Statistics for Data Science

MSc Data Science WiSe 2019/20

Prof. Dr. Dirk Ostwald
(5) Expectation, covariance, inequalities, limits
Bibliographic remarks

The material presented in this section is not to be understood as a comprehensive introduction to the respective topics, but merely serves as a collection of results that will be put to use in later sections. The treatment of expectations, variances, and covariances follows Wasserman (2004, Sections 3.1 - 3.3) and DeGroot and Schervish (2012, Sections 4.1 - 4.3, 4.6). The discussion of the Markov and Chebyshev inequalities follows DeGroot and Schervish (2012, Section 6.2). The discussion of Jensen’s inequality, as well as the Weak and Strong Laws of Large Number follow Casella and Berger (2002, Section 4.7 and 5.5), respectively. The discussion of the Central limit theorem follows DeGroot and Schervish (2012, Section 6.3).
Expectation, covariance, inequalities, limits

• Expectation
• Variance
• Covariance
• Inequalities
• Limits
Expectation, covariance, inequalities, limits

- **Expectation**
- **Variance**
- **Covariance**
- **Inequalities**
- **Limits**
Expectation

Definition

The *expected value* or *expectation* of a random variable $X$ is defined as

$$
\mathbb{E}(X) := \int x \, d\mathbb{P}(x) = \begin{cases} 
\sum_{x \in X} x p(x), & \text{if } X \text{ is discrete} \\
\int_{X} x p(x) \, dx, & \text{if } X \text{ is continuous.}
\end{cases}
$$

(1)

We say that $\mathbb{E}(X)$ exists, if and only if $\int |x| \, d\mathbb{P}(x) < \infty$. 

Remarks

• The expectation is a one-number summary of a distribution.

• Intuitively, $\mathbb{E}(X) \approx \frac{1}{n} \sum_{i=1}^{n} X_i$ for a large number $n$ of i.i.d. draws $X_i$.

• $\int x \, d\mathbb{P}(x)$ serves as a summary notation for $\sum_{x \in X} x p(x)$ and $\int_{X} x p(x) \, dx$.

• The formal basis for $\int x \, d\mathbb{P}(x)$ is the Lebesgue integral.
Expectation

Definition

The *expected value* or *expectation* of a random variable $X$ is defined as

$$
\mathbb{E}(X) := \int x \, d\mathbb{P}(x) = \begin{cases} 
\sum_{x \in \mathcal{X}} x \, p(x), & \text{if } X \text{ is discrete} \\
\int_{\mathcal{X}} x \, p(x) \, dx, & \text{if } X \text{ is continuous.}
\end{cases}
$$

(1)

We say that $\mathbb{E}(X)$ exists, if and only if $\int |x| \, d\mathbb{P}(x) < \infty$.

Remarks

- The expectation is a one-number summary of a distribution.
Expectation

Definition

The *expected value* or *expectation* of a random variable $X$ is defined as

$$
\mathbb{E}(X) := \int x \, d\mathbb{P}(x) = \begin{cases} 
\sum_{x \in \chi} x \, p(x), & \text{if } X \text{ is discrete} \\
\int_{\chi} x \, p(x) \, dx, & \text{if } X \text{ is continuous.}
\end{cases}
$$

(1)

We say that $\mathbb{E}(X)$ exists, if and only if $\int |x| \, d\mathbb{P}(x) < \infty$.

Remarks

- The expectation is a one-number summary of a distribution.
- Intuitively, $\mathbb{E}(X) \approx \frac{1}{n} \sum_{i=1}^{n} X_i$ for a large number $n$ of i.i.d. draws $X_i$. 
Definition

The expected value or expectation of a random variable $X$ is defined as

$$
\mathbb{E}(X) := \int x \, d\mathbb{P}(x) = \begin{cases} 
\sum_{x \in \mathcal{X}} x \, p(x), & \text{if } X \text{ is discrete} \\
\int_{\mathcal{X}} x \, p(x) \, dx, & \text{if } X \text{ is continuous.}
\end{cases}
$$

(1)

We say that $\mathbb{E}(X)$ exists, if and only if $\int |x| \, d\mathbb{P}(x) < \infty$.

Remarks

- The expectation is a one-number summary of a distribution.
- Intuitively, $\mathbb{E}(X) \approx \frac{1}{n} \sum_{i=1}^{n} X_i$ for a large number $n$ of i.i.d. draws $X_i$.
- $\int x \, d\mathbb{P}(x)$ serves as a summary notation for $\sum_{x \in \mathcal{X}} x \, p(x)$ and $\int_{\mathcal{X}} x \, p(x) \, dx$. 

Definition

The *expected value* or *expectation* of a random variable $X$ is defined as

$$
\mathbb{E}(X) := \int x \, d\mathbb{P}(x) = \begin{cases} 
\sum_{x \in \mathcal{X}} x \, p(x), & \text{if } X \text{ is discrete} \\
\int_{\mathcal{X}} x \, p(x) \, dx, & \text{if } X \text{ is continuous.}
\end{cases}
$$

We say that $\mathbb{E}(X)$ exists, if and only if $\int |x| \, d\mathbb{P}(x) < \infty$.

Remarks

- The expectation is a one-number summary of a distribution.
- Intuitively, $\mathbb{E}(X) \approx \frac{1}{n} \sum_{i=1}^{n} X_i$ for a large number $n$ of i.i.d. draws $X_i$.
- $\int x \, d\mathbb{P}(x)$ serves as a summary notation for $\sum_{x \in \mathcal{X}} x \, p(x)$ and $\int_{\mathcal{X}} x \, p(x) \, dx$.
- The formal basis for $\int x \, d\mathbb{P}(x)$ is the *Lebesgue integral*. 
Example (Bernoulli variable expectation)

Let $X \sim \text{Bern} (\mu)$. Then

$$\mathbb{E}(X) = \mu. \quad (2)$$
Expectation

Example (Bernoulli variable expectation)

Let $X \sim \text{Bern}(\mu)$. Then

$$\mathbb{E}(X) = \mu.$$  \hspace{1cm} (2)

Proof

$X$ is discrete with $\mathcal{X} = \{0, 1\}$. Thus,

$$\mathbb{E}(X) = \sum_{x \in \{0, 1\}} x \text{ Bern}(x; \mu)$$

$$= 0 \cdot \mu^0 (1 - \mu)^{1-0} + 1 \cdot \mu^1 (1 - \mu)^{1-1}$$

$$= 0 \cdot \mu^0 (1 - \mu)^1 + 1 \cdot \mu^1 (1 - \mu)^0$$

$$= \mu.$$

\hspace{1cm} (3)
**Expectation**

**Example (Gaussian variable expectation)**

Let $X \sim N(\mu, \sigma^2)$. Then

$$\mathbb{E}(X) = \mu. \quad (4)$$

**Proof**

We assume that

$$\int_{\mathbb{R}} x \exp \left(-x^2\right) \, dx = 0 \quad \text{and} \quad \int_{\mathbb{R}} \exp \left(-x^2\right) \, dx = \sqrt{\pi} \quad (5)$$

are known. We have

$$\mathbb{E}(X) = \int_{\mathbb{R}} x \frac{1}{\sigma \sqrt{2\pi}} \exp \left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \, dx. \quad (6)$$

Setting $\xi := (x-\mu)/\sqrt{2}\sigma$ and thus $x = \sqrt{2}\sigma\xi + \mu$ then yields

$$\mathbb{E}(X) = \frac{\sqrt{2}\sigma}{\sigma \sqrt{2\pi}} \int_{\mathbb{R}} \left(\sqrt{2}\sigma\xi + \mu\right) \exp \left(-\xi^2\right) \, d\xi$$

$$= \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\sigma \int_{\mathbb{R}} \xi \exp \left(-\xi^2\right) \, d\xi + \mu \int_{\mathbb{R}} \exp \left(-\xi^2\right) \, d\xi\right)$$

$$= \frac{1}{\sqrt{\pi}} \left(0 + \mu \sqrt{\pi}\right)$$

$$= \frac{\mu \sqrt{\pi}}{\sqrt{\pi}}$$

$$= \mu. \quad (7)$$

$\square$
Theorem (Properties of expectations)

- **Linearity.** If $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants, then

$$
\mathbb{E}\left(\sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i \mathbb{E}(X_i). \tag{8}
$$
Theorem (Properties of expectations)

- **Linearity.** If $X_1, ..., X_n$ are random variables and $a_1, ..., a_n$ are constants, then
  \[ E \left( \sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i E(X_i). \] (8)

- **Factorization under independence.** If $X_1, ..., X_n$ are independent random variables, then
  \[ E \left( \prod_{i=1}^{n} X_i \right) = \prod_{i=1}^{n} E(X_i). \] (9)

Remarks

- These properties are helpful when evaluating expectations.
- Proofs of these properties are discussed in the ®ubung.
Theorem (Properties of expectations)

- **Linearity.** If $X_1, ..., X_n$ are random variables and $a_1, ..., a_n$ are constants, then
  \[
  \mathbb{E} \left( \sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i \mathbb{E}(X_i). \quad (8)
  \]

- **Factorization under independence.** If $X_1, ..., X_n$ are independent random variables, then
  \[
  \mathbb{E} \left( \prod_{i=1}^{n} X_i \right) = \prod_{i=1}^{n} \mathbb{E}(X_i). \quad (9)
  \]
### Theorem (Properties of expectations)

- **Linearity.** If $X_1, ..., X_n$ are random variables and $a_1, ..., a_n$ are constants, then
  \[
  \mathbb{E} \left( \sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i \mathbb{E}(X_i). \tag{8}
  \]

- **Factorization under independence.** If $X_1, ..., X_n$ are independent random variables, then
  \[
  \mathbb{E} \left( \prod_{i=1}^{n} X_i \right) = \prod_{i=1}^{n} \mathbb{E}(X_i). \tag{9}
  \]

### Remarks

- These properties are helpful when evaluating expectations.
Theorem (Properties of expectations)

- **Linearity.** If $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants, then
  \[
  \mathbb{E} \left( \sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i \mathbb{E}(X_i). \tag{8}
  \]

- **Factorization under independence.** If $X_1, \ldots, X_n$ are independent random variables, then
  \[
  \mathbb{E} \left( \prod_{i=1}^{n} X_i \right) = \prod_{i=1}^{n} \mathbb{E}(X_i). \tag{9}
  \]

Remarks

- These properties are helpful when evaluating expectations.
- Proofs of these properties are discussed in the Übung.
Expectation, covariance, inequalities, limits

• Expectation
• Variance
• Covariance
• Inequalities
• Limits
Definition (Variance and standard deviation)

Let \( X \) be a random variable with expectation \( \mathbb{E}(X) \). The variance of \( X \) is defined by

\[
\mathbb{V}(X) := \mathbb{E}\left( (X - \mathbb{E}(X))^2 \right),
\]

assuming that this expectation exists.

Remarks
• The variance measures the spread of a distribution.
• The square is necessitated by \( \mathbb{E}(X - \mathbb{E}(X)) = \mathbb{E}(X) - \mathbb{E}(X) = 0 \).
• An alternative measure of distribution spread is \( \mathbb{E}(\left| X - \mathbb{E}(X) \right|) \).
• Another alternative measure is the entropy of a distribution.
Definition (Variance and standard deviation)

Let $X$ be a random variable with expectation $\mathbb{E}(X)$. The variance of $X$ is defined by

$$\mathbb{V}(X) := \mathbb{E}\left((X - \mathbb{E}(X))^2\right),$$

assuming that this expectation exists. The standard deviation is defined as

$$\mathbb{S}(X) := \sqrt{\mathbb{V}(X)}.$$
Definition (Variance and standard deviation)

Let $X$ be a random variable with expectation $\mathbb{E}(X)$. The variance of $X$ is defined by

$$\mathbb{V}(X) := \mathbb{E} ((X - \mathbb{E}(X))^2), \quad (10)$$

assuming that this expectation exists. The standard deviation is defined as

$$\mathbb{S}(X) := \sqrt{\mathbb{V}(X)}. \quad (11)$$

Remarks

- The variance measures the spread of a distribution.
Definition (Variance and standard deviation)

Let $X$ be a random variable with expectation $\mathbb{E}(X)$. The variance of $X$ is defined by

$$\mathbb{V}(X) := \mathbb{E}\left((X - \mathbb{E}(X))^2\right),$$  \hspace{1cm} (10)

assuming that this expectation exists. The standard deviation is defined as

$$\mathbb{S}(X) := \sqrt{\mathbb{V}(X)}.$$  \hspace{1cm} (11)

Remarks

• The variance measures the spread of a distribution.

• The square is necessitated by $\mathbb{E}(X - \mathbb{E}(X)) = \mathbb{E}(X) - \mathbb{E}(X) = 0$. 
Variance

Definition (Variance and standard deviation)

Let $X$ be a random variable with expectation $\mathbb{E}(X)$. The variance of $X$ is defined by

$$\mathbb{V}(X) := \mathbb{E}((X - \mathbb{E}(X))^2),$$

(10)

assuming that this expectation exists. The standard deviation is defined as

$$\mathbb{S}(X) := \sqrt{\mathbb{V}(X)}.$$  

(11)

Remarks

- The variance measures the spread of a distribution.
- The square is necessitated by $\mathbb{E}(X - \mathbb{E}(X)) = \mathbb{E}(X) - \mathbb{E}(X) = 0$.
- An alternative measure of distribution spread is $\mathbb{E}(|X - \mathbb{E}(X)|)$.
Variance

Definition (Variance and standard deviation)

Let $X$ be a random variable with expectation $\mathbb{E}(X)$. The variance of $X$ is defined by

$$\mathbb{V}(X) := \mathbb{E} \left( (X - \mathbb{E}(X))^2 \right),$$

(10)

assuming that this expectation exists. The standard deviation is defined as

$$\mathbb{S}(X) := \sqrt{\mathbb{V}(X)}.$$  

(11)

Remarks

- The variance measures the spread of a distribution.
- The square is necessitated by $\mathbb{E}(X - \mathbb{E}(X)) = \mathbb{E}(X) - \mathbb{E}(X) = 0$.
- An alternative measure of distribution spread is $\mathbb{E}(|X - \mathbb{E}(X)|)$.
- Another alternative measure is the entropy of a distribution.
Variance

Example (Bernoulli variable variance)

Let $X \sim \text{Bern}(\mu)$. Then the variance of $X$ is

$$\mathbb{V}(X) = \mu(1 - \mu).$$  \hspace{1cm} (12)
Variance

Example (Bernoulli variable variance)

Let $X \sim \text{Bern}(\mu)$. Then the variance of $X$ is

$$\text{V}(X) = \mu(1 - \mu).$$ (12)

Proof

$X$ is discrete and we have $\mathbb{E}(X) = \mu$. Thus

$$\text{V}(X) = \mathbb{E}((X - \mu)^2)$$

$$= \sum_{x \in \{0, 1\}} (x - \mu)^2 \text{Bern}(x; \mu)$$

$$= (0 - \mu)^2 \mu^0 (1 - \mu)^{1-0} + (1 - \mu)^2 \mu^1 (1 - \mu)^{1-1}$$

$$= \mu^2 (1 - \mu) + (1 - \mu)^2 \mu$$

$$= (\mu^2 + (1 - \mu)\mu) (1 - \mu)$$

$$= (\mu^2 + \mu - \mu^2) (1 - \mu)$$

$$= \mu (1 - \mu).$$ (13)
Example (Gaussian variable variance)

Let $X \sim N(\mu, \sigma^2)$. Then

$$ \text{Var}(X) = \sigma^2. \quad (14) $$

Proof

Theorem (Variance translation theorem)

It holds that

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$  \hspace{1cm} (15)
Theorem (Variance translation theorem)

It holds that

$$\text{V}(X) = \mathbb{E} (X^2) - \mathbb{E}(X)^2.$$ (15)

Proof

With the definition of the variance and the linearity of expectations, we have

$$\text{V}(X) = \mathbb{E} \left( (X - \mathbb{E}(X))^2 \right)$$

$$= \mathbb{E} \left( X^2 - 2X\mathbb{E}(X) + \mathbb{E}(X)^2 \right)$$

$$= \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + \mathbb{E}(X)^2$$

$$= \mathbb{E}(X^2) - 2\mathbb{E}(X)^2 + \mathbb{E}(X)^2$$

$$= \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$ (16)

□
Variance

Theorem (Variance translation theorem)

It holds that

\[ \forall(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2. \]  \(15\)

Proof

With the definition of the variance and the linearity of expectations, we have

\[
\forall(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right)
= \mathbb{E}\left(X^2 - 2X\mathbb{E}(X) + \mathbb{E}(X)^2\right)
= \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + \mathbb{E}(X)^2
= \mathbb{E}(X^2) - 2\mathbb{E}(X)^2 + \mathbb{E}(X)^2
= \mathbb{E}(X^2) - \mathbb{E}(X)^2.
\]  \(16\)

Remark

- The theorem is useful, if computing \(\mathbb{E}(X^2)\) and \(\mathbb{E}(X)\) is easy.
Variance

Theorem (Variance properties)

- Variance of linear-affine functions. If $X$ is a random variable and $a$ and $b$ are constants, then

$$\text{Var}(aX + b) = a^2 \text{Var}(X).$$  \hfill (17)
Variance

Theorem (Variance properties)

• *Variance of linear-affine functions.* If $X$ is a random variable and $a$ and $b$ are constants, then

$$\text{V}(aX + b) = a^2 \text{V}(X).$$

(17)

• *Variance of linear combinations.* If $X_1, \ldots, X_n$ are independent and $a_1, \ldots, a_n$ are constants, then

$$\text{V} \left( \sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i^2 \text{V}(X_i).$$

(18)
**Theorem (Variance properties)**

- **Variance of linear-affine functions.** If $X$ is a random variable and $a$ and $b$ are constants, then
  \[
  \text{Var}(aX + b) = a^2 \text{Var}(X). \tag{17}
  \]

- **Variance of linear combinations.** If $X_1, \ldots, X_n$ are independent and $a_1, \ldots, a_n$ are constants, then
  \[
  \text{Var} \left( \sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i). \tag{18}
  \]
### Theorem (Variance properties)

- **Variance of linear-affine functions.** If $X$ is a random variable and $a$ and $b$ are constants, then
  \[
  \text{V}(aX + b) = a^2 \text{V}(X).
  \] (17)

- **Variance of linear combinations.** If $X_1, ..., X_n$ are independent and $a_1, ..., a_n$ are constants, then
  \[
  \text{V} \left( \sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i^2 \text{V}(X_i).
  \] (18)

### Remark

- Proofs of these properties are discussed in the Übung.
Expectation, covariance, inequalities, limits

- Expectation
- Variance
- Covariance
- Inequalities
- Limits
Definition (Covariance and correlation)

The covariance of two random variables $X$ and $Y$ with finite expectations is defined as

$$C(X, Y) = \mathbb{E}((X - \mathbb{E}(X)) (Y - \mathbb{E}(Y))).$$  \hspace{1cm} (19)

if this expectation exists.

Remarks

• The covariance of $X$ with itself is the variance of $X$.

• Correlation is a normalized measure of stochastic dependence.

• $\rho(X,Y) \in [-1, 1]$.

• Correlation measures the degree of linear dependence.
Covariance

Definition (Covariance and correlation)

The covariance of two random variables $X$ and $Y$ with finite expectations is defined as

$$\mathbb{C}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))).$$  \hspace{1cm} (19)

if this expectation exists. The correlation of two random variables $X$ and $Y$ with finite expectations is defined as

$$\rho(X, Y) = \frac{\mathbb{C}(X, Y)}{\sqrt{\mathbb{V}(X)}\sqrt{\mathbb{V}(Y)}}.$$  \hspace{1cm} (20)

Remarks

- The covariance of $X$ with itself is the variance of $X$. 


Covariance

Definition (Covariance and correlation)

The covariance of two random variables $X$ and $Y$ with finite expectations is defined as

$$ C(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))). \quad (19) $$

if this expectation exists. The correlation of two random variables $X$ and $Y$ with finite expectations is defined as

$$ \rho(X, Y) = \frac{C(X, Y)}{\sqrt{\text{V}(X)} \sqrt{\text{V}(Y)}}. \quad (20) $$

Remarks

- The covariance of $X$ with itself is the variance of $X$.
- Correlation is a normalized measure of stochastic dependence.
Definition (Covariance and correlation)

The covariance of two random variables $X$ and $Y$ with finite expectations is defined as

$$ C(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))). $$

(19)

if this expectation exists. The correlation of two random variables $X$ and $Y$ with finite expectations is defined as

$$ \rho(X, Y) = \frac{C(X, Y)}{\sqrt{\text{V}(X)}\sqrt{\text{V}(Y)}}. $$

(20)

Remarks

- The covariance of $X$ with itself is the variance of $X$.
- Correlation is a normalized measure of stochastic dependence.
- $\rho(X, Y) \in [-1, 1]$. 
Definition (Covariance and correlation)

The covariance of two random variables $X$ and $Y$ with finite expectations is defined as

$$C(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))).$$  \hspace{1cm} (19)

if this expectation exists. The correlation of two random variables $X$ and $Y$ with finite expectations is defined as

$$\rho(X, Y) = \frac{C(X, Y)}{\sqrt{\mathbb{V}(X)} \sqrt{\mathbb{V}(Y)}}.$$  \hspace{1cm} (20)

Remarks

- The covariance of $X$ with itself is the variance of $X$.
- Correlation is a normalized measure of stochastic dependence.
- $\rho(X, Y) \in [-1, 1]$.
- Correlation measures the degree of linear dependence.
## Theorem (Covariance translation theorem)

It holds that

\[ C(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \]  

(21)

### Remarks

- The theorem is useful if computing \( \mathbb{E}(XY) \), \( \mathbb{E}(X) \), and \( \mathbb{E}(Y) \) is easy.
- For \( Y = X \), we recover \( \mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \).
- For independent \( X, Y \) we have \( \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \) and hence \( C(X, Y) = 0 \).
Covariance

Theorem (Covariance translation theorem)

It holds that

\[ C(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \]  \hspace{1cm} (21)

Proof

With the definition of the covariance, we have

\[ C(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \]
\[ = \mathbb{E}(XY - X\mathbb{E}(Y) + \mathbb{E}(X)Y - \mathbb{E}(X)\mathbb{E}(Y)) \]
\[ = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) \]
\[ = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \] \hspace{1cm} (22)

\[ \square \]
Theorem (Covariance translation theorem)

It holds that

\[ C(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \]  

(21)

Proof

With the definition of the covariance, we have

\[
C(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \\
= \mathbb{E}(XY - X\mathbb{E}(Y) + \mathbb{E}(X)Y - \mathbb{E}(X)\mathbb{E}(Y)) \\
= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) \\
= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).
\]  

(22)

Remarks

- The theorem is useful, if computing \( \mathbb{E}(XY), \mathbb{E}(X) \) and \( \mathbb{E}(Y) \) is easy.
### Theorem (Covariance translation theorem)

It holds that

\[
C(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).
\]  

(21)

### Proof

With the definition of the covariance, we have

\[
C(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))
\]

\[
= \mathbb{E}(XY - X\mathbb{E}(Y) + \mathbb{E}(X)Y - \mathbb{E}(X)\mathbb{E}(Y))
\]

\[
= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y)
\]

\[
= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).
\]  

(22)

### Remarks

- The theorem is useful, if computing \(\mathbb{E}(XY), \mathbb{E}(X)\) and \(\mathbb{E}(Y)\) is easy.

- For \(Y = X\), we recover \(\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2\).
Covariance

Theorem (Covariance translation theorem)

It holds that

$$C(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$  \hfill (21)

Proof

With the definition of the covariance, we have

$$C(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$$

$$= \mathbb{E}(XY) - X\mathbb{E}(Y) + \mathbb{E}(X)Y - \mathbb{E}(X)\mathbb{E}(Y))$$

$$= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y)$$

$$= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$  \hfill (22)

Remarks

• The theorem is useful, if computing $\mathbb{E}(XY), \mathbb{E}(X)$ and $\mathbb{E}(Y)$ is easy.

• For $Y = X$, we recover $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.

• For independent $X, Y$ we have $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ and hence $C(X, Y) = 0$. 
Theorem (Variances of sums of random variables)

For two random variables $X$ and $Y$ it holds that

$$V(X + Y) = V(X) + V(Y) + 2C(X, Y) \tag{23}$$

and

$$V(X - Y) = V(X) + V(Y) - 2C(X, Y) \tag{24}$$
**Theorem (Variances of sums of random variables)**

For two random variables $X$ and $Y$ it holds that

$$\text{\Var}(X + Y) = \text{\Var}(X) + \text{\Var}(Y) + 2\text{\Cov}(X, Y)$$

(23)

and

$$\text{\Var}(X - Y) = \text{\Var}(X) + \text{\Var}(Y) - 2\text{\Cov}(X, Y)$$

(24)

Moreover, for $n$ random variables $X_1, ..., X_n$ it holds that

$$\text{\Var} \left( \sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i^2 \text{\Var}(X_i) + 2 \sum_{j=1}^{n} \sum_{i \leq j} \text{\Cov}(X_i, X_j).$$

(25)
Covariance

**Theorem (Variances of sums of random variables)**

For two random variables $X$ and $Y$ it holds that

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$  \hspace{1cm} (23)

and

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$$  \hspace{1cm} (24)

Moreover, for $n$ random variables $X_1, \ldots, X_n$ it holds that

$$\text{Var} \left( \sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + 2 \sum_{j=1}^{n} \sum_{i \leq j} \text{Cov}(X_i, X_j).$$  \hspace{1cm} (25)

**Remark**

- Proofs of these properties are discussed in the Übung.
Expectation, covariance, inequalities, limits

• Expectation
• Variance
• Covariance
• Inequalities
• Limits
Theorem (Markov inequality)

Let $X$ denote a random variable with $\mathbb{P}(X \geq 0) = 1$. Then for all $x \in \mathbb{R}$ it holds that

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}(X)}{x}.$$  \hspace{1cm} (26)
## Inequalities

### Theorem (Markov inequality)

Let $X$ denote a random variable with $\mathbb{P}(X \geq 0) = 1$. Then for all $x \in \mathbb{R}$ it holds that

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}(X)}{x}. \quad (26)$$

### Proof

We consider the case of a continuous $X$ with PDF $p$. We first note that

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} \xi p(\xi) \, d\xi = \int_{-\infty}^{x} \xi p(\xi) \, d\xi + \int_{x}^{\infty} \xi p(\xi) \, d\xi. \quad (27)$$

With $\mathbb{P}(X \geq 0) = 1$ it then follows that

$$\mathbb{E}(X) \geq \int_{x}^{\infty} \xi p(\xi) \, d\xi \geq \int_{x}^{\infty} x p(\xi) \, d\xi = x \int_{x}^{\infty} p(\xi) \, d\xi = x \mathbb{P}(X \geq x). \quad (28)$$

Hence

$$\mathbb{E}(X) \geq x \mathbb{P}(X \geq x) \Leftrightarrow \mathbb{P}(X \geq x) \leq \frac{\mathbb{E}(X)}{x}. \quad (29)$$

□
Theorem (Markov inequality)

Let $X$ denote a random variable with $P(X \geq 0) = 1$. Then for all $x \in \mathbb{R}$ it holds that

$$P(X \geq x) \leq \frac{E(X)}{x}. \quad (26)$$

Proof

We consider the case of a continuous $X$ with PDF $p$. We first note that

$$E(X) = \int_{-\infty}^{\infty} \xi p(\xi) \, d\xi = \int_{-\infty}^{x} \xi p(\xi) \, d\xi + \int_{x}^{\infty} \xi p(\xi) \, d\xi. \quad (27)$$

With $P(X \geq 0) = 1$ it then follows that

$$E(X) \geq \int_{x}^{\infty} \xi p(\xi) \, d\xi \geq \int_{x}^{\infty} x \, p(\xi) \, d\xi = x \int_{x}^{\infty} p(\xi) \, d\xi = x \, P(X \geq x). \quad (28)$$

Hence

$$E(X) \geq x \, P(X \geq x) \Leftrightarrow P(X \geq x) \leq \frac{E(X)}{x}. \quad (29)$$

Remarks

- The Markov inequality relates exceedance probabilities and expectations.
Inequalities

**Theorem (Markov inequality)**

Let $X$ denote a random variable with $\mathbb{P}(X \geq 0) = 1$. Then for all $x \in \mathbb{R}$ it holds that

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}(X)}{x}. \quad (26)$$

**Proof**

We consider the case of a continuous $X$ with PDF $p$. We first note that

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} \xi p(\xi) \, d\xi = \int_{-\infty}^{x} \xi p(\xi) \, d\xi + \int_{x}^{\infty} \xi p(\xi) \, d\xi. \quad (27)$$

With $\mathbb{P}(X \geq 0) = 1$ it then follows that

$$\mathbb{E}(X) \geq \int_{x}^{\infty} \xi p(\xi) \, d\xi \geq \int_{x}^{\infty} x p(\xi) \, d\xi = x \int_{x}^{\infty} p(\xi) \, d\xi = x \mathbb{P}(X \geq x). \quad (28)$$

Hence

$$\mathbb{E}(X) \geq x \mathbb{P}(X \geq x) \Leftrightarrow \mathbb{P}(X \geq x) \leq \frac{\mathbb{E}(X)}{x}. \quad (29)$$

**Remarks**

- The Markov inequality relates exceedance probabilities and expectations.
- For example, if $\mathbb{E}(X) = 1$, then $\mathbb{P}(X \geq 100) \leq 0.01$. 

Theorem (Chebychev inequality)

Let $X$ denote a random variable with variance $\mathbb{V}(X)$. Then for all $x \in \mathbb{R}$

$$
P(|X - \mathbb{E}(X)| \geq x) \leq \frac{\mathbb{V}(X)}{x^2}.
$$

(30)

Proof

We first note that for $a, b \in \mathbb{R}$, it holds that

$$
a^2 \geq b^2 \iff |a| \geq b
$$

(31)

Next, set $Y := X - \mathbb{E}(X)$. Then with the Markov inequality

$$
P\left(Y \geq x^2\right) \leq \frac{\mathbb{E}(Y)}{x^2} \iff P\left(\left(X - \mathbb{E}(X)\right)^2 \geq x^2\right) \leq \frac{\mathbb{E}\left((X - \mathbb{E}(X))^2\right)}{x^2} \iff P(|X - \mathbb{E}(X)| \geq x) \leq \frac{\mathbb{V}(X)}{x^2}.
$$

□
Theorem (Chebychev inequality)

Let $X$ denote a random variable with variance $\mathbb{V}(X)$. Then for all $x \in \mathbb{R}$

$$
\mathbb{P}(\left| X - \mathbb{E}(X) \right| \geq x) \leq \frac{\mathbb{V}(X)}{x^2}.
$$

(30)

Proof

We first note that for $a, b \in \mathbb{R}$, it holds that

$$a^2 \geq b^2 \iff |a| \geq b
$$

(31)

Next, set $Y := X - \mathbb{E}(X)$. Then with the Markov inequality

$$
\mathbb{P}(Y \geq x^2) \leq \frac{\mathbb{E}(Y)}{x^2} \iff \mathbb{P}((X - \mathbb{E}(X))^2 \geq x^2) \leq \frac{\mathbb{E}((X - \mathbb{E}(X))^2)}{x^2} \iff \mathbb{P}(|X - \mathbb{E}(X)| \geq x) \leq \frac{\mathbb{V}(X)}{x^2}.
$$

Remarks

- The Chebychev inequality relates deviations from expectations to variances.
Theorem (Chebychev inequality)

Let $X$ denote a random variable with variance $\text{V}(X)$. Then for all $x \in \mathbb{R}$

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq x) \leq \frac{\text{V}(X)}{x^2}. \quad (30)$$

Proof

We first note that for $a, b \in \mathbb{R}$, it holds that

$$a^2 \geq b^2 \iff |a| \geq b \quad (31)$$

Next, set $Y := X - \mathbb{E}(X)$. Then with the Markov inequality

$$\mathbb{P}(Y \geq x^2) \leq \frac{\mathbb{E}(Y)}{x^2} \iff \mathbb{P}((X - \mathbb{E}(X))^2 \geq x^2) \leq \frac{\mathbb{E}((X - \mathbb{E}(X))^2)}{x^2} \iff \mathbb{P}(|X - \mathbb{E}(X)| \geq x) \leq \frac{\text{V}(X)}{x^2}. \quad \square$$

Remarks

- The Chebychev inequality relates deviations from expectations to variances.
- Note that

$$\mathbb{P}\left(|X - \mathbb{E}(X)| \geq 3\sqrt{\text{V}(X)}\right) \leq \frac{\text{V}(X)}{(3\sqrt{\text{V}(X)})^2} = \frac{1}{9}.$$
Inequalities

**Theorem (Jensen’s inequality)**

Let $X$ be a random variable and $g$ be a convex function, i.e.

$$g(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda g(x_1) + (1 - \lambda)g(x_2).$$  \hspace{1cm} (32)

Then

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}(X)).$$  \hspace{1cm} (33)

Conversely, let $g$ be a concave function, i.e.

$$g(\lambda x_1 + (1 - \lambda)x_2) \geq \lambda g(x_1) + (1 - \lambda)g(x_2).$$  \hspace{1cm} (34)

Then

$$\mathbb{E}(g(X)) \leq g(\mathbb{E}(X)).$$  \hspace{1cm} (35)

**Remarks**

- For convex $g$ the function’s graph lies below the straight line $g(x_1)$ to $g(x_2)$. 
Inequalities

Theorem (Jensen’s inequality)

Let $X$ be a random variable and $g$ be a convex function, i.e.

$$g(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda g(x_1) + (1 - \lambda)g(x_2).$$  \hfill (32)

Then

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}(X)).$$  \hfill (33)

Conversely, let $g$ be a concave function, i.e.

$$g(\lambda x_1 + (1 - \lambda)x_2) \geq \lambda g(x_1) + (1 - \lambda)g(x_2).$$  \hfill (34)

Then

$$\mathbb{E}(g(X)) \leq g(\mathbb{E}(X)).$$  \hfill (35)

Remarks

- For convex $g$ the function’s graph lies below the straight line $g(x_1)$ to $g(x_2)$.
- For concave $g$ the function’s graph lies above the straight line $g(x_1)$ to $g(x_2)$. 

## Inequalities

**Theorem (Jensen’s inequality)**

Let $X$ be a random variable and $g$ be a convex function, i.e.

$$g(\lambda x_1 + (1 - \lambda) x_2) \leq \lambda g(x_1) + (1 - \lambda) g(x_2).$$  \hfill (32)

Then

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}(X)).$$  \hfill (33)

Conversely, let $g$ be a concave function, i.e.

$$g(\lambda x_1 + (1 - \lambda) x_2) \geq \lambda g(x_1) + (1 - \lambda) g(x_2).$$  \hfill (34)

Then

$$\mathbb{E}(g(X)) \leq g(\mathbb{E}(X)).$$  \hfill (35)

**Remarks**

- For convex $g$ the function’s graph lies below the straight line $g(x_1)$ to $g(x_2)$.
- For concave $g$ the function’s graph lies above the straight line $g(x_1)$ to $g(x_2)$.
- The logarithm is a concave function, hence $\mathbb{E}(\ln X) \leq \ln \mathbb{E}(X)$. 
Inequalities

Proof

By adapting the proof of Casella and Berger (2002, Theorem 4.7.8), we show the inequality for the concave case. Let $f$ be a tangent line at the point $g(\mathbb{E}(X))$, i.e. is a linear-affine function of the form $f(X) := aX + b$ for some $a, b \in \mathbb{R}$ with $f(\mathbb{E}(X)) = g(\mathbb{E}(X))$. Because $g$ is concave, we have $g(x) \leq ax + b$ for all $x \in \mathbb{R}$ and thus also $g(X) \leq aX + b$. Hence,

$$\mathbb{E}(g(X)) \leq \mathbb{E}(aX + b) = a\mathbb{E}(X) + b = f(\mathbb{E}(X)) = g(\mathbb{E}(X)).$$  \hspace{1cm} (36)
Theorem (Cauchy-Schwarz inequality)

Let $X, Y$ be two random variables such that $\mathbb{E}(X, Y)$ exists. Then

$$\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2) \mathbb{E}(Y^2)$$  \hspace{1cm} (37)

Remark

- This is $|\langle x, y \rangle|^2 \leq \langle x, x \rangle \cdot \langle y, y \rangle$ for random variables.
Proof
We consider the case that $0 < \mathbb{E}(X^2) < \infty$ and $0 < \mathbb{E}(Y^2) < \infty$. Then for all $a, b \in \mathbb{R}$, it holds that

$$0 \leq \mathbb{E}((aX + bY)^2) \quad \text{and} \quad 0 \leq \mathbb{E}((aX - bY)^2).$$

(38)
Proof
We consider the case that $0 < \mathbb{E}(X^2) < \infty$ and $0 < \mathbb{E}(Y^2) < \infty$. Then for all $a, b \in \mathbb{R}$, it holds that

$$0 \leq \mathbb{E}((aX + bY)^2) \quad \text{and} \quad 0 \leq \mathbb{E}((aX - bY)^2).$$

(38)

Hence,

$$0 \leq a^2\mathbb{E}(X^2) + b^2\mathbb{E}(Y^2) + 2ab\mathbb{E}(XY) \quad \text{and} \quad 0 \leq a^2\mathbb{E}(X^2) + b^2\mathbb{E}(Y^2) - 2ab\mathbb{E}(XY).$$

(39)
Proof
We consider the case that $0 < \mathbb{E}(X^2) < \infty$ and $0 < \mathbb{E}(Y^2) < \infty$. Then for all $a, b \in \mathbb{R}$, it holds that

$$0 \leq \mathbb{E} \left( (aX + bY)^2 \right) \quad \text{and} \quad 0 \leq \mathbb{E} \left( (aX - bY)^2 \right). \quad (38)$$

Hence,

$$0 \leq a^2 \mathbb{E}(X^2) + b^2 \mathbb{E}(Y^2) + 2ab \mathbb{E}(XY) \quad \text{and} \quad 0 \leq a^2 \mathbb{E}(X^2) + b^2 \mathbb{E}(Y^2) - 2ab \mathbb{E}(XY). \quad (39)$$

Setting $a := \sqrt{\mathbb{E}(Y^2)}$ and $b := \sqrt{\mathbb{E}(X^2)}$ then yields
Proof
We consider the case that $0 < \mathbb{E}(X^2) < \infty$ and $0 < \mathbb{E}(Y^2) < \infty$. Then for all $a, b \in \mathbb{R}$, it holds that

$$0 \leq \mathbb{E}\left((aX + bY)^2\right) \quad \text{and} \quad 0 \leq \mathbb{E}\left((aX - bY)^2\right). \quad (38)$$

Hence,

$$0 \leq a^2\mathbb{E}(X^2) + b^2\mathbb{E}(Y^2) + 2ab\mathbb{E}(XY) \quad \text{and} \quad 0 \leq a^2\mathbb{E}(X^2) + b^2\mathbb{E}(Y^2) - 2ab\mathbb{E}(XY). \quad (39)$$

Setting $a := \sqrt{\mathbb{E}(Y^2)}$ and $b := \sqrt{\mathbb{E}(X^2)}$ then yields

$$0 \leq \mathbb{E}(Y^2)\mathbb{E}(X^2) + \mathbb{E}(X^2)\mathbb{E}(Y^2) + 2\sqrt{\mathbb{E}(Y^2)}\sqrt{\mathbb{E}(X^2)}\mathbb{E}(XY)$$

$$\Leftrightarrow -2\sqrt{\mathbb{E}(Y^2)}\sqrt{\mathbb{E}(X^2)}\mathbb{E}(XY) \leq 2\mathbb{E}(X^2)\mathbb{E}(Y^2)$$

$$\Leftrightarrow -\sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)\mathbb{E}(XY)} \leq \sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)\mathbb{E}(X^2)}$$

$$\Leftrightarrow -\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)}, \quad (40)$$

Together, the above imply the Cauchy-Schwarz inequality. See DeGroot and Schervish (2012, Theorem 4.2.6) for a full proof.
Proof
We consider the case that \(0 < \mathbb{E}(X^2) < \infty\) and \(0 < \mathbb{E}(Y^2) < \infty\). Then for all \(a, b \in \mathbb{R}\), it holds that

\[
0 \leq \mathbb{E}\left( (aX + bY)^2 \right) \quad \text{and} \quad 0 \leq \mathbb{E}\left( (aX - bY)^2 \right). \tag{38}
\]

Hence,

\[
0 \leq a^2\mathbb{E}(X^2) + b^2\mathbb{E}(Y^2) + 2ab\mathbb{E}(XY) \quad \text{and} \quad 0 \leq a^2\mathbb{E}(X^2) + b^2\mathbb{E}(Y^2) - 2ab\mathbb{E}(XY). \tag{39}
\]

Setting \(a := \sqrt{\mathbb{E}(Y^2)}\) and \(b := \sqrt{\mathbb{E}(X^2)}\) then yields

\[
0 \leq \mathbb{E}(Y^2)\mathbb{E}(X^2) + \mathbb{E}(X^2)\mathbb{E}(Y^2) + 2\sqrt{\mathbb{E}(Y^2)}\sqrt{\mathbb{E}(X^2)}\mathbb{E}(XY)
\]

\[
\Leftrightarrow -2\sqrt{\mathbb{E}(Y^2)}\sqrt{\mathbb{E}(X^2)}\mathbb{E}(XY) \leq 2\mathbb{E}(X^2)\mathbb{E}(Y^2) \tag{40}
\]

\[
\Leftrightarrow -\sqrt{\mathbb{E}(Y^2)}\mathbb{E}(X^2)\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(Y^2)}\mathbb{E}(X^2)\sqrt{\mathbb{E}(Y^2)}\mathbb{E}(X^2)
\]

\[
\Leftrightarrow -\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(Y^2)}\mathbb{E}(X^2),
\]

and similarly

\[
0 \leq \mathbb{E}(Y^2)\mathbb{E}(X^2) + \mathbb{E}(X^2)\mathbb{E}(Y^2) - 2\sqrt{\mathbb{E}(Y^2)}\sqrt{\mathbb{E}(X^2)}\mathbb{E}(XY)
\]

\[
\Leftrightarrow 2\sqrt{\mathbb{E}(Y^2)}\sqrt{\mathbb{E}(X^2)}\mathbb{E}(XY) \leq 2\mathbb{E}(X^2)\mathbb{E}(Y^2) \tag{41}
\]

\[
\Leftrightarrow \sqrt{\mathbb{E}(Y^2)}\mathbb{E}(X^2)\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(Y^2)}\mathbb{E}(X^2)\sqrt{\mathbb{E}(Y^2)}\mathbb{E}(X^2)
\]

\[
\Leftrightarrow \mathbb{E}(XY) \leq \sqrt{\mathbb{E}(Y^2)}\mathbb{E}(X^2).
\]
Proof
We consider the case that \(0 < \mathbb{E}(X^2) < \infty\) and \(0 < \mathbb{E}(Y^2) < \infty\). Then for all \(a, b \in \mathbb{R}\), it holds that
\[
0 \leq \mathbb{E}\left((aX + bY)^2\right) \quad \text{and} \quad 0 \leq \mathbb{E}\left((aX - bY)^2\right).
\]
(38)

Hence,
\[
0 \leq a^2\mathbb{E}(X^2) + b^2\mathbb{E}(Y^2) + 2ab\mathbb{E}(XY) \quad \text{and} \quad 0 \leq a^2\mathbb{E}(X^2) + b^2\mathbb{E}(Y^2) - 2ab\mathbb{E}(XY).
\]
(39)

Setting \(a := \sqrt{\mathbb{E}(Y^2)}\) and \(b := \sqrt{\mathbb{E}(X^2)}\) then yields
\[
0 \leq \mathbb{E}(Y^2)\mathbb{E}(X^2) + \mathbb{E}(X^2)\mathbb{E}(Y^2) + 2\sqrt{\mathbb{E}(Y^2)}\sqrt{\mathbb{E}(X^2)}\mathbb{E}(XY)
\]
\[
\Leftrightarrow -2\sqrt{\mathbb{E}(Y^2)}\sqrt{\mathbb{E}(X^2)}\mathbb{E}(XY) \leq 2\mathbb{E}(X^2)\mathbb{E}(Y^2)
\]
\[
\Leftrightarrow -\sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)\mathbb{E}(XY)} \leq \sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)\mathbb{E}(XY)}
\]
\[
\Leftrightarrow -\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)},
\]
(40)

and similarly
\[
0 \leq \mathbb{E}(Y^2)\mathbb{E}(X^2) + \mathbb{E}(X^2)\mathbb{E}(Y^2) - 2\sqrt{\mathbb{E}(Y^2)}\sqrt{\mathbb{E}(X^2)}\mathbb{E}(XY)
\]
\[
\Leftrightarrow 2\sqrt{\mathbb{E}(Y^2)}\sqrt{\mathbb{E}(X^2)}\mathbb{E}(XY) \leq 2\mathbb{E}(X^2)\mathbb{E}(Y^2)
\]
\[
\Leftrightarrow \sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)\mathbb{E}(XY)} \leq \sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)\mathbb{E}(XY)}
\]
\[
\Leftrightarrow \mathbb{E}(XY) \leq \sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)}.
\]
(41)

Together, the above imply the Cauchy-Schwarz inequality. See DeGroot and Schervish (2012, Theorem 4.2.6) for a full proof.
### Theorem (Correlation inequality)

Let $X$ and $Y$ denote two random variables with $\mathbb{V}(X), \mathbb{V}(Y) > 0$. Then

$$\rho(X, Y)^2 = \frac{\mathbb{C}(X, Y)^2}{\mathbb{V}(X)\mathbb{V}(Y)} \leq 1.$$  \hfill (42)
Inequalities

Theorem (Correlation inequality)

Let $X$ and $Y$ denote two random variables with $\mathbb{V}(X), \mathbb{V}(Y) > 0$. Then

$$\rho(X,Y)^2 = \frac{\mathbb{C}(X,Y)^2}{\mathbb{V}(X)\mathbb{V}(Y)} \leq 1.$$  \hfill (42)

Proof

According to the Cauchy-Schwarz inequality for random variables $U, V$, it holds that

$$\left(\mathbb{E}(UV)\right)^2 \leq \mathbb{E} \left( U^2 \right) \mathbb{E} \left( V^2 \right).$$ \hfill (43)

Set $U := X - \mathbb{E}(X)$ and $V := Y - \mathbb{E}(Y)$. Then according to the Cauchy-Schwarz inequality

$$\mathbb{E} \left( (X - \mathbb{E}(X))(Y - \mathbb{E}(Y))^2 \right) \leq \mathbb{E} \left( (X - \mathbb{E}(X))^2 \right) \mathbb{E} \left( (Y - \mathbb{E}(Y))^2 \right)$$ \hfill (44)

Hence,

$$\mathbb{C}(X,Y)^2 \leq \mathbb{V}(X)\mathbb{V}(Y) \Leftrightarrow \frac{\mathbb{C}(X,Y)^2}{\mathbb{V}(X)\mathbb{V}(Y)} \leq 1.$$ \hfill (45)

\[\square\]
Inequalities

Theorem (Correlation inequality)

Let \( X \) and \( Y \) denote two random variables with \( \mathbb{V}(X), \mathbb{V}(Y) > 0 \). Then

\[
\rho(X, Y)^2 = \frac{\mathbb{C}(X, Y)^2}{\mathbb{V}(X)\mathbb{V}(Y)} \leq 1.
\] (42)

Proof

According to the Cauchy-Schwarz inequality for random variables \( U, V \), it holds that

\[
(\mathbb{E}(UV))^2 \leq \mathbb{E}(U^2)\mathbb{E}(V^2).
\] (43)

Set \( U := X - \mathbb{E}(X) \) and \( V := Y - \mathbb{E}(Y) \). Then according to the Cauchy-Schwarz inequality

\[
\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))^2 \leq \mathbb{E}((X - \mathbb{E}(X))^2)\mathbb{E}((Y - \mathbb{E}(Y))^2)
\] (44)

Hence,

\[
\mathbb{C}(X, Y)^2 \leq \mathbb{V}(X)\mathbb{V}(Y) \iff \frac{\mathbb{C}(X, Y)^2}{\mathbb{V}(X)\mathbb{V}(Y)} \leq 1.
\] (45)

Remark

\( \bullet \) \( \rho(X, Y)^2 \leq 1 \iff |\rho(X, Y)| \leq 1 \), that is \( \rho(X, Y) \in [-1, 1] \).
Expectation, covariance, inequalities, limits

- Expectation
- Variance
- Covariance
- Inequalities
- Limits
Limits

Overview

- Intuitively, both the Weak and the Strong Law of Large Numbers state that for i.i.d. random samples of a distribution, the sample mean approximates the expectation of that distribution for large sample sizes. The “Weak” and the “Strong” form of the Law of Large Numbers differ in the form of random variable convergence considered.
Limits

Overview

• Intuitively, both the Weak and the Strong Law of Large Numbers state that for i.i.d. random samples of a distribution, the sample mean approximates the expectation of that distribution for large sample sizes. The “Weak” and the “Strong” form of the Law of Large Numbers differ in the form of random variable convergence considered.

• Intuitively, the Central Limit Theorem states that the sum of very many independent and arbitrarily distributed random variables is normally distributed. The “Lindenberg and Lévy” form assumes independent and identically distributed random variables and is easier to proof than the “Liapunov” form, which only assumes independent random variables.
Limits

**Definition (Convergence in probability)**

A sequence $X_1, X_2, \ldots$ of random variables *converges to a random variable $X$ in probability*, written as

$$X \xrightarrow{P} X,$$  \hspace{1cm} (46)

if for every $\varepsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$  \hspace{1cm} (47)

Remarks

• $X \xrightarrow{P} X$ means that the probability that $X_n$ lies outside the random interval $[X - \varepsilon, X + \varepsilon]$, no matter how small this interval may be, approaches 0 as $n \to \infty$.

• Intuitively, for a constant random variable $X := x$, this means that the probability distribution of $X_i$ becomes increasingly concentrated around $x$ as $n \to \infty$.
Limits

Definition (Convergence in probability)

A sequence $X_1, X_2, \ldots$ of random variables converges to a random variable $X$ in probability, written as

$$X \xrightarrow{P} X,$$  \hspace{1cm} (46)

if for every $\varepsilon > 0$

$$\lim_{n \to \infty} P(|X_n - X| > \varepsilon) = 0.$$  \hspace{1cm} (47)

Remarks

- $X \xrightarrow{P} X$ means that the probability that $X_n$ lies outside the random interval $[X - \varepsilon, X + \varepsilon]$, no matter how small this interval may be, approaches 0 as $n \to \infty$. 

### Definition (Convergence in probability)

A sequence $X_1, X_2, \ldots$ of random variables *converges to a random variable $X$ in probability*, written as

$$X \xrightarrow{P} X,$$

if for every $\varepsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$  

### Remarks

- $X \xrightarrow{P} X$ means that the probability that $X_n$ lies outside the random interval $[X - \varepsilon, X + \varepsilon]$, no matter how small this interval may be, approaches 0 as $n \to \infty$.
- Intuitively, for a constant random variable $X := x$, this means that the probability distribution of $X_i$ becomes increasingly concentrated around $x$ as $n \to \infty$. 


Theorem (Weak Law of Large Numbers)

Let $X_1, \ldots, X_n$ denote a random sample from a distribution with expectation $\mu$. Let

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

(48)

denote the sample mean. Then $\bar{X}_n$ converges to $\mu$ in probability,

$$\bar{X}_n \xrightarrow{P} \mu.$$ 

(49)
Theorem (Weak Law of Large Numbers)

Let \( X_1, \ldots, X_n \) denote a random sample from a distribution with expectation \( \mu \). Let

\[
\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i
\]  

(48)
denote the sample mean. Then \( \bar{X}_n \) converges to \( \mu \) in probability,

\[
\bar{X}_n \xrightarrow{P} \mu.
\]  

(49)

Remark

- \( \bar{X}_n \xrightarrow{P} \mu \) means that there is a high probability that the sample mean is close to the expectation of the \( X_i, i = 1, \ldots, n \), if the sample size \( n \) is large.
Definition (Almost sure convergence)

A sequence $X_1, X_2, \ldots$ of random variables converges almost surely to a random variable $X$, written as

$$X \xrightarrow{\text{a.s.}} X,$$  

(50)

if for every $\varepsilon > 0$

$$\mathbb{P}\left(\lim_{n \to \infty} |X_n - \mu| < \varepsilon\right) = 1.$$  

(51)
Limits

Definition (Almost sure convergence)

A sequence $X_1, X_2, \ldots$ of random variables converges almost surely to a random variable $X$, written as

$$X \xrightarrow{a.s.} X,$$

if for every $\varepsilon > 0$

$$\mathbb{P} \left( \lim_{n \to \infty} |X_n - \mu| < \varepsilon \right) = 1.$$ 

Remarks

- Recall that for $(\Omega, \mathcal{A}, \mathbb{P})$ random variables are functions $X : \Omega \to \mathcal{X}$. 
Limits

Definition (Almost sure convergence)

A sequence $X_1, X_2, \ldots$ of random variables converges almost surely to a random variable $X$, written as

$$X \xrightarrow{\text{a.s.}} X,$$

if for every $\varepsilon > 0$

$$\mathbb{P} \left( \lim_{n \to \infty} |X_n - \mu| < \varepsilon \right) = 1.$$  

Remarks

- Recall that for $(\Omega, \mathcal{A}, \mathbb{P})$ random variables are functions $X : \Omega \to \mathcal{X}$.
- Let $N \subset \Omega$ be a null set, i.e. $\mathbb{P}(N) = 0$. 

## Limits

### Definition (Almost sure convergence)

A sequence $X_1, X_2, ...$ of random variables *converges almost surely to a random variable* $X$, written as

$$X \xrightarrow{\text{a.s.}} X,$$

if for every $\varepsilon > 0$

$$\mathbb{P}\left( \lim_{n \to \infty} |X_n - \mu| < \varepsilon \right) = 1. \quad (51)$$

### Remarks

- Recall that for $(\Omega, \mathcal{A}, \mathbb{P})$ random variables are functions $X : \Omega \to \mathcal{X}$.
- Let $N \subset \Omega$ be a null set, i.e. $\mathbb{P}(N) = 0$.
- A.s. convergence implies $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega \setminus N$. 

Limits

Definition (Almost sure convergence)

A sequence \( X_1, X_2, \ldots \) of random variables \( \text{converges almost surely to a random variable} \ X \), written as

\[
X \xrightarrow{\text{a.s.}} X, \tag{50}
\]

if for every \( \varepsilon > 0 \)

\[
P \left( \lim_{n \to \infty} |X_n - \mu| < \varepsilon \right) = 1. \tag{51}
\]

Remarks

- Recall that for \((\Omega, \mathcal{A}, \mathbb{P})\) random variables are functions \( X : \Omega \to \mathcal{X} \).
- Let \( N \subset \Omega \) be a null set, i.e. \( \mathbb{P}(N) = 0 \).
- A.s. convergence implies \( X_n(\omega) \to X(\omega) \) for all \( \omega \in \Omega \setminus N \).
- A.s. convergence corresponds to pointwise convergence of function sequences
Limits

Definition (Almost sure convergence)

A sequence $X_1, X_2, \ldots$ of random variables converges almost surely to a random variable $X$, written as

$$X \overset{a.s.}{\rightarrow} X,$$  \hspace{1cm} (50)

if for every $\varepsilon > 0$

$$\mathbb{P} \left( \lim_{n \to \infty} |X_n - \mu| < \varepsilon \right) = 1.$$  \hspace{1cm} (51)

Remarks

- Recall that for $(\Omega, \mathcal{A}, \mathbb{P})$ random variables are functions $X : \Omega \rightarrow \mathcal{X}$.
- Let $N \subset \Omega$ be a null set, i.e. $\mathbb{P}(N) = 0$.
- A.s. convergence implies $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega \setminus N$.
- A.s. convergence corresponds to pointwise convergence of function sequences.
- A.s. convergence implies convergence in probability, but not vice versa.
Limits

Definition (Almost sure convergence)

A sequence $X_1, X_2, \ldots$ of random variables converges almost surely to a random variable $X$, written as

$$X \xrightarrow{a.s.} X,$$

if for every $\varepsilon > 0$

$$\mathbb{P}\left(\lim_{n \to \infty} |X_n - \mu| < \varepsilon\right) = 1. \quad (51)$$

Remarks

• Recall that for $(\Omega, \mathcal{A}, \mathbb{P})$ random variables are functions $X : \Omega \to \mathcal{X}$.

• Let $N \subset \Omega$ be a null set, i.e. $\mathbb{P}(N) = 0$.

• A.s. convergence implies $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega \setminus N$.

• A.s. convergence corresponds to pointwise convergence of function sequences.

• A.s. convergence implies convergence in probability, but not vice versa.

• A.s. convergence is a strong form of random variable convergence.
### Theorem (Strong Law of Large Numbers)

Let $X_1, \ldots, X_n$ denote a random sample from a distribution with expectation $\mu$. Let

$$
\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \quad (52)
$$

denote the sample mean. Then $\bar{X}_n$ converges almost surely to $\mu$,

$$
\bar{X}_n \xrightarrow{a.s.} \mu. \quad (53)
$$
Theorem (Strong Law of Large Numbers)

Let \( X_1, \ldots, X_n \) denote a random sample from a distribution with expectation \( \mu \). Let

\[
\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \tag{52}
\]

denote the sample mean. Then \( \bar{X}_n \) converges almost surely to \( \mu \),

\[
\bar{X}_n \xrightarrow{a.s.} \mu. \tag{53}
\]

Remark

- \( \bar{X}_n \xrightarrow{a.s.} \mu \) means that the probability that \(|X_n - X|\) is smaller than some arbitrarily small \( \epsilon > 0 \) is 1 as \( n \) goes to infinity.
## Limits

**Definition (Convergence in distribution)**

A sequence $X_1, X_2, \ldots$ of random variables *converges to a random variable $X$ in distribution*, written as

$$X \xrightarrow{D} X,$$  \hspace{1cm} (54)

if

$$\lim_{n \to \infty} P_{X_n}(x) = P_X(x).$$ \hspace{1cm} (55)

at all points where $P_X$ is continuous.
### Limits

#### Definition (Convergence in distribution)

A sequence $X_1, X_2, \ldots$ of random variables *converges to a random variable $X$ in distribution*, written as

$$X \xrightarrow{D} X,$$

if

$$\lim_{n \to \infty} P_{X_n}(x) = P_X(x).$$

(55)

at all points where $P_X$ is continuous.

#### Remarks

- $X \xrightarrow{D} X$ is a statement about the convergence of CDFs.
Limits

Definition (Convergence in distribution)

A sequence $X_1, X_2, \ldots$ of random variables *converges to a random variable $X$ in distribution*, written as

$$X \overset{D}{\rightarrow} X,$$

if

$$\lim_{n \to \infty} P_{X_n}(x) = P_X(x).$$

at all points where $P_X$ is continuous.

Remarks

• $X \overset{D}{\rightarrow} X$ is a statement about the convergence of CDFs.

• Convergence in probability implies convergence in distribution.
## Limits

### Definition (Convergence in distribution)

A sequence $X_1, X_2, \ldots$ of random variables *converges to a random variable $X$ in distribution*, written as

$$X \xrightarrow{D} X,$$

if

$$\lim_{n \to \infty} P_{X_n}(x) = P_X(x).$$

at all points where $P_X$ is continuous.

### Remarks

- $X \xrightarrow{D} X$ is a statement about the convergence of CDFs.
- Convergence in probability implies convergence in distribution.
- Almost sure convergence implies convergence in distribution.
Limits

Definition (Convergence in distribution)

A sequence $X_1, X_2, \ldots$ of random variables converges to a random variable $X$ in distribution, written as

$$X \overset{D}{\to} X,$$

(54)

if

$$\lim_{n \to \infty} P_{X_n}(x) = P_X(x).$$

(55)

at all points where $P_X$ is continuous.

Remarks

- $X \overset{D}{\to} X$ is a statement about the convergence of CDFs.
- Convergence in probability implies convergence in distribution.
- Almost sure convergence implies convergence in distribution.
- Convergence in distribution is a weak form of convergence.
Limits

Theorem (Central limit theorem (Lindenberg and Lévy))

Let the random variables $X_1, \ldots, X_n$ form an independent and identically distributed random sample of size $n$ from a given distribution with expectation $\mu$ and variance $0 < \sigma^2 < \infty$.

Then for each fixed number $x$

$$\lim_{n \to \infty} P\left( \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x \right) = \Phi(x) \quad (56)$$

where $\Phi$ is the CDF of the standard normal distribution.

Remarks

• For large $n$, the distribution of $\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ is approximately $\mathcal{N}(0, 1)$.

• For large $n$, the distribution of $\bar{X}_n$ is approximately $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

• For large $n$, the distribution of $\sum_{i=1}^{n} X_i$ is approximately $\mathcal{N}(n\mu, n\sigma^2)$. 

Theorem (Central limit theorem (Lindenberg and Lévy))

Let the random variables $X_1, \ldots, X_n$ form an independent and identically distributed random sample of size $n$ from a given distribution with expectation $\mu$ and variance $0 < \sigma^2 < \infty$. Then for each fixed number $x$

\[
\lim_{n \to \infty} P \left( \frac{\bar{X}_n - \mu}{\sigma / n^{1/2}} \leq x \right) = \Phi(x) \tag{56}
\]

where $\Phi$ is the CDF of the standard normal distribution.
Theorem (Central limit theorem (Lindenberg and Lévy))

Let the random variables $X_1, \ldots, X_n$ form an independent and identically distributed random sample of size $n$ from a given distribution with expectation $\mu$ and variance $0 < \sigma^2 < \infty$. Then for each fixed number $x$

$$\lim_{n \to \infty} \mathbb{P} \left( \frac{\bar{X}_n - \mu}{\sigma/n^{1/2}} \leq x \right) = \Phi(x)$$

(56)

where $\Phi$ is the CDF of the standard normal distribution.

Remarks

- For large $n$, the distribution of $n^{1/2}(\bar{X}_n - \mu)/\sigma$ is approximately $N(0, 1)$. 

Theorem (Central limit theorem (Lindenberg and Lévy))

Let the random variables $X_1, \ldots, X_n$ form an independent and identically distributed random sample of size $n$ from a given distribution with expectation $\mu$ and variance $0 < \sigma^2 < \infty$. Then for each fixed number $x$

$$
\lim_{n \to \infty} P \left( \frac{\bar{X}_n - \mu}{\sigma / n^{1/2}} \leq x \right) = \Phi(x) \quad (56)
$$

where $\Phi$ is the CDF of the standard normal distribution.

Remarks

- For large $n$, the distribution of $n^{1/2}(\bar{X}_n - \mu)/\sigma$ is approximately $N(0, 1)$.
- For large $n$, the distribution of $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$. 
Theorem (Central limit theorem (Lindenberg and Lévy))

Let the random variables $X_1, ..., X_n$ form an independent and identically distributed random sample of size $n$ from a given distribution with expectation $\mu$ and variance $0 < \sigma^2 < \infty$. Then for each fixed number $x$

$$\lim_{n \to \infty} \mathbb{P} \left( \frac{\bar{X}_n - \mu}{\sigma / n^{1/2}} \leq x \right) = \Phi(x)$$

where $\Phi$ is the CDF of the standard normal distribution.

Remarks

- For large $n$, the distribution of $n^{1/2}(\bar{X}_n - \mu)/\sigma$ is approximately $N(0, 1)$.
- For large $n$, the distribution of $\bar{X}_n$ is approximately $N(\mu, \sigma^2 / n)$.
- For large $n$, the distribution of $\sum_{i=1}^{n} X_i$ is approximately $N(n\mu, n\sigma^2)$. 
Theorem (Central limit theorem (Liapounov))

Let $X_1, X_2, \ldots$ be a sequence of independent, but not necessarily identically, distributed random variables, such that

$$\mathbb{E}(\left|X_i - \mu_i\right|^3) < \infty \quad \text{and} \quad \lim_{n \to \infty} \frac{\sum_{i=1}^{n} \mathbb{E}(\left|X_i - \mu_i\right|^3)}{(\sum_{i=1}^{n} \sigma_i^2)^{3/2}} = 0.$$  \hfill (57)

Let $\mu_i := \mathbb{E}(X_i)$ and $\sigma_i^2 := \mathbb{V}(X_i)$ for $i = 1, \ldots, n$ and define

$$Y_n := \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i \sqrt{\sum_{i=1}^{n} \sigma_i^2},$$

such that $\mathbb{E}(Y_n) = 0$ and $\mathbb{V}(Y_n) = 1$.

Remarks

- For large $n$, the distribution of $\sum_{i=1}^{n} X_i$ is approximately $\mathcal{N}(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$.
- The sum of many independent random factors is approximately normally distributed.
- This justifies the ubiquitous assumption of normally distributed observation errors.
Theorem (Central limit theorem (Liapounov))

Let $X_1, X_2, \ldots$ be a sequence of independent, but not necessarily identically, distributed random variables, such that

$$\mathbb{E}(|X_i - \mu_i|^3) < \infty \quad \text{and} \quad \lim_{n \to \infty} \frac{\sum_{i=1}^{n} \mathbb{E}(|X_i - \mu_i|^3)}{(\sum_{i=1}^{n} \sigma_i^2)^{3/2}} = 0. \quad (57)$$

Let $\mu_i := \mathbb{E}(X_i)$ and $\sigma_i^2 = \mathbb{V}(X_i)$ for $i = 1, \ldots, n$ and define

$$Y_n := \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}, \quad (58)$$

such that $\mathbb{E}(Y_n) = 0$ and $\mathbb{V}(Y_n) = 1$. 
Theorem (Central limit theorem (Liapounov))

Let \( X_1, X_2, \ldots \) be a sequence of independent, but not necessarily identically, distributed random variables, such that

\[
\mathbb{E}(|X_i - \mu_i|^3) < \infty \quad \text{and} \quad \lim_{n \to \infty} \frac{\sum_{i=1}^{n} \mathbb{E}(|X_i - \mu_i|^3)}{(\sum_{i=1}^{n} \sigma_i^2)^{3/2}} = 0.
\]  

(57)

Let \( \mu_i := \mathbb{E}(X_i) \) and \( \sigma_i^2 = \mathbb{V}(X_i) \) for \( i = 1, \ldots, n \) and define

\[
Y_n := \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}},
\]  

(58)

such that \( \mathbb{E}(Y_n) = 0 \) and \( \mathbb{V}(Y_n) = 1 \). Then, for each fixed number \( x \),

\[
\lim_{n \to \infty} \mathbb{P}(Y_n \leq x) = \Phi(x),
\]  

(59)

where \( \Phi \) is the CDF of the standard normal distribution.
Limits

Theorem (Central limit theorem (Liapounov))

Let $X_1, X_2, \ldots$ be a sequence of independent, but not necessarily identically, distributed random variables, such that

$$
E(|X_i - \mu_i|^3) < \infty \quad \text{and} \quad \lim_{n \to \infty} \frac{\sum_{i=1}^{n} E(|X_i - \mu_i|^3)}{\left(\sum_{i=1}^{n} \sigma_i^2\right)^{3/2}} = 0. \quad (57)
$$

Let $\mu_i := E(X_i)$ and $\sigma_i^2 := V(X_i)$ for $i = 1, \ldots, n$ and define

$$
Y_n := \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}, \quad (58)
$$

such that $E(Y_n) = 0$ and $V(Y_n) = 1$. Then, for each fixed number $x$,

$$
\lim_{n \to \infty} P(Y_n \leq x) = \Phi(x), \quad (59)
$$

where $\Phi$ is the CDF of the standard normal distribution.

Remarks

- For large $n$, the distribution of $\sum_{i=1}^{n} X_i$ is approximately $N \left( \sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2 \right)$. 


Limits

**Theorem (Central limit theorem (Liapounov))**

Let $X_1, X_2, ...$ be a sequence of independent, but not necessarily identically, distributed random variables, such that

$$\mathbb{E}(|X_i - \mu_i|^3) < \infty \quad \text{and} \quad \lim_{n \to \infty} \frac{\sum_{i=1}^{n} \mathbb{E}(|X_i - \mu_i|^3)}{(\sum_{i=1}^{n} \sigma_i^2)^{3/2}} = 0. \quad (57)$$

Let $\mu_i := \mathbb{E}(X_i)$ and $\sigma_i^2 = \mathbb{V}(X_i)$ for $i = 1, ..., n$ and define

$$Y_n := \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}, \quad (58)$$

such that $\mathbb{E}(Y_n) = 0$ and $\mathbb{V}(Y_n) = 1$. Then, for each fixed number $x$,

$$\lim_{n \to \infty} \mathbb{P}(Y_n \leq x) = \Phi(x), \quad (59)$$

where $\Phi$ is the CDF of the standard normal distribution.

**Remarks**

- For large $n$, the distribution of $\sum_{i=1}^{n} X_i$ is approximately $N \left( \sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2 \right)$.
- The sum of many independent random factors is approximately normally distributed.
Theorem (Central limit theorem (Liapounov))

Let $X_1, X_2, \ldots$ be a sequence of independent, but not necessarily identically, distributed random variables, such that

$$
\mathbb{E}(|X_i - \mu_i|^3) < \infty \quad \text{and} \quad \lim_{n \to \infty} \frac{\sum_{i=1}^{n} \mathbb{E}(|X_i - \mu_i|^3)}{(\sum_{i=1}^{n} \sigma_i^2)^{3/2}} = 0.
$$

(57)

Let $\mu_i := \mathbb{E}(X_i)$ and $\sigma_i^2 = \mathbb{V}(X_i)$ for $i = 1, \ldots, n$ and define

$$
Y_n := \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}},
$$

(58)
such that $\mathbb{E}(Y_n) = 0$ and $\mathbb{V}(Y_n) = 1$. Then, for each fixed number $x$,

$$
\lim_{n \to \infty} \mathbb{P}(Y_n \leq x) = \Phi(x),
$$

(59)

where $\Phi$ is the CDF of the standard normal distribution.

Remarks

- For large $n$, the distribution of $\sum_{i=1}^{n} X_i$ is approximately $N \left( \sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2 \right)$.
- The sum of many independent random factors is approximately normally distributed.
- This justifies the ubiquitous assumption of normally distributed observation errors.
Frequentist inference
(6) Foundations and maximum likelihood
Bibliographic remarks

The material presented in this section follows Wasserman (2004, Sections 6.1 - 6.3, 9.1, 9.3) and Held and Sabanés Bové (2014, Sections 2.2 - 2.3, C.1.3).
Foundations

Maximum likelihood estimation

• Foundations

• Analytical maximum likelihood estimation
  • Bernoulli parameter estimation
  • Gaussian parameter estimation

• Numerical maximum likelihood estimation
  • The Newton-Raphson method
  • Fisher scoring
Foundations

Maximum likelihood estimation

• Foundations

• Analytical maximum likelihood estimation
  • Bernoulli parameter estimation
  • Gaussian parameter estimation

• Numerical maximum likelihood estimation
  • The Newton-Raphson method
  • Fisher scoring
A statistical model $\mathcal{P}$ is a set of probability distributions.
A statistical model $\mathcal{P}$ is a set of probability distributions.

• A *parametric statistical model* is a statistical model that can be parameterized by a finite number of parameters.
A statistical model $\mathcal{P}$ is a set of probability distributions.

- A *parametric statistical model* is a statistical model that can be parameterized by a finite number of parameters.

- A *nonparametric statistical model* is a statistical model that cannot be parameterized by a finite number of parameters.
Parametric statistical models

Typical parametric statistical models have the form

$$\mathcal{P} = \{p_\theta(x) | \theta \in \Theta\}, \quad (60)$$

where

- $p_\theta$ is a PMF or PDF parameterized by $\theta$,
- $\theta$ is a parameter (vector), and
- $\Theta$ is the parameter space.
Foundations

Parametric statistical models

Typical parametric statistical models have the form

\[ \mathcal{P} = \{ p_\theta(x) | \theta \in \Theta \} \]

(60)

where

- \( p_\theta \) is a PMF or PDF parameterized by \( \theta \),
- \( \theta \) is a parameter (vector), and
- \( \Theta \) is the parameter space.

Examples

- \( X_1, \ldots, X_n \sim p_\mu \) and \( p_\mu \in \mathcal{P} := \{ B(x; \mu) | \mu \in [0, 1] \} \)
- \( X_1, \ldots, X_n \sim p_{\mu, \sigma^2} \) and \( p_{\mu, \sigma^2} \in \mathcal{P} := \{ N(x; \mu, \sigma^2) | (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} \} \)
The standard problems of frequentist inference

(1) *Parameter estimation*

The aim of parameter estimation is to find a best guess for the true, but unknown, parameter value of the model, typically based on the observation of $X_1, \ldots, X_n \sim p$. 
The standard problems of frequentist inference

(1) *Parameter estimation*

The aim of parameter estimation is to find a best guess for the true, but unknown, parameter value of the model, typically based on the observation of $X_1, \ldots, X_n \sim p$.

(2) *Confidence interval evaluation*

The aim of confidence interval evaluation is to provide a quantitative uncertainty statement about a parameter estimate based on the parameter estimator’s sampling distribution.
Foundations

The standard problems of frequentist inference

(1) Parameter estimation
   The aim of parameter estimation is to find a best guess for the true, but unknown, parameter value of the model, typically based on the observation of \(X_1, \ldots, X_n \sim p\).

(2) Confidence interval evaluation
   The aim of confidence interval evaluation is to provide a quantitative uncertainty statement about a parameter estimate based on the parameter estimator’s sampling distribution.

(3) Hypothesis testing
   The aim of hypothesis testing is to decide, based on the observations \(X_1, \ldots, X_n\) and in a sensible fashion whether, the true, but unknown, parameter is in one of usually two mutually exclusive subsets of the parameter space.
Foundations

The standard problems of frequentist inference

1. **Parameter estimation**
   The aim of parameter estimation is to find a best guess for the true, but unknown, parameter value of the model, typically based on the observation of $X_1, \ldots, X_n \sim p$.

2. **Confidence interval evaluation**
   The aim of confidence interval evaluation is to provide a quantitative uncertainty statement about a parameter estimate based on the parameter estimator’s sampling distribution.

3. **Hypothesis testing**
   The aim of hypothesis testing is to decide, based on the observations $X_1, \ldots, X_n$ and in a sensible fashion whether, the true, but unknown, parameter is in one of usually two mutually exclusive subsets of the parameter space.
The standard problems of frequentist inference

1) *Parameter estimation*
   The aim of parameter estimation is to find a best guess for the true, but unknown, parameter value of the model, typically based on the observation of \(X_1, \ldots, X_n \sim p\).

2) *Confidence interval evaluation*
   The aim of confidence interval evaluation is to provide a quantitative uncertainty statement about a parameter estimate based on the parameter estimator’s sampling distribution.

3) *Hypothesis testing*
   The aim of hypothesis testing is to decide, based on the observations \(X_1, \ldots, X_n\) and in a sensible fashion whether, the true, but unknown, parameter is in one of usually two mutually exclusive subsets of the parameter space.

Confidence interval evaluation and hypothesis testing make extensive use of statistics \(h(X_1, \ldots, X_n)\) and their distributional properties.
Foundations

Maximum likelihood estimation

- Foundations
- Analytical maximum likelihood estimation
  - Bernoulli parameter estimation
  - Gaussian parameter estimation
- Numerical maximum likelihood estimation
  - The Newton-Raphson method
  - Fisher scoring
Definition (Point estimator)

Let $X_1, \ldots, X_n$ be $n$ independent and identically distributed data points from some distribution with outcome space $\mathcal{X}$ defined in terms of a PMF or PDF $p_\theta(x)$. A point estimator $\hat{\theta}_n$ is a function of $X_1, \ldots, X_n$, written as $\hat{\theta}_n(X_1, \ldots, X_n)$.

Remarks

- A point estimator provides a single best guess of some quantity of interest.
Definition (Point estimator)

Let $X_1, ..., X_n$ be $n$ independent and identically distributed data points from some distribution with outcome space $\mathcal{X}$ defined in terms of a PMF or PDF $p_\theta(x)$. A point estimator $\hat{\theta}_n$ is a function of $X_1, ..., X_n$, written as $\hat{\theta}_n(X_1, ..., X_n)$.

Remarks

- A point estimator provides a single best guess of some quantity of interest.
- As a function of random variables, a point estimator is a random variable.
## Definition (Point estimator)

Let $X_1, ..., X_n$ be $n$ independent and identically distributed data points from some distribution with outcome space $\mathcal{X}$ defined in terms of a PMF or PDF $p_\theta(x)$. A point estimator $\hat{\theta}_n$ is a function of $X_1, ..., X_n$, written as $\hat{\theta}_n(X_1, ..., X_n)$.

### Remarks

- A point estimator provides a single best guess of some quantity of interest.
- As a function of random variables, a point estimator is a random variable.
- A value taken on by point estimator is referred to as *point estimate*. 
Definition (Point estimator)

Let $X_1, \ldots, X_n$ be $n$ independent and identically distributed data points from some distribution with outcome space $\mathcal{X}$ defined in terms of a PMF or PDF $p_\theta(x)$. A point estimator $\hat{\theta}_n$ is a function of $X_1, \ldots, X_n$, written as $\hat{\theta}_n(X_1, \ldots, X_n)$.

Remarks

- A point estimator provides a single best guess of some quantity of interest.
- As a function of random variables, a point estimator is a random variable.
- A value taken on by point estimator is referred to as point estimate.
- Frequentist inference is concerned with the properties of point estimators.
Foundations

**Definition (Point estimator)**

Let $X_1, ..., X_n$ be $n$ independent and identically distributed data points from some distribution with outcome space $\mathcal{X}$ defined in terms of a PMF or PDF $p_\theta(x)$. A point estimator $\hat{\theta}_n$ is a function of $X_1, ..., X_n$, written as $\hat{\theta}_n(X_1, ..., X_n)$.

**Remarks**

- A point estimator provides a single best guess of some quantity of interest.
- As a function of random variables, a point estimator is a random variable.
- A value taken on by point estimator is referred to as *point estimate*.
- Frequentist inference is concerned with the properties of point estimators.
- *Maximum likelihood estimation* is the most common method for point estimation in parametric statistical methods.
Foundations

Definition (Likelihood function and log likelihood function)

Let $\mathcal{P}$ denote a parametric statistical model with PMF or PDF $p_\theta(x)$ and let $X_1, \ldots, X_n \sim p_\theta$. The likelihood function is defined as

$$L_n : \Theta \rightarrow [0, \infty[, \theta \mapsto L_n(\theta) := \prod_{i=1}^{n} p_\theta(x_i).$$

(61)
### Definition (Likelihood function and log likelihood function)

Let $\mathcal{P}$ denote a parametric statistical model with PMF or PDF $p_\theta(x)$ and let $X_1, \ldots, X_n \sim p_\theta$. The *likelihood function* is defined as

$$L_n : \Theta \rightarrow [0, \infty[, \theta \mapsto L_n(\theta) := \prod_{i=1}^{n} p_\theta(x_i).$$  \hspace{1cm} (61)  

The *log likelihood function* is defined as

$$\ell_n : \Theta \rightarrow \mathbb{R}, \theta \mapsto \ell_n(\theta) := \ln L_n(\theta).$$  \hspace{1cm} (62)  

### Remarks

- The value of the likelihood function is the joint density of the data.
### Definition (Likelihood function and log likelihood function)

Let $\mathcal{P}$ denote a parametric statistical model with PMF or PDF $p_\theta(x)$ and let $X_1, \ldots, X_n \sim p_\theta$. The *likelihood function* is defined as

$$L_n : \Theta \to [0, \infty[, \theta \mapsto L_n(\theta) := \prod_{i=1}^{n} p_\theta(x_i). \quad (61)$$

The *log likelihood function* is defined as

$$\ell_n : \Theta \to \mathbb{R}, \theta \mapsto \ell_n(\theta) := \ln L_n(\theta). \quad (62)$$

### Remarks

- The value of the likelihood function is the joint density of the data.
- The likelihood function is a function of the parameter.
Definition (Likelihood function and log likelihood function)

Let $\mathcal{P}$ denote a parametric statistical model with PMF or PDF $p_{\theta}(x)$ and let $X_1, \ldots, X_n \sim p_{\theta}$. The 
\textit{likelihood function} is defined as

$$L_n : \Theta \rightarrow [0, \infty[, \theta \mapsto L_n(\theta) := \prod_{i=1}^{n} p_{\theta}(x_i).$$  \hspace{1cm} (61)

The \textit{log likelihood function} is defined as

$$\ell_n : \Theta \rightarrow \mathbb{R}, \theta \mapsto \ell_n(\theta) := \ln L_n(\theta).$$ \hspace{1cm} (62)

Remarks

- The value of the likelihood function is the joint density of the data.
- The likelihood function is a function of the parameter.
- In general the likelihood function does not integrate to 1 with respect to $\theta$. 


Definition (Maximum likelihood estimator)

The \textit{maximum likelihood estimator} (MLE) is defined as

\[ \hat{\theta}_n^{ML} = \arg \max_{\theta \in \Theta} L_n(\theta) = \arg \max_{\theta \in \Theta} \ell_n(\theta). \]  

(63)
Definition (Maximum likelihood estimator)

The *maximum likelihood estimator* (MLE) is defined as

\[
\hat{\theta}^\text{ML}_n = \arg \max_{\theta \in \Theta} L_n(\theta) = \arg \max_{\theta \in \Theta} \ell_n(\theta). \tag{63}
\]

Remarks

- Because \( \ln \) is monotonically increasing, the maximum of \( \ell_n \) occurs at the same place as does the maximum of \( L_n(\theta) \).
Definition (Maximum likelihood estimator)

The maximum likelihood estimator (MLE) is defined as

$$\hat{\theta}_n^{ML} = \arg\max_{\theta \in \Theta} L_n(\theta) = \arg\max_{\theta \in \Theta} \ell_n(\theta).$$  

Remarks

- Because \(\ln\) is monotonically increasing, the maximum of \(\ell_n\) occurs at the same place as does the maximum of \(L_n(\theta)\).
- Working with the log likelihood function is often easier.
Definition (Maximum likelihood estimator)

The *maximum likelihood estimator* (MLE) is defined as

\[
\hat{\theta}_n^{ML} = \arg \max_{\theta \in \Theta} L_n(\theta) = \arg \max_{\theta \in \Theta} \ell_n(\theta).
\]  

(63)

Remarks

- Because \( \ln \) is monotonically increasing, the maximum of \( \ell_n \) occurs at the same place as does the maximum of \( L_n(\theta) \).

- Working with the log likelihood function is often easier.

- Multiplying \( L_n \) by a positive constant not depending on \( \theta \) does not change the MLE, for maximum likelihood estimation, constant contributions to likelihood functions can thus be neglected.
Foundations

Definition (Maximum likelihood estimator)

The *maximum likelihood estimator* (MLE) is defined as

\[
\hat{\theta}_n^{\text{ML}} = \arg \max_{\theta \in \Theta} L_n(\theta) = \arg \max_{\theta \in \Theta} \ell_n(\theta).
\]  

(63)

Remarks

- Because \(\ln\) is monotonically increasing, the maximum of \(\ell_n\) occurs at the same place as does the maximum of \(L_n(\theta)\).
- Working with the log likelihood function is often easier.
- Multiplying \(L_n\) by a positive constant not depending on \(\theta\) does not change the MLE, for maximum likelihood estimation, constant contributions to likelihood functions can thus be neglected.
- Maximum likelihood is a standard problem in function maximization.
The general maximum likelihood procedure for parametric statistical models

1. Formulation of the log likelihood function.
2. Evaluation of the log likelihood function’s derivative and setting to zero.
3. Solving for critical points, verification of maximum.
The general maximum likelihood procedure for parametric statistical models

1. Formulation of the log likelihood function.
2. Evaluation of the log likelihood function’s derivative and setting to zero.
3. Solving for critical points, verification of maximum.

Two approaches for maximum likelihood estimation

1. Analytical function maximization in classical examples.
2. Numerical function maximization in most applied scenarios.
Foundations

Maximum likelihood estimation

- Foundations

- Analytical maximum likelihood estimation
  - Bernoulli parameter estimation
  - Gaussian parameter estimation

- Numerical maximum likelihood estimation
  - The Newton-Raphson method
  - Fisher scoring
Bernoulli parameter estimation

Example (Bernoulli distribution)

Let $X_1, ..., X_n \sim \text{Bern}(\mu)$ be $n$ i.i.d. Bernoulli distributed random variables.
Bernoulli parameter estimation

Example (Bernoulli distribution)

Let \( X_1, ..., X_n \sim \text{Bern}(\mu) \) be \( n \) i.i.d. Bernoulli distributed random variables.

(1) Formulation of the log likelihood function

We have

\[
L_n : [0, 1] \rightarrow [0, 1], \mu \mapsto L_n(\mu) := \prod_{i=1}^{n} \mu^{x_i} (1-\mu)^{1-x_i} = \mu \sum_{i=1}^{n} x_i (1-\mu)^{n-\sum_{i=1}^{n} x_i}. \tag{64}
\]
Example (Bernoulli distribution)

Let $X_1, \ldots, X_n \sim \text{Bern} (\mu)$ be $n$ i.i.d. Bernoulli distributed random variables.

(1) Formulation of the log likelihood function

We have

$$L_n : [0, 1] \to [0, 1], \mu \mapsto L_n(\mu) := \prod_{i=1}^{n} \mu^{x_i} (1-\mu)^{1-x_i} = \mu \sum_{i=1}^{n} x_i (1-\mu)^{n-\sum_{i=1}^{n} x_i}. \quad (64)$$

Taking the logarithm yields

$$\ell_n : [0, 1] \to \mathbb{R}, \mu \mapsto \ell_n(\mu) = \ln \mu \sum_{i=1}^{n} x_i + \ln(1-\mu) \left( n - \sum_{i=1}^{n} x_i \right). \quad (65)$$
Example (Bernoulli distribution)

(2) Evaluation of the log likelihood function derivative, setting to zero

We have

$$
\frac{d}{d\mu} \ell_n(\mu) = \frac{d}{d\mu} \ln \mu \sum_{i=1}^{n} x_i + \ln(1 - \mu) \left( n - \sum_{i=1}^{n} x_i \right)
$$

$$
= \frac{d}{d\mu} \ln \mu \sum_{i=1}^{n} x_i + \frac{d}{d\mu} \ln(1 - \mu) \left( n - \sum_{i=1}^{n} x_i \right)
$$

$$
= \frac{1}{\mu} \sum_{i=1}^{n} x_i - \frac{1}{1 - \mu} \left( n - \sum_{i=1}^{n} x_i \right)
$$

and hence the maximum likelihood equation takes the form

$$
\frac{1}{\hat{\mu}} \sum_{i=1}^{n} x_i - \frac{1}{1 - \hat{\mu}} \left( n - \sum_{i=1}^{n} x_i \right) = 0.
$$
Bernoulli parameter estimation

Example (Bernoulli distribution)

(3) Solving for critical points

We have

\[
\frac{1}{\hat{\mu}} \sum_{i=1}^{n} x_i - \frac{1}{1 - \hat{\mu}} \left( n - \sum_{i=1}^{n} x_i \right) = 0
\]

\[
\Leftrightarrow \hat{\mu}(1 - \hat{\mu}) \left( \frac{1}{\hat{\mu}} \sum_{i=1}^{n} x_i - \frac{1}{1 - \hat{\mu}} \left( n - \sum_{i=1}^{n} x_i \right) \right) = 0
\]

\[
\Leftrightarrow \sum_{i=1}^{n} x_i - \hat{\mu} \sum_{i=1}^{n} x_i - n\hat{\mu} + \hat{\mu} \sum_{i=1}^{n} x_i = 0
\]

(68)

\[
\Leftrightarrow n\hat{\mu} = \sum_{i=1}^{n} x_i
\]

\[
\Leftrightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i.
\]

Hence \( \hat{\mu} = n^{-1} \sum_{i=1}^{n} x_i \) is a candidate for an MLE of \( \mu \).
Foundations

Maximum likelihood estimation

• Foundations

• **Analytical maximum likelihood estimation**
  • Bernoulli parameter estimation
  • Gaussian parameter estimation

• Numerical maximum likelihood estimation
  • The Newton-Raphson method
  • Fisher scoring
Example (Gaussian distribution)

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be $n$ i.i.d. Gaussian random variables.
Gaussian parameter estimation

Example (Gaussian distribution)

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be $n$ i.i.d. Gaussian random variables.

(1) Formulation of the log likelihood function

We have

$$L_n : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}_{>0}, (\mu, \sigma^2) \mapsto L_n(\mu, \sigma^2) := \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right)$$

$$= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right).$$
Gaussian parameter estimation

Example (Gaussian distribution)

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be $n$ i.i.d. Gaussian random variables.

(1) Formulation of the log likelihood function

We have

$$L_n : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}_{>0}, (\mu, \sigma^2) \mapsto L_n(\mu, \sigma^2) := \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right)$$

$$= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right). \tag{69}$$

Taking the logarithm yields

$$\ell_n : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}, (\mu, \sigma^2) \mapsto \ell_n(\mu, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2. \tag{70}$$
Gaussian parameter estimation

Example (Gaussian distribution)

(2) Evaluation of the log likelihood function derivative, setting to zero

We have

\[
\frac{d}{d\mu} \ell_n(\mu) = -\frac{d}{d\mu} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{d}{d\mu} (x_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu).
\]  

(71)
Gaussian parameter estimation

Example (Gaussian distribution)

(2) Evaluation of the log likelihood function derivative, setting to zero

We have

\[
\frac{d}{d\mu} \ell_n(\mu) = -\frac{d}{d\mu} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{d}{d\mu} (x_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu). \tag{71}
\]

Similarly,

\[
\frac{d}{d\sigma^2} \ell_n(\mu) = -\frac{n}{2} \frac{d}{d\sigma^2} \ln \sigma^2 - \frac{d}{d\sigma^2} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2. \tag{72}
\]
Gaussian parameter estimation

Example (Gaussian distribution)

(2) Evaluation of the log likelihood function derivative, setting to zero

We have

\[
\frac{d}{d\mu} \ell_n(\mu) = -\frac{d}{d\mu} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{d}{d\mu} (x_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu). \tag{71}
\]

Similarly,

\[
\frac{d}{d\sigma^2} \ell_n(\mu) = -\frac{n}{2} \frac{d}{d\sigma^2} \ln \sigma^2 - \frac{d}{d\sigma^2} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2. \tag{72}
\]

Hence the maximum likelihood equations take the form

\[
\sum_{i=1}^{n} (x_i - \hat{\mu}) = 0
\]

\[
-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \tag{73}
\]
Gaussian parameter estimation

Example (Gaussian distribution)

(3) Solving for critical points

We have

\[ \sum_{i=1}^{n} (x_i - \hat{\mu}) = 0 \iff \sum_{i=1}^{n} x_i = n\hat{\mu} \iff \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i. \]  

(74)

Hence \( \hat{\mu} = n^{-1} \sum_{i=1}^{n} x_i \) is a candidate for an MLE of \( \mu \).
Example (Gaussian distribution)

(3) Solving for critical points

We have

\[ \sum_{i=1}^{n} (x_i - \hat{\mu}) = 0 \iff \sum_{i=1}^{n} x_i = n \hat{\mu} \iff \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i. \]  

(74)

Hence \( \hat{\mu} = n^{-1} \sum_{i=1}^{n} x_i \) is a candidate for an MLE of \( \mu \).

Substitution yields

\[ -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = 0 \]

\[ \iff -n\hat{\sigma}^2 + \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = 0 \]

(75)

\[ \iff \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2. \]

Hence \( \hat{\sigma} = n^{-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \) is a candidate for an MLE of \( \sigma^2 \).
Foundations

Maximum likelihood estimation

• Foundations

• Analytical maximum likelihood estimation
  • Bernoulli parameter estimation
  • Gaussian parameter estimation

• Numerical maximum likelihood estimation
  • The Newton-Raphson method
  • Fisher scoring
References


