Random variables and probability distributions

Random variables (RVs) are functions of the outcomes.

In the case of a throw of a fair die the outcomes are $\omega \in \Omega = \{1, 2, 3, 4, 5, 6\}$. Examples of random variables $X : \Omega \to F$, F a measureable space:

- $X = 1$ if the thrown number is even, $X = 0$ otherwise:

$$X(\omega) = \begin{cases} 1 & \omega \in \{2, 4, 6\} \\ 0 & \omega \in \{1, 3, 5\} \end{cases}$$

- X is a 3d unit vector pointing in the +x, -x, +y, -y, +z, -z directions:

$$X(\omega) = \begin{cases} \vec{e}_x & \omega = 1 \\ -\vec{e}_x & \omega = 2 \\ \vec{e}_y & \omega = 3 \\ \dots \end{cases}$$

- X is the outcome of the throw: $X(\omega) = \omega$

Probability distributions

Discrete random variables

If the random variable X is discrete, the probability of X taking the value x is given by the probability mass function $p_X(x)$:

- $p_X(x) = \Pr(X = x)$
- $0 \leq p_X(x) \leq 1 \; \forall x$
- $\sum_x p_X(x) = 1$

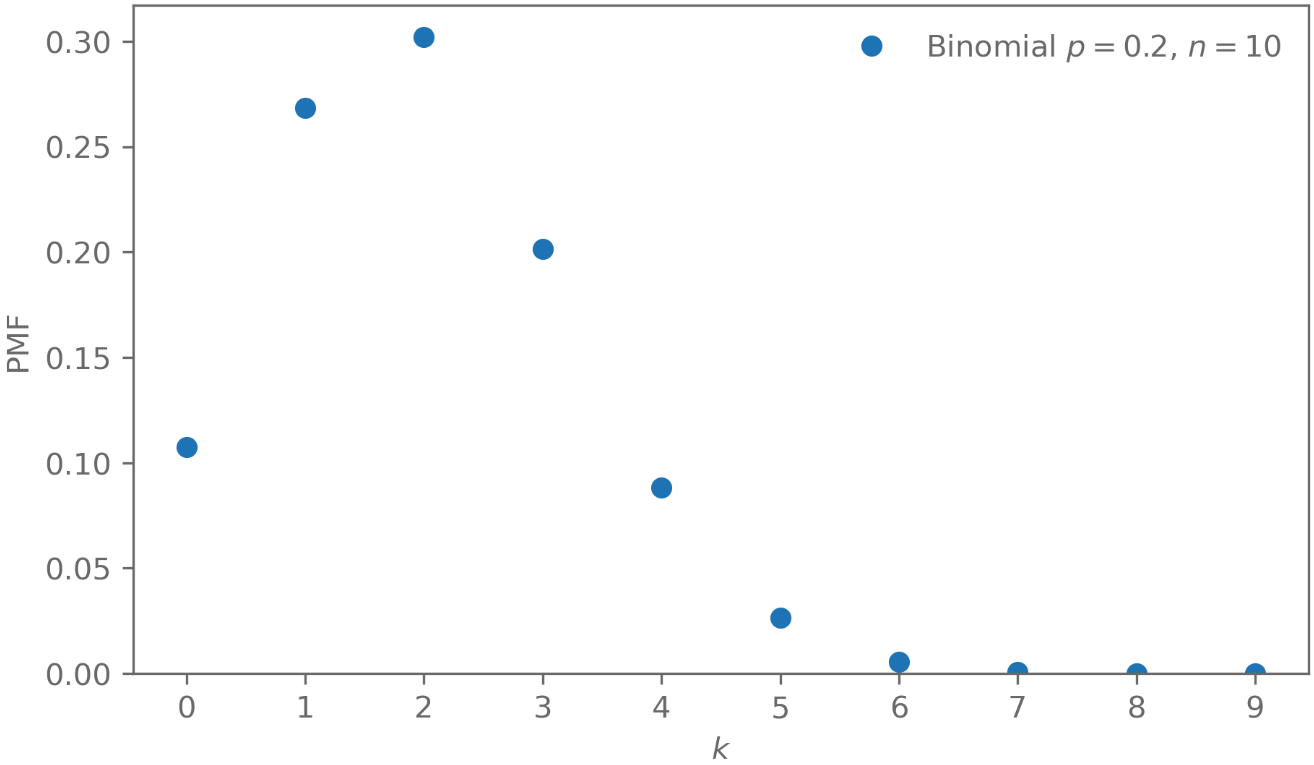
Example: binomial distribution

Consider an experiment that can have two outcomes: success and failure, with probability of success being p and failure being $q = 1 - p$.

The binomial distribution gives the probability of having success k times in n independent trials:

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Parameters:
  - p: probability of success
  - n: number of trials



Continuous random variables

For continuous X, the probability density function $f_X(x)\mathrm{d}x$ is the probability of x falling into the interval $[x, x + \mathrm{d}x]$. More formally:

- $\Pr(a \leq X \leq b) = \int_a^b f_X(x)\mathrm{d}x$
- $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$
- $f_X(x) \geq 0 \; \forall x$

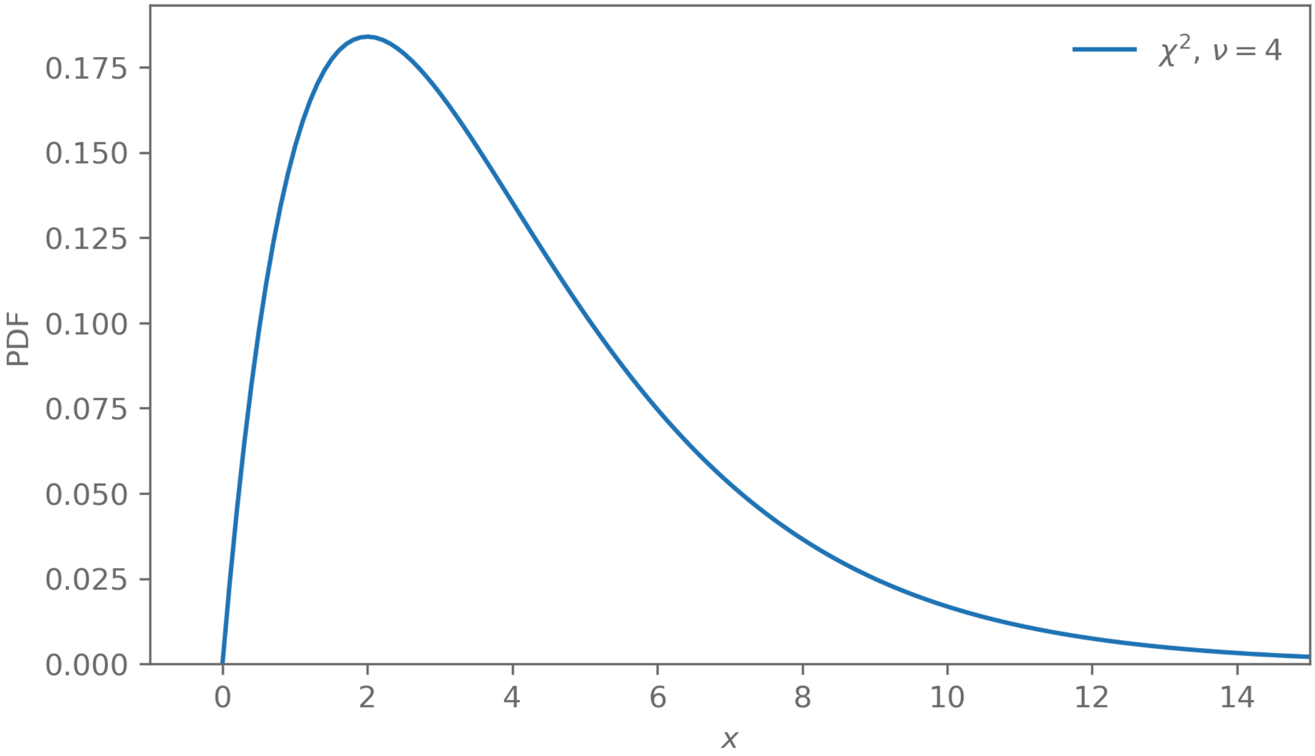
Going forward, we will often abuse notation a bit and use p, such as $p(x)$, to refer to probabilities, probability density functions, probability distributions etc. From the context it is usually clear which random variable they refer to. If not, we will add the subscript for clarity.

Example: chi-squared distribution

The pdf of the chi-squared distribution with ν degrees of freedom is

$$p(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} \mathrm{e}^{-\frac{x}{2}}$$

- Parameter:
  - ν: number of degrees of freedom



Cumulative distribution function (CDF)

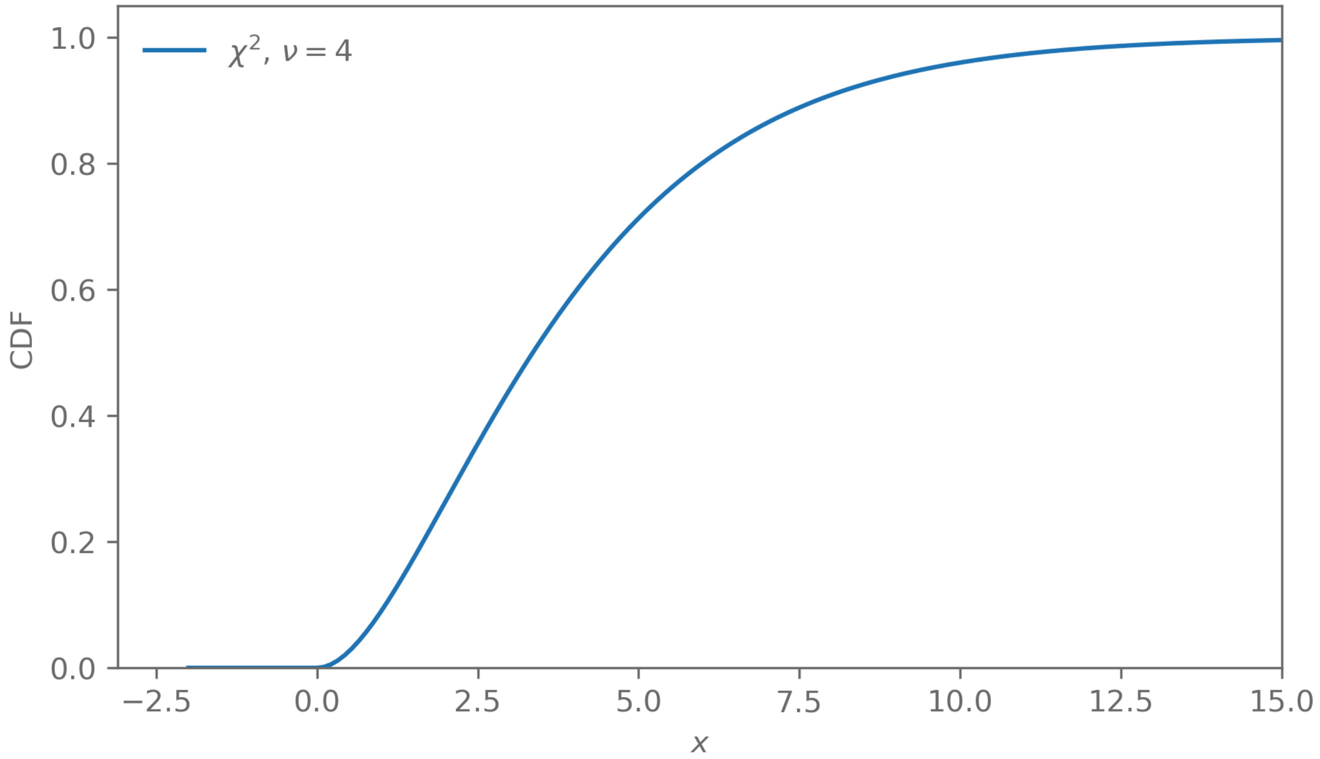
The cumulative distribution function (CDF) $F_X(x)$ is the probability of X being at most x:

$\Pr(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(x)\mathrm{d}x$

The fundamental theorem of calculus gives the relationship between the probability density function and the cumulative distribution function:

- $f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)$
- $\Pr(a \leq X \leq b) = F_X(b) - F_X(a)$

Because $f_X(x) \geq 0$, it follows that $F_X(x)$ is monotonic.



Independent and identically distributed random variables

Often random variables are assumed to be independent and identically distributed (i.i.d.):

- They are mutually independent
- They are all drawn from the same probability distribution

Many theorems assume i.i.d random variables but real data often violate either or both of these assumptions!

Change of variables

Let $g : \mathbb{R} \to \mathbb{R}$ be a monotonic function that maps the random variable X to Y.

What is the PDF $f_Y(y)$, $y = g(x)$ of Y?

The probability of Y in a small interval $[y, y + \mathrm{d}y]$ should be the same as that of X in a small interval $[x, x + \mathrm{d}x]$: $f_X(x)\mathrm{d}x = f_Y(y)\mathrm{d}y$

From this follows that

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right|$$

If g is not monotonic,

$$f_Y(y) = \sum_i f_X(g_i^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g_i^{-1}(y) \right| ,$$

where $g_i^{-1}(y)$ are the solutions to $g(x) = y$ at y.

Inverse transform sampling

An important application of the change of variables formula is inverse transform sampling:

Let $U \sim \mathcal{U}(0, 1)$ and $Y = g(U)$

$$f_U(u)\mathrm{d}u = f_Y(y)\mathrm{d}y \tag{1}$$

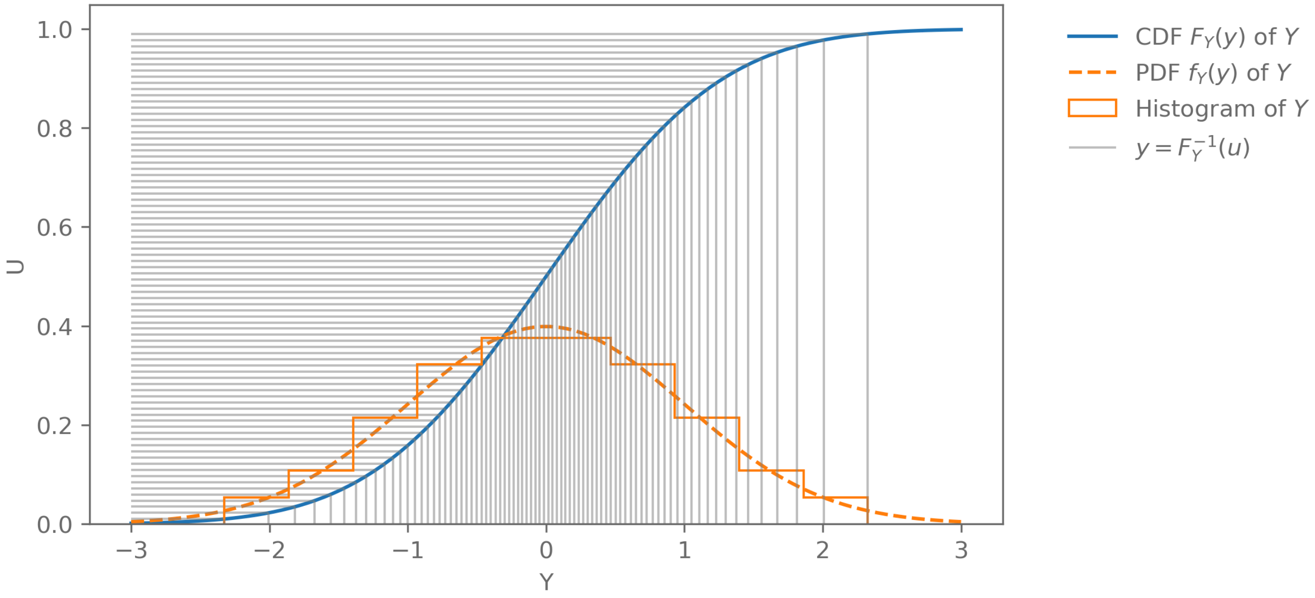
$$\int^u f_U(u')\mathrm{d}u' = \int^{y=g(u)} f_Y(y')\mathrm{d}y' \tag{2}$$

$$F_U(u) = u = F_Y(y) \tag{3}$$

$$y = g(u) = F_Y^{-1}(u) \tag{4}$$

If we know the inverse of the CDF of Y, then we can sample from the distribution of Y.

$$y = F_Y^{-1}(u)$$



Exercise

Sample from the distribution with PDF

$$p(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2}\sin x & \text{otherwise} \\ 0 & x > \pi \end{cases}$$

- Plot the PDF and CDF
- Sample from the distribution using invere transform sampling
- Compare the histogram of the samples to the PDF

Expectation

We usually want to know what values a random variable "typically" takes. There are many ways to define "typically" but a common one is the expectation, or mean, of the random variable X with respect to its probability distribution $p_X(x)$:

$$\mathrm{E}_{p_X(x)}[x] = \int_{-\infty}^{\infty} x\, p_X(x) \mathrm{d}x \tag{5}$$

We usually drop the subscripts if it is clear from context: $\mathrm{E}[x] = \int_{-\infty}^{\infty} x\, p(x) \mathrm{d}x$

More generally, the expectation can be with respect to functions of random variables:

$$\mathrm{E}[f(x)] = \int_{-\infty}^{\infty} f(x)\, p(x) \mathrm{d}x \tag{6}$$

We treat this as a definition but it is a consequence of the definition of the expectation and the change-of-variables formula.

Mean and variance

Some of these functions are used often and have their own names:

Mean $\mathrm{E}[x]$, sometimes also written as $< x >$

- $f(x) = x$
- The value of the mean is often written as μ: $\mathrm{E}[x] = \mu$
- For some constant a: $\mathrm{E}[x + a] = \mu + a$
- For some constant c: $\mathrm{E}[cx] = c\mu$

Variance $\mathrm{Var}[x]$

- $f(x) = (x - \mathrm{E}[x])^2$
- The value of the variance is often written as σ^2: $\mathrm{Var}[x] = \sigma^2$
- $\mathrm{Var}[x] = \mathrm{E}[(x - \mathrm{E}[x])^2] = \mathrm{E}[x^2] - \mathrm{E}[x]^2$
- For some constant a: $\mathrm{Var}[x + a] = \sigma^2$
- For some constant c: $\mathrm{Var}[cx] = c^2\sigma^2$

Moments

More generally, the n-th (raw) moment is given by $\mathrm{E}[X^n]$.

- The first raw moment is the mean

The n-th central moment is given by $\mathrm{E}[(X-\mu)^n]$.

- The second central moment is the variance

The n-th standardised moment is given by $\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^n\right]$.

- The 3rd standardised moment is called the skewness. The skewness is a measure of the asymmetry in a distribution
- The 4th standardised moment is called the kurtosis. The kurtosis measures how heavy the tails of a distribution are

Mode

Often we are also interested in what value a random variable is most likely to take. This is given by the mode, which is location of the maximum of the distribution:

$$x^* = \underset{x}{\operatorname{argmax}}\, p(x)\,.$$

Exercise

Show the expressions for the mean and variance.

Joint distributions

The definitions so far were for univariate distributions but they can be generalised to joint distributions $f_{X_1, \ldots X_n}(x_1, \ldots, x_n)$ of multiple random variables:

$$\Pr(X_1, \ldots, X_n \in D) = \int_D f_{X_1, \ldots X_n}(x_1, \ldots, x_n) \mathrm{d}x_1 \ldots \mathrm{d}x_n \tag{7}$$

And similarly for the CDF:

$$\begin{aligned} F_{X_1, \ldots X_n}(x_1, \ldots, x_n) &= \Pr(X_1 \leq x_1, \ldots, X_n \leq x_n) \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1, \ldots X_n}(x_1', \ldots, x_n') \mathrm{d}x_1' \ldots \mathrm{d}x_n' \end{aligned} \tag{8}$$

with

$$f_{X_1, \ldots X_n}(x_1, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F_{X_1, \ldots X_n}(x_1, \ldots, x_n)$$

Marginal distributions

We often only care about some subset $X_1, \dots X_j$ of random variables out of all $X_1, \dots X_n$, $j < n$. For example, the first j RVs might correspond to physical parameters we want to infer and the others are parameters that we require in our model but do not care about their values, so-called nuisance parameters.

The marginal density function of X, given the joint density of X and Y is

$$f_X(x) = \int f_{X,Y}(x, y) \mathrm{d}y \tag{9}$$

For discrete random variables we recover the formula for law of total probability from earlier:

$$p_X(x) = \sum_y p_{X,Y}(x, y) = \sum_y p_{X,Y}(x|y) p_Y(y)$$

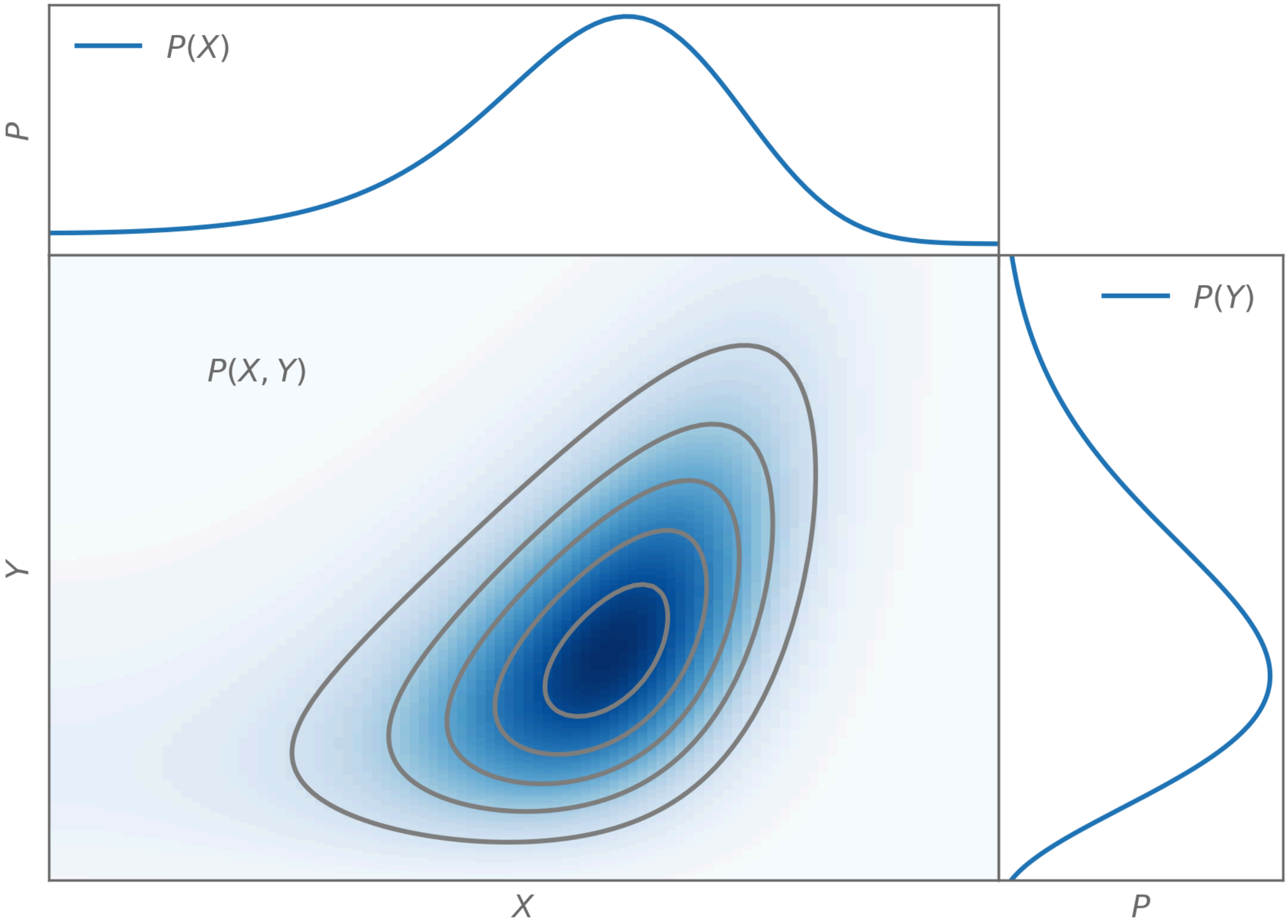
This is in contrast to the conditional distribution

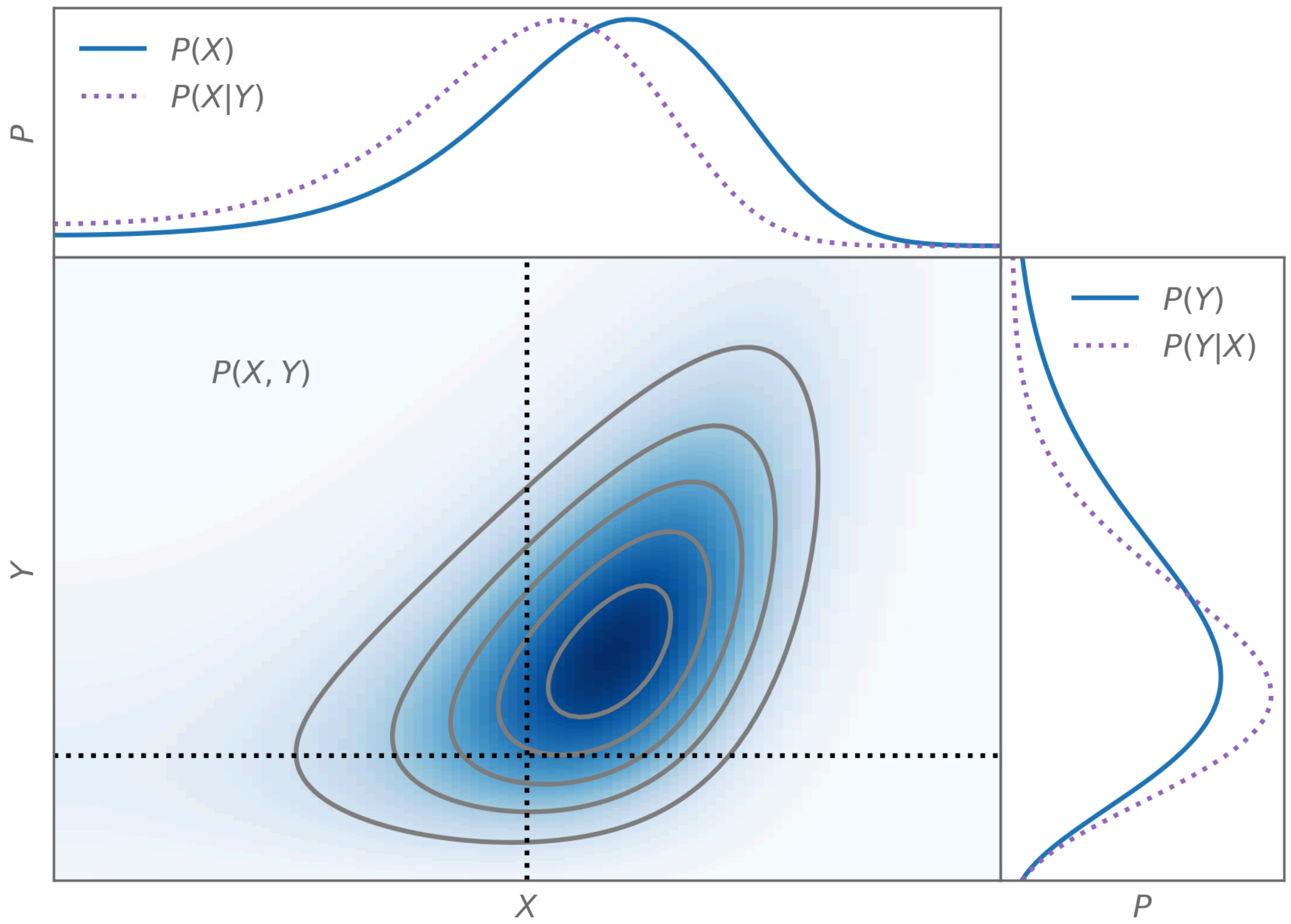
$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

Both of these marginal and conditional distributions are distributions of X.

The marginal distribution is the distribution of X, with the dependence on Y integrated (marginalised) out.

The conditional distribution is the distribution of X, given $Y = y$.


P(X)
P
P(X, Y)
Y
X
P(Y)
P



Change of variables

In the multivariate case, $\vec{y} = g(\vec{x})$ and $g : \mathbb{R}^n \to \mathbb{R}^n$.

The PDF $f_Y(\vec{y})$ is given by

$$f_Y(\vec{y}) = f_X\left(g^{-1}(\vec{y})\right)|J| \tag{10}$$

where $|J|$ is the determinant of the Jacobian of $g^{-1}(\vec{y})$

$$J_{ij} = \frac{\partial g_i^{-1}(\vec{y})}{\partial y_j}$$

The deep learning method of normalising flows is based on this formula.

Common probability distributions

Uniform

$$p(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

- Parameters:
  - a: lower bound
  - b: upper bound
- Symbol: $X \sim \mathcal{U}(a, b)$
- Mean: $\frac{a+b}{2}$
- Variance: $\frac{(b-a)^2}{12}$

Binomial

Consider an experiment that can have two outcomes: success and failure, with probability of success being p and failure being $q = 1 - p$.

The binomial distribution gives the probability of having success k times in n independent trials:

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Parameters:
  - p: probability of success
  - n: number of trials
- Mean: $\mathrm{E}[X] = np$
- Variance: $\mathrm{Var}[X] = np(1-p)$

As $n \to \infty$, with p fixed, the binomial distribution approaches a normal distribution.

As $n \to \infty$, with np fixed, the binomial distribution approaches a Poisson distribution.

Multinomial

Instead of only having two outcomes, consider an experiment that can have k outcomes with probabilities $p_1, \ldots, p_k$, $\sum_i p_i = 1$.

The multinomial distribution gives the probability that in n trials, outcome $i \in \{1, \ldots, k\}$ occurred x_i times. Each trial only has one outcome, so $\sum_i x_i = n$.

$$\Pr(X = x_1, \ldots, x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}$$

- Parameters:
  - p_i: probabilities of outcomes $i \in \{1, \ldots, k\}$
  - n: number of trials
- Mean: $\mathrm{E}[X_i] = np_i$

- Covariance: $\mathrm{Cov}[X_i, X_j] = \delta_{ij} n p_i - n p_i p_j$

Poisson

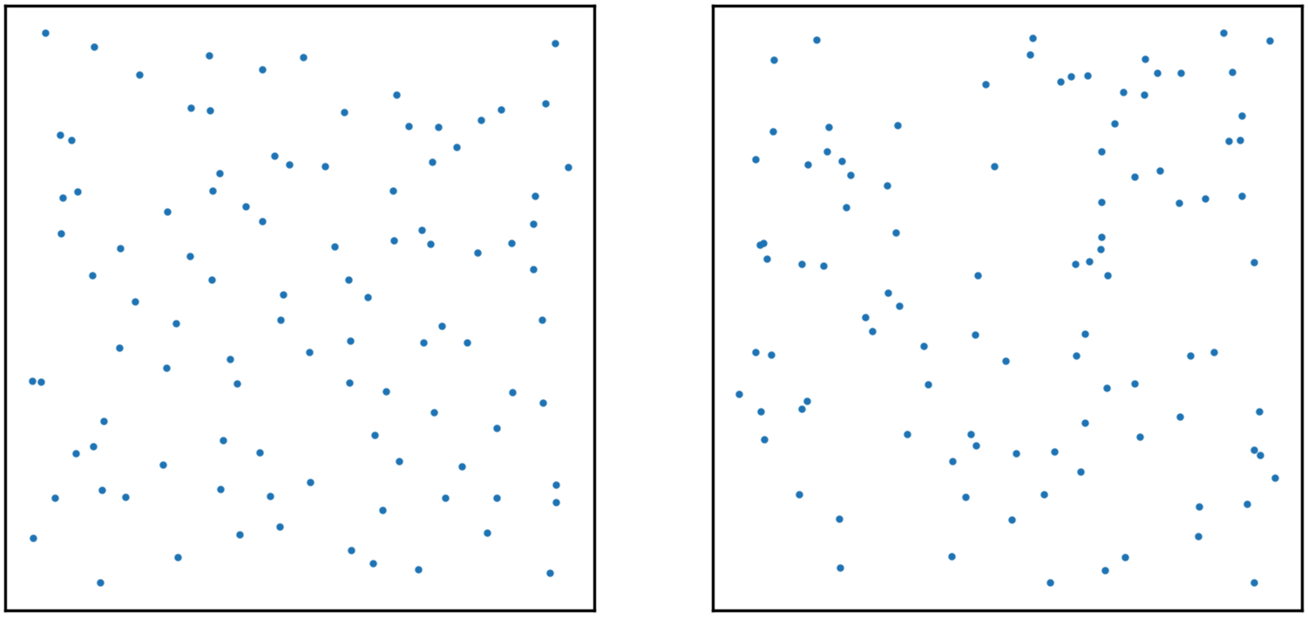
The Poisson distribution gives the probability of k (independent) events happening in a time interval, with expected rate λ.

$$\Pr(X = k) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}$$

- Parameter:
  - λ: event rate
- Mean: λ
- Variance: λ

The Poisson distribution appears commonly when events are rare and independent. Examples are decay of nuclei, arrival of photons in telescopes. Note that the interval for which the rate is defined does not need to be a time interval, it can also be spatial.

As $\lambda \to \infty$, the Poisson distribution approaches a normal distribution.



Exercises

- Confirm by simulation that the probability of k out of n iid $X_i \sim \mathcal{U}(0,1)$ being in a small interval Δx follows a Poisson distribution.
- Derive the Poisson distribution from the binomial distribution.

Normal

Also called Gaussian distribution. If there is only one distribution you need to remember, it is this one. It appears *everywhere*, and has many fascinating properties.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Parameters:
  - μ: mean
  - σ^2: variance
- Symbol: $X \sim \mathcal{N}(\mu, \sigma^2)$
- Mean: μ
- Variance: σ^2

Some properties:

- Sums of Gaussian RVs are Gaussian
- Jointly Gaussian RVs that are uncorrelated are independent. This is not true in general!
- Sample mean and sample variance are independent
- For a given mean and variance, the normal distribution is the distribution with the highest entropy

Multivariate normal

Multivariate generalisation of the normal distribution. Let $\vec{x} \in \mathbb{R}^n$, then

$$p(\vec{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \det \Sigma} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})\right)$$

- Parameters:
  - $\vec{\mu}$: mean
  - Σ: covariance
- Symbol: $\vec{X} \sim \mathcal{N}(\vec{\mu}, \Sigma)$
- Mean: $\vec{\mu}$
- Covariance: Σ, positive definite symmetric matrix

Chi-squared

The square of a normally distributed RV $X \sim \mathcal{N}(0, 1)$ is chi-squared distributed with 1 degree of freedom.

The sum of the squares of n iid $X_i \sim \mathcal{N}(0, 1)$ is chi-squared distributed with n degrees of freedom: $\sum_i^n X_i^2 \sim \chi_n^2$

The pdf of the chi-squared distribution with ν degrees of freedom is

$$p(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} \mathrm{e}^{-\frac{x}{2}}$$

- Parameter:
  - ν: number of degrees of freedom
- Symbol: χ^2_ν
- Mean: ν
- Variance: 2ν

Approaches a normal distribution for large ν.

An important application is that for a n-dimensional multivariate RV $\vec{X} \sim \mathcal{N}(\vec{\mu}, \Sigma)$, we have

$$(\vec{X} - \vec{\mu})^T \Sigma^{-1} (\vec{X} - \vec{\mu}) \sim \chi^2_n$$

We will come back to this when we look at the goodness of fit in Gaussian likelihoods.

Cauchy

The Cauchy distribution comes up occationally and sets itself apart by how pathological it is. The PDF looks innocuous but it has no defined mean or variance.

$$p(x) = \frac{1}{\pi\gamma \left(1 + \frac{(x-x_0)^2}{\gamma}\right)}$$

- Parameters:
  - x_0: location
  - γ: scale
- Mean: undefined!
- Variance: undefined!

Power law

Scale invariance is a property that is common in the physical sciences. Consider the function $f(x) = ax^{-k}$. The function does not change it shape when scaling the argument, it is scale invariant:

$$f(cx) = c^{-k} f(x) \propto f(x)\ .$$

Depending on the exponent, such function cannot be a probability distribution. In practice, there are physical lower and/or upper limits that make the distribution well-defined. The PDF is

$$f(x) = \begin{cases} \frac{1-\alpha}{b^{1-\alpha}-a^{1-\alpha}} x^{-\alpha} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}\ ,$$

with CDF

$$F(x) = \begin{cases} 0 & x < a \\ \frac{1}{b^{1-\alpha} - a^{1-\alpha}} \left(x^{1-\alpha} - a^{1-\alpha} \right) & a \leq x \leq b \\ 1 & x > b \end{cases} .$$

- Parameters:
  - a, b: lower and upper limits
  - α: power low exponent

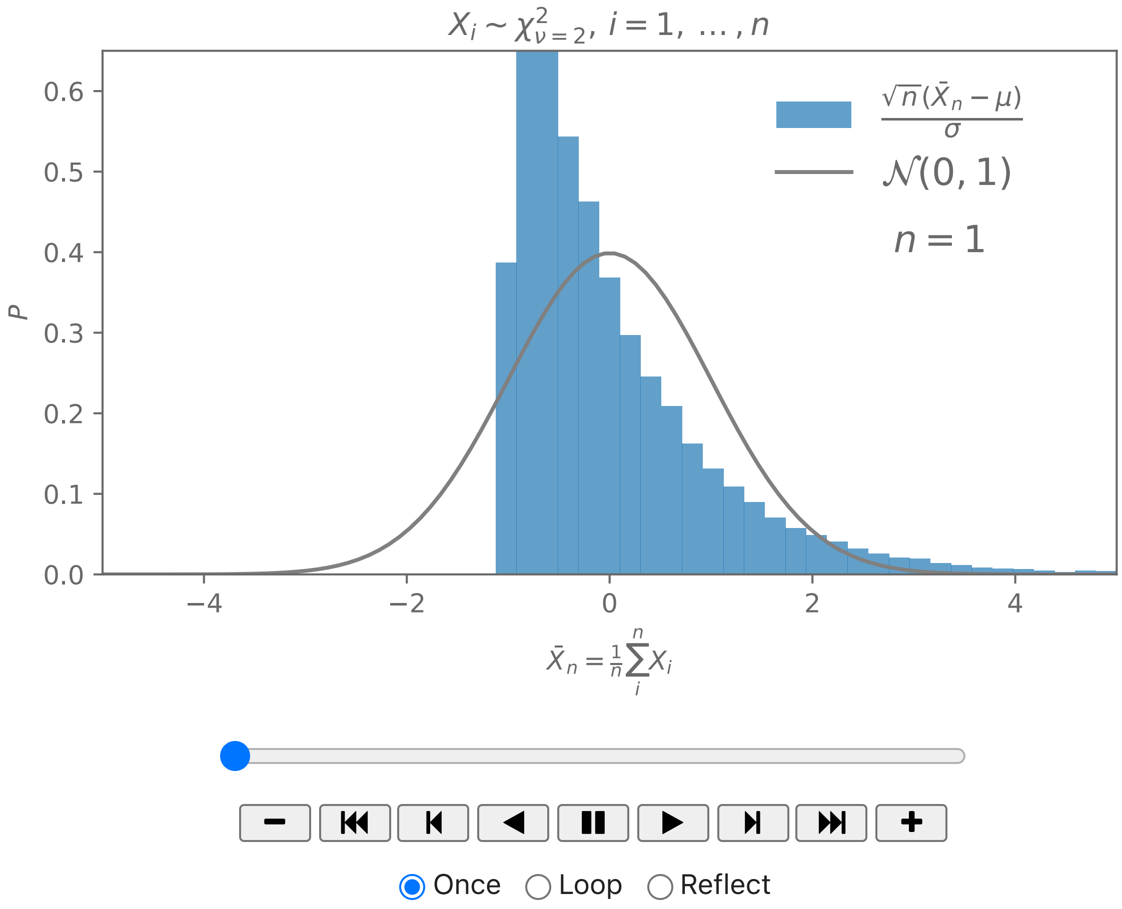
The mean and variance are formally defined for appropriate lower/upper limits and exponents. They are usually poor descriptors for the distribution, however. For steep power laws, these distributions can behave quite unintuitively.

Central limit theorem

One reason the normal distribution is so ubiquitous is the central limit theorem:

The mean $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$ of n iid RVs with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$ tends to a normal distribution.

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0,1) \quad (n \to \infty)$$



- The CLT can be generalised for non-iid random variables in certain cases but care needs to be taken when trying to apply it to non-iid data.

- The "limit" part in the CLT is important - for finite n (which is always the case for real data) the distribution of the mean might not have sufficiently converged.

Exercises

- Make plots of the PDFs of the distributions we covered so far, varying their parameters
- Compute the distribution of the sum of two independent Gaussian RVs.
- What is the distribution of the sum of two independent RVs $X \sim f$ and $Y \sim g$?
- Derive the PDF of the $\chi^2_{\nu=1}$ distribution
- Bonus:
  - What is the product of two independent RVs U and V?
  - Compute the distribution of the ratio of two Gaussian RVs.