# A Few Reminders on the Probability Theory and Statistics

Olga Klopp

## 1 The law of a random variable

A random variable is a variable whose possible values are outcomes of a random phenomenon. As a function, a random variable is required to be measurable, which rules out certain pathological cases. Here is the definition of the law of a random variable:

**Definition 1.** *A random variable $\Omega \longrightarrow \mathbb{R}$ is a measurable function from a set of possible outcomes $\Omega$ to $\mathbb{R}$. Then, the* **law of** $X$*, denoted by $\mathbb{P}^X$ , is a probability measure on $(\mathbb{R}, \mathcal{B})$ such that*

$$\mathbb{P}^X\left[A\right] = \mathbb{P}\left[X^{-1}(A)\right] = \mathbb{P}[X \in A], \ \ \forall A \in \mathcal{B}.$$

*where $\mathcal{B}$ denotes the Borel sets of $\mathbb{R}$*

The probability distribution of a random variable is often characterised by a small number of parameters, which also have a practical interpretation. For example, it is often enough to know what its "average value" is. This is captured by the **expected value** of a random variable (also called the first moment). Moments can be defined for real-valued functions of random variables: for any measurable function $\varphi$

$$\boxed{\mathbb{E}\left[\varphi(X)\right] = \int_{\Omega} \varphi\big(X(\omega)\big)\, \mathbb{P}(d\omega) = \int_{\mathbb{R}} \varphi(x)\, \mathbb{P}^X(dx)}.$$

**Example 1 :** $X$ follows Bernoulli's law of parameter $1/3$

- The law of $X$ is described by
$$\mathbb{P}\left[X = 1\right] = \tfrac{1}{3} = 1 - \mathbb{P}\left[X = 0\right]$$

- Writing of $\mathbb{P}^X$ :
$$\boxed{\mathbb{P}^X = \tfrac{1}{3}\delta_1 + \tfrac{2}{3}\delta_0}$$

  Here $\delta_1$ ($\delta_0$) is the Dirac delta function at 1 (at 0). Delta function is a generalized function that is equal to zero everywhere except for one (for zero) and whose integral over the entire real line is equal to one

- Computing the expectation:

$$\begin{aligned}
\mathbb{E}\left[\varphi(X)\right] &= \int_{\mathbb{R}} \varphi(x)\, \mathbb{P}^X(dx) \\
&= \tfrac{1}{3} \int_{\mathbb{R}} \varphi(x)\delta_1(dx) + \tfrac{2}{3} \int_{\mathbb{R}} \varphi(x)\delta_0(dx) \\
&= \tfrac{1}{3}\, \varphi(1) + \tfrac{2}{3}\, \varphi(0)
\end{aligned}$$

### 1.1 Binomial probability distribution

There is an entire class of problems that are characterized by the feature that there are exactly two possible outcomes (for each trial) of interest. These problems are called binomial experiments, or Bernoulli experiments.

Features of a binomial experiment:

- There are a fixed number of trials. We denote this number by $n$.
- The $n$ trials are independent and repeated under identical conditions.
- Each trial has only two outcomes: success and failure.
- For each individual trial, the probability of success is the same. We denote the probability of success by $p$ and that of failure by $q$. Since each trial results in either success or failure, $p + q = 1$.

Anytime we make selections from a population without replacement, we do not have independent trials. However, replacement is often not practical. If the number of trials is quite small with respect to the population, we almost have independent trials, and we can say the situation is closely approximated by a binomial experiment.

**Example 2** Suppose we select 20 tuition bills at random from a collection of $10,000$ bills issued at one college and observe if each bill is in error or not. If 600 of the $10,000$ bills are in error, then the probability that the first one selected is in error is 0.06. If the first is in error, then the probability that the second is in error is 0.0599. Even if the first 19 bills selected are in error, the probability that the 20th is also in error is 0.0582. All these probabilities round to 0.06, and we can say that the independence condition is approximately satisfied.

If the population is relatively small and we draw samples without replacement, the assumption of independent trials is not valid and we should not use the binomial distribution (if the sample size is so small that sampling without replacement results in trials that are not even approximately independent $\rightarrow$ the hypergeometric distribution).

**Formula for the binomial probability distribution:**

$$\boxed{\mathbb{P}^X(X = k) = \frac{n!}{k!(n-k)!} p^k q^{n-k} = \binom{n}{k} p^k q^{n-k}}$$

- $n =$ number of trials
- $p =$probability of success on each trial
- $q = 1 - p$probability of failure on each trial
- $k =$the number of successes out of $n$ trials
- $\binom{n}{k}$ =the binomial coefficient (the number of combinations of $n$ distinct objects taken $k$ at a time)

In the formula for $\mathbb{P}^X(X = k)$ there are two parts: the expression $p^k q^{n-k}$ is the probability of getting one outcome with $k$ successes and $n-k$ failures; the binomial coefficient $\binom{n}{k}$ counts the number of outcomes that have $k$ successes and $n - k$ failures.

**Example 2 :** $X$ follows Binomial law of parameters $n = 6$ and $p = 0.5$

Computing the expectation:

$$\begin{aligned}\mathbb{E}\left[X\right] &= \sum_{k=1}^{6} k \binom{6}{k} (0.5)^k (0.5)^{6-k} \\ &= n \times p = 6 \times 0.5 = 3\end{aligned}$$

## 1.2 Geometric probability distribution

In many real-life situations, we keep on trying until we achieve success. Suppose we have an experiment in which we repeat binomial trials until we get our first success, and then we stop. Let $n$ be the number of the trial on which we get our first success. In this context, $n$ is not a fixed number. In fact, $n$ could be any of the numbers $1, 2, 3$, and so on. What is the probability that our first success comes on the nth trial? The answer is given by the geometric probability distribution.

**Formula for the geometric probability distribution:**

$$\boxed{\mathbb{P}^X(X = n) = p(1-p)^{n-1}}$$

- $n$ =number of trial on which the first success occurs
- $p$ =probability of success on each trial

### 1.3 Poisson Probability Distribution

If for the binomial distribution the number of trials $n$ gets larger and larger while the probability of success $p$ gets smaller and smaller, we obtain the Poisson distribution.

**Formula for the Poisson probability distribution:**

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

- $\lambda$ =the mean number of successes over time, volume, area, ...
- $k$ =the number of successes
- $\binom{n}{k}$ =the binomial coefficient (the number of combinations of $n$ distinct objects taken $k$ at a time)

Example 2 : $X\sim$ follows Poisson law of parameter 2

- The law of $X$ is described by
$$\mathbb{P}\left[X=k\right]=\frac{2^k}{k!}e^{-2},\ \ k=0,1,\dots$$
- Writing of $\mathbb{P}^X$ :
$$\boxed{\mathbb{P}^X=e^{-2}\sum_{k\in\mathbb{N}}\tfrac{2^k}{k!}\delta_k}$$
- Computing expectation:
$$\mathbb{E}\left[X\right]=\int_{\mathbb{R}}x\,\mathbb{P}^X(dx)=e^{-2}\sum_{k\in\mathbb{N}}k\tfrac{2^k}{k!}$$

There are many applications of the Poisson distribution. For example, if we take the point of view that waiting time can be subdivided into many small intervals, then the actual arrival (of whatever we are waiting for) during any one of the very short intervals could be thought of as an infrequent (or rare) event. This means that the Poisson distribution can be used as a mathematical model to describe the probabilities of arrivals such as cars to a gas station, planes to an airport, calls to a fire station ...

**Poisson Approximation to the Binomial Distribution:**

Consider binomial distribution. If the number of trials $n$ is large while the probability $p$ of success is quite small, we call the event (success) a "rare" event. For most practical purposes, the Poisson distribution with $\lambda=np$ will be a very good approximation to the binomial distribution provided the number of trials $n\geq 100$ and $\lambda n\leq 10$. As $n$ gets larger and $p$ gets smaller, the approximation becomes better and better.

### 1.4 Normal distribution

One of the most important examples of a continuous probability distribution is the normal distribution.

Important properties of a normal curve:

1. The curve is bell-shaped, with the highest point over the mean $\mu$
2. The curve is symmetrical about a vertical line through $\mu$
3. The inflection points between cupping upward and downward occur above $\mu+\sigma$ and $\mu-\sigma$.
4. The parameter $\sigma$ controls the spread of the curve: if the standard deviation $\sigma$ is large, the curve will be more spread out; if it is small, the curve will be more peaked.

Exemple 3 : $X \sim \mathcal{N}(0,1)$ (standard normal law).

- The law of $X$ is described by
$$\mathbb{P}\left[X \in [a,b]\right] = \int_{[a,b]} e^{-x^2/2} \frac{dx}{\sqrt{2\pi}}$$

- Writing of $\mathbb{P}^X$ : density
$$\boxed{f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}}$$

- Computing expectation
$$\mathbb{E}\left[\varphi(X)\right] = \int_{\mathbb{R}} \varphi(x)\, \mathbb{P}^X(dx) = \int_{\mathbb{R}} \varphi(x) e^{-x^2/2} \frac{dx}{\sqrt{2\pi}}$$

**Normal approximation to the binomial distribution:**

**Example** The probability that a new vaccine will protect adults from cholera is known to be 0.85. The vaccine is administered to 300 adults who must enter an area where the disease is prevalent. What is the probability that more than 280 of these adults will be protected from cholera by the vaccine?

This question falls into the category of a binomial experiment with the number of trials n equal to 300, the probability of success $p$ equal to 0.85, and the number of successes $k$ greater than 280. It is possible to use the formula for the binomial distribution to compute the probability that $k$ is greater than 280. However, this approach would involve a number of tedious and long calculations. An easier way to do this problem using normal approximation to the binomial distribution.

The Normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$ will be a good approximation to the binomial distribution provided $np > 5$ and $nq > 5$. As $n$ increases, the approximation becomes better.

## 1.5 Distribution Function

- The law of a random variable $X$ is a "complicated object":
  - it can be discreet (e.g. Bernoulli)
  - it can be (absolutely) continuous with respect to the Lebesgue measure $\lambda$ (e.g. Normal distribution)
  - or a combination of both ...
- We can **characterize the law** of $X$ by a simpler object: an increasing bounded function, the **distribution function**.
- Easier to study in a **statistical context**.
- This approach has some limitations...

**Definition 2.** *The cumulative distribution function of a real-valued random variable $X$, evaluated at $x$, is the probability that $X$ will take a value less than or equal to $x$:*
$$F(x) := \mathbb{P}\left[X \leq x\right],\ \forall x \in \mathbb{R}.$$

- $F$ is increasing, continuous to the right (only the limit from the right is required to equal the value of the function), $F(-\infty) = 0$, $F(+\infty) = 1$
- $F$ **characterizes** the law $\mathbb{P}^X$ :
$$\boxed{\mathbb{P}^X\left[(a,b]\right] = \mathbb{P}\left[a < X \leq b\right] = F(b) - F(a)}$$
- From now on, the **law** of $X$ will indicate indifferently $F$ or $\mathbb{P}^X$.

### 1.6 Properties of Variance

**Definition 3.** *Variance is the expectation of the squared deviation of a random variable from its mean:*

$$\mathrm{Var}(X) = \mathbb{E}\left(X - \mathbb{E}\,X\right)^2$$

1. $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$
2. $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$
3. In general: $\mathrm{Var}(X + Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$
4. If $X$ and $Y$ are **independent**, then $\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$ and $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

## 2 Type of convergence of sequence of random variables

The convergence of sequences of random variables to some limit random variable is an important concept. There exist several different notions of convergence of random variables. They formalize the idea that a sequence of random events can sometimes be expected to settle down into a "pattern" when items are far enough into the sequence. For example, the sequence eventually takes a constant value or the values in the sequence continue to change but can be described by an unchanging probability distribution.

### 2.1 Convergence in distribution

**Idea:** we increasingly expect to see the next outcome in a sequence becoming better and better modeled by a given probability distribution.

Convergence in distribution is the weakest form of convergence (it is implied by all other types of convergence). It is very used in practice; most often it arises from application of the central limit theorem.

**Definition 4.** *A sequence $(X_n)$ of real-valued random variables is said to converge in distribution (or converge in law) to a random variable $X$ if*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

*for every number $x \in \mathbb{R}$ at which $F$ is continuous. Here $F_n$ and $F$ are the cumulative distribution functions of random variables $X_n$ and $X$, respectively.*

**Notation:** $X_n \xrightarrow{d} X$

1. The requirement that **only the continuity points** of $F$ should be considered is essential.

   **Example** If $X_n$ are distributed uniformly on intervals $(0, 1/n)$ then this sequence converges in distribution to a degenerate random variable $X = 0$. We have $F(0) = 1$ and $F_n(0) = 0$ for all $n$.

2. We have that

   $$\mathbb{E}\,f(X_n) \to \mathbb{E}\,f(X)$$

   for all continuous bounded functions $f : \mathbb{R} \mapsto \mathbb{R}$

**Example: Tossing coins.** Let $X_n$ be the fraction of heads after tossing up an unbiased coin $n$ times. Then $X_1$ has the Bernoulli distribution with expected value $\mu = 0.5$ and variance $\sigma^2 = 0.25$. The subsequent random variables $X_2, X_3, ...$ will all be distributed binomially.

As $n$ grows larger, this distribution will gradually start to take shape more and more similar to the bell curve of the normal distribution. If we shift and rescale $X_n$ appropriately, then $Z_n = \frac{\sqrt{n}}{\sigma}(X_n - \mu)$ will be converging in distribution to the standard normal (follows from the central limit theorem).

## 2.2 Convergence in probability

**Idea:** the probability of an "unusual" outcome becomes smaller and smaller as the sequence progresses.

Convergence in probability is used very often in statistics. For example, an estimator is called consistent if it converges in probability to the quantity being estimated.

**Definition 5.** *A sequence $X_n$ of random variables converges in probability towards the random variable $X$ if for all $\epsilon > 0$*

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| > \varepsilon\right) = 0$$

*That is, pick any $\epsilon > 0$ and any $\delta > 0$. Let $P_n$ be the probability that $X_n$ is outside the ball of radius $\epsilon$ centered at $X$. Then for $X_n$ to converge in probability to $X$ there should exist a number $N$ (which will depend on $\epsilon$ and $\delta$) such that for all $n \geq N, P_n < \delta$.*

**Notation:** $X_n \xrightarrow{\mathbb{P}} X$

1. Convergence in probability implies convergence in distribution.
2. In the opposite direction, convergence in distribution implies convergence in probability when the limiting random variable $X$ is a constant.

**Example archer:** Suppose a person takes a bow and starts shooting arrows at a target. Let $X_n$ be his score in $n$-th shot. Initially he will be very likely to score zeros, but as the time goes and his archery skill increases, he will become more and more likely to hit the bullseye and score 10 points. After years of practice the probability that he hit anything but 10 will be getting increasingly smaller and smaller and will converge to 0. Thus, the sequence $X_n$ converges in probability to $X = 10$.

Note that $X_n$ does not converge almost surely however. No matter how professional the archer becomes, there will always be a small probability of making an error. Thus the sequence $X_n$ will never turn stationary: there will always be non-perfect scores in it, even if they are becoming increasingly less frequent.

## 2.3 Almost sure convergence

This is the type of convergence that is most similar to the usual pointwise convergence.

**Definition 6.** *$X_n$ converges almost surely (or with probability 1) towards $X$ if*

$$\mathbb{P}\Big(\lim_{n\to\infty} X_n = X\Big) = 1$$

*This means that the values of $X_n$ approach the value of $X$, in the sense that events for which $X_n$ does not converge to $X$ have probability* $0$.

**Notation:** $X_n \xrightarrow{a.s.} X$

1. Almost sure convergence implies convergence in probability (by Fatou's lemma), and hence implies convergence in distribution.

**Example** You start with \$1 and each day you play a game in which you either lose all your money or triple it with equal probability. Let $X_n$ be a random variable that represents the amount of money you have after $n$ days. This sequence of random variables almost surely converges to the random variable $X = 0$.