Chapter 1: Basics of Probability Theory

Mathematical Statistics

UIC-DMS

February 27, 2024

Overview

Section 1

Events and their Probabilities

Sample space and events

- The set of all possible outcomes of a random experiment is known as the sample space of the experiment and is denoted by Ω.
- An "event" is a property which can be observed either to hold or not to hold *after* the experiment is done. Mathematically, an event is identified with a subset of Ω.

Example 1.1.1

- If the experiment consists of tossing two coins, then the sample space consists of the following four outcomes:
$$\Omega = \{HH, HT, TH, TT\}.$$
- Let E be the event that a head appears on the first coin. The event E occurs if and only if the outcome HH or HT appears. Thus we can describe E by the subset
$$E = \{HH, HT\}.$$

σ-Fields

- One collects "good" subsets of Ω, the events, in a class $\mathcal{F}$, say.
- In probability theory we require $\mathcal{F}$ to be a σ-field (also called σ-algebra). Such a class is supposed to contain all interesting events and is thus closed under usual set operations.

Definition 1.1.2 (σ-field)

Let $\mathcal{F}$ be a collection of subsets of Ω. We call $\mathcal{F}$ a σ-**field over** Ω, if

- $\emptyset \in \mathcal{F}$;
- If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$, where A^c denotes the complement of A;
- $\mathcal{F}$ is closed under countable unions: that is, if $A_1, A_2, A_3, \ldots$ is a countable sequence of events in $\mathcal{F}$, then $\cup_{i=1}^{\infty} A_i$ is also in $\mathcal{F}$.

- Roughly speaking, we would like that elementary operations such as $\cap, \cup$ and complement on the events of $\mathcal{F}$ should not lead outside the class $\mathcal{F}$. This is the intuitive meaning of a σ-field $\mathcal{F}$.

cont'd

Theorem 1.1.3

Suppose $\mathcal{F}$ is a σ-field, $A_1, A_2, \ldots$ are in $\mathcal{F}$, and $m \in \mathbb{N}$. Then each of the sets
$$\Omega,\ A_1 \backslash A_2,\ \bigcup_{j=1}^{m} A_j,\ \bigcap_{j=1}^{m} A_j,\ \bigcap_{j=1}^{\infty} A_j$$
also belongs to $\mathcal{F}$.

Probability measure

- To each event $A \in \mathcal{F}$ we assign a number $\mathbb{P}(A) \in [0, 1]$. This number is the expected fraction of occurrences of the event A in a long series of experiments where A are observed.

Definition 1.1.4 (Probability measure)

A **probability measure** *defined on a σ-field $\mathcal{F}$ over Ω is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ that satisfies*:

- $\mathbb{P}(\Omega) = 1$;
- $\mathbb{P}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$ *whenever the A_i are in $\mathcal{F}$ and are pairwise disjoint (i.e. $A_n \cap A_m = \emptyset$ if $n \neq m$).*

- We call $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space.

Elementary properties of probability measures

Theorem 1.1.5

1. $\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i)$, *if* $A_1, \ldots, A_n$ *are pairwise disjoint.*
2. $\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B)$.
3. $\mathbb{P}(B \backslash A) = \mathbb{P}(B) - \mathbb{P}(A)$ *if* $A \subset B$.
4. $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$.
5. *If* $A \subset B$, *then* $\mathbb{P}(A) \leq \mathbb{P}(B)$, *that is,* $\mathbb{P}$ *is monotone.*
6. $\mathbb{P}(A_n) \uparrow \mathbb{P}(A)$ *if* $A_n \uparrow A$. *Here* $A_n \uparrow A$ *means that* $A_1 \subset A_2 \subset \ldots$ *and* $\cup_{n=1} A_n = A$.
7. $\mathbb{P}(A_n) \downarrow \mathbb{P}(A)$ *if* $A_n \downarrow A$. *Here* $A_n \downarrow A$ *means that* $A_1 \supset A_2 \supset \ldots$ *and* $\cap_{n=1} A_n = A$.

Conditional probability and product rule

- Conditional probability: if $\mathbb{P}(A) > 0$, define

$$\mathbb{P}(B|A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

- All basic formulas of probability remain true, conditionally, e.g.:

$$\mathbb{P}(B^c|A) = 1 - \mathbb{P}(B|A),$$
$$\mathbb{P}(B \cup C|A) = \mathbb{P}(B|A) + \mathbb{P}(C|A) - \mathbb{P}(B \cap C|A).$$

Product rule

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \mathbb{P}(A) \cdot \mathbb{P}(B|A) \\
\mathbb{P}(A \cap B \cap C) &= \mathbb{P}(A) \cdot \mathbb{P}(B|A) \cdot \mathbb{P}(C|A \cap B) \\
\mathbb{P}(A \cap B \cap C \cap D) &= \mathbb{P}(A) \cdot \mathbb{P}(B|A) \cdot \mathbb{P}(C|A \cap B) \cdot \mathbb{P}(D|A \cap B \cap C) \\
&\vdots
\end{aligned}$$

Law of total probability and Bayes' Theorem

- A partition represents chopping the sample space into several smaller events, say $A_1, A_2, A_3, \ldots, A_n$, so that they
  - are mutually exclusive (i.e. don't overlap): $A_i \cap A_j = \emptyset$ for any $i \neq j$
  - cover the whole Ω (i.e. 'no gaps'): $A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n = \Omega$.

Law of total probability

For any partition, and any event B, we have

$$\mathbb{P}(B) = \mathbb{P}(B|A_1) \cdot \mathbb{P}(A_1) + \mathbb{P}(B|A_2) \cdot \mathbb{P}(A_2) + \ldots + \mathbb{P}(B|A_n) \cdot \mathbb{P}(A_n).$$

Bayes' Theorem

Conditional probabilities can be inverted. That is,

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)}$$

Overview

Section 2

Random Variables

Example 1.2.6 (Random Variables)

- A fair coin is tossed twice: $\Omega = \{HH, HT, TH, TT\}$.
  For $\omega \in \Omega$, let $X(\omega)$ be the number of heads, so that
  $$X(HH) = 2, \; X(HT) = X(TH) = 1, \; X(TT) = 0.$$
- Now suppose that a gambler wagers his fortune of \$1 on the result of this experiment. He gambles cumulatively so that his fortune is doubled each time a head appears, and is annihilated on the appearance of a tail. His subsequent fortune W is a random variable given by
  $$W(HH) = 4, \; W(HT) = W(TH) = W(TT) = 0.$$

Random Variables

We need the random variables to link sample spaces and events to data.

Definition 1.2.7 (Random Variables)

A **random variable** is a mapping $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome $\omega \in \Omega$, with the property that X is $\mathcal{F}$**-measurable**, that is, $\{\omega \in \Omega : X(\omega) \leq c\} \in \mathcal{F}$ for each $c \in \mathbb{R}$.

This mapping induces probability on $\mathbb{R}$ from Ω as follows: for $A \subset \mathbb{R}$ define
$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$$
and let
$$\mathbb{P}(X \in A) = \mathbb{P}\left(X^{-1}(A)\right) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\})$$

Definition 1.2.8 (Cumulative Distribution Function)

The **cumulative distribution function (CDF)** $F_X : \mathbb{R} \to [0, 1]$ is defined by
$$F_X(x) = \mathbb{P}(X \leq x).$$

Properties of CDFs

Theorem 1.2.9

A function $F : \mathbb{R} \to [0, 1]$ is a CDF for some random variable if and only if it satisfies the following three conditions:
(1) F is non-decreasing:
$$x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$$
(2) F is normalized:
$$\lim_{x \to -\infty} F(x) = 0 \text{ and } \lim_{x \to +\infty} F(x) = 1$$
(3) F is right-continuous:
$$\lim_{y \downarrow x} F(y) = F(x)$$

Discrete random variables

Definition 1.2.10 (Probability Mass Function)

X is discrete if it takes countable many values $\{x_1, x_2, \ldots\}$. We define the **probability mass function (PMF)** for X by

$$f_X(x) = \mathbb{P}(X = x)$$

Relationships between CDF and PMF:

- The CDF of X is related to the PMF f_X by

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$$

- The PMF f_X is related to the CDF F_X by

$$f_X(x) = F_X(x) - F_X(x^-) = F_X(x) - \lim_{y \uparrow x} F(y).$$

  Here $F_X(x^-)$ denots the left-limit of F_X at x.

Continuous random variables

Definition 1.2.11

A random variable is continuous if there exists a function f_X such that

- $f_X(x) \geq 0$ for all x
- $\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = 1$, and
- For every $A \subset \mathbb{R}$,

$$\mathbb{P}(X \in A) = \int_A f_X(x)\,\mathrm{d}x$$

- The function $f_X(x)$ is called the probability density function (PDF)
- Relationship between the CDF $F_X(x)$ and PDF $f_X(x)$:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \quad f_X(x) = F_X'(x)$$

Common distributions

(a) *Bernoulli*. A random variable is Bernoulli if $\mathbb{P}(X = 1) = p, \mathbb{P}(X = 0) = 1 - p$ for some $p \in [0, 1]$.

(b) *Binomial*. This is defined by $\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$, where n is a positive integer, $0 \leq k \leq n$, and $p \in [0, 1]$.

(c) *Geometric*. For $p \in (0, 1)$ we set $\mathbb{P}(X = k) = (1 - p)^k p$. Here k is a nonnegative integer.

(d) *Poisson*. For $\lambda > 0$ we set $\mathbb{P}(X = k) = e^{-\lambda} \lambda^k / k!$. Again k is a nonnegative integer.

(e) *Uniform*. For some positive integer n, set $\mathbb{P}(X = k) = 1/n$ for $1 \leq k \leq n$.

(f) *Uniform on* (a, b). Define $f(x) = (b - a)^{-1} \mathbf{1}_{(a,b)}(x)$, where $\mathbf{1}_{(a,b)}$ is the indicator function of the interval (a, b), i.e., $\mathbf{1}_{(a,b)}(x) = 1$ if $x \in (a, b)$ and $\mathbf{1}_{(a,b)}(x) = 0$ if $x \notin (a, b)$. If X has a uniform distribution, then

$$\mathbb{P}(X \in A) = \int_A \frac{1}{b - a} \mathbf{1}_{(a,b)}(x) dx.$$

Common distributions cont'd

(g) *Exponential*. For $x > 0$ let $f(x) = \beta e^{-\beta x}$ and otherwise $f(x) = 0$.

(h) *Standard normal*. Define $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. So

$$\mathbb{P}(X \in A) = \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2} dx.$$

(i) $N(\mu, \sigma^2)$. We shall see later that a standard normal has mean zero and variance one. If Z is a standard normal, then a $N(\mu, \sigma^2)$ random variable has the same distribution as $\mu + \sigma Z$. It is an exercise in calculus to check that such a random variable has density

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R}.$$

(j) Gamma(α, β). Here

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0.$$

Transformation of Random Variables

Suppose that X is a random variable with PDF f_X and CDF F_X.
Let $Y = r(X)$ be a function of X.

Q: How to compute the PDF and CDF of Y?

1. For each y, find the set $A_y = \{x : r(x) \leq y\}$
2. Find the CDF $F_Y(y)$

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(r(X) \leq y) = \mathbb{P}(X \in A_y) = \int_{A_y} f_X(x)\, \mathrm{d}x$$

3. The PDF is then $f_Y(y) = F_Y'(y)$

Important Fact: When r is strictly monotonic, then r has an inverse $s = r^{-1}$ and

$$\boxed{f_Y(y) = f_X(s(y)) \left| \frac{\mathrm{d}\, s(y)}{\mathrm{d}\, y} \right|}$$

Overview

Section 3

Bivariate Distributions

Joint Distributions

- Discrete Case

Definition 1.3.12

Given a pair of discrete random variables X and Y, their joint PMF is defined by

$$f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$$

- Continuous Case

Definition 1.3.13

A function $f_{X,Y}(x, y)$ is called the **joint PDF** of continuous random variables X and Y if

- $f_{X,Y}(x, y) \geq 0, \quad \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y = 1$
- For any set $A \subset \mathbb{R} \times \mathbb{R}$

$$\mathbb{P}((X, Y) \in A) = \iint_A f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y$$

The **joint CDF** of X and Y is defined as $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$

Marginal Distributions

- Discrete Case

If X and Y have joint PMF $f_{X,Y}$, then the **marginal PMF** of X is

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f_{X,Y}(x, y)$$

Similarly, the **marginal PMF** of Y is

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = y) = \sum_x f_{X,Y}(x, y)$$

- Continuous Case

If X and Y have joint PDF $f_{X,Y}$, then the **marginal PDFs** of X and Y are

$$f_X(x) = \int f_{X,Y}(x, y)\, \mathrm{d}y \quad \text{and} \quad f_Y(y) = \int f_{X,Y}(x, y)\, \mathrm{d}x$$

Independent Random Variables

Definition 1.3.14

Two random variables X and Y are **independent** if, for every $A, B \subset \mathbb{R}$

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

Criterion of independence:

Theorem 1.3.15

Let X and Y have joint PDF/PMF $f_{X,Y}$. Then X and Y are independent if and only if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

Conditional Distributions

- Discrete Case

The **conditional PMF**:

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

- Continuous Case

The **conditional PDF** is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Then,

$$\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y)\, \mathrm{d}x$$

Overview

Section 4

Expected Values

The expectation (or mean) of a random variable X is the average value of X.

Definition 1.4.16 (Expectation)

The **expectation**, or **mean**, or **first moment** of X is

$$\boxed{\mu_X \equiv \mathbb{E}[X] = \begin{cases} \sum_x x f_X(x), & \text{if } X \text{ is discrete} \\ \int x f_X(x)\, \mathrm{d}x, & \text{if } X \text{ is continuous} \end{cases}}$$

assuming that the sum (or integral) is well-defined.

- Let $Y = r(X)$, then $\mathbb{E}[Y] = \mathbb{E}[r(X)] = \sum_x r(x) f_X(x)$ or $\int r(x) f_X(x)\, \mathrm{d}x$
- If $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants, then

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}\left[X_i\right]$$

- Let X, Y be independent random variables. Then,

$$\mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$$

The variance measures the "spread" of a distribution.

Definition 1.4.17 (Variance)

Let X be a random variable with mean μ_X.
The **variance** of X, denoted $\mathrm{Var}[X]$ or σ_X^2, is defined by

$$\boxed{\sigma_X^2 \equiv \mathrm{Var}[X] = \mathbb{E}\left[(X-\mu_X)^2\right] = \begin{cases} \sum_x (x-\mu_X)^2 f_X(x), & \text{if } X \text{ is discrete} \\ \int (x-\mu_X)^2 f_X(x)\, \mathrm{d}x, & \text{if } X \text{ is continuous} \end{cases}}$$

The standard deviation is $\sigma_X = \sqrt{\mathrm{Var}[X]}$

Important Properties of $\mathrm{Var}[X]$

- $\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$
- If a and b are constants, then $\mathrm{Var}[aX+b] = a^2\,\mathrm{Var}[X]$
- If $X_1, \ldots, X_n$ are independent and $a_1, \ldots, a_n$ are constants, then

$$\mathrm{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2\,\mathrm{Var}\left[X_i\right]$$

Expectation and Variance of Important Random Variables

Distribution	Mean	Variance
Point mass at a	a	0
Bernoulli(p)	p	$p(1-p)$
Bin(n, p)	np	$np(1-p)$
Geom(p)	$1/p$	$(1-p)/p^2$
Poisson(λ)	λ	λ
Uniform(a, b)	$(a+b)/2$	$(b-a)^2/12$
$\mathcal{N}(\mu, \sigma^2)$	μ	σ^2
Exp(β)	$1/\beta$	$1/\beta^2$
Gamma(α, β)	α/β	α/β^2

Covariance and Correlation

If X and Y are random variables, then the covariance and correlation between X and Y measure how strong the linear relationship is between X and Y.

Definition 1.4.18 (Covariance)

Let X and Y be random variables with means μ_X and μ_Y and standard deviations σ_X and σ_Y. Define the **covariance** between X and Y by

$$\boxed{\operatorname{Cov}(X, Y) = \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]}$$

and the **correlation** (also called correlation coefficient) by

$$\boxed{\rho(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sigma_X \sigma_Y}}$$

Properties of Covariance and Correlation

- The covariance satisfies (useful in computations):

$$\operatorname{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

- The correlation satisfies:

$$-1 \leq \rho(X, Y) \leq 1$$

- If $Y = aX + b$ for some constants a and b, then

$$\rho(X, Y) = \begin{cases} 1, & \text{if } a > 0 \\ -1, & \text{if } a < 0 \end{cases}$$

- If X and Y are independent, then $\operatorname{Cov}(X, Y) = \rho(X, Y) = 0$. The converse is not true.
- For random variables $X_1, \ldots, X_n$

$$\operatorname{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \operatorname{Var}\left[X_i\right] + 2 \sum_{i<j} a_i a_j \operatorname{Cov}\left(X_i, X_j\right)$$

Conditional Expectation

- The conditional expectation of X given $Y = y$ is

$$\mathbb{E}[X|Y = y] = \begin{cases} \sum_X x f_{X|Y}(x|y), & \text{discrete case;} \\ \int x f_{X|Y}(x|y)\,\mathrm{d}x, & \text{continuous case.} \end{cases}$$

  - $\mathbb{E}[X]$ is a number
  - $\mathbb{E}[X|Y = y]$ is a function of y
  - $\mathbb{E}[X|Y]$ is the random variable whose value is $\mathbb{E}[X|Y = y]$ when $Y = y$
- The Rule of Iterated Expectations or Law of Total Expectation

$$\mathbb{E}\left(\mathbb{E}[X|Y]\right) = \mathbb{E}[X]$$

Conditional Variance

- The conditional variance of X given $Y = y$ is

$$\operatorname{Var}[X|Y = y] = \mathbb{E}\left[(X - \mathbb{E}[X|Y = y])^2 | Y = y\right]$$

  - $\operatorname{Var}[X]$ is a number
  - $\operatorname{Var}[X|Y = y]$ is a function of y
  - $\operatorname{Var}[X|Y]$ is the random variable whose value is $\operatorname{Var}[X|Y = y]$ when $Y = y$
- For random variables X and Y

$$\operatorname{Var}[X] = \mathbb{E}\operatorname{Var}[X|Y] + \operatorname{Var}\mathbb{E}[X|Y]$$

Moment-generating functions

Definition 1.4.19 (Moment-Generating Function)

The moment-generating function (MGF) of a random variable X is

$$M(t) = \mathbb{E}\left[e^{tx}\right]$$

(if the expectation is defined)

Important Properties of MGFs:

- If $\exists \varepsilon > 0$ such that $M(t)$ exists for all $t \in (-\varepsilon, \varepsilon)$, then $M(t)$ uniquely determines the probability distribution, and we write $M(t) \rightsquigarrow f(x)$.
- If $M(t)$ exists in an open interval containing zero, then

$$M^{(r)}(0) = \mathbb{E}\left[X^r\right] \qquad \text{(hence the name)}$$

  To find moments $\mathbb{E}\left[X^r\right]$, we must do integration or calculate a sum. Knowing the MGF allows to replace integration or sum by differentiation.

Moment-generating functions

Important Properties of MGFs: (continuation)

- If X has the MGF $M_X(t)$ and $Y = a + bX$, then

$$M_Y(t) = e^{at} M_X(bt)$$

- If X and Y are independent, then

$$M_{X+Y}(t) = M_X(t) M_Y(t)$$

- If X and Y have a joint distribution, then their joint MGF is defined as

$$M_{X,Y}(s,t) = \mathbb{E}\left[e^{sX+tY}\right]$$

  X and Y are independent if and only if

$$M_{X,Y}(s,t) = M_X(s) M_Y(t)$$

Moment-generating functions: Limitations and Examples

The major limitation of the moment-generating function is that it may not exist. In this case, the characteristic function may be used:

$$\phi(t) = \mathbb{E}\left[e^{itX}\right]$$

Examples:

- $\mathcal{N}\left(\mu, \sigma^2\right)$:

$$M(t) = e^{\mu t} e^{\sigma^2 t^2/2}$$

- Gamma(α, β):

$$\begin{aligned} M(t) &= \int_0^\infty e^{tx} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx \\ &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x(1-t/\beta)} dx \qquad [y := x(1-t/\beta)] \\ &= \frac{1}{(1-t/\beta)^\alpha} \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} dx \\ &= \frac{1}{(1-t/\beta)^\alpha}, \quad t < \beta. \end{aligned}$$

Inequalities

- Chebyshev inequality: If X is a non-negative random variable, then for any $a > 0$

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

- Cauchy-Schwarz inequality: If X and Y have finite variances, then

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]}$$

- Jensen Inequality:
  - Recall that a function g on $\mathbb{R}$ is said to be **convex**, if for any $x, y \in \mathbb{R}$ and any $0 \leq \lambda \leq 1$,
  $$g(\lambda x + (1-\lambda)y) \geq \lambda g(x) + (1-\lambda) g(y).$$
  E.g., $g(x) = x^2$ or $g(x) = |x|$.
  - If g is convex, then $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$

Law of Large Numbers

The LLN says that the mean of a large sample is close to the mean of the distribution.

Theorem 1.4.20 (The Weak Law of Large Numbers)

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\boxed{\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mu} \quad \text{as } n \to \infty$$

The notation $\xrightarrow{\mathbb{P}}$ means **convergence in probability**, whose more precise definition is as follows: for every $\epsilon > 0$,

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| > \epsilon\right) \to 0 \quad \text{as } n \to \infty$$

.

Central Limit Theorem

The CLT says that $\bar{X}_n$ has a distribution which is approximately Normal with mean μ and variance σ^2/n. This is remarkable since nothing is assumed about the distribution of X_i, except the existence of the mean and variance.

Theorem 1.4.21 (The Central Limit Theorem)

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Then

$$\boxed{Z_n \equiv \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim N(0,1)} \quad \text{as } n \to \infty$$

The notation $\xrightarrow{\mathcal{D}}$ means **convergence in distribution**, and it holds if and only if

$$\mathbb{P}\left(a \le Z_n \le b\right) \to \mathbb{P}(a \le Z \le b) \quad \text{as } n \to \infty$$

for every a and b.

Overview

Section 5

Random Vectors

Random Vector

- Let $\mathbf{X} = (X_1, \ldots, X_n)^T$ denote an n-dimensional **random vector** if its components $X_1, \ldots, X_n$ are one-dimensional random variables.
- The **space** of of this random vector is the set of ordered n-tuples

$$\mathcal{D} = \{(x_1, x_2, \cdots, x_n) : x_1 = X_1(\omega), \cdots, x_n = X_n(\omega),\ \omega \in \mathcal{C}\}.$$

Notation: $\mathbf{y}^T$, or $\mathbf{y}'$, denotes the transpose of $\mathbf{y}$, where $\mathbf{y}$ can be a matrix or a vector.
We denote $(X_1, \ldots, X_n)$ by the n-dimensional column vector $\mathbf{X}$ and the observed values $(x_1, \ldots, x_n)$ of the random vector by $\mathbf{x}$.

Joint CDF of a Random Vector

- The joint cumulative distribution function of a random vector $\mathbf{X}$ is defined as

$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) =& \mathbb{P}(X_1 \le x_1, \ldots, X_n \le x_n) \\ =& \mathbb{P}(\{\omega \in \mathcal{C} : X_1(\omega) \le x_1, \ldots, X_n(\omega) \le x_n\}), \end{aligned}$$

where

$$\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

For simplicity, the continuous random vectors are taken as examples for the following text. As for the discrete case, the analog is simple.

Continuous Random Vectors and Joint pdf

A random vector $\mathbf{X}$ is **continuous**, if it has a **joint probability density function** $f_{\mathbf{X}}$, that is, for every $A \subset \mathbb{R}^n$,

$$\mathbb{P}(\mathbf{X} \in A) = \int \cdots \int_A f_{\mathbf{X}}(x_1, \ldots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n,$$

where the density is a function satisfying

$$f_{\mathbf{X}}(\mathbf{x}) \ge 0 \quad \text{for every } \mathbf{x} \in \mathbb{R}^n$$

and

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \ldots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n = 1.$$

Marginal Densities

If a vector $\mathbf{X}$ has density $f_{\mathbf{X}}$, all its components X_i, the vectors of the pairs $(X_i, X_j)^T$, triples $(X_i, X_j, X_k)^T$, etc., have their own **marginal densities**.

Example 1.5.22

We consider the case $n = 3$. Then the marginal densities are obtained as follows:

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})\mathrm{d}x_2\mathrm{d}x_3, \quad f_{X_1,X_2}(x_1, x_2) = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})\mathrm{d}x_3$$

$f_{X_2}(x_2)$ is obtained by integrating $f_{\mathbf{X}}(\mathbf{x})$ with respect to x_1 and x_3, f_{X_1,X_3} by integrating $f_{\mathbf{X}}(\mathbf{x})$ with respect to x_2, etc.

Mean Vector and Covariance Matrix

- Consider an n-dimensional random vector $\mathbf{X} = (X_1, X_2, \cdots, X_n)^T$. The **mean vector** of $\mathbf{X}$ is

$$\mathbb{E}(\mathbf{X}) = (\mathbb{E}X_1, \mathbb{E}X_2, \cdots, \mathbb{E}X_n)^T$$

- The **covariance matrix** of $\mathbf{X}$ is defined as

$$\begin{aligned} \mathrm{Var}(\mathbf{X}) &= \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T\right] \\ &= \begin{pmatrix} \mathrm{Var}\,(X_1) & \cdots & \mathrm{Cov}\,(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}\,(X_n, X_1) & \cdots & \mathrm{Var}\,(X_n) \end{pmatrix}, \end{aligned}$$

where

$$\begin{aligned} \mathrm{Cov}\,(X_i, X_j) &= \mathbb{E}\left[(X_i - \mu_{X_i})\,(X_j - \mu_{X_j})\right] \\ &= \mathbb{E}\,(X_i X_j) - \mu_{X_i}\mu_{X_j} \end{aligned}$$

is the covariance of X_i and X_j. Notice that $\mathrm{Cov}\,(X_i, X_i) = \sigma^2_{X_i}$.

Mean vector and covariance matrix under linear transform

Theorem 1.5.23

Let $\mathbf{X}$ be an n-dimensional random vector. Suppose $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{b} \in \mathbb{R}^m$. Then

$$\mathbb{E}(\mathbf{AX} + \mathbf{b}) = \mathbf{A}\mathbb{E}(\mathbf{X}) + \mathbf{b} \quad \textit{and} \quad \mathrm{Var}\,(\mathbf{AX} + \mathbf{b}) = \mathbf{A}\,\mathrm{Var}\,(\mathbf{X})\,\mathbf{A}^T.$$

Covariance matrix is positive semi-definite

Theorem 1.5.24

Let $\mathbf{X}$ be a random vector. Then $Var(\mathbf{X})$ is symmetric and positive semi-definite.

Proof.

- $\mathrm{Var}(\mathbf{X})$ is obviously symmetric.
- For any $\mathbf{c} \in \mathbb{R}^n$, define $Y := \mathbf{c}^T\mathbf{X}$, which is a random variable. Then

$$\begin{aligned} 0 \le \mathrm{Var}(Y) = \mathrm{Var}\left(\mathbf{c}^T\mathbf{X}\right) &= \mathbb{E}\left[\left(\mathbf{c}^T\mathbf{X} - \mathbb{E}\mathbf{c}^T\mathbf{X}\right)^2\right] \\ &= \mathbb{E}\left[\left(\mathbf{c}^T\mathbf{X} - \mathbb{E}\mathbf{c}^T\mathbf{X}\right)\left(\mathbf{c}^T\mathbf{X} - \mathbb{E}\mathbf{c}^T\mathbf{X}\right)^T\right] \\ &= \mathbf{c}^T\mathbb{E}\left[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^T\right]\mathbf{c} \\ &= \mathbf{c}^T\,\mathrm{Var}(\mathbf{X})\mathbf{c}, \end{aligned}$$

showing that $\mathrm{Var}(\mathbf{X})$ is positive semi-definite.

Independence for multiple events or RVs

- The definition of independence can be extended to an arbitrary finite number of events and random variables.
- The events $A_1, \ldots, A_n$ are independent if, for every choice of indices $1 \le i_1 < \cdots < i_k \le n$ and integers $1 \le k \le n$

$$\mathbb{P}\,(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}\,(A_{i_1}) \cdots \mathbb{P}\,(A_{i_k}).$$

- The random variables $X_1, \ldots, X_n$ are independent if, for every choice of indices $1 \le i_1 < \cdots < i_k \le n$, integers $1 \le k \le n$ and all subsets $B_1, \ldots, B_n$ of $\mathbb{R}$,

$$\mathbb{P}\,(X_{i_1} \in B_{i_1}, \ldots, X_{i_k} \in B_{i_k}) = \mathbb{P}\,(X_{i_1} \in B_{i_1}) \cdots \mathbb{P}\,(X_{i_k} \in B_{i_k})$$

This means that the events $\{X_1 \in B_1\}, \ldots, \{X_n \in B_n\}$ are independent.

  - Notice that independence of the components of a random vector implies the independence of each pair of its components, but the converse is in general not true.

Random vector with independent components

- The random variables $X_1, \ldots, X_n$ are independent if and only if their joint CDF can be written as follows:

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = F_{X_1}(x_1)\cdots F_{X_n}(x_n), \quad (x_1,\ldots,x_n) \in \mathbb{R}^n$$

- If the random vector $\mathbf{X} = (X_1,\ldots,X_n)^T$ has density $f_{\mathbf{X}}$, then $X_1,\ldots,X_n$ are independent if and only if

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_{X_1}(x_1)\cdots f_{X_n}(x_n), \quad (x_1,\ldots,x_n) \in \mathbb{R}^n.$$

- If the random variable X_i has the marginal mgf $M(0,\ldots,0,t_i,0,\ldots,0)$, then X_1, ..., X_n are mutually independent if and only if

$$M(t_1,t_2,\ldots,t_n) = \prod_{i=1}^{n} M(0,\ldots,0,t_i,0,\ldots,0).$$

Properties of independent RVs

An important consequence of the independence of random variables is the following property:

Theorem 1.5.25

If $X_1,\ldots,X_n$ are independent, then for any real-valued functions $g_1,\ldots,g_n$, the random variables $g_1(X_1),\ldots,g_n(X_n)$ are again independent and moreover,

$$\mathbb{E}\left[g_1(X_1)\cdots g_n(X_n)\right] = \mathbb{E}g_1(X_1)\cdots\mathbb{E}g_n(X_n),$$

provided the considered expectations are well defined.

Overview

Section 6

Normal Random Vectors

Standard normal random vector

- Consider $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$, where $Z_1, \ldots, Z_n$ are i.i.d. $N(0,1)$ random variables. Then the density of $\mathbf{Z}$ is

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z_i^2\right\} = \left(\frac{1}{2\pi}\right)^{n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} z_i^2\right\}$$
$$= \left(\frac{1}{2\pi}\right)^{n/2} \exp\left\{-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right\}$$

  for $\mathbf{z} \in \mathbb{R}^n$.
- Obviously,
$$E[\mathbf{Z}] = \mathbf{0} \text{ and } \mathrm{Cov}[\mathbf{Z}] = \mathbf{I}_n,$$
  where $\mathbf{I}_n$ denotes the identity matrix of order n.
- We call $\mathbf{Z}$ an n-dimensional standard normal random vector and write
$$\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_n)$$

Centered normal random vector

Definition 1.6.26

A random vector $\mathbf{X} = (X_1, \ldots, X_n)^T$ is called a centered normal random vector if there exists a deterministic $n \times \ell$ matrix $\mathbf{A}$ such that

$$\mathbf{X} = \mathbf{A}\mathbf{Z},$$

where $\mathbf{Z}$ is a standard normal random vector with ℓ components.

Normal random vector

Definition 1.6.27

A random vector $\mathbf{X} = (X_1, \ldots, X_n)^T$ is called a normal random vector if there exists an ℓ-dimensional standard normal random vector $\mathbf{Z}$, an n-vector $\boldsymbol{\mu}$, and an $n \times \ell$ matrix $\mathbf{A}$, such that

$$\mathbf{X} = \mathbf{A}\mathbf{Z} + \boldsymbol{\mu}.$$

In this case we write

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma), \quad \text{with} \quad \Sigma = \mathbf{A}\mathbf{A}^{\mathrm{T}}.$$

Formally:

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma) \quad \Longleftrightarrow \quad \exists\, \boldsymbol{\mu} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n\times\ell}, \mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_\ell) \text{ such that}$$
$$\mathbf{X} = \mathbf{A}\mathbf{Z} + \boldsymbol{\mu} \quad \text{and} \quad \Sigma = \mathbf{A}\mathbf{A}^T$$

- If $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, then it's easy to see that
$$\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}, \quad \mathrm{Cov}(\mathbf{X}) = \Sigma.$$

Density function: non-degenerate case

Theorem 1.6.28

Suppose $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ is positive definite. Then $\mathbf{X}$ has probability density function given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \quad \mathbf{x} \in \mathbb{R}^n.$$

Proof: We only show the 2-dimensional case, the general case is similar. By results from linear algebra, $\exists\ \mathbf{A} \in \mathbb{R}^{2\times 2}$ such that

$$\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T.$$

Thus for $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_2)$, it follows that

$$\mathbf{A}\mathbf{Z} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \Sigma).$$

Without loss of generality, assume that

$$\mathbf{X} = \mathbf{A}\mathbf{Z} + \boldsymbol{\mu}.$$

Proof cont'd

Consider the map $g : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$g(\mathbf{z}) = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}, \quad \mathbf{z} \in \mathbb{R}^2.$$

Let $h = g^{-1}$ be the inverse map. Note that

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu} \iff \mathbf{z} = \mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

So $h(\mathbf{x}) = \mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu})$. It follows that the Jacobian of h is

$$J = \begin{pmatrix} \frac{\partial h_1}{\partial x_1} & \frac{\partial h_1}{\partial x_2} \\ \frac{\partial h_2}{\partial x_1} & \frac{\partial h_2}{\partial x_2} \end{pmatrix} = \mathbf{A}^{-1}.$$

Since $\mathbf{Z}$ has density

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{2\pi} \exp\left\{ -\frac{1}{2}\mathbf{z}^T\mathbf{z} \right\},$$

it follows that $\mathbf{X}$ has density

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2\pi} \exp\left\{ -\frac{1}{2} h(\mathbf{x})^T h(\mathbf{x}) \right\} |J|.$$

Proof cont'd

It remains to note that

$$\begin{aligned} h(\mathbf{x})^T h(\mathbf{x}) &= \left(\mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)^T \mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \\ &= (\mathbf{x} - \boldsymbol{\mu})^T \left(\mathbf{A}^{-1}\right)^T \mathbf{A}^{-1}(\mathbf{x} - \mu) \\ &= (\mathbf{x} - \mu)^T \left(\mathbf{A}\mathbf{A}^T\right)^{-1} (\mathbf{x} - \mu) \\ &= (\mathbf{x} - \mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu). \end{aligned}$$

and

$$|J| = |\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|} = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}},$$

where the last equality follows from

$$|\boldsymbol{\Sigma}| = |\mathbf{A}\mathbf{A}^T| = |\mathbf{A}|^2.$$

Density function: bivariate case

Let $(X, Y)^T$ be a 2-dimensional normal random vector with parameters

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix},$$

where ρ is the correlation between X and Y.

Theorem 1.6.29

Suppose $\boldsymbol{\Sigma}$ is non-degenerate, i.e., $\sigma_X > 0$, $\sigma_Y > 0$ and $|\rho| \neq 1$. Then $(X, Y)^T$ has density

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 \right] \right)$$

Proof

It's easy to see that

$$|\boldsymbol{\Sigma}| = \sigma_X^2\sigma_Y^2\left(1-\rho^2\right) \text{ and } \boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)} \begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix}.$$

So, for $\mathbf{z} = (x, y)^T$,

$$\begin{aligned} &(\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}) \\ &= \frac{1}{\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)} (x-\mu_X, y-\mu_Y)^T \begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix} \begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix} \\ &= \frac{1}{\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)} \left[\sigma_Y^2\left(x-\mu_X\right)^2 - 2\rho\sigma_X\sigma_Y\left(x-\mu_X\right)\left(y-\mu_Y\right) + \sigma_X^2\left(y-\mu_Y\right)^2 \right] \\ &= \frac{1}{\left(1-\rho^2\right)} \left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 \right]. \end{aligned}$$

The assertion now follows from the previous theorem.

Decomposition into independent components

If the covariance matrix of a normal random vector $\mathbf{X}$ is of block diagonal form, i.e.,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & 0 \\ 0 & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \qquad (1)$$

where $\boldsymbol{\Sigma}_{11}$ is an $m \times m$ matrix with $m < n$, then we can write

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix},$$

where $\mathbf{X}_1 := (X_1, \ldots, X_m)^T$, $\mathbf{X}_2 := (X_{m+1}, \ldots, X_n)^T$ and $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ are similarly defined.

Theorem 1.6.30

Let $\boldsymbol{\Sigma}$ be of block diagonal form as in (1). Then $\mathbf{X}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii})$, $i = 1, 2$, and $\mathbf{X}_1, \mathbf{X}_2$ are independent.

Proof

Since $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ are both positive semi-definite, there exist $m \times m$ matrix $\mathbf{A}_1$ and $(n-m) \times (n-m)$ matrix $\mathbf{A}_2$ such that

$$\boldsymbol{\Sigma}_{11} = \mathbf{A}_1 \mathbf{A}_1^T, \quad \boldsymbol{\Sigma}_{22} = \mathbf{A}_2 \mathbf{A}_2^T.$$

Then

$$\mathbf{A} := \begin{pmatrix} \mathbf{A}_1 & 0 \\ 0 & \mathbf{A}_2 \end{pmatrix} \quad \text{satisfies} \quad \mathbf{A}\mathbf{A}^T = \begin{pmatrix} \mathbf{A}_1 \mathbf{A}_1^T & 0 \\ 0 & \mathbf{A}_2 \mathbf{A}_2^T \end{pmatrix} = \boldsymbol{\Sigma}.$$

We can find i.i.d. $Z_i \sim N(0,1)$ with

$$\mathbf{Z}_1 := (Z_1, \ldots, Z_m)^T, \quad \mathbf{Z}_2 := (Z_{m+1}, \ldots, Z_n)^T$$

such that

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}_1 & 0 \\ 0 & \mathbf{A}_2 \end{pmatrix} \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}_1 \mathbf{Z}_1 + \boldsymbol{\mu}_1 \\ \mathbf{A}_2 \mathbf{Z}_2 + \boldsymbol{\mu}_2 \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

since $\mathbf{A}\mathbf{A}^T = \boldsymbol{\Sigma}$. Thus

$$\mathbf{X}_i = \mathbf{A}_i \mathbf{Z}_i + \boldsymbol{\mu}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii})$$

and

$$\mathbf{Z}_1 \perp\!\!\!\perp \mathbf{Z}_2 \implies \mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2. \quad [\text{"} \perp\!\!\!\perp \text{" means independence}]$$

Linear transform of a normal random vector

Theorem 1.6.31

Let $\mathbf{c}$ be a $m \times n$ matrix and $\mathbf{d} \in \mathbb{R}^m$. If

$$\mathbf{X} = (X_1, \cdots, X_n)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

then

$$\mathbf{c}\mathbf{X} + \mathbf{d} \sim N(\mathbf{c}\boldsymbol{\mu} + \mathbf{d}, \mathbf{c}\boldsymbol{\Sigma}\mathbf{c}^T).$$

Linear transform of a normal random vector cont'd

Corollary 1.6.32

Suppose $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{c} \in \mathbb{R}^n$. Then

$$\mathbf{c}^T \mathbf{X} \sim N(\mathbf{c}^T \boldsymbol{\mu}, \mathbf{c}^T \boldsymbol{\Sigma} \mathbf{c}).$$

- In other words, any linear combination of the one-dimensional components of a normal random vector is again normal.

Example

Let $X_1, \cdots, X_n$ be i.i.d. random variables each having a normal distribution with mean μ and variance σ^2. Then

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Example

Suppose $\mathbf{X} = (X_1, \cdots, X_n)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

$$X_i \sim N(\mu_i, \sigma_{ii}), \quad i = 1, \ldots, n.$$

Overview

Section 7

Distributions Derived from the Normal Distribution: χ^2, t, F Distributions

Definition of Chi-square distribution

The random variable X has a chi-square distribution with n degrees of freedom if X has the same distribution as

$$\sum_{i=1}^{n} Z_i^2,$$

where $Z_1, \ldots, Z_n$ are independent standard normal random variables.

- Note that Z_i^2 has mgf $(1-2t)^{-1/2}$. In fact, for $t < \frac{1}{2}$,

$$\begin{aligned}
\mathbb{E}(e^{tZ_i^2}) &= \int e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
&= \int \frac{1}{\sqrt{2\pi}} e^{-(1-2t)z^2/2} dz \qquad [\tilde{z} := \sqrt{1-2t}z] \\
&= \left(\sqrt{1-2t}\right)^{-1} \int \frac{1}{\sqrt{2\pi}} e^{-\tilde{z}^2} d\tilde{z} = \left(\sqrt{1-2t}\right)^{-1}.
\end{aligned}$$

The mgf of Chi-square distribution

It follows that the mgf of X is

$$\begin{aligned}
M(t) &= \mathbb{E}(e^{t\sum_{i=1}^n Z_i^2}) = \mathbb{E}(e^{tZ_1^2} \cdots e^{tZ_n^2}) \\
&= \mathbb{E}(e^{tZ_1^2}) \cdots \mathbb{E}(e^{tZ_n^2}) = (1-2t)^{-n/2}, \quad t < \frac{1}{2}.
\end{aligned} \tag{2}$$

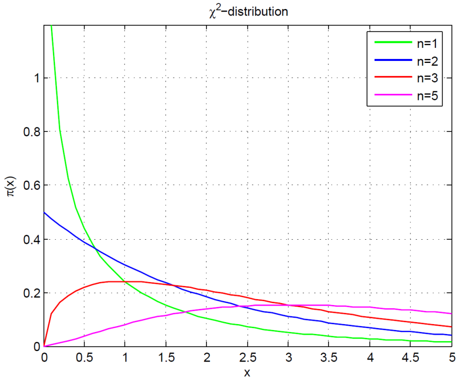
- Hence,

$$\mathbb{E}X = M'(0) = n$$

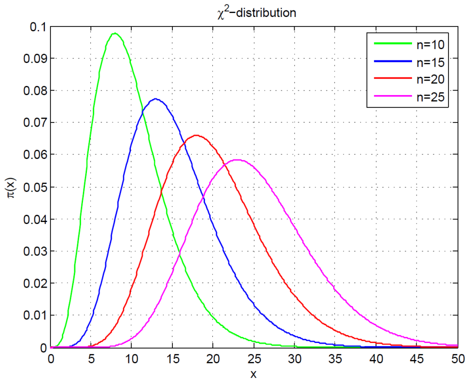
and

$$\mathrm{Var}\, X = M''(0) - n^2 = 2n.$$

Graph of the χ_n^2 PDF: small n



Graph of the χ_n^2 PDF: large n



- CLT: χ_n^2 converges to a normal distribution as $n \to \infty$
- $\chi_n^2 \to \mathcal{N}(n, 2n)$, as $n \to \infty$
- When $n > 50$, for many practical purposes, $\chi_n^2 \approx N(n, 2n)$

Gamma distribution

Recall that a Gamma distribution has a pdf with two parameters $\alpha > 0$ and $\beta > 0$:

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & 0 < x < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

This distribution is usually denoted by $\mathrm{Gamma}(\alpha, \beta)$.

Equivalent definition as a Gamma distribution

- The mgf of a gamma distribution is

$$M(t) = \frac{1}{(1 - t/\beta)^\alpha}, \quad t < \beta.$$

- Compared with (2), we see that the chi-square distribution with n degrees of freedom is identical to the gamma distribution with parameters $(n/2, 1/2)$.

Properties of Chi-square distribution

Theorem 1.7.33

Suppose X_1, X_2 are independent and $X_1 \sim \chi^2_n, X_2 \sim \chi^2_m$. Then

$$X_1 + X_2 \sim \chi^2_{n+m}.$$

Proof: It suffices to show that $X_1 + X_2$ has mgf $(1-2t)^{-(m+n)/2}$. But, by independence,

$$\begin{aligned} \mathbb{E}(e^{t(X_1+X_2)}) &= \mathbb{E}(e^{tX_1} e^{tX_2}) \\ &= \mathbb{E}(e^{tX_1})\mathbb{E}(e^{tX_2}) \\ &= (1-2t)^{-n/2}(1-2t)^{-m/2} \\ &= (1-2t)^{-(m+n)/2}, \quad t < \frac{1}{2}. \end{aligned}$$

The theorem is proved.

The t-distribution

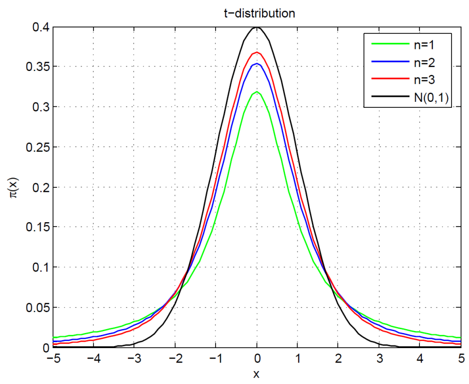
If Z is a standard normal random variable and V is a chi-square random variable with r degrees of freedom, and Z and V are independent, then

$$T = \frac{Z}{\sqrt{V/r}}$$

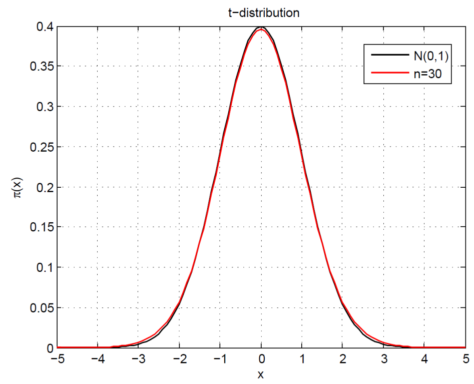
is a random variable following a t-distribution with r degrees of freedom. Its pdf is

$$f_T(t) = \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r}\Gamma(r/2)} \frac{1}{(1 + t^2/r)^{(r+1)/2}}, \quad -\infty < t < \infty.$$

Graph of the t-distribution PDF: small n



Graph of the t-distribution PDF: large n



The F-distribution

If U is a $\chi^2(r_1)$ random variable and V is a $\chi^2(r_2)$ random variable, and U and V are independent, then

$$F = \frac{U/r_1}{V/r_2}$$

is a random variable following an F-distribution with r_1 and r_2 degrees of freedom. Its pdf is

$$f_F(x) = \begin{cases} \frac{\Gamma[(r_1+r_2)/2](r_1/r_2)^{r_1/2}}{\Gamma(r_1/2)\Gamma(r_2/2)} \frac{x_1^{r_1/2-1}}{(1+r_1x/r_2)^{(r_1+r_2)/2}} & 0 < x < \infty \\ 0 & \text{elsewhere.} \end{cases}$$

Student's Theorem

Theorem 1.7.34

Let $X_1, \cdots, X_n$ be i.i.d random variables with $X_i \sim N(\mu, \sigma^2)$. Define

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Then

1. $\bar{X}$ *has a* $N\left(\mu, \frac{\sigma^2}{n}\right)$ *distribution.*
2. $\bar{X}$ *and* S^2 *are independent.*
3. $(n-1)S^2/\sigma^2$ *has a* $\chi^2(n-1)$ *distribution.*
4. *The random variable*
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$
has a t-distribution with $n-1$ degrees of freedom.

Proof: (1) and (2)

We have already shown (1) earlier. Let's turn to (2). Note that for each i

$$\begin{pmatrix} \bar{X} \\ X_i - \bar{X} \end{pmatrix}$$

is a linear transform of $(X_1, \ldots, X_n)^T$, so it is a 2-dimensional normal random vector. But

$$\begin{aligned} \operatorname{Cov}\left(\bar{X}, X_i - \bar{X}\right) &= \operatorname{Cov}\left(\bar{X}, X_i\right) - \operatorname{Cov}(\bar{X}, \bar{X}) \\ &= \operatorname{Cov}\left(\frac{1}{n} X_i, X_i\right) - \operatorname{Var}(\bar{X}) \\ &= \frac{1}{n} - \frac{1}{n} = 0 \end{aligned}$$

Therefore, $\bar{X}$ and $X_i - \bar{X}$ are independent for all i. Because S^2 is a function of $X_i - \bar{X}, i = 1, \cdots, n$, it follows that S^2 is independent of $\bar{X}$.

Proof: (3) and (4)

We first note that

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_n^2$$

Also,

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left[(X_i - \bar{X}) + (\bar{X} - \mu)\right]^2$$

Expanding the square and using the fact that $\sum_{i=1}^{n} \left(X_i - \bar{X}\right) = 0$, we obtain

$$W := \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 =: U + V$$

This is a relation of the form $W = U + V$. Independence of U and V implies $M_W(t) = M_U(t) M_V(t)$. W and V both follow chi-square distributions, so

$$M_U(t) = \frac{M_W(t)}{M_V(t)} = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-(n-1)/2},$$

which is mgf of a χ_{n-1}^2 distribution. So (3) is true. The assertion (4) now follows from (2) and (3) .