Table of Contents

1. Sequence. Convergence of sequences.

Let $x : \mathbb{N} \to \mathbb{R}$. Then we can say that sequence was defined and there is a valid notation: $x(n) = x_n$.

Definition

Let $\{x_n\}_{n=1}^{\infty} \subset \mathbb{R}$ some sequence, we can say that it converge to $l \in \mathbb{R}$ (or $l = \lim\limits_{n\to\infty} x$), iff:

$$(\forall \varepsilon > 0)(\exists N \in \mathbb{N})(\forall n > N)\ |x_n - l| < \varepsilon$$

Example:

$$\lim_{n\to\infty} \frac{1}{n} = 0 \Longleftrightarrow$$
$$\Leftrightarrow (\forall \varepsilon > 0)(\exists N \in \mathbb{N})(\forall n > N) : \left|\frac{1}{n}\right| < \varepsilon \Longleftrightarrow$$
$$\Leftrightarrow n > \frac{1}{\varepsilon},\ N = \left[\frac{1}{\varepsilon}\right] + 1.$$
$$(\forall \varepsilon > 0)\left(N = \left[\frac{1}{\varepsilon}\right] + 1 \in \mathbb{N}\right)(\forall n > N) n > N \longrightarrow$$
$$\to n > \frac{1}{\varepsilon} \Rightarrow \frac{1}{n} < \varepsilon.$$

Theorem

Numeric sequence can't have more than one limit.

Theorem: Properties of limit of consequence

Let $\{x_n\}_{n=1}^{\infty}$ some sequence. We can define some properties of it:

- if $\{x_n\}_{n=1}^{\infty}$ converges then $\{x_n\}_{n=1}^{\infty}$ is bounded;
- if $\lim\limits_{n\to\infty} x_n = l \neq 0$, then
$$(\exists N \in \mathbb{N})(\forall n > N)$$
$$(\operatorname{sgn}(x_n) = \operatorname{sgn}(l)) \wedge |x_n| > \frac{|l|}{2};$$
- if $\lim\limits_{n\to\infty} x_n = l_1$, $\lim\limits_{n\to\infty} y_n = l_2$:
$$(\forall n \in \mathbb{N})\ x_n \leq y_n \Rightarrow l_1 \leq l_2$$
- if $\lim\limits_{n\to\infty} x_n = \lim\limits_{n\to\infty} z_n = l$:
$$(\forall n \in \mathbb{N}) x_n \leq y_n \leq z_n$$
then $\lim\limits_{n\to\infty} y_n = 1$.

Theorem: Arithmetic operations with limits

If $\lim\limits_{n\to\infty} x_n = l_1$, $\lim\limits_{n\to\infty} y_n = l_2$, then:

- $x_n \pm y_n$ converges to $l_1 \pm l_2$;
- $x_n \cdot y_n$ converges to $l_1 \cdot l_2$;
- if in addition $y_n \neq 0$, then $(\forall n \in \mathbb{N})$, $l_2 \neq 0$, then $\dfrac{x_n}{y_n}$ converges to $\dfrac{l_1}{l_2}$.

Definition: Infinitesimal

Infinitesimal sequence is called sequence converged to zero.

Theorem

Product of an infinitesimal sequence and bounded one is infinitesimal.

Definition

Infinitely large sequence is called a sequence with infinite limit.

Theorem

$\{x_n\}_{n=1}^{\infty} \subset \mathbb{R}\backslash\{0\}$ infinitesimal iff $\left\{\dfrac{1}{x_n}\right\}_{n=1}^{\infty}$ infinitely large.

Lemma

$$(\forall x \geq -1)(\forall n \in \mathbb{N})\ (1 + x)^n \geq 1 + nx$$

Theorem

Sequence $x_n = \left(1 + \dfrac{1}{n}\right)^n$ converges and its limit named e.

2. The exponential. Logarithms.

Let's define exponential and logarithm:

Definition: (Exponential)

Let function $f : \mathbb{R} \to \mathbb{R}$ such that:

$$f(x) = b^{kx},$$

with base $b > 0$ and constant k is called the exponential function.

Definition: (Logarithm)

A logarithm is an exponent which indicates to what power a base must be raised to produce a given number, i.e.:

$$\begin{aligned} y &= b^x && \text{exponential form} \\ x &= \log_b y && \text{logarithmic form} \end{aligned}$$

Note

x is the logarithm of y to the base b, $\log_b y$ is the power to wich we have to raise b to get y.

Common (Briggsian) logarithms (logarithms with base 10). Notation:

$$\log_{10} y = \log y$$

Natural (Naperian) logarithms (logarithms with base e). Notation:

$$\log_e x = \ln x$$

Theorem: (Properties of exponents and logarithms)

$$b^m \cdot b^n \equiv b^{m+n}$$
$$\frac{b^m}{b^n} \equiv b^{m-n}$$
$$(b^m)^n$$

$$\log_b(yz) = \log_b y + \log_b z$$
$$\log_b\left(\frac{y}{z}\right) = \log_b y - \log_b z$$

Note

Other properties:

$$\ln x^y = y \cdot \ln x$$
$$\ln e^x = x, \quad e^{\ln x} = x$$
$$\log_b a = \frac{\ln a}{\ln b} = \frac{\log a}{\log b}$$

3. Derivative and basic differential skills

Definition of derivative:

Definition: (Derivative)

Let function f differentiable at a point a of its domain, if its domain contains an open interval containing a, and the limit:

$$\lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(a + \Delta x) - f(a)}{\Delta x} = f'(a),$$

where Δx is an increment of the argument and Δf the same for function, exists. And $f'(a)$ is called a derivative.

Theorem

If $\exists f'(a)$ then f is continuous function at a point.

Note

But the converse is not generally true.

Example: $f(x) = |x|$, $a = 0$. The limit does not exist, because:

$$\lim_{\Delta x \to +0} \frac{f(\Delta x)}{\Delta x} = 1; \quad \lim_{\Delta x \to -0} \frac{f(\Delta x)}{\Delta x} = -1.$$

Theorem: (Arythmetic operations with derivatives)

If $\exists f'(x_0)$ and $g'(x_0)$, then $\exists$ at the point x_0 : $f \pm g$, $f \cdot g$ and $\dfrac{1}{g}$ with additional condition

$g(x_0) \neq 0$, such that:

$$(f \pm g)'(x_0) = f'(x_0) + g'(x_0);$$
$$(f \cdot g)'(x_0) = f'(x_0) \cdot g(x_0) + f(x_0) \cdot g'(x_0)$$
$$\left(\frac{f}{g}\right)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{g^2(x_0)}$$

Theorem: (Basic derivatives)

$$(\sin x)' = \cos x \qquad (\sinh x)' = \cosh x$$
$$(\cos x)' = -\sin x \qquad (\cosh x)' = \sinh x$$
$$(\tan x)' = \frac{1}{\cos^2 x} = \sec^2 x$$
$$(\tan x)' = \frac{1}{\cos^2 x} = \sec^2 x$$
$$(\tanh x)' = \frac{1}{\cosh^2 x}$$
$$(\cot x)' = -\frac{1}{\sin^2 x} = -\csc^2 x$$
$$(\coth x)' = -\frac{1}{\sinh^2 x}$$
$$(x^a)' = ax^{a-1} \qquad (a^x)' = a^x \ln a$$

Theorem: (Some n-th derivatives)

$$(a^x)^{(n)} = a^x \ln^n a$$
$$(\sin x)^{(n)} = \sin\left(x + \frac{\pi n}{2}\right)$$
$$(\cos x)^{(n)} = \cos\left(x + \frac{\pi n}{2}\right)$$
$$(x^a)^{(n)} = a \cdot (a-1) \dots (a-n+1) \cdot x^{a-n},$$
$$(a \notin \mathbb{N}) \vee (a \in \mathbb{N},\ a \geq n)$$
$$(\ln(1+x))^{(n)} = (-1)^{n+1}(n-1)!(1+x)^{-1}.$$

Note

Chain rule:

$$(f(g(x)))' = f'(g(x))\, g'(x)$$

In the below, $u = f(x)$ is a function of x. These rules are all generalizations of the above rules using the chain rule:

1. $\dfrac{d}{dx}(u^n) = nu^{n-1}\dfrac{du}{dx};$
2. $\dfrac{d}{dx}(a^u) = a^u \ln(a)\dfrac{du}{dx};$
3. $\dfrac{d}{dx}(e^u) = e^u\dfrac{du}{dx};$
4. $\dfrac{d}{dx}(\log_a(u)) = \dfrac{1}{x\ln(u)}\dfrac{du}{dx};$
5. $\dfrac{d}{dx}(\ln(u)) = \dfrac{1}{u}\dfrac{du}{dx};$
6. $\dfrac{d}{dx}(\sin(u)) = \cos(u)\dfrac{du}{dx};$
7. $\dfrac{d}{dx}(\cos(u)) = -\sin\dfrac{du}{dx};$
8. $\dfrac{d}{dx}(\tan(u)) = \sec^2(u)\dfrac{du}{dx};$
9. $\dfrac{d}{dx}\left(\tan^{-1}(u)\right) = \dfrac{1}{1+u^2}\dfrac{du}{dx}$

Implicit differentiation

Use whenever you need to take the derivative of a function that is implicity defined. Steps for solving:

1. Differentiate both sides of the equation with respect to x;
2. When taking the derivative of any term that has a y in it multiply the term by y';
3. Solve for y'.

When finding the second derivative y'', remember to replace any y' terms in your final answer with the equation for y' you already found. In other words, your final answer should not have any y' terms in it.

Log differentiation

Two cases when this method is used:

- Use whenever you can take advantage of log laws to make a hard problem easier:
  - Examples: $\frac{(x^3+x)\cos x}{x^2+1}$ or $\ln(x^2+1)\cos(x)\tan^{-1}(x)$, etc.
  - Note that in the above examples, log differentiation is not required but makes taking the derivative easier.
- Use whenever you are to differentiate:

$$\frac{d}{dx}\left(f(x)^{g(x)}\right)$$

There is no other way to take such derivatives.

Steps:

1. Take the ln of both sides;
2. Simplify the problem using log laws;
3. Take the derivative of both sides;
4. Solve for y'.

4. Sum of the series. Integral

5. Basic integration skills

Antiderivatives of basic functions

Power rule

$$\int x^n dx = \begin{cases} \frac{x^{n+1}}{n+1} + C, & \text{if } n \neq -1 \\ \ln|x| + C, & \text{otherwise.} \end{cases}$$

Exponential functions

With base a:

$$\int a^x dx = \frac{a^x}{\ln(a)} + C.$$

With base e, this becomes:

$$\int e^x dx = e^x + C.$$

If we have base e and a linear function in the exponent, then:

$$\int e^{ax+b} dx = \frac{1}{a} e^{ax+b} + C.$$

Trigonometric functions

$$\int \sin x dx = -\cos x + C, \qquad \int \cos x dx = \sin x + C,$$

$$\int \sec^2 x dx = \tan x + C, \qquad \int \csc^2 x dx = -\cot x + C,$$

$$\int \sec x \tan x dx = \sec x + C, \qquad \int \csc x \cot x dx = -\csc x + C.$$

Inverse trigonometric functions

$$\int \frac{1}{\sqrt{1-x^2}} dx = \arcsin x + C,$$

$$\int \frac{1}{x\sqrt{x^2-1}} dx = \operatorname{arcsec} x + C,$$

$$\int \frac{1}{1+x^2} dx = \arctan x + C,$$

$$\int \frac{1}{a^2+x^2} dx = \frac{1}{a} \arctan\left(\frac{x}{a}\right) + C.$$

Hyperbolic functions

$$\int \sinh x dx \cosh x + C, \qquad \int \cosh x dx = \sinh x + C,$$

$$\int \operatorname{sech}^2 x dx = \tanh x + C, \qquad \int -\operatorname{csch}^2 x dx = \coth x + C,$$

$$\int -\operatorname{csch} x \coth x dx = \operatorname{csch} x + C,$$

$$\int \operatorname{sech} x \tanh x dx = \operatorname{sech} x + C.$$

Integration techniques

u-substitution

If $u = g(x)$ is a differentiable function whose range is an interval I and f is continuous on I, then:

$$\int f(g(x))\, g'(x) dx = \int f(u) du.$$

Integration by parts

Recall the product rule:

$$\frac{d}{dx}\left[u(x)v(x)\right] = v(x)\frac{du}{dx} + u(x)\frac{dv}{dx}.$$

Integrating both sides leads to the following equation:

$$uv = \int u dv + \int v du,$$

from which one we can obtain the standard formula for integration by parts:

$$\int u dv = uv - \int v du.$$

If exists some troubles deciding what u and dv should be to accomplish an integral simplification, we can use rules "LIATE" to choose u:

- Logarithmic;
- Inverse trigonometric;
- Algebraic, i.e. polynomials and rational functions;
- Trigonometric;
- Exponential,

and then whatever is left is dv.

Trigonometric integrals

For integrals involving only powers of sine and cosine (both with the same argument):

- If at least one of them is raised to an odd power, pull of one to save for a u-substitution, use a Pythagorean identity ($\cos^2 x + \sin^2 x = 1$) to convert the remaining (not even) power to the other trigonometric function, then make a u-substitution with $u =$ (whichever trigonometric function you didn't save) and the trigonometric function you set aside will be part of du;
- If they are both raised to an even power, use a half-angle formulae $\cos^2 x = \dfrac{1+\cos 2x}{2}$ or $\sin^2 x = \dfrac{1-\cos 2x}{2}$ to convert to cosines, expand the result and apply half-angle formulas again if needed (keep doing this until you no longer have any powers of cosine), then integrate (may need a simple u-sub).

For integrals involving only powers of secant and tangent (both with the same argument):

- If the secant is raised an even power, pull off two of them to save for a u-substitution, use the Pythagorean identity ($\sec^2 x = 1 + \tan^2 x$) to convert the remaining powers to tangents, then make a u-substitution with $u = \tan x$ and the $\sec^2 x$ you set aside earlier will be part of du;
- If the tangent is raised to an odd power, pull off one of each to save for a u-substitution, use the Pythagorean identity ($\tan^2 x = \sec^2 x - 1$) to convert the remaining powers to tangent, then make a u-substitution with $u = \sec x$ and the $\sec x \tan x$ you set aside earlier will be part of du.

Trigonometric substitutions

With certain integrals we can use right triangles to help us determine a helpful substitutions:

If the integral contains an expression of the form

1. $\sqrt{a^2 - x^2}$, then make a substitution:
$$\begin{array}{c} x = a\sin\theta \\ dx = a\cos\theta d\theta \end{array};$$
2. $\sqrt{a^2 + x^2}$, then make a substitution:
$$\begin{array}{c} x = a\tan\theta \\ dx = a\sec^2\theta d\theta \end{array};$$
3. $\sqrt{x^2 - a^2}$, then make a substitution:
$$\begin{array}{c} x = a\sec\theta \\ dx = a\sec\theta\tan\theta d\theta \end{array}$$

Partial fraction decomposition

Given a rational function to integrate, follow these steps:

1. If the degree of the numerator is greater than or equal to that of the denominator perform long division;
2. Factor the denominator into unique linear factors or irreducible quadratics;

3. Split the rational function into a sum of partial fractions with unknown constants on top as follows:

$$\underbrace{\frac{A}{ax+b}}_{\text{for a linear factor}} + \underbrace{\frac{B}{cx+d} + \frac{C}{(cx+d)^2} + \ldots +}_{\text{for a repeated linear factor}}$$
$$+ \underbrace{\frac{Dx+E}{ex^2+fx+g}}_{\text{for an irreducible quadratic}};$$

4. Multiply both sides by the entire denominator and simplify;
5. Solve for the unknown constants by using a system of equations or picking appropriate numbers to substitute in for x;
6. Integrate each partial fraction.

Note

Helpful substitution:
$$\int \frac{1}{x^2+a^2}dx = \frac{1}{a}\tan^{-1}\left(\frac{x}{a}\right) + C.$$

Euler substitution

Euler substitution is a method for evaluating integrals of the form:

$$\int R(x, \sqrt{ax^2+bx+c})dx,$$

where R is a rational function.

Euler's first substitution

The first substitution of Euler is used when $a > 0$. We substitute:

$$\sqrt{ax^2+bx+c} = \pm x\sqrt{a} + t$$

and solve the resulting expression for x. We have that $x = \dfrac{c-t^2}{\pm 2t\sqrt{a} - b}$ and that the dx term is expressible rationally in t.

Euler's second substitution

If $c > 0$, we take:

$$\sqrt{ax^2+bx+c} = xt \pm \sqrt{c}.$$

We solve for x similarly as above and find:

$$x = \frac{\pm 2t\sqrt{c} - b}{a - t^2}.$$

Euler's third substitution

If the polynomial $ax^2 + bx + c$ has real roots α and β, we may choose:

$$\sqrt{ax^2+bx+c} = \sqrt{a(x-\alpha)(x-\beta)} = (x-\alpha)t.$$

This yields
$$x = \frac{a\beta - \alpha t^2}{a - y^2},$$

and as in the preceding cases, we can express the entire integrand rationally in t.

6. Indicator function

Let A be any event. Define the indicator function

$$I_A = \begin{cases} 1, & \text{if event } A \text{ occurs} \\ 0, & \text{otherwise.} \end{cases}$$

7. Continuous and discrete random variables

There are two main types of r.v.-s (random variables): discrete and continuous. Let's start from the definition of discrete random variable:

Definition: (Discrete random variable)

A random variable X is said to be discrete if there is a finite list of values $a_1, a_2, \ldots, a_n$ or an infinite list $a_1, a_2, \ldots$ such that $P(X = a_j \text{ for some } j) = 1$. If X is a discrete r.v., then this finite or countably infinite set of values it takes and such that $P(X = x) > 0$ is called the support of X.

If $X \in \mathbb{R}$ is a real-valued quantity, it is called a continuous random variable. In this case, we can no longer create a finite (or countable) set of distinct possible values it can take on. However, there are a countable number of intervals which we can partition the real line into. If we associate events with X being in each one of these intervals, we can use the methods discussed above for discrete random variables. Informally speaking, we can represent the probability of X taking on a specific real value by allowing the size of the intervals to shrink to zero, as we show below.

8. Independent random variables. Conditions of independency

9. Expectation and variance

Expectation

The mean, expected value or expectation of a random variable X is written as $\mathbb{E}(X)$ or μ_x. If we observe N random variables of X, then the mean of the N values will be approximately equal to $\mathbb{E}(X)$ for large N. The expectation is defined differently for continuous and discrete random variables.

Definition: (Expectation for continuous r.v.-s)

Let X be a continuous random variable with p.d.f. f_X. The expected value of X is:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx.$$

Definition: (Expectation for discrete r.v.-s)

Let X be a discrete random variable with probability mass function $f_X(x)$. The expected value of X is:

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i f_X(x_i)$$

Properties of expectation

Theorem: (Properties of Expectation)

- For any r.v. X and $\forall a, b \in \mathbb{R}$:
  $\mathbb{E}[aX + b] = a\mathbb{E}X + b$;
- Let X and Y be any random variables. Then:
  $$\mathbb{E}[X + Y] = \mathbb{E}X + \mathbb{E}Y$$
- Let X and Y be independent random variables. Then:
  $$\mathbb{E}[XY] = \mathbb{E}X\mathbb{E}Y.$$

Note

(for last property) The converse is not generally true.

Probability as an expectation

Let A be any event. We can write $P(A)$ as an expectation, as follows. Define the indicator function

$$I_A = \begin{cases} 1, & \text{if event } A \text{ occurs} \\ 0, & \text{otherwise.} \end{cases}$$

Then I_A is a random variable, and

$$\mathbb{E}I_A = \sum_{r=0}^{1} rP(I_A = r) = 0 \cdot P(I_A = 0) + 1 \cdot P(I_A = 1) = \\ = P(I_A = 1) = P(A).$$

Thus for any event A:

$$P(A) = \mathbb{E}I_A$$

Variance

The variance of a random variable X is a measure of how spread out it is. The variance measures how far the values of X are from their mean, on average.

Definition

Let X be any random variable. The variance of X is:

$$\mathrm{var}(X) = \mathbb{E}\left((X - \mu_X)^2\right) = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$

Theorem: (Properties of variance)

- For any r.v. X and $\forall a, b \in \mathbb{R}$:

$$\mathrm{var}\,(aX + b) = a^2\,\mathrm{var}\,X.$$

- Let X and Y be independent random variables. Then:

$$\mathrm{var}(X + Y) = \mathrm{var}\,X + \mathrm{var}\,Y$$

- If X and Y are not independent, then:

$$\mathrm{var}(X+Y) = \mathrm{var}\,X + \mathrm{var}\,Y + 2\,\mathrm{cov}(X, Y)$$

10. Covariance

Covariance is a measure of the association or dependence between two random variables X and Y. Covariance can be either positive or negative. (Variance is always positive.)

Definition: (Covariance)

Let X and Y be any r.v.-s. The covariance between X and Y is given by:

$$\mathrm{cov}(X, Y) = \mathbb{E}\left\{(X - \mu_x)\left(Y - \mu_y\right)\right\} = \\ = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y.$$

11. Random vector and covariance matrix

12. Mean. Mode. Median

Mode

The mode of a distribution is the value with the highest probability mass or probability density function:

$$x^* = \underset{x}{\mathrm{argmax}}\; p(x)$$

This may not be unique, in such cases the distribution is called multimodal Furthermore, even if there is a unique mode, this point may not be a good summary of the distribution.

13. Conditional probability. Conditional independence

Definition: (Conditional probability)

If A and B are events with $P(B) > 0$, then the conditional probability of A given B:

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

where $P(A)$ – prior probability of A, $P(A|B)$ – posterior probability of A.

Note

(Posterior is equivalent to updated based on evidence, prior – before this update)

For any event A, $P(A|A) = \dfrac{P(A \cap A)}{P(A)} = 1$ – if A occurred, our updated probability for A is 1.

Theorem: (Probability of intersection)

For any two events A and B with positive probabilities:

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A).$$

Applying this repeatedly, we get:

Theorem: (Probability of intersection of n events)

For any events $A_1, \ldots, A_n$ with

$P(A_1, A_2, \dots, A_{n-1}) > 0$:

$$P(A_1, A_2, \dots, A_n) = \\ = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdot \dots \\ \dots \cdot P(A_n|A_1, \dots, A_{n-1}),$$

where $, \equiv \cap$, e.g. $P(A_3|A_1, A_2) \equiv P(A_3, |A_1 \cap A_2)$

Definition: (Conditional independence)

Events A and B are conditionally independent given E if

$$P(A \cap B|E) = P(A|E)P(B|E).$$

Note

Conditional independence does not imply independence, nor does independence imply conditional independence.

14. Law of total expectation

Conditional expectation and conditional variance

Suppose that X and Y are discrete r.v.-s, possibly dependent on each other (the same results hold for continuous r.v.-s too, but will assume for simplicity the first one case). Suppose that we fix Y at the value y. This gives us a set of conditional probabilities $P(X = x|Y = y)$. This is called the conditional distribution of X, given that $Y = y$.

Definition

Let X and Y be discrete random variables. The conditional probability function of X, given that $Y = y$, is:

$$P(X = x|Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

Note

Notation:

$$f_{X|Y}(x|y) = P(X = x|Y = y).$$

Definition: (Conditional expectation)

Let X and Y be discrete random variables. The conditional expectation of X, given that $Y = y$, is:

$$\mu_{X|Y=y} = \mathbb{E}[X|Y = y] = \sum_x x f_{X|Y}(x|y).$$

Note

Intuition: $E[X|Y = y]$ is the mean value of X, when Y is fixed at y.

Note

Conditional expectation, $\mathbb{E}(X|Y)$, is a random variable with randomness inherited from Y, not X.

The conditional variance is similar to the conditional expectation:

- var$(X|Y = y)$ is the variance of X, when Y is fixed at the value $Y = y$;
- var$(X|Y)$ is a random variable, giving the variance of X when Y is fixed at a value to be selected randomly.

Definition: (Conditional variance)

Let X and Y be random variables. The conditional variance of X, given Y, is given by:

$$\text{var}(X|Y) = \mathbb{E}(X^2|Y) - \{\mathbb{E}(X|Y)\}^2 = \\ = \mathbb{E}\left\{(X - \mu_{X|Y})^2 |Y\right\}$$

Note

Like expectation, var $(X|Y = y)$ is a number depending on y, while var$(X|Y)$ is a random variable with randomness inherited from Y.

Law of total expectation

$$\mathbb{E}X = \mathbb{E}_Y\left[X|Y\right],$$

where $\mathbb{E}_Y$ is denoted by expectation over Y, i.e. the expectation is computed over the distribution of the random variable Y.

Note

The law of total expectation says that the total average is the average of case-by-case averages.

15. Law of total variance

Theorem: (Law of total variance)

$$\operatorname{var} X = \mathbb{E}_Y\left[\operatorname{var}(X|Y)\right] + \operatorname{var}_Y\left(\mathbb{E}[X|Y]\right),$$

where $\mathbb{E}_Y$ and var_Y denote expectation over Y and variance over Y.

The variance is computed over the distribution of the r.v. Y.

Let's rationale about the terms:

> What is $\mathbb{E}_Y\left[\operatorname{var}\left(X|Y\right)\right]$?

Is the average of the variance of X over all possible values of the random variable Y. In other words: take the variance of X in each conditional space of $Y = y$. Then, take the average of the variances. This is called the average within-sample variance.

> What is $\operatorname{var}_Y(\mathbb{E}[X|Y])$?

Note that the first term $\mathbb{E}_Y\left[\operatorname{var}\left(X|Y\right)\right]$, only considers the average of the variances of $X|Y$. That term does not take into account the movement of the mean itself, just the variation about each, possibly varying, mean.

If we treat each $Y = y$ as a separate "treatment", then the first term is measuring the average within-sample variance, while the second is measuring the between-sample variance.

16. Dirac delta function and its connection with simple constant

17. Difference between cdf and pdf

18. Sum rule

Suppose we have two r.v.-s X and Y. We can define the joint distribution of two r.v.-s using $P(x,y) = P(X = x \cap Y = y)$ for all possible values of X and Y. Given a joint distribution, we define the marginal distribution of an r.v. as follows:

$$P(X = x) = \sum_y P(X = x \cap Y = y),$$

where we are summing over all possible states of Y. This is sometimes called sum rule or the rule of total probability.

19. Product rule and chain rule of probability

Product rule

We define the conditional distribution of an r.v. using:

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

We can rearrange this equation to get:

$$P(x,y) = P(x)P(y|x)$$

Chain rule of probability

By extending the product rule to D variables, we obtain the chain rule of probability:

$$P(x_{1:D}) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\cdot\ldots\cdot P(x_D|x_{1:D-1})$$

20. Bayes theorem

Bayes' rule is a formula for computing the probability distribution over possible values of an unknown quantity H given some observed data $Y = y$:

$$P(H = h|Y = y) = \frac{P(H = h)P(Y = y|H = h)}{P(Y = y)}.$$

It easily follows from the product rule of probability:

$$P(h|y)P(y) = P(h)P(y|h) = P(h, y).$$

The term $P(H)$ represents what we know about possible values of H before we see any data: this is called the prior distribution. The term $P(Y|H = h)$ represents the distribution over the possible outcomes Y we expect to see if $H = h$; this is called the observation distribution. When we evaluate this at point corresponding to the actual observations, y, we get the function $P(Y = y|H = h)$, which is called the likelihood. Multiplying the prior distribution $P(H = h)$ by the likelihood function $P(Y|H = h)$ for each h gives the unnormalized joint distribution $P(H = h, Y = y)$. We can convert this into normalized one by dividing by $P(Y = y)$, which is known as the marginal likelihood, since it is computed by marhinalizing over unknown H:

$$\begin{aligned} P(Y = y) &= \sum_{h' \in \mathcal{H}} P(H = h')P(Y = y|H = h') = \\ &= \sum_{h' \in \mathcal{H}} P(H = h', Y = y). \end{aligned}$$

Normalizing the joint distribution by computing $\frac{P(H = h, Y = y)}{P(Y = y)}$ for each h gives the posterior distribution $P(H = h|Y = y)$; this represents our new belief state about the possible values of H. To summarize:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

21. Central limit theorem

22. Law of large numbers

23. Differences between quantiles and percentiles

24. Vectors. Vector spaces and vector fields

25. Metric axioms

26. Relationship between metrics, norms and distances

27. Metric space

28. Orthogonal vectors

29. Affine transformation

30. Linear subspace

31. Projection onto a subspace

32. Linear operator

33. Convex function