Probability for machine learning

Rokhlin D.B.

IMMCS, SFU, 2021

Probability space $(\Omega, \mathscr{F}, \mathsf{P})$

- Ω: sample space
- $\mathscr{F}$: σ-algebra of events (subsets of Ω)
- P: probability measure

$\mathscr{F}$ satisfies the axioms

- $\emptyset \in \mathscr{F}$;
- the condition $A \in \mathscr{F}$ implies that $A^c := \Omega \backslash A \in \mathscr{F}$;
- the condition $A_i \in \mathscr{F}$, $i \in \mathbb{Z}_+$ implies that $\cup_{i=1}^{\infty} A_i \in \mathscr{F}$.

A function $\mathsf{P} : \mathscr{F} \mapsto [0, 1]$ is called a probability measure, if

- $\mathsf{P}(\emptyset) = 0$, $\mathsf{P}(\Omega) = 1$;
- for any sequence of pairwise *disjoint* sets $A_i \in \mathscr{F}$, $i = 1, \ldots, \infty$ we have the equality

$$\mathsf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathsf{P}(A_i).$$

Examples

- $\Omega = \{H, T\}$, $\mathscr{F} = \{\varnothing, \Omega, \{H\}, \{T\}\} = 2^{\Omega}$,

$$\mathsf{P}(\{H\}) = p \in [0, 1], \quad \mathsf{P}(\{T\}) = 1 - p.$$

- $\Omega = [0, 1]$, $\mathscr{F} = \mathscr{B}(0, 1)$, $\mathsf{P}((a, b)) = b - a$.
- $\Omega = \mathbb{R}$, $\mathscr{F} = \mathscr{B}(\mathbb{R})$, $\mathsf{P}(a, b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx$.

Random variables

A function $\xi : \Omega \mapsto \mathbb{R}^d$ is called $\mathscr{F}$-measurable if

$$\xi^{-1}(B) = \{\omega : \xi(\omega) \in B\} \in \mathscr{F}.$$

Such functions are called random variables.

- $\Omega = \{(a_1, a_2) : a_i \in \{H, T\}\}$, $\mathscr{F}_1 = \sigma(\{H\}, \{T\})$,

$$\xi_1 = I_{\{a_1 = H\}}, \quad \xi_2 = I_{\{a_2 = H\}},$$

$$I_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}$$

ξ_1 is $\mathscr{F}_1$-measurable, ξ_2 is not $\mathscr{F}_1$-measurable.

- $(\Omega, \mathscr{F}) = ([0,1], (\varnothing, [0,1]))$.
  $\xi : [0,1] \mapsto \mathbb{R}$ is $\mathscr{F}$-measurable $\iff$ $\xi = \text{const}$.
- $(\Omega, \mathscr{F}) = ([0,1], \mathscr{B}(0,1))$.
  $\xi : [0,1] \mapsto \mathbb{R}$ is $\mathscr{F}$-measurable $\iff$ ξ is Borel.
- $\Omega = \{(a_1, \ldots, a_m) : a_i \in \{H, T\}\}$,
  $\mathscr{F} = \sigma\{\{H,H\}, \{H,T\}, \{T,H\}, \{T,T\}\}$.
  $\xi : \Omega \mapsto \mathbb{R}$ is $\mathscr{F}$-measurable $\iff$ ξ does not depend on $a_3, \ldots, a_m$.

Integral

Simple functions:

$$\xi(\omega) = \sum_{j=1}^{n} c_j I_{A_j}(\omega), \quad A_j \in \mathscr{F}, \quad c_j \in \mathbb{R}$$

Clearly, such functions are $\mathscr{F}$-measurable. Integral of a simple function:

$$\int_\Omega \xi(\omega)\,\mathsf{P}(d\omega) = \int_\Omega \xi\,d\mathsf{P} := \sum_{j=1}^{n} c_j \mu(A_j).$$

Integral of a non-negative measurable function:

$$\int_\Omega \xi\,d\mathsf{P} = \sup\left\{\int_\Omega \varphi\,d\mathsf{P},\ \text{where } \varphi \text{ simple and } 0 \le \varphi \le \xi\right\}.$$

Infinite values are acceptable.

Furthermore, any $\mathscr{F}$-measurable function can be represented as a difference of two nonnegative $\mathscr{F}$-measurable functions: $\xi = \xi^+ - \xi^-$, where

$$\xi^+(\omega) := \max\{\xi, 0\}, \qquad \xi^-(\omega) := -\max\{-\xi(\omega), 0\}.$$

If both values are $\int_\Omega \xi^+ \, d\mathsf{P}$, $\int_\Omega \xi^- \, d\mathsf{P}$ are finite, then ξ is called *integrable* and the integral is defined by the formula

$$\int_\Omega \xi \, d\mu = \int_\Omega \xi^+ \, d\mu - \int_\Omega \xi^- \, d\mu.$$

- expectation: $\mathsf{E}\xi = \int_\Omega \xi \, d\mathsf{P}$,
- variance: $\mathrm{Var}(\xi) = \mathsf{E}(\xi - \mathsf{E}\xi)^2$,
- standard deviation: $\sqrt{\mathrm{Var}(\xi)}$.

Distribution

An random variable $\xi : \Omega \mapsto \mathbb{R}^d$ induces the probability space $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d), \mathsf{P}_\xi)$, where P_ξ is the image measure:

$$\mathsf{P}_\xi(B) = \mathsf{P}(\xi^{-1}(B)), \quad B \in \mathscr{B}(\mathbb{R}^d),$$

which is called the distribution of ξ.

- ξ is called discrete if there exist a finite or countable set Z and a probability mass function (p.m.f.) $p_\xi : Z \mapsto [0, 1]$ such that

$$\mathsf{P}_\xi(B) = \sum_{z_i \in Z \cap B} p_\xi(z_i), \quad B \in \mathscr{B}(\mathbb{R}^d).$$

- ξ is called continuous if there exists a Borel function $p_\xi : \mathbb{R}^d \mapsto \mathbb{R}_+$ such that

$$\mathsf{P}_\xi(B) = \int_B p_\xi(x)\, dx, \quad B \in \mathscr{B}(\mathbb{R}^d).$$

Change of variables formula

Let g be a (bounded) Borel function. Then

$$\mathsf{E}g(\xi) = \int_{\mathbb{R}} g(x)\,\mathsf{P}_\xi(dx).$$

- for discrete ξ: $\mathsf{E}g(\xi) = \sum_{x\in Z} g(x)p_\xi(x)$,
- for continuous ξ: $\mathsf{E}g(\xi) = \int_{\mathbb{R}} g(x)f_\xi(x)\,dx$.

In particular,

- for discrete ξ: $\mathsf{E}\xi = \sum_{x\in Z} xp_\xi(x)$,
- for continuous ξ: $\mathsf{E}\xi = \int_{\mathbb{R}} xf_\xi(x)\,dx$.

Examples

- Bernoulli distribution: $\xi \sim Ber(p)$, $Z = \{0, 1\}$, $p_\xi(0) = 1$, $p_\xi(1) = 1 - p$,

$$\mathsf{E}\xi = 0 \cdot p_\xi(0) + 1 \cdot p_\xi(1) = p,$$

$$\mathsf{E}(\xi^2) = 0 \cdot (1 - p) + 1 \cdot p = p,$$

$$\mathrm{Var}(\xi) = \mathsf{E}(\xi^2) - (\mathsf{E}\xi)^2 = p - p^2 = p(1 - p).$$

- Poisson distribution: $\xi \sim \Pi(\lambda)$, $Z = \mathbb{Z}_+$,

$$p_\xi(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \in \mathbb{Z}_+.$$

$$\mathsf{E}\xi = \sum_{k=0}^{\infty} k p_\xi(k) = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = \lambda \sum_{k=0}^{\infty} p_\xi(k) = \lambda,$$

$$\mathrm{Var}(\xi) = \mathsf{E}\xi^2 - (\mathsf{E}\xi)^2 = \sum_{k=0}^{\infty} k^2 p_\xi(k) - (\mathsf{E}\xi)^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda.$$

- Uniform distribution: $\xi \sim U(a,b)$, $a < b$,

$$f_\xi(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b], \\ 0, & x \notin [a,b]. \end{cases}$$

$$\mathsf{E}\xi = \frac{a+b}{2}, \quad \mathrm{Var}(\xi) = \frac{(b-a)^2}{12}.$$

- Gaussian (normal) distribution: $\xi \sim N(0,1)$,

$$p_\xi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad x \in \mathbb{R}.$$

$$\mathsf{E}\xi = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2}\, dx = 0.$$

$$\mathrm{Var}(\xi) = \mathsf{E}\xi^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\, dx = -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x\, de^{-x^2/2}$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = 1.$$

Distribution functions

$F_\xi(x) = \mathsf{P}(\xi \le x)$ is called a cumulative distribution function (cdf) of ξ.

$$F_\xi(x) = \begin{cases} \sum_{y \le x} p_\xi(y), & \xi \text{ discrete}, \\ \int_{-\infty}^{x} f_\xi(y)\, dy, & \xi \text{ continuous}. \end{cases}$$

$$F_\xi(x) \le F_\xi(y), \quad \text{if } x \le y,$$

$$\lim_{x \to -\infty} F_\xi(x) = 0, \quad \lim_{x \to +\infty} F_\xi(x) = 1,$$

Moreover, the distribution functions are right-continuous:

$$\lim_{y \downarrow x} F_\xi(y) = \lim_{Q \ni y \downarrow x} \mathsf{P}(\xi \le y) = \mathsf{P}(\cap_{y \ge x, y \in \mathbb{Q}} \{\xi \le y\}) = \mathsf{P}(\xi \le x) = F(x).$$

The second equality uses the continuity of the probability measure.
For a continuous random variable $F'_\xi(x) = f_\xi(x)$.

Let $\xi \sim N(0,1)$, then

$$\Phi(x) := F_\xi(x) = \mathsf{P}(\xi \le x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy.$$

"Three-sigma" rule. $\Phi(3) \approx 0.9987$. Let $\eta \in N(\mu, \sigma^2)$. Then

$$\xi = \frac{\eta - \mu}{\sigma} \in N(0,1).$$

$$\begin{aligned}
\mathsf{P}(|\eta - \mu| \le 3\sigma) &= \mathsf{P}(|\xi| \le 3) = \mathsf{P}(\xi \le 3) - \mathsf{P}(\xi \le -3) \\
&= \mathsf{P}(\xi \le 3) - (1 - \mathsf{P}(\xi \le 3)) = 2\mathsf{P}(\xi \le 3) - 1 \\
&= 2\Phi(3) - 1 \approx 0.997
\end{aligned}$$

Conditional probability

Let $\mathsf{P}(B) > 0$. The *conditional probability* of A given B is the number

$$\mathsf{P}(A|B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}.$$

This formula shows what portion of the probability attributed to B relates to A.

Example

Roll 2 symmetric 6-sided dices. The first die shows 5. Find the conditional probability that the sum of the obtained points will be 10.

$$\Omega = \{(a_1, a_2) : a_i \in \{1, \ldots, 6\}\}, \quad \mathscr{F} = 2^{\Omega},$$

$$B = \{\omega : a_1 = 5\}, \quad A = \{\omega : a_1 + a_2 = 10\},$$

$$\mathsf{P}(A|B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(\{\omega : a_1 = 5, a_2 = 5\})}{\mathsf{P}(\{\omega : a_1 = 5\})} = \frac{1/36}{1/6} = 1/6.$$

Multiplication rule

Let $\mathsf{P}(A_1 \cap \cdots \cap A_m) > 0$. The equality

$$\mathsf{P}(A_1 \cap \cdots \cap A_m) = \mathsf{P}(A_1)\mathsf{P}(A_2|A_1) \ldots \mathsf{P}(A_m|A_1 \cap \cdots \cap A_{m-1})$$

is called the multiplication rule.

Proof.

$(m = 3)$

$$\begin{aligned}\mathsf{P}(A_1 \cap A_2 \cap A_3) =& \mathsf{P}(A_3|A_1 \cap A_2)\mathsf{P}(A_1 \cap A_2) \\ =& \mathsf{P}(A_3|A_1 \cap A_2)\mathsf{P}(A_2|A_1)\mathsf{P}(A_1)\end{aligned}$$

Example

There is a 36-card deck. 3 cards are drawn at random. What is the probability that there will get no spades?

Let A_i be the event that the card drawn at step i is not a spade. We are interested in the probability of the event $A_1 \cap A_2 \cap A_3$. By the multiplication rule,

$$\mathrm{P}(A_1 \cap A_2 \cap A_3) = \mathrm{P}(A_1)\mathrm{P}(A_2|A_1)\mathrm{P}(A_3|A_1 \cap A_2).$$

Clearly,

$$\mathrm{P}(A_1) = \frac{27}{36}, \quad \mathrm{P}(A_2|A_1) = \frac{26}{35}, \quad \mathrm{P}(A_3|A_1 \cap A_2) = \frac{25}{34}.$$

$$\mathrm{P}(A_1 \cap A_2 \cap A_3) \approx 0.41.$$

Example

There is a 36-card deck. 3 cards are drawn at random. What is the probability that there will get no spades?

Let A_i be the event that the card drawn at step i is not a spade. We are interested in the probability of the event $A_1 \cap A_2 \cap A_3$. By the multiplication rule,

$$\mathsf{P}(A_1 \cap A_2 \cap A_3) = \mathsf{P}(A_1)\mathsf{P}(A_2|A_1)\mathsf{P}(A_3|A_1 \cap A_2).$$

Clearly,

$$\mathsf{P}(A_1) = \frac{27}{36}, \quad \mathsf{P}(A_2|A_1) = \frac{26}{35}, \quad \mathsf{P}(A_3|A_1 \cap A_2) = \frac{25}{34}.$$

$$\mathsf{P}(A_1 \cap A_2 \cap A_3) \approx 0.41.$$

Total probability formula

Consider the partition $\mathscr{D} = \{A_1, \ldots, A_m\}$. The equality

$$\mathsf{P}(B) = \sum_{i=1}^{m} \mathsf{P}(B|A_i)\mathsf{P}(A_i)$$

is called the *total probability formula*.

Proof.

$$\sum_{i=1}^{m} \mathsf{P}(B|A_i)\mathsf{P}(A_i) = \sum_{i=1}^{m} \mathsf{P}(B \cap A_i) = \mathsf{P}\left(B \bigcap \bigcup_{i=1}^{m} A_i\right) = \mathsf{P}(B).$$

Example

There are three urns, each of which can contain white and black balls:

- in the first urn there are 3 white and 2 black ones;
- in the second urn there are 0 white and 4 black;
- in the third urn there are 5 white and 2 black ones.

In the first step of the experiment, one of the urns is selected at random. In the second step, one ball is drawn from it at random. What is the probability to get a white ball?

Example (continued)

Let A_i be the event that the urn with the number i is selected, and B be the event that the white ball is drawn. The probabilities that B will occur under the condition A_i are easy to calculate since we know the number of white and black balls in the corresponding urn:

$$\mathsf{P}(B|A_1) = 3/5, \quad \mathsf{P}(B|A_2) = 0, \quad \mathsf{P}(B|A_3) = 5/7.$$

By the total probability formula:

$$\mathsf{P}(B) = \sum_{i=1}^{3} \mathsf{P}(B|A_i)\mathsf{P}(A_i) = \frac{3}{5} \cdot \frac{1}{3} + \frac{5}{7} \cdot \frac{1}{3} = \frac{46}{105} \approx 0.44.$$

Example (Monty Hall)

There is a prize behind one of the three closed doors, there is nothing behind the other two.

- The player selects one of the doors.
- The host opens one of the two remaining doors, behind which there is nothing.
- The player can keep or change his original choice.

Which strategy is better?

Example (continued)

If the player retains his original choice, then the probability that he will receive the prize is $1/3$.
Denote by A the event that the prize is behind the door originally chosen by the player. Let B be the event that a prize is won. If the player changes his original choice, then

$$\mathsf{P}(B|A) = 0, \quad \mathsf{P}(B|A^c) = 1.$$

By the total probability formula,

$$\mathsf{P}(B) = \mathsf{P}(B|A)\mathsf{P}(A) + \mathsf{P}(B|A^c)\mathsf{P}(A^c) = 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3}.$$

Thus, the second strategy is much better.

The Bayes formula

Consider a partition $\mathscr{D} = \{A_1, \ldots, A_m\}$. The equality

$$\mathsf{P}(A_i|B) = \frac{\mathsf{P}(B|A_i)\mathsf{P}(A_i)}{\sum_{j=1}^{m} \mathsf{P}(B|A_j)\mathsf{P}(A_j)}$$

is called *the Bayes formula*.

Proof.

$$\mathsf{P}(A_i|B) = \frac{\mathsf{P}(A_i \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(B|A_i)\mathsf{P}(A_i)}{\sum_{j=1}^{m} \mathsf{P}(B|A_j)\mathsf{P}(A_j)} \quad \square$$

Bayes' formula allows to take into account information concerning the occurrence of an event B to re-evaluate the probabilities of other events.

- $\mathsf{P}(A_i)$ — prior probabilities (before an experiment)
- $\mathsf{P}(A_i|B) = \mathsf{P}_B(A_i)$ — posterior probabilities (after the experiment)

Example

A patient has a rare desease with probability 0.001. A medical test gives the correct answer in 95% of cases. What is the probability that the patient is ill if the test showed the positive result?

- H the hypothesis that the patient is ill, $\mathscr{D} = \{H, H^c\}$
- B the event, which is that the test result is positive.

The Bayes formula gives a surprising result:

$$\begin{aligned} \mathsf{P}(H|B) &= \frac{\mathsf{P}(B|H)\mathsf{P}(H)}{\mathsf{P}(B|H)\mathsf{P}(H) + \mathsf{P}(B|H^c)\mathsf{P}(H^c)} \\ &= \frac{0.95 \cdot 0.001}{0.95 \cdot 0.001 + 0.05 \cdot 0.999} \approx 0.0187, \end{aligned}$$

that is, the patient is ill less than in 2% of cases. This means that a much more accurate test is required to determine such a rare disease or test should be repeated.

Example

A patient has a rare desease with probability 0.001. A medical test gives the correct answer in 95% of cases. What is the probability that the patient is ill if the test showed the positive result?

- H the hypothesis that the patient is ill, $\mathscr{D} = \{H, H^c\}$
- B the event, which is that the test result is positive.

The Bayes formula gives a surprising result:

$$\begin{aligned} \mathsf{P}(H|B) &= \frac{\mathsf{P}(B|H)\mathsf{P}(H)}{\mathsf{P}(B|H)\mathsf{P}(H) + \mathsf{P}(B|H^c)\mathsf{P}(H^c)} \\ &= \frac{0.95 \cdot 0.001}{0.95 \cdot 0.001 + 0.05 \cdot 0.999} \approx 0.0187, \end{aligned}$$

that is, the patient is ill less than in 2% of cases. This means that a much more accurate test is required to determine such a rare disease or test should be repeated.

Independence

Events A, B are independent if

$$\mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B).$$

If $\mathsf{P}(B) > 0$, then the independence is equivalent to the condition

$$\mathsf{P}(A|B) = \mathsf{P}(A).$$

Similarly, if $\mathsf{P}(A) > 0$, then the independence is equivalent to the condition

$$\mathsf{P}(B|A) = \mathsf{P}(B).$$

Example

One card is taken at random from the deck of 36 cards. Are the events $A = \{\text{queen}\}$, $B = \{\text{spades}\}$ independent? Will the answer change for a 52-card deck?

- 36 cards

$$\mathsf{P}(A) = 1/9, \quad \mathsf{P}(B) = 1/4,$$

$$\mathsf{P}(A \cap B) = \mathsf{P}(\text{queen of spades}) = 1/36 = \mathsf{P}(A)\mathsf{P}(B).$$

- 52 cards

$$\mathsf{P}(A) = 1/13, \quad \mathsf{P}(B) = 1/4,$$

$$\mathsf{P}(A \cap B) = \mathsf{P}(\text{queen of spades}) = 1/52 = \mathsf{P}(A)\mathsf{P}(B).$$

Example

One card is taken at random from the deck of 36 cards. Are the events $A = \{\text{queen}\}$, $B = \{\text{spades}\}$ independent? Will the answer change for a 52-card deck?

- 36 cards

$$\mathsf{P}(A) = 1/9, \quad \mathsf{P}(B) = 1/4,$$

$$\mathsf{P}(A \cap B) = \mathsf{P}(\text{queen of spades}) = 1/36 = \mathsf{P}(A)\mathsf{P}(B).$$

- 52 cards

$$\mathsf{P}(A) = 1/13, \quad \mathsf{P}(B) = 1/4,$$

$$\mathsf{P}(A \cap B) = \mathsf{P}(\text{queen of spades}) = 1/52 = \mathsf{P}(A)\mathsf{P}(B).$$

Independence

σ-algebras $\mathscr{F}_1, \ldots, \mathscr{F}_n$ are called independent, if

$$\mathsf{P}(A_1 \cap \cdots \cap A_n) = \mathsf{P}(A_1) \ldots \mathsf{P}(A_n).$$

ξ is independent of $\mathscr{H}$, if $\mathscr{F}_\xi$ and $\mathscr{H}$ are independent

$$\mathsf{P}(\{\xi \in B\} \cap A) = \mathsf{P}(\xi \in B)\mathsf{P}(A), \quad B \in \mathscr{B}(\mathbb{R}), A \in \mathscr{H}.$$

$(\xi_i)_{i=1}^n$ are independent, if σ-algebras $\mathscr{F}_{\xi_i}$ are independent:

$$\mathsf{P}(\xi_1 \in B_1, \ldots, \xi_n \in B_n) = \mathsf{P}(\xi_1 \in B_1) \ldots \mathsf{P}(\xi_n \in B_n)$$

In other words,

$$\mathsf{P}_{\xi_1,\ldots,\xi_n}(B_1, \ldots, B_n) = \mathsf{P}_{\xi_1}(B_1) \ldots \mathsf{P}_{\xi_n}(B_n).$$

For discrete ξ_i:

$$p_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n) = \mathsf{P}_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n) = \mathsf{P}_{\xi_1}(x_1)\ldots\mathsf{P}_{\xi_n}(x_n)$$
$$= p_{\xi_1}(x_1)\ldots p_{\xi_n}(x_n).$$

For continuous ξ_i:

$$\int_{B_1}\cdots\int_{B_n} p_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n)\,dx_1\ldots dx_n = \mathsf{P}_{\xi_1,\ldots,\xi_n}(B_1,\ldots,B_n)$$
$$=\mathsf{P}_{\xi_1}(B_1)\ldots\mathsf{P}_{\xi_n}(B_n) = \int_{B_1} p_{\xi_1}(x_1)\,dx_1\cdots\int_{B_n} p_{\xi_n}(x_n)\,dx_n$$
$$= \int_{B_1}\cdots\int_{B_n} p_{\xi_1}(x_1)\ldots p_{\xi_n}(x_n)\,dx_1\ldots dx_n.$$

Thus,

$$p_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n) = p_{\xi_1}(x_1)\ldots p_{\xi_n}(x_n).$$

Let ξ, η be independent, then $\mathsf{E}(\xi\eta) = \mathsf{E}(\xi)\mathsf{E}(\eta)$.
In particular, if $\xi = I_A$, $\eta = I_B$, then

$$\mathsf{P}(A \cap B) = \mathsf{E}I_{A\cap B} = \mathsf{E}(I_A I_B) = \mathsf{E}I_A \mathsf{E}I_B = \mathsf{P}(A)\mathsf{P}(B).$$

Furthermore,

$$\begin{aligned}\mathrm{Var}(\xi+\eta) &= \mathrm{Var}(\xi - \mathsf{E}\xi + \eta - \mathsf{E}\eta) = \mathsf{E}(\xi - \mathsf{E}\xi + \eta - \mathsf{E}\eta)^2 \\ &= \mathsf{E}(\xi - \mathsf{E}\xi)^2 + 2\mathsf{E}\left[(\xi - \mathsf{E}\xi)(\eta - \mathsf{E}\eta)\right] + \mathsf{E}(\eta - \mathsf{E}\eta)^2 \\ &= \mathrm{Var}(\xi - \mathsf{E}\xi) + \mathrm{Var}(\eta - \mathsf{E}\eta) = \mathrm{Var}(\xi) + \mathrm{Var}(\eta).\end{aligned}$$

Example: variance of the binomial distribution

$\xi \sim B(n,p)$: $p_\xi(k) = C_n^k p^k (1-p)^{n-k}$. Find $\mathrm{Var}(\xi)$.

Sum of independent Bernoulli random variables

$$\eta = \sum_{i=1}^{n} \xi_i, \quad \xi_i = I_{\{a_i = H\}}, \quad \xi_i \sim \mathrm{Ber}(p)$$

has the binomial distribution: $\eta \sim B(n,p)$. Hence,

$$\mathrm{Var}(\xi) = \mathrm{Var}(\eta) = \sum_{i=1}^{n} \mathrm{Var}(\xi_i) = np(1-p).$$

Covariance

$$\mathrm{Cov}(\xi,\eta) := \mathsf{E}[(\xi-\mathsf{E}\xi)(\eta-\mathsf{E}\eta)] = \int_{\mathbb{R}^2}(x-\mathsf{E}\xi)(y-\mathsf{E}\eta)\mathsf{P}_\xi(dxdy)$$
$$= \mathsf{E}(\xi\eta) - \mathsf{E}\xi\cdot\mathsf{E}\eta.$$

The covariance is positive (negative) if the deviations of ξ and η from the mean values "more often " have the same sign (different signs).
If ξ, η are independent, then

$$\mathrm{Cov}(\xi,\eta) := \mathsf{E}[\xi-\mathsf{E}\xi]\cdot\mathsf{E}[\eta-\mathsf{E}\eta] = 0.$$

The converse is false.

Properties of the covariance

$$\mathrm{Cov}(\xi, \xi) = \mathrm{Var}(\xi),$$

$$\mathrm{Cov}(\xi, \eta) = \mathrm{Cov}(\eta, \xi),$$

$$\mathrm{Cov}(\alpha\xi + \beta\eta, \zeta) = \alpha\mathrm{Cov}(\xi, \zeta) + \beta\mathrm{Cov}(\eta, \zeta).$$

$$\mathrm{Var}(\xi + \eta) = \mathrm{Cov}(\xi + \eta, \xi + \eta) = \mathrm{Var}(\xi) + \mathrm{Var}(\eta) + 2\mathrm{Cov}(\xi, \eta)$$

$$\mathrm{Var}\left(\sum_{i=1}^{n} \xi_i\right) = \sum_{i=1}^{n} \mathrm{Var}(\xi_i) + \sum_{i \neq j} \mathrm{Cov}(\xi_i, \xi_j)$$

Correlation coefficient

Cauchy-Schwarz inequality

$$|\mathsf{E}(\xi\eta)| \le \sqrt{\mathsf{E}(\xi^2)}\sqrt{\mathsf{E}(\eta^2)}.$$

Puy $\sigma(\xi) = \sqrt{\mathrm{Var}(\xi)}$. Correlation coefficient

$$\rho(\xi,\eta) := \frac{\mathrm{Cov}(\xi,\eta)}{\sigma(\xi)\sigma(\eta)}, \quad \sigma(\xi),\sigma(\eta) > 0.$$

is a dimensionless version of the covariance coefficient. By the Cauchy-Schwarz inequality,

$$|\mathrm{Cov}(\xi,\eta)| \le \sqrt{\mathsf{E}(\xi-\mathsf{E}\xi)^2}\sqrt{\mathsf{E}(\eta-\mathsf{E}\eta)^2} = \sigma(\xi)\sigma(\eta).$$

Hence, $\rho(\xi,\eta) \in [-1,1]$.

Density of transformation of a multidimensional random variable

Let $g : \mathbb{R}^d \mapsto \mathbb{R}^d$ be a one-to-one differentiable mapping. Let $\eta = (\eta_1, \ldots, \eta_d) = g(\xi_1, \ldots, \xi_d) = g(\xi)$. Then

$$\begin{aligned} \mathsf{E}h(\eta) &= \int_{\mathbb{R}^d} h(y) f_\eta(y)\, dy = \mathsf{E}h(g(\xi)) = \int_{\mathbb{R}^d} h(g(x)) f_\xi(x)\, dx \\ &= \int_{\mathbb{R}^d} h(y) f_\xi(g^{-1}(y)) |\det\left[(g^{-1})'\right]|\, dy, \end{aligned}$$

where $(g^{-1})'$ is the Jacobian of the mapping g^{-1}. Thus,

$$f_\eta(y) = f_\xi(g^{-1}(y)) \left|\det\left[(g^{-1})'\right]\right|.$$

In particular, if $\eta = A\xi + b$, where A is an invertible $d \times d$ matrix and $b \in \mathbb{R}^d$, then

$$f_\eta(y) = f_\xi(A^{-1}(y-b)) \left|\det\left[A^{-1}\right]\right| = \frac{1}{|\det A|} f_\xi(A^{-1}(y-b)).$$

Multidimensional normal distribution

Let $\xi_i \sim N(0,1)$, $i = 1, \ldots, d$ be independent. *Multivariate standard normal (Gaussian)* distribution is the joint distribution of ξ_i:

$$f_{\xi_1,\ldots,\xi_d}(x_1,\ldots,x_d) = \prod_{i=1}^{d} f_{\xi_i}(x_i) = \frac{1}{(\sqrt{2\pi})^d} \prod_{i=1}^{d} e^{-x_i^2/2}$$
$$= \left(\frac{1}{\sqrt{2\pi}}\right)^d e^{-\langle x,x\rangle/2}$$

Multidimensional normal (Gaussian) distribution is the distribution of the vector $\eta = A\xi + \mu$ (the matrix A is assumed to be invertible for now). According to the density transformation formula,

$$f_\eta(y) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \frac{1}{|\det A|} e^{-\langle A^{-1}(y-\mu), A^{-1}(y-\mu)\rangle/2}$$
$$= \left(\frac{1}{\sqrt{2\pi}}\right)^d \frac{1}{|\det A|} e^{-\langle (A^{-1})^T A^{-1}(y-\mu), y-\mu\rangle/2}$$

$$(A^{-1})^T A^{-1} = (A^T)^{-1} A^{-1} = (AA^T)^{-1}.$$

Matrix AA^T is symmetric and positive definite:

$$\langle AA^T x, x \rangle = \|A^T x\|^2 > 0, \quad x \neq 0.$$

Put $\Sigma = AA^T$. Since

$$\det \Sigma = \det(AA^T) = (\det A)^2,$$

we have

$$f_\eta(y) = \left(\frac{1}{\sqrt{2\pi}} \right)^d \frac{1}{\sqrt{\det \Sigma}} e^{-\langle \Sigma^{-1}(y-\mu), y-\mu \rangle / 2}$$

For the distribution of a Gaussian vector η, the notation $\eta \sim N(\mu, \Sigma)$ is used. Consider the vector of expectations μ:

$$\mathsf{E}\eta = \mathsf{E}(\mu + A\xi) = \mu$$

and the *covariance matrix*

$$\begin{aligned}\operatorname{Cov}(\eta_i, \eta_j) = \mathsf{E}(\langle a_i, \xi\rangle \cdot \langle a_j, \xi\rangle) = \mathsf{E}\sum_{kr} a_{ik}a_{jr}\xi_k\xi_r \\ = \sum_k a_{ik}a_{jk} = (AA^T)_{ij} = \Sigma_{ij}.\end{aligned}$$

These formulas indicate the meaning of the parameters μ, Σ.

If η_i are uncorrelated, then the matrix Σ is diagonal and

$$f_\eta(y) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \frac{1}{\sqrt{\prod_{i=1}^d \Sigma_{ii}}} \exp\left(-\sum_{i=1}^d \frac{(y_i-\mu_i)^2}{2\Sigma_{ii}}\right)$$
$$= \prod_{i=1}^d \sqrt{\frac{1}{2\pi\Sigma_{ii}}} \exp\left(-\frac{(y_i-\mu_i)^2}{2\Sigma_{ii}}\right)$$

Thus, $f_\eta(y)$ decomposes into the product of marginal densities of one-dimensional normal random variables $\eta_i \sim N(\mu_i, \Sigma_{ii})$. Hence the components of η are independent. Thus, for a Gaussian vector, the the components are uncorrelated if and only if they are independent.

Characteristic functions

Characteristic function of a random variable $\xi : \Omega \mapsto \mathbb{R}^d$ is the Fourier transform of its distribution:

$$\varphi_\xi(u) = \mathsf{E}e^{i\langle u,\xi\rangle} = \int_{\mathbb{R}^d} e^{i\langle u,x\rangle}\, \mathsf{P}_\xi(dx).$$

Here $\langle u, x\rangle$ is the usual scalar product in $\mathbb{R}^d$.

- The distribution P_ξ is uniquely determined by its characteristic function (uniqueness theorem).

Since the characteristic function is uniquely determined by the distribution, it can also be called the characteristic function of the distribution .
The characteristic function can be used to calculate the mean and variance, since (in the one-dimensional case)

$$\varphi'_\xi(u) = i\mathsf{E}\left(\xi e^{iu\xi}\right), \quad \varphi''_\xi(0) = -\mathsf{E}\left(\xi^2 e^{iu\xi}\right),$$

$$\varphi'_\xi(0) = i\mathsf{E}\xi, \quad \varphi''_\xi(0) = -\mathsf{E}\xi^2.$$

The characteristic function of the Poisson distribution

Let us find the expectation and variance of the Poisson distribution using the characteristic functions. Let $\xi \sim \Pi(\lambda)$. Then

$$\varphi_\xi(u) = \mathsf{E}e^{iu\xi} = \sum_{k=0}^{\infty} e^{iuk}\frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^{iu})^k}{k!} = \exp(-\lambda + \lambda e^{iu}),$$

$$\varphi'_\xi(u) = \lambda i e^{iu}\exp(\lambda(e^{iu}-1)),$$

$$\varphi''_\xi(u) = (-\lambda e^{iu} + (\lambda i e^{iu})^2)\exp(\lambda(e^{iu}-1)).$$

Hence,

$$\mathsf{E}\xi = -i\varphi'_\xi(0) = \lambda, \quad \mathsf{E}\xi^2 = -\varphi''_\xi(0) = \lambda + \lambda^2,$$

$$\mathrm{Var}(\xi) = \mathsf{E}(\xi^2) - (\mathsf{E}\xi)^2 = \lambda.$$

The characteristic function of the normal distribution

Ler $\eta \sim N(0,1)$. Then

$$\varphi_\eta(u) = \mathsf{E}e^{iu\eta} = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{iux} e^{-x^2/2}\, dx,$$

$$\varphi'_\eta(u) = i\mathsf{E}(\eta e^{iu\eta}) = \frac{i}{\sqrt{2\pi}} \int_{\mathbb{R}} x e^{iux} e^{-x^2/2}\, dx = -\frac{i}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{iux} de^{-x^2/2}$$
$$= -\frac{u}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{iux} e^{-x^2/2}\, dx = -u\varphi_\eta(u).$$

Solving the differential equation $\varphi'_\eta(u) = -u\varphi_\eta(u)$ with the initial condition $\varphi_\eta(0) = 1$, we get

$$\varphi_\eta(u) = e^{-u^2/2}.$$

To find the characteristic function of an arbitrary normal distribution, consider a random variable $\xi = \mu + \sigma\eta \sim N(\mu, \sigma^2)$:

$$\varphi_\xi(u) = e^{iu\mu}\mathsf{E}e^{iu\sigma\eta} = e^{i\mu u - \sigma^2 u^2/2}.$$

$$\varphi'_\xi(u) = (i\mu - \sigma^2 u)e^{i\mu u - \sigma^2 u^2/2}, \quad \varphi''_\xi(u) = [(i\mu - \sigma^2 u)^2 - \sigma^2]e^{i\mu u - \sigma^2 u^2/2}.$$

Hence,

$$\mathsf{E}\xi = -i\varphi'_\xi(0) = \mu, \quad \mathsf{E}\xi^2 = -\varphi''_\xi(0) = \mu^2 + \sigma^2,$$
$$\mathrm{Var}(\xi) = \mathsf{E}\xi^2 - (\mathsf{E}\xi)^2 = \sigma^2.$$

Characteristic function of the sum of independent random variables

The characteristic function of the joint distribution of a random vector is equal to the product of the characteristic functions of the components:

$$\varphi_\xi(u) = \mathsf{E}e^{i\langle u,\xi\rangle} = \mathsf{E}\prod_{k=1}^{n} e^{iu_k\xi_k} = \prod_{k=1}^{n} \mathsf{E}e^{iu_k\xi_k} = \prod_{k=1}^{n} \varphi_{\xi_k}(u_k).$$

From the mentioned uniqueness theorem it follows that the converse statement is also true, i.e. this equality is a criterion of independence of $(\xi_1, \ldots, \xi_n)$.

Characteristic function of a multivariate normal distribution

$\eta = \mu + A\xi$:

$$\varphi_\eta(u) = e^{i\langle u,\mu\rangle}\mathsf{E}e^{i\langle u,A\xi\rangle} = e^{i\langle u,\mu\rangle}\mathsf{E}e^{i\langle A^T u,\xi\rangle} = e^{i\langle u,\mu\rangle}\prod_{k=1}^{d}\mathsf{E}e^{i(A^T u)_k\xi_k}$$

$$= e^{i\langle u,\mu\rangle}\prod_{i=1}^{d}e^{-(A^T u)_i^2/2} = e^{i\langle u,\mu\rangle}e^{-\|A^T u\|^2/2} = e^{i\langle u,\mu\rangle-\langle\Sigma u,u\rangle/2},$$

where $\Sigma = AA^T$. Thus, the vector η is *Gaussian*, if and only if

$$\varphi_\eta(u) = e^{i\langle u,\mu\rangle-\langle\Sigma u,u\rangle/2},$$

where $\mu \in \mathbb{R}^d$ ïS Σ is a symmetric nonnegative definite matrix. This property can be taken as a definition of a multidimensional Gaussian random variable. Moreover, it suffices to assume that Σ is non-negative definite: its invertibility is not required.

Distribution of a sum of one-dimensional independent normal r.v.

$\xi_i \sim N(\mu_i, \sigma_i^2)$ are independent, $\zeta = \xi_1 + \xi_2$.

$$\begin{aligned}\varphi_\zeta(u) &= \varphi_{\xi_1}(u)\varphi_{\xi_2}(u) = e^{iu\mu_1 - \sigma_1^2 u^2/2} e^{iu\mu_2 - \sigma_2^2 u^2/2} \\ &= e^{iu(\mu_1+\mu_2) - (\sigma_1^2+\sigma_2^2)u^2/2}.\end{aligned}$$

The uniqueness theorem implies that $\zeta \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. It is much more difficult to obtain this result by calculating convolution. It was formulated above, but calculations were not performed.

Convergence of random variables

A sequence of random variables ξ_n converges to ξ

- *in probability*: $\xi_n \xrightarrow{\mathsf{P}} \xi$, if

$$\lim_{n\to\infty} \mathsf{P}\{\omega : |\xi_n(\omega) - \xi(\omega)| \geq \varepsilon\} = 0;$$

- *in distribution*: $\xi_n \xrightarrow{d} \xi$, if $\lim_{n\to\infty} \mathsf{E}g(\xi_n) = \mathsf{E}g(\xi)$ for any bounded continuous function $g : \mathbb{R}^d \mapsto \mathbb{R}$.

- convergence in probability $\Longrightarrow$ convergence in distribution;
- *Continuity theorem*: if $\varphi_{\xi_n}(u) \to \varphi(u)$, $n \to \infty$ for all u, then $\xi_n \xrightarrow{d} \xi$.

Weak law of large numbers

Let $(\xi_i)_{i=1}^{\infty}$ be i.i.d. random variables, $\mathrm{E}\xi_i = \mu$, $\mathrm{Var}(\xi_i) = \sigma^2$.

- *Weak law of large numbers:*

$$M_n = \frac{1}{n}\sum_{i=1}^{n} \xi_i \xrightarrow{\mathrm{P}} \mu, \quad n \to \infty.$$

Proof.

$$\mathrm{E}M_n = \mu, \quad \mathrm{Var}(M_n) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(\xi_i) = \frac{\sigma^2}{n}.$$

$$\mathrm{P}(|M_n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(M_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0, \quad n \to \infty. \quad \square$$

In particular, the frequency of occurrence of an event in a series of independent experiments approaches its probability: A_i are independent, $\mathrm{P}(A_i) = p$,

$$\nu_n = \frac{1}{n}\sum_{i=1}^{n} I_{A_i} \xrightarrow{\mathrm{P}} p$$

Central limit theorem

Let $(\xi_k)_{k=0}^{\infty}$ be i.i.d. random variables, $\mathsf{E}\xi_n = 0$, $\mathrm{Var}(\xi_n) = 1$.

- *Central limit theorem*:

$$\frac{S_n}{\sqrt{n}} := \frac{1}{\sqrt{n}}(\xi_1 + \cdots + \xi_n) \xrightarrow{d} \eta \sim N(0,1).$$

Proof. Let $\xi_j \sim \xi$. Then

$$\varphi_{S_n/\sqrt{n}}(u) = \mathsf{E}\prod_{j=1}^{n} e^{iu\xi_j/\sqrt{n}} = \prod_{j=1}^{n} \mathsf{E}e^{iu\xi_j/\sqrt{n}} = (\mathsf{E}e^{iu\xi/\sqrt{n}})^n$$

$$= \left(\mathsf{E}\left(1 + \frac{iu\xi}{\sqrt{n}} - \frac{1}{2}\frac{u^2\xi^2}{n} + o\left(\frac{1}{n}\right)\right)\right)^n = \left(1 - \frac{1}{2}\frac{u^2}{n} + o\left(\frac{1}{n}\right)\right)^n$$

$$= \exp\left(n \ln\left(1 - \frac{u^2}{2n} + o\left(\frac{1}{n}\right)\right)\right) \to e^{-u^2/2}, \quad n \to \infty \quad \square$$

- $\eta_n \xrightarrow{d} \eta$, if and only if

$$F_{\eta_n}(x) \to F_\eta(x)$$

at all points where F_η is continuous.

In particular, in the central limit theorem

$$\mathsf{P}\left(\frac{\xi_1 + \cdots + \xi_n}{\sqrt{n}} \le u\right) \to \Phi(x) := \mathsf{P}(\zeta \le x),$$

$$\zeta \in N(0,1), \quad \mathsf{E}\xi_i = 0, \quad \mathrm{Var}(\xi_i) = 1.$$

Monte-Carlo integration

Let $X_i \sim U(0,1)$. Then

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{\mathsf{P}} \mathsf{E}g(X_1) = \int_0^1 g(x)\,dx,$$

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \int_0^1 g(x)\,dx\right) \xrightarrow{d} N(0, \mathrm{Var}(g(X_1))).$$

If X_i has a p.d.f. f on $\mathbb{R}$, then

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{\mathsf{P}} \mathsf{E}g(X_1) = \int_{\mathbb{R}} g(x)f(x)\,dx.$$

Thus, if a computer can generate a realization of a sample with a given distribution, then we get a very simple method for the approximate calculation of integrals.

It is actually enough to have a generator for the uniform distribution: $\zeta \sim U(0,1)$. Let $F_\xi(x) = \mathsf{P}(\xi \le x)$ be the c.d.f. of ξ. Put

$$F_\xi^{-1}(u) = \min\{x : F_\xi(x) \ge u\}, \quad u \in (0,1)$$

(the minimum is attained, since F_ξ is right-continuous).

- Inverse-transform method:

$$F_\xi^{-1}(\zeta) \sim \mathsf{P}_\xi.$$

Proof for a continuous and strictly increasing F_ξ. The equation

$$F_\xi(x) = u$$

has a unique solution for any $u \in (0,1)$. In this case F_ξ^{-1} coincides with the ordinary inverse function and

$$\mathsf{P}(F_\xi^{-1}(\zeta) \le x) = \mathsf{P}(\zeta \le F_\xi(x)) = F_\xi(x).$$

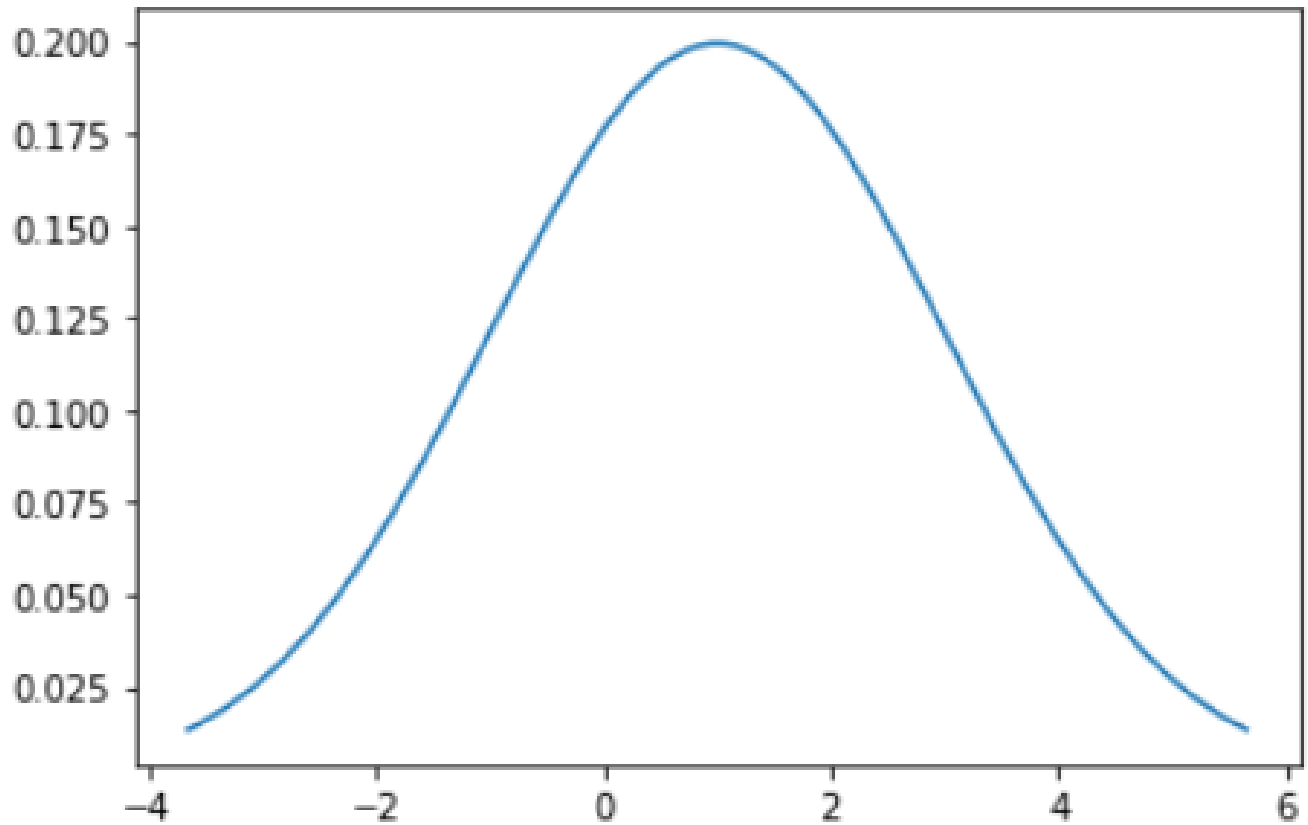
In [10]:
```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
rv = stats.norm(1,2)
```

In [11]:
```
print('mean:', rv.mean())
print('variance:', rv.var())
print('standard deviation:', rv.std())
print('median:', rv.median())
print('0.01 percentile:', rv.ppf(0.01))
print('0.99 percentile:', rv.ppf(0.99))
```

```
mean: 1.0
variance: 4.0
standard deviation: 2.0
median: 1.0
0.01 percentile: -3.6526957480816815
0.99 percentile: 5.6526957480816815
```

In [12]:
```
x=np.linspace(rv.ppf(0.01),rv.ppf(0.99),100)
plt.plot(x,rv.pdf(x))
```

Out[12]:
```
[<matplotlib.lines.Line2D at 0x1d0295ca128>]
```



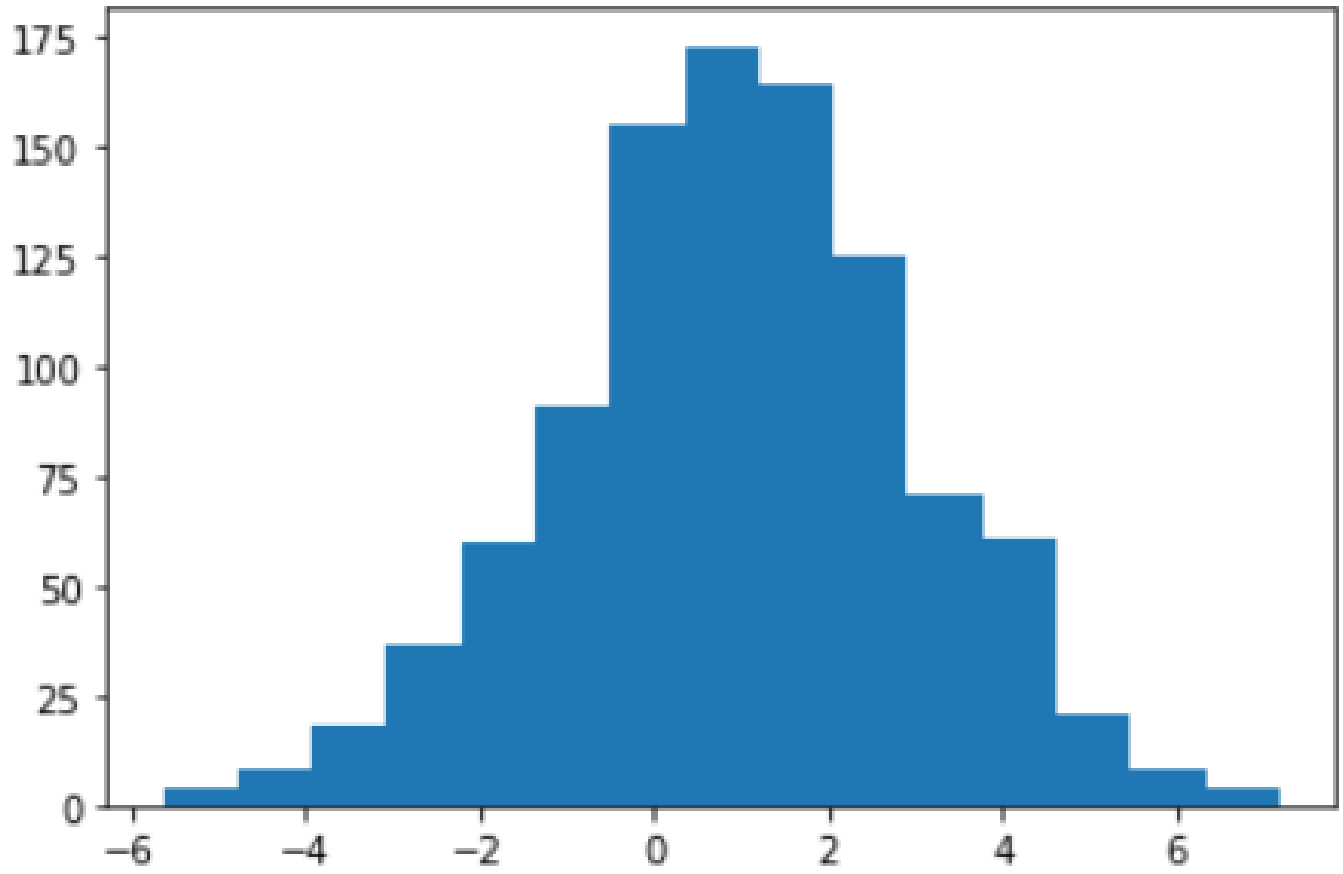
In [14]:
```
X_samples=rv.rvs(1000)
print(X_samples[:4])
```

```
[-1.26321354 -1.12273014  0.94553942  1.94721138]
```

In [6]:
```
print('sample mean:', X_samples.mean())
print('sample variance:', X_samples.var())
print('sample standard deviation:', X_samples.std())
print('sample median:', np.median(X_samples))
print('0.01 percentile:', np.percentile(X_samples,1.))
print('0.99 percentile:', np.percentile(X_samples,99.))
```

```
sample mean: 1.023470817992147
sample variance: 4.20289766675017
sample standard deviation: 2.050096895958994
sample median: 1.031712699806072 8
0.01 percentile: -3.719701467370804 4
0.99 percentile: 5.67401027625593
```

In [15]:
```
plt.hist(X_samples, bins=15)
plt.show()
```



Approximate calculation of the integral $\int_0^1 x^2\,dx = 1/3$

by the Monte-Carlo method:

In [17]:
```
X_samples=stats.uniform.rvs(size=1000)
np.mean(X_samples**2)
```

Out[17]:
```
0.3434236797277684
```

Approximate calculation of the integral $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2}\,dx = 1$

by the Monte-Carlo method:

In [18]:
```
X_samples=stats.norm.rvs(size=1000)
np.mean(X_samples**2)
```

Out[18]:
```
0.9752834638543979
```

Parameter estimation in classical statistics

Data $x_1, \ldots, x_n$ are assumed to be the realizations of i.i.d. random variables $X_1, \ldots, X_n \sim \mathcal{P}$, where P is an unknown probability measure from a known set $\mathcal{P}$. The aim is to identify P using the data.
Parametric model: $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$. Usually one considers families of p.m.f or p.d.f.:

$$\{x \mapsto f(x;\theta) : \theta \in \Theta\},$$

which we will denote by the same letter $\mathcal{P}$. For example, the family of normal p.d.f.:

$$f(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}, \quad \theta = (\mu, \sigma) \in \mathbb{R} \times (0, \infty).$$

We want to construct a statistics $\widehat{\theta}$ (any function of data) such that

$$\theta \approx \widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n).$$

Maximum likelihood method

Likelihood function

$$L_n(x;\theta) = \prod_{i=1}^{n} p(x_i;\theta).$$

For a p.m.f. family, $L_n(x;\theta) = \mathsf{P}_\theta(X_1 = x_1, \ldots, X_n = x_n)$.
Logarithmic likelihood function:

$$l_n(x;\theta) = \ln L_n(x;\theta).$$

MLE (maximum likelihood estimate):

$$\widehat{\theta} \in \arg\max_\theta L_n(x;\theta) = \arg\max_\theta l_n(x;\theta).$$

In connection with this optimization problem, we can assume that the likelihood function is determined up to a positive factor $C(x)$, independent of θ:

$$L_n(x;\theta) \propto C(x) L_n(x;\theta).$$

Example. $X_i \sim \mathrm{Ber}(\theta)$,

$$p_\theta(x) = \begin{cases} \theta, & x = 1, \\ 1-\theta, & x = 0. \end{cases}$$

$$L_n(x;\theta) = \prod_{i=1}^{n} p_\theta(x_i) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^S(1-\theta)^{n-S}, \quad S = \sum_{i=1}^{n} x_i,$$

$$l_n(x;\theta) = S\ln\theta + (n-S)\ln(1-\theta),$$

$$\frac{\partial l_n(x;\theta)}{\partial\theta} = \frac{S}{\theta} - \frac{n-S}{1-\theta} = 0, \quad S = n\theta,$$

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

Example. $X \sim N(\mu, 1)$, $\theta = \mu$,

$$f_\theta(y) = \frac{1}{\sqrt{2\pi}} e^{-(y-\mu)^2/2},$$

$$L_n(x;\mu) \propto \prod_{i=1}^{n} e^{-(x_i-\mu)^2/2} \propto e^{-n(\overline{x}-\mu)^2/2}, \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Indeed,

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n} x_i^2 - 2n\mu\overline{x} + n\mu^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 + n(\overline{x}-\mu)^2.$$

Hence (neglecting terms independent of μ),

$$l_n(x;\mu) = -n(\overline{x}-\mu)^2/2,$$

Thus,

$$\widehat{\mu} = \overline{X}.$$

Example. $X_i \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma)$,

$$L_n(x;\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/(2\sigma^2)} \propto \frac{1}{\sigma^n} \prod_{i=1}^{n} e^{-(x_i-\mu)^2/(2\sigma^2)},$$

$$l_n(x;\theta) = \ln L_n(x;\theta) = -n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$(l_n)_\mu = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} x_i - n\mu \right) = 0,$$

$$(l_n)_\sigma = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 = 0.$$

It follows that

$$\widehat{\mu} = \overline{X}, \quad \widehat{\sigma} = \left(\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 \right)^{1/2}.$$

By the invariance principle for MLE for σ^2 coincides with $\widehat{\sigma}^2$. The same estimates were obtained by the method of moments.

Example. $X_i \sim U(0,\theta)$,

$$f(y;\theta) = \begin{cases} 1/\theta, & y \in (0,\theta), \\ 0, & \text{otherwise}, \end{cases}$$

$$L_n(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \begin{cases} \frac{1}{\theta^n}, & x_i < \theta, i = 1,\dots,n, \\ 0, & \text{otherwise}. \end{cases}.$$

The MLE is the smallest θ, which is larger than all x_i:

$$\widehat{\theta} = X_{(n)} := \max_{1 \le i \le n} X_i.$$

The method of moments gave another estimate:

$$\widehat{\theta} = 2\overline{X}.$$

Example. $X_i \sim \Pi(\lambda)$,

$$f(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x \in \mathbb{Z}_+,$$

$$L_n(x;\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}\prod_{i=1}^{n}\frac{1}{x_i!},$$

$$l_n(x;\lambda) = -n\lambda + \sum_{i=1}^{n} x_i \ln\lambda + \text{Const},$$

$$\frac{\partial l_n}{\partial\lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0.$$

MLE:

$$\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

Conditional expectation

Consider a σ-algebra $\mathscr{H} \subset \mathscr{F}$. Let $\xi \in L^1(\mathscr{F}) := L^1(\Omega, \mathscr{F}, \mathsf{P}, \mathbb{R})$.

- A random variable $\eta \in L^1(\mathscr{H})$ is called a *conditional expectation* of ξ given $\mathscr{H}$, if
$$\mathsf{E}(\xi I_A) = \mathsf{E}(\eta I_A), \quad A \in \mathscr{H}.$$
Notation: $\eta = \mathsf{E}(\xi|\mathscr{H})$.
- If $\eta' = \eta$ a.s.:ïS. $\mathsf{P}(\eta = \eta') = 1$, then
$$\mathsf{E}(\eta I_A) = \mathsf{E}(\eta' I_A).$$
Hence, the conditional expectation can be unique only up to a set of measure zero.

Theorem

For any $\xi \in L^1(\mathscr{F})$ the conditional expectation

$$\eta = \mathsf{E}(\xi|\mathscr{H}) \in L^1(\mathscr{H})$$

exists and is unique (as an equivalence class).

Evident properties:

- Let $\xi = C$, then $\mathsf{E}(\xi|\mathscr{H}) = \xi$ (by the definition).
- Let $\xi \in L^1(\mathscr{H})$, then $\mathsf{E}(\xi|\mathscr{H}) = \xi$ (by the definition).
- Let $\mathscr{H} = \{\emptyset, \Omega)$. Then $\mathsf{E}(\xi|\mathscr{H}) = \mathsf{E}\xi$.

Example

Consider a partition $(A_i)_{i=1}^m$ with $\mathsf{P}(A_i) > 0$, $i = 1, \ldots, m$. Assume that $\mathscr{H}$ is generated by this partition. Then

$$\mathsf{E}(\xi|\mathscr{H}) = \sum_{i=1}^{m} \frac{\mathsf{E}(\xi I_{A_i})}{\mathsf{P}(A_i)} I_{A_i}, \qquad \xi \in L^1(\mathscr{F}).$$

Indeed, denote the left-hand side by η. For any atom A_j we have

$$\mathsf{E}(\eta I_{A_j}) = \mathsf{E}\left(\sum_{i=1}^{m} \frac{\mathsf{E}(\xi I_{A_i})}{\mathsf{P}(A_i)} I_{A_i} I_{A_j}\right) = \mathsf{E}\left(\frac{\mathsf{E}(\xi I_{A_j})}{\mathsf{P}(A_j)} I_{A_j}\right) = \mathsf{E}(\xi I_{A_j}).$$

But any $A \in \mathscr{H}$ is the union of some A_j.

Example

Coin tossing model.

- $\Omega = \{(a_1, \ldots, a_n) : a_i \in \{H, T\}\}$,
- $\mathscr{F}_t = \sigma(\{\omega : (a_1, \ldots, a_t) \text{ fixed}\})$,
- $\mathsf{P}(\omega) = p^{\nu(\omega)} q^{n-\nu(\omega)}$, $\nu(\omega)$ — the number of H in the sequence $\omega = (a_1, \ldots, a_n)$, $p \in (0, 1)$, $q = 1 - p$.
- $A_H := H_1 = \{\omega : a_1 = H\}$,
- $A_{HH} := H_1 \cap H_2 = \{\omega : a_1 = a_2 = H\}$,
- $A_{HHH} := H_1 \cap H_2 \cap H_3 = \{\omega : a_1 = a_2 = a_3 = H\}$.

$$\mathsf{E} I_{A_{HHH}} = \mathsf{P}(A_{HHH}) = p^3,$$

$$\mathsf{E}(I_{A_{HHH}} | \mathscr{F}_1) = \frac{\mathsf{E}(I_{A_{HHH}} I_{A_H})}{\mathsf{P}(A_H)} I_{A_H} + \frac{\mathsf{E}(I_{A_{HHH}} I_{A_T})}{\mathsf{P}(A_T)} I_{A_T}$$

$$= \frac{\mathsf{P}(A_{HHH})}{\mathsf{P}(A_H)} I_{A_H} = \frac{p^3}{p} I_{A_H} = p^2 I_{A_H}.$$

Example (continued)

$$
\begin{aligned}
\mathsf{E}(I_{A_{HHH}}|\mathscr{F}_2) &= \frac{\mathsf{E}(I_{A_{HHH}} I_{A_{HH}})}{\mathsf{P}(A_{HH})} I_{A_{HH}} + \frac{\mathsf{E}(I_{A_{HHH}} I_{A_{HT}})}{\mathsf{P}(A_{HT})} I_{A_{HT}} \\
&\quad + \frac{\mathsf{E}(I_{A_{HHH}} I_{A_{TH}})}{\mathsf{P}(A_{TH})} I_{A_{TH}} + \frac{\mathsf{E}(I_{A_{HHH}} I_{A_{TT}})}{\mathsf{P}(A_{TT})} I_{A_{TT}} \\
&= \frac{\mathsf{E}(I_{A_{HHH} \cap A_{HH}})}{\mathsf{P}(A_{HH})} I_{A_{HH}} = \frac{\mathsf{P}(A_{HHH})}{\mathsf{P}(A_{HH})} I_{A_{HH}} = \frac{p^3}{p^2} I_{A_{HH}} \\
&= p I_{A_{HH}}, \\
\mathsf{E}(I_{A_{HHH}}|\mathscr{F}_3) &= I_{A_{HHH}}.
\end{aligned}
$$

Example

On the same probability space consider the stochastic process

$$S_{t+1} = S_t(uI_{\{a_{t+1}=H\}} + dI_{\{a_{t+1}=T\}}),$$

where S_0 is a constant and $u > d$. This is the simplest model of a risky asset (binomial model, Cox-Ross-Rubinstein model). Let us find $\mathsf{E}(S_2|\mathscr{F}_1)$:

$$\mathsf{E}(S_2|\mathscr{F}_1) = \frac{\mathsf{E}(S_2 I_{A_H})}{\mathsf{P}(A_H)} I_{A_H} + \frac{\mathsf{E}(S_2 I_{A_T})}{\mathsf{P}(A_T)} I_{A_T}.$$

Since $S_2 = S_0(u^2 I_{A_{HH}} + udI_{A_{HT}} + duI_{A_{TH}} + d^2 I_{A_{TT}})$, we get

$$\mathsf{E}(S_2 I_{A_H}) = S_0 \mathsf{E}(u^2 I_{A_{HH}} + udI_{A_{HT}}) = S_0(u^2 p^2 + udpq),$$

$$\mathsf{E}(S_2 I_{A_T}) = S_0 \mathsf{E}(duI_{A_{TH}} + d^2 I_{A_{TT}}) = S_0(udpq + d^2 q^2).$$

Example (continued)

Thus,

$$\begin{aligned}
\mathsf{E}(S_2|\mathscr{F}_1) &= \frac{S_0(u^2p^2+udpq)}{p} I_{A_H} + \frac{S_0(udpq+d^2q^2)}{q} I_{A_T} \\
&= S_0u(up+dq)I_{A_H} + S_0d(up+dq)I_{A_T} \\
&= (up+dq)S_1
\end{aligned}$$

Properties of conditional expectation

(1) Conditional expectation is a linear operator from $L^1(\mathscr{F})$ to $L^1(\mathscr{H})$:

$$\mathsf{E}(a_1\xi_1 + a_2\xi_2|\mathscr{H}) = a_1\mathsf{E}(\xi_1|\mathscr{H}) + a_2\mathsf{E}(\xi_2|\mathscr{H}).$$

(2) Let $\xi \in L^1(\mathscr{F})$ and let η be $\mathscr{H}$-measurable and bounded. Then

$$\mathsf{E}(\xi\eta|\mathscr{H}) = \eta\mathsf{E}(\xi|\mathscr{H}).$$

Proof for indicators. Let $\eta = I_B$, $B \in \mathscr{H}$. Then the left and right sides are equal to

$$\mathsf{E}(\xi\eta I_A) = \mathsf{E}(\xi I_B I_A) = \mathsf{E}(\xi I_{A\cap B})$$

$$\mathsf{E}(\eta\mathsf{E}(\xi|\mathscr{H})I_A) = \mathsf{E}(\mathsf{E}(\xi|\mathscr{H})I_{A\cap B}) = \mathsf{E}(\xi I_{A\cap B})$$

respectively.

Properties of conditional expectation

(3) If $\xi \in L^1(\mathscr{F})$ is independent from $\mathscr{H}$, then

$$\mathsf{E}(\xi|\mathscr{H}) = \mathsf{E}\xi.$$

Proof.

$$\mathsf{E}(\xi I_A) = \mathsf{E}\xi \cdot \mathsf{E}I_A = \mathsf{E}(I_A \mathsf{E}\xi), \quad A \in \mathscr{H}.$$

(4) Let $\xi \in L^1(\mathscr{F})$. Then

$$\mathsf{E}\mathsf{E}(\xi|\mathscr{H}) = \mathsf{E}\xi$$

(law of total expectation).
Prof. Put $A = \Omega$ in the definition of conditional expectation:

$$\mathsf{E}\xi = \mathsf{E}(\xi I_\Omega) = \mathsf{E}(\mathsf{E}(\xi|\mathscr{H})I_\Omega) = \mathsf{E}\mathsf{E}(\xi|\mathscr{H}).$$

Properties of conditional expectation

(5) Let $\xi \in L^1(\mathscr{F})$ and $\mathscr{H} \subset \mathscr{G} \subset \mathscr{F}$. Then

$$\mathsf{E}(\mathsf{E}(\xi|\mathscr{G})|\mathscr{H}) = \mathsf{E}(\xi|\mathscr{H})$$

Telescopic property, (tower property, law of iterated expectations): generalizes (4).

Proof. We need to proof that

$$\mathsf{E}(\xi I_A) = \mathsf{E}(\mathsf{E}(\mathsf{E}(\xi|\mathscr{G})|\mathscr{H})I_A), \quad A \in \mathscr{H}.$$

Use (2) and (4) in right-hand side:

$$\mathsf{E}(\mathsf{E}(\mathsf{E}(\xi|\mathscr{G})|\mathscr{H})I_A) \overset{(2)}{=} \mathsf{E}(\mathsf{E}(\mathsf{E}(\xi|\mathscr{G})I_A|\mathscr{H})) \overset{(4)}{=} \mathsf{E}(\mathsf{E}(\xi|\mathscr{G})I_A)$$
$$\overset{(2)}{=}\mathsf{E}(\mathsf{E}(\xi I_A|\mathscr{G})) \overset{(4)}{=} \mathsf{E}(\xi I_A).$$

Properties of conditional expectation

(6) Let $\xi \in L^2(\mathscr{F})$. Then $\eta = \mathsf{E}(\xi|\mathscr{H}) \in L^2(\mathscr{H})$ is the best L^2-approximation of ξ on the basis of information contained in $\mathscr{H}$:

$$\mathsf{E}(\xi - \eta)^2 = \min_{\zeta \in L^2(\mathscr{H})} \mathsf{E}(\xi - \zeta)^2.$$

In other words, $\xi \mapsto \mathsf{E}(\xi|\mathscr{H})$ is the orthogonal projection on $L^2(\mathscr{H})$.

Proof.

$$\begin{aligned}
\mathsf{E}((\xi - \zeta)^2|\mathscr{H}) &= \mathsf{E}(\xi^2|\mathscr{H}) - 2\zeta\mathsf{E}(\xi|\mathscr{H}) + \zeta^2 \\
&= \mathsf{E}(\xi^2|\mathscr{H}) + (\mathsf{E}(\xi|\mathscr{H}) - \zeta)^2 - (\mathsf{E}(\xi|\mathscr{H}))^2, \\
\mathsf{E}\mathsf{E}((\xi - \zeta)^2|\mathscr{H}) &= \mathsf{E}(\mathsf{E}(\xi|\mathscr{H}) - \zeta)^2 + \mathsf{E}\xi^2 - \mathsf{E}(\mathsf{E}(\xi|\mathscr{H}))^2.
\end{aligned}$$

The minimal value is attained at $\zeta^* = \mathsf{E}(\xi|\mathscr{H})$.

Example

Let us find $\mathsf{E}(S_{t+1}|\mathscr{F}_t)$, $\mathsf{E}S_{t+1}$ in the binomial model.
In the formual

$$S_{t+1} = S_t(uI_{\{a_{t+1}=H\}} + dI_{\{a_{t+1}=T\}}),$$

the first multiplier is $\mathscr{F}_t$-measurable, and the second one does not depend on $\mathscr{F}_t$. Hence,

$$\begin{aligned}
\mathsf{E}(S_{t+1}|\mathscr{F}_t) &\overset{(2)}{=} S_t\mathsf{E}(uI_{\{a_{t+1}=H\}} + dI_{\{a_{t+1}=T\}}|\mathscr{F}_t) \\
&\overset{(3)}{=} S_t\mathsf{E}(uI_{\{a_{t+1}=H\}} + dI_{\{a_{t+1}=T\}}) \\
&= S_t(u\mathsf{E}I_{\{a_{t+1}=H\}} + d\mathsf{E}I_{\{a_{t+1}=T\}}) = (up+dq)S_t. \\
\mathsf{E}S_{t+1} &\overset{(4)}{=} \mathsf{E}\mathsf{E}(S_{t+1}|\mathscr{F}_t) = (up+dq)\mathsf{E}S_t = (up+dq)^2\mathsf{E}S_{t-1} \\
&= (up+dq)^{t+1}S_0.
\end{aligned}$$

Conditional expectation with respect to a random variable

Let $\xi \in L^1(\mathscr{F})$. Be the definition,

$$\mathsf{E}(\xi|\eta) = \mathsf{E}(\xi|\mathscr{F}_\eta),$$

where $\mathscr{F}_\eta := \sigma(\eta) = \{\eta^{-1}(B), B \in \mathscr{B}(\mathbb{R})\}$ is the σ-algebra, generated by η.

Example

Let $\eta : \Omega \mapsto \{z_1, \ldots, z_m\}$. Then

$$\mathscr{F}_\eta = \sigma(A_1, \ldots, A_m), \quad A_i = \eta^{-1}(\{z_i\}).$$

$$\mathsf{E}(\xi|\eta) = \sum_{i=1}^{m} \frac{\mathsf{E}(\xi I_{A_i})}{\mathsf{P}(A_i)} I_{A_i} = \sum_{i=1}^{m} \frac{\mathsf{E}(\xi I_{\{\eta=z_i\}})}{\mathsf{P}(\eta = z_i)} I_{\{\eta=z_i\}}.$$

- For a random variable η with finite number of values we obtained the formula
$$\mathsf{E}(\xi|\eta) = \sum_{i=1}^{m} c_i I_{\{\eta=z_i\}} = g(\eta).$$
- It is known that if ζ is $\sigma(\eta)$-measurable, then $\zeta = g(\eta)$ for some Borel function g (*Doob-Dynkin lemma*).
- It follows that for any r.v. η the there exists a Borel function g such that $\mathsf{E}(\xi|\eta) = g(\eta)$. The following notation is used:
$$\mathsf{E}(\xi|\eta = x) = g(x).$$
So we obtain the possibility to calculate the conditional expectation with respect to events with zero probabilities.
- g is not unique. We could defined it in an arbitrary way outside the set $\{z_1, \ldots, z_m\}$.

Example

Let $\xi_1, \ldots, \xi_n$ be i.i.d integrable random variables. Put $\xi = \xi_1 + \cdots + \xi_n$ and find $\mathsf{E}(\xi_1|\xi)$, $\mathsf{E}(\xi|\xi_1)$. From the symmetry,

$$\mathsf{E}(\xi_1|\xi) = \cdots = \mathsf{E}(\xi_n|\xi).$$

From the definition it follows that

$$\mathsf{E}(\xi_1|\xi) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}(\xi_i|\xi) = \frac{1}{n}\mathsf{E}(\xi_1 + \cdots + \xi_n|\xi) = \frac{1}{n}\mathsf{E}(\xi|\xi) = \frac{\xi}{n}.$$

Furthermore,

$$\begin{aligned}\mathsf{E}(\xi|\xi_1) &= \mathsf{E}(\xi_2 + \cdots + \xi_n|\xi_1) + \mathsf{E}(\xi_1|\xi_1) = \mathsf{E}(\xi_2 + \cdots + \xi_n) + \xi_1 \\ &= (n-1)\mathsf{E}\xi_1 + \xi_1.\end{aligned}$$

Example

Let $\xi_1, \ldots, \xi_n$ be i.i.d integrable random variables. Put $\xi = \xi_1 + \cdots + \xi_n$ and find $\mathsf{E}(\xi_1|\xi)$, $\mathsf{E}(\xi|\xi_1)$. From the symmetry,

$$\mathsf{E}(\xi_1|\xi) = \cdots = \mathsf{E}(\xi_n|\xi).$$

From the definition it follows that

$$\mathsf{E}(\xi_1|\xi) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}(\xi_i|\xi) = \frac{1}{n}\mathsf{E}(\xi_1 + \cdots + \xi_n|\xi) = \frac{1}{n}\mathsf{E}(\xi|\xi) = \frac{\xi}{n}.$$

Furthermore,

$$\begin{aligned}\mathsf{E}(\xi|\xi_1) &= \mathsf{E}(\xi_2 + \cdots + \xi_n|\xi_1) + \mathsf{E}(\xi_1|\xi_1) = \mathsf{E}(\xi_2 + \cdots + \xi_n) + \xi_1 \\ &= (n-1)\mathsf{E}\xi_1 + \xi_1.\end{aligned}$$

Example

Let $\xi_1, \ldots, \xi_n$ be i.i.d integrable random variables. Put $\xi = \xi_1 + \cdots + \xi_n$ and find $\mathsf{E}(\xi_1|\xi)$, $\mathsf{E}(\xi|\xi_1)$. From the symmetry,

$$\mathsf{E}(\xi_1|\xi) = \cdots = \mathsf{E}(\xi_n|\xi).$$

From the definition it follows that

$$\mathsf{E}(\xi_1|\xi) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}(\xi_i|\xi) = \frac{1}{n}\mathsf{E}(\xi_1 + \cdots + \xi_n|\xi) = \frac{1}{n}\mathsf{E}(\xi|\xi) = \frac{\xi}{n}.$$

Furthermore,

$$\begin{aligned}\mathsf{E}(\xi|\xi_1) &= \mathsf{E}(\xi_2 + \cdots + \xi_n|\xi_1) + \mathsf{E}(\xi_1|\xi_1) = \mathsf{E}(\xi_2 + \cdots + \xi_n) + \xi_1 \\ &= (n-1)\mathsf{E}\xi_1 + \xi_1.\end{aligned}$$

The independence (freezing) lemma

Let ξ be independent from η and $g : \mathbb{R}^2 \mapsto \mathbb{R}$ be a (bounded) Borel function. Then

$$\mathsf{E}(g(\xi,\eta)|\eta) = h(\eta), \quad h(y) = \mathsf{E}g(\xi,y).$$

Proof for indicators. $g = I_{B_1 \times B_2}$:

$$\begin{aligned}
\mathsf{E}(g(\xi,\eta)|\eta) &= \mathsf{E}(I_{B_1\times B_2}(\xi,\eta)|\eta) = \mathsf{E}(I_{B_1}(\xi)I_{B_2}(\eta)|\eta) \\
&= I_{B_2}(\eta)\mathsf{E}(I_{B_1}(\xi)|\eta) = I_{B_2}(\eta)\mathsf{E}(I_{B_1}(\xi)) \\
&= \mathsf{E}(I_{B_1}(\xi)I_{B_2}(y))|_{y=\eta} = h(\eta), \\
h(y) &= \mathsf{E}(I_{B_1}(\xi)I_{B_2}(y)) = \mathsf{E}g(\xi,y).
\end{aligned}$$

Example

Let ξ and η be independent exponential r.v. with parameters λ and μ respectively. Let us find $\mathsf{P}(\xi \le \eta)$:

$$\mathsf{P}(\xi \le \eta) = \mathsf{E}\mathsf{E}(I_{\{\xi \le \eta\}} | \eta) = \mathsf{E}h(\eta),$$

$$h(y) = \mathsf{E}I_{\{\xi \le y\}} = \mathsf{P}(\xi \le y) = \int_0^y \lambda e^{-\lambda x}\, dx = -e^{-\lambda x}\Big|_{x=0}^{x=y} = 1 - e^{-\lambda y},$$

$$\mathsf{P}(\xi \le \eta) = 1 - \mathsf{E}e^{-\lambda\eta} = 1 - \int_0^\infty e^{-\lambda y}\mu e^{-\mu y}\, dy = 1 - \frac{\mu}{\lambda+\mu} = \frac{\lambda}{\lambda+\mu}.$$

Transition probability

A function $Q : \mathbb{R}^m \times \mathscr{B}(\mathbb{R}^d) \mapsto [0,1]$ is called the *transition probability* from $\mathbb{R}^m$ to $\mathbb{R}^d$, if

- for any $B \in \mathscr{B}(\mathbb{R}^d)$ the function $y \mapsto Q(y, B)$ is $\mathscr{B}(\mathbb{R}^m)$-measurable;
- for any $y \in \mathbb{R}^m$ the function $B \mapsto Q(y, B)$ is a probability measure on $\mathscr{B}(\mathbb{R}^d)$.

Conditional distribution

Consider the random variables $\xi : \Omega \mapsto \mathbb{R}^d$, $\eta : \Omega \mapsto \mathbb{R}^m$. A transition probability $\mathsf{P}_{\xi|\eta}(B|y) = Q(y, B)$ from $\mathbb{R}^m$ to $\mathbb{R}^d$ is called a *(regular) conditional distribution of* ξ *with respect to* η, if

$$\mathsf{E}(g(\xi)|\eta)(\omega) = \int_{\mathbb{R}^d} g(x)\, \mathsf{P}_{\xi|\eta}(dx|\eta) \quad \mathsf{P}\text{-a.s.}$$

for any bounded Borel function on $\mathbb{R}^d$. Regular conditional distribution always exists.
The definition above also can be written as follows:

$$\mathsf{E}(g(\xi, \eta)|\eta) = \int_{\mathbb{R}^d} g(x, \eta)\mathsf{P}_{\xi|\eta}(dx|\eta).$$

Analysis of two-dimensional discrete random variables

Consider two dimensional random variable $(\xi, \eta) : \Omega \mapsto \mathbb{R}^2$ with the distribution

$$\mathsf{P}_{\xi,\eta}(B) = \mathsf{P}((\xi, \eta) \in B).$$

Assume that there exists a finite or countable set $Z \subset \mathbb{R}^2$ such that

$$\mathsf{P}((\xi, \eta) \in Z) = \mathsf{P}_{\xi,\eta}(Z) = 1.$$

Let $p_{\xi,\eta}(x, y) = \mathsf{P}_{\xi,\eta}(x, y)$ be the correspondent p.m.f. By the change of variables formula

$$\mathsf{E}g(\xi, \eta) = \sum_{(x,y)\in Z} g(x, y)p_{\xi,\eta}(x, y).$$

Let us find $\mathsf{E}(g(\xi)|\eta)$ and the conditional expectation of ξ with respect to η.

By the general theory,

$$\mathsf{E}(g(\xi)|\eta) = r(\eta).$$

From the definition of conditional expectation it follows that

$$\mathsf{E}(g(\xi)h(\eta)) = \mathsf{E}(r(\eta)h(\eta))$$

for any bounded Borel function h. By the change of variables formula,

$$\mathsf{E}(g(\xi)h(\eta)) = \sum_{x,y} g(x)h(y)p_{\xi,\eta}(x,y)$$

$$\mathsf{E}(r(\eta)h(\eta)) = \sum_{y} r(y)h(y)p_{\eta}(y),$$

where $p_\eta(y)$ is the marginal distribution of η. Since h is arbitrary, we get

$$r(y)p_{\eta}(y) = \sum_{x} g(x)p_{\xi,\eta}(x,y).$$

Hence,

$$\mathsf{E}(g(\xi)|\eta) = r(\eta) = \sum_x g(x) \frac{p_{\xi,\eta}(x,\eta)}{p_\eta(\eta)}.$$

Introduce the *conditional probability mass*

$$p_{\xi|\eta}(x|y) = \frac{p_{\xi,\eta}(x,y)}{p_\eta(y)} \left(= \frac{\mathsf{P}(\xi = x, \eta = y)}{\mathsf{P}(\eta = y)} \right).$$

Then

$$\mathsf{E}(g(\xi)|\eta) = \sum_x g(x) p_{\xi|\eta}(x|\eta).$$

Hence,

$$\mathsf{P}_{\xi|\eta}(A|y) = \sum_{x \in A} p_{\xi|\eta}(x|y)$$

is the conditional distribution of ξ with respect to η.

Total probability formula

$$p_\xi(x) = \mathsf{P}_\xi(x) = \mathsf{E} I_{\{\xi=x\}} = \mathsf{E}\mathsf{E}(I_{\{\xi=x\}}|\eta) = \mathsf{E} p_{\xi|\eta}(x|\eta)$$
$$= \sum_y p_{\xi|\eta}(x|y)p_\eta(y).$$

Total expectation formula

$$\mathsf{E} g(\xi) = \mathsf{E}\mathsf{E}(g(\xi)|\eta) = \mathsf{E}\sum_x g(x)p_{\xi|\eta}(x|\eta) = \sum_y \sum_x g(x)p_{\xi|\eta}(x|y)p_\eta(y).$$

The Bayes formula

$$p_{\xi,\eta}(x,y) = p_{\xi|\eta}(x|y)p_\eta(y) = p_{\eta|\xi}(y|x)p_\xi(x),$$

$$p_{\eta|\xi}(y|x) = \frac{p_{\xi|\eta}(x|y)p_\eta(y)}{p_\xi(x)} = \frac{p_{\xi|\eta}(x|y)p_\eta(y)}{\sum_{y'} p_{\xi|\eta}(x|y')p_\eta(y')}.$$

Analysis of two dimensional continuous random variables

Consider some two dimensional continuous random variable $(\xi, \eta) : \Omega \mapsto \mathbb{R}^2$ with the distribution

$$\mathsf{P}_{\xi,\eta}(B) = \mathsf{P}((\xi, \eta) \in B).$$

Assume that there exists a p.d.f. $f_{\xi,\eta} : \mathbb{R}^2 \mapsto \mathbb{R}_+$ such that

$$\mathsf{P}((\xi, \eta) \in B) = \int_B f_{\xi,\eta}(x, y)\, dxdy.$$

By the change of variables formula

$$\mathsf{E}g(\xi, \eta) = \int_{\mathbb{R}^2} g(x, y) f_{\xi,\eta}(x, y)\, dxdy.$$

Let us find $\mathsf{E}(g(\xi)|\eta)$ and the conditional distribution ξ with respect to η.

$$\mathsf{E}(g(\xi)|\eta) = r(\eta)$$

$$\mathsf{E}(g(\xi)h(\eta)) = \mathsf{E}(r(\eta)h(\eta)),$$

for any bounded Borel function h. By the change of variables formula,

$$\begin{aligned}\mathsf{E}(g(\xi)h(\eta)) &= \int_{\mathbb{R}^2} g(x)h(y)f_{\xi,\eta}(x,y)\,dxdy \\ &= \int_{\mathbb{R}} h(y)\left(\int_{\mathbb{R}} g(x)f_{\xi,\eta}(x,y)\,dx\right)\,dy\end{aligned}$$

$$\mathsf{E}(r(\eta)h(\eta)) = \int_{\mathbb{R}} r(y)h(y)f_\eta(y)\,dy,$$

where $f_\eta(y)$ is the marginal density of η. Since h is arbitrary, we get

$$r(y)f_\eta(y) = \int_{\mathbb{R}} g(x)f_{\xi,\eta}(x,y)\,dx.$$

Hence,

$$\mathsf{E}(g(\xi)|\eta) = r(\eta) = \int_{\mathbb{R}} g(x) \frac{f_{\xi,\eta}(x,\eta)}{f_\eta(\eta)}\, dx.$$

Let us introduce the *conditional densitпiS*

$$f_{\xi|\eta}(x|y) = \frac{f_{\xi,\eta}(x,y)}{f_\eta(y)}.$$

Then

$$\mathsf{E}(g(\xi)|\eta) = \int_{\mathbb{R}} g(x) f_{\xi|\eta}(x|\eta)\, dx.$$

Hence,

$$\mathsf{P}_{\xi|\eta}(A|y) = \int_{x\in A} f_{\xi|\eta}(x|y)\, dx$$

is the coditional distribution of ξ with respect to η.

Total probability formula

$$f_\xi(x) = \int_{\mathbb{R}} f_{\xi,\eta}(x,y)\,dy = \int_{\mathbb{R}} f_{\xi|\eta}(x|y) f_\eta(y)\,dy$$

Total expectation formula

$$\begin{aligned}
\mathsf{E}g(\xi) &= \mathsf{E}\mathsf{E}(g(\xi)|\eta) = \mathsf{E}\int_{\mathbb{R}} g(x) f_{\xi|\eta}(x|\eta)\,dx \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x) f_{\xi|\eta}(x|y)\,dx \right) f_\eta(y)\,dy \\
&= \int_{\mathbb{R}} \mathsf{E}(g(\xi)|\eta = y) f_\eta(y)\,dy
\end{aligned}$$

The Bayes formula

$$f_{\xi,\eta}(x,y) = f_{\xi|\eta}(x|y)f_\eta(y) = f_{\eta|\xi}(y|x)f_\xi(x),$$

$$f_{\eta|\xi}(y|x) = \frac{f_{\xi|\eta}(x|y)f_\eta(y)}{f_\xi(x)} = \frac{f_{\xi|\eta}(x|y)f_\eta(y)}{\int_{\mathbb{R}} f_{\xi|\eta}(x|y')f_\eta(y')\,dy'}.$$

Freezing lemma for conditional distributions

- Let ξ, η be independent, then the conditional distribution $\zeta = \psi(\xi, \eta)$ with respect to $\eta = y$ coincides with the distribution of the random variable $\psi(\xi, y)$:

$$\mathsf{P}_{\zeta|\eta}(dz|y) = \mathsf{P}_{\psi(\xi,y)}(dz).$$

Proof.

$$\mathsf{E}(g(\psi(\xi,\eta))|\eta) = \mathsf{E}g(\psi(\xi,y))\Big|_{y=\eta} = \int_{\mathbb{R}} g(z)\mathsf{P}_{\psi(\xi,y)}(dz)\Big|_{y=\eta}$$
$$= \int_{\mathbb{R}} g(z)\mathsf{P}_{\zeta|\eta}(dz|y)\Big|_{y=\eta}.$$

Example: signal detection

One-dimensional signal $\eta \sim N(0,1)$. Noise $w \sim N(0,\sigma^2)$ is independent from η. Only the sum $\xi = \eta + w$ is observable. Find the conditional distribution η with respect to ξ.

Solution. To find $f_{\xi|\eta}(x|y)$ fix (freeze) η and wright down the distribution of $y + w$:

$$f_{\xi|\eta}(x|y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-y)^2/(2\sigma^2)}.$$

By the Bayes formula,

$$f_{\eta|\xi}(y|x) = c_1(x) f_{\xi|\eta}(x|y) f_\eta(y) = c_2(x) \exp\left(-\frac{1}{2}\left(y^2 + \frac{(x-y)^2}{\sigma^2}\right)\right)$$

$$= c_3(x) \exp\left(-\frac{1+\sigma^2}{2\sigma^2}\left(y - \frac{x}{1+\sigma^2}\right)^2\right),$$

Hence for fixed $\xi = x$ the signal $\eta \sim N\left(\frac{x}{1+\sigma^2}, \frac{\sigma^2}{1+\sigma^2}\right)$.

ξ is continuous, η is discrete

Let ξ be continuous with the p.d.f. $f_\xi(x)$, and let η be discrete with the p.m.f. $p_\eta(y)$. *Introduce* the conditional density $f_{\xi|\eta}(x|y)$,

$$f_{\xi|\eta}(x|y) \geq 0, \quad \int_{\mathbb{R}} f_{\xi|\eta}(x|y)\,dx = 1.$$

In other words, assume that the conditional distribution of ξ with respect to η is determined by the formula

$$\mathsf{E}(g(\xi,\eta)|\eta) = \int_{\mathbb{R}} g(x,\eta) f_{\xi|\eta}(x|\eta)\,dx.$$

Let us find the joint distribution:

$$\begin{aligned}\mathsf{E}g(\xi,\eta) = \mathsf{E}\mathsf{E}(g(\xi,\eta)|\eta) &= \mathsf{E}\int_{\mathbb{R}} g(x,\eta) f_{\xi|\eta}(x|\eta)\,dx \\ &= \sum_y \int_{\mathbb{R}} g(x,y) f_{\xi|\eta}(x|y)\,dx \cdot p_\eta(y).\end{aligned}$$

In particular we get the total expectation formula:

$$\mathsf{E}g(\xi) = \sum_y \int_{\mathbb{R}} g(x) f_{\xi|\eta}(x|y)\, dx \cdot p_\eta(y) = \int_{\mathbb{R}} g(x) \sum_y f_{\xi|\eta}(x|y) p_\eta(y)\, dx$$

and the total probability formula:

$$f_\xi(x) = \sum_y f_{\xi|\eta}(x|y) p_\eta(y).$$

Let us find $\mathsf{E}(g(\eta)|\xi) = r(\xi)$. By the definition,

$$\mathsf{E}(g(\eta)h(\xi)) = \mathsf{E}(r(\xi)h(\xi))$$

for any bounded Borel function h.

$$\mathsf{E}(g(\eta)h(\xi)) = \sum_y \int_{\mathbb{R}} g(y)h(x)f_{\xi|\eta}(x|y)\,dx \cdot p_\eta(y)$$

$$= \int_{\mathbb{R}} \left(\sum_y g(y)f_{\xi|\eta}(x|y)p_\eta(y) \right) h(x)\,dx,$$

$$\mathsf{E}(r(\xi)h(\xi)) = \int_{\mathbb{R}} r(x)h(x)f_\xi(x)\,dx.$$

Hence,

$$r(x) = \sum_y g(y) \frac{f_{\xi|\eta}(x|y)p_\eta(y)}{f_\xi(x)},$$

$$\mathsf{E}(g(\eta)|\xi) = r(\xi) = \sum_y g(y) \frac{f_{\xi|\eta}(\xi|y)p_\eta(y)}{f_\xi(x)} = \sum_y g(y)p_{\eta|\xi}(y|\xi),$$

where the conditional density $p_{\eta|\xi}(y|x)$ is determined by the Bayes formula:

$$p_{\eta|\xi}(y|x) = \frac{f_{\xi|\eta}(x|y)p_\eta(y)}{f_\xi(x)} = \frac{f_{\xi|\eta}(x|y)p_\eta(y)}{\sum_{y'} f_{\xi|\eta}(x|y')p_\eta(y')}.$$

Example: signal detection

A binary signal is described by the random variable η:

$$p_\eta(1) = p, \quad p_\eta(-1) = q = 1 - p.$$

The signal is affected by the standard Gaussian noise $w \sim N(0, 1)$, which does not depend on η. The received signal has the form

$$\xi = \eta + w.$$

Find the probability that $\eta = 1$ under the condition that the observed value of ξ equals to x.

Solution. $f_{\xi|\eta}(x|y)$ coincides with the density of the random variable $y + w \sim N(y, 1)$:

$$f_{\xi|\eta}(x|y) = \frac{1}{\sqrt{2\pi}} e^{-(x-y)^2/2}.$$

By the Bayes formula

$$p_{\eta|\xi}(y|x) = \frac{f_{\xi|\eta}(x|y)p_\eta(y)}{\sum_{y'} f_{\xi|\eta}(x|y')p_\eta(y')}.$$

In particular,

$$p_{\eta|\xi}(1|x) = \frac{e^{-(x-1)^2/2}p}{e^{-(x-1)^2/2}p + e^{-(x+1)^2/2}q} = \frac{p}{p + qe^{-2x}}.$$

Note that the function $x \mapsto p_{\eta|\xi}(1|x)$ is increasing and

$$\lim_{x\to-\infty} p_{\eta|\xi}(1|x) = 0, \quad \lim_{x\to\infty} p_{\eta|\xi}(1|x) = 1.$$

ξ is discrete, η is continuous

Let ξ be discrete with the p.m.f. $p_\xi(x)$, and η is continuous with the density $f_\eta(y)$. *Introduce* the conditional p.m.f. $p_{\xi|\eta}(x|y)$:

$$p_{\xi|\eta}(x|y) \geq 0, \quad \sum_x p_{\xi|\eta}(x|y) = 1.$$

In other words, assume that the conditional distribution of ξ with respect to η is determined by the formula

$$\mathsf{E}(g(\xi,\eta)|\eta) = \sum_x g(x,\eta) p_{\xi|\eta}(x|\eta).$$

Let us find the joint distribution:

$$\mathsf{E}g(\xi,\eta) = \mathsf{E}\mathsf{E}(g(\xi,\eta)|\eta) = \mathsf{E}\sum_x g(x,\eta) p_{\xi|\eta}(x|\eta)$$

$$= \int_{\mathbb{R}} \sum_x g(x,y) p_{\xi|\eta}(x|y) f_\eta(y)\, dy.$$

A special case is the total expectation formula:

$$\mathsf{E}g(\xi) = \sum_x g(x) \int_{\mathbb{R}} p_{\xi|\eta}(x|y) f_\eta(y)\, dy$$

which implies the total probability formula

$$p_\xi(x) = \int_{\mathbb{R}} p_{\xi|\eta}(x|y) f_\eta(y)\, dy.$$

Let us find $\mathsf{E}(g(\eta)|\xi) = r(\xi)$. By the definition,

$$\mathsf{E}(g(\eta)h(\xi)) = \mathsf{E}(r(\xi)h(\xi))$$

for any bounded Borel function h.

$$\begin{aligned}
\mathsf{E}(g(\eta)h(\xi)) &= \int_{\mathbb{R}} \sum_x g(y)h(x)p_{\xi|\eta}(x|y)f_\eta(y)\,dy \\
&= \sum_x \left(\int_{\mathbb{R}} g(y)p_{\xi|\eta}(x|y)f_\eta(y)\,dy \right) h(x), \\
\mathsf{E}(r(\xi)h(\xi)) &= \sum_x r(x)h(x)p_\xi(x).
\end{aligned}$$

Hence,

$$r(x) = \int_{\mathbb{R}} g(y) \frac{p_{\xi|\eta}(x|y)f_\eta(y)}{p_\xi(x)}\,dy,$$

$$\mathsf{E}(g(\eta)|\xi) = r(\xi) = \int_{\mathbb{R}} g(y) \frac{p_{\xi|\eta}(\xi|y)f_\eta(y)}{p_\xi(\xi)}\,dy = \int_{\mathbb{R}} g(y) f_{\eta|\xi}(y|\xi)\,dy,$$

where the conditional density $f_{\eta|\xi}(y|x)$ is determined by the Bayes formula:

$$f_{\eta|\xi}(y|x) = \frac{p_{\xi|\eta}(x|y)f_\eta(y)}{p_\xi(x)} = \frac{p_{\xi|\eta}(x|y)f_\eta(y)}{\int_{\mathbb{R}} p_{\xi|\eta}(x|y')f_\eta(y')\,dy'}.$$