Estimation Theory

Hebrew University of Jerusalem

Spring, 2022

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

1. Estimators
2. Estimator Properties
3. Maximum Likelihood
4. Estimation Quality - Measure of Concentration
5. Multivariate Distributions
6. Summary

1. Estimators
2. Estimator Properties
3. Maximum Likelihood
4. Estimation Quality - Measure of Concentration
5. Multivariate Distributions
6. Summary

- Suppose we wish to know the current time but we do not have the means of telling it. How should we find out what is the time?

- Suppose we wish to know the current time but we do not have the means of telling it. How should we find out what is the time?
- We can ask someone but will the answer be correct/accurate?

- Suppose we wish to know the current time but we do not have the means of telling it. How should we find out what is the time?
- We can ask someone but will the answer be correct/accurate?
- We can ask multiple individuals and *estimate* the actual time as the average of their answers.

Basic Terminology

- The answers are referred to as *observations* and are denoted by $x_1, \ldots, x_m$. Our estimated time is then:

$$\overline{x} = \frac{1}{m} \sum_i x_i$$

Basic Terminology

- The answers are referred to as *observations* and are denoted by $x_1, \ldots, x_m$. Our estimated time is then:

$$\overline{x} = \frac{1}{m} \sum_i x_i$$

- The set of observations is called a *sample*.

Basic Terminology

- The answers are referred to as *observations* and are denoted by $x_1, \ldots, x_m$. Our estimated time is then:

$$\overline{x} = \frac{1}{m} \sum_i x_i$$

- The set of observations is called a *sample*.
- We consider $x_1, \ldots, x_m$ as the *realizations* of the random variables $X_1, \ldots, X_m$.

Basic Terminology

- The answers are referred to as *observations* and are denoted by $x_1, \ldots, x_m$. Our estimated time is then:

$$\overline{x} = \frac{1}{m} \sum_i x_i$$

- The set of observations is called a *sample*.
- We consider $x_1, \ldots, x_m$ as the *realizations* of the random variables $X_1, \ldots, X_m$.
- We assume the samples $x_1, \ldots, x_m$ follow some underlying probability distribution $\mathcal{P}$.

Definition

We say that samples $x_1, \ldots, x_m$ are *identically distributed* if they are all the realizations of random variables over the sample probability distribution

$$X_1, \ldots, X_m \sim \mathcal{P}, \quad \forall i \; X_i = x_i$$

Basic Definitions

Definition

We say that samples $x_1, \ldots, x_m$ are *identically distributed* if they are all the realizations of random variables over the sample probability distribution

$$X_1, \ldots, X_m \sim \mathcal{P}, \quad \forall i \; X_i = x_i$$

Definition

We say that samples $x_1, \ldots, x_m$ are *independently identically distributed* (i.i.d) if they are identically distributed and are independent

$$X_1, \ldots, X_m \overset{i.i.d}{\sim} \mathcal{P}, \quad \forall i \; X_i = x_i$$

Basic Definitions

- Assume $\mathcal{P}$ is some *parametric* distribution characterized by a set of parameters $\boldsymbol{\theta} \in \Theta$.
  - $\boldsymbol{\theta}$ a vector of parameters
  - Θ the set of all possible values for the parameters

Basic Definitions

- Assume $\mathcal{P}$ is some *parametric* distribution characterized by a set of parameters $\boldsymbol{\theta} \in \Theta$.
    - $\boldsymbol{\theta}$ a vector of parameters
    - Θ the set of all possible values for the parameters

- For $\mathcal{P} := Poisson(\lambda)$ then $\boldsymbol{\theta} := \{\lambda\}$ and $\Theta := \mathbb{R}_+$
- For $\mathcal{P} := \mathcal{N}(\mu, \sigma^2)$ then $\boldsymbol{\theta} := \{\mu, \sigma^2\}$ and $\Theta := \mathbb{R} \times \mathbb{R}_+$

Basic Definitions

- Assume $\mathcal{P}$ is some *parametric* distribution characterized by a set of parameters $\boldsymbol{\theta} \in \Theta$.
  - $\boldsymbol{\theta}$ a vector of parameters
  - Θ the set of all possible values for the parameters

- For $\mathcal{P} := Poisson(\lambda)$ then $\boldsymbol{\theta} := \{\lambda\}$ and $\Theta := \mathbb{R}_+$
- For $\mathcal{P} := \mathcal{N}(\mu, \sigma^2)$ then $\boldsymbol{\theta} := \{\mu, \sigma^2\}$ and $\Theta := \mathbb{R} \times \mathbb{R}_+$

- Given a sample $x_1, \ldots, x_m$, drawn i.i.d according to $\mathcal{P}(\boldsymbol{\theta})$, where we **do not know** $\boldsymbol{\theta}$, we wish to choose $\boldsymbol{\theta}^*$ that "best" fits the true $\boldsymbol{\theta}$

Formalization

- We formulate this by defining a *decision function/rule* $\delta : \mathbb{R}^m \rightarrow \Theta$

Formalization

- We formulate this by defining a *decision function/rule* $\delta : \mathbb{R}^m \to \Theta$
    - For $\mathcal{P} := Poisson(\lambda)$:

$$(x_1, \ldots, x_m) \xrightarrow{\delta} \lambda$$

Formalization

- We formulate this by defining a *decision function/rule* $\delta : \mathbb{R}^m \to \Theta$
  - For $\mathcal{P} := Poisson(\lambda)$:

$$(x_1, \ldots, x_m) \xrightarrow{\delta} \lambda$$

  - For $\mathcal{P} := \mathcal{N}(\mu, \sigma^2)$:

$$(x_1, \ldots, x_m) \xrightarrow{\delta} (\mu, \sigma^2)$$

Formalization

- We formulate this by defining a *decision function/rule* $\delta : \mathbb{R}^m \to \Theta$
    - For $\mathcal{P} := Poisson(\lambda)$:

      $$(x_1, \ldots, x_m) \xrightarrow{\delta} \lambda$$

    - For $\mathcal{P} := \mathcal{N}(\mu, \sigma^2)$:

      $$(x_1, \ldots, x_m) \xrightarrow{\delta} (\mu, \sigma^2)$$

- Denote $\Delta := \{\delta : \mathbb{R}^m \to \Theta\}$ the set of possible decision functions (referred to as *hypothesis class*. Also denoted as $\mathcal{H}$)

Formalization

- We formulate this by defining a *decision function/rule* $\delta : \mathbb{R}^m \to \Theta$
  - For $\mathcal{P} := Poisson(\lambda)$:

$$(x_1, \ldots, x_m) \xrightarrow{\delta} \lambda$$

  - For $\mathcal{P} := \mathcal{N}(\mu, \sigma^2)$:

$$(x_1, \ldots, x_m) \xrightarrow{\delta} (\mu, \sigma^2)$$

- Denote $\Delta := \{\delta : \mathbb{R}^m \to \Theta\}$ the set of possible decision functions (referred to as *hypothesis class*. Also denoted as $\mathcal{H}$)
- We wish to choose the *optimal* $\delta \in \Delta$: $\delta \in \Delta$ such that $\boldsymbol{\theta}^* := \delta(x_1, \ldots, x_m)$ best describes $\boldsymbol{\theta}$.

Formalization

- We formulate this by defining a *decision function/rule* $\delta : \mathbb{R}^m \to \Theta$
  - For $\mathcal{P} := Poisson(\lambda)$:
  $$(x_1, \dots, x_m) \xrightarrow{\delta} \lambda$$
  - For $\mathcal{P} := \mathcal{N}(\mu, \sigma^2)$:
  $$(x_1, \dots, x_m) \xrightarrow{\delta} (\mu, \sigma^2)$$
- Denote $\Delta := \{\delta : \mathbb{R}^m \to \Theta\}$ the set of possible decision functions (referred to as *hypothesis class*. Also denoted as $\mathcal{H}$)

Definition

Let $\delta \in \Delta$ be a decision function. Then $\delta(X_1, \dots, X_m)$ is called a point statistical estimator of a parameter $\boldsymbol{\theta}$, or simply an estimator.

Estimators of Gaussians

Suppose we have obtained a sample $x_1, \ldots, x_m \overset{i.i.d}{\sim} \mathcal{P}$ for $\mathcal{N}(\mu, \sigma^2)$, and we wish to estimate μ, σ^2. How should we do so? Which estimators should we define for μ, σ^2?

Estimators of Gaussians

Suppose we have obtained a sample $x_1, \ldots, x_m \overset{i.i.d}{\sim} \mathcal{P}$ for $\mathcal{N}(\mu, \sigma^2)$, and we wish to estimate μ, σ^2. How should we do so? Which estimators should we define for μ, σ^2?

- The *sample mean* estimator is defined as:

$$\hat{\mu}_X := \frac{1}{m} \sum x_i$$

Estimators of Gaussians

Suppose we have obtained a sample $x_1, \dots, x_m \overset{i.i.d}{\sim} \mathcal{P}$ for $\mathcal{N}(\mu, \sigma^2)$, and we wish to estimate μ, σ^2. How should we do so? Which estimators should we define for μ, σ^2?

- The *sample mean* estimator is defined as:

$$\hat{\mu}_X := \frac{1}{m} \sum x_i$$

- The *sample variance* estimator is defined as:

$$\hat{\sigma}^2_X := \frac{1}{m-1} \sum (x_i - \hat{\mu}_X)^2$$

1. Estimators

2. Estimator Properties

3. Maximum Likelihood

4. Estimation Quality - Measure of Concentration

5. Multivariate Distributions

6. Summary

Estimator Properties

Defining the sample mean and -variance estimators, we could have considered many other functions

$$\hat{\sigma}_1^2 := \frac{1}{m-1}\sum\left(x_i-\hat{\mu}\right)^2, \quad \hat{\sigma}_2^2 := \frac{1}{m}\sum\left(x_i-\hat{\mu}\right)^2, \quad \hat{\sigma}_3^2 := \frac{1}{m}\sum\left|x_i-\hat{\mu}\right|$$

Estimator Properties

Defining the sample mean and -variance estimators, we could have considered many other functions

$$\hat{\sigma}_1^2 := \frac{1}{m-1}\sum\left(x_i-\hat{\mu}\right)^2, \quad \hat{\sigma}_2^2 := \frac{1}{m}\sum\left(x_i-\hat{\mu}\right)^2, \quad \hat{\sigma}_3^2 := \frac{1}{m}\sum\left|x_i-\hat{\mu}\right|$$

- Which is the “right”/“best” estimator?
- What does “right”/“best” even mean?

Estimator Properties

Defining the sample mean and -variance estimators, we could have considered many other functions

$$\hat{\sigma}_1^2 := \frac{1}{m-1}\sum\left(x_i - \hat{\mu}\right)^2, \quad \hat{\sigma}_2^2 := \frac{1}{m}\sum\left(x_i - \hat{\mu}\right)^2, \quad \hat{\sigma}_3^2 := \frac{1}{m}\sum\left|x_i - \hat{\mu}\right|$$

- Which is the “right”/“best” estimator?
- What does “right”/“best” even mean?

We answer the above questions in two steps:

- Defining/deriving properties of estimators
- We need to define some criterion by which we define what “best” means

Biasness of Estimators

Observation: An estimator δ as a function of random variables $X_1, \ldots, X_m$, it itself is a *random variable*. We can therefore ask what is the *expectation* of the estimator, its variance *variance* or its distribution

Biasness of Estimators

Observation: An estimator δ as a function of random variables $X_1, \ldots, X_m$, it itself is a *random variable*. We can therefore ask what is the *expectation* of the estimator, its variance *variance* or its distribution

When defining an estimator, one logical property we might want to have is that it will be *unbiased*. That is, that on average the estimated value will be equal to the true value.

Biasness of Estimators

Observation: An estimator δ as a function of random variables $X_1, \dots, X_m$, it itself is a *random variable*. We can therefore ask what is the *expectation* of the estimator, its variance *variance* or its distribution

When defining an estimator, one logical property we might want to have is that it will be *unbiased*. That is, that on average the estimated value will be equal to the true value.

Definition

Let δ be an estimator for a parameter θ. The difference $d := \delta(X_1, \dots, X_m) - \theta$ is called the *error* of δ.

Biasness of Estimators

Definition

Let δ be an estimator for a parameter θ. The quantity

$$\begin{aligned}\text{Bias}_\theta\left[\delta(X_1,\ldots,X_m)\right] &:= \mathbb{E}_{X_1,\ldots,X_m|\theta}\left[d\right]\\ &= \mathbb{E}_{X_1,\ldots,X_m|\theta}\left[\delta(X_1,\ldots,X_m)-\theta\right]\end{aligned}$$

is called the *bias* (or systemic error) of δ.

Biasness of Estimators

Definition

Let δ be an estimator for a parameter θ. The quantity

$$\begin{aligned}\text{Bias}_{\theta}\left[\delta(X_1,\ldots,X_m)\right] &:= \mathbb{E}_{X_1,\ldots,X_m|\theta}\left[d\right] \\ &= \mathbb{E}_{X_1,\ldots,X_m|\theta}\left[\delta(X_1,\ldots,X_m)-\theta\right]\end{aligned}$$

is called the *bias* (or systemic error) of δ.

Definition

Let δ be an estimator for a parameter θ. δ is said to be *unbiased* iff

$$\forall\theta\in\Theta \quad \text{Bias}_{\theta}\left[\delta(X_1,\ldots,X_m)\right]=0$$

Biasness of sample mean and -variance

The sample mean is an unbiased estimator:

$$\mathbb{E}_{x_1,\ldots,x_m|\mu}\left[\hat{\mu}(x_1,\ldots,x_m)\right] = \mathbb{E}_{x_1,\ldots,x_m|\mu}\left[\frac{1}{m}\sum x_i\right]$$

Biasness of sample mean and -variance

The sample mean is an unbiased estimator:

$$\begin{aligned}\mathbb{E}_{x_1,\ldots,x_m|\mu}\left[\hat{\mu}(x_1,\ldots,x_m)\right] &= \mathbb{E}_{x_1,\ldots,x_m|\mu}\left[\frac{1}{m}\sum x_i\right] \\ &= \frac{1}{m}\sum \mathbb{E}_{x_i|\mu}\left[x_i\right]\end{aligned}$$

Biasness of sample mean and -variance

The sample mean is an unbiased estimator:

$$
\begin{aligned}
\mathbb{E}_{x_1,\ldots,x_m|\mu}\left[\hat{\mu}(x_1,\ldots,x_m)\right] &= \mathbb{E}_{x_1,\ldots,x_m|\mu}\left[\tfrac{1}{m}\sum x_i\right] \\
&= \frac{1}{m}\sum \mathbb{E}_{x_i|\mu}\left[x_i\right] \\
&= \frac{1}{m}\sum \mathbb{E}_{X|\mu}\left[X\right]
\end{aligned}
$$

Biasness of sample mean and -variance

The sample mean is an unbiased estimator:

$$
\begin{aligned}
\mathbb{E}_{x_1,\ldots,x_m|\mu}\left[\hat{\mu}(x_1,\ldots,x_m)\right] &= \mathbb{E}_{x_1,\ldots,x_m|\mu}\left[\tfrac{1}{m}\sum x_i\right] \\
&= \frac{1}{m}\sum \mathbb{E}_{x_i|\mu}\left[x_i\right] \\
&= \frac{1}{m}\sum \mathbb{E}_{X|\mu}\left[X\right] \\
&= \mathbb{E}_{X|\mu}\left[X\right]
\end{aligned}
$$

Biasness of sample mean and -variance

The sample mean is an unbiased estimator:

$$\begin{aligned}
\mathbb{E}_{x_1,\ldots,x_m|\mu}\left[\hat{\mu}(x_1,\ldots,x_m)\right] &= \mathbb{E}_{x_1,\ldots,x_m|\mu}\left[\frac{1}{m}\sum x_i\right] \\
&= \frac{1}{m}\sum \mathbb{E}_{x_i|\mu}\left[x_i\right] \\
&= \frac{1}{m}\sum \mathbb{E}_{X|\mu}\left[X\right] \\
&= \mathbb{E}_{X|\mu}\left[X\right] \\
&= \mu
\end{aligned}$$

Biasness of sample mean and -variance

The sample variance is an unbiased estimator:

$$\mathbb{E}\left[\hat{\sigma}^2\right] \quad = \quad \tfrac{1}{m-1}\sum_i \mathbb{E}\left[\left(x_i - \hat{\mu}\right)^2\right]$$

Biasness of sample mean and -variance

The sample variance is an unbiased estimator:

$$\begin{aligned}\mathbb{E}\left[\hat{\sigma}^2\right] &= \tfrac{1}{m-1}\textstyle\sum_i \mathbb{E}\left[\left(x_i - \hat{\mu}\right)^2\right] \\ &= \tfrac{1}{m-1}\textstyle\sum_i \mathbb{E}\left[x_i^2 - 2x_i \cdot \tfrac{1}{m}\sum_j x_j + \tfrac{1}{m^2}\sum_{j,k} x_j x_k\right]\end{aligned}$$

Biasness of sample mean and -variance

The sample variance is an unbiased estimator:

$$
\begin{aligned}
\mathbb{E}\left[\hat{\sigma}^2\right] &= \tfrac{1}{m-1}\sum_i \mathbb{E}\left[\left(x_i - \hat{\mu}\right)^2\right] \\
&= \tfrac{1}{m-1}\sum_i \mathbb{E}\left[x_i^2 - 2x_i \cdot \tfrac{1}{m}\sum_j x_j + \tfrac{1}{m^2}\sum_{j,k} x_j x_k\right] \\
&= \tfrac{1}{m-1}\left(\sum_i \mathbb{E}\left[x_i^2\right] - \tfrac{2}{m}\sum_{i,j}\mathbb{E}\left[x_i x_j\right] + \tfrac{1}{m}\sum_{j,k}\mathbb{E}\left[x_j x_k\right]\right)
\end{aligned}
$$

Biasness of sample mean and -variance

The sample variance is an unbiased estimator:

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}^2\right] &= \tfrac{1}{m-1}\sum_i \mathbb{E}\left[(x_i-\hat{\mu})^2\right] \\
&= \tfrac{1}{m-1}\sum_i \mathbb{E}\left[x_i^2 - 2x_i\cdot\tfrac{1}{m}\sum_j x_j + \tfrac{1}{m^2}\sum_{j,k} x_j x_k\right] \\
&= \tfrac{1}{m-1}\left(\sum_i \mathbb{E}\left[x_i^2\right] - \tfrac{2}{m}\sum_{i,j}\mathbb{E}\left[x_i x_j\right] + \tfrac{1}{m}\sum_{j,k}\mathbb{E}\left[x_j x_k\right]\right) \\
&= \tfrac{1}{m-1}\left(\sum_i \mathbb{E}\left[x_i^2\right] - \tfrac{1}{m}\sum_{i,j}\mathbb{E}\left[x_i x_j\right]\right)
\end{aligned}$$

Biasness of sample mean and -variance

The sample variance is an unbiased estimator:

$$
\begin{aligned}
\mathbb{E}\left[\hat{\sigma}^2\right] &= \tfrac{1}{m-1}\sum_i \mathbb{E}\left[\left(x_i-\hat{\mu}\right)^2\right] \\
&= \tfrac{1}{m-1}\sum_i \mathbb{E}\left[x_i^2 - 2x_i\cdot\tfrac{1}{m}\sum_j x_j + \tfrac{1}{m^2}\sum_{j,k} x_j x_k\right] \\
&= \tfrac{1}{m-1}\left(\sum_i \mathbb{E}\left[x_i^2\right] - \tfrac{2}{m}\sum_{i,j}\mathbb{E}\left[x_i x_j\right] + \tfrac{1}{m}\sum_{j,k}\mathbb{E}\left[x_j x_k\right]\right) \\
&= \tfrac{1}{m-1}\left(\sum_i \mathbb{E}\left[x_i^2\right] - \tfrac{1}{m}\sum_{i,j}\mathbb{E}\left[x_i x_j\right]\right) \\
&= \tfrac{1}{m-1}\left(\left(1-\tfrac{1}{m}\right)\sum_i \mathbb{E}\left[x_i^2\right] - \tfrac{1}{m}\sum_{i\neq j}\mathbb{E}\left[x_i x_j\right]\right)
\end{aligned}
$$

Biasness of sample mean and -variance

The sample variance is an unbiased estimator:

$$
\begin{aligned}
\mathbb{E}\left[\hat{\sigma}^2\right] &= \tfrac{1}{m-1}\sum_i \mathbb{E}\left[\left(x_i-\hat{\mu}\right)^2\right]\\
&= \tfrac{1}{m-1}\sum_i \mathbb{E}\left[x_i^2-2x_i\cdot\tfrac{1}{m}\sum_j x_j+\tfrac{1}{m^2}\sum_{j,k}x_jx_k\right]\\
&= \tfrac{1}{m-1}\left(\sum_i \mathbb{E}\left[x_i^2\right]-\tfrac{2}{m}\sum_{i,j}\mathbb{E}\left[x_ix_j\right]+\tfrac{1}{m}\sum_{j,k}\mathbb{E}\left[x_jx_k\right]\right)\\
&= \tfrac{1}{m-1}\left(\sum_i \mathbb{E}\left[x_i^2\right]-\tfrac{1}{m}\sum_{i,j}\mathbb{E}\left[x_ix_j\right]\right)\\
&= \tfrac{1}{m-1}\left(\left(1-\tfrac{1}{m}\right)\sum_i \mathbb{E}\left[x_i^2\right]-\tfrac{1}{m}\sum_{i\neq j}\mathbb{E}\left[x_ix_j\right]\right)
\end{aligned}
$$

Since the samples are i.i.d then:

- For any i: $\mathbb{E}\left[x_i\right]=\mathbb{E}\left[X\right], \mathbb{E}\left[x_i^2\right]=\mathbb{E}\left[X^2\right]$
- For any $i\neq j$: $\mathbb{E}\left[x_ix_j\right]=\mathbb{E}\left[x_i\right]\mathbb{E}\left[x_j\right]=\mathbb{E}^2\left[X\right]$

Biasness of sample mean and -variance

The sample variance is an unbiased estimator:

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}^2\right] &= \tfrac{1}{m-1}\sum_i \mathbb{E}\left[\left(x_i-\hat{\mu}\right)^2\right] \\
&= \tfrac{1}{m-1}\sum_i \mathbb{E}\left[x_i^2 - 2x_i\cdot\tfrac{1}{m}\sum_j x_j + \tfrac{1}{m^2}\sum_{j,k} x_j x_k\right] \\
&= \tfrac{1}{m-1}\left(\sum_i \mathbb{E}\left[x_i^2\right] - \tfrac{2}{m}\sum_{i,j}\mathbb{E}\left[x_i x_j\right] + \tfrac{1}{m}\sum_{j,k}\mathbb{E}\left[x_j x_k\right]\right) \\
&= \tfrac{1}{m-1}\left(\sum_i \mathbb{E}\left[x_i^2\right] - \tfrac{1}{m}\sum_{i,j}\mathbb{E}\left[x_i x_j\right]\right) \\
&= \tfrac{1}{m-1}\left(\left(1-\tfrac{1}{m}\right)\sum_i \mathbb{E}\left[x_i^2\right] - \tfrac{1}{m}\sum_{i\neq j}\mathbb{E}\left[x_i x_j\right]\right) \\
&= \tfrac{1}{m-1}\left((m-1)\,\mathbb{E}\left[X^2\right] - \tfrac{m(m-1)}{m}\mathbb{E}^2\left[X\right]\right) \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}^2\left[X\right]
\end{aligned}$$

Biasness of sample mean and -variance

The sample variance is an unbiased estimator:

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}^2\right] &= \tfrac{1}{m-1}\sum_i \mathbb{E}\left[\left(x_i-\hat{\mu}\right)^2\right] \\
&= \tfrac{1}{m-1}\sum_i \mathbb{E}\left[x_i^2 - 2x_i\cdot\tfrac{1}{m}\sum_j x_j + \tfrac{1}{m^2}\sum_{j,k} x_j x_k\right] \\
&= \tfrac{1}{m-1}\left(\sum_i \mathbb{E}\left[x_i^2\right] - \tfrac{2}{m}\sum_{i,j}\mathbb{E}\left[x_i x_j\right] + \tfrac{1}{m}\sum_{j,k}\mathbb{E}\left[x_j x_k\right]\right) \\
&= \tfrac{1}{m-1}\left(\sum_i \mathbb{E}\left[x_i^2\right] - \tfrac{1}{m}\sum_{i,j}\mathbb{E}\left[x_i x_j\right]\right) \\
&= \tfrac{1}{m-1}\left(\left(1-\tfrac{1}{m}\right)\sum_i \mathbb{E}\left[x_i^2\right] - \tfrac{1}{m}\sum_{i\neq j}\mathbb{E}\left[x_i x_j\right]\right) \\
&= \tfrac{1}{m-1}\left((m-1)\,\mathbb{E}\left[X^2\right] - \tfrac{m(m-1)}{m}\mathbb{E}^2\left[X\right]\right) \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}^2\left[X\right] \\
&= Var\left(X\right) \\
&= \sigma^2
\end{aligned}$$

Variance of Estimators

Definition

Let δ be an estimator for a parameter θ. The *variance* of δ is

$$Var(\delta) := \mathbb{E}_{X_1,\dots,X_m|\theta}\left[\left(\delta(X_1,\dots,X_m) - \mathbb{E}_{X_1,\dots,X_m|\theta}\left[\delta(X_1,\dots,X_m)\right]\right)^2$$

Variance of Estimators

Let us calculate the variance of the sample mean estimator. Let $x_1 \ldots, x_m \overset{i.i.d}{\sim} \mathcal{P}$ with variance of σ^2.

Variance of Estimators

Let us calculate the variance of the sample mean estimator. Let $x_1 \ldots, x_m \overset{i.i.d}{\sim} \mathcal{P}$ with variance of σ^2.

$$Var(\hat{\mu}) \quad = \quad Var\left(\tfrac{1}{m}\sum x_i\right)$$

Variance of Estimators

Let us calculate the variance of the sample mean estimator. Let $x_1 \ldots, x_m \overset{i.i.d}{\sim} \mathcal{P}$ with variance of σ^2.

$$\begin{aligned} Var(\hat{\mu}) &= Var\left(\frac{1}{m}\sum x_i\right) \\ &= \frac{1}{m^2} Var\left(\sum x_i\right) \end{aligned}$$

Variance of Estimators

Let us calculate the variance of the sample mean estimator. Let $x_1 \ldots, x_m \overset{i.i.d}{\sim} \mathcal{P}$ with variance of σ^2.

$$\begin{array}{rcl} Var(\hat{\mu}) & = & Var\left(\frac{1}{m}\sum x_i\right) \\ & = & \frac{1}{m^2} Var\left(\sum x_i\right) \\ & = & \frac{1}{m^2}\sum Var\left(x_i\right) \end{array}$$

Variance of Estimators

Let us calculate the variance of the sample mean estimator. Let $x_1 \ldots, x_m \overset{i.i.d}{\sim} \mathcal{P}$ with variance of σ^2.

$$\begin{aligned} Var(\hat{\mu}) &= Var\left(\frac{1}{m}\sum x_i\right) \\ &= \frac{1}{m^2}Var\left(\sum x_i\right) \\ &= \frac{1}{m^2}\sum Var\left(x_i\right) \\ &= \frac{1}{m^2}\cdot m\cdot\sigma^2 \\ &= \frac{\sigma^2}{m} \end{aligned}$$

Variance of Estimators

Let us calculate the variance of the sample mean estimator. Let $x_1 \ldots, x_m \overset{i.i.d}{\sim} \mathcal{P}$ with variance of σ^2.

$$\begin{aligned} Var(\hat{\mu}) &= Var\left(\tfrac{1}{m}\sum x_i\right) \\ &= \tfrac{1}{m^2} Var\left(\sum x_i\right) \\ &= \tfrac{1}{m^2} \sum Var\left(x_i\right) \\ &= \tfrac{1}{m^2} \cdot m \cdot \sigma^2 \\ &= \tfrac{\sigma^2}{m} \end{aligned}$$

The variance of an estimator provides us with an indication of how well is the estimator performing.

1. Estimators
2. Estimator Properties
3. Maximum Likelihood
4. Estimation Quality - Measure of Concentration
5. Multivariate Distributions
6. Summary

Maximum Likelihood

Recall, given the set of possible estimators, the hypothesis class, we wanted to answer the following questions:

- Which is the “right”/“best” estimator?
- What does “right”/“best” even mean?

The properties defined above, though interesting in themselves and provide us with tools to assess our estimators, can also be used to define optimality criteria. We can decide for example that the “optimal” estimator is $\delta \in \Delta$ which is *unbiased* and with the *minimial variance* out of all unbiased estimators.

Maximum Likelihood

Another approach is by looking for the *Maximum Likelihood Estimator*, that is, the estimator under which the observed data is *most likely*.

Definition

Let $X \sim \mathcal{P}\left(\boldsymbol{\theta}\right)$ and f be the probability density function of $\mathcal{P}$. The *likelihood* function is:

$$\mathcal{L}\left(\boldsymbol{\theta}|x\right) \coloneqq f_{\boldsymbol{\theta}}\left(x\right)$$

for x the realization of X.

Maximum Likelihood of Gaussian

Consider the case of a univariate Gaussian distribution with $x_1, \ldots, x_m \overset{i.i.d}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ and denote $\boldsymbol{\theta} = \{\mu, \sigma^2\}$.

The likelihood function is:

$$\mathcal{L}\left(\boldsymbol{\theta}|x_i\right) = f_{\boldsymbol{\theta}}(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right), \quad i \in [m]$$

Maximum Likelihood of Gaussian

Consider the case of a univariate Gaussian distribution with $x_1, \dots, x_m \overset{i.i.d}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ and denote $\boldsymbol{\theta} = \{\mu, \sigma^2\}$.

The likelihood function is:

$$\mathcal{L}\left(\boldsymbol{\theta}|x_i\right) = f_{\boldsymbol{\theta}}(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right), \quad i \in [m]$$

Since $x_1, \dots, x_m$ are i.i.d then:

$$\mathcal{L}\left(\boldsymbol{\theta}|x_1, \dots, x_m\right) \quad = \quad f_{\boldsymbol{\theta}}(x_1, \dots, x_m)$$

Maximum Likelihood of Gaussian

Consider the case of a univariate Gaussian distribution with $x_1, \ldots, x_m \overset{i.i.d}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ and denote $\boldsymbol{\theta} = \{\mu, \sigma^2\}$.

The likelihood function is:

$$\mathcal{L}\left(\boldsymbol{\theta}|x_i\right) = f_{\boldsymbol{\theta}}(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right), \quad i \in [m]$$

Since $x_1, \ldots, x_m$ are i.i.d then:

$$\begin{array}{rcl} \mathcal{L}\left(\boldsymbol{\theta}|x_1, \ldots, x_m\right) & = & f_{\boldsymbol{\theta}}(x_1, \ldots, x_m) \\ & = & \prod_{i=1}^{m} f_{\boldsymbol{\theta}}(x_i) \end{array}$$

Maximum Likelihood of Gaussian

Consider the case of a univariate Gaussian distribution with $x_1, \ldots, x_m \overset{i.i.d}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ and denote $\boldsymbol{\theta} = \{\mu, \sigma^2\}$.

The likelihood function is:

$$\mathcal{L}\left(\boldsymbol{\theta} | x_i\right) = f_{\boldsymbol{\theta}}(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right), \quad i \in [m]$$

Since $x_1, \ldots, x_m$ are i.i.d then:

$$\begin{array}{rcl} \mathcal{L}\left(\boldsymbol{\theta} | x_1, \ldots, x_m\right) & = & f_{\boldsymbol{\theta}}(x_1, \ldots, x_m) \\ & = & \prod_{i=1}^m f_{\boldsymbol{\theta}}(x_i) \\ & = & \frac{1}{\left(2\pi\sigma^2\right)^{m/2}} \exp\left(-\frac{1}{2\sigma^2} \sum \left(x_i - \mu\right)^2\right) \end{array}$$

Maximum Likelihood of Gaussian

Using the likelihood function we can now provide each $\boldsymbol{\theta}$ with a quantity (the likelihood) of how likely is it to have generated the observed data.
Let $x_1, x_2, x_3, x_4 \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $x_1 = -1$, $x_2 = 0$, $x_3 = 0$, $x_4 = 1$ and $\sigma^2 = 1$.

Maximum Likelihood of Gaussian

Using the likelihood function we can now provide each $\boldsymbol{\theta}$ with a quantity (the likelihood) of how likely is it to have generated the observed data.

Let $x_1, x_2, x_3, x_4 \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $x_1 = -1$, $x_2 = 0$, $x_3 = 0$, $x_4 = 1$ and $\sigma^2 = 1$.

- The likelihood of $\mu = 0$ is:

$$\begin{aligned} \mathcal{L}(\mu = 0, \sigma^2 = 1 | x_1, x_2, x_3, x_4) &= \frac{1}{(2\pi)^2} \exp(-\frac{1}{2} \sum_{i=1}^{4} x_i^2) \\ &= \frac{1}{4\pi^2} \exp(-\frac{2}{2}) \\ &\approx 0.00931 \end{aligned}$$

Maximum Likelihood of Gaussian

Using the likelihood function we can now provide each $\boldsymbol{\theta}$ with a quantity (the likelihood) of how likely is it to have generated the observed data.

Let $x_1, x_2, x_3, x_4 \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $x_1 = -1$, $x_2 = 0$, $x_3 = 0$, $x_4 = 1$ and $\sigma^2 = 1$.

- The likelihood of $\mu = 0$ is:

$$\begin{aligned} \mathcal{L}(\mu = 0, \sigma^2 = 1 | x_1, x_2, x_3, x_4) &= \frac{1}{(2\pi)^2} \exp(-\tfrac{1}{2} \textstyle\sum_{i=1}^4 x_i^2) \\ &= \frac{1}{4\pi^2} \exp(-\tfrac{2}{2}) \\ &\approx 0.00931 \end{aligned}$$

- The likelihood of $\mu = 1$ is:

$$\begin{aligned} \mathcal{L}(\mu = 1, \sigma^2 = 1 | x_1, x_2, x_3, x_4) &= \frac{1}{(2\pi)^2} \exp(-\tfrac{1}{2} \textstyle\sum_{i=1}^4 (x_i - 1)^2) \\ &= \frac{1}{4\pi^2} \exp(-\tfrac{2^2+1+1}{2}) \\ &\approx 0.00126 \end{aligned}$$

Maximum Likelihood Estimator

Definition

Let $\mathcal{L}$ be the likelihood function of some probability distribution $\mathcal{P}$ characterized by $\boldsymbol{\theta} \in \Theta$. Let $X \sim \mathcal{P}(\boldsymbol{\theta})$ be a random variable and x its realization. The *Maximum Likelihood Estimator* (MLE) for $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}^{MLE} := \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \ \mathcal{L}\left(\boldsymbol{\theta} | x\right)$$

Let us derive the MLE of a univariate Gaussian distribution's mean $\hat{\mu}^{MLE}$, when σ^2 is known. Let $x_1, \ldots, x_m \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$. We wish to find

$$\hat{\mu}^{MLE} = \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \ \mathcal{L}(\mu | x_1, \ldots, x_m, \sigma^2)$$

Maximum Likelihood Estimator

$$\begin{aligned}\hat{\mu}^{MLE} &= \operatorname*{argmax}_{\mu\in\mathbb{R}} \mathcal{L}(\mu|x_1,\ldots,x_m,\sigma^2) \\ &= \operatorname*{argmax}_{\mu\in\mathbb{R}} \textstyle\prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\end{aligned}$$

Maximum Likelihood Estimator

$$\begin{aligned}
\hat{\mu}^{MLE} &= \operatorname*{argmax}_{\mu\in\mathbb{R}} \mathcal{L}(\mu|x_1,\ldots,x_m,\sigma^2) \\
&= \operatorname*{argmax}_{\mu\in\mathbb{R}} \textstyle\prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \operatorname*{argmax}_{\mu\in\mathbb{R}} \textstyle\prod_i \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)
\end{aligned}$$

Maximum Likelihood Estimator

$$\begin{aligned}
\hat{\mu}^{MLE} &= \underset{\mu\in\mathbb{R}}{\operatorname{argmax}}\ \mathcal{L}(\mu|x_1,\ldots,x_m,\sigma^2) \\
&= \underset{\mu\in\mathbb{R}}{\operatorname{argmax}}\ \prod_i \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \underset{\mu\in\mathbb{R}}{\operatorname{argmax}}\ \prod_i \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \underset{\mu\in\mathbb{R}}{\operatorname{argmax}}\ \exp\left(-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right)
\end{aligned}$$

Maximum Likelihood Estimator

$$\begin{aligned}
\hat{\mu}^{MLE} &= \operatorname*{argmax}_{\mu\in\mathbb{R}} \mathcal{L}(\mu|x_1,\ldots,x_m,\sigma^2) \\
&= \operatorname*{argmax}_{\mu\in\mathbb{R}} \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \operatorname*{argmax}_{\mu\in\mathbb{R}} \prod_i \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \operatorname*{argmax}_{\mu\in\mathbb{R}} \exp\left(-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right)
\end{aligned}$$

Notice that the maximizer of the likelihood is also the maximizer of the *log-likelihood* as the logarithm is a monotonous increasing transformation. Thus:

Maximum Likelihood Estimator

$$\begin{aligned}\hat{\mu}^{MLE} &= \underset{\mu\in\mathbb{R}}{\operatorname{argmax}}\ \log\left(\exp\left(-\tfrac{1}{2\sigma^2}\textstyle\sum_i(x_i-\mu)^2\right)\right)\\ &= \underset{\mu\in\mathbb{R}}{\operatorname{argmax}}\ -\tfrac{1}{2\sigma^2}\textstyle\sum_i(x_i-\mu)^2\end{aligned}$$

Maximum Likelihood Estimator

$$\begin{aligned}
\hat{\mu}^{MLE} &= \underset{\mu\in\mathbb{R}}{\operatorname{argmax}}\ \log\left(\exp\left(-\tfrac{1}{2\sigma^2}\textstyle\sum_i (x_i-\mu)^2\right)\right) \\
&= \underset{\mu\in\mathbb{R}}{\operatorname{argmax}}\ -\tfrac{1}{2\sigma^2}\textstyle\sum_i (x_i-\mu)^2 \\
&= \underset{\mu\in\mathbb{R}}{\operatorname{argmax}}\ -\textstyle\sum_i (x_i-\mu)^2
\end{aligned}$$

Maximum Likelihood Estimator

$$\begin{aligned}
\hat{\mu}^{MLE} &= \operatorname*{argmax}_{\mu\in\mathbb{R}} \log\left(\exp\left(-\tfrac{1}{2\sigma^2}\textstyle\sum_i (x_i-\mu)^2\right)\right) \\
&= \operatorname*{argmax}_{\mu\in\mathbb{R}} -\tfrac{1}{2\sigma^2}\textstyle\sum_i (x_i-\mu)^2 \\
&= \operatorname*{argmax}_{\mu\in\mathbb{R}} -\textstyle\sum_i (x_i-\mu)^2
\end{aligned}$$

To find the maximizer, we derive (with respect to μ) and equate to zero:

$$\frac{\partial}{\partial\mu}\left(-\textstyle\sum_{i=1}^m (x_i-\mu)^2\right) = -\textstyle\sum_{i=1}^m \frac{\partial(x_i-\mu)^2}{\partial\mu} = \textstyle\sum_{i=1}^m 2\left(x_i-\mu\right) = 0$$

$$\Downarrow$$

$$\hat{\mu}^{MLE} = \tfrac{1}{m}\textstyle\sum_{i=1}^m x_i$$

Maximum Likelihood Estimator

$$\begin{aligned}\hat{\mu}^{MLE} &= \underset{\mu\in\mathbb{R}}{\text{argmax}}\ \log\left(\exp\left(-\tfrac{1}{2\sigma^2}\textstyle\sum_i(x_i-\mu)^2\right)\right)\\ &= \underset{\mu\in\mathbb{R}}{\text{argmax}}\ -\tfrac{1}{2\sigma^2}\textstyle\sum_i(x_i-\mu)^2\\ &= \underset{\mu\in\mathbb{R}}{\text{argmax}}\ -\textstyle\sum_i(x_i-\mu)^2\end{aligned}$$

To find the maximizer, we derive (with respect to μ) and equate to zero:

$$\frac{\partial}{\partial\mu}\left(-\sum_{i=1}^m (x_i-\mu)^2\right) = -\sum_{i=1}^m \frac{\partial(x_i-\mu)^2}{\partial\mu} = \sum_{i=1}^m 2(x_i-\mu) = 0$$

$$\Downarrow$$

$$\hat{\mu}^{MLE} = \frac{1}{m}\sum_{i=1}^m x_i$$

We therefore also conclude that the sample mean minimizes the sum of squared distances

1. Estimators
2. Estimator Properties
3. Maximum Likelihood
4. Estimation Quality - Measure of Concentration
5. Multivariate Distributions
6. Summary

The Distribution of An Estimator

- Since an estimator is a *random variable*, and as we have seen that we could calculate it's expected value and variance, a natural question is to ask “how does the estimator distribute”.

The Distribution of An Estimator

- Since an estimator is a *random variable*, and as we have seen that we could calculate it's expected value and variance, a natural question is to ask “how does the estimator distribute”.
- In the case of the sample mean estimator, We have shown that for $x_1, \ldots, x_m \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$ it holds that:

$$\mathbb{E}\left[\hat{\mu}\right] = \mu, \quad Var\left(\hat{\mu}\right) = \frac{\sigma^2}{m}$$

It can further be shown that: $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{m}\right)$

The Distribution of An Estimator

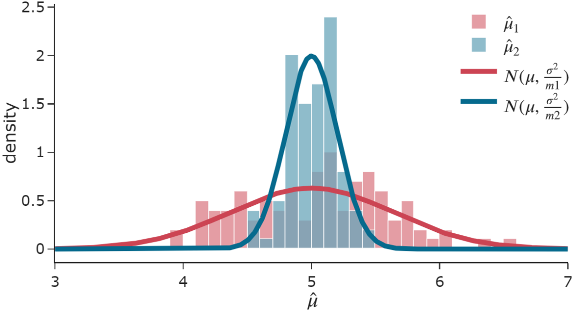
- Since an estimator is a *random variable*, and as we have seen that we could calculate it's expected value and variance, a natural question is to ask "how does the estimator distribute".
- In the case of the sample mean estimator, We have shown that for $x_1, \ldots, x_m \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$ it holds that:

$$\mathbb{E}\left[\hat{\mu}\right] = \mu, \quad Var\left(\hat{\mu}\right) = \frac{\sigma^2}{m}$$

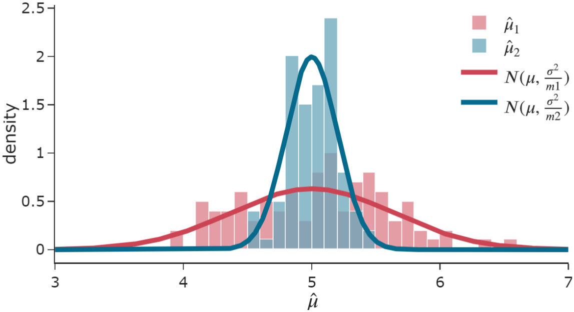
It can further be shown that: $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{m}\right)$

- Namely, this estimator follows a Gaussian distribution centered at the true value μ (unbiased) and with a variance proportional to the variance of the data, and that decays linearly in the number of samples.

The Distribution of An Estimator



The Distribution of An Estimator



We can therefore ask questions such as:

- What is the probability of estimating a certain value?
- How certain are we in our prediction? (variance? confidence-interval?)
- What is the probability of deviating from the true value, and how does it depend on the number of samples?

Coin Prediction

- Consider the task of estimating the bias of a coin. That is, with what probability p flipping the coin will result with “Heads”.

Coin Prediction

- Consider the task of estimating the bias of a coin. That is, with what probability p flipping the coin will result with “Heads”.
- Formally, we think of (i.e *model*) a single coin flip as a Bernoulli random variable:

$$\mathcal{D}_p\left(X\right) := \begin{cases} p & X=1 \\ 1-p & X=0 \end{cases}, \quad p \in [0,1]$$

Coin Prediction

- Consider the task of estimating the bias of a coin. That is, with what probability p flipping the coin will result with "Heads".
- Formally, we think of (i.e *model*) a single coin flip as a Bernoulli random variable:

$$\mathcal{D}_p\left(X\right) := \begin{cases} p & X = 1 \\ 1 - p & X = 0 \end{cases}, \quad p \in [0,1]$$

- Given a sample of m coin tosses $S := \{x_1, \ldots, x_m\}$ we denote the probability distribution of m coin tosses by $\mathcal{D}_p^m$ and the probability of obtaining S by $\mathcal{D}_p^m(S)$.

Coin Prediction

- Given such a sample we would like to devide a *learning algorithm* $\mathcal{A}$ to estimate/predict p and then we will ask how accurate is our algorithm.

Coin Prediction

- Given such a sample we would like to devide a *learning algorithm* $\mathcal{A}$ to estimate/predict p and then we will ask how accurate is our algorithm.

- The coin prediction learning algorithm takes S as input, drawn i.i.d according to $\mathcal{D}^m$ and outputs an estimation of p. We denote this by $\mathcal{A}(S)$ or $\hat{p}(S)$ or simply $\hat{p}$.

Coin Prediction

- Given such a sample we would like to devide a *learning algorithm* $\mathcal{A}$ to estimate/predict p and then we will ask how accurate is our algorithm.
- The coin prediction learning algorithm takes S as input, drawn i.i.d according to $\mathcal{D}^m$ and outputs an estimation of p. We denote this by $\mathcal{A}(S)$ or $\hat{p}(S)$ or simply $\hat{p}$.
- The learning algorithm we will use is to simply calculate the *empirical* proportion of heads, i.e the sample mean:

$$\hat{p}(S) = \frac{1}{m} \sum_i x_i$$

Coin Prediction

- We already know that this algorithm (estimator) is unbiased. Namely, if we sample many sets $S_j = \{x_1^{(j)}, \ldots, x_m^{(j)}\}$ and for each obtain $\hat{p}_j$, their average will be p. Formally $\mathbb{E}_S\left[\hat{p}\right] = p$.

Coin Prediction

- We already know that this algorithm (estimator) is unbiased. Namely, if we sample many sets $S_j = \{x_1^{(j)}, \ldots, x_m^{(j)}\}$ and for each obtain $\hat{p}_j$, their average will be p. Formally $\mathbb{E}_S\left[\hat{p}\right] = p$.

- **Problem:** Finite sample sets and usually only one set. Our estimation is not likely to be exactly p. We therefore settle for $\mathcal{A}$ to be accurate enough: $|\hat{p} - p| \leq \varepsilon$ for some $\varepsilon \in (0, 1)$.

Coin Prediction

- We already know that this algorithm (estimator) is unbiased. Namely, if we sample many sets $S_j = \{x_1^{(j)}, \ldots, x_m^{(j)}\}$ and for each obtain $\hat{p}_j$, their average will be p. Formally $\mathbb{E}_S\left[\hat{p}\right] = p$.

- **Problem:** Finite sample sets and usually only one set. Our estimation is not likely to be exactly p. We therefore settle for $\mathcal{A}$ to be accurate enough: $|\hat{p} - p| \leq \varepsilon$ for some $\varepsilon \in (0, 1)$.

- **Problem:** Even then, unless p equals exactly to 0 or 1, there is always *some* chance to obtain S that is highly non-representative. We will therefore ask how often is $\mathcal{A}$ accurate enough, that is calculate $\mathbb{P}\left(|\hat{p} - p| \leq \varepsilon\right)$.

Coin Prediction - Markov Bounding

Recall Markov's Inequality which states that for a non-negative random variable X, with a finite expectation, for any $a > 0$ then:

$$\mathbb{P}\left(X \geq a\right) \leq \frac{\mathbb{E}\left[X\right]}{a}$$

Coin Prediction - Markov Bounding

Recall Markov's Inequality which states that for a non-negative random variable X, with a finite expectation, for any $a > 0$ then:

$$\mathbb{P}\left(X \geq a\right) \leq \frac{\mathbb{E}\left[X\right]}{a}$$

As $|\hat{p} - p|$ is a non-negative random variable, for any accuracy level $\varepsilon \in (0, 1)$ we can bound the probability of deviating from p by more than ε by:

$$\mathcal{D}^m\left[|\hat{p} - p| \geq \varepsilon\right] \leq \frac{\mathbb{E}\left[|\hat{p} - p|\right]}{\varepsilon}$$

Coin Prediction - Markov Bounding

We bound the expectation from above as follows. Since $Var(A) = \mathbb{E}\left[A^2\right] - \mathbb{E}^2\left[A\right]$ then:

$$Var\left(|\hat{p} - p|\right) = \mathbb{E}\left[|\hat{p} - p|^2\right] - \mathbb{E}^2\left[|\hat{p} - p|\right]$$

Coin Prediction - Markov Bounding

We bound the expectation from above as follows. Since $Var(A) = \mathbb{E}\left[A^2\right] - \mathbb{E}^2\left[A\right]$ then:

$$Var\left(\left|\hat{p}-p\right|\right) = \mathbb{E}\left[\left|\hat{p}-p\right|^2\right] - \mathbb{E}^2\left[\left|\hat{p}-p\right|\right]$$
$$\Downarrow$$
$$\mathbb{E}^2\left[\left|\hat{p}-p\right|\right] \leq \mathbb{E}\left[\left|\hat{p}-p\right|^2\right]$$

Coin Prediction - Markov Bounding

We bound the expectation from above as follows. Since $Var(A) = \mathbb{E}\left[A^2\right] - \mathbb{E}^2\left[A\right]$ then:

$$Var\left(|\hat{p} - p|\right) = \mathbb{E}\left[|\hat{p} - p|^2\right] - \mathbb{E}^2\left[|\hat{p} - p|\right]$$
$$\Downarrow$$
$$\mathbb{E}^2\left[|\hat{p} - p|\right] \leq \mathbb{E}\left[|\hat{p} - p|^2\right]$$

Next:

$$\mathbb{E}^2\left[|\hat{p} - p|\right] \quad \leq \quad \mathbb{E}\left[|\hat{p} - p|^2\right] \quad = \quad \mathbb{E}\left[(\hat{p} - p)^2\right]$$

Coin Prediction - Markov Bounding

We bound the expectation from above as follows. Since $Var(A) = \mathbb{E}\left[A^2\right] - \mathbb{E}^2\left[A\right]$ then:

$$Var\left(|\hat{p}-p|\right) = \mathbb{E}\left[|\hat{p}-p|^2\right] - \mathbb{E}^2\left[|\hat{p}-p|\right]$$
$$\Downarrow$$
$$\mathbb{E}^2\left[|\hat{p}-p|\right] \le \mathbb{E}\left[|\hat{p}-p|^2\right]$$

Next:

$$\begin{aligned}
\mathbb{E}^2\left[|\hat{p}-p|\right] &\le \mathbb{E}\left[|\hat{p}-p|^2\right] = \mathbb{E}\left[(\hat{p}-p)^2\right] \\
&= \mathbb{E}\left[(\hat{p}-\mathbb{E}\left[\hat{p}\right])^2\right]
\end{aligned}$$

Coin Prediction - Markov Bounding

We bound the expectation from above as follows. Since $Var(A) = \mathbb{E}\left[A^2\right] - \mathbb{E}^2\left[A\right]$ then:

$$Var\left(|\hat{p}-p|\right) = \mathbb{E}\left[|\hat{p}-p|^2\right] - \mathbb{E}^2\left[|\hat{p}-p|\right]$$
$$\Downarrow$$
$$\mathbb{E}^2\left[|\hat{p}-p|\right] \leq \mathbb{E}\left[|\hat{p}-p|^2\right]$$

Next:

$$\begin{aligned} \mathbb{E}^2\left[|\hat{p}-p|\right] &\leq \mathbb{E}\left[|\hat{p}-p|^2\right] &&= \mathbb{E}\left[(\hat{p}-p)^2\right] \\ &= \mathbb{E}\left[(\hat{p}-\mathbb{E}\left[\hat{p}\right])^2\right] &&= Var\left(\hat{p}\right) \end{aligned}$$

Coin Prediction - Markov Bounding

We bound the expectation from above as follows. Since $Var(A) = \mathbb{E}\left[A^2\right] - \mathbb{E}^2\left[A\right]$ then:

$$Var\left(|\hat{p}-p|\right) = \mathbb{E}\left[|\hat{p}-p|^2\right] - \mathbb{E}^2\left[|\hat{p}-p|\right]$$
$$\Downarrow$$
$$\mathbb{E}^2\left[|\hat{p}-p|\right] \leq \mathbb{E}\left[|\hat{p}-p|^2\right]$$

Next:

$$\begin{aligned}
\mathbb{E}^2\left[|\hat{p}-p|\right] &\leq \mathbb{E}\left[|\hat{p}-p|^2\right] &&= \mathbb{E}\left[(\hat{p}-p)^2\right] \\
&= \mathbb{E}\left[(\hat{p}-\mathbb{E}\left[\hat{p}\right])^2\right] &&= Var\left(\hat{p}\right) \\
&= Var\left(\tfrac{1}{m}\sum x_i\right)
\end{aligned}$$

Coin Prediction - Markov Bounding

We bound the expectation from above as follows. Since $Var(A) = \mathbb{E}\left[A^2\right] - \mathbb{E}^2\left[A\right]$ then:

$$Var\left(|\hat{p}-p|\right) = \mathbb{E}\left[|\hat{p}-p|^2\right] - \mathbb{E}^2\left[|\hat{p}-p|\right]$$
$$\Downarrow$$
$$\mathbb{E}^2\left[|\hat{p}-p|\right] \leq \mathbb{E}\left[|\hat{p}-p|^2\right]$$

Next:

$$\begin{aligned} \mathbb{E}^2\left[|\hat{p}-p|\right] &\leq \mathbb{E}\left[|\hat{p}-p|^2\right] &&= \mathbb{E}\left[(\hat{p}-p)^2\right] \\ &= \mathbb{E}\left[(\hat{p}-\mathbb{E}\left[\hat{p}\right])^2\right] &&= Var\left(\hat{p}\right) \\ &= Var\left(\frac{1}{m}\sum x_i\right) &&= \frac{1}{m^2}Var\left(\sum x_i\right) \end{aligned}$$

Coin Prediction - Markov Bounding

We bound the expectation from above as follows. Since $Var(A) = \mathbb{E}\left[A^2\right] - \mathbb{E}^2\left[A\right]$ then:

$$Var\left(|\hat{p}-p|\right) = \mathbb{E}\left[|\hat{p}-p|^2\right] - \mathbb{E}^2\left[|\hat{p}-p|\right]$$
$$\Downarrow$$
$$\mathbb{E}^2\left[|\hat{p}-p|\right] \leq \mathbb{E}\left[|\hat{p}-p|^2\right]$$

Next:

$$\begin{array}{rclcl}
\mathbb{E}^2\left[|\hat{p}-p|\right] & \leq & \mathbb{E}\left[|\hat{p}-p|^2\right] & = & \mathbb{E}\left[(\hat{p}-p)^2\right] \\
 & = & \mathbb{E}\left[(\hat{p}-\mathbb{E}\left[\hat{p}\right])^2\right] & = & Var\left(\hat{p}\right) \\
 & = & Var\left(\frac{1}{m}\sum x_i\right) & = & \frac{1}{m^2}Var\left(\sum x_i\right) \\
 & = & \frac{p(1-p)}{m} & \leq & \frac{1}{4m}
\end{array}$$

Coin Prediction - Markov Bounding

Therefore:

$$\mathbb{E}^2\left[|\hat{p} - p|\right] \leq \frac{1}{4m} \quad \Rightarrow \quad \mathbb{E}\left[|\hat{p} - p|\right] \leq \frac{1}{\sqrt{4m}}$$

Coin Prediction - Markov Bounding

Therefore:

$$\mathbb{E}^2\left[|\hat{p}-p|\right] \leq \frac{1}{4m} \quad \Rightarrow \quad \mathbb{E}\left[|\hat{p}-p|\right] \leq \frac{1}{\sqrt{4m}}$$

Put all together then:

$$\mathcal{D}^m\left[|\hat{p}-p| \geq \varepsilon\right] \leq \frac{\mathbb{E}[|\hat{p}-p|]}{\varepsilon} \leq \frac{1}{\sqrt{4m\varepsilon^2}}$$

$$\Updownarrow$$

$$\mathcal{D}^m\left[|\hat{p}-p| \leq \varepsilon\right] \geq 1-\frac{1}{\sqrt{4m\varepsilon^2}}$$

Coin Prediction - Markov Bounding

As a last step, denote $\delta = \frac{1}{\sqrt{4m\varepsilon^2}}$. Solving for m then: $m = \frac{1}{4\varepsilon^2\delta^2}$

Coin Prediction - Markov Bounding

As a last step, denote $\delta = \frac{1}{\sqrt{4m\varepsilon^2}}$. Solving for m then: $m = \frac{1}{4\varepsilon^2\delta^2}$
This means that for any $\varepsilon, \delta \in (0,1)$, if we use a sample of size at least $m = \lceil \frac{1}{4\varepsilon^2\delta^2} \rceil$ then

$$\mathcal{D}^m \left[|\hat{p} - p| \leq \varepsilon \right] \geq 1 - \delta$$

Coin Prediction - Markov Bounding

As a last step, denote $\delta = \frac{1}{\sqrt{4m\varepsilon^2}}$. Solving for m then: $m = \frac{1}{4\varepsilon^2\delta^2}$ This means that for any $\varepsilon, \delta \in (0,1)$, if we use a sample of size at least $m = \lceil \frac{1}{4\varepsilon^2\delta^2} \rceil$ then

$$\mathcal{D}^m \left[|\hat{p} - p| \leq \varepsilon \right] \geq 1 - \delta$$

Namely, by providing *enough* samples, we can guarantee with *confidence* of $1 - \delta$ that our estimation is *accurate* (ε) enough.

Coin Prediction - Chebyshev Bounding

We can improve the upper bound achieved using Markov's Inequality by including the estimator's second moment (i.e its variance).

- That is, achieve a tighter bound on the needed number of samples to achieve a given accuracy and confidence levels.

Definition

Let X be a random variable with a finite mean and variance. Then, for every $\varepsilon \geq 0$:

$$\mathbb{P}\left(|X - \mathbb{E}\left[X\right] \geq \varepsilon|\right) \leq \frac{Var(X)}{\varepsilon^2}$$

Coin Prediction - Chebyshev Bounding

So, given a sample $x_1 \dots, x_m \overset{i.i.d}{\sim} Ber(p)$ we bound the deviance of $\hat{p}$ from its expected value as follows:

$$\mathbb{P}\left[|\hat{p} - \mathbb{E}\left[\hat{p}\right]| \geq \varepsilon\right] \overset{unbiased}{=} \mathbb{P}\left[|\hat{p} - p| \geq \varepsilon\right] \overset{Cheb.}{\leq} \frac{Var(\hat{p})}{\varepsilon^2}$$

Coin Prediction - Chebyshev Bounding

So, given a sample $x_1 \ldots, x_m \overset{i.i.d}{\sim} Ber(p)$ we bound the deviance of $\hat{p}$ from its expected value as follows:

$$\begin{aligned}
\mathbb{P}\left[\left|\hat{p}-\mathbb{E}\left[\hat{p}\right]\right| \geq \varepsilon\right] \quad &\overset{unbiased}{=} \quad \mathbb{P}\left[\left|\hat{p}-p\right| \geq \varepsilon\right] \quad \overset{Cheb.}{\leq} \quad \frac{Var(\hat{p})}{\varepsilon^2} \\
&= \quad \frac{1}{\varepsilon^2} Var\left(\frac{1}{m}\sum x_i\right) \quad \overset{i.i.d}{=} \quad \frac{1}{m^2\varepsilon^2}\sum Var\left(x_i\right)
\end{aligned}$$

Coin Prediction - Chebyshev Bounding

So, given a sample $x_1 \ldots, x_m \overset{i.i.d}{\sim} Ber(p)$ we bound the deviance of $\hat{p}$ from its expected value as follows:

$$\begin{array}{lclcl}
\mathbb{P}\left[\left|\hat{p} - \mathbb{E}\left[\hat{p}\right]\right| \geq \varepsilon\right] & \overset{unbiased}{=} & \mathbb{P}\left[\left|\hat{p} - p\right| \geq \varepsilon\right] & \overset{Cheb.}{\leq} & \frac{Var(\hat{p})}{\varepsilon^2} \\
 & = & \frac{1}{\varepsilon^2} Var\left(\frac{1}{m}\sum x_i\right) & \overset{i.i.d}{=} & \frac{1}{m^2\varepsilon^2}\sum Var\left(x_i\right) \\
 & = & \frac{p(1-p)}{m\varepsilon^2} & \leq & \frac{1}{4m\varepsilon^2}
\end{array}$$

Coin Prediction - Chebyshev Bounding

So, given a sample $x_1 \ldots, x_m \overset{i.i.d}{\sim} Ber(p)$ we bound the deviance of $\hat{p}$ from its expected value as follows:

$$\begin{aligned}
\mathbb{P}\left[\left|\hat{p}-\mathbb{E}\left[\hat{p}\right]\right| \geq \varepsilon\right] \quad & \overset{unbiased}{=} \quad \mathbb{P}\left[\left|\hat{p}-p\right| \geq \varepsilon\right] & \overset{Cheb.}{\leq} \quad & \frac{Var(\hat{p})}{\varepsilon^2} \\
& = \quad \tfrac{1}{\varepsilon^2} Var\left(\tfrac{1}{m}\sum x_i\right) & \overset{i.i.d}{=} \quad & \frac{1}{m^2\varepsilon^2}\sum Var\left(x_i\right) \\
& = \quad \tfrac{p(1-p)}{m\varepsilon^2} & \leq \quad & \frac{1}{4m\varepsilon^2}
\end{aligned}$$

- Bounds tend to zero as: Chebyshev $\frac{1}{m}$, Markov $\frac{1}{\sqrt{m}}$

Coin Prediction - Chebyshev Bounding

So, given a sample $x_1 \dots, x_m \overset{i.i.d}{\sim} Ber(p)$ we bound the deviance of $\hat{p}$ from its expected value as follows:

$$\begin{array}{llll} \mathbb{P}\left[\left|\hat{p}-\mathbb{E}\left[\hat{p}\right]\right| \geq \varepsilon\right] & \overset{unbiased}{=} & \mathbb{P}\left[\left|\hat{p}-p\right| \geq \varepsilon\right] & \overset{Cheb.}{\leq} & \frac{Var(\hat{p})}{\varepsilon^2} \\ & = & \frac{1}{\varepsilon^2} Var\left(\frac{1}{m}\sum x_i\right) & \overset{i.i.d}{=} & \frac{1}{m^2\varepsilon^2}\sum Var\left(x_i\right) \\ & = & \frac{p(1-p)}{m\varepsilon^2} & \leq & \frac{1}{4m\varepsilon^2} \end{array}$$

- Bounds tend to zero as: Chebyshev $\frac{1}{m}$, Markov $\frac{1}{\sqrt{m}}$
- As before, we denote $\delta = \frac{1}{4m\varepsilon^2}$ and solve for m: $m = \frac{1}{4\delta\varepsilon^2}$

Coin Prediction - Chebyshev Bounding

So, given a sample $x_1 \dots, x_m \overset{i.i.d}{\sim} Ber(p)$ we bound the deviance of $\hat{p}$ from its expected value as follows:

$$\begin{aligned}
\mathbb{P}\left[|\hat{p} - \mathbb{E}\left[\hat{p}\right]| \geq \varepsilon\right] &\overset{unbiased}{=} \mathbb{P}\left[|\hat{p} - p| \geq \varepsilon\right] &&\overset{Cheb.}{\leq} \frac{Var(\hat{p})}{\varepsilon^2} \\
&= \frac{1}{\varepsilon^2} Var\left(\frac{1}{m}\sum x_i\right) &&\overset{i.i.d}{=} \frac{1}{m^2\varepsilon^2}\sum Var\left(x_i\right) \\
&= \frac{p(1-p)}{m\varepsilon^2} &&\leq \frac{1}{4m\varepsilon^2}
\end{aligned}$$

- Bounds tend to zero as: Chebyshev $\frac{1}{m}$, Markov $\frac{1}{\sqrt{m}}$
- As before, we denote $\delta = \frac{1}{4m\varepsilon^2}$ and solve for m: $m = \frac{1}{4\delta\varepsilon^2}$

We conclude that for any $\varepsilon, \delta \in (0,1)$, if we provide $m \geq \lceil \frac{1}{4\delta\varepsilon^2} \rceil$ samples then:

$$\mathcal{D}^m\left[|\hat{p} - p| \leq \varepsilon\right] \geq 1 - \delta$$

1. Estimators
2. Estimator Properties
3. Maximum Likelihood
4. Estimation Quality - Measure of Concentration
5. Multivariate Distributions
6. Summary

Multivariate Distributions

- Up to this point, we have dealt with univariate estimation tasks, i.e. estimating a single value and thus modeling the problem with random variables taking a single value.

Multivariate Distributions

- Up to this point, we have dealt with univariate estimation tasks, i.e. estimating a single value and thus modeling the problem with random variables taking a single value.
- How should we deal with a situation of estimating multiple values, where these values may (or may not) influence one another?

Multivariate Distributions

- Up to this point, we have dealt with univariate estimation tasks, i.e. estimating a single value and thus modeling the problem with random variables taking a single value.
- How should we deal with a situation of estimating multiple values, where these values may (or may not) influence one another?

Consider for example describing human weight and height.

Multivariate Distributions

- Up to this point, we have dealt with univariate estimation tasks, i.e. estimating a single value and thus modeling the problem with random variables taking a single value.
- How should we deal with a situation of estimating multiple values, where these values may (or may not) influence one another?

Consider for example describing human weight and height.

- It is possible to model (i.e describe how the data behaves) each push property independently. For example:

$$w_1, \ldots, w_m \overset{i.i.d}{\sim} \mathcal{N}\left(75, 3\right)$$

$$h_1, \ldots, h_m \overset{i.i.d}{\sim} \mathcal{N}\left(170, 5\right)$$

Multivariate Distributions

- We know however that these properties are linked. There exists (“in the data”/ “in the nature of the problem”) a connection between the two such that, in general, the taller an individual the higher the weight.

Multivariate Distributions

- We know however that these properties are linked. There exists (“in the data”/ “in the nature of the problem”) a connection between the two such that, in general, the taller an individual the higher the weight.
- To model such connection we must first define *multivarite* probabilities.

Multivariate Distributions

- We know however that these properties are linked. There exists (“in the data”/ “in the nature of the problem”) a connection between the two such that, in general, the taller an individual the higher the weight.
- To model such connection we must first define *multivarite* probabilities.

Definition

Let $X_1, \ldots, X_d$ be a finite set of random variables defined over the same probability space. Then $X := (X_1, \ldots, X_d)^\top$ is called a *random vector*

That is, a random vector is a map from the probability space to $\mathbb{R}^d$.

Multivariate Distributions

Definition

Let $X := (X_1, \ldots, X_d)^\top$ be a random vector. The *joint probability distribution distribution* of $X_1, \ldots, X_d$ is the corresponding probability distribution on all possible outputs of $X_1, \ldots, X_d$.

Multivariate Distributions

Definition

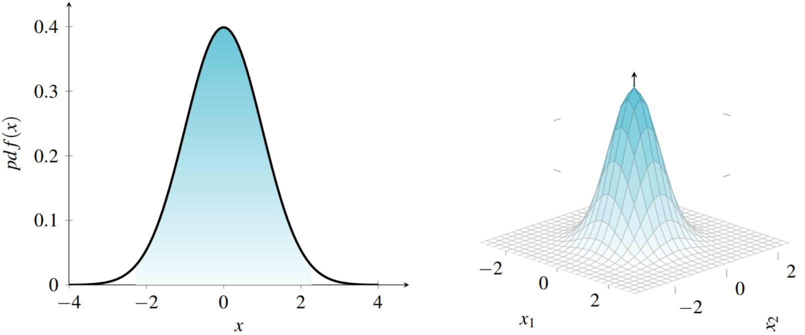
Let $X := (X_1, \ldots, X_d)^\top$ be a random vector. The *joint probability distribution distribution* of $X_1, \ldots, X_d$ is the corresponding probability distribution on all possible outputs of $X_1, \ldots, X_d$.

Definition

Let $X := (X_1, \ldots, X_d)^\top$ be a random vector. The *covariance matrix* Σ is a $d \times d$ matrix whose (i, j) entry is the covariance

$$\Sigma_{ij} := COV(X_i, X_j) = \mathbb{E}\left[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\right]$$

Multivariate Gaussian Distribution

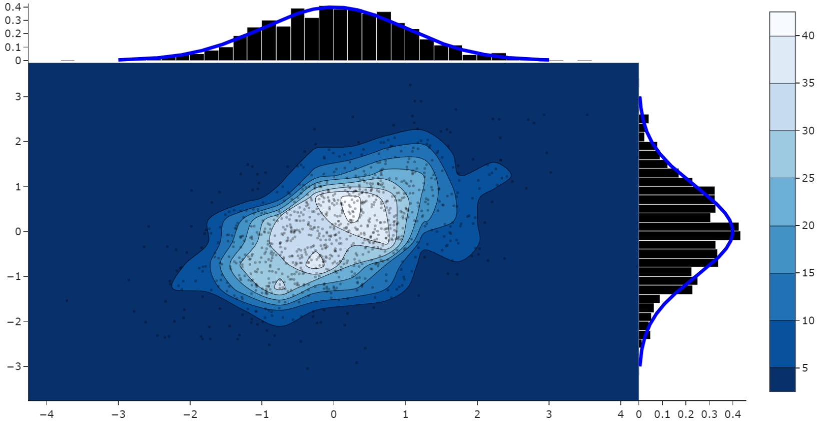


***(a)** Univariate Gaussian* $\mathcal{N}(0,1)$

***(b)** Bivariate Gaussian* $\mathcal{N}(0,I_2)$

Note: the contour lines of a Gaussian distribution as ellipsoids.

Multivariate Gaussian Distribution - Empirical



Note: the contour lines simply capture the *empirical* density of the data.

Multivariate Gaussian Distribution

Definition

A random vector $X := (X_1, \ldots, X_d)^\top$ follows a *multivariate normal distribution* with expectation $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ if it has a joint probability density function of the form:

$$f\left(X\right) = \frac{1}{\sqrt{\left(2\pi\right)^d \left|\Sigma\right|}} \exp\left\{-\frac{1}{2}(X - \mu)^\top \Sigma^{-1} \left(X - \mu\right)\right\}$$

In this case we write $X \sim \mathcal{N}\left(\mu, \Sigma\right)$

Definition

Let $X := (X_1, \ldots, X_d)^\top$ be a random vector with a joint probability distribution function. The *marginal distribution* of a subset of coordinates $A \in [d]$ and $B = [d] \backslash A$, is the probability distribution of the coordinates in the set:

$$f(X_A) := \int_{X_B} f(X_A, X_B)\, dX_B$$

Multivariate Gaussian Distribution

In the case of the Gaussian distribution, the marginal distributions (of any coordinate subset size) is also a Gaussian distribution

Multivariate Gaussian Distribution

In the case of the Gaussian distribution, the marginal distributions (of any coordinate subset size) is also a Gaussian distribution

Claim: Let $X \sim \mathcal{N}(\mu, \Sigma)$ be a bivariate Gaussian, i.e $X = (X_1, X_2)^\top$, with a diagonal covariance matrix:

$$X \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right)$$

Multivariate Gaussian Distribution

In the case of the Gaussian distribution, the marginal distributions (of any coordinate subset size) is also a Gaussian distribution

Claim: Let $X \sim \mathcal{N}(\mu, \Sigma)$ be a bivariate Gaussian, i.e $X = (X_1, X_2)^\top$, with a diagonal covariance matrix:

$$X \sim \mathcal{N}\left(\left[\begin{array}{c} \mu_1 \\ \mu_2 \end{array}\right], \left[\begin{array}{cc} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{array}\right]\right)$$

The marginal distribution of coordinate i is

$$f(X_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2\right)$$

Estimators of Multivariate Gaussian Distribution

Returning to the task of describing human weight and height, we can now model each observation as a *vector* $\mathbf{x} \in \mathbb{R}^d$, being the realization of a *random vector* $X := (X_{weight}, X_{height})^\top$.

- Therefore a sample is $S := \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, where each $\mathbf{x}_i$ is the realization of X_i.

Estimators of Multivariate Gaussian Distribution

Returning to the task of describing human weight and height, we can now model each observation as a *vector* $\mathbf{x} \in \mathbb{R}^d$, being the realization of a *random vector* $X := (X_{weight}, X_{height})^\top$.

- Therefore a sample is $S := \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, where each $\mathbf{x}_i$ is the realization of X_i.
- The j'th property (coordinate/feature/variate) of sample $\mathbf{x}_i$ is $\mathbf{x}_i(j)$ or $\mathbf{x}_{ij}$.

Estimators of Multivariate Gaussian Distribution

Returning to the task of describing human weight and height, we can now model each observation as a *vector* $\mathbf{x} \in \mathbb{R}^d$, being the realization of a *random vector* $X := (X_{weight}, X_{height})^\top$.

- Therefore a sample is $S := \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, where each $\mathbf{x}_i$ is the realization of X_i.
- The j'th property (coordinate/feature/variate) of sample $\mathbf{x}_i$ is $\mathbf{x}_i(j)$ or $\mathbf{x}_{ij}$.
- The j'th property $\mathbf{x}_i(j)$ is the realization of the j'th coordinate of X_i: $X_i^{(j)}$.

Estimators of Multivariate Gaussian Distribution

Now, given a sample of m i.i.d observations we wish to estimate the distribution parameters

$$\mathbf{x}_1, \ldots, \mathbf{x}_m \overset{i.i.d}{\sim} \mathcal{N}\left(\begin{bmatrix} 75 \\ 170 \end{bmatrix}, \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix} \right)$$

- The multivariate *sample mean* estimator is simply the univariate estimator of each variable:

$$\hat{\mu} := \begin{bmatrix} \vdots \\ \hat{\mu}_j \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \frac{1}{m} \sum_i x_{i,j} \\ \vdots \end{bmatrix}$$

Estimators of Multivariate Gaussian Distribution

- To define an estimator for the covariance matrix we first define an estimator of the sample covariance between two random variables. The *unbised* sample covariance estimator between X_i, X_j is given by

$$\hat{\sigma}\left(X_i, X_j\right) := \frac{1}{m-1}\sum_k \left(x_{ki} - \hat{\mu}_i\right)\left(x_{kj} - \hat{\mu}_j\right)$$

- Then, the *sample covariance matrix* estimator is a $d \times d$ matrix $\hat{\Sigma}$ such that

$$\hat{\Sigma}_{ij} := \hat{\sigma}\left(X_i, X_j\right), \quad i, j \in [d]$$

Estimators of Multivariate Gaussian Distribution

In matrix notation, for $\mathbf{X} \in \mathbb{R}^{m \times d}$ whose rows are the samples $\mathbf{x}_1, \ldots, \mathbf{x}_m$ the:

- biased sample covariance matrix is given by

$$\hat{\Sigma} := \frac{1}{m} \sum_{i=1}^{m} \left(\mathbf{x}_i - \hat{\mu}\right) \left(\mathbf{x}_i - \hat{\mu}\right)^\top = \frac{1}{m} \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}$$

for $\widetilde{\mathbf{X}}$ being the centered matrix: $\widetilde{\mathbf{X}}_{\cdot,i} := \mathbf{X}_{\cdot,i} - \hat{\mu}$.

Estimators of Multivariate Gaussian Distribution

In matrix notation, for $\mathbf{X} \in \mathbb{R}^{m \times d}$ whose rows are the samples $\mathbf{x}_1, \ldots, \mathbf{x}_m$ the:

- biased sample covariance matrix is given by

$$\hat{\Sigma} := \frac{1}{m} \sum_{i=1}^{m} \left(\mathbf{x}_i - \hat{\mu}\right) \left(\mathbf{x}_i - \hat{\mu}\right)^\top = \frac{1}{m} \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}$$

  for $\widetilde{\mathbf{X}}$ being the centered matrix: $\widetilde{\mathbf{X}}_{\cdot,i} := \mathbf{X}_{\cdot,i} - \hat{\mu}$.

- unbiased sample covariance matrix is given by

$$\hat{\Sigma} := \frac{1}{m-1} \sum_{i=1}^{m} \left(\mathbf{x}_i - \hat{\mu}\right) \left(\mathbf{x}_i - \hat{\mu}\right)^\top = \frac{1}{m-1} \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}$$

1. Estimators
2. Estimator Properties
3. Maximum Likelihood
4. Estimation Quality - Measure of Concentration
5. Multivariate Distributions
6. Summary

Summary

- We have defined the basic terminology of estimation theory: **observation, sample, i.i.d, decision rule, hypothesis class, estimator**

Summary

- We have defined the basic terminology of estimation theory: **observation, sample, i.i.d, decision rule, hypothesis class, estimator**
- We have laid out the general scheme of estimation/learning:

Summary

- We have defined the basic terminology of estimation theory: **observation, sample, i.i.d, decision rule, hypothesis class, estimator**
- We have laid out the general scheme of estimation/learning:
  - Given a sample $x_1, \ldots, x_m$, drawn i.i.d according to $\mathcal{P}(\boldsymbol{\theta})$, where we **do not know $\boldsymbol{\theta}$**, we wish to choose $\boldsymbol{\theta}^*$ that "best" fits the true $\boldsymbol{\theta}$

Summary

- We have defined the basic terminology of estimation theory: **observation, sample, i.i.d, decision rule, hypothesis class, estimator**
- We have laid out the general scheme of estimation/learning:
  - Given a sample $x_1, \dots, x_m$, drawn i.i.d according to $\mathcal{P}(\boldsymbol{\theta})$, where we **do not know** $\boldsymbol{\theta}$, we wish to choose $\boldsymbol{\theta}^*$ that "best" fits the true $\boldsymbol{\theta}$
- We have defined properties of estimators: **bias** and **variance**

Summary

- We have defined the basic terminology of estimation theory: **observation, sample, i.i.d, decision rule, hypothesis class, estimator**
- We have laid out the general scheme of estimation/learning:
    - Given a sample $x_1, \ldots, x_m$, drawn i.i.d according to $\mathcal{P}(\boldsymbol{\theta})$, where we **do not know** $\boldsymbol{\theta}$, we wish to choose $\boldsymbol{\theta}^*$ that "best" fits the true $\boldsymbol{\theta}$
- We have defined properties of estimators: **bias** and **variance**
- We have seen our first learning principle: **Maximum Likelihood**

Summary

- We have defined the basic terminology of estimation theory: **observation, sample, i.i.d, decision rule, hypothesis class, estimator**
- We have laid out the general scheme of estimation/learning:
  - Given a sample $x_1, \ldots, x_m$, drawn i.i.d according to $\mathcal{P}(\boldsymbol{\theta})$, where we **do not know $\boldsymbol{\theta}$**, we wish to choose $\boldsymbol{\theta}^*$ that "best" fits the true $\boldsymbol{\theta}$
- We have defined properties of estimators: **bias** and **variance**
- We have seen our first learning principle: **Maximum Likelihood**
- We have seen ways to measure the quality of our estimation using Markov and Chebyshev (and Hoeffding) bounds

Summary

- We have defined the basic terminology of estimation theory: **observation, sample, i.i.d, decision rule, hypothesis class, estimator**
- We have laid out the general scheme of estimation/learning:
    - Given a sample $x_1, \ldots, x_m$, drawn i.i.d according to $\mathcal{P}(\boldsymbol{\theta})$, where we **do not know** $\boldsymbol{\theta}$, we wish to choose $\boldsymbol{\theta}^*$ that "best" fits the true $\boldsymbol{\theta}$
- We have defined properties of estimators: **bias** and **variance**
- We have seen our first learning principle: **Maximum Likelihood**
- We have seen ways to measure the quality of our estimation using Markov and Chebyshev (and Hoeffding) bounds
- We have seen how to **estimate univariate and multivariate Gaussian** distributions