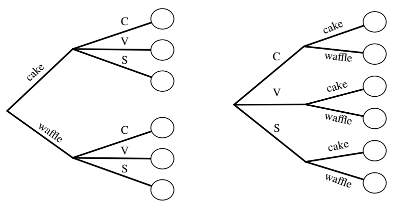
Probability Cheatsheet v2.0

Original work at **http://github.com/wzchen/probability_cheatsheet**, modified by Pablo Angulo for ESTIN@UPM.ES

Last Updated August 3, 2020

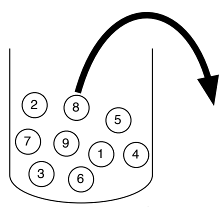
Counting

Multiplication Rule



Let's say we have a compound experiment (an experiment with multiple components). If the 1st component has n_1 possible outcomes, the 2nd component has n_2 possible outcomes, $\dots$, and the rth component has n_r possible outcomes, then overall there are $n_1 n_2 \dots n_r$ possibilities for the whole experiment.

Sampling Table



The sampling table gives the number of possible samples of size k out of a population of size n, under various assumptions about how the sample is collected.

	Order Matters	**Not Matter**
With Replacement	n^k	$\binom{n+k-1}{k}$
Without Replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Cardano's Definition of Probability

If the number of outcomes is finite and all outcomes are equally likely, the probability of an event A happening is:

$$P_{\text{Cardano}}(A) = \frac{\text{number of outcomes favorable to } A}{\text{number of outcomes}}$$

Set algebra

Unions, Intersections, and Complements

Complements - The following are true.

$$\mathbf{A} \cup \mathbf{A}^c = \Omega$$

$$\mathbf{A} \cap \mathbf{A}^c = \emptyset$$

De Morgan's Laws

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Probability

Axioms of probability

Any assignment from subsets of E to real numbers is a *probability measure* if the following holds:

Probabilities are positive $P(A) \geq 0$.

The probability of the whole space is 1 $P(E) = 1$.

Probabilities of a union of disjoint sets $P(A \cup B) = P(A) + P(B)$, *provided* $A \cap B = \emptyset$.

Consequences

For any probability measure, the following are true:

Probability of the empty set $P(\emptyset) = 0$.

Probability of the complement $P(A^C) = 1 - P(A)$.

Conditional probability

Conditional Probability

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

Probability of A, given that B occurred.

Conditional Probability *is* Probability $P(A|B)$ is a probability function for any fixed B. Any theorem that holds for probability also holds for conditional probability.

Probability of an Intersection or Union

Intersections via Conditioning

$$P(A, B) = P(A)P(B|A)$$
$$P(A, B, C) = P(A)P(B|A)P(C|A, B)$$

Unions via Inclusion-Exclusion

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\quad + P(A \cap B \cap C). \end{aligned}$$

Law of Total Probability

Assume the n events A_i are pairwise disjoint ($A_i \cap A_j = \emptyset$ for any $i \neq j$) and their union is the whole sample space, and let B be any event. Then:

$$\begin{aligned} P(B) &= P(B|A_1)P(A_1) + \dots + P(B|A_n)P(A_n) \\ &= \sum_{i=1}^{n} P(B|A_i)P(A_i) \end{aligned}$$

Bayes' Rule

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

Independence

2 Independent Events A and B are independent if knowing whether A occurred gives no information about whether B occurred. More formally, A and B (which have nonzero probability) are independent if and only if one of the following equivalent statements holds:

$$P(A \cap B) = P(A)P(B), \quad P(A|B) = P(A), \quad P(B|A) = P(B)$$

3 Independent Events A, B and C are independent if information about two of them gives no information about whether the third one occurred. In other words, $P(A|E_B \cap E_C) = P(A)$, where E_B is either B, B^C, or E, and E_C is either C, C^C or E. The relations obtained by permuting A, B and C must also hold.

Conditional Independence A and B are conditionally independent given C if $P(A \cap B|C) = P(A|C)P(B|C)$. Conditional independence does not imply independence, and independence does not imply conditional independence.

Random Variables

A **Random Variable (RV)** is a function form a probability space into the real numbers:

$$X : \Omega \to \mathbb{R}$$

The **support** of X is the smallest closed set S such that $P(X \in S) = 1$ (morally, the set of values that X can take).
The **distribution** of the RV is not the probability distribution of Ω, but the induced distribution in the real numbers $A \to P(X \in A)$. E.g. Ω has a continuous uniform distribution in the interval $[0, 1]$, $X(w)$ is 1 if $w > 1/2$ and 0 otherwise. Then X has a Bernoulli distribution with $p = 1/2$.

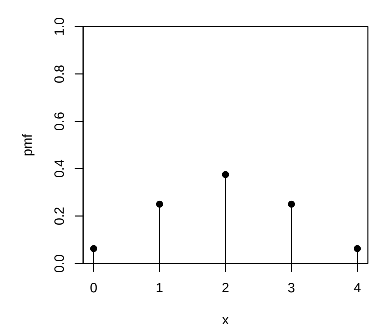
Discrete Random Variables

A RV is *discrete* if its support is finite, or infinite countable (the integer, the positive integers, etc). Its support is $\{x \in \Omega : \; P(X = x) > 0\}$.

PMF, CDF, and Independence

Probability Mass Function (PMF) Gives the probability that a *discrete* random variable takes on the value x.

$$p_X(x) = P(X = x)$$

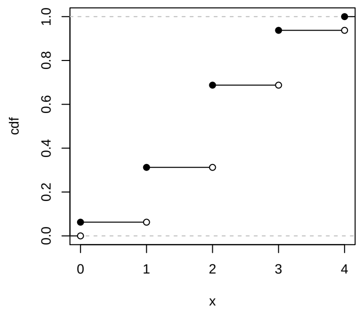


The PMF satisfies

$$p_X(x) \geq 0 \text{ and } \sum_x p_X(x) = 1$$

Cumulative Distribution Function (CDF) Gives the probability that a random variable is less than or equal to x.

$$F_X(x) = P(X \le x)$$



The CDF is an increasing, right-continuous function with

$$F_X(x) \to 0 \text{ as } x \to -\infty \text{ and } F_X(x) \to 1 \text{ as } x \to \infty$$

Independence Intuitively, two random variables are independent if knowing the value of one gives no information about the other. Discrete r.v.s X and Y are independent if for *all* values of x and y

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Continuous Random Variables (CRVs)

Probability density function (PDF)

What's the probability that a CRV is in an interval? Take the difference in CDF values (or use the PDF as described later).

$$P(a \le X \le b) = P(X \le b) - P(X \le a) = F_X(b) - F_X(a)$$

For $X \sim \mathcal{N}(\mu, \sigma^2)$, this becomes

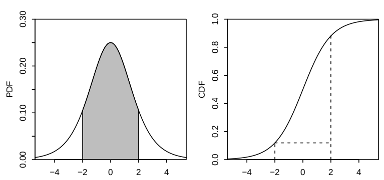
$$P(a \le X \le b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

What is the Probability Density Function (PDF)? The PDF f is the derivative of the CDF F.

$$F'(x) = f(x)$$

A PDF is nonnegative and integrates to 1. By the fundamental theorem of calculus, to get from PDF back to CDF we can integrate:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$



To find the probability that a CRV takes on a value in an interval, integrate the PDF over that interval.

$$F(b) - F(a) = \int_a^b f(x)dx$$

Expected Value and Indicators

Expected Value and Linearity

Expected Value (a.k.a. *mean*, *expectation*, or *average*) is a weighted average of the possible outcomes of our random variable. Mathematically, if $x_1, x_2, x_3, \ldots$ are all of the distinct possible values that a discrete random variable X can take, the expected value of X is

$$E(X) = \sum_i x_i P(X = x_i)$$

Expected value of a CRV Analogous to the discrete case, where you sum x times the PMF, for CRVs you integrate x times the PDF.

$$E(X) = \int_{-\infty}^{\infty} x f(x)dx$$

X	Y	$X + Y$
3	4	7
2	2	4
6	8	14
10	23	33
1	–3	–2
1	0	1
5	9	14
4	1	5
...	...	...

$$\frac{1}{n}\sum_{i=1}^{n} x_i \quad + \quad \frac{1}{n}\sum_{i=1}^{n} y_i \quad = \quad \frac{1}{n}\sum_{i=1}^{n} (x_i + y_i)$$

$$E(X) \quad + \quad E(Y) \quad = \quad E(X+Y)$$

Linearity For any r.v.s X and Y, and constants a, b, c,

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$

Same distribution implies same mean If X and Y have the same distribution, then $E(X) = E(Y)$ and, more generally,

$$E(g(X)) = E(g(Y))$$

Conditional Expected Value is defined like expectation, only conditioned on any event A.

$$E(X|A) = \sum_x x P(X = x|A)$$

Indicator Random Variables

Indicator Random Variable is a random variable that takes on the value 1 or 0. It is always an indicator of some event: if the event occurs, the indicator is 1; otherwise it is 0. They are useful for many problems about counting how many events of some kind occur. Write

$$I_A = \begin{cases} 1 & \text{if } A \text{ occurs,} \\ 0 & \text{if } A \text{ does not occur.} \end{cases}$$

Note that $I_A^2 = I_A, I_A I_B = I_{A \cap B}$, and $I_{A \cup B} = I_A + I_B - I_A I_B$.

Distribution $I_A \sim \text{Bern}(p)$ where $p = P(A)$.

Fundamental Bridge The expectation of the indicator for event A is the probability of event A: $E(I_A) = P(A)$.

Variance and Standard Deviation

$$\text{Var}(X) = E\left(X - E(X)\right)^2 = E(X^2) - (E(X))^2$$

$$\text{SD}(X) = \sqrt{\text{Var}(X)}$$

LOTUS, UoU

LOTUS

Expected value of a function of an r.v. The expected value of X is defined this way:

$$E(X) = \sum_x x P(X = x) \text{ (for discrete } X\text{)}$$

$$E(X) = \int_{-\infty}^{\infty} x f(x)dx \text{ (for continuous } X\text{)}$$

The **Law of the Unconscious Statistician (LOTUS)** states that you can find the expected value of a *function of a random variable*, $g(X)$, in a similar way, by replacing the x in front of the PMF/PDF by $g(x)$ but still working with the PMF/PDF of X:

$$E(g(X)) = \sum_x g(x)P(X = x) \text{ (for discrete } X\text{)}$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx \text{ (for continuous } X\text{)}$$

What's a function of a random variable? A function of a random variable is also a random variable. For example, if X is the number of bikes you see in an hour, then $g(X) = 2X$ is the number of bike wheels you see in that hour and $h(X) = \binom{X}{2} = \frac{X(X-1)}{2}$ is the number of *pairs* of bikes such that you see both of those bikes in that hour.

What's the point? You don't need to know the PMF/PDF of $g(X)$ to find its expected value. All you need is the PMF/PDF of X.

Universality of Uniform (UoU)

When you plug any CRV into its own CDF, you get a Uniform(0,1) random variable. When you plug a Uniform(0,1) r.v. into an inverse CDF, you get an r.v. with that CDF. For example, let's say that a random variable X has CDF

$$F(x) = 1 - e^{-x}, \text{ for } x > 0$$

By UoU, if we plug X into this function then we get a uniformly distributed random variable.

$$F(X) = 1 - e^{-X} \sim \text{Unif}(0, 1)$$

Similarly, if $U \sim \text{Unif}(0, 1)$ then $F^{-1}(U)$ has CDF F. The key point is that for any continuous random variable X, we can transform it into a Uniform random variable and back by using its CDF.

Moments

Moments describe the shape of a distribution. Let X have mean μ and standard deviation σ, and $Z = (X - \mu)/\sigma$ be the *standardized* version of X. The kth moment of X is $\mu_k = E(X^k)$ and the kth standardized moment of X is $m_k = E(Z^k)$. The mean, variance, skewness, and kurtosis are important summaries of the shape of a distribution.

Mean $E(X) = \mu_1$

Variance $\text{Var}(X) = \mu_2 - \mu_1^2$

Skewness $\text{Skew}(X) = m_3$

Kurtosis $\text{Kurt}(X) = m_4 - 3$

Joint PDFs and CDFs

Joint Distributions

The **joint CDF** of X and Y is

$$F(x, y) = P(X \le x, Y \le y)$$

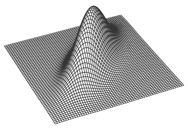
In the discrete case, X and Y have a **joint PMF**

$$p_{X,Y}(x, y) = P(X = x, Y = y).$$

In the continuous case, they have a **joint PDF**

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

The joint PMF/PDF must be nonnegative and sum/integrate to 1.



Conditional Distributions

Conditioning and Bayes' rule for discrete r.v.s

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

Conditioning and Bayes' rule for continuous r.v.s

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

Hybrid Bayes' rule

$$f_X(x|A) = \frac{P(A|X = x)f_X(x)}{P(A)}$$

Marginal Distributions

To find the distribution of one (or more) random variables from a joint PMF/PDF, sum/integrate over the unwanted random variables.

Marginal PMF from joint PMF

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

Marginal PDF from joint PDF

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$

Independence of Random Variables

Random variables X and Y are independent if and only if any of the following conditions holds:

- Joint CDF is the product of the marginal CDFs
- Joint PMF/PDF is the product of the marginal PMFs/PDFs
- Conditional distribution of Y given X is the marginal distribution of Y

Write $X \perp\!\!\!\perp Y$ to denote that X and Y are independent.

Multivariate LOTUS

LOTUS in more than one dimension is analogous to the 1D LOTUS. For discrete random variables:

$$E(g(X, Y)) = \sum_x \sum_y g(x, y)P(X = x, Y = y)$$

For continuous random variables:

$$E(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f_{X,Y}(x, y)dxdy$$

Covariance and Transformations

Covariance and Correlation

Covariance is the analog of variance for two random variables.

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$$

Note that

$$\text{Cov}(X, X) = E(X^2) - (E(X))^2 = \text{Var}(X)$$

Correlation is a standardized version of covariance that is always between -1 and 1.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Covariance and Independence If two random variables are independent, then they are uncorrelated. The converse is not necessarily true (e.g., consider $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$).

$$X \perp\!\!\!\perp Y \longrightarrow \text{Cov}(X, Y) = 0 \longrightarrow E(XY) = E(X)E(Y)$$

Covariance and Variance The variance of a sum can be found by

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j)$$

If X and Y are independent then they have covariance 0, so

$$X \perp\!\!\!\perp Y \Longrightarrow \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

Covariance Properties For random variables W, X, Y, Z and constants a, b:

$$\begin{aligned} \text{Cov}(X, Y) &= \text{Cov}(Y, X) \\ \text{Cov}(X + a, Y + b) &= \text{Cov}(X, Y) \\ \text{Cov}(aX, bY) &= ab\text{Cov}(X, Y) \\ \text{Cov}(W + X, Y + Z) &= \text{Cov}(W, Y) + \text{Cov}(W, Z) + \text{Cov}(X, Y) \\ &\quad + \text{Cov}(X, Z) \end{aligned}$$

Correlation is location-invariant and scale-invariant For any constants a, b, c, d with a and c nonzero,

$$\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$$

If correlation is 1 If $\text{Corr}(X, Y) = 1$, there are constants $a > 0, b$ such that $X = aY + b$

If correlation is -1 If $\text{Corr}(X, Y) = -1$, there are constants $a < 0, b$ such that $X = aY + b$

Transformations

One Variable Transformations Let's say that we have a random variable X with PDF $f_X(x)$, but we are also interested in some function of X. We call this function $Y = g(X)$. Also let $y = g(x)$. If g is differentiable and strictly increasing (or strictly decreasing), then the PDF of Y is

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

The derivative of the inverse transformation is called the **Jacobian**.

Two Variable Transformations Similarly, let's say we know the joint PDF of U and V but are also interested in the random vector (X, Y) defined by $(X, Y) = g(U, V)$. Let

$$\frac{\partial(u, v)}{\partial(x, y)} = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix}$$

be the **Jacobian matrix**. If the entries in this matrix exist and are continuous, and the determinant of the matrix is never 0, then

$$f_{X,Y}(x, y) = f_{U,V}(u, v)\left|\left|\frac{\partial(u, v)}{\partial(x, y)}\right|\right|$$

The inner bars tells us to take the matrix's determinant, and the outer bars tell us to take the absolute value. In a 2×2 matrix,

$$\left|\left|\begin{matrix} a & b \\ c & d \end{matrix}\right|\right| = |ad - bc|$$

Convolutions

Convolution Integral If you want to find the PDF of the sum of two independent CRVs X and Y, you can do the following integral:

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x)f_Y(t - x)dx$$

Example Let $X, Y \sim \mathcal{N}(0, 1)$ be i.i.d. Then for each fixed t,

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\frac{1}{\sqrt{2\pi}}e^{-(t-x)^2/2}dx$$

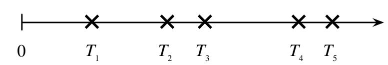
By completing the square and using the fact that a Normal PDF integrates to 1, this works out to $f_{X+Y}(t)$ being the $\mathcal{N}(0, 2)$ PDF.

Poisson Process

Definition We have a **Poisson process** of rate λ arrivals per unit time if the following conditions hold:

1. The number of arrivals in a time interval of length t is $\text{Pois}(\lambda t)$.
2. Numbers of arrivals in disjoint time intervals are independent.

For example, the numbers of arrivals in the time intervals $[0, 5]$, $(5, 12)$, and $[13, 23)$ are independent with $\text{Pois}(5\lambda), \text{Pois}(7\lambda), \text{Pois}(10\lambda)$ distributions, respectively.



Count-Time Duality Consider a Poisson process of emails arriving in an inbox at rate λ emails per hour. Let T_n be the time of arrival of the nth email (relative to some starting time 0) and N_t be the number of emails that arrive in $[0, t]$. Let's find the distribution of T_1. The event $T_1 > t$, the event that you have to wait more than t hours to get the first email, is the same as the event $N_t = 0$, which is the event that there are no emails in the first t hours. So

$$P(T_1 > t) = P(N_t = 0) = e^{-\lambda t} \longrightarrow P(T_1 \le t) = 1 - e^{-\lambda t}$$

Thus we have $T_1 \sim \text{Expo}(\lambda)$. By the memoryless property and similar reasoning, the interarrival times between emails are i.i.d. $\text{Expo}(\lambda)$, i.e., the differences $T_n - T_{n-1}$ are i.i.d. $\text{Expo}(\lambda)$.

Discrete Distributions

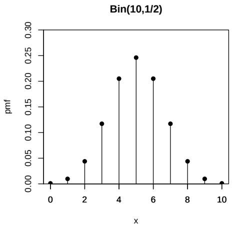
Bernoulli Distribution

The Bernoulli distribution is the simplest case of the Binomial distribution, where we only have one trial ($n = 1$). Let us say that X is distributed $\text{Bern}(p)$. We know the following:

Story A trial is performed with probability p of "success", and X is the indicator of success: 1 means success, 0 means failure.

Example Let X be the indicator of Heads for a fair coin toss. Then $X \sim \text{Bern}(\frac{1}{2})$. Also, $1 - X \sim \text{Bern}(\frac{1}{2})$ is the indicator of Tails.

Binomial Distribution



Let us say that X is distributed $\text{Bin}(n, p)$. We know the following:

Story X is the number of "successes" that we will achieve in n independent trials, where each trial is either a success or a failure, each with the same probability p of success. We can also write X as a sum of multiple independent $\text{Bern}(p)$ random variables. Let $X \sim \text{Bin}(n, p)$ and $X_j \sim \text{Bern}(p)$, where all of the Bernoullis are independent. Then

$$X = X_1 + X_2 + X_3 + \cdots + X_n$$

Example If Jeremy Lin makes 10 free throws and each one independently has a $\frac{3}{4}$ chance of getting in, then the number of free throws he makes is distributed $\text{Bin}(10, \frac{3}{4})$.

Properties Let $X \sim \text{Bin}(n, p), Y \sim \text{Bin}(m, p)$ with $X \perp\!\!\!\perp Y$.

- **Redefine success** $n - X \sim \text{Bin}(n, 1-p)$
- **Sum** $X + Y \sim \text{Bin}(n + m, p)$
- **Conditional** $X|(X + Y = r) \sim \text{HGeom}(n, m, r)$
- **Binomial-Poisson Relationship** $\text{Bin}(n, p)$ is approximately $\text{Pois}(\lambda)$ if p is small.
- **Binomial-Normal Relationship** $\text{Bin}(n, p)$ is approximately $\mathcal{N}(np, np(1-p))$ if n is large and p is not near 0 or 1.

Geometric Distribution

Let us say that X is distributed $\text{Geom}(p)$. We know the following:

Story X is the number of "trials" that we will repeat before we observe our first success. Our successes have probability p.

Example If each pokeball we throw has probability $\frac{1}{10}$ to catch Mew, the number of pokeballs thrown will be distributed $\text{Geom}(\frac{1}{10})$.

Poisson Distribution

Let us say that X is distributed $\text{Pois}(\lambda)$. We know the following:

Story There are rare events (low probability events) that occur many different ways (high possibilities of occurences) at an average rate of λ occurrences per unit space or time. The number of events that occur in that unit of space or time is X.

Example A certain busy intersection has an average of 2 accidents per month. Since an accident is a low probability event that can happen many different ways, it is reasonable to model the number of accidents in a month at that intersection as $\text{Pois}(2)$. Then the number of accidents that happen in two months at that intersection is distributed $\text{Pois}(4)$.

Properties Let $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$, with $X \perp\!\!\!\perp Y$.

1. **Sum** $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
2. **Conditional** $X|(X + Y = n) \sim \text{Bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$
3. **Chicken-egg** If there are $Z \sim \text{Pois}(\lambda)$ items and we randomly and independently "accept" each item with probability p, then the number of accepted items $Z_1 \sim \text{Pois}(\lambda p)$, and the number of rejected items $Z_2 \sim \text{Pois}(\lambda(1-p))$, and $Z_1 \perp\!\!\!\perp Z_2$.

Continuous Distributions

Uniform Distribution

Let us say that U is distributed $\text{Unif}(a, b)$. We know the following:

Properties of the Uniform For a Uniform distribution, the probability of a draw from any interval within the support is proportional to the length of the interval. See *Universality of Uniform* and *Order Statistics* for other properties.

Example William throws darts really badly, so his darts are uniform over the whole room because they're equally likely to appear anywhere. William's darts have a Uniform distribution on the surface of the room. The Uniform is the only distribution where the probability of hitting in any specific region is proportional to the length/area/volume of that region, and where the density of occurrence in any one specific spot is constant throughout the whole support.

Normal Distribution

Let us say that X is distributed $\mathcal{N}(\mu, \sigma^2)$. We know the following:

Central Limit Theorem The Normal distribution is ubiquitous because of the Central Limit Theorem, which states that the sample mean of i.i.d. r.v.s will approach a Normal distribution as the sample size grows, regardless of the initial distribution.

Location-Scale Transformation Every time we shift a Normal r.v. (by adding a constant) or rescale a Normal (by multiplying by a constant), we change it to another Normal r.v. For any Normal $X \sim \mathcal{N}(\mu, \sigma^2)$, we can transform it to the standard $\mathcal{N}(0, 1)$ by the following transformation:

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Standard Normal The Standard Normal, $Z \sim \mathcal{N}(0, 1)$, has mean 0 and variance 1. Its CDF is denoted by Φ.

Exponential Distribution

Let us say that X is distributed $\text{Expo}(\lambda)$. We know the following:

Story You're sitting on an open meadow right before the break of dawn, wishing that airplanes in the night sky were shooting stars, because you could really use a wish right now. You know that shooting stars come on average every 15 minutes, but a shooting star is not "due" to come just because you've waited so long. Your waiting time is memoryless; the additional time until the next shooting star comes does not depend on how long you've waited already.

Example The waiting time until the next shooting star is distributed $\text{Expo}(4)$ hours. Here $\lambda = 4$ is the **rate parameter**, since shooting stars arrive at a rate of 1 per 1/4 hour on average. The expected time until the next shooting star is $1/\lambda = 1/4$ hour.

Expos as a rescaled Expo(1)

$$Y \sim \text{Expo}(\lambda) \rightarrow X = \lambda Y \sim \text{Expo}(1)$$

Memorylessness The Exponential Distribution is the only continuous memoryless distribution. The memoryless property says that for $X \sim \text{Expo}(\lambda)$ and any positive numbers s and t,

$$P(X > s + t | X > s) = P(X > t)$$

Equivalently,

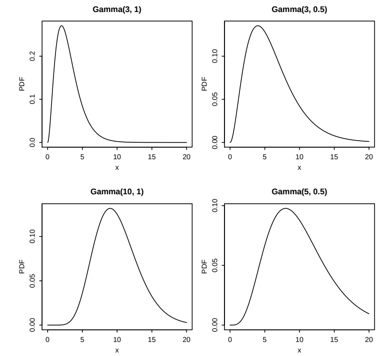
$$X - a|(X > a) \sim \text{Expo}(\lambda)$$

For example, a product with an $\text{Expo}(\lambda)$ lifetime is always "as good as new" (it doesn't experience wear and tear). Given that the product has survived a years, the additional time that it will last is still $\text{Expo}(\lambda)$.

Min of Expos If we have independent $X_i \sim \text{Expo}(\lambda_i)$, then $\min(X_1, \ldots, X_k) \sim \text{Expo}(\lambda_1 + \lambda_2 + \cdots + \lambda_k)$.

Max of Expos If we have i.i.d. $X_i \sim \text{Expo}(\lambda)$, then $\max(X_1, \ldots, X_k)$ has the same distribution as $Y_1 + Y_2 + \cdots + Y_k$, where $Y_j \sim \text{Expo}(j\lambda)$ and the Y_j are independent.

Gamma Distribution



Let us say that X is distributed $\text{Gamma}(a, \lambda)$. We know the following:

Story You sit waiting for shooting stars, where the waiting time for a star is distributed $\text{Expo}(\lambda)$. You want to see n shooting stars before you go home. The total waiting time for the nth shooting star is $\text{Gamma}(n, \lambda)$.

Example You are at a bank, and there are 3 people ahead of you. The serving time for each person is Exponential with mean 2 minutes. Only one person at a time can be served. The distribution of your waiting time until it's your turn to be served is $\text{Gamma}(3, \frac{1}{2})$.

χ^2 (Chi-Square) Distribution

Let us say that X is distributed χ^2_n. We know the following:

Story A Chi-Square(n) is the sum of the squares of n independent standard Normal r.v.s.

Properties and Representations

$$X \text{ is distributed as } Z_1^2 + Z_2^2 + \cdots + Z_n^2 \text{ for i.i.d. } Z_i \sim \mathcal{N}(0, 1)$$

$$X \sim \text{Gamma}(n/2, 1/2)$$

LLN, CLT

Law of Large Numbers (LLN)

Let $X_1, X_2, X_3 \ldots$ be i.i.d. with mean μ. The **sample mean** is

$$\bar{X}_n = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

The **Law of Large Numbers** states that as $n \to \infty$, $\bar{X}_n \to \mu$ with probability 1. For example, in flips of a coin with probability p of Heads, let X_j be the indicator of the jth flip being Heads. Then LLN says the proportion of Heads converges to p (with probability 1).

Central Limit Theorem (CLT)

Approximation using CLT

We use $\dot{\sim}$ to denote *is approximately distributed*. We can use the **Central Limit Theorem** to approximate the distribution of a random variable $Y = X_1 + X_2 + \cdots + X_n$ that is a sum of n i.i.d. random variables X_i. Let $E(Y) = \mu_Y$ and $\text{Var}(Y) = \sigma^2_Y$. The CLT says

$$Y \dot{\sim} \mathcal{N}(\mu_Y, \sigma^2_Y)$$

If the X_i are i.i.d. with mean μ_X and variance σ^2_X, then $\mu_Y = n\mu_X$ and $\sigma^2_Y = n\sigma^2_X$. For the sample mean $\bar{X}_n$, the CLT says

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \dot{\sim} \mathcal{N}(\mu_X, \sigma^2_X/n)$$

Asymptotic Distributions using CLT

We use $\xrightarrow{D}$ to denote *converges in distribution to* as $n \to \infty$. The CLT says that if we standardize the sum $X_1 + \cdots + X_n$ then the distribution of the sum converges to $\mathcal{N}(0,1)$ as $n \to \infty$:

$$\frac{1}{\sigma\sqrt{n}}(X_1 + \cdots + X_n - n\mu_X) \xrightarrow{D} \mathcal{N}(0,1)$$

In other words, the CDF of the left-hand side goes to the standard Normal CDF, Φ. In terms of the sample mean, the CLT says

$$\frac{\sqrt{n}(\bar{X}_n - \mu_X)}{\sigma_X} \xrightarrow{D} \mathcal{N}(0,1)$$

Continuous Multivariate Distributions

Joint Probability density $f(x,y) \rightsquigarrow P((X,Y) \in A) = \int_A f(x,y)$.

Marginal density $f_x(x) = \int_{\mathbb{R}} f(x,y)\, dy \rightsquigarrow P(X \in C) = \int_C f_x(x)\, dx$.

Multivariate Uniform Distribution

See the univariate Uniform for stories and examples. For the 2D Uniform on some region, probability is proportional to area. Every point in the support has equal density, of value $\frac{1}{\text{area of region}}$. For the 3D Uniform, probability is proportional to volume.

Multivariate Normal (MVN) Distribution

A vector $\vec{X} = (X_1, X_2, \ldots, X_d)$ is Multivariate Normal if every linear combination is Normally distributed, i.e., $t_1X_1 + t_2X_2 + \cdots + t_dX_d$ is Normal for any constants $t_1, t_2, \ldots, t_d$. The parameters of the Multivariate Normal are the **mean vector** $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)$ and the **covariance matrix** Σ where the (i,j) entry is $\text{Cov}(X_i, X_j)$.

Properties The Multivariate Normal has the following properties.

- Any subvector is also MVN.
- If any two elements within an MVN are uncorrelated, then they are independent.
- The joint PDF of a Multivariate Normal is: $f(x) = \det((2\pi)^d \mathbf{\Sigma})^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$

Distribution Properties

Convolutions of Random Variables

A convolution of n random variables is simply their sum. For the following results, let X and Y be *independent*.

1. $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2) \longrightarrow X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
2. $X \sim \text{Bin}(n_1, p)$, $Y \sim \text{Bin}(n_2, p) \longrightarrow X + Y \sim \text{Bin}(n_1 + n_2, p)$. $\text{Bin}(n,p)$ can be thought of as a sum of i.i.d. $\text{Bern}(p)$ r.v.s.
3. $X \sim \text{Gamma}(a_1, \lambda)$, $Y \sim \text{Gamma}(a_2, \lambda)$ $\longrightarrow X + Y \sim \text{Gamma}(a_1 + a_2, \lambda)$. $\text{Gamma}(n, \lambda)$ with n an integer can be thought of as a sum of i.i.d. $\text{Expo}(\lambda)$ r.v.s.
4. $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ $\longrightarrow X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Special Cases of Distributions

1. $\text{Bin}(1,p) \sim \text{Bern}(p)$
2. $\text{Beta}(1,1) \sim \text{Unif}(0,1)$
3. $\text{Gamma}(1,\lambda) \sim \text{Expo}(\lambda)$

Inequalities

1. **Cauchy-Schwarz** $|E(XY)| \le \sqrt{E(X^2)E(Y^2)}$
2. **Markov** $P(X \ge a) \le \frac{E|X|}{a}$ for $a > 0$
3. **Chebyshev** $P(|X - \mu| \ge a) \le \frac{\sigma^2}{a^2}$ for $E(X) = \mu, \text{Var}(X) = \sigma^2$
4. **Jensen** $E(g(X)) \ge g(E(X))$ for g convex; reverse if g is concave

Miscellaneous Definitions

Precision The *precision* of a distribution is the inverse of the variance $\tau = \frac{1}{\sigma^2}$.

Mode The mode of a *discrete* distribution is the point in the support that maximizes the *PMF*. The mode of a *continuous* distribution is the point in the support that maximizes the *PDF*.

Medians and Quantiles Let X have CDF F. Then X has median m if $F(m) \ge 0.5$ and $P(X \ge m) \ge 0.5$. For X continuous, m satisfies $F(m) = 1/2$. In general, the ath quantile of X is $\min\{x : F(x) \ge a\}$; the median is the case $a = 1/2$.

log Statisticians generally use log to refer to natural log (i.e., base e).

i.i.d r.v.s Independent, identically-distributed random variables.

Gamma and Beta Integrals

You can sometimes solve complicated-looking integrals by pattern-matching to a gamma or beta integral:

$$\int_0^\infty x^{t-1}e^{-x}\, dx = \Gamma(t) \qquad \int_0^1 x^{a-1}(1-x)^{b-1}\, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Also, $\Gamma(a+1) = a\Gamma(a)$, and $\Gamma(n) = (n-1)!$ if n is a positive integer.

Maximum likelihood

The RV X follows a parametric distribution $X \sim \mathcal{D}(\lambda)$. We don't know λ, but we have n independent observations $\{x_j\}_{j=1}^n$ from X. The **likelihood** of λ is:

- If $\mathcal{D}(\lambda)$ is discrete with mass function $p_\lambda(x)$:

$$\mathcal{L}\left(\lambda |\, \{x_j\}_{j=1}^n\right) = P\left(\{x_j\}_{j=1}^n |\lambda\right) = \prod_{j=1}^n \mathrm{p}_\lambda(x_j)$$

- If $\mathcal{D}(\lambda)$ es continuous with density $f_\lambda(x)$:

$$\mathcal{L}\left(\lambda |\, \{x_j\}_{j=1}^n\right) = \prod_{j=1}^n f_\lambda(x_j)$$

The *maximum likelihood estimator* of λ is the value λ^* that maximizes the likelihood:

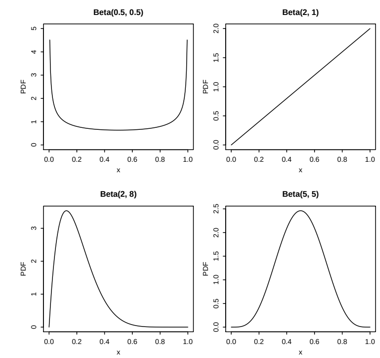
$$\lambda^* = \text{argmax}_\lambda\, \mathcal{L}\left(\lambda |\, (x_j)\right)$$

Conjugate families

In the Bayesian approach to statistics, parameters are uncertain, so we assign a probability distribution to them. The **prior** for a parameter is its distribution before observing data. The **posterior** is the distribution for the parameter after observing data.

The Beta family The Beta is a parametric family of distributions depending on two parameters a, b, used to represent uncertainty about a real number p known to lie in the interval $[0,1]$ (for instance, a probability).

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad x \in (0,1)$$



Beta is the Conjugate Prior of Bernoulli experiments Beta is the *conjugate* prior of the Binomial because if you have a Beta-distributed prior on p in a Binomial, then the posterior distribution on p given the Binomial data is also Beta-distributed. Consider the following two-level model:

$$X|p \sim \text{Bin}(n,p)$$
$$p \sim \text{Beta}(a,b)$$

Then after observing $X = x$, we get the posterior distribution

$$p|(X = x) \sim \text{Beta}(a + x, b + n - x).$$

Beta is also the *conjugate* prior of the Geometric: if you have a Beta-distributed prior on p, and the experiment follows a Geometric distribution based on p:

$$Y|p \sim \text{Geom}(p)$$
$$p \sim \text{Beta}(a,b)$$

Then after observing $Y = y$, we get the posterior distribution

$$p|(Y = y) \sim \text{Beta}(a + 1, b + x - 1).$$

Gamma is the Conjugate Prior of a Poisson Process If our uncertainty for the rate λ of a Poisson process is modelled with a $\text{Gamma}(\alpha, \beta)$, and we count x observations on a time interval of length T, then our posterior follows $\lambda \sim \text{Gamma}(\alpha + x, \beta + T)$.

Maximum A Posteriori (MAP) The MAP estimator is the mode of the posterior. It can be regarded as a *smoothed* version of the maximum likelihood estimator.

Objective priors

In absence of prior information, it is customary to use a prior that carries as little information as possible. These are called **objective priors**. There are several notions of objective prior, but most of them are **improper**: they are not true probability distributions, so we don't really have prior probabilistic information.
However, *an improper prior can be updated with data* to provide a *proper posterior*: a true probability distribution that can answer probabilistic questions and give expected values. Usually, the objective prior can be interpreted as a limiting case of the conjugate family, and *the updating rule for the conjugate family still holds.*

Uniform prior

The **uniform prior** for a real parameter assigns equal probability density to all admissible values of the parameter. This is called the *principle of indifference*. Hence, it is improper whenever the support is infinite.

Uniform priors for:

a probability p: the uniform distribution on $[0,1]$, which is also the Beta$(1,1)$ distribution.

the rate λ of a Poisson process: the uniform distribution on $\mathbb{R}^+$, which is *improper*, and is also the Gamma$(1,0)$ distribution.

the mean μ of a Normal $\mathcal{N}$ experiment with known precision τ: the uniform distribution on $\mathbb{R}$, which is *improper*, and is also the $\mathcal{N}(\mu=0,\tau=0)$ distribution.

the precision τ of Normal $\mathcal{N}$ experiment with known mean μ: the uniform distribution on $\mathbb{R}^+$, which is *improper*, and is also the Gamma$(1,0)$ distribution.

The *main drawback* of the principle of indifference is that *the prior associated to σ is different from the prior associated to σ^2 or the prior associated to $\tau=\frac{1}{\sigma^2}$* (it depends on the parameterization).

Jeffreys prior

The **Jeffreys prior** is invariant by reparameterization: the Jeffreys prior for σ, σ^2 and τ are all equivalent.

For a single parameter, the Jeffreys prior is also a **reference prior**: it maximizes the expected information gain from the data.

Jeffreys prior for:

a probability p: density $f(p)\propto\frac{1}{\sqrt{p(1-p)}}$: the Beta$(\frac{1}{2},\frac{1}{2})$ distribution.

the rate λ of a Poisson process: "pseudo-density" $f(\lambda)\propto\frac{1}{\sqrt{\lambda}}$: the Gamma$(\frac{1}{2},0)$ *improper* distribution.

the mean μ of a Normal $\mathcal{N}$ experiment with known precision τ: "pseudo-density" $f(\mu)\propto 1$: the uniform distribution on $\mathbb{R}$, which is *improper*, and is also the $\mathcal{N}(\mu=0,\tau=0)$ distribution.

the precision τ of Normal $\mathcal{N}$ experiment with known mean μ: "pseudo-density" $f(\tau)\propto\frac{1}{\tau}$: the Gamma$(0,0)$ distribution for the precision τ.

Regression

General regression model: some variables $X_1,\dots,X_n$ are known (not random), others $\varepsilon_1,\dots,\varepsilon_k$ are random. The goal is to understand better, and make predictions for the target variable Y. The function f is unknown:

$$Y=f(X_1,\dots,X_n,\varepsilon_1,\dots,\varepsilon_k)$$

Linear Regression

$$\begin{aligned}Y&=f(X_1,\dots,X_n,\varepsilon)\\&=\beta_0+\beta_1\cdot X_1+\dots+\beta_n\cdot X_n+\varepsilon\\\varepsilon&\sim\mathcal{N}(0,\sigma^2).\end{aligned}$$

Least squares regression: mininize $RSS=\sum(y_j-f(x_j))^2$. It's also the maximum likelihood estimation for the unknown parameters $\beta_0,\dots,\beta_n,\sigma$.

Software

A simulation in python

In a simulation, we write different code for the random experiment and the filters that check if the outcome belongs to different events. We freeze a random independent sample, and approximate all probabilities and means using the same random sample. In this way, we approximate the whole probability space by a sample, in which the Cardano rule applies, and we are guaranteed to work with a probability distribution (all probabilities lie between 0 and 1, correlations lie between -1 and 1, etc).

```
import random

# a random experiment
def one_dice():
    return random.randint(1,6)

# a "filter" decides if an outcome belongs to an event
def five_or_more(w):
    return w>=5
def is_even(w):
    return w%2==0

#sample size
N = 10000
#independent random sample
sample = [one_dice() for _ in range(N)]

#approximate the probability that dice is 5 or 6
probA = sum(1 for w in sample if five_or_more(w))/N
#approximate the probability that dice is even
probB = sum(1 for w in sample if is_even(w))/N
#approximate the probability of the intersection
probAB = sum(1 for w in sample
             if five_or_more(w) and is_even(w) )/N

#approximate the conditional probability
probA_cond_B = probAB/probB

#check the two events are independent (approximately)
print(probAB, probA*probB)

#approximate the expected value
mean = sum(w for w in sample)/N
#approximate the variance
var = sum((w-mean)**2 for w in sample)/N
```

A simulation with Random Variables

```
import random

# Omega is "throw three coins"
def threecoins():
    '''Returns a list with three elements, each one of
       them
    is either 1 (head) or 0 (tails)'''
    return [random.randint(0,1) for _ in range(3)]

# Two random variables are defined on Omega: X and Y
def X(w):
    'counts total number of heads'
    return sum(w)

def Y(w):
    'True if first toss is the same as the last'
    return w[0]==w[-1]
```

```
N = 1000
sample = [threecoins() for _ in random(N)]
# Aproximate E[X*Y]
eXY = sum(X(w)*Y(w) for w in sample)/N
```

Plotting

Histogram of discrete data

```
P = st.poisson(mu=1.8)
N = 1000
#A sample that takes values in the integers
sample = P.rvs(N)
max_integer = 10
plt.hist(sample, bins=[k+0.5 for k in range(-1,
    max_integer+1)], density=1,alpha=0.8)
```

Histogram of continuous data

```
E = st.expon(scale=3)
N = 100
sample = E.rvs(N)
plt.hist(sample, density=1,alpha=0.8)
```

Bar plot of a mass function

```
n0=5
B = st.binom(n=n0, p=0.25)
plt.bar(range(n0+1),[B.pmf(k) for k in range(n0+1)],
    fill=False)
```

Line plot of a density, or distribution function

```
E = st.expon(scale=3)
xmin, xmax = -3, 3 #plotting interval
N = 100 # number of subdivions
xs = np.linspace(xmin, xmax, N)
ys = E.pdf(xs) #use E.cdf(xs) for cumulative
    distribution function
plt.plot(xs, ys, 'g')
```

Combine plots, with labels and titles

```
E1 = st.expon(scale=1)
E2 = st.expon(scale=2)

xmin, xmax = -1,8
N = 100
xs = np.linspace(xmin, xmax, N)
ys1 = E1.pdf(xs)
plt.plot(xs, ys1, 'g', label='density function of
    Exponential(1)')

ys2 = E2.pdf(xs)
plt.plot(xs, ys2, 'b-', label='density function of
    Exponential(2)')

plt.title('Density functions of two Exponential
    distributions')
plt.xlabel('x')
plt.legend()
plt.show()
```

scipy.stats

A frozen distribution `N = scipy.stats.norm(loc=mean, scale=std)`

Random sample of size `M` `N.rvs(M)`

Mean `N.mean()`

Variance `N.var()`

Distribution function at points `xs` **(array)** `N.cdf(xs)`

Density function at points `xs` **(if continuous)** `N.pdf(xs)`

Mass function at points `xs` **(if discrete)** `N.pmf(xs)`

Percentiles `ps` `N.ppf(ps)`

pandas

Create a dataframe:

```
df = pd.DataFrame(data = {
    "calculus": [10,5,8,7],
    "algebra": [8,7,6,5],
    "probability": [7,6,6,8],
    },
    index = ["Jaimita", "Fulanito", "Menganito", "
        Zutanita"],
)
```

Browse first rows `df.head(2)`

Summary of column types `df.info()`

Column statistics `df.describe(include="all")`

Selecting a column `df["calculus"]` (the result is a `Series`)

max of a `Series` `df["calculus"].max()`

mean of a `Series` `df["calculus"].mean()`

std of a `Series` `df["calculus"].std()`

Selecting a column `df["calculus"]` (the result is a `Series`)

Selecting columns `df[["calculus", "probability"]]`

Selecting rows by index `df.loc[["Jaimita", "Fulano"]]`

Selecting rows by row number `df.iloc[1:3]`

Selecting rows and columns `df.loc[ list_of_indices, list_of_columns]`

Selecting rows by condition `df[df["calculus"]>7]`

Plot histogram `df["calculus"].hist()`

Scatter plot `df.plot.scatter("algebra", "calculus")`

Drop rows `df.drop(["Jaimita", "Fulano"], inplace=True)`

Drop columns `df.drop(["calculus", "probability"], inplace=True)`

Read a csv file `advertising = pd.read_csv("advertising.csv", usecols=[1,2,3,4])`

scikit-learn

Fit a linear model, print R^2 score:

```
import sklearn.linear_model as skl_lm
regr = skl_lm.LinearRegression()
X = advertising[["TV", "Radio", "Newspaper"]]
y = advertising["Sales"]
regr.fit(X,y)
print(regr.score())
```

Make predictions

```
advertising_future = pd.DataFrame(
    [   [100,30,30],
        [100,40,30],
    ],
    columns=["TV", "Radio", "Newspaper"]
)
regr.predict(advertising_future)
```

Fit a polinomial model, split randomly into train and test sets:

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import PolynomialFeatures
poly = PolynomialFeatures(degree=2)
X = poly.fit_transform(auto[["horsepower"]])
y = auto["mpg"]
Xtrain, Xtest, ytrain, ytest = train_test_split(X, y,
     test_size=0.25)
regr = skl_lm.LinearRegression()
regr.fit(Xtrain,ytrain)
print(regr.score(Xtest, ytest))
regr.predict(poly.fit_transform([[250]]))
```

statsmodels

```
import statsmodels.formula.api as smf
regr = smf.ols("Sales ~ TV + Radio", advertising).fit()
est.predict(advertising_future)
regr.summary()
```

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.897
Model:	OLS	**Adj. R-squared:**	0.896
Method:	Least Squares	**F-statistic:**	570.3
Date:	Tue, 09 Apr 2019	**Prob (F-statistic):**	1.58e-96
Time:	12:31:14	**Log-Likelihood:**	-386.18
No. Observations:	200	**AIC:**	780.4
Df Residuals:	196	**BIC:**	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
Radio	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

`R-squared` $R^2 = 1 - \frac{RSS}{TSS}$, where $RSS = \sum(y_j - f(x_j))^2$, $TSS = \sum(y_j - \bar{y})^2$. Always smaller than 1. The larger the better.

`adjusted R-squared` Adjusted - $R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$, where n is the number of data points, and p is the number of explanatory variables. The larger the better.

`AIC` Akaike Information Criterion. The smaller the better. The absolute value is not important. A difference of ≈ 1.4 between the AIC of model A and the AIC of model B means that model A is twice as likely as model B to minimize information loss, regardless of the magnitude of the AIC.

`BIC` Bayes Information Criterion. The smaller the better. As for AIC, only the differences in BIC matter, and not their absolute value.

`Intercept` Independent term in the linear model.

`P>|t|` p-value for t-statistic for each coefficient. If one of them is greater than 0.05, you should consider removing that explanatory variable.

`[0.025 0.975]` confidence interval for each coefficient. Values in the interval are not "unreasonable".

Computations with one dimensional Random Variables

Object of interest	**Finite**	**Infinite discrete**	**Continuous**	**Sample**
Support	A finite set F	An infinite but countable set I e.g. $\mathbb{N}, \mathbf{Z}, \dots$	A subset S of $\mathbb{R}$ e.g. $\mathbb{R}, (0,\infty), (a,b)$	A sample of size N $\{x_1,\dots,x_N\}$
$P(X \in A)$ **probability** of A	$\sum_{k\in A\cap F} p_X(k)$ p_X is the **mass function**	$\sum_{k\in A\cap I} p_X(k)$ p_X is the **mass function**	$\int_{A\cap S} f_X(x)\,dx$ f_X is the **density function**	$P(A) \approx P_{\text{sample}}(A)$ $P_{\text{sample}}(A) = \frac{\text{number of } x_i \text{ that lie in } A}{N}$
$P(X \le t)$		$F_X(t)$ F_X is the **distribution function**		$P(X \le t) \approx F_{\text{sample}}(A)$ $F_{\text{sample}}(A) = \frac{\text{number of } x_i \text{ smaller than } t}{N}$ faster to compute if the sample is ordered
$g(X)$ **transformation** of X by g g is *inyective*	$g(X)$ is finite $p_{g(X)}(k) = p_X(g^{-1}(k))$	$g(X)$ is discrete infinite $p_{g(X)}(k) = p_X(g^{-1}(k))$	$g(X)$ is continuous if g is smooth $f_{g(X)}(k) = f_X(g^{-1}(x))\left(g^{-1}\right)'(x)$	$\{g(x_1),\dots,g(x_N)\}$ is a sample of $g(X)$ of size N
$g(X)$ **transformation** of X by g g is **not** *inyective*		can get complicated		$\{g(x_1),\dots,g(x_N)\}$ is a sample of $g(X)$ of size N
$X\|A$ **conditioning** the RV X by the event A	$X\|A$ is finite $p_{X\|A}(k) = \frac{p_X(k)}{P(A)}$	$X\|A$ is discrete $p_{X\|A}(k) = \frac{p_X(k)}{P(A)}$	$X\|A$ is continuous $f_{X\|A}(x) = \frac{f_X(x)}{P(A)}$	filter $\{x_1,\dots,x_N\}$ keep only the x_j that lie in A get a sample of $X\|A$ of size *smaller than* N
$E[X]$ **expectation** of X	$\sum_{k\in F} k\, p_X(k)$ a finite **sum**	$\sum_{k\in I} k\, p_X(k)$ an infinite **series**	$\int_S x\, f_X(x)\,dx$ an **integral**	$E[X] \approx \text{sample mean} = \frac{\Sigma_{i=1}^N x_i}{N}$
$E[g(X)]$ **expectation** of $g(X)$	$\sum_{k\in F} g(k)\, p_X(k)$ a finite **sum**	$\sum_{k\in I} g(k)\, p_X(k)$ an infinite **series**	$\int_S g(x)\, f_X(x)\,dx$ an **integral**	$E[g(X)] \approx \frac{\Sigma_{i=1}^N g(x_i)}{N}$
$X+Y$ **sum** of RVs X and Y		can get complicated (involves "convolutions") except in a few special cases		$\{x_1+y_1,\dots,x_N+y_N\}$ is a sample of $X+Y$ of size N
X follows a parametric distribution $X \sim \mathcal{D}(Y,Z,\dots)$ the parameters $Y,Z,\dots$ are RVs		rather complicated except in a few special cases		first sample $y_j \in Y, z_j \in Z, \dots$ then sample x_j from $\mathcal{D}(y_j,z_j,\dots)$ $\{x_1,\dots,x_N\}$ is a sample of X of size N

Table of Distributions

Distribution	PMF/PDF and Support	Expected Value	Variance	scipy.stats
Discrete Uniform DisUniform$(1, \ldots, n)$	$P(X = k) = 1/n$ $k = 1, \ldots, n$	$\frac{1+n}{2}$	$\frac{n^2-1}{12}$	`randint(low=1, high=n+1)`
Bernoulli Bern(p)	$P(X = 1) = p$ $P(X = 0) = q = 1 - p$	p	pq	`bernoulli(p=p0)`
Binomial Bin(n, p)	$P(X = k) = \binom{n}{k} p^k q^{n-k}$ $k \in \{0, 1, 2, \ldots n\}$	np	npq	`binom(n=n0, p=p0)`
Geometric Geom(p)	$P(X = k) = (1-p)^{k-1} p$ $k \in \{1, 2, \ldots\}$	$1/p$	$\frac{1-p}{p^2}$	`geom(p=p0)`
Poisson Pois(μ)	$P(X = k) = \frac{e^{-\mu}\mu^k}{k!}$ $k \in \{0, 1, 2, \ldots\}$	μ	μ	`poisson(mu=mu0)`
Uniform Unif(a, b)	$f(x) = \frac{1}{b-a}$ $x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	`uniform(loc=a, scale=b-a)`
Normal $\mathcal{N}(\mu, \sigma^2)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$ $x \in (-\infty, \infty)$	μ	σ^2	`norm(loc=mu,scale=sigma)`
Exponential Expo(λ)	$f(x) = \lambda e^{-\lambda x}$ $x \in (0, \infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	`expon(scale=1/lambd)`
Gamma Gamma(α, β)	$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ $x \in (0, \infty)$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	`gamma(a=alpha, scale=1/beta)`
Beta Beta(a, b)	$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$ $x \in (0, 1)$	$\mu = \frac{a}{a+b}$	$\frac{\mu(1-\mu)}{(a+b+1)}$	`beta(a=a0, b=b0)`
Multivariate Normal $\mathcal{N}(\mu, \Sigma)$	$f(x) = \det((2\pi)\boldsymbol{\Sigma})^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$	μ	Σ	`multivariate_normal(mean=mu, cov=Sigma)`

Maximum likelihood and Conjugate distributions

Data	Likelihood	Unknown Parameters	Max Likelihood	Conjugate prior	Conjugate posterior	MAP
x is 0 or 1 a single Bernoulli trial	Bernoulli $X \sim \text{Bern}(p)$	a *probability* $p \in [0,1]$	$\hat{p} = x$	$p \sim \text{Beta}(a,b)$	$p \sim \text{Beta}(a, b+1)$ if $x = 0$ $p \sim \text{Beta}(a+1, b)$ if $x = 1$	$p = \frac{a+x-1}{a+b-1}$
x_j is 0 or 1 $j \in \{1, \dots, n\}$ n Bernoulli trials	Bernoulli $X_j \sim \text{Bern}(p)$	a *probability* $p \in [0,1]$	$\hat{p} = \frac{\sum_{j=1}^n x_j}{n}$	$p \sim \text{Beta}(a,b)$	$p \sim \text{Beta}(a+e, b+f)$ $e = \sum_{j=1}^n x_j$ successes $f = n - \sum_{j=1}^n x_j$ failures	$p = \frac{a + \sum_{j=1}^n x_j - 1}{a+b+n-2}$
$x \in \{0, \dots, n\}$ a binomial experiment with n items	Binomial $X \sim \text{Bin}(p,n)$	a *probability* $p \in [0,1]$	$\hat{p} = \frac{x}{n}$	$p \sim \text{Beta}(a,b)$	$p \sim \text{Beta}(a+x, b+f)$ x successes, $f = n - x$ failures	$p = \frac{a+x-1}{a+b+n-2}$
$x \in \{1, 2 \dots\}$ a single geometric experiment	Geometric $X \sim \text{Geom}(p)$	a *probability* $p \in [0,1]$	$\hat{p} = \frac{1}{x}$	$p \sim \text{Beta}(a,b)$	$p \sim \text{Beta}(a+1, b+f)$ 1 success, $f = x - 1$ failures	$p = \frac{a}{a+b+x-2}$
$x \in \{1, 2 \dots\}$ $j \in \{1, \dots, n\}$ n geometric experiments	Geometric $X_j \sim \text{Geom}(p)$	a *probability* $p \in [0,1]$	$\hat{p} = \frac{n}{\sum_{j=1}^n x_j}$	$p \sim \text{Beta}(a,b)$	$p \sim \text{Beta}(a+n, b+f)$ n successes $f = \sum x_j - n$ failures	$p = \frac{a+n-1}{a+b+n+f-2}$
$x \in \{1, 2 \dots\}$ a Poisson experiment on a time interval of length T	Poisson $X \sim \text{Pois}(T\lambda)$	the process *rate* $\lambda > 0$	$\hat{\lambda} = \frac{x}{T}$	$\lambda \sim \text{Gamma}(\alpha, \beta)$	$\lambda \sim \text{Gamma}(\alpha + x, \beta + T)$ x observations, time T	$\lambda = \frac{\alpha+x-1}{\beta+T}$
$t_j \in \mathbb{R}^+$ time between observations of Poisson process $j \in \{1, \dots, n\}$	Exponential $X_j \sim \text{Expo}(\lambda)$	a *rate* $\lambda > 0$	$\hat{\lambda} = \frac{n}{\sum_{j=1}^n t_j}$	$\lambda \sim \text{Gamma}(\alpha, \beta)$	$\lambda \sim \text{Gamma}(\alpha + n, \beta + T)$ n observations total time $T = \sum_{j=1}^n t_j$	$\lambda = \frac{\alpha+n-1}{\beta + \sum_{j=1}^n t_j}$
$x_j \in \mathbb{R}$ a Gaussian with *known mean* μ $j \in \{1, \dots, n\}$	Gaussian $X_j \sim \mathcal{N}(\mu, \sigma)$	the Gaussian *variance* or the Gaussian *precision* $\tau = \frac{1}{\sigma^2} > 0$	$\hat{\sigma^2} = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$ $\hat{\tau} = \frac{n}{\sum_{i=1}^n (x_i - \mu)^2}$	$\tau \sim \text{Gamma}(\alpha, \beta)$	$\tau \sim \text{Gamma}(\tilde{\alpha}, \tilde{\beta})$ $\tilde{\alpha} = \alpha + \frac{n}{2}$ $\tilde{\beta} = \beta + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2}$	$\frac{\tilde{\alpha}-1}{\tilde{\beta}}$
$x_j \in \mathbb{R}$ a Gaussian with *known precision* $\tau = \frac{1}{\sigma^2}$ $j \in \{1, \dots, n\}$	Gaussian $X_j \sim \mathcal{N}(\mu, \sigma)$	the Gaussian *mean* $\mu \in \mathbb{R}$	$\hat{\mu} = \frac{\sum_{j=1}^n x_j}{n}$	$\mu \sim \mathcal{N}(m, t = \frac{1}{s^2})$	$\mu \sim \mathcal{N}(\tilde{m}, \tilde{t})$ $\tilde{m} = \frac{t\, m + \tau \sum_{i=1}^n x_i}{t + n\tau}$ $\tilde{t} = t + n\tau$	$\tilde{m}$
$x_j \in \mathbb{R}$ a Gaussian with *unknown parameters*	Gaussian $X_j \sim \mathcal{N}(\mu, \sigma)$	the Gaussian *mean and variance* $\mu \in \mathbb{R}, \sigma \in \mathbb{R}$	$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$ $\hat{\sigma^2} = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2$		... Normal-Gamma(m, t, α, β) ...	
$\mathbf{x}_j \in \mathbb{R}^p$ a Gaussian vector with *unknown parameters*	Gaussian $X_j \sim \mathcal{N}(\mu, \Sigma)$	the Gaussian parameters $\mu \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p \times p}$	$\hat{\mu} = \bar{\mathbf{x}} = \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j$ $\hat{\Sigma} = \frac{1}{n} \sum_{j=1}^n (\mathbf{x}_j - \bar{\mathbf{x}}) \cdot (\mathbf{x}_j - \bar{\mathbf{x}})^t$		... Normal-Wishart ...	