Probability (2)

STAT 315

Contents

Chapter 1 : Events and random variables

1 Probability axioms

Example 1: Roll a die. Suppose the outcomes $1, \ldots, 6$, have probabilities $1/4, 1/4, 1/8, 1/8, 1/8, 1/8$ respectively. What is the probability of (a) an even number, (b) a prime number?

Example 2: Same experiment but we are now given the following information:

$$P(\{1,2\}) = P(\{3,4\}) = P(\{5,6\}) = 1/3, \quad P(\{1,2,3\}) = 1/2.$$

What is (a) $P(\{3\})$, (b) P({ 4 }), (c) P({ 6 })?

Example 3: Same experiment but we are now given the following information:

$$P(\{1,2\}) = P(\{3,4\}) = P(\{5,6\}) = 1/2, \quad P(\{1\}) = 1/4.$$

What is $P(\{2\})$?

Formalise this intuition.

The sample space Ω is an arbitrary set, thought of as the set ofpossible outcomes. Events are subsets of the sample space, and probabilities are numbers assigned to events. The collection of events, $\mathcal{F}$, to which we assign probabilities has a certain structure:

1. $\Omega \in \mathcal{F}$
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
3. If $A_n \in \mathcal{F}$ for $n = 1, 2, 3, ...$, then $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$.

A collection of subsets of Ω having these properties is called a σ-algebra.

Probability axioms: A function $P : \mathcal{F} \to \mathbb{R}$ is called a probability if

1. $0 \leq P(A) \leq 1$ for all $A \in \mathcal{F}$, and $P(\Omega) = 1$.
2. If $A_1, A_2, \ldots$ are disjoint sets, then $P(\cup_{n=1}^{\infty} A_n = \sum_{n=1}^{\infty} P(A_n)$.

Example 4: A dart is thrown at a circular dartboard of radius 1m and is equally likely to land anywhere on it. (What does this mean?) How likely is it to land in (a) the top half, (b) the bull's eye, which is a central circle of radius 1cm, (c) the point with co-ordinates $(+0.2, -0.3)$, (d) the horizontal diameter?

The point of this example is to illustrate why it is not possible to extend the axioms to uncountable sums.

2 Conditional probability and independence

Example: Let's go back to the example of rolling a die, where the outcomes $1, \ldots, 6$, have probabilities $1/4, 1/4, 1/8, 1/8, 1/8, 1/8$ respectively. Given that an even number was rolled, how likely are the events (a) $\{2\}$, (b) $\{3, 4, 5\}$?

Definitions: For events A and B, if $P(B) > 0$, then $P(A|B) := P(A \cap B)/P(B)$. Note: $P(\cdot|B)$ is a probability.

We say that events A and B are independent of each other if $P(A \cap B) = P(A)P(B)$. If $P(B) > 0$, this is the same as saying that $P(A|B) = P(A)$.

Events $A_1, \ldots, A_n$ are mutually independent if

$$P(\cap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i). \tag{1}$$

An infinite sequence of events is said to be mutually independent if every finite subcollection of them is mutually independent. (Why would it be a bad idea to define it analogous to (1)?)

Total probability formula and Bayes' formula: Let $A_1, A_2 \ldots, A_n$ be a partition of Ω, i.e., the sets are mutually disjoint and their union is Ω.

(I mean measurable partition, but I'll omit the qualifier henceforth on the grounds that we will only consider measurable sets.) Then, for any event B,

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B).$$

Therefore,

$$\boxed{P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i \cap B)}{\sum_{j=1}^{n} P(A_j \cap B)}.}$$

This formula can be used to compute all the $P(A_i|B)$ if we are given all the $P(A_i)$ and $P(B|A_i)$.

We can define conditional independence just like independence. We say that A and B are conditionally independent given C if

$$P(A \cap B|C) \;=\; P(A|C)P(B|C).$$

3 Random variables

Probability spaces become a little more interesting when we define random variables on them. A real valued function X defined on Ω is said to be a **random variable** if for every Borel set $B \subset \mathbf{R}$ we have $X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathcal{F}$. When we need to emphasize the σ-field, we will say that X is $\mathcal{F}$-**measurable** or write $X \in \mathcal{F}$. If Ω is a discrete probability space (see Example 1.1.1), then any function $X : \Omega \to \mathbf{R}$ is a random variable. A second trivial, but useful, type of example of a random variable is the **indicator function** of a set $A \in \mathcal{F}$:

$$1_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases} .$$

The notation is supposed to remind us that this function is 1 on A . Analysts call this object the characteristic function of A .

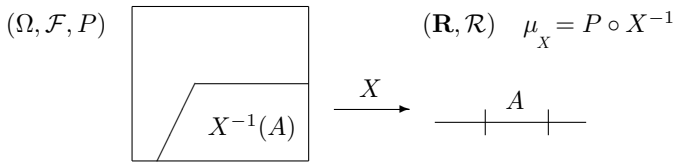

Figure 3.1: Definition of the distribution of X

Definition A random variable is a (measurable) function from the sample space to the real numbers.

Example: $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F}$ = all subsets,

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

Often, the sample space will be implicit and we'll just write X instead of $X(\omega)$.

If X is a random variable, then X induces a probability measure on $\mathbf{R}$ called its **distribution** by setting $\mu_X(A) = P(X \in A)$ for Borel sets A. Using the notation introduced above, the right-hand side can be written as $P(X^{-1}(A))$. In words, we pull $A \in \mathcal{R}$ back to $X^{-1}(A) \in \mathcal{F}$ and then take P of that set.

To check that μ is a probability measure we observe that if the A_i are disjoint then using the definition of μ; the fact that X lands in the union if and only if it lands in one of the A_i; the fact that if the sets $A_i \in \mathcal{R}$ are disjoint then the events $\{X \in A_i\}$ are disjoint; and the definition of μ again; we have:

$$\mu\left(\cup_i A_i\right) = P\left(X \in \cup_i A_i\right) = P\left(\cup_i \{X \in A_i\}\right) = \sum_i P(X \in A_i) = \sum_i \mu(A_i).$$

4 Discrete and Continuous random variables

Definition: A random variable X is said to be discrete if there is a a set $\{x_1, x_2, \ldots\}$, and a positive sequence $p_1, p_2, \cdots$ such that

$$\boxed{P(X = x_n) \;=\; p_n \;.}$$

Definition: A random variable X is said to becontinuous if there is a non-negative function f such that, for any interval (x, y),

$$\boxed{P(X \in (x, y)) = \int_x^y f(u)du.}$$

The function f is called the probability density function of X.

We write f_X if we need to make clear which random variable we are talking about. Observe that

Distribution functions

The cumulative distribution function (C.D.F) has the advantage of being completely general, and applying to both discrete and continuous random variables (and mixtures of the two).

The distribution of a random variable X is usually described by giving its **distribution function**, $F(x) = P(X \leq x)$.

Clearly, F must be non-decreasing and right-continuous (why?), and we must have $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. The function F is called the distribution function (or cumulative distribution function, cdf) of the random variable X. We write F_X when we want to make it clear which random variable we are talking about.

Theorem: *Any distribution function F has the following properties:*

(i) F is nondecreasing.

(ii) $\lim_{x\to\infty} F(x) = 1$, $\lim_{x\to-\infty} F(x) = 0$.

(iii) F is right continuous, i.e. $\lim_{y\downarrow x} F(y) = F(x)$.

(iv) If $F(x-) = \lim_{y\uparrow x} F(y)$ *then* $F(x-) = P(X < x)$.

(v) $P(X = x) = F(x) - F(x-)$.

Theorem: *If F satisfies (i), (ii), and (iii) in Theroem 1.2.1, then it is the distribution function of some random variable.*

In the case of discrete random variables, we were able to specify the probability of each possible outcome. That isn't possible for continuous random variables. What we want is to be able to specify the probability of every "measurable" subset of the real numbers.

Observe that $P(X \in (x, y)) = P(x \in [x, y])$ for continuous r.v.s.
How is F related to f? $F(x) = P(X \leq x)$ has the form

$$\boxed{F(x) = \int_{-\infty}^{x} f(y)\, dy}$$

Examples of discrete random variables:

1. Bernoulli(p): Models the outcome of a coin toss.
$$P(X = 1) = p, \quad P(X = 0) = 1 - p.$$
2. Binomial(n, p): Models the number of heads in n coin tosses.
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \le k \le n.$$
(Is this a valid probability distribution/ probability mass function?) Note: Can construct X as $X = Y_1 + \ldots + Y_n$, where the Y_i are iid Bernoulli(p).
3. Poisson(λ):
$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots$$
4. Geometric(p): Models the number of coin tosses until seeing the first head. $P(X = k) = (1-p)^{k-1} p$, $k = 1, 2, 3, \ldots$

Where does the Poisson distribution come from? Consider random variables $X_1, X_2, \ldots$ where X_n is Binomial($n, \lambda/n$). Fix $k \ge 0$ and look at $P(X_n = k)$ as n tends to infinity. We have

$$\begin{aligned} P(X_n = k) &= \binom{n}{k} \Big(\frac{\lambda}{n}\Big)^k \Big(1 - \frac{\lambda}{n}\Big)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n \cdot n \cdots n} \frac{1}{k!} \lambda^k \Big(1 - \frac{\lambda}{n}\Big)^n \Big(1 - \frac{\lambda}{n}\Big)^{-k} \\ &\to \frac{1}{k!} \lambda^k e^{-\lambda}. \end{aligned}$$

Roughly speaking, the Poisson distribution models the number of occurences of an event which is *individually rare* but where there is a *large population* of individuals where it could occur. An example is the number of life insurance policy holders of a given age who die in a given year. (This is a bit of a simplification, a compound Poisson would be a better model.) Another example is the number of atoms in a sample undergoing radioactive decay in a given time period. Poisson apparently arrived at this model by studying the number of deaths in the Prussian army due to being kicked by horses.

Examples of continuous distributions

1. Uniform($[a, b]$), $a < b$:
$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}, \quad F(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b. \end{cases}$$

2. Exponential(λ), $\lambda > 0$: $f(x) = \lambda e^{-\lambda x} 1(x \geq 0)$,

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

3. Gamma(α, λ), $\alpha, \lambda > 0$:

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} \lambda e^{-\lambda x} (\lambda x)^{\alpha - 1}, & x \geq 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $\Gamma(\alpha) := \int_0^\infty x^{\alpha-1} e^{-x} dx$. Here, α is called the shape parameter and λ is called the scale parameter.

4. Normal(μ, σ^2):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big).$$

The parameters μ and σ^2 are in fact the mean and variance of this distribution (to be defined).

The exponential distribution is used to model the lifetime of things whose "frailty" doesn't change with age. What do we mean by this? Let X be an Exp(λ) random variable denoting the lifetime of a light bulb, say. Conditional on the light bulb having survived up to time t, what is the probability that it will survive until time $t + s$? We can calculate this using Bayes' formula. We have

$$\begin{aligned} P(X > t + s | X > t) &= \frac{P(\{X > t + s\} \cap \{X > t\})}{P(X > t)} = \frac{P(X > t + s)}{P(X > t)} \\ &= \frac{\exp(-\lambda(t + s))}{\exp(-\lambda t)} = e^{-\lambda s} = P(X > s). \end{aligned} \tag{2}$$

In other words, the probability that the light bulb will survive for another s time units is the same no matter how old the light bulb is.

Examples of the exponential distribution in nature include the radioactive decay of nuclei, where the probability that a nucleus decays in some time interval $(s, t]$ doesn't depend on how old the nucleus is at time s. (The residual lifetime is independent of the age.)

Suppose Y is a Gamma random variable with parameter (α, λ) and α is a whole number. Then, we can obtain Y as

$$Y = X_1 + X_2 + \ldots + X_\alpha,$$

where the X_i are iid Exponential random variables with parameter λ.

The normal distribution is possibly the most famous distribution in all of probability. It is also known as the Gaussian distribution, after Carl Friedrich Gauss, who used it to model errors in astronomical observations. It owes its ubiquity in probability and statistics to the Central Limit Theorem, which we'll see later.

Chapter 2 : Expectation and distribution of random variables

1 Expectation and variance

Intuitively, the average of a data set is one way of describing the "centre" of the data set. It is not the only way; the median is another example. The average, or mean, is defined for a data set $x_1, \ldots, x_n$ as

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

There is a related quantity defined for random variables, called their expectation.

If X is discrete and takes values in the set $\{x_1, x_2, \ldots\}$, it is defined as

$$\boxed{E[X] = \sum_{n \in \mathbb{N}} x_n P(X = x_n),}$$

whever the sum is absolutely convergent. If X is continuous , it is defined as

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

whever the integral makes sense (integrability). Using the Riemann-Stieltjes integral, both cases can be combined by defining it as

$$E[X] = \int_{-\infty}^{\infty} x dF(x).$$

What is the connection between the mean of a data set and the expectation of a random variable? If the random variable were defined as a uniform random sample from the data set, they would be the same.

Alternatively, if we were to generate a large data set by considering repeated, independent realisations of the random variable, then the mean of this data set would be close to the expectation of the random variable. This statement is called the Law of Large Numbers.

Expectation of functions of a random variable: Let X be a random variable and g a function. Then, $Y = g(X)$ is another random variable (it is

a function on the sample space defined by $Y(\omega) = (g \circ X)(\omega))$. To compute its expectation as above, we would first have to compute the distribution of the random variable Y. In fact, it turns out that there is an easier way:

$$E[Y] = \int_{-\infty}^{\infty} g(x) dF(x).$$

This should be intuitively obvious, at least in the discrete case.

Example: Let X be the outcome of the roll of a fair die, and define

$$Y = g(X) = \begin{cases} 1, & \text{if } X \text{ is even,} \\ 0, & \text{if } X \text{ is odd.} \end{cases}$$

Then, it is clear that $P(Y = 1) = 1/2$ and $P(Y = 0) = 1/2$, so $E[Y] = 0.5$. Alternatively, we have

$$E[Y] = \sum_{n=1}^{6} g(n) P(X = n) = (0 + 1 + 0 + 1 + 0 + 1)\frac{1}{6} = \frac{1}{2}.$$

Properties of the expectation:

1. Expectation is linear: For any two random variables X and Y defined on the same sample space, and any constants a and b, $E[aX + bY] = aE[X] + bE[Y]$. This is a very important property of expectations. Again, it is pretty easy to see in the discrete case. The result extends in the obvious way to sums of finitely many random variables.

2. The expectation of a constant is equal to that constant. (Think of a constant as a random variable which takes only one value, with probability 1).

Variance: Let X be a random variable and let us denote $E[X]$ by μ. The variance of X is defined as $\mathrm{Var}(X) = E[(X - \mu)^2]$. Another way to write this is

$$\begin{aligned} \mathrm{Var}(X) &= E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - (E[X])^2. \end{aligned}$$

Note that the variance has to be non-negative, because it is the expectation of a non-negative random variable. Thus, we have shown that $E[X^2] \geq (EX)^2$ for any random variable X. It is also clear from the definition that, for any real numbers a and b,

$$\boxed{\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X),}$$

and that the variance of a constant is zero.

Computing expected Values

Example 1 If X has an **exponential distribution** with rate 1 then

$$EX^k = \int_0^\infty x^k e^{-x} dx = k!$$

So the mean of X is 1 and variance is $EX^2 - (EX)^2 = 2 - 1^2 = 1$. If we let $Y = X/\lambda$, then Y has the **exponential density** with parameter λ. We conclude that Y has expectation $1/\lambda$ and variance $1/\lambda^2$.

Example 2 If X has a **standard normal distribution**,

$$EX = \int x(2\pi)^{-1/2} \exp(-x^2/2)\, dx = 0 \quad \text{(by symmetry)}$$

$$\text{var}\,(X) = EX^2 = \int x^2(2\pi)^{-1/2} \exp(-x^2/2)\, dx = 1\,.$$

If we let $\sigma > 0$, $\mu \in \mathbf{R}$, and $Y = \sigma X + \mu$, then (b) of Theorem 1.6.1 and (1.6.4), imply $EY = \mu$ and $\text{var}\,(Y) = \sigma^2$. Actually, the random variable Y has density

$$(2\pi\sigma^2)^{-1/2} \exp(-(y-\mu)^2/2\sigma^2)$$

the **normal distribution** with mean μ and variance σ^2.

Example 3 Recall X has a **Bernoulli distribution** with parameter p if $P(X = 1) = p$ and $P(X = 0) = 1 - p$. Clearly,

$$EX = p \cdot 1 + (1-p) \cdot 0 = p\,.$$

Since $X^2 = X$, we have $EX^2 = EX = p$ and

$$\text{var}\,(X) = EX^2 - (EX)^2 = p - p^2 = p(1-p).$$

Example 4 Recall X has a **Poisson distribution** with parameter λ if

$$P(X = k) = e^{-\lambda}\lambda^k/k! \text{ for } k = 0, 1, 2, \ldots$$

To evaluate the moments of the Poisson random variable, we use a little inspiration to observe that for $k \geq 1$

$$\begin{aligned} E(X(X-1)\cdots(X-k+1)) &= \sum_{j=k}^{\infty} j(j-1)\cdots(j-k+1)e^{-\lambda}\frac{\lambda^j}{j!} \\ &= \lambda^k \sum_{j=k}^{\infty} e^{-\lambda}\frac{\lambda^{j-k}}{(j-k)!} = \lambda^k \end{aligned}$$

where the equalities follow from the facts that (i) $j(j-1)\cdots(j-k+1) = 0$ when $j < k$, (ii) cancelling part of the factorial, (iii) the fact that Poisson distribution has total mass 1. Using the last formula, it follows that $EX = \lambda$ while

$$\text{var}\,(X) = EX^2 - (EX)^2 = E(X(X-1)) + EX - \lambda^2 = \lambda\,.$$

2 Joint and marginal distributions

Let us build up our intuition starting with discrete random variables. Roll two "independent" fair dice. (We haven't yet defined independence for random variables but just use your intuition.) Let X be the number shown on the first die, Y on the second and Z their sum. (Note that all three random variables are defined on the same sample space, $\{1,\ldots,6\}\times\{1,\ldots,6\}$.) The joint distribution of (X,Y) can be specified by writing down the probability of each of the 36 possible outcomes (i,j), $1\le i,j\le 6$. Likewise, the joint distribution of (X,Z) can be specified by specifying probabilities of 36 different events of the form (i,j), $1\le i\le 6$, $i+1\le j\le i+6$. In both cases, the function $(i,j)\mapsto p(i,j)$ is called the probability mass function.

Given the joint pmf for (X,Z), we can compute the "marginal" pmf for one of them, say Z. For example, suppose we are given the following joint probabilities:

$$p_{X,Z}(1,4)=p_{X,Z}(2,4)=p_{X,Z}(3,4)=\frac{1}{36},$$
$$p_{X,Z}(4,4)=p_{X,Z}(5,4)=p_{X,Z}(6,4)=0,$$

where $p_{X,Z}(i,j)$ denotes $P(X=i,Z=j)$. Then we can compute

$$p_Z(4):=P(Z=4)=\sum_{i=1}^{6}p_{X,Z}(i,4)=\frac{3}{36}.$$

Definitions: Let $X_1,\ldots,X_n$ be random variables defined on the same sample space. Their joint distribution function (or joint cdf) is defined as

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)=P(X_1\le x_1,\ldots,X_n\le x_n)$$
$$=P(\{\omega\in\Omega:(X_1(\omega),\ldots,X_n(\omega))\in(-\infty,x_1]\times\cdots\times(-\infty,x_n]\}).$$

(What properties should F satisfy?) Note that the cdf is defined for both discrete and continuous random variables.

We say that $X_1,\ldots,X_n$ have joint density f if f is a non-negative function such that

$$\boxed{F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}f(u_1,\ldots,u_n)du_1\cdots du_n.}$$

Likewise, we can find the probability that $(X_1,\ldots,X_n)$ lie in a (measurable) subset of $\mathbb{R}^n$ by integrating the joint density over this set.

Covariance: Let X and Y be random variables on the same sample space. Their covariance is defined as

$$\boxed{\mathrm{Cov}(X,Y) = E[(X-EX)(Y-EY)].}$$

Since EX and EY are numbers, $(X-EX)(Y-EY)$ is a function of (X,Y), i.e., it is also a random variable. Its expectation can be computed as

$$E[(X-EX)(Y-EY)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-EX)(y-EY)f(x,y)dxdy,$$

assuming that (X,Y) have a joint density f.

The covariance can also be expressed as follows:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E[XY - X\cdot(EY) - Y\cdot(EX) + (EX)\cdot(EY)] \\
&= E[XY] - E[X\cdot(EY)] - E[Y\cdot(EX)] + E[(EX)\cdot(EY)] \\
&= E[XY] - (EY)\cdot(EX)] - (EX)\cdot(EY)] + (EX)\cdot(EY) \\
&= \underline{E[XY] - (EX)(EY)}.
\end{aligned}$$

The covariance of any two random variables has the following property:

$$(\mathrm{Cov}(X,Y))^2 \le \mathrm{Var}(X)\mathrm{Var}(Y).$$

This will follow from the fact that, for any two random variables,

$$(E[XY])^2 \le (E[X^2])(E[Y^2]), \tag{1}$$

for which we now give a proof. Note that

$$0 \le E[(X+aY)^2] = E[X^2] + 2aE[XY] + a^2E[Y^2] \tag{2}$$

for all $a \in \mathbb{R}$. The minimum of the RHS is attained at $a = -E[XY]/E[Y^2]$. Substituting this for a above, we get

$$0 \le E[X^2] - 2\frac{(E[XY])^2}{E[Y^2]} + \frac{(E[XY])^2}{E[Y^2]}.$$

Re-arranging this yields (1) provided $E[Y^2] \neq 0$. If $E[Y^2] = 0$ but $E[X^2] \neq 0$, the proof still works with X and Y interchanged. If $E[X^2]$ and $E[Y^2]$ are both zero, then (2) tells us that $2aE[XY] \ge 0$ for all $a \in \mathbb{R}$, which is only possible if $E[XY] = 0$.

Independence: We discussed earlier what it means for events to be independent. What does it mean for two or more random variables defined on the same sample space to be independent? Loosely speaking, random variables $X_1, \ldots, X_n$ are mutually independent if *any* events involving each of them individually are independent. More precisely, they are mutually independent if, for any measurable subsets $B_1, \ldots, B_n$ of $\mathbb{R}$, we have

$$\boxed{P(X_1 \in B_1, \ldots, X_n \in B_n) = \prod_{i=1}^{n} P(X_i \in B_i).}$$

This is not an operationally useful definition (at least for continuous random variables) because it is not feasible to check this equality for all measurable subsets!

We have the following alternative characterisation of independence. Random variables $X_1, \ldots, X_n$ are mutually independent if, and only if,

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n), \quad \forall x_1, \ldots, x_n \in \mathbb{R}.$$

Here, F_{X_i} the marginal distribution of X_i. (Given the joint distribution, how do you compute the marginal distribution?) Equivalently, if the random variables possess a joint density, then they are mutually independent if, and only if,

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n), \quad \forall x_1, \ldots, x_n \in \mathbb{R}.$$

Note that the existence of marginal densities is guaranteed if there is a joint density. (How do you compute the marginal densities from the joint?)

Examples:

1. Suppose U_1 and U_2 are uniform on $[0, 1]$, and independent of each other. Let $X_1 = \min\{U_1, U_2\}$ and $X_2 = \max\{U_1, U_2\}$. Let us compute the joint distribution of (X_1, X_2), denoted F. First, it is clear that $F(x_1, x_2)$ is equal to zero if either x_1 or x_2 is negative, and equal to 1 if both are bigger than 1.

   Let us consider the case where both are between 0 and 1. (You might want to work out the other cases for yourself.) First, if $x_1 > x_2$, it is clear that $F(x_1, x_2) = F(x_2, x_2)$, so it suffices to consider $x_1 \leq x_2$. In that case, we have

$$\begin{aligned} &P(X_1 \leq x_1, X_2 \leq x_2) = P(U_1 \leq x_1, U_2 \leq x_1) \\ &\quad + P(U_1 \leq x_1, x_1 < U_2 \leq x_2) + P(U_2 \leq x_1, x_1 < U_1 \leq x_2) \\ &= x_1^2 + 2x_1(x_2 - x_1) = 2x_1x_2 - x_1^2). \end{aligned}$$

From this, we can calculate the density. On the region $0 \leq x_1 \leq x_2 \leq 1$, the density is given by

$$f(x_1, x_2) = \frac{\partial^2 F(x, y)}{\partial x \partial y} = 2.$$

The density is zero outside this region.

2. Let us return to the example of a dart thrown at a circular dartboard of unit radius, and equally likely to fall anywhere on it. In this case, it is easy to see that the dart's position has density

$$f(x, y) = \frac{1}{\pi} 1(x^2 + y^2 \leq 1).$$

Are the co-ordinates X and Y independent? Can you compute the marginal densities of X and Y?

3. Multivariate normal distribution: The random variables $X_1, \ldots, X_n$ are said to be jointly normally distributed with mean vector $\mu = (\mu_1, \ldots, \mu_n)$ and covariance matrix C if C is a positive definite matrix, and they have the joint density function

$$f(\mathbf{x}) = \frac{1}{\sqrt{2\pi |\det(C)|}} \exp\Big(-\frac{1}{2}(\mathbf{x} - \mu)^T C^{-1} (\mathbf{x} - \mu)\Big),$$

where $\mathbf{x} = (x_1, \ldots, x_n)^T$.

3 Conditional distributions and conditional expectations

Let us go back to the example of rolling two fair dice. Let X and Y denote the number showing on the individual dice, and Z their sum. What do we mean by the conditional distribution of X given Z? Earlier, we defined conditional probability for events. What it means to specify the above is to specify the conditional probability of every event involving X given any event involving Z. How would this work in the above example?

Consider first the event $Z = 2$. Conditioning on this, we have

$$P(X = 1 | Z = 2) = \frac{P(X = 1, Z = 2)}{P(Z = 2)} \stackrel{\cdot}{=} \frac{P(X = 1, Y = 1)}{P(X = 1, Y = 1)} = 1.$$

Also, it is clear that $P(X = j|Z = 2) = 0$ for all $j \in \{2, \ldots, 6\}$. Likewise, conditioning on $Z = 3$, we have $P(X = 1|Z = 3) = P(X = 2|Z = 3) = 1/2$ and $P(X = j|Z = 3) = 0$ for $j \notin \{1, 2\}$. Similarly, we can compute conditional probabilities conditioning on each of the "elementary" events $Z = k$. We shall use the notation $p_{X|Z}(\cdot|k)$ to denote the probability mass function of X conditional on the event $Z = k$. (Recall that a conditional probability is also a probability, hence this is a probability mass function.)

To complete the description, we also have to specify probabilities conditional on events of the form $Z \in \{1, 2\}$, $Z \in \{2, 3, 5\}$ etc. But this is not necessary, because we can compute all such conditional probabilities from the marginal distribution of Z, and the conditional pmf described above, using Bayes' formula. For example,

$$\begin{aligned} P(X = 1|Z \in \{2, 3\}) &= \frac{P(X = 1, Z \in \{2, 3\})}{P(Z \in \{2, 3\})} \\ &= \frac{P(X = 1, Z = 2) + P(X = 1, Z = 3)}{P(Z \in \{2, 3\})} \\ &= \frac{p_Z(2)p_{X|Z}(1|2) + p_Z(3)p_{X|Z}(1|3)}{p_Z(2) + p_Z(3)}. \end{aligned}$$

The last quantity above can be computed given the marginal pmf of Z and the conditional pmf of X conditioned on elementary events for Z.

It is not obvious how to extend this idea to continuous distributions because, if X and Z are continuous random variables, then $P(Z = z)$ will be zero for any z. Hence, we can't compute conditional probabilities conditional on this event. But let's do it heuristically anyway. Suppose (X, Z) have a joint density $f_{X,Z}$ and marginals f_X and f_Z. Thus, for an infinitesimal dz, the probability that Z is in $(z, z+dz)$ is $f(z)dz$. Now, conditional on this, what is the probability that X lies in $(x, x + dx)$. We can compute this using Bayes' formula:

$$\begin{aligned} &P(X \in (x, x + dx)|Z \in (z, z + dz)) = \\ &\quad \frac{P((X, Z) \in (x, x + dx) \times (z, z + dz))}{P(Z \in (z, z + dz))} = \frac{f(x, z)dxdz}{f(z)dz}. \end{aligned}$$

This motivates us to define the conditional density as follows:

$$f_{X|Z}(x|z) = \frac{f_{X,Z}(x, z)}{f_Z(z)}.$$

Note that, for each fixed z, this defines a density function. We can use it to define the conditional cdf

$$F_{X|Z}(x|z) = \int_{-\infty}^{x} f_{X|Z}(u|z)du.$$

Having defined conditional distributions, we can define the property of conditional independence. We say that random variables X and Y are conditionally independent given Z if

$$F_{X,Y|Z}(x, y|z) = F_{X|Z}(x|z)F_{Y|Z}(y|z) \quad \text{for all } x, y, z.$$

Next, we turn to conditional expectations. Their definition is very similar to that of conditional distributions. As usual, we start with the discrete space. Let us go back to the example of the two dice, where X and Y denote the outcomes on the individual dice, and Z their sum. Say we are interested in the expectation of X conditional on Z. As in the case of conditional distributions, it suffices to specify $E[X|Z = z]$ for each possible value of z. Thus, for example,

$$E[X|Z = 3] = 1 \cdot P(X = 1|Z = 3) + 2 \cdot P(X = 2|Z = 3) = \frac{1}{2} + \frac{2}{2} = \frac{3}{2}.$$

In general, in the discrete case,

$$E[X|Z = z] = \sum xP(X = x|Z = z) = \frac{\sum xP(X = x, Z = z)}{P(Z = z)}.$$

This is a number, one for each possible value of Z. We can think of these numbers as describing a function of Z. In other words, $E[X|Z] = g(Z)$, where g is the function defined by $g(z) = E[X|Z = z]$. Thus, $E[X|Z]$ is itself a random variable.

The definition of conditional expectations in the continuous case is analogous. We shall only be interested in cases where the joint (and hence, conditional) densities exist. If that is so, we can define

$$E[X|Y = y] = \int_{-\infty}^{\infty} xf_{X|Y}(x|y)dx = \int_{-\infty}^{\infty} x\frac{f_{X,Y}(x, y)}{f_Y(y)}dx,$$

and think of $E[X|Y]$ as a function of Y, whose value at y is specified by the equation above.

Conditional expectation satisfies the same linearity property as expectation, i.e., $E[aX + bY|Z] = aE[X] + bE[Y]$. The analogue of the second property, namely that the expectation of a constant is constant, is somewhat different. It is that the expectation of any function of Z, conditional on Z, behaves like a constant. In particular, it is conditionally independent of every random variable. In other words,

$$E[Xh(Z)|Z] = E[X|Z]E[h(Z)|Z] = h(Z)E[X|Z],$$

for any random variables X and Z, and any measurable function h.

Conditional expectation satisfies one more property, which doesn't have an analogue for expectations. Recall that $E[X|Z]$ is itself a random variable. What is the expected value of this random variable? It turns out that

$$E[E[X|Z]] = E[X].$$

This is easy to prove, at least in the discrete case. It can also be extended in the form of a chain rule, as follows. Observe that $E[X|Y, Z]$ is a function of (Y, Z) and itself a random variable. If we compute its conditional expectation given Z, then we get another random variable, which is a function of Z. And we have,

$$E[E[E[X|Y, Z]|Z]] = E[X].$$

Chapter 3 : Fundamental results in probability

1 Transformation of random variables

Example: Consider the probability space $\Omega = \{1, \ldots, 6\}$, $\mathcal{F}$ = all subsets of Ω, with probabilities $P(\omega) = 1/6$ for all $\omega \in \Omega$.
(a) On this space, define the random variable $X(\omega) = \omega$. Then the pmf of X is $\{1/6, \ldots, 1/6\}$ on the set $\{1, \ldots, 6\}$. Suppose $Y = X^2$. Then what is the pmf of Y?
(b) On the same space, suppose that X is defined instead as $X(\omega) = \omega - 2$, and that again $Y = X^2$. What are the pmfs of X and Y?

The idea can be extended to continuous random variables, but there is one subtlety involved.

Example: Suppose X is Uniform($[0, 1]$) and $Y = 2X$. What are the cdf and pdf of Y? We first compute the cdf. It is obvious that $F_Y(y) = 0$ for $y < 0$. Also,

$$P(Y \leq y) = P(2X \leq y) = P(X \leq y/2) = y/2 \text{ for } y \in [0, 2).$$

Finally, $F_Y(y) = 1$ for $y \geq 2$. Differentiating the above cdf, we get $f_Y(y) = 1/2$ for $y \in (0, 1)$ and $f_Y(y) = 0$ otherwise.

Could we have guessed this? Intuitively, for an infinitesimal dy,

$$P(Y \in (y, y + dy)) = P(2X \in (y, y + dy)) = P\Big(X \in \Big(\frac{y}{2}, \frac{y}{2} + dy2\Big)\Big),$$

so that

$$f_Y(y)dy = f_X\Big(\frac{y}{2}\Big)\frac{1}{2}dy,$$

which gives the same answer. This intuition can be extended.

Let X be a random variable, g be a differentiable and strictly monotone function, and let $Y = g(X)$. Then, by the same reasoning as above,

$$f_Y(y)dy = f_X(x)dx,$$

where $y = g(x)$. How are dy and dx related? We want $y + dy = g(x + dx)$, so we must have $dy = g'(x)dx$. We are almost there, except that the sign of $g'(x)$ doesn't matter. (It may be the interval $(x - dx, x)$ that gets mapped to $(y, y + dy)$.) So, we have

$$f_Y(y) = f_X(g^{-1}(y))\frac{1}{|g'(g^{-1}(y)|}, \tag{1}$$

where the inverse g^{-1} of the function g is well-defined by the assumption that g is strictly monotone. (The domain of g^{-1} is the range of g.)

What if g isn't monotone? Then the equation $y = g(x)$ may have many solutions for x, and we have to add up the probability contributions from all of them. If there are only countably many solutions, then (1) changes to

$$f_Y(y) = \sum_{x:g(x)=y} f_X(x)\frac{1}{|g'(x)|}. \tag{2}$$

The same idea extends to joint distributions. Suppose $X_1, \ldots, X_n$ are random variables on the same sample space and $(Y_1, \ldots, Y_n) = g(X_1, \ldots, X_n)$ for some differentiable function $g : \mathbb{R}^n \to \mathbb{R}^n$. Then, using boldface to denote vectors,

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{\mathbf{x}:g(\mathbf{x})=\mathbf{y}} f_{\mathbf{X}}(\mathbf{x})\frac{1}{|\det(J_g(\mathbf{x}))|}. \tag{3}$$

Here, $\det(J_g(\mathbf{x}))$ denotes the determinant of the Jacobian matrix

$$J_g(\mathbf{x}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial g_n}{\partial x_1}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial x_n}(\mathbf{x}) & \cdots & \frac{\partial g_n}{\partial x_n}(\mathbf{x}) \end{pmatrix}$$

2 Sums of independent random variables

Example: Suppose X and Y are the numbers obtained by rolling two dice, and suppose $Z = X + Y$. What is $P(Z = 4)$?

If you have written that out in full, then you will see that for arbitrary discrete random variables X and Y taking only integer values, if we define Z as $X + Y$, then

$$P(Z = n) = \sum_{k=-\infty}^{\infty} P(X = k, Y = n - k).$$

If, moreover, X and Y are independent, then we can rewrite this as

$$P(Z = n) = \sum_{k=-\infty}^{\infty} P(X = k)P(Y = n - k). \tag{4}$$

If X and Y are continuous random variables, we get an analogous equation for the density of $Z = X + Y$:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx. \tag{5}$$

The expressions on the RHS of (4) and (5) are called convolutions.

3 Generating functions and characteristic functions

Let X be a discrete random variable. Its generating function G_X is defined as

$$G_X(z) = E[z^X] = \sum_x z^x P(X = x).$$

If X only takes values in $\{0, 1, 2, \dots, \}$, then the above is a power series in z and always converges for all z (real or complex) such that $|z| \leq 1$. The radius of convergence of a power series is defined as the largest value of r such that the power series converges whenever $|z| \leq r$. Thus, for generating functions, the radius of convergence is at least 1, and could be bigger (possibly infinite).

Generating functions have the following properties:

1. $G_X(1) = E[1^X] = 1$.
2. If $|z| < r$, where r is the radius of convergence, then $G'_X(z) = E[Xz^{X-1}]$, $G''_X(z) = E[X(X-1)z^{X-2}]$, and so on. In particular, $G'_X(1) = E[X]$, $G''_X(1) = E[X(X-1)]$ etc., provided that G_X is twice differentiable at 1; this will be the case if the radius of convergence is strictly bigger than one. If not, we need to take a limit as z increases to 1.

3. If X and Y are independent, and $Z = X + Y$, then

   $$G_Z(z) = E[z^Z] = E[z^{X+Y}] = E[z^X z^Y] = E[z^X]E[z^Y] = G_X(z)G_Y(z).$$

   (Which equality in the chain above uses independence?)

There is a closely related function called the moment generating function (mgf), which we'll denote ϕ. It is defined as

$$\phi_X(s) = E[e^{sX}].$$

If X has a density f_X, then

$$\phi_X(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x)dx.$$

(The integral is well-defined for all real s but could take the value $+\infty$.)

We can obtain the properties of mgfs analogous to those of generating functions. In particular,

1. $\phi_X(0) = 1$.
2. If ϕ_X is finite in a neighbourhood of zero, then

   $$\phi_X^{(k)}(0) = E[X^k].$$

3. If X and Y are independent and $Z = X + Y$, then

   $$\phi_Z(s) = \phi_X(s)\phi_Y(s).$$

Finally, characteristic functions are just like generating functions, expect that they are defined on the imaginary axis instead of the real axis. We'll use ψ to denote the characteristic function, defined for a random variable X as $\psi_X(\theta) = E[e^{i\theta X}]$. If X has a density f_X, this implies that

$$\psi_X(\theta) = \int_{-\infty}^{\infty} e^{i\theta x} f_X(x)dx.$$

You might recognise this as the Fourier transform of f_X. It can be inverted to obtain the density of X:

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\theta x} \psi_X(\theta)d\theta.$$

4 Probability inequalities

Markov's inequality: Suppose X is a positive random variable, i.e., $P(X \geq 0) = 1$. Then, for any $a > 0$,

$$P(X > a) \leq \frac{E[X]}{a}.$$

This follows from the fact that

$$X \geq X \cdot 1(X > a) \geq a1(X > a),$$

and so the expectations of these random variables obey the same inequalities. Here $1(X > a)$ denotes the random variable which takes the value 1 on $\{\omega \in \Omega : X(\omega) > a\}$ and takes the value 0 on $\{\omega \in \Omega : X(\omega) \leq a\}$. It is called the indicator of the event $\{X > a\}$. Note that $E[1(X > a)] = P(X > a)$. In general, the expectation of the indicator of an event is the probability of that event.

Chebyshev's inequality: Let X be any random variable. Take Y to be the random variable $Y = (X - E[X])^2$. Then Y is positive and $E[Y] = \mathrm{Var}(X)$. Applying Markov's inequality to Y (and then restating it in terms of X), we get

$$P(|X - E[X]| > a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

Chernoff's inequality: Let X be any random variable and take $Y = e^{\theta X}$, which is positive for all real θ. Applying Markov's inequality to Y yields

$$P(X > a) \leq e^{-\theta a} E[e^{\theta X}] = e^{-\theta a}\phi(\theta) \quad \forall\, \theta \geq 0.$$

Why only for $\theta \geq 0$ and not all real θ? Can you state a corresponding inequality for $P(X < a)$?

5 Laws of large numbers and the central limit theorem

Convergence of random variables Let X and $X_1, X_2, \ldots$ be random variables defined on the same sample space. We say that the sequence X_n converges to X in probability if

$$P(|X_n - X| > \delta) \to 0 \quad \forall \delta > 0. \tag{6}$$

Go back to thinking of random variables as functions on the sample space. We say that the functions X_n converge pointwise to X if $X_n(\omega)$ converges to $X(\omega)$ for all $\omega \in \Omega$. Is convergence in probability the same as pointwise convergence? The answer is no. But there is a notion of convergence which is closely related to pointwise convergence.

We say that the sequence X_n converges to X almost surely (a.s.) if

$$P(\{\omega : X_n(\omega) \to X(\omega)\}) = 1. \tag{7}$$

Almost sure convergence implies convergence in probability but not the other way round.

Suppose now that the random variables $X_1, X_2, \ldots$ are independent and identically distributed (iid), and also that they have finite mean μ. Define $S_n = X_1 + \ldots + X_n$. Then,

$$\frac{S_n}{n} \to \mu \text{ in probability} \qquad \text{(weak law of large numbers)}$$
$$\frac{S_n}{n} \to \mu \text{ almost surely} \qquad \text{(strong law of large numbers)}$$

We now give a proof of the WLLN under the stronger assumption that the X_i have finite variance, denoted σ^2. First observe that

$$\mathrm{Var}\Big(\frac{S_n}{n}\Big) = \frac{1}{n^2}(\mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)) = \frac{\sigma^2}{n}.$$

On the other hand, $E[S_n/n] = \mu$. Hence, by Chebyshev's inequality,

$$P\Big(\Big|\ \frac{S_b}{n} - \mu\ \Big| > \delta\Big) \le \frac{\sigma^2}{n\delta},$$

which tends to zero as n tends to infinity.

Central Limit Theorem: Suppose as before that $X_1, X_2, \ldots$ are iid random variables, and assume that they have both finite mean μ and finite variance σ^2. Define S_n as before, and Z_n as $(S_n - n\mu)/\sigma^2$. Then the sequence of random variables Z_n converges in distribution to a standard normal random variable Z.

I haven't defined convergence in distribution. A formal definition is that, for all bounded continuous functions g, $E[g(Z_n)]$ converges to $E[g(Z)]$. In the context of the CLT, it means that for all intervals (a, b), $P(Z_n \in (a, b))$ converges to $P(Z \in (a, b))$. (If the limiting distribution was not continuous, then we'd have to be careful about points of discontinuity of the cdf. The definition in terms of bounded continuous functions avoids this technicality.)

6 More on convergences

Links between the modes of convergence

Reminder: A sequence of r.v. $(X_n)_{n\geq 1}$ is said to converge to a r.v. X,

(i) in the *almost sure* sense, and we denote $X_n \xrightarrow{a.s.} X$ if
$$P\big(\{\omega \in \Omega\,,\, X_n(\omega) \to X(\omega)\}\big) = 1;$$

(ii) in the *probability* sense, and we denote $X_n \xrightarrow{P} X$ if
$$P\big(|X_n - X| > \varepsilon\}\big) \to 0, \qquad \forall \varepsilon > 0;$$

(iii) in L^p sense ($p > 0$), and we denote $X_n \xrightarrow{L^p} X$ if
$$E\big(|\,X_n - X|^p\big) \to 0;$$

(iv) in the *law* sense or the *distribution* sense, and we denote $X_n \xrightarrow{d} X$ if
$$E\big(h(X_n)\big) \to E\big(h(X)\big),$$
for each continous bounded function h .

We have the following diagramm for the modes of convergence:

$$\begin{array}{ccccc} X_n \xrightarrow{a.s.} X & \Rightarrow & X_n \xrightarrow{\mathbb{P}} X & \Rightarrow & X_n \xrightarrow{d} X \\ & & \Uparrow & & \\ X_n \xrightarrow{L^p} X,\ (p>1) & \Rightarrow & X_n \xrightarrow{L^1} X & & \end{array}$$

Integration to the Limit

Theorem : Monotone Convergence Theorem
If $0 \leq X_n \uparrow X$ *a.s., then* $EX_n \uparrow EX$.

Theorem 1.6.6. Dominated Convergence Theorem
If $X_n \to X$ *a.s.,* $|X_n| \leq Y$ *for all* n, *and* $EY < \infty$, *then* $EX_n \to EX$.

Chapter 4 : **Order statistics**

1 Sample median and population median

The picture is that there is a very large (theoretically infinite) population. Each member of the population has some characteristic quantity X. Consider a number α between zero and one. Then there is supposed to be a number t_α such that the proportion of the population for which the X is less than or equal to t_α is α.

One can think of taking a single random member of the population and measuring this quantity X_1. The assumption is that X_1 is a continuous random variable. Then the cumulative distribution function $F(t) = P[X \leq t]$ is continuous. It follows that there is a t_α such that $F(t_\alpha) = \alpha$.

There are several common examples. The most important is the value such that half the population is above this value and half the population is below this value. Thus when $\alpha = 1/2$ the corresponding $t_{\frac{1}{2}}$ is called the population median m.

Similarly, when $\alpha = 1/4$ the $t_{\frac{1}{4}}$ is called the lower population quartile. In the same way, when $\alpha = 3/4$ the $t_{\frac{3}{4}}$ is called the upper population quartile. In statistics the function F characterizing the population is unknown. Therefore all these t_α are unknown quantities associated with the population.

Now consider the experiment of taking a random sample of size n and measuring the corresponding quantities $X_1, \ldots, X_n$. Thus again we have independent random variables all with the same distribution. We are assuming that the distribution is continuous. Thus the probability is one that for all $i \neq j$ the quantities $X_i \neq X_j$ are unequal.

The order statistics $X_{(1)}, \ldots, X_{(n)}$ are the quantities obtained by arranging the random variables $X_1, \ldots, X_n$ in increasing order. Thus by definition

$$X_{(1)} < X_{(2)} < \cdots < X_{(i)} < \cdots < X_{(n-1)} < X_{(n)}.$$

The joint density of $X_1, \ldots, X_n$ is $f(x_1) \cdots f(x_n)$. This product structure is equivalence to the independence of the random variables. On the other hand, the joint density of the order statistics $X_{(1)}, \ldots, X_{(n)}$ is $n! f(x_1) \cdots f(x_n)$ for $x_1 < x_2 < \cdots < x_n$ and zero otherwise. There is no way to factor this. The order statistics are far from independent.

The order statistics are quite useful for estimation. Take $\alpha = i/(n+1)$. Then it seems reasonable to use the order statistics $X_{(i)}$ to estimate t_α.

Thus, for instance, if n is odd and $i = (n+1)/2$ and $\alpha = 1/2$, then $X_{(i)}$ is the sample median. This estimates the population median $m = t_{\frac{1}{2}}$.

The fundamental theorem on order statistics is the following. It shows that questions about order statistics reduce to questions about binomial random variables.

Theorem 5.1 *Let $X_1, \ldots, X_n$ be independent random variables with a common continuous distribution. Let $X_{(1)}, \ldots, X_{(n)}$ be their order statistics. For each x, let $N_n(x)$ be the number of i such that $X_i \le x$. Then $N_n(x)$ is a binomial random variable with parameters n and $F(x)$. Furthermore,*

$$P[X_{(j)} \le x] = P[N_n(x) \ge j].$$

This result can be stated even more explicitly in terms of the binomial probabilities. In this form it says that if $P[X_i \le x] = F(x)$, then

$$P[X_{(j)} \le x] = \sum_{k=j}^{n} \binom{n}{k} F(x)^k (1 - F(x))^{n-k}.$$

This theorem is remarkable, in that it gives a rather complete description of order statistics for large sample sizes. This is because one can use the central limit theorem for the corresponding binomial random variables.

Theorem 5.2 *Let $X_1, \ldots, X_n$ be independent random variables with a common continuous distribution. Let $X_{(1)}, \ldots, X_{(n)}$ be their order statistics. Fix α and let $F(t_\alpha) = \alpha$. Let $n \to \infty$ and let $j \to \infty$ so that $\sqrt{n}(j/n - \alpha) \to 0$. Then the order statistics $X_{(j)}$ is approximately normally distributed with mean*

$$E[X_{(j)}] \approx t_\alpha$$

and standard deviation

$$\sigma_{X_{(j)}} \approx \frac{\sqrt{\alpha(1-\alpha)}}{f(t_\alpha)\sqrt{n}}.$$