Lecture 8: Statistical Inference

Yunwen Lei

School of Computer Science, University of Birmingham

1 Descriptive Statistic

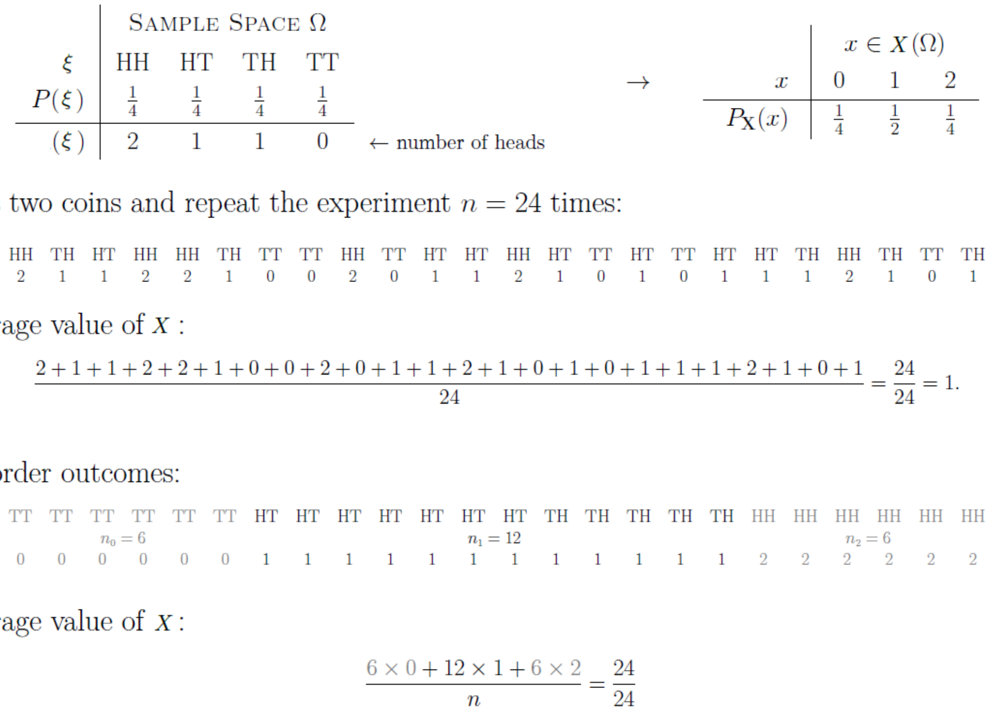
Descriptive Statistics refers to a summary that quantitatively describes features from a collection of information, which does not assume that the data come from a larger population – We simply describe our data in hand, the data we currently have.

- Central tendency: **expectation** (mean), median, mode, etc. For example, the median is the value separating the higher half from the lower half of a data sample.
- Dispersion: the range and quartiles of the dataset. For example, a quartile is a statistical term that describes a division of observations into four defined intervals based on the values of the data and how they compare to the entire set of observations.
- Spread: **variance** and standard deviation. If a random variable has a small variance, then it concentrates sharply around its mean.
- Shape of the distribution: skewness and kurtosi

1.1 Expectation

We consider an example of sample mean to motivate the definition of expectation.

Example 1 (Sample Mean: Toss Two Coins Many Times)**.** We consider the experiment of tossing two coins and define X as the number of heads in these two tosses.



	Sample Space Ω			
ξ	HH	HT	TH	TT
$P(\xi)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
(ξ)	2	1	1	0

← number of heads

→

	$x \in X(\Omega)$		
x	0	1	2
$P_X(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Toss two coins and repeat the experiment $n = 24$ times:

HH	TH	HT	HH	HH	TH	TT	TT	HH	TT	HT	HT	HH	HT	TT	HT	TT	HT	HT	TH	HH	TH	TT	TH
2	1	1	2	2	1	0	0	2	0	1	1	2	1	0	1	0	1	1	1	2	1	0	1

Average value of X :

$$\frac{2+1+1+2+2+1+0+0+2+0+1+1+2+1+0+1+0+1+1+1+2+1+0+1}{24} = \frac{24}{24} = 1.$$

Re-order outcomes:

TT	TT	TT	TT	TT	TT	HT	HT	HT	HT	HT	HT	HT	TH	TH	TH	TH	TH	HH	HH	HH	HH	HH	HH
		$n_0=6$									$n_1=12$									$n_2=6$			
0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2

Average value of X :

$$\frac{6\times 0 + 12\times 1 + 6\times 2}{n} = \frac{24}{24}$$

Suppose we repeat this experiment $n = 24$ times and are interested in the sample mean.

- We can first take the average of X for these 24 experiments. This gives the first way to compute the average value of X.
- We can counter the number n_0 of experiments with 0 heads, n_1 of experiments with 1 head and n_2 of experiments with 2 heads. Then the sample average is

$$\frac{1}{n}\Big(n_0 \times 0 + n_1 \times 1 + n_2 \times 2\Big) = \frac{6 \times 0 + 12 \times 1 + 6 \times 2}{24} = 1.$$

In the above example, n_0/n is the empirical frequency of observing 0 heads. If $n \to \infty$, this empirical frequency approaches to the probability $P_X(0)$. This motivates the introduction of expectation.

TT	TT	TT	TT	TT	TT	HT	HT	HT	HT	HT	HT	HT	TH	TH	TH	TH	TH	HH	HH	HH	HH	HH	HH
		$n_0 = 6$									$n_1 = 12$									$n_2 = 6$			
0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2

Average value of $\mathbf{X}$:

$$\begin{aligned}\frac{n_0 \times 0 + n_1 \times 1 + n_2 \times 2 + n_3 \times 3}{n} &= \underset{\approx P_{\mathbf{X}}(0)}{\underset{\uparrow}{\frac{n_0}{n}}} \times 0 + \underset{\approx P_{\mathbf{X}}(1)}{\underset{\uparrow}{\frac{n_1}{n}}} \times 1 + \underset{\approx P_{\mathbf{X}}(2)}{\underset{\uparrow}{\frac{n_2}{n}}} \times 2 \\ &\approx P_{\mathbf{X}}(0) \times 0 + P_{\mathbf{X}}(1) \times 1 + P_{\mathbf{X}}(2) \times 2 \\ &= \sum_{x \in \mathbf{X}(\Omega)} x \cdot P_{\mathbf{X}}(x)\end{aligned}$$

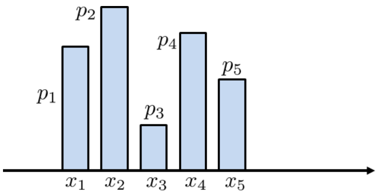
The expectation is defined by replacing the empirical frequency with the probability in the computation of sample average.

Definition 1 (Expectation)**.** The Expectation of a random variable X is

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x P_X(x). \tag{1.1}$$

Intuition: it gives expected value before experiment

$$\mathbb{E}[X] = \underbrace{\sum_{x \in X(\Omega)}}_{\text{sum over all states}} \underbrace{x}_{\text{a state } X \text{ takes}} \underbrace{P_X(x)}_{\text{the percentage}}$$



$\mathbb{E}[X] = p_1 x_1 + \ldots + p_5 x_5$

For two coins the expected value is $0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$.

Example 2 (3 coin tosses)**.** Let X be the number of heads from 3 coin tosses:

	Sample Space Ω							
ω	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
$P(\omega)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
$\mathbf{X}(\omega)$	3	2	2	1	2	1	1	0

→

	$x \in \mathbf{X}(\Omega)$			
x	0	1	2	3
$P_{\mathbf{X}}(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

We can compute the expectation as follows

$$\mathbb{E}[\text{number heads}] = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{12}{8}.$$

Let X_i be the number of heads for the i-th coin toss. Then $\mathbb{E}[X_i] = \frac{1}{2}(0+1)$. Note in this example we have

$$\mathbb{E}[X] = \mathbb{E}[X_1 + X_2 + X_3] = \frac{3}{2} = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \mathbb{E}[X_3].$$

The above equation is not a coincidence. It holds for any random variables. This is called the linearity of expectation. This linearity shows that we can exchange the expectation with the summation and multiplication by a constant, which is very useful for us to compute the expectation of a complicated random variable.

Theorem 1 (Linearity of Expectation)**.** *Let $X_1, X_2, \ldots, X_k$ be random variables. Let $a_1, \ldots, a_k$ be constants. Then*

$$\mathbb{E}[\sum_{i=1}^{k} a_i X_i] = \sum_{i=1}^{k} a_i \mathbb{E}[X_i].$$

Remark 1 (Sample Mean and Expectation)**.** Sample mean and expectation are two important concepts.

- Suppose we repeatedly do the same random experiment. This leads to n random variables $X_i, i \in [n]$ with the same PMF

  $$\mathbb{P}[X_i = a] = \mathbb{P}[X = a], \quad \forall a \in X(\Omega), i \in [n].$$

- We can define the **Sample Mean** $\bar{X}_n$ as

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

  Note that sample mean depends on the implementation of experiments. We now show the difference between sample mean and expectation.

Expectation $\mathbb{E}[X]$

- A statistical property of a random variable
- A deterministic number independent of implementation of the experiment.
- Often unknown, or is the center question of estimation. It is a quantity about the population.

Sample Mean $\bar{X}_n$

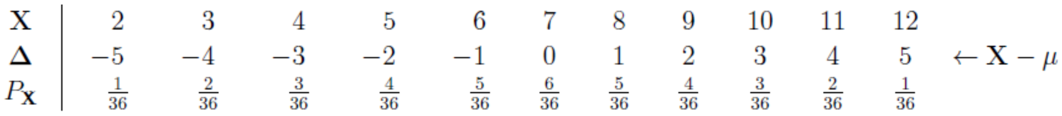
- A numerical value. Calculated from data
  - suppose we toss coins 4 times and get “H”, “T”, “T”, “H”
  - The sample mean for the number of heads is $(1 + 0 + 0 + 1)/4 = 1/2$
- Itself is a random variable. Note it is an average of X_i, each of which is a random variable.
- It has uncertainty (i.e., it takes different values for different repetitions of experiments). The uncertainty reduces as more samples are used. The reason is that the randomization is likely to offset each other with more experiments.
- We can use sample mean to estimate the expectation: the expectation cannot be computed from the data, while the sample mean can.

1.2 Variance

The expectation gives the value we expect for a random variable before implementing an experiment. Now we introduce another key concept called variance. Let us consider the experiment of tossing 2 dice and X = sum of 2 dice

$$\mathbb{E}[X] = \frac{1}{36} \cdot 2 + \frac{2}{36} \cdot 3 + \frac{3}{36} \cdot 4 + \frac{4}{36} \cdot 5 + \cdots + \frac{1}{36} \cdot 12 = 7 \leftarrow \mu$$

Let $\Delta = X - \mu$, which measures the deviation from the mean.



$\mathbf{X}$	2	3	4	5	6	7	8	9	10	11	12	
$\boldsymbol{\Delta}$	-5	-4	-3	-2	-1	0	1	2	3	4	5	$\leftarrow \mathbf{X} - \mu$
$P_\mathbf{X}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	

Sometimes we are interested in the variation of a random variable, i.e., how a random variable would concentrate around its expectation. Whether it concentrates sharply or not? The variance is a concept to describe this behavior.

Definition 2 (Variance and Standard Deviation)**.** Variance, $\mathrm{Var}(X)$, is the expected value of the squared deviations

$$\mathrm{Var}(X) = \mathbb{E}[\Delta^2] = \mathbb{E}\big[(X-\mu)^2\big] = \mathbb{E}\big[(X-\mathbb{E}[X])^2\big].$$

The Standard Deviation, σ, is the square-root of the variance: $\sigma = \sqrt{\mathbb{E}[\Delta^2]}$. If the variance is large, then this means that the deviation Δ can be large with high probability. If the variance is small, then the deviation Δ would be small with high probability. Therefore, the variance measures the uncertainty of a random variable.

We can compute the variance for the random variable X defined above

$$\begin{aligned}\mathrm{Var}(X) = \frac{1}{36}\cdot(-5)^2 + \frac{2}{36}(-4)^2 + \frac{3}{36}(-3)^2 + \frac{4}{36}(-2)^2 + \frac{5}{36}(-1)^2 + \frac{6}{36}0^2 \\ + \frac{1}{36}\cdot(5)^2 + \frac{2}{36}(4)^2 + \frac{3}{36}(3)^2 + \frac{4}{36}(2)^2 + \frac{5}{36}(1)^2 = \frac{35}{6}.\end{aligned}$$

Remark 2 (Variance is a Measure of Risk)**.** We consider two random variables with the same mean but different variance

	Game 1			Game 2	
$\mathbf{X}_1$:	win \$2	probability $= \frac{2}{3}$;	$\mathbf{X}_2$:	win \$102	probability $= \frac{2}{3}$;
	lose \$1	probability $= \frac{1}{3}$.		lose \$201	probability $= \frac{1}{3}$.

Expectation

$$\begin{aligned}\mathbb{E}[X_1] &= \frac{2}{3}\cdot 2 + \frac{1}{3}\cdot(-1) = 1 \\ \mathbb{E}[X_2] &= \frac{2}{3}\cdot 102 + \frac{1}{3}\cdot(-201) = 1\end{aligned}$$

Variance

$$\begin{aligned}\mathrm{Var}(X_1) &= \frac{2}{3}\cdot(2-1)^2 + \frac{1}{3}(-1-1)^2 = 2 \\ \mathrm{Var}(X_2) &= \frac{2}{3}\cdot(102-1)^2 + \frac{1}{3}(-201-1)^2 \approx 20,000\end{aligned}$$

Since variance of X_2 is much larger than the variance of X_1, it has more uncertainty. In different running of the game, you can either win or lose a lot of money. For a small expected profit you might consider Game 1 with risking a small loss not a huge loss.

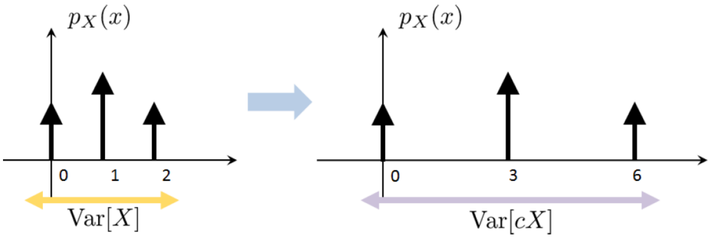
Theorem 2 (Properties of Variance)**.** *Let X be a random variable. Then we have the following property.*

(a) Variance is the expectation of square minus the square of expectation

$$Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2. \tag{1.2}$$

(b) Scale. For any constant c

$$Var(cX) = c^2\, Var(X)$$



(c) Shift. For any constant c

$$\mathit{Var}(X+c) = \mathit{Var}(X)$$

(d) If X and Y are independent then

$$\mathit{Var}(X+Y) = \mathit{Var}(X) + \mathit{Var}(Y).$$

Remark 3. Part (a) shows that the variance of a random variable X is the expectation of X^2 minus the square of the expectation. Part (b) shows that if we multiply a random variable by a constant c, the variance will change by a factor of c^2. Part (c) shows that shifting a random variable does not change the variance. This is reasonable since shifting by a constant changes both the random variable and expectation, and therefore the deviation would be the same. Part (d) shows that the variance is additive for two independent random variables. This identity does not hold if X and Y are dependent.

Example 3 (Expectation and Variance)**.** If $X \sim \text{Bernoulli}(\theta)$, find its expectation and variance.

Answer: For the Bernoulli distribution, we have $P_X(1) = \mathbb{P}(X=1) = \theta$ and $P_X(0) = \mathbb{P}(X=0) = 1-\theta$. Then the expectation is

$$\begin{aligned}\mathbb{E}[X] &= 0 \cdot P_X(0) + 1 \cdot P_X(1) \\ &= 0 \cdot (1-\theta) + 1 \cdot \theta = \theta\end{aligned}$$

and the variance is

$$\begin{aligned}\text{Var}(X) &= \sum_{x \in \{0,1\}} P_X(x)(x - \mathbb{E}[X])^2 \\ &= P_X(0)(0-\theta)^2 + P_X(1)(1-\theta)^2 = (1-\theta)\theta^2 + \theta(1-\theta)^2 = (1-\theta)\theta\end{aligned}$$

1.3 Continuous Random Variables

In the above discussions, we only consider expectation and variance for discrete random variables. We can also define the expectation and variance for continuous random variables. We only need to replace the summation there by integral with the help of PDF.

Definition 3 (Expectation for continuous random variables)**.** For a continuous random variable X with the PDF $f_X(x)$, the expectation is defined as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx. \tag{1.3}$$

Definition 4 (Variance for continuous random variables)**.** For a continuous random variable X with the PDF $f_X(x)$ and the expectation $\mathbb{E}[X] = \mu_X$, the variance is defined as

$$\text{Var}(X) = \mathbb{E}[(\underbrace{X - \mu_X}_{\text{deviation}})^2] = \int_{-\infty}^{\infty} f_X(x)(x-\mu_X)^2 dx. \tag{1.4}$$

Similar to the case with discrete random variables, the expectation for continuous random variable is a quantity we expect before the implementation of the experiment, while the variance measures the concentration behavior of a random variable around its expectation.

Example 4 (Uniform random variable)**.** Let X be a continuous random variable with PDF $f_X(x) = \frac{1}{b-a}$ for $x \in [a,b]$, and 0 otherwise.

Expectation:

$$\begin{aligned}\mathbb{E}[X] &= \int_a^b x f_X(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a}\int_a^b x dx \\ &= \frac{1}{b-a}\frac{b^2-a^2}{2} = \frac{a+b}{2}.\end{aligned}$$

Variance:

$$\mathbb{E}[X^2] = \int_a^b x^2 f_X(x)dx = \frac{1}{b-a}\int_a^b x^2 dx = \frac{a^2+b^2+ab}{3}.$$

Then according to Part (a) of Theorem 2 we know

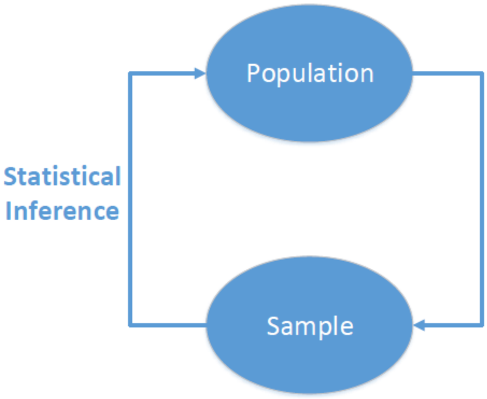
$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{a^2+b^2+ab}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

2 Statistical Inference

In this section, we consider statistical inference. **Statistical inference** refers to the process that collects data to estimate the desired quantity about the **population**.

Population: the universe to which we wish to generalise, e.g., the weights of all adults in the UK. In the population, we have parameters which are the true (fixed) unknown values we wish to estimate (infer).

Sample: the finite study we perform to collect data.



Process of statistical inference.

Note the population can be extremely large or infinite, therefore it is impossible to get tractable inference by going through the whole population. Instead, we often draw finite sample and then try to analyze the sample to get useful inference.

Example 5 (Statistical Inference)**.** Here are some examples on statistical inference.

- **USA president election prediction**: Randomly choose a number of people (called random sampling) and ask them who they plan to vote for. Based on the sampled data we want to make prediction.
- **Wireless communication**: a message is transmitted from a transmitter to a receiver. Because of the noisy channel, the message is corrupted. The receiver needs to infer the original message from the received corrupted version.
- **Disease diagnosis**: Given a set of measures or data from a patient, e.g., some protein concentration, because of the noise, e.g. measurement error, and uncertainty, e.g., our limited knowledge about the molecular biology and the patient, we need to diagnose whether the patient has the disease or not.

2.1 Point Estimation

A fundamental inference method is the point estimation. It tries to estimate the unknown parameter, e.g., mean or variance.

Definition 5 (Point estimation)**.** In statistics, point estimation is a process that samples data to calculate **a single value** as a "best estimate" of an unknown population parameter.

Example 6. We would like to estimate the average height of all 3-year old babies in the UK

- Population: heights of all 3-year old babies in the UK
- Population parameter: expectation or mean
- Sample: we conduct a finite study to measure the heights of 3-year old babies in the UK

To do statistical inference, we need to build sample of population. How to collect data?

Randon Sampling: We choose a random sample of size n with replacement from the population. More specifically,

- We choose n person uniformly and independently from the population and record their heights, denoted as random variables X_i, $i = 1, \dots, n$.
- Sampling **with** replacement: we allow one person to be chosen twice
- Sampling **without** replacement: we do NOT allow one person to be chosen twice

Why use sampling with replacement: Random variables X_i are independent which simplifies the analysis.

When to use sampling with replacement: when the population is large enough that the probability of choosing one person twice is extremely low.

Remark 4 (Point Estimation of Expectation)**.** After collecting data, i.e., $X_1, X_2, \dots, X_n$ using sampling with replacement, we estimate the average height by the average height estimator as

$$\hat{\Theta} = \overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

The above example consists of

- **Random samples**: $X_1, X_2, \dots, X_n$ are independent and identically distributed (i.i.d.) continuous random variables, which have the same distributions:

$$F_{X_1}(x) = F_{X_2}(x) = \cdots = F_{X_n}(x),$$

- **Point estimator**: the average height estimator $\hat{\Theta}$ is a random variable, called a point estimator

Estimate: After performing the above experiment, we will obtain $\hat{\Theta} = \hat{\theta}$, $\hat{\theta}$ is called an estimate of the average height in the population.

Underlying assumption: to estimate an unknown parameter θ we assume that θ is a fixed (non-random) quantity.

Definition 6 (Point Estimator)**.** Suppose we have a random sample X_i with the same distribution $i = 1, \dots, n$. A point estimator $\hat{\Theta}$ for an unknown parameter θ is defined as a function of the random sample:

$$\hat{\Theta} = h(X_1, \dots, X_n).$$

where h can be any function, e.g., mean.

Remark 5 (Estimate versus Estimator)**.** We now explain the difference between estimate and estimator.

- An estimate is a number, e.g., $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$. It is the random realization of a random variable. This value can be computed from the empirical data.
- An estimator is a random variable, e.g., $\hat{\Theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$. It takes a set of random variables as inputs and generates another random variable. It gives a formula to estimate the parameter. We can view estimate as a realization of the estimator.

Example 7 (Point Estimation of Expectation)**.** Let $X_1, \dots, X_n$ be Gaussian i.i.d. random variables with unknown mean θ and known variance σ^2.

There are several different ways to define estimators. Here are two examples:

$$\hat{\Theta}_1(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$\hat{\Theta}_2(X_1, \dots, X_n) = X_1.$$

If $x_1, \dots, x_n$ are realizations of $X_1, \dots, X_n$, then we get the corresponding two estimates

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\theta}_2 = x_1.$$

Below we show how to estimate the variance. For example, how to estimate the variance σ^2 of the heights of all 3-year old toddlers in the UK?

Example 8 (Point Estimation of Variance)**.** By definition:

$$\sigma^2 = \mathbb{E}[(X-\mu)^2],$$

that is, variance itself is the mean of the random variable $Y = (X-\mu)^2$. Thus, we have following estimator for the variance (replace the expectation by empirical average)

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2. \tag{2.1}$$

However, in practice we do not know the true μ. We substitute μ with our estimate or the sample mean $\overline{X}$, which results in the variance estimator:

$$\overline{S}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2 = \frac{1}{n}\sum_{k=1}^{n}\left(X_k^2 + \overline{X}^2 - 2X_k\overline{X}\right) = \frac{1}{n}\left(\sum_{k=1}^{n} X_k^2 - n\overline{X}^2\right) = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \overline{X}^2, \tag{2.2}$$

where we have used the identity $\sum_{k=1}^{n} X_k = n\overline{X}$. That is, we can use the empirical average of X_k square minus the square of empirical average as a variance estimator. This is consistent with Part (a) of Theorem 2: the variance is the expectation of X^2 minus the square of expectation.

2.2 Evaluating Point Estimation

There are various ways to define point estimation. Different ways have different quality. A basic question is how to evaluate different point estimation methods. Since estimator is a random variable, we can compute its expectation and see how the expectation would differ from the parameter θ we are interested in.

Definition 7 (Bias)**.** Let $\hat{\Theta} = h(X_1, \dots, X_n)$ be a point estimator for θ. The bias of point estimator $\hat{\Theta}$ is defined as

$$B(\hat{\Theta}) = \mathbb{E}[\hat{\Theta}] - \theta.$$

An estimator $\hat{\Theta}$ is unbiased iff $B(\hat{\Theta}) = 0$

We can compute the deviation of the estimator from θ, which itself is a random variable. The mean squared error is the expected square of this deviation.

Definition 8 (Mean Squared Error (MSE))**.** The mean squared error (MSE) of a point estimator $\hat{\Theta}$, denoted as $\mathrm{MSE}(\hat{\Theta})$, is defined as (prove the second identity in an exercise)

$$\mathrm{MSE}(\hat{\Theta}) := \mathbb{E}[(\hat{\Theta} - \theta)^2] = \mathrm{Var}(\hat{\Theta}) + B(\hat{\Theta})^2 \tag{2.3}$$

According to Eq. (2.3), the mean square error can be decomposed into two components: the first is the variance of the estimator, while the second is the square of the bias. Roughly speaking, the variance considers its fluctuation while the bias consider the difference between the expectation and the true parameter. We say an estimator $\hat{\Theta}$ is good if the MSE is small.

Below we give an example to compute the MSE for an estimator.

Example 9. Let $X_1, \ldots, X_n$ be i.i.d. Gaussian random variables with an unknown mean μ and variance σ^2. Define the estimator as $\hat{\Theta} = \frac{1}{n}\sum_{i=1}^n X_i$.

Bias. We get

$$\mathbb{E}[\hat{\Theta}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^n \theta = \theta \Longrightarrow B(\hat{\Theta}) = 0.$$

Variance. We get

$$\mathrm{Var}(\hat{\Theta}) = \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^n X_i\Big) = \frac{1}{n^2}\mathrm{Var}\Big(\sum_{i=1}^n X_i\Big) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{1}{n}\mathrm{Var}(X_1) = \frac{\sigma^2}{n}.$$

MSE. According to Eq. (2.3) we can compute MSE as follows

$$\mathrm{MSE} = \mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^n X_i - \mu\Big)^2\Big] = \mathrm{Var}(\hat{\Theta}).$$

Example 10. In Eq. (2.2), we provide an estimator of the variance. Let $X_1, X_2, \ldots, X_n$ be a random sample with mean $\mathbb{E}[X_i] = \mu$, and variance $\mathrm{Var}(X_i) = \sigma^2$. Suppose we use the variance estimator:

$$\overline{S}^2 = \frac{1}{n}\sum_{k=1}^n (X_k - \overline{X})^2 = \frac{1}{n}\left(\sum_{k=1}^n X_k^2 - n\overline{X}^2\right) \tag{2.4}$$

to estimate σ^2. We now show that this variance estimator $\overline{S}$ is a biased estimator of σ^2.

To evaluate the bias, we use the equation: $B(\hat{\Theta}) = \mathbb{E}[\hat{\Theta}] - \theta = \mathbb{E}[\overline{S}^2] - \sigma^2$. Then it follows from the linearity of expectation that

$$\mathbb{E}[\overline{S}^2] = \mathbb{E}\Big[\frac{1}{n}\Big(\sum_{k=1}^n X_k^2 - n\overline{X}^2\Big)\Big] = \frac{1}{n}\Big(\sum_{k=1}^n \mathbb{E}[X_k^2] - n\mathbb{E}[\overline{X}^2]\Big)$$

Since

$$\mathbb{E}[\overline{X}^2] = \mathbb{E}[\overline{X}]^2 + \mathrm{Var}(\overline{X}) = \mu^2 + \frac{\sigma^2}{n}$$

and

$$\mathbb{E}[X_k^2] = \sigma^2 + \mu^2,$$

we know

$$\mathbb{E}[\overline{S}^2] = \frac{1}{n}\left(n(\mu^2 + \sigma^2) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right) = \frac{n-1}{n}\sigma^2.$$

Then the bias is not zero:

$$B(\overline{S}^2) = \mathbb{E}[\overline{S}^2] - \sigma^2 = -\frac{\sigma^2}{n} \neq 0$$

From the above calculation, we know that

$$\frac{n}{n-1}\overline{S}^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k - \overline{X})^2$$

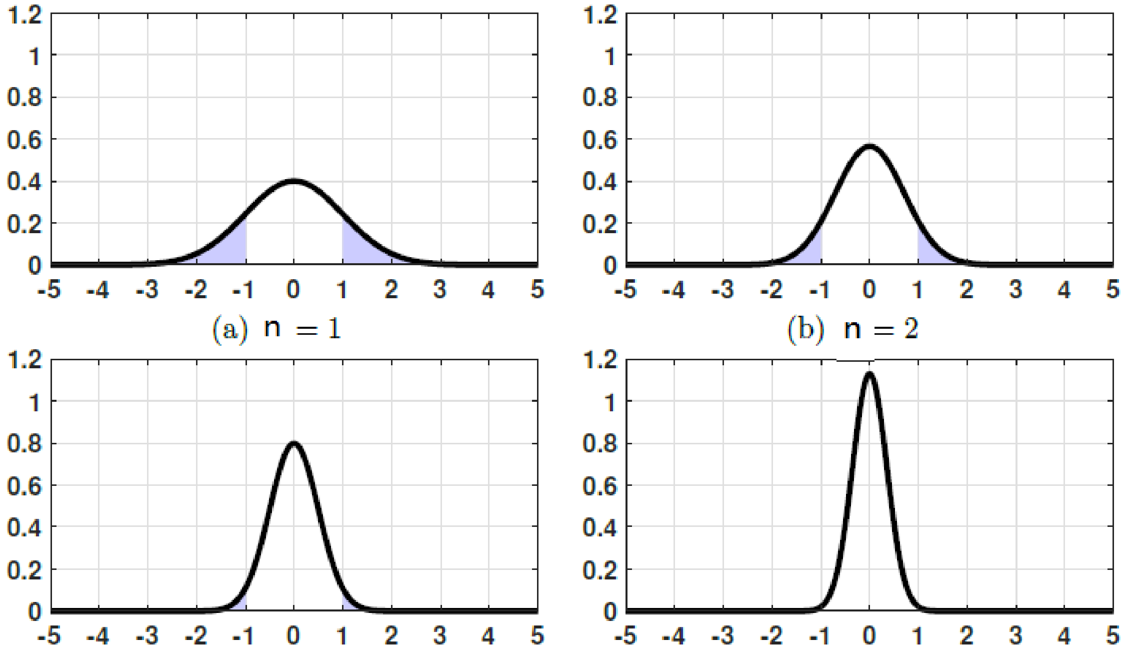
is an unbiased estimator of the variance!

Another measurement of an estimator is the consistency. We wish as the sample size increases, we get more and more precise estimators. The following definition means that the probability of the deviation greater than ϵ goes to 0 as the sample size goes to infinity. Here we use $\hat{\Theta}_n$ to emphasize its dependency on the sample size.

Definition 9 (Consistency)**.** Let $\hat{\Theta}_i$, $i = 1, \ldots, n, \ldots$ be a sequence of point estimators of θ. We say that $\hat{\Theta}_n$ is a consistent estimator of θ if

$$\lim_{n \to \infty} \mathbb{P}(|\hat{\Theta}_n - \theta| \geq \epsilon) = 0, \quad \text{for all } \epsilon > 0 \tag{2.5}$$

Below is an illustration of the probability of error $\mathbb{P}(|\hat{\Theta}_n - \theta| \geq \epsilon)$ with $\epsilon = 1$ for an estimator $\hat{\Theta}_n$. The shaded area corresponds to the event $|\hat{\Theta}_n - \theta| \geq \epsilon$. As n grows, we see that the probability of error diminishes. This shows that this estimator is consistent.



(a) n = 1

(b) n = 2

(c) n = 4

(d) n = 8