Chapter 1
A brief recap on probability

1 Probability and probability distributions

Following the use in probability theory, we shall denote with capital letters, for instance X, a random variable, and with small letters, for instance x, a particular realization of this random variable.

Definition 1 (Probability density functions) The probability density function (PDF), or density, is a non-negative function that describes the relative likelihood for this random variable to take on a given value such that

$$\int p_X(x)dx = 1 \qquad \text{Normalization}$$

$$p_X(x)dx = (X \in xx + dx)$$

$$p_X(x) \geq 0$$

Often, it is written that $X \sim p_X(x)$.

Remark 1 Careful: with continuous variable $p(x)dx$ is a probability, but NOT $p(x)$. In particular, $p(x)$ can be LARGER than 1.

Definition 2 (Cumulative density functions) Let X be a random variable. The cumulative density function is a real-valued function given by:

$$F_X(x) = \int_{-\infty}^{x} p_X(u)u = (X \leq x)$$

Definition 3 (Expectation or expected value) Let X be a random variable of probability density function $p_X(x)$, then the expected value can be computed as:

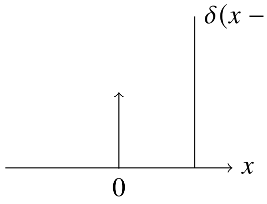
$$\mathrm{e}_X[g(X)] = \int_{-\infty}^{\infty} g(x)p_X(x)x = \overline{g(x)}$$

$\mathrm{e}_X(X)$, often noted in physics $\bar{X}$ or $\langle X \rangle$, is named the “mean”. $M_n = \mathrm{e}_X(X^n)$ is named the “n-th moment”.

Definition 4 (Variance and Standard deviation) The variance Δ and the standard deviation σ of a random variable X are defined as:

$$\Delta = \mathrm{e}[X^2] - \mathrm{e}[X]^2 \qquad \sigma = \sqrt{\Delta}$$

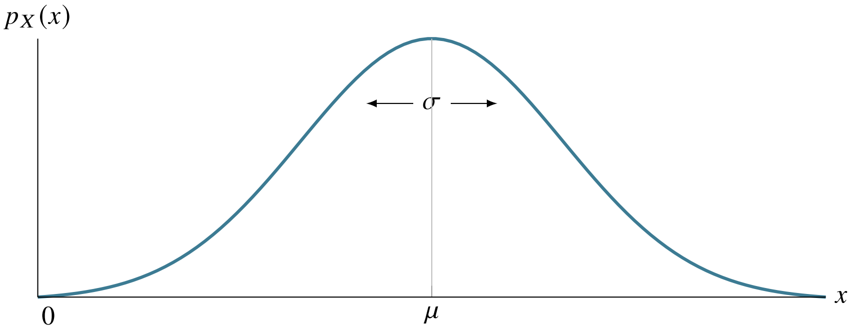
Example 1 (Dirac distribution (or point-mass)
)



In this example, the mean is 2 and the variance is 0.

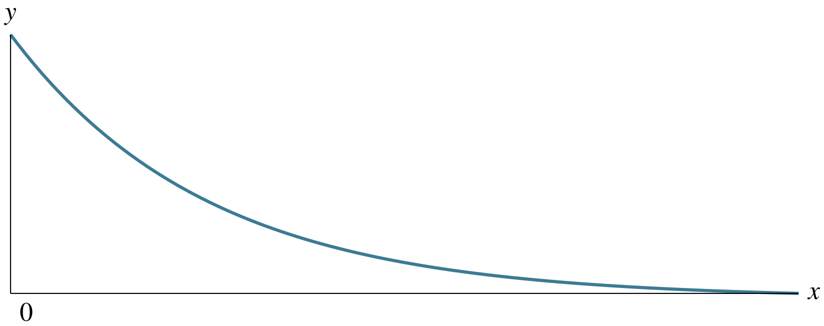
Example 2 (Gaussian (or Normal) distribution
) The gaussian distribution of mean μ and variance $\Delta = \sigma^2$ has the following form:

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu)^2}{2\sigma^2} \sim \mathcal{N}(\mu, \sigma^2)$$



Example 3 (Exponential distribution
) The exponential distribution of mean μ and a variance μ^2 has the following form:

$$p_X(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}$$



Remark 2 We can add several PDF between them, but in this case we need to care about normalization. For example:

$$p_X(x) = \frac{1}{6}\sum_{i=1}^{6}\delta(x-i)$$

The exponential distribution is a probability distribution that describes the time between events in a Poisson process.
The probability that a bulb breaks between t and $t + t$ is λt with $\lambda = \frac{1}{\mu}$. Thus, we can obtain:

$$(\text{not broken between } t \text{ and } t + t) = 1 - \lambda t$$

$$\begin{aligned}(\text{not broken between } 0 \text{ and } T, \text{ breaking between } T \text{ and } T + dt) &\propto (1-\lambda t)^{\frac{T}{t}}\lambda t \\ &\simeq e^{-\lambda T}\lambda t\end{aligned}$$

Thus, we obtain the probability distribution with $\mu = \frac{1}{\lambda}$ when modeling such events.

Remark 3 The mean and variance do not always exist. A power-law x^{-k} has a well-defined mean over $x \in [1, \infty)$ only if $k > 2$, and it has a finite variance only if $k > 3$.

2 Some basic properties

2.1 Densities of Transformations of Random Variables Using delta-function

It is very useful to use delta functions to transform and change variables. For instance, if X is distributed as $p_X(x)$, and one wondering the distribution $p_Y(y)$ of the variable $Y = g(X)$, then one can write

$$p_Y(y) = \int dx p_X(x) \delta(y - g(x))$$

Combining this with the classic result for the delta function

$$\delta(f(x)) = \sum_i \frac{\delta(x - x_i)}{|f'(x_i)|} \quad \forall x_i \text{ such that } f(x_i) = 0$$

one finds that for all x_i solutions of $g(x_i) = y$

$$p_Y(y) = \sum_i \int dx p_X(x) \frac{\delta(x - x_i)}{|g'(x_i)|} = \sum_i \frac{p_X(x_i)}{|g'(x_i)|}$$

Proposition 1 (Change of variables) *Let g be a monotonous function, $p_X(x)$ a PDF and $y = f(x)$. Then the probability distribution of y is expressed as:*

$$p_X(x) = p_Y(y) \frac{y}{x} \tag{1}$$

This proof is based on the fact that the probability contained in a differential area must be invariant under change of variables. Or in other words:

$$p_Y(y) y = p_X(x) x$$

Further calculations lead to another formula:

$$p_Y(y) = \frac{x}{y} p_X(x) = \frac{}{y}(x) p_X(x) = \frac{}{y}(g^{-1}(y)) p_X(g^{-1}(y)) = \frac{p_X(g^{-1}(y))}{g'(g^{-1}(y))}$$

A simple example is given by the following transformation: Say for instance that X is distributed uniformly on $]0, 1]$, then the distribution of $Y = -\ln X$ is given by

$$p_Y(y) = |\frac{dx}{dy}| = e^{-y}$$

with Y in $]0, \infty[$.

Remark 4 If f is not monotonic, we need to sum on all the x variable.

$$\sum_{k=1}^{n(y)} \frac{}{y} f_k^{-1}(y) \dot{p}_X(f_k^{-1}(y))$$

with $n(y)$ the number of solutions of $f(x) = y$.

Another classic use of the delta function is to add variables. For instance if $Z = X + Y$ then

$$p_Z(z) = \int dx dy p_X(x) p_Y(y) \delta(z - (x + y))$$

3 Basic bounds

Proposition 2 (Markov inequality) *This inequality gives an upper bound for the probability that a non-negative function of a random variable is greater than or equal to some positive constant. If X is a non-negative random variable and $a > 0$.*

$$(X \geq a) \leq \frac{\mathrm{e}[X]}{a} \tag{2}$$

The proof of this inequality is obtained directly with the definition of cumulative function:

$$\begin{aligned}
(X \geq a) &= \int_a^{+\infty} p_X(x)x \\
&\leq \int_a^{\infty} \frac{x}{a} p_X(x)x \quad \text{because } x \geq a \\
&\leq \frac{\mathrm{e}[X]}{a} \quad \text{because } x p_X(x) \geq 0
\end{aligned}$$

Proposition 3 (Chebyshev's inequality) *This inequality uses the variance to bound the probability that a random variable, with mean and variance, deviates far from the mean:*

$$(X - \mathrm{e}[X] \geq k\sigma) \leq \frac{1}{k^2} \tag{3}$$

The Chebyshev's inequality follows from Markov's inequality by considering the random variable $(X - \mathrm{e}[X])^2$ and $k^2\Delta$:

$$\begin{aligned}
(X - \mathrm{e}[X] \geq k\sigma) &= ((X - \mathrm{e}[X])^2 \geq k^2\Delta) \\
&\leq \frac{\mathrm{e}[(X - \mathrm{e}[X])^2]}{k^2\Delta} = \frac{\Delta}{k^2\Delta} \quad \text{using Markov inequality} \\
&\leq \frac{1}{k^2}
\end{aligned}$$

4 Cumulants, Connected moments, and moment generative functional

Definition 5 (Moment generating function (MGF)) Let X be a random variable. The moment generating function is defined as:

$$f(t) = \mathrm{e}[e^{tX}] = \underbrace{\int_{-\infty}^{\infty} x p_X(x) e^{tx}}_{\text{Laplace transform}} \tag{4}$$

This function is called MGF because of the following property:

Proposition 4

$$\left.\frac{\partial^n f}{\partial t^n}\right|_{t=0} = \int_{-\infty}^{\infty} x\ x^n p_x(x) = \mathrm{e}[X^n]$$

Definition 6 (Characteristic function) The characteristic function of any real-valued random variable completely defines its probability distribution. The function is defined as:

$$\tilde{f}(t) = \mathrm{e}[e^{itX}] = \underbrace{\int_{-\infty}^{\infty} x p_X(x) e^{itx}}_{\text{Fourier transform}} \tag{5}$$

Proposition 5 *As for the MGF, we have the following property:*

$$\left.\frac{\partial^n \tilde{f}}{\partial t^n}\right|_{t=0} = i^n \mathrm{e}[X^n]$$

In statistical physics the main quantity of interest is $Z = \sum e^{-\beta H}$, but we usually work with $\ln Z$, it may be interesting to work with the logarithm of characteristic and MG functions.

Definition 7 (Cumulant) The cumulant K_n of a probability distribution is a set of quantities that provides alternatives to the moments of the distribution. They are defined via the cumulant generating function $K(t)$:

$$K(t) = \ln \mathrm{e}[e^{tX}] = \ln f(t)$$

Indeed, the cumulant is the power expansion of the cumulant generating function:

$$K(t) = \sum_n K_n \frac{t^n}{n!}$$

Or in other words:

$$K_n = \frac{\partial^n K(t)}{\partial t^n}|_{t=0} = \frac{\partial^n \ln f(t)}{\partial t^n}|_{t=0} \tag{6}$$

With this definition we can compute the cumulants of any probability distribution. As examples, we will give the two first:

$$K_1 = \frac{f'}{f}|_{t=0} = \mathrm{e}[X]$$

$$K_2 = \frac{f''(0)f(0) - f'(0)^2}{f(0)^2} = \mathrm{e}[X^2] - \mathrm{e}[X]^2 = \Delta$$

However, then becomes more complicated after this. For instance $K_3 = M_3 - 3M_2M_1 + 2M_1{}^3$ and $K_4 = M_4 - 4M_3M_1 - 3M_2{}^2 + 12M_2M_1{}^2 - 6M_1{}^4$ with $M_n = \mathrm{e}[X^n]$. Working over cumlants instead of moments has an advantage because of the following property:

Proposition 6 *Let X and Y be independent random variables. Then the cumulant of the sum is the sum of the corresponding cumulants, i.e.*

$$K_n(X+Y) = K_n(X) + K_n(Y)$$

Let $Z = X + Y$, with X and Y defined as before. Then, we have:

$$p_Z(z) = \int x \int y p_X(x) p_Y(y) \delta(z - x - y)$$

We now want to evaluate the MGF:

$$\begin{aligned} f_Z(t) &= \int z p_Z(z) e^{tz} \\ &= \int z \int x \int y p_X(x) p_Y(y) e^{t(x+y)} \delta(z - x - y) \\ &= f_X(t) f_Y(t) \quad \text{integrate over z} \end{aligned} \tag{7}$$

If we take the logarithm of 7:

$$\ln f_Z = \ln f_X + \ln f_Y \Leftrightarrow K_n(Z) = K_n(X) + K_n(Y)$$

This is not true for moments.

$$M_n(X+Y) \neq M_n(X) + M_n(Y)$$

4.1 A first introduction to statistics

The theory we just develop for probability can be applied to a lot of different subjects. One of the main results will be the *Central Limit Theorem* and will be presented later on. Now, we are going to make a short presentation of the link between statistics and probability.

This figure resumes in a very short way the link between probabilities and statistics. For example, let's take an unknown $p_X(x)$ generating $\{X_1, ..., X_n\}$. How to estimate $\mathrm{e}[X]$?
The idea is to introduce an estimator

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

This estimator is said to be *unbiased* because of the following property:

$$\mathrm{e}[\hat{m}] = \frac{1}{n}\sum_{i} \mathrm{e}[X_i] = \mathrm{e}[X]$$

The variance of the estimator is: $\mathrm{Var}(\hat{m}) = n \times \frac{\Delta}{n^2} = \frac{\Delta}{n}$. This leads to the following results, adopting physicists' notations:

$$\hat{m} \approx m \pm \sqrt{\frac{\Delta}{n}}$$

The good news is that doing a mistake greater than $\sqrt{\frac{\Delta}{n}}$ is very unlikely: thanks to the Chebyshev inequality, we know that the probability to find a result further than k times σ away from the mean decays AT LEAST quadratically with k ($\propto \frac{1}{k^2}$) (in fact, we will soon see that in many cases, it is decaying even faster). This is an example of what is called PAC estimation: Probably Approximatively Correct (This means that very probably, the result we get is approximatively correct).

The estimator was kind of trivial for the mean of the distribution, but what if we want to estimate the variance of the distribution ?

- As a first idea we can use the estimator:

$$\hat{\Delta} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i} X_i^2 - 2X_i\bar{X}_n + (\bar{X}_n)^2$$

But this estimator is biased. Indeed we have the following decomposition:

$$n\hat{\Delta} = \sum_{i=1}^{n} X_i^2 + n(\bar{X}_n)^2 - 2n(\bar{X}_n)^2 = \sum_{i} X_i^2 - n(\bar{X}_n)^2$$

This gives the following result:

$$\begin{aligned} n\mathrm{e}[\hat{\Delta}] &= n\mathrm{e}[X_1^2] - n\mathrm{e}[\bar{X}_n^2] \\ &= n\Delta + n\mathrm{e}^2[X_1] - n\mathrm{Var}(\bar{X}_n) - n\mathrm{e}^2[\bar{X}_n] \\ &= (n-1)\Delta \quad \text{because } \mathrm{Var}(\bar{X}_n) = \Delta/n \end{aligned}$$

Finally, we obtain:

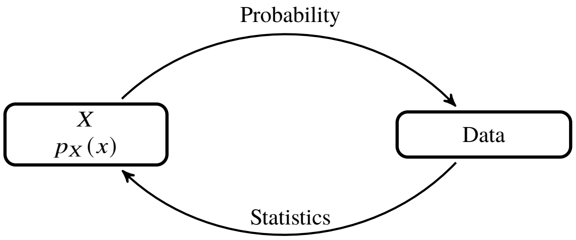
$$\mathrm{e}[\hat{\Delta}] = \frac{n-1}{n}\Delta$$

- But the last estimator is biased[1]. So instead, we use usually:

$$\hat{\Delta}^{\star} = \frac{n}{n-1}\bar{\Delta} = \frac{1}{n-1}\sum_i (X_i - \bar{X}_n)^2$$

  This one is an unbiased estimator.

[1] An unbiased estimator is not necessarily bad, it is better to use a biased estimator with a small variance than an unbiased one with a large variance. There is a *bias-variance tradeoff*: bias is the error coming from wrong assumptions in learning algorithm, variance is the error coming from sensitivity to fluctuations in the data set.


Probability
X
$p_X(x)$
Data
Statistics