Quantitative Tools - Level 2 - Fall 2017
Louis de Charsonville

Lecture 6

Probability (2/2)*

---

1 Random variables, Expected Value and Variance

1.1 Random variable

Definition 1.1. Random variable
Suppose we have an experiment whose outcome depends on chance. We called the outcome of the experiment a *random variable*, usually denoted by the roman letter X. If the sample space of the experiment is either finite or countably infinite[1], the random variable is said to be *discrete*.

Example 1.1. A die is rolled once. The sample space, Ω, of the experiment is:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Let X be the outcome of this experiment. We can compute the following probability:

$$P(X \leq 3) = \frac{1}{2}$$

Definition 1.2. Uniform distribution
The uniform distribution on a sample space Ω containing n elements is the function m defined by:

$$\forall \omega \in \Omega,\ m(\omega) = \frac{1}{n}$$

1.2 Expected value

So far, we have studied the computation of the probability of a single event but usually, we are usually interested in descriptive quantities such as the average, the median or the standard-deviation.

Definition 1.3. Expected value
Let X be a discrete random variable with a sample space Ω and distribution function $m(x)$. The expected value, denoted $E(X)$, is defined by:

$$E(X) = \sum_{x \in \Omega} x m(x)$$

We often refer to the expected value as the *mean*[2].

---

*The notes draws extensively on the Grinstead and Snell's "Introduction to Probability", 2006, which is freely accessible on the world wide web.

[1] A set is said to be *countably infinite* if one can count set's elements, in other words, that for each item in the set, there is an natural number associated. If you can't think of what could be an *uncoutable infinite* set, think of the set of real numbers between 0 and 1.

[2] The sum might be equal to the infinite. In such a case, we say that the sum *does not converge* and X does not have an expected value.

Example 1.2. A dice is rolled. If the dice falls on an odd number, we win an amount equal to the number, otherwise we loose the amount equal to the number. Let X the payout from the roll of dice, then $\Omega = \{-6, -4, -2, 1, 3, 5\}$.

$$\begin{aligned} E(X) &= \sum_{x \in \Omega} x * P(X = x) \\ &= \frac{1}{6}(-6 - 4 - 2 + 1 + 3 + 5) \\ &= -\frac{1}{2} \end{aligned}$$

Theorem 1.1. Sum of two random variables
Let X, Y be two random varibles with finite expected values. Then

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

Properties 1.1. Let X be a random variable, and c any constant, then

$$E(cX) = cE(X)$$

Theorem 1.2. Product of two independent random variables
Let X, Y be two *independent* random varibles with finite expected values. Then:

$$E(XY) = E(X)E(Y)$$

Example 1.3. Link between expected value and average value We flip a coin. If it is a head, we earn €1, if it is a tail, we loose €1 . The expected value is:

$$\begin{aligned} E(X) &= -1 * \frac{1}{2} + 1 * \frac{1}{2} \\ &= 0 \end{aligned}$$

We simulate 200 games where the coin is flipped 1000 times per game. Figure 1 is the plot of the outcome. Each blue line is one of the 200 games, the *x-axis* is the number of tosses and the red line the average of the blue lines.

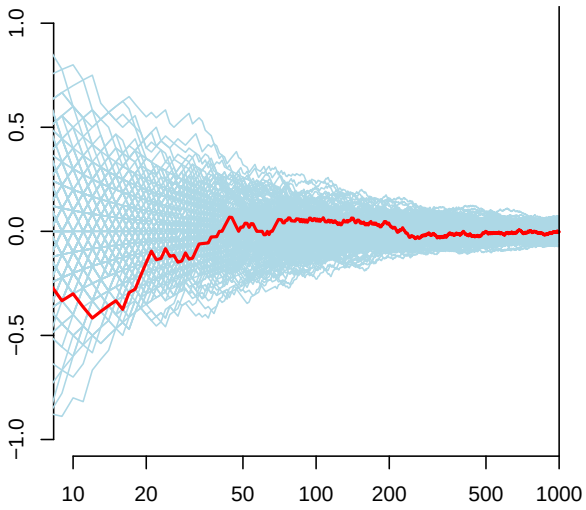

Figure 1: 200 simulations of coin flipped 1000 times

Example 1.4. St Petersburg Paradox
Let's consider the following game: a coin is flipped n times until a heads comes out. The player earns 2^n. How much should the player be willing too pay ?

Answer:
The player should play the game if the expected value is positive. Let's denote Q the price of the game. We want:

$$\begin{aligned} E(X) - Q &= 0 \\ Q &= E(X) \\ Q &= \sum_{n=0} P(X = n) * X \end{aligned}$$

What is the probability that a heads comes out after n flips ? It is the probability that the coin falls on tail $n-1$ times which is $\frac{1}{2^{n-1}}$ and that the coin then falls on head:

$$P(X = n) = \frac{1}{2^n}$$

Then:

$$\begin{aligned} Q &= \sum_{n=0}^{+\infty} \frac{1}{2^n} 2^n \\ &= \sum_{n=0} 1 \end{aligned}$$

1.3 Variance

The expected value measures the average outcome but another interesting metrics considering probability is by how much the outcome is likely to deviate from the average value. This is what variance measure.

Definition 1.4. Variance
Let X be a random variable, then the *variance* of X, denoted by $V(X)$ or σ_X^2, is:

$$\begin{aligned} V(X) &= E((X - E(X))^2) \\ &= E(X^2) - E(X)^2 \end{aligned}$$

Definition 1.5. Standard deviation
The *standard deviation* of X, denoted σ_X, is $\sigma_X = \sqrt{V(X)}$

Example 1.5. Let's compute the variance in the setup of example 1.2

$$\begin{aligned} V(X) &= E(X^2) - E(X)^2 \\ &= \frac{1}{6}(36 + 16 + 4 + 1 + 9 + 25) - \frac{1}{4} \\ &= \frac{91}{6} - \frac{1}{4} \\ &= \frac{179}{12} \end{aligned}$$

Properties 1.2. For any X a random variable, and c a constant:

$$V(cX) = c^2 V(X)$$

Properties 1.3. For any X,Y two *independent* random variables[3], we have:

$$V(X + Y) = V(X) + V(Y)$$

Definition 1.6. Covariance
Let X and Y be two random variables. Then the *covariance* of X and Y, denoted by $Cov(X, Y)$ is:

$$Cov(X, Y) = E((X - E(X))(Y - E(Y)))$$

Properties 1.4. Let X and Y two independent random variables. Then $Cov(X, Y) = 0$.

[3] This property can be generalized for n independent random variables.

Proof
We use the expected value of the product of two *independent* random variables is the product if their expected values, i.e. $E(XY) = E(X)E(Y)$.

$$\begin{aligned} Cov(X,Y) &= E((X - E(X))(Y - E(Y))) \\ &= E(XY - XE(Y) - YE(X) + E(X)E(Y)) \\ &= E(XY) - E(X)E(Y) \\ &= E(X)E(Y) - E(X)E(Y) \\ &= 0 \end{aligned}$$

Properties 1.5. Sum of two random variables
For any X,Y two random variables, we have:

$$V(X+Y) = V(X) + V(Y) - 2Cov(X,Y)$$

2 Discrete probability distributions

In this section, we describe the discrete probability distributions that occur the most often in applied statistics.

2.1 Uniform distribution

The uniform distribution is the distribution whereby a finite number of values are equally likely to be observed. For instance, a dice roll or a flipped coin follows uniform distribution. If n is the number of possible outcomes, and $\omega_1, \ldots, \omega_n$ the outcomes, then:

$$\forall \omega_i \in \{\omega_1, \ldots, \omega_n\},\ P(X = \omega_i) = \frac{1}{n}$$

The discrete uniform distribution itself is non-parametric (i.e. does not depend of exogenous parameters). However, it is convenient to represent its values by all integers in an interval $[a, \ldots, b]$.
The **expected value** and the **variance** write:

$$\begin{aligned} E(X) &= \frac{a+b}{n} \\ V(X) &= \frac{(b-a+1)^2 - 1}{12} \end{aligned}$$

2.2 Binomial distribution

Definition 2.1. Bernoulli Trial
A *bernoulli* trial is a random experiment with *exactly* two possible outcomes. One is usually named "success" and occurs with probability p. It is named after Jacob Bernoulli, a Swiss mathematician.

Definition 2.2. Binomial Distribution
The binomial distribution, with parameters $\{k, n, p\}$, is the probability that an event occurs k times after n of Bernoulli trial with probability p.

$$b(n,p,k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Example 2.1. The binomial distribution describes the number k of heads after n tosses.

Properties 2.1. Expected value and variance

- The **expected value** of a binomial distribution of parameters,n et p, $(B)(n,p)$ is $E(X) = np$.
- The **variance** of $(B)(n,p)$ is $np(1-p$.

2.3 Poisson distribution

The Poisson distribution is one of the most famous distribution. It is mostly used to model modelling the number of times an event occurs in an interval of time or space. Notably, it is used to model the frequency of rare events, like claim modelling by insurance companies, the number of airplane accidents, or waiting times at a bus station.

2.3.1 Intuition

Suppose[4] that we have a situation in which a certain kind of occurrence happens at random over a period of time. For example, the occurrences that we are interested in might be incoming telephone calls to a police station in a large city. We assume that the average rate, i.e. the average number of occurrences per minute, is a constant, denoted λ. Our next assumption is that the number of occurrences in two non-overlapping time intervals are independent. In our example, this means that the events that there are j calls between 5:00 and 5:15 P.M. and k calls between 6:00 and 6:15 P.M. on the same day are independent.

We can use the binomial distribution to model this situation. We imagine that a given time interval is broken up into n subintervals of equal length. If the subin- tervals are sufficiently short, we can assume that two or more occurrences happen in one subinterval with a probability which is negligible in comparison with the probability of at most one occurrence. Thus, in each subinterval, we are assuming that there is either 0 or 1 occurrence. This means that the sequence of subintervals can be thought of as a sequence of Bernoulli trials, with a success corresponding to an occurrence in the subinterval.

To decide upon the proper value of p, the probability of an occurrence in a given subinterval, we reason as follows. On the average, there are λt occurrences in a time interval of length t. If this time interval is divided into n subintervals, then we would expect, using the Bernoulli trials interpretation, that there should be np occurrences. Thus, we want:

$$\begin{aligned} \lambda t &= np \\ p &= \frac{\lambda t}{n} \end{aligned}$$

We now consider the number of occurences in a given time interval, X. We want to calculate the distribution of X. To ease calculations, let's assume that $t = 1$.
We want to compute $P(X = k)$. We are going to do it by induction. First, using the binomial distribution, we know that:

$$\begin{aligned} P(X = 0) &= b(n, p, 0) \\ &= (1 - p)^n \\ &= \left(1 - \frac{\lambda}{n}\right)^n \\ &\approx e^{-\lambda} \quad \text{when n is big} \end{aligned}$$

Then:

$$\begin{aligned} P(X = k) &= b(n, p, k) \\ &= \binom{n}{k} p^k (1 - p)^{n-k} \\ &= \frac{\lambda - (k-1)p}{k(1-p)} \binom{n}{k} p^{k-1} (1 - p)^{n-k+1} \\ &= \frac{\lambda - (k-1)p}{k(1-p)} b(n, p, k-1) \\ &\approx \frac{\lambda}{k}\, b(n, p, k-1) \quad \text{when n is big} \end{aligned}$$

When n is big and therefore p is small:

$$P(X = k) \approx \frac{\lambda^k}{k!} e^{-\lambda}$$

The above distribution is the *Poisson distribution*. The Poisson distribution is an approximation to the binomial distribution when the parameter n are large, and p is small.

Definition 2.3. Poisson distribution
An event occur λ times in an interval. We call λ the event-rate. The probability of observing k events in an interval is given by:

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

[4] Example stolen from Chapter 5, Distribution and Densities, Grinstead and Snell's Introduction to Probability, 2009.

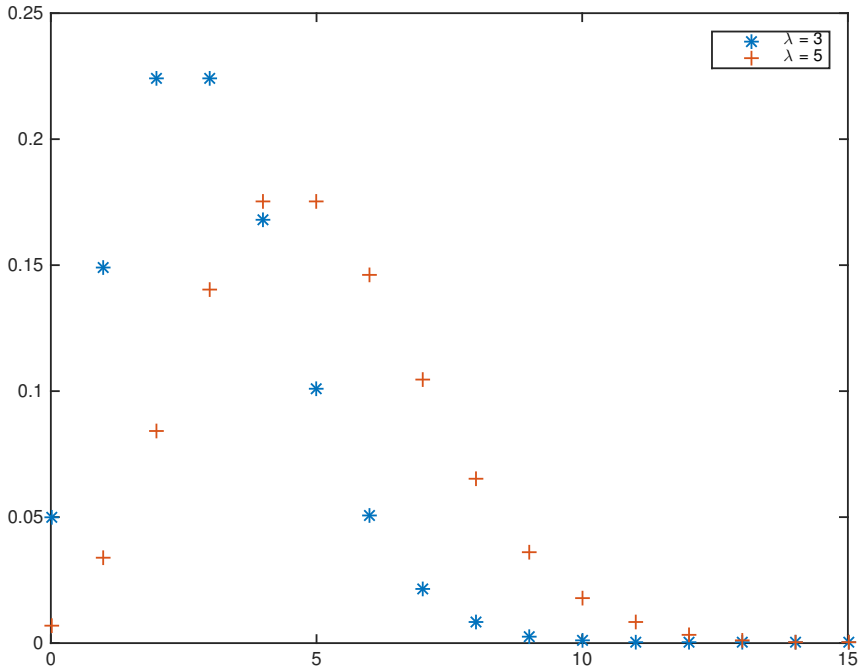

Figure 2: Poisson processes

Properties 2.2. Expected value
The expected value of a Poisson distribution of parameter λ, denoted $\mathcal{P}(k)$, is λ.

Properties 2.3. Variance
The *variance* of a Poisson distribution of parameter *lambda*, denoted $\mathcal{P}(k)$, is λ.

2.4 Hypergeometric distribution

The **hypergeometric distribution** is used to model draws without replacement (whereas the geometric distribution models draws with replacement).

Definition 2.4. Hypergeometric distribution
Suppose that we have a set of N balls, of which p % are red. We choose n of these balls without replacement and define k the number of red balls among our sample. The probability to observe k red balls in a sample of n draws from a population of size N, denoted $\mathcal{H}(N, n, k)$:

$$\mathcal{H}(N, n, k) = \frac{\binom{Np}{k}\binom{N - Np}{n - k}}{\binom{N}{n}}$$

Properties 2.4. Expected value
The expected value of an hypergeometric distribution of parameter N, n, k:

$$E(X)_{\mathcal{H}(N,n,k)} = n * p$$

Properties 2.5. Variance
The variance of an hypergeometric distribution of parameter N, n, k:

$$V(X)_{\mathcal{H}(N,n,k)} = \frac{np(1-p)(N-n)}{N-1}.$$

3 Continuous probability distributions

3.1 Continuous random variables

So far we have studied discrete random variables, i.e. the sample space was a discrete set. We shall now study continuous random variables for which the sample space is real-valued, which means the sample space is an

uncountable set.

Example 3.1. Spinner
We begin by constructing a spinner, which consists of a circle of unit circumference and a pointer. We pick a point on the circle and label it 0, and then label every other point on the circle with the distance, denoted x, from 0 to that point, measured counterclockwise. The experiment consists of spinning the pointer and recording the label of the point at the tip of the pointer. We let the random variable X denote the value of this outcome. The sample space is the interval $[0, 1]$, which is *uncountable*.
A consequence of the sample space being *uncountable* is that:

$$\forall \omega \in \Omega,\ P(X = \omega) = 0$$

on the contrary:

$$P(0 \leq X \leq 1) = 1$$

since the spinner comes to rest *somewhere* in the circle.
We also have :

$$P(0 \leq X < \frac{1}{2}) = P(\frac{1}{2} \leq X < 1) = \frac{1}{2}$$

More generally:

$$P(a \leq X \leq b) = b - a$$

Definition 3.1. Density function
Let X be a continuous real-valued random variable. A *density function* for X is a real-valued function f which satisfies:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Definition 3.2. Cumulative distribution function (CDF)
The *cumulative distribution function* of a real-valued random variable X is the probability that X will take a value less than or equal to x. We usually denote it $F_X(x)$:

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= \int_{-\infty}^{x} f_X(t)dt \end{aligned}$$

Definition 3.3. Expected value
Let X be real-valued random variable. We say that X has an expected value if $\int_{-\infty}^{x} t f_X(t)dt$ converges. We denote the expected value, $E(X)$:

$$E(X) = \int_{-\infty}^{x} t f_X(t)dt$$

Definition 3.4. Variance
The variance of a random variable X, when it exists, is:

$$V(X) = E((X - E(X)^2)$$

3.2 Normal distribution

The normal distribution is the most important density function.

Definition 3.5. Normal density
The normal density function with parameters μ and σ is defined as:

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

We usually denote the normal distribution $\mathcal{N}(\mu, \sigma)$ and say that a random variable follows a normal distribution $X \sim \mathcal{N}(\mu, \sigma)$.

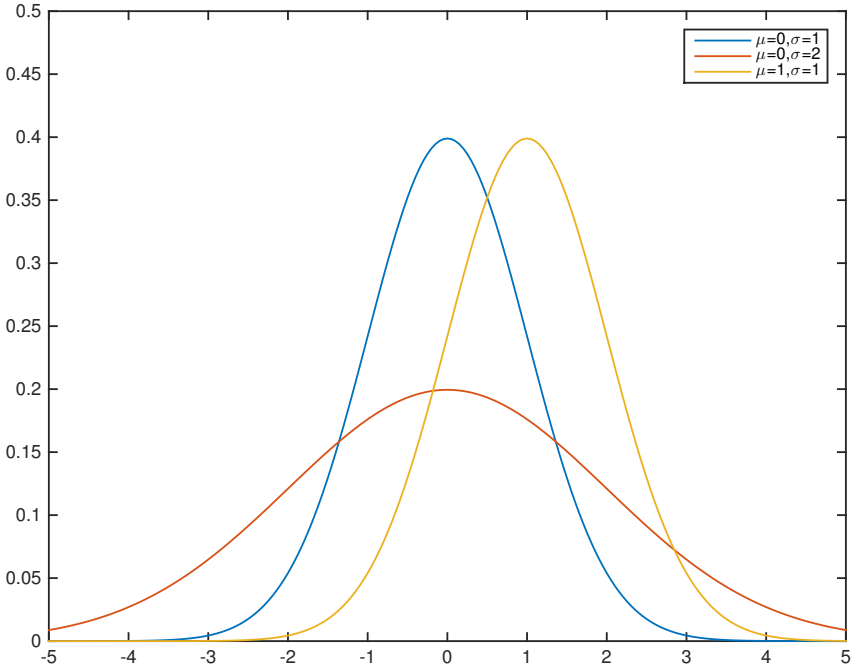

Figure 3: Normal densities

Properties 3.1. Expected value and variance

- The expected value of a normal distribution $\mathcal{N}(\mu, \sigma)$ is μ.
- The expected value of a normal distribution $\mathcal{N}(\mu, \sigma)$ is σ^2.

Definition 3.6. Standard normal distribution The standard normal distribution is the normal distribution with parameters $\mu = 0$ and $\sigma = 1$.

4 Fundamental theorem of probabilities

4.1 Law of Large numbers

Intuition Sometimes coined as the *law of averages*, the law of large numbers states the average of the results obtained from a large number of trials is close to the expected value. It is to be related with the *frequency interpretation of probability*: the probability of a certain outcome can be seen as the frequency with which that outcome occurs in the long run.

Theorem 4.1. Markov Inequality
Let X be a *discrete* random variable with expected value $\mu = E(X)$, and let ϵ ¿ 0 be any positive real number. Then:

$$P(|X| \geq \epsilon) \leq \frac{E(|X|)}{\epsilon} \tag{1}$$

Theorem 4.2. Chebyshev Inequality
Let X be a *discrete* random variable with expected value $\mu = E(X)$, and let ϵ ¿ 0 be any positive real number. Then

$$P(|X - \mu| \geq \epsilon) \leq \frac{V(X)}{\epsilon^2}$$

Theorem 4.3. Law of Large Numbers
Let $X_1, X_2, \ldots, X_n$ be an independent trials process, with finite expected value $\mu = E(X_j)\ \forall j \in [1, n]$ and finite variance $\mu = V(X_j)\ \forall j \in [1, n]$. We denote $S_n = X_1 + X_2 + \cdots + X_n$. Then $\forall \epsilon > 0$:

$$P(|\frac{S_n}{n} - \mu| \geq \epsilon) \xrightarrow[n \to \infty]{} 0$$

And equivalently:

$$P(|\frac{S_n}{n} - \mu| < \epsilon) \xrightarrow[n \to \infty]{} 1$$

4.2 Central Limit Theorem

The second fundamental theorem of probability is the *Central Limit Theorem*. This theorem says that if S_n is the sum of n mutually independent random variables, then the distribution function of S_n is well-approximated by a certain type of continuous function known as a normal density function.

Theorem 4.4. Central Limit Theorem
Let $(X_n)_{n\geq 1}$, a sequence of n independent and indentically distributed random variables. We denote the standardized sum S_n^* by :

$$S_n^* = \frac{\frac{\sum_{i=1}^n X_i}{n} - \mu}{\sigma/\sqrt{n}}$$

Then:

$$S_n^* \xrightarrow[n\to\infty]{} \mathcal{N}(0,1) \tag{2}$$

Application to statistics The central limit theorem is widely used in statistics to compute interval of confidence, p-values and other statistics of interests.

* * *

Bibliography

These notes draw heavily on the following books that are worth reading by the interested reader:

- Grinstead and Snell's Introduction to Probability, Peter G. Doyle, 2009.
- Introduction to Statistics, David M. Lane, 2013.
- De l'analyse à la prévision, Volume 3, Didier Schlacther, 2009.
- Analysis of Economic Data, Third Edition, Gary Koop, 2009.