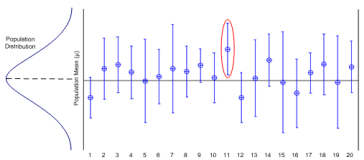
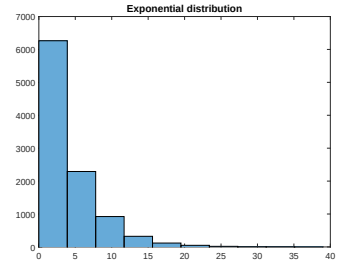
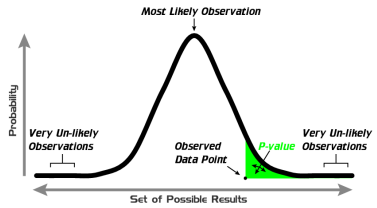
Machine Learning

Linear Algebra and Statistics Recap



Francesco Trovò





Outline

Outline

Practical Information

- Course materials (slides, pdfs, links to recorded lectures) uploaded to Beep
- For questions or clarifications, ask after (or before) the lecture or ask by email for an appointment for a call

Suggested literature to fill the gaps:

- Bishop, C.M., “Pattern recognition and machine learning”, 2006, Springer
- Montgomery, D.C., Runger, G.C., “Applied statistics and probability for engineers”, 2010, John Wiley & Sons

Used Software: Python

- We will use *Colab* as repository for the code
- We will use some libraries which are commonly used in ML projects:
  - `numpy`
  - `scipy`
  - `pandas`
  - `stats`

Outline

Outline

Problem Definition

Consider:

- a dataset composed of a set of N inputs $\mathbf{x}_i := (x_{i1}, \ldots, x_{iD})$, with $x_{ij} \in \mathbb{R}$, each of which has dimension D
- a target $t_i \in \mathbb{R}$ for each input $\mathbf{x}_i$

We want to compute a prediction of the target t_i by considering a linear combination of the input $\mathbf{x}_i$, i.e., generate a vector of parameter $\mathbf{w} := (w_1, \ldots, w_D)^\top$ minimizing some loss function

- If the loss function considers the summation of the squared prediction errors, we are using Least Square minimization

Loss Function

Consider the following loss function:

$$L(\mathbf{w}) := \frac{1}{2}\sum_{i=1}^{N}\left(t_i - \sum_{j=1}^{D} x_{ij}w_j\right)^2$$

Defining $X = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1D} \\ \vdots & & \vdots \\ x_{N1} & \cdots & x_{ND} \end{pmatrix}$, we can equivalently write the loss as:

$$L(\mathbf{w}) := \frac{1}{2}||\mathbf{t} - X\mathbf{w}||_2^2 = \frac{1}{2}(\mathbf{t} - X\mathbf{w})^\top(\mathbf{t} - X\mathbf{w})$$

Loss Function

Consider the following loss function:

$$L(\mathbf{w}) := \frac{1}{2}\sum_{i=1}^{N}\left(t_i - \sum_{j=1}^{D} x_{ij}w_j\right)^2$$

Defining $X = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1D} \\ \vdots & & \vdots \\ x_{N1} & \cdots & x_{ND} \end{pmatrix}$, we can equivalently write the loss as:

$$L(\mathbf{w}) := \frac{1}{2}||\mathbf{t} - X\mathbf{w}||_2^2 = \frac{1}{2}(\mathbf{t} - X\mathbf{w})^\top(\mathbf{t} - X\mathbf{w})$$

Minimize the Loss

If we want to minimize the loss we can compute its derivatives w.r.t. each component of $\mathbf{w}$:

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} := \left(\frac{\partial L(\mathbf{w})}{\partial w_1}, \dots, \frac{\partial L(\mathbf{w})}{\partial w_D} \right)$$

Even in this case we have two different formulation for the derivative

Derivatives

Traditional (elementwise) derivative:

$$\left(\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}}\right)_h = \frac{\partial L(\mathbf{w})}{\partial w_h} = \frac{\partial}{\partial w_h}\left[\frac{1}{2}\sum_{i=1}^{N}\left(t_i - \sum_{j=1}^{D} x_{ij} w_j\right)^2\right]$$

Matrix derivative:

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}\left[\frac{1}{2}(\mathbf{t} - X\mathbf{w})^\top(\mathbf{t} - X\mathbf{w})\right]$$

Outline

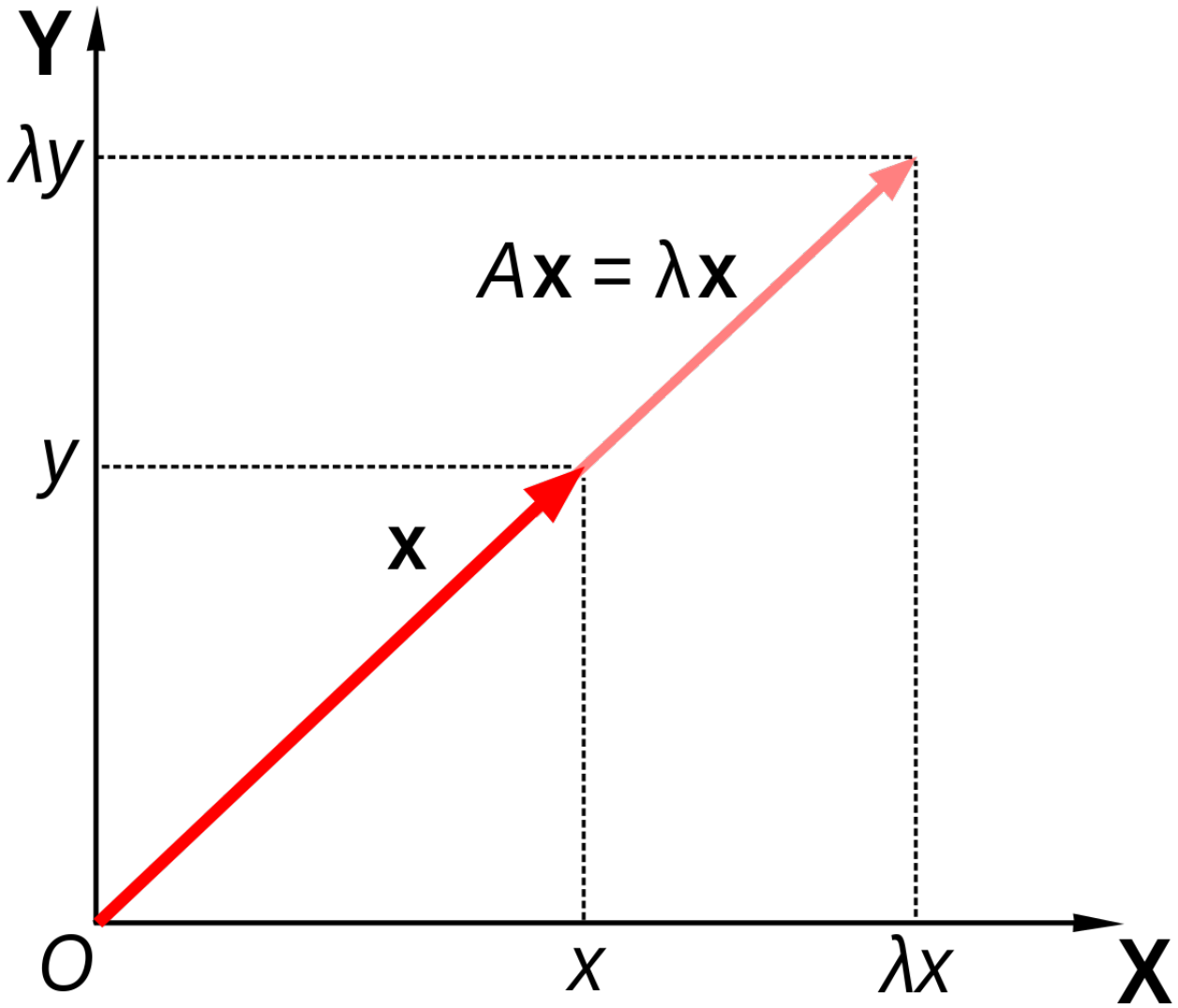
Eigenvector Equations

Given a a square matrix $A \in \mathbb{R}^{N \times N}$, the corresponding eigenvectors equations are:

$$A\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

- *eigenvectors* $\mathbf{v}_1, \ldots, \mathbf{v}_N$, which is the direction along which the vectors are not deviated by the transformation A
- *eigenvalues* $\lambda_1, \ldots, \lambda_N$, which is the stretch factor for the directions $\mathbf{v}_i$

Graphical Representation



Eigenvector Problem

Using the matricial form we can write for a generic pair $(\lambda, \mathbf{v})$:

$$A\mathbf{v} = \lambda\mathbf{v}$$
$$(A - \lambda I_N)\mathbf{v} = 0$$

The previous problem has a non-null solution only if the rank of the matrix $A - \lambda I_N$ is full, or, equivalently:

$$|A - \lambda I_N| = 0$$

"Eigen" Properties

- The rank of A is equal to the number of non-zero eigenvalues
- The determinant of A is equal to the product of its eigenvalues

$$|A| = \prod_{i=1}^{N} \lambda_i$$

- The trace of A is equal to the sum of its eigenvalues:

$$Tr(A) = \sum_{i=1}^{N} \lambda_i$$

Two Useful Definitions

- a matrix A is said to be *positive definite*, if $\mathbf{x}^\top A\mathbf{x} > 0$ for all possible vectors $\mathbf{x} \in \mathbb{R}^N \setminus \{\mathbf{0}\}$
- a matrix A is said to be *semi-positive definite*, if $\mathbf{x}^\top A\mathbf{x} \geq 0$ for all possible vectors $\mathbf{x} \in \mathbb{R}^N \setminus \{\mathbf{0}\}$

Easier way of checking these properties:

- A positive definite matrix has all positive eigenvalues, i.e., $\lambda_i > 0 \; \forall i$
- A semi-positive definite matrix has all non-negative eigenvalues, i.e., $\lambda_i \geq 0 \; \forall i$

Two Useful Definitions

- a matrix A is said to be *positive definite*, if $\mathbf{x}^\top A\mathbf{x} > 0$ for all possible vectors $\mathbf{x} \in \mathbb{R}^N \setminus \{\mathbf{0}\}$
- a matrix A is said to be *semi-positive definite*, if $\mathbf{x}^\top A\mathbf{x} \geq 0$ for all possible vectors $\mathbf{x} \in \mathbb{R}^N \setminus \{\mathbf{0}\}$

Easier way of checking these properties:

- A positive definite matrix has all positive eigenvalues, i.e., $\lambda_i > 0 \; \forall i$
- A semi-positive definite matrix has all non-negative eigenvalues, i.e., $\lambda_i \geq 0 \; \forall i$

Outline

Outline

Discrete Random Variables

- A discrete random variable X is a variable with values in a discrete set E whose value is determined by a stochastic phenomenon
- We define a *probability function* $\mathbb{P} : E \to [0, 1]$ which tells you how often an event in E occurs, i.e.,

$$\mathbb{P}(X = i) = \frac{|i|}{|E|},$$

For instance, in a 20-faced dice:

- $E = \{1, \ldots, 20\}$
- $\mathbb{P}(X = i) = \frac{1}{20}$

Discrete Random Variables

- A discrete random variable X is a variable with values in a discrete set E whose value is determined by a stochastic phenomenon
- We define a *probability function* $\mathbb{P} : E \to [0, 1]$ which tells you how often an event in E occurs, i.e.,

$$\mathbb{P}(X = i) = \frac{|i|}{|E|},$$

For instance, in a 20-faced dice:

- $E = \{1, \ldots, 20\}$
- $\mathbb{P}(X = i) = \frac{1}{20}$

Probability Function Properties

A properly defined probability function should satisfy the following properties:

- $\forall i \in E, \;\; 0 \leq \mathbb{P}(X = i) \leq 1$
- $\sum_{i \in E} \mathbb{P}(X = i) = 1$

Cumulative Function

Assuming that the events are ordered, a function defining the probability of multiple events $F : E \to [0, 1]$, formally:

$$F(i) := \mathbb{P}(X \leq i) = \sum_{h=1}^{i} \mathbb{P}(X = h) = \sum_{h \in E, h \leq i} \frac{|h|}{|E|}$$

For instance, in a 20-faced dice:

$$F(i) = \sum_{h=1}^{i} \frac{1}{20} = \frac{i}{20}$$

Cumulative Function

Assuming that the events are ordered, a function defining the probability of multiple events $F : E \to [0, 1]$, formally:

$$F(i) := \mathbb{P}(X \leq i) = \sum_{h=1}^{i} \mathbb{P}(X = h) = \sum_{h \in E, h \leq i} \frac{|h|}{|E|}$$

For instance, in a 20-faced dice:

$$F(i) = \sum_{h=1}^{i} \frac{1}{20} = \frac{i}{20}$$

Cumulative Function Properties

A properly defined cumulative function should satisfy the following properties:

- $0 \leq F(i) \leq 1$
- $F(i) = 0, \ \forall i < \min_{h \in E} h$
- $F(i) = 1, \ \forall i \geq \max_{h \in E} h$

Characterizing the Distribution

Quantities which characterize a random variable the *expected value* and the *variance* (first two moments):

$$\mathbb{E}[X] := \sum_{i \in E} i \, P(X = i)$$

$$Var(X) := \sum_{i \in E} (\mathbb{E}[X] - i)^2 P(X = i)$$

For instance, in a 20-faced dice:

$$\mathbb{E}[X] := \sum_{i=1}^{20} \frac{i}{20} = \frac{1}{20} \frac{20(20+1)}{2} = \frac{21}{2}$$

$$Var(X) := \sum_{i=1}^{20} \frac{\left(\frac{21}{2} - i\right)^2}{20} = \frac{57}{4}$$

Sometimes we compute the *standard deviation* instead of the variance:

$$std(X) = \sqrt{Var(X)}$$

Characterizing the Distribution

Quantities which characterize a random variable the *expected value* and the *variance* (first two moments):

$$\mathbb{E}[X] := \sum_{i \in E} i \, P(X = i)$$

$$Var(X) := \sum_{i \in E} (\mathbb{E}[X] - i)^2 P(X = i)$$

For instance, in a 20-faced dice:

$$\mathbb{E}[X] := \sum_{i=1}^{20} \frac{i}{20} = \frac{1}{20} \frac{20(20+1)}{2} = \frac{21}{2}$$

$$Var(X) := \sum_{i=1}^{20} \frac{\left(\frac{21}{2} - i\right)^2}{20} = \frac{57}{4}$$

Sometimes we compute the *standard deviation* instead of the variance:

$$std(X) = \sqrt{Var(X)}$$

Characterizing the Distribution

Quantities which characterize a random variable the *expected value* and the *variance* (first two moments):

$$\mathbb{E}[X] := \sum_{i \in E} i \, P(X = i)$$

$$Var(X) := \sum_{i \in E} (\mathbb{E}[X] - i)^2 P(X = i)$$

For instance, in a 20-faced dice:

$$\mathbb{E}[X] := \sum_{i=1}^{20} \frac{i}{20} = \frac{1}{20} \frac{20(20+1)}{2} = \frac{21}{2}$$

$$Var(X) := \sum_{i=1}^{20} \frac{\left(\frac{21}{2} - i\right)^2}{20} = \frac{57}{4}$$

Sometimes we compute the *standard deviation* instead of the variance:

$$std(X) = \sqrt{Var(X)}$$

Continuous Random Variables

Similarly, if the set E is not discrete, i.e., $E \subseteq \mathbb{R}$, we define the *probability density function* (pdf):

$$f(x) := \lim_{\delta x \to 0} \frac{\mathbb{P}(x \leq X \leq x + \delta x)}{\delta x}$$

with properties:

$$f(x) \geq 0 \ \ \forall x \in \Omega$$

$$\int_{x \in \Omega} f(x) dx = 1$$

Continuous Random Variables

The *Cumulative Density Function* (CDF) is defined as:

$$F(x) := \int_{s \in \Omega, s \leq x} f(s)ds$$

with properties:

$$0 \leq F(x) \leq 1 \; \forall x \in \Omega$$

$$F\left(\min_{x \in \Omega} x - \varepsilon\right) = 0 \forall \varepsilon > 0$$

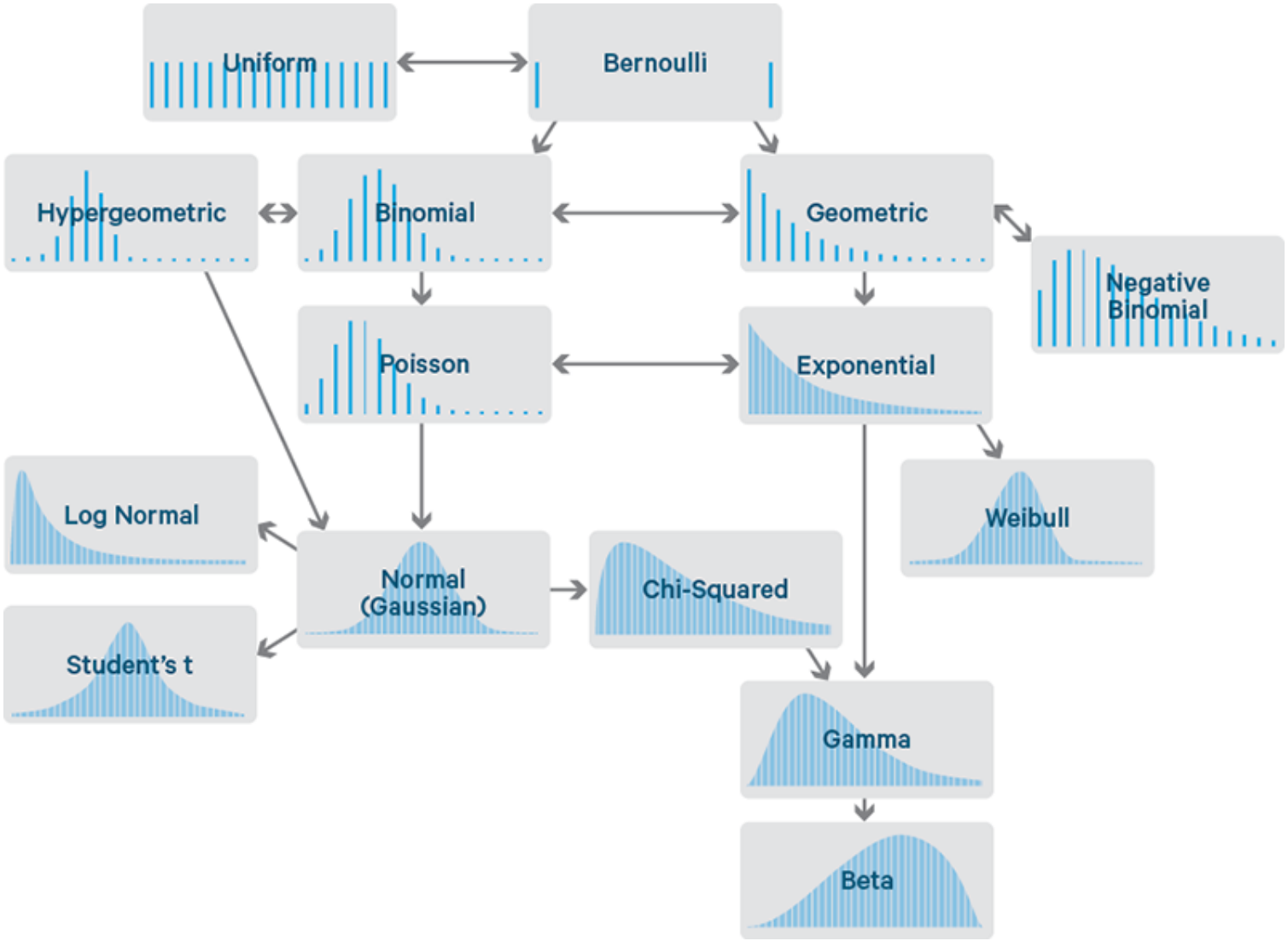
$$F\left(\max_{x \in \Omega} x\right) = 1$$

A quantile of order α is defined as a point $z_\alpha \in \Omega$ s.t.:

$$F(z_\alpha) = 1 - \alpha$$

or a point of the domain leaving to his left a cumulated probability of $1 - \alpha$

Examples of Distributions



Characterizing the Continuous Random Variables

The expected value and the variance for a continuous random variable are defined as follows:

$$\mu = \mathbb{E}[X] := \int_{x \in \Omega} x f(x) dx;$$

$$\sigma^2 = Var(X) := \int_{x \in \Omega} (\mathbb{E}[X] - x)^2 f(x) dx.$$

For the Gaussian case:

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$F(x; \mu, \sigma) = \int_{-\infty}^{x} f(t; \mu, \sigma) dt$$

Characterizing the Continuous Random Variables

The expected value and the variance for a continuous random variable are defined as follows:

$$\mu = \mathbb{E}[X] := \int_{x \in \Omega} x f(x) dx;$$
$$\sigma^2 = Var(X) := \int_{x \in \Omega} (\mathbb{E}[X] - x)^2 f(x) dx.$$

For the Gaussian case:

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$$F(x; \mu, \sigma) = \int_{-\infty}^{x} f(t; \mu, \sigma) dt$$

Univariate Distributions in Python

Present in the package `scipy.stats`, among the others:

- `binom` Binomial distribution (discrete, `n` and `p` parameters)
- `norm` Gaussian distribution (continuous, `loc` and `scale` parameters)
- `unif` Uniform distribution (continuous, `loc` and `scale` parameters)
- `beta` Beta distribution (continuous, `a` and `b` parameters)

each of which will have specific parameters

Estimating the Distributions

- Usually we do not have any information about the distributions
- We need to estimate their mean and variance:

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

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

- They are consistent estimator for the expected value and for the variance, respectively

Central Limit Theorem

Let us consider the empirical expected value $\bar{X}$. We recall the *Central Limit Theorem* (CLT):

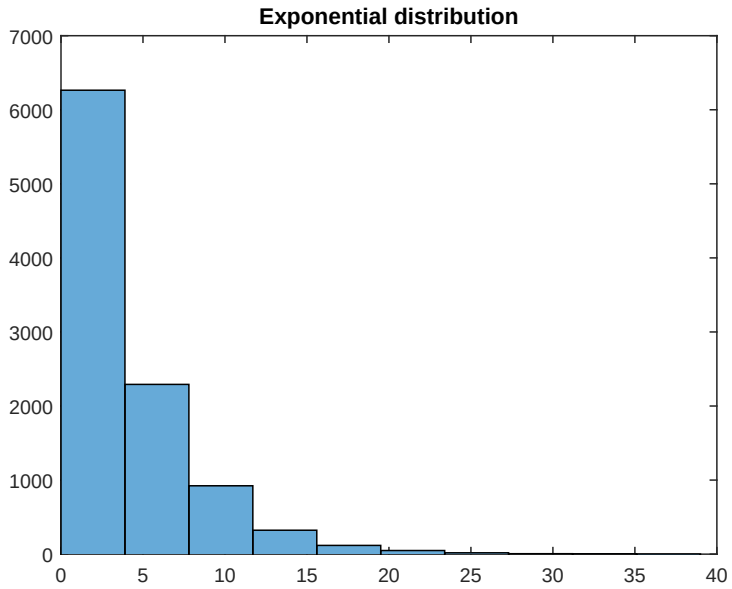
Theorem (Central Limit Theorem)

Assume $\{X_1, \dots X_n\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$, then:

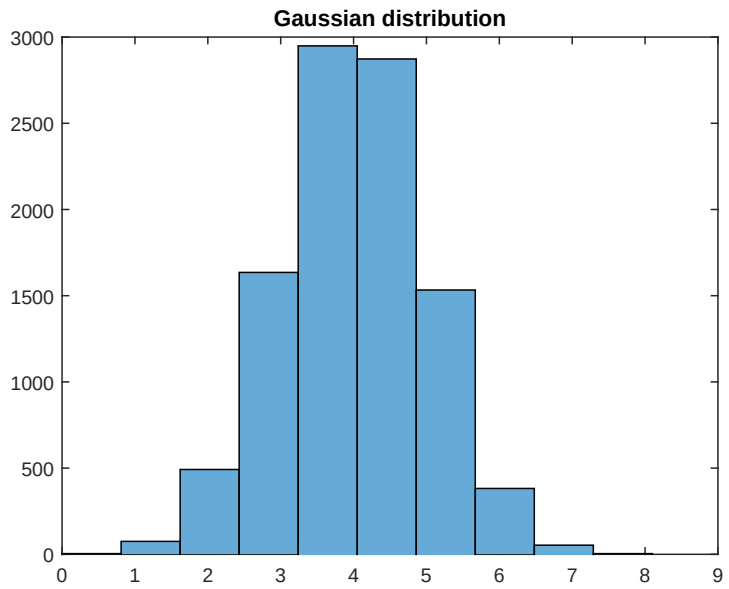
$$\sqrt{n}\left(\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right) \to \mathcal{N}(0; \sigma^2),$$

where the convergence holds in distribution

Empirical Distributions



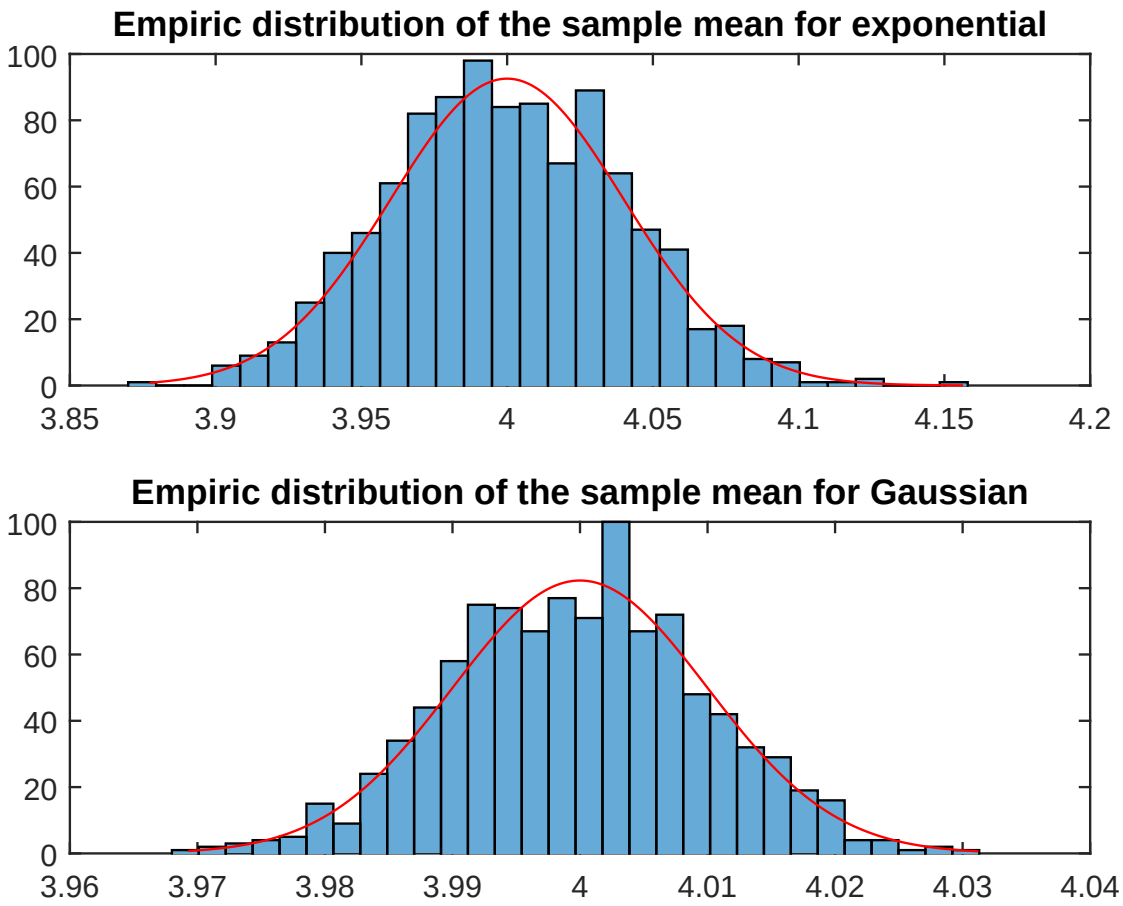
(a) Exponential



(b) Gaussian

10000 samples from an exponential and a Gaussian distribution

Empirical Evidence of the CLT

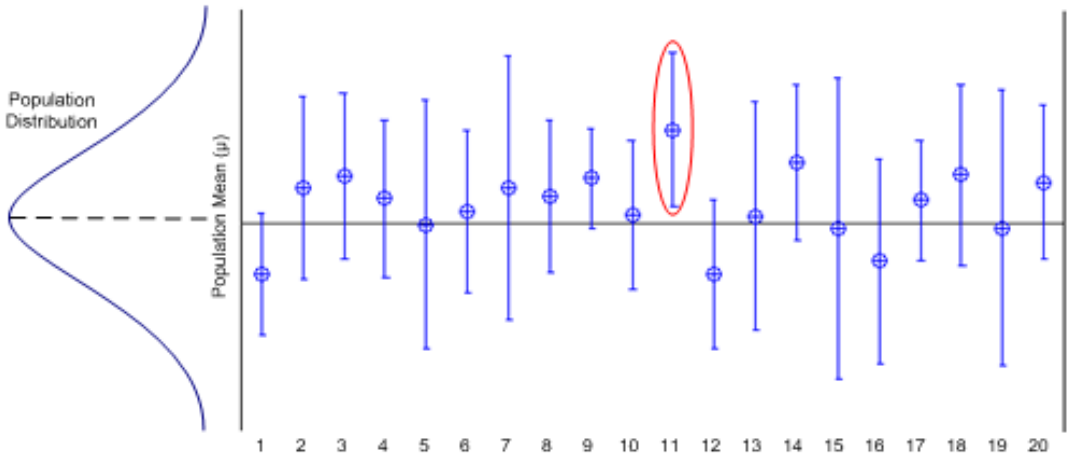


Even if the samples are coming from two distributions we have the same distribution for the expected value

Outline

Confidence Intervals

- We need to set a level identifying if our estimator is "good enough"
- The probability that $\bar{X} = \mathbb{E}[X]$ is zero, since the realization of the expected value is a continuous random variable itself
- We need to built some intervals, where we have high confidence that the true mean $\mathbb{E}[X]$ is in



- A 95%-confidence interval works 95% of the times

Options for the Confidence Intervals

- Gaussian Approximation

$$\bar{X} - \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}}$$

- Chebichev Inequality (requires $\mathbb{E}[X] = \mu < \infty$ and $Var[X] = \sigma^2 < \infty$)

$$\bar{X} - \frac{\sigma}{\sqrt{n}\sqrt{\alpha}} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}\sqrt{\alpha}}$$

- Hoeffding Inequality (finite support)

$$\bar{X} - (b-a)\sqrt{\frac{-\log(\alpha/2)}{2n}} \leq \mu \leq \bar{X} + (b-a)\sqrt{\frac{-\log(\alpha/2)}{2n}}$$

Outline

Hypothesis Testing

Scientific statement: we want to show

- *The estimated parameter $\bar{X}$ is equal to μ*
- *The estimated parameter $\bar{X}'$ is different from another estimated parameter $\bar{X}''$*

Answer provided by statistics: define null H_0 and alternative H_1 hypotheses

$$H_0 \ : \ \mu = \mu_0 \quad vs. \quad H_1 \ : \ \mu \neq \mu_0$$

and use data to provide evidence that either one or the other holds

Hypothesis Testing

Scientific statement: we want to show

- *The estimated parameter $\bar{X}$ is equal to μ*
- *The estimated parameter $\bar{X}'$ is different from another estimated parameter $\bar{X}''$*

Answer provided by statistics: define null H_0 and alternative H_1 hypotheses

$$H_0 \ : \ \mu = \mu_0 \quad vs. \quad H_1 \ : \ \mu \neq \mu_0$$

and use data to provide evidence that either one or the other holds

Basic Gaussian Test

Given the data $\{x_1, \ldots, x_n\}$ we have:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \rightarrow t = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

Fixing a confidence $1 - \alpha$, $\alpha \in (0, 1)$, the test statistic t should be close to the real mean μ with "high" probability, formally:

$$\mathbb{P}(t < z_{\alpha/2} \vee t > z_{1-\alpha/2}) = \alpha$$

		Decision	
		Fail to reject H_0	Reject H_0
True	H_0	Correct	Type I error (α)
	H_1	Type II error	Correct

Basic Gaussian Test

Given the data $\{x_1, \ldots, x_n\}$ we have:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \rightarrow t = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

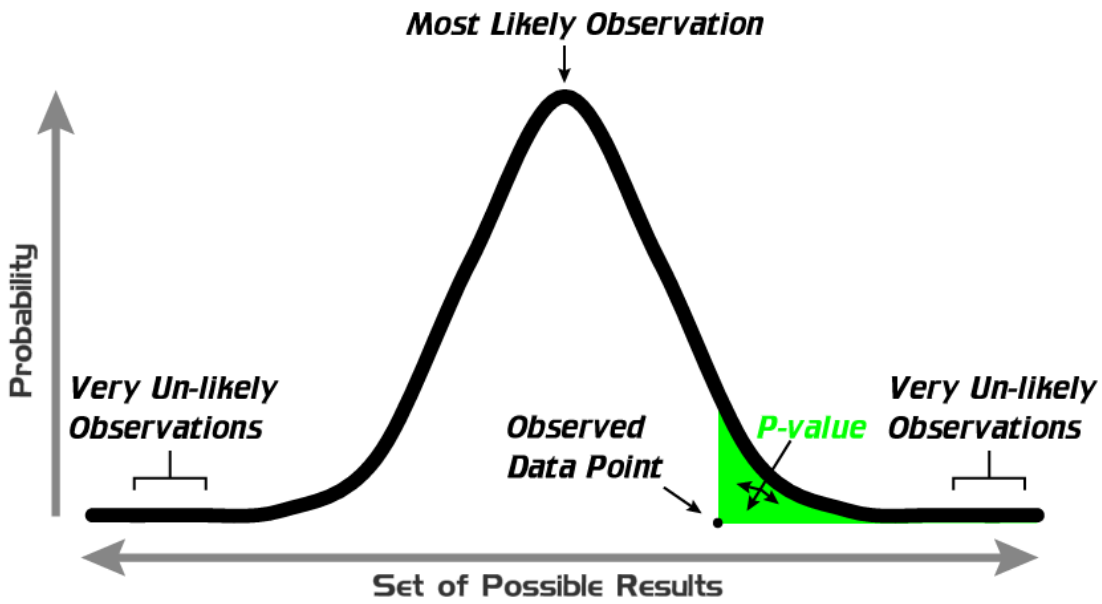
Fixing a confidence $1 - \alpha$, $\alpha \in (0, 1)$, the test statistic t should be close to the real mean μ with "high" probability, formally:

$$\mathbb{P}(t < z_{\alpha/2} \vee t > z_{1-\alpha/2}) = \alpha$$

		Decision	
		Fail to reject H_0	Reject H_0
True	H_0	Correct	Type I error (α)
	H_1	Type II error	Correct

P-Value

To avoid specifying the confidence α, we can let the data tell us how much we might be confident about their correspondence to a specific hypothesis:



- Small p-Values imply that we are confident that the H_1 hypothesis holds
- Large p-Values imply we are not able to tell that H_0 hypothesis does not hold

Outline

Different Model for the Parameters

- The bounds include some of the information about the data
- They do not allow to incorporate information one has about the distribution parameters

New idea: the expected value of the random variable μ is a random variable itself

We use the Bayes formula to update this information:

$$\mathbb{P}(a|b) = \frac{\mathbb{P}(b|a)\mathbb{P}(a)}{\mathbb{P}(b)}$$

Different Model for the Parameters

- The bounds include some of the information about the data
- They do not allow to incorporate information one has about the distribution parameters

New idea: the expected value of the random variable μ is a random variable itself

We use the Bayes formula to update this information:

$$\mathbb{P}(a|b) = \frac{\mathbb{P}(b|a)\mathbb{P}(a)}{\mathbb{P}(b)}$$

Example: Bernoulli Variables

Consider the Bayes formula:

$$\begin{aligned}\mathbb{P}(\mu|x_1,\ldots,x_t) &= \frac{\mathbb{P}(x_1,\ldots,x_{t-1},x_t|\mu)\mathbb{P}(\mu)}{\mathbb{P}(x_1,\ldots,x_t)}\\ &\propto \mathbb{P}(x_t|\mu)\mathbb{P}(x_1,\ldots,x_{t-1}|\mu)\mathbb{P}(\mu)\\ &= \mathbb{P}(x_t|\mu)\mathbb{P}(x_{t-1}|\mu)\mathbb{P}(x_1,\ldots,x_{t-2}|\mu)\mathbb{P}(\mu)\\ &= \mathbb{P}(\mu)\prod_{h=1}^{t}\mathbb{P}(x_h|\mu),\end{aligned}$$

We incorporate incrementally information from a prior distribution $\mathbb{P}(\mu)$

Example: if we have a prior telling us hat it had 3 successes over 10 trials, using a Beta distribution as prior for μ ($\mu \sim \text{Beta}(3,7)$), the posterior is still a Beta (i.e., Bernoulli and Beta are conjugate prior-posterior)

Example: Bernoulli Variables

Consider the Bayes formula:

$$\begin{aligned}\mathbb{P}(\mu|x_1,\ldots,x_t) &= \frac{\mathbb{P}(x_1,\ldots,x_{t-1},x_t|\mu)\mathbb{P}(\mu)}{\mathbb{P}(x_1,\ldots,x_t)}\\ &\propto \mathbb{P}(x_t|\mu)\mathbb{P}(x_1,\ldots,x_{t-1}|\mu)\mathbb{P}(\mu)\\ &= \mathbb{P}(x_t|\mu)\mathbb{P}(x_{t-1}|\mu)\mathbb{P}(x_1,\ldots,x_{t-2}|\mu)\mathbb{P}(\mu)\\ &= \mathbb{P}(\mu)\prod_{h=1}^{t}\mathbb{P}(x_h|\mu),\end{aligned}$$

We incorporate incrementally information from a prior distribution $\mathbb{P}(\mu)$

Example: if we have a prior telling us hat it had 3 successes over 10 trials, using a Beta distribution as prior for μ ($\mu \sim \text{Beta}(3,7)$), the posterior is still a Beta (i.e., Bernoulli and Beta are conjugate prior-posterior)