a population with a  $N(0, 1)$ . The sample mean and sample standard deviation of **each sample of each 10 observations** are computed.

Sample	$\bar{x}$	$s$
1	0.0757	0.8914
2	0.5549	1.4599
3	-0.066	0.7606
4	0.7433	0.6039
5	-0.5988	1.0473

Note that all these samples each have different values since the observations vary from sample to sample, so does any value computed from a sample. Capturing this uncertainty is a key component of **statistical analysis**.

**Statistic:** Any quantity calculated from sample data.

Prior to obtaining the data, there is a level of uncertainty as to which value of the statistic will occur. This uncertainty is modelled as a random variable thus **a statistic is a random variable**.

Since a statistic is a random variable, it has a probability distribution called the **sampling distribution**. This distribution reflects how the statistic value varies across the possible samples that could possibly be selected.

The **expected value** and **theoretical variance** of the sampling distribution will be important values.

## 2 Random samples

The sampling distribution of a statistic depends on several factors:

- Parent population size.
- Sample size.
- Sampling mechanism used.

**Random sample** (of size  $n$  from a distribution with probability mass/-density function  $f_X(x)$ ): A set of independently and identically distributed random variables  $X_1, X_2, \dots, X_n$  each with mass/density function  $f_X(x)$ .

**Example 1:** Consider the random variable  $X$ , with PMF:

$x$	1	2	3
$f_X(x)$	0.2	0.3	0.5

Two numbers  $X_1$  and  $X_2$  are drawn independently according to  $f_X(x)$ . Find the mean defined as  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ . The small range of  $X$  makes it practical to consider every pair  $(X_1, X_2)$  and the corresponding mean.

Recall:

- **Independent:** Can compute  $f_{X_1, X_2}(x_1, x_2) = P(X_1 = x_1 \cap X_2 = x_2)$ .

Process:

1. Calculate the expected value of  $\bar{X}$ :

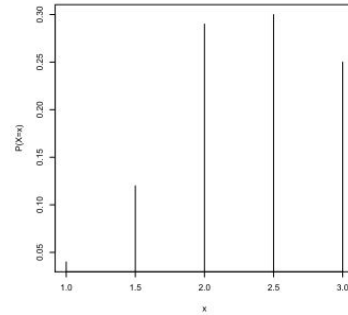
$$\bar{X} = \mu = 1(0.2) + 2(0.3) + 3(0.5) = 2.3$$

2. Calculate the variance:

$$\text{Variance} = \frac{\sigma^2}{n} = \frac{\sum x^2 p - \mu^2}{2} = 0.305$$

3. Find the sampling distribution.

$x_1$	$x_2$	$p(x_1, x_2)$	$\bar{x}$
1	1	0.04	1.0
2	1	0.06	1.5
3	1	0.10	2.0
1	2	0.06	1.5
2	2	0.09	2.0
3	2	0.15	2.5
1	3	0.10	2.0
2	3	0.15	2.5
3	3	0.25	3.0



(a) Each pair  $(X_1, X_2)$  and the corresponding mean.

(b) Sampling distribution

### 3 Distribution of sample mean

Many important statistical procedures are concerned with making statements about the **population mean** using the properties of the distribution of the **sample mean**.

If  $X_1, X_2, \dots, X_n$  is a random sample from a population with mean  $\mu$ , variance  $\sigma^2$  and sample mean  $\bar{X}$ :

- Expected value  $E[\bar{X}]$ :

$$\begin{aligned}
 E[\bar{X}] &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = E\left(\frac{1}{n} \sum_{i=1}^n E(X_i)\right) \\
 &= \frac{n\mu}{n} = \mu
 \end{aligned}$$

- Variance  $\text{Var}(\bar{X})$ :

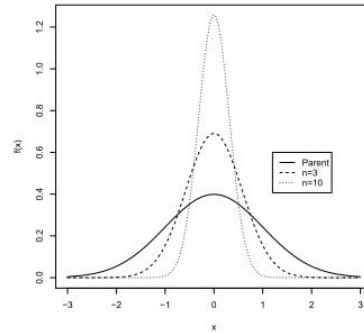
$$\begin{aligned}
 \text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\
 &= \frac{1}{n^2} \left(\sum_{i=1}^n \text{Var}(X_i)\right) \\
 &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}
 \end{aligned}$$

- Standard deviation:

$$\frac{\sigma}{\sqrt{n}}$$

Commonly known as the **standard error** of the sample mean. This quantity decreases as the sample size increases.

A common relationship between parent normal distribution and the corresponding sampling distributions for  $\bar{X}$



is that as the sample size increases, the sampling distribution becomes increasingly concentrated about its mean. When the parent population is normal, such that  $X_1, X_2, \dots, X_n$  are independent and share the identical base distribution  $N(\mu, \sigma^2)$ :

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

**Example 1:** The total daily output of a titanium oxide facility is well modelled by a normal distribution, with **mean** ( $\mu$ ) 500 tons, and **standard deviation** ( $\sigma$ ) 50 tons.

On 25 randomly selected days per year the output is precisely measured and the sample mean of the 25 measurements is recorded as an annual *QC* measure.

What is the probability that this sample mean is between 490 and 510?

1. Denote the daily output as  $X$ :

$$X \sim N(500, 50^2) \text{ where sample size } n = 25$$

2. Find  $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$  and the standardised version:

$$\bar{X} \sim N\left(500, \frac{50^2}{25}\right) \text{ and } Z = \frac{\bar{X} - 500}{10} \sim N(0, 1)$$

3. Compute the probability:

$$P(490 \leq X \leq 510) = P(-1 \leq Z \leq 1) = \phi(1) - \phi(-1) = 0.682$$

## 4 Central Limit Theorem

**Statistical inference** is usually based on the sampling distribution of a statistic, but there are few choices of parent distribution and statistic for which the sampling distribution can be computed exactly. Most will require approximations.

Previously, the CLT can be used to approximate the sum of independent random variables from any distribution. Using the properties of the normal distribution, the CLT can be extended to the sample mean:

Given a random sample of size  $n$  from a distribution with mean  $\mu$  and variance  $\sigma^2$ , the distribution of  $\bar{X}$  is **approximately**  $N\left(\mu, \frac{\sigma^2}{n}\right)$  when  $n$  is large enough.

The quality of the approximation improves with increasing  $n$ .

## 5 Statistical analysis

Probability theory provides the tools and models to reason about random experiments. When observing a random variable  $x$ , the variable is treated as only one possible value of  $X$ . The objective of statistical inference is to learn the aspects of the probability distribution which produced  $x$ , to allow inferences to be drawn about the nature of the process under study.

The inferences drawn can be about various aspects. In this context, the inference drawn concerns with making generalisations from the specific (the sample) to the general (the population). The notation used to represent the probability model for  $x$ :

$$f(x|\theta)$$

where:

- $\theta$  - Parameters of the model i.e.  $n$  and  $p$  for a binomial random variable and  $\lambda$  for an exponential.

The purpose of this notation is to make explicit that **the value of  $\theta$  needs to be known in order to evaluate the probability of a particular observation**. In all the previous examples, the parameters and distribution were known and could easily evaluate probabilities such as  $P(X > 3)$ .

Statistical problems often involve the reverse situation: **realisations** i.e. sample observations for some unknown probability model are obtained. The goal is to estimate the values of the model parameters. The questions involved will consider situations involving a known distribution with unknown parameters.

**Statistic:** A function that refers to a random sample.

**Estimator:** A statistic used to guess the value of some unknown parameter.

Observing the sample  $x_1, x_2, \dots, x_n$ , two point of interest:

- **Point estimation:** Estimating the value of unknown parameters.
- **Interval estimation:** Estimating an interval that contains the unknown parameters with high probability.

## 5.1 Point estimation

Suppose it is of interest the true average lifetime of a specific type of spark plug i.e. population mean value  $\mu$ . Obtain a random sample  $x_1, x_2, \dots, x_9$  of  $n = 9$  spark plugs to make an observation. The sample mean  $\bar{x}$  can be used to make statements about  $\mu$ . Other statistics can also be used to make related statements e.g. the sample variance  $s^2$  provides information about the population variance  $\sigma^2$ .

Sample realisations and statistics are usually denoted with Roman symbols, like  $\bar{x}$  and  $s$ , whereas unknown population parameters are denoted with Greek symbols, like  $\mu$  and  $\theta$ .

The objective of point estimation is to use a sample of data to provide a guess of an unknown population parameter.

**Point estimate** (of an unknown population parameter  $\theta$ ): A single number which represents the “best guess” of  $\theta$  that arises from selecting a suitable statistic and computing its value from a sample.

The selected statistic is called a **point estimator** for  $\theta$  denoted by  $\hat{\theta}$ .

**Example 1:** A chemical plant produces a compound that contains a particular component B. A random sample of size  $n = 11$  batches of this compound was obtained to measure the percentages of component B in each. The results are as follows:

5.3, 4.9, 6.2, 5.7, 4.8, 5.4, 6.1, 6.3, 5.6, 5.5, 7.2

Assuming that the normal distribution with mean  $\mu$  provides a plausible model for the percentage of B and that this distribution is symmetric,  $\mu$  can be estimated several ways:



- **Sample mean:**

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{11} x_i = 5.72$$

- **Sample median:**

$$\tilde{x} = \frac{(x_5 + x_6)}{2} = 5.6$$

- **Midpoint between the extremes:**

$$\frac{x_1 + x_{11}}{2} = 6$$

The quality cannot be directly assessed since the value of the parameter is unknown, thus estimations have to be selected among estimators using other criteria.

## 5.2 Properties of estimators

Recall that statistics are functions of random samples and all have a sampling distribution. The sampling distribution provides a description of the random behaviour of the statistic thus can be used to determine choice of estimator used.

Two criteria that sometimes lead to accurate estimators are:

- Unbiasedness
- Minimum variance

### 5.2.1 Estimator bias

Given a point estimator  $\hat{\theta}$ :

**Bias** (of  $\hat{\theta}$ ): The difference between expected value **of estimator** and a given value  $\theta$ :

$$\text{Bias of } \theta = E(\hat{\theta}) - \theta$$

A point estimator  $\hat{\theta}$  is an **unbiased estimator** of  $\theta$  if  $E(\hat{\theta}) = \theta$ , for all  $\theta$ . The bias for an unbiased estimator is 0.

The definition of an unbiased estimator states that the sampling distribution of an unbiased estimator has a expected value equal to the population value  $\theta$ . Thus, if the experiment is repeated an infinite number of times, on average an unbiased estimator would yield the correct answer.

If the parent population is known, the bias of an estimator  $\hat{\theta}$  can be found even if  $\theta$  is unknown.

**Example 1:** Components of a certain type are either operable or defective. It is of interest in the random variable  $X$  that counts the number of defective components in a batch of 50 components.

This is a binomial experiment. A batch is randomly selected and the components tested.  $X = 3$  defective is found. Use different ways techniques to estimate the unknown probability that a component is defective denoted  $p$ .

Estimate 1: The basic probability:

$$\begin{aligned}\hat{p} &= \frac{X}{n} \\ &= \frac{3}{50} \\ &= 0.06\end{aligned}$$

Estimate 2: Finding the bias and determining a better estimate:

1. Find the expected value of the estimator:

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{E(x)}{n} = \frac{np}{n} = p$$

2. The estimator is an **unbiased estimator** of  $p$ .

### 5.2.2 Minimum variance estimators

When selecting among unbiased estimators of  $\theta$ , the estimator with **minimum variance** is chosen since the minimum variance estimator has the most concentrated distribution.

There is a special term for this: **Minimum variance unbiased estimator** (MVUE) of  $\theta$ .

### 5.2.3 Mean square error

While the MVUE is desirable, it is often possible to obtain an estimator with small bias, but appreciably smaller variance.

For this reason, choose the estimator which minimises the **mean squared error** (MSE) defined:

$$MSE_{\theta}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + \left(\text{Bias}_{\theta}(\hat{\theta})\right)^2$$

## 6 Standard errors

The standard error of the sample mean i.e. the standard deviation of the sampling distribution of the mean can be extended to any estimator.

**Standard error** of estimator  $\theta$  -  $\sigma_{\hat{\theta}}$ : The standard deviation of estimator:

$$\text{Standard error} = \sqrt{\text{Var}(\hat{\theta})}$$

Sometimes, the standard error will depend on unknown parameters that must be estimated. In this case, the **estimated standard error** of the estimator  $\hat{\sigma}_{\hat{\theta}}$  is used.

**Example 1:** A chemical plant produces a compound that contains a particular component B. A random sample of size  $n = 11$  batches of this compound was obtained to measure the percentages of component B in each. The results are as follows:

$$5.3, 4.9, 6.2, 5.7, 4.8, 5.4, 6.1, 6.3, 5.6, 5.5, 7.2$$

Assuming that the normal distribution with unknown mean  $\mu$  and unknown variance  $\sigma^2$ :

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

find the estimated standard error.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 0.476$$

Estimated error is:

$$s_{\bar{X}} = 0.2$$

## 7 Point estimation methods

The previous section listed the desirable characteristics of estimators without specifying how to construct estimators. There are a number of methods available for automatically constructing estimators:

- Method of moments
- Least squares
- Maximum likelihood (ML)

In general, ML estimation is preferred in statistics because of desirable theoretical properties, although in realistic contexts, it sometimes require significant computation.

### 7.1 Method of moments

The method of moments attempts to equate sample characteristics with corresponding population theoretical values. Its estimators are the solutions to such equations that satisfies the ideal characteristics.

Consider a random sample  $X_1, X_2, \dots, X_n$  from PDF or PMF  $f(x)$ . Recall that the  $k^{th}$  moment of  $X$  is defined  $m_k = E(X^k)$ . The **sample moments** - sample equivalent of theoretical moments are defined:

$$\frac{1}{n} \sum_{i=1}^n X_i^k \text{ for } k = 1, 2, \dots$$

The process gets complicated when there are more parameters that need to be estimated.

Some common theoretical moments are:

- First theoretical moment  $E(X) = \mu$  and corresponding sample moment is  $\frac{1}{n} \sum_{i=1}^n X_i$
- Second theoretical moment  $E[X^2]$  and corresponding sample moment is  $\frac{1}{n} \sum_{i=1}^n X_i^2$

**Example 1:**

## 7.2 Maximum likelihood (ML)

Maximum-likelihood estimation is usually the recommended approach, since the estimators have desirable properties. This approach is built around the joint probability or density of the collection of random variables in a sample.

Given same  $X_1, X_2, \dots, X_n$  with joint density or PMF:

$$L(\theta) = f(X_1 = x_1, \dots, X_n = x_n | \theta)$$

where  $\theta$  refers to unknown parameters.

**Likelihood function:** The joint probability is regarded as a function of  $\theta$ . The value of  $\theta$  that maximises the likelihood function is known as the **maximum-likelihood estimators** (MLE) of  $\theta$ . The estimates evaluated for observed values yield **maximum-likelihood estimates**.

The maximum-likelihood estimators give the values of  $\theta$  that agree mostly closely with the observed data. Due to the property of independence, the joint density/mass function of a random sample decomposes into a product of the marginal density/mass functions.

General ML process:

1. Write and simplify the likelihood.
2. Take logs.
3. Differentiate and equate the resulting equations with 0 and solve.
4. Examine the second derivatives.