7. Continuous random variables (RVs with densities)

All models are wrong but some are useful.
George Box

All things flow.
Heraclitus

7.1. Density and distribution functions

Discrete random variables take only a countable set of values. But there are many important questions in which we must consider random variables not subject to such a restriction. This means that we need a sample space that is not countable. Technical questions of “measurability” then arise which cannot be treated satisfactorily without more advanced mathematics. This kind of difficulty stems from the impossibility of assigning a probability to every subset of the sample space when it is uncountable. The matter is resolved by confining ourselves to sample sets belonging to an adequate class called a σ-algebra (also sigma-algebra, σ-field, sigma-field); see Appendix 4. Here we will take up a particular but very important situation that covers most applications and requires little mathematical abstraction. This is the case of a RV with a “density” (continuous RV).

Consider a function $f(\cdot)$ defined on $R^1 = (-\infty, +\infty)$ and satisfying two conditions:

$$\text{(i) } \forall\, u\colon f(u) \geq 0;$$
$$\text{(ii) } \int_{-\infty}^{\infty} f(u)du = 1. \qquad (7.1)$$

Such a function is called a **probability density function (PDF)** or shortly **density function** on R^1. The integral in (ii) is the Riemann integral taught in calculus. You may recall that if $f(\cdot)$ is continuous or just piecewise continuous, then the definite integral $\int_a^b f(u)\, du$ exists for any interval $[a, b]$. But in order that the integral over the infinite range $(-\infty, +\infty)$ should exist, further conditions are needed to make sure that $f(u)$ is pretty small for large $|u|$. In general, such a function is said to be “integrable over R^1”. The requirement that the total integral be equal to 1 is less serious than it might appear, because if $\int_{-\infty}^{\infty} f(u)du = M < \infty$, we can just divide through by M and use $f(\cdot)/M$ instead of $f(\cdot)$.

Density functions can be a great variety. The only constraints are that the curve should not lie below the x-axis anywhere, and the area under the curve should be equal to 1.

We can now define a class of random variables on a general sample space as follows: X is a function on Ω: $\omega \to X(\omega) \in R^1$, but its probabilities are prescribed by means of a density function so that for any interval $[a, b]$ we have

$$P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b) = \int_a^b f(u)\,du\,. \tag{7.2}$$

More generally, if A is the union of intervals not necessarily disjoint and some of which may be infinite, we have

$$P(X \in A) = \int_A f(u)\,du. \tag{7.3}$$

Such a random variable is said to have a density, and its density function is $f(\cdot)$.

If A is a finite union of intervals, then it can be split up into disjoint ones, some of which may abut on each other, such as $A = \bigcup_{j=1}^{k} [a_j, b_j]$, and then the right-hand side of (7.3) may be written as $\sum_{j=1}^{k} \int_{a_j}^{b_j} f(u)\,du$. This is a property of integrals which is geometrically obvious when you consider them as areas.

Next if $A = (-\infty, x]$, then we can write

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(u)du\,. \tag{7.4}$$

Such functions $F(\cdot)$ are called **absolutely continuous.**

Formula (7.4) defines the **distribution function** $F(\cdot)$ of X. Sometimes $F(\cdot)$ is called **cumulative distribution function (CDF).** It follows from the fundamental theorem of calculus that if $f(\cdot)$ is continuous, then $f(\cdot)$ is the derivative of $F(\cdot)$:

$$F'(x) = f(x). \tag{7.5}$$

Thus in this case the two functions $f(\cdot)$ and $F(\cdot)$ mutually determine each other.

If $f(\cdot)$ is not continuous everywhere, (7.5) is still true for every x at which $f(\cdot)$ is continuous. These things are proved in calculus.

It has to be said that for many purposes, the detailed information about what exactly the outcome space Ω is where $X \in \Omega$ is defined is actually irrelevant. For example, normal RVs arise in a great variety of models in statistics, but what matters is that they are jointly or individually Gaussian, i.e. have a prescribed PDF. Also, an exponential RV arises in many models and may be associated with a lifetime of an item or a time between subsequent changes of a state in a system, or in a purely geometric context. It is essential to be able to think of such RVs without referring to a particular Ω.

Let us observe that in the definition above of a random variable with a density, it is implied that the sets $\{a \le X \le b\}$ and $\{X \in A\}$ have probabilities assigned to them; in fact, they are specified in (7.2) and (7.3) by means of the density function (for more details see Appendix 4). If CDF has the form (7.5) one says that the corresponding RV has an absolutely continuous distribution (with a PDF $f(\cdot)$).

For the discrete distributions, the CDF is locally constant, with positive jumps at the points of a discrete set.

So far we have encountered two types of RVs: either

(i) with a discrete set of values (finite or countable) or

(ii) with a PDF (on a subset of R^1).

These types do not exhaust all occurring situations. In particular, a number of applications require consideration of an RV X that represents a "mixture" of the two above types where a positive portion of a probability mass is sitting at a point (or points) and another portion is spread out with a PDF over an interval in R^1. Then the corresponding CDF F_X has jumps at the points x_j where probability $P(X = x_j) > 0$, of a size equal to the probability, and is absolutely continuous outside these points.

7.2. Continuous as compared to discrete RVs

Rather, let us remark on the close resemblance between the formulas above and the corresponding ones for discrete RVs. This will be amplified by a definition of mathematical expectation in the present case and listed below for comparison.

	Countable case	Density case
Range	x_n, $n = 1,2,\dots$	$-\infty < u < \infty$
Element of probability	p_n	$f(u)\,du$
$P(a \le X \le b)$	$\sum_{a \le x_n \le b} p_n$	$\int_a^b f(u)du$
$P(X \le x)$	$\sum_{x_n \le x} x_n p_n$	$\int_{-\infty}^x f(u)du$
$E(X)$	$\sum_n x_n p_n$	$\int_{-\infty}^{\infty} f(u)udu$
proviso	$\sum_n \lvert x_n \rvert p_n < \infty$	$\int_{-\infty}^{\infty} f(u)\lvert u\rvert du < \infty$

More generally,

$$E(\phi(X)) = \int_{-\infty}^{\infty} \phi(u) f(u) du. \tag{7.6}$$

The expectation $E(X)$ of a continuous random variable X has the same useful basic properties as in the discrete case:

let X be a random variable with finite mean $E(X)$, a and b be constants, and let g and h be functions, then:

(i) $E(aX + b) = aE(X) + b$;

(ii) and generally if $g(X)$ and $h(X)$ have finite mean, then

$$E(ag(X) + bh(X)) = aE(g(X)) + bE(h(X));$$

(iii) if $P(a \leq X \leq b) = 1$, then $a \leq E(X) \leq b$.

Prove the properties (first, don't forget to establish the necessary absolute convergence).

Further insight into the analogy is gained by looking at the following picture:

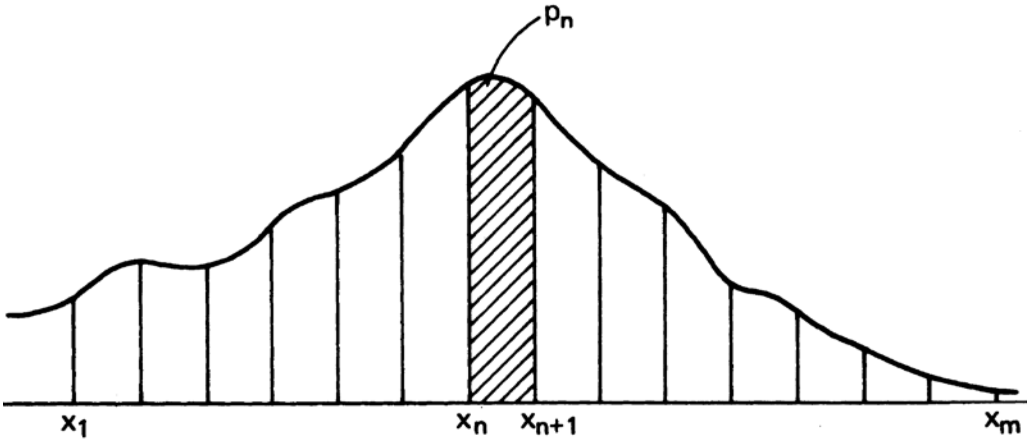


Figure 7.1

The curve is the graph of a density function $f(\cdot)$. We have divided the x axis into $m + 1$ pieces, not necessarily equal and not necessarily small, and denote the area under the curve between x_n and x_{n+1} by p_n, thus: $p_n = \int_{x_n}^{x_{n+1}} f(u)\, du$, $0 \leq n \leq m,$ where $x_0 = -\infty$, $x_{m+1} = +\infty$. It is clear that we have $\forall n$: $p_n \geq 0$ and $\sum_{n=0}^{\infty} p_n = 1$.

Hence the numbers p_n satisfy the normalized conditions. Instead of a finite partition we may have a countable one by suitable labeling such as . . . , p_{-2}, p_{-1}, p_0, p_1, ... Thus we can derive a set of "elementary probabilities" from a density function in infinitely many ways. This process may be called **discretization**. If X has the density $f(\cdot)$, we may consider a random variable Y such that $P(Y = x_n) = p_n$, where we may replace x_n by any other number in the subinterval $[x_n, x_{n+1})$.

Now if $f(\cdot)$ is continuous and the partition is sufficiently fine, namely if the pieces are sufficiently small, then it is geometrically evident that Y is in some sense a discrete approximation of X. For instance, $E(Y) = \sum_n p_n x_n$ will be an approximation of $E(X) = \int_{-\infty}^{\infty} u f(u) du$. Remember the Riemann sums defined in calculus lead to a Riemann integral? There the strips with curved tops in Figure 7.1 are replaced by flat tops (rectangles).

From a practical point of view, it is the discrete approximations that can really be measured, whereas the continuous density is only a mathematical idealization.

Having dwelled on the similarity of the two cases of random variable, we will pause to stress a fundamental difference between them. If X has a density, then by (7.2) with $a = b = x$, we have

$$P(X = x) = \int_x^x f(u)\, du = 0. \tag{7.7}$$

Geometrically speaking, this merely states the trivial fact that a line segment has zero area. Since x is arbitrary in (7.7), it follows that X takes any preassigned value with probability zero. This is in direct contrast to a random variable taking a countable set of values, for then it must take some of these values with positive probability. It seems paradoxical that on the one hand, $X(\omega)$ must be some number for every ω, and on the other hand any given number has probability zero.

The following simple example should clarify this point.

Example 7.1. Spin a needle on a circular dial. When it stops it points at a random angle θ (measured from the horizontal, say). Under normal conditions it is reasonable to suppose that θ is uniformly distributed between 0◦and 360 degrees. This means it has the following density function:

$$f(u) = 1/360 \text{ for } 0 \le u \le 360, \text{ and } f(u) = 0 \text{ otherwise.}$$

Thus for any $\theta_1 < \theta_2$ we have

$$P(\theta_1 \le \theta \le \theta_2) = \int_{\theta_1}^{\theta_2} 1/360\, du = (\theta_2 - \theta_1)/360. \tag{7.8}$$

This formula says that the probability of the needle pointing between any two directions is proportional to the angle between them. If the angle $\theta_2 - \theta_1$ shrinks to zero, then so does the probability. Hence in the limit the probability of the needle pointing exactly at θ is equal to zero. From an empirical point of view, this event does not really make sense because the needle itself must have a width. So in the end it is the mathematical fiction or idealization of a “line without width” that is the root of the paradox.

There is a deeper way of looking at this situation which is very rich. It should be clear that instead of spinning a needle we may just as well “pick a number at random” from the interval [0, 1]. This can be done by bending the circle into a line segment and changing the unit. Now every point in [0, 1] can be represented by a decimal such as

$$0.141592653589793\ldots. \tag{7.9}$$

There is no real difference if the decimal terminates because then we just have all digits equal to 0 from a certain place on, and 0 is no different from any other digit. Thus, to pick a number in [0, 1] amounts to picking all its decimal digits one after another. Now the chance of picking any prescribed digit, say the first digit “1” above, is equal to 1/10 and the successive pickings are totally independent trials. Hence the chance of picking the 15 digits shown in (7.9) is equal to 0.1^{15}.

If we remember that 10^9 is 1 billion, this probability is already so small that according to Emile Borel (1871–1956; great French mathematician and one of the founders of modern probability theory), it is terrestrially negligible and should be equated to zero! But we have only gone 15 digits in the decimals of the number $\pi - 3$, so there can be no question whatsoever of picking this number itself. So here again we are up against a mathematical fiction − the real number system.

We may generalize example 1 as follows. Let $[a, b]$ be any finite, nondegenerate interval in R^1 and put

$$f(u) = 1/(b - a) \text{ for } a \le u \le b, f(u) = 0 \text{ otherwise.}$$

This is a density function, and the corresponding distribution is called the **uniform distribution** on $[a, b]$.

7.3. Bertrand's paradox

A chord is drawn at random in a circle. What is the probability that its length exceeds that of a side of an inscribed equilateral triangle?

Let us draw such a triangle in a circle with center 0 and radius R, and make the following observations. The side is at distance $R/2$ from 0; its midpoint is on a concentric circle of radius $R/2$; it subtends an angle of 120 degrees at 0. You ought to know how to compute the length of the side, but this will not be needed. Let us denote by A the desired event that a random chord be longer than that side. Now the length of any chord is determined by any one of the three quantities: its distance d from 0; the location of its midpoint M; the angle θ it subtends at 0. See Figure 7.2. We are going to assume in turn that each of these has a uniform distribution over its range and compute the probability of A under each assumption.

(1) Suppose that d is uniformly distributed in $[0, R]$. This is a plausible assumption if we move a ruler parallel to itself with constant speed from a tangential position toward the center, stopping somewhere to intersect the circle in a chord. It is geometrically obvious that the event A will occur if and only if (shortly iff) $d < R/2$. Hence $P(A) = 1/2$.

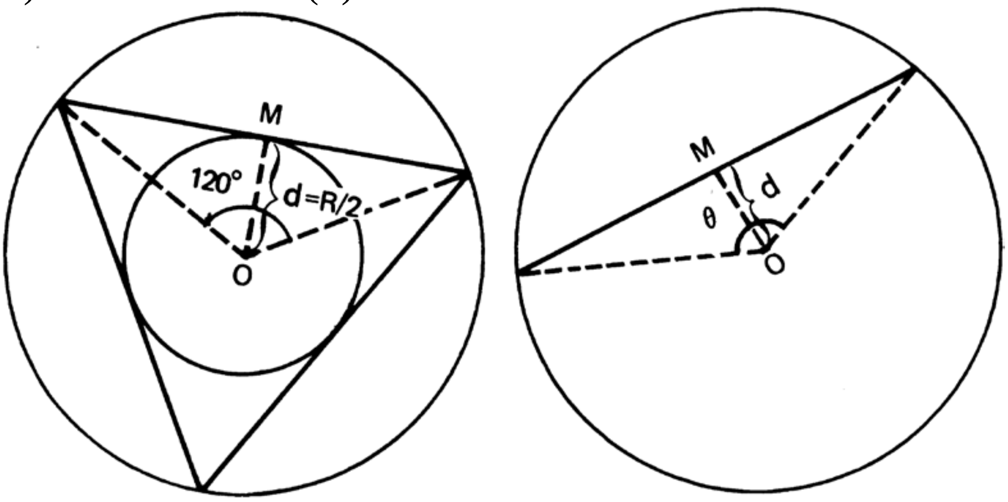


Figure 7.2

(2) Suppose that M is uniformly distributed over the disk D formed by the given circle. This is a plausible assumption if a tiny dart is thrown at D and a chord is then drawn perpendicular to the line joining the hitting point to 0. Let D' denote the concentric disk of radius $R/2$. Then the event A will occur iff M falls within D'.

Hence $P(A) = P(M \in D') =$ (area of D')/(area of D) = 1/4.

(3) Suppose that θ is uniformly distributed between 0 and 360 degrees. This is plausible if one endpoint of the chord is arbitrarily fixed and the other is obtained by rotating a radius at constant speed to stop somewhere on the circle. Then it is clear from the picture (Figure 7.3) that A will occur iff θ is between 120 and 240 degrees. Hence $P(A) = (240-120)/360 = 1/3$.

Thus the answer to the problem is 1/2, 1/4, or 1/3 according to the different hypotheses made. It follows that these hypotheses are not compatible with one another. Other hypotheses are possible and may lead to still other answers. This problem was known as **Bertrand's paradox** in the earlier days of discussions of probability theory.

But of course the paradox is due only to the fact that the problem is not well posed without specifying the underlying nature of the randomness.

It is not surprising that the different ways of randomization should yield different probabilities, which can be verified experimentally by the mechanical procedures described.

Here is a facile analogy. Suppose that you are asked how long it takes to go from your dormitory to the classroom without specifying whether we are talking about "walking," "biking," or "driving" time. Would you call it paradoxical that there are different answers to the question?

7.4. Exponential distribution

Suppose you station yourself at a spot on a relatively serene country road and watch the cars that pass by that spot. With your stopwatch you can clock the time before the first car passes. This is a random variable T called the **waiting time**. Under certain circumstances it is a reasonable hypothesis that T has the density function below with a certain $\lambda > 0$:

$$f(u) = \lambda e^{-\lambda u}, \; u \geq 0. \tag{7.10}$$

It goes without saying that $f(u) = 0$ for $u < 0$. We see that $f(\cdot)$ satisfies the conditions in (7.1), so it is indeed a density function.

The corresponding distribution function is obtained by integrating $f(\cdot)$ as in (7.4):

$$P(T \leq x) = F(x) = \int_{-\infty}^{x} \lambda e^{-\lambda u} du = 1 - e^{-\lambda x}. \tag{7.11}$$

The distribution defined by the PDF and the CDF is called the **exponential distribution** with parameter λ.

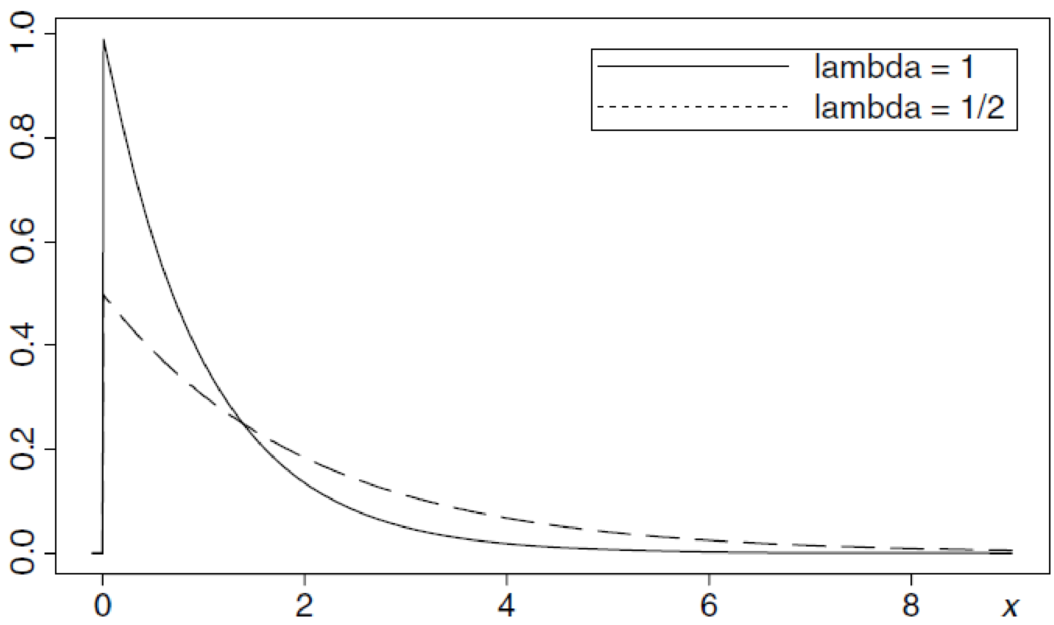


Figure 7.3. The exponential distribution with parameter λ (lambda)

In this case it is often more convenient to use the so called **tail probability**:

$$P(T > x) = 1 - F(x) = e^{-\lambda x}. \qquad (7.12)$$

For every given x, say 5 (seconds), the probability $e^{-5\lambda}$ in (7.12) decreases as λ increases. This means that your waiting time tends to be shorter if λ is larger. On a busy highway λ will indeed be large. The expected waiting time is given by

$$E(T) = \int_0^\infty u\lambda e^{-\lambda u} du = 1/\lambda. \qquad (7.13)$$

This result supports our preceding observation that T tends on the average to be smaller when λ is larger.

The exponential distribution is a very useful model for various types of waiting time problems such as telephone calls, service times, splitting of radioactive particles, etc. It also plays a central role in reliability, where the exponential distribution is one of the most important failure laws. In reliability studies, the time to failure for a physical component or a system is expected to be exponentially distributed if the unit fails as soon as some single event, such as malfunction of a component, occurs, assuming such events happen independently.

Exponential distribution has a strong connection with Poisson distribution.

Let random variable $X(0, t)$ be the number of arrivals in the time interval $[0, t]$ and assume that it is Poisson distributed. Our interest now is in the time between two successive arrivals, which is, of course, also a random variable. Let this interarrival time be denoted by T. Its probability distribution function, $F_T(t)$, is, by definition, $F_T(t) = P(T \le t) = 1 - P(T > t)$ for $t \ge 0$, $F_T(t) = =0$ elsewhere.

In terms of $X(0, t)$, the event $T > t$ is equivalent to the event that there are no arrivals during time interval $[0, t]$, or $X(0, t) = 0$. Hence, since $P(X(0, t) = 0) = e^{-\lambda x}$, we have

$$F_T(t) = 1 - e^{-\lambda x}.$$

Comparing this expression with (7.11), we can establish the result that the interarrival time between Poisson arrivals has an exponential distribution; the parameter λ in the distribution of T is the mean arrival rate associated with Poisson arrivals.

7.5. Multivariate continuous distribution. Independence

The random vector (X, Y) is said to have a joint density function $f(\cdot,\cdot)$ in case

$$P(X \le x,\ Y \le y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\, du\, dv\,. \qquad (7.14)$$

for all (x, y). It then follows that for any "reasonable" subset S of the Cartesian plane (called a **Borel set**, see Appendix 5 for more detail), we have

$$P((X, Y) \in S) = \iint_S f(u, v)\, du\, dv. \qquad (7.15)$$

For example, S may be polygons, disks, ellipses, and unions of such shapes.

Note that (7.15) contains (7.14) as a very particular case and we can, at a pinch, accept the more comprehensive condition (7.15) as the definition of f as density for (X, Y). However, here is a heuristic argument from (7.14) to (7.15). Let us denote by $R(x, y)$ the infinite rectangle in the plane with sides parallel to the coordinate axes and lying to the southwest of the point (x, y). The picture below shows that for any $\delta > 0$ and $\delta' > 0$:

$R(x + \delta, y + \delta') \setminus R(x + \delta, y) \setminus R(x, y + \delta')$ is the shaded rectangle

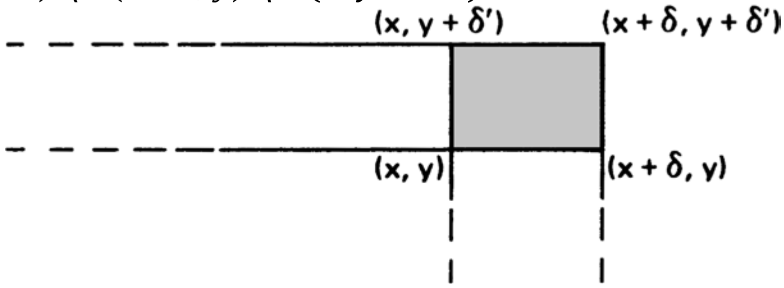


Figure 7.4

It follows that if we manipulate the relation (7.14) in the same way, we get

$$P(x \le X \le x + \delta,\ y \le Y \le y + \delta') = \int_{x}^{x+\delta} \int_{y}^{y+\delta'} f(u, v)\, du\, dv\,.$$

This means (7.15) is true for the shaded rectangle. By varying x, y as well as δ, δ', we see that the formula is true for any rectangle of this shape. Now any reasonable figure can be approximated from inside and outside by a number of such small rectangles (even just squares) − a fact known already to the ancient Greeks. Hence in the limit we can get (7.15) as asserted.

The joint density function $f(\cdot,\cdot)$ satisfies the following conditions:

(i) $f(u, v) \geq 0$ for all (u, v);

(ii) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u,v)\, du\, dv = 1$.

Of course, (ii) implies that $f(\cdot,\cdot)$ is integrable over the whole plane. Frequently we also assume that $f(\cdot,\cdot)$ is continuous. Now the formulas analogous to discrete case (marginal distributions) are

$$P(X \leq x) = \int_{-\infty}^{x} f(u,*)\, du, \text{ where } f(u,*) = \int_{-\infty}^{\infty} f(u,v) dv,$$

$$P(Y \leq y) = \int_{-\infty}^{y} f(*,v)\, dv, \text{ where } f(*, v) = \int_{-\infty}^{\infty} f(u,v) du. \quad (7.16)$$

The functions $f(u, *)$ and $f(*, v)$ are respectively called the **marginal density functions** of X and Y. They are derived from the joint density function after "integrating out" the variable that is not in question.

For any "reasonable" (i.e. Borel) function ϕ:

$$E(\phi(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi(u,v) f(u,v)\, du\, dv. \quad (7.17)$$

The class of reasonable functions includes all bounded continuous functions in (u, v), indicators of reasonable sets, and functions that are continuous except across some smooth boundaries, for which the integral above exists, etc.

In the most general case the **joint distribution function** $F(\cdot,\cdot)$ of (X, Y) is defined by

$$F(x, y) = P(X \leq x, Y \leq y) \text{ for all } (x, y). \quad (7.18)$$

If we denote $\lim_{y\to\infty} F(x,y)$ by $F(x, \infty)$, we have $F(x, \infty) = P(X \leq x, Y < \infty) = P(X \leq x) = F_X(x)$ since "$Y < \infty$" puts no restriction on Y. Thus $F(x, \infty)$ is the **marginal distribution function** of X. The marginal distribution function of Y $F_Y(y)$ is similarly defined.

Independence is an extremely important property; its definition is by now familiar.

Jointly distributed random variables are **independent** if, for all x and y,

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y).$$

In terms of distributions, this is equivalent to the statement that

$$F(x, y) = F_X(x)F_Y(y). \quad (7.19)$$

For random variables with a density, it follows immediately by differentiating that

$$f(x, y) = f_X(x)\, f_Y(y)$$

if X and Y are independent. The converse is obviously true (If X and Y have density $f(x, y)$, and for all x and y it is true that $f(x, y) = f_X(x)\, f_Y(y)$, then X and Y are independent).

If $C = ((x, y)\colon x \in A, y \in B)$ and X and Y are independent, then

$$\iint_C f(x,y)dxdy = \int_A f_X(x)dx \int_B f_Y(y)dy, \quad (7.20)$$

assuming of course that the integrals exist.

Finally, if the random variables U and V satisfy $U = g(X)$, $V = h(Y)$, and X and Y are independent, then U and V are independent too. To see this, just let

$A = (x\colon g(x) \le u)$ and $B = (g\colon h(y) \le v)$, and the independence follows from (7.19) and (7.20).

7.6. Transformation of random variables

We have interpreted the random vector (X, Y) as a random point Q picked in R^2 according to some density $f(x, y)$, where (x, y) are the Cartesian coordinates of Q. Of course, the choice of coordinate system is arbitrary; we may for some very good reasons choose to represent Q in another system of coordinates (u, v), where (x, y) and (u, v) are related by $u = u(x, y)$ and $v = v(x, y)$. What now is the joint density of $U = u(X, Y)$ and $V = v(X, Y)$?

Equally, given a pair of random variables X and Y, our real interest may well lie in some function or functions of X and Y. What is their (joint) distribution?

At a symbolic or formal level, the answer is straightforward.

For U and V above, and $A = \{x, y\colon u(x, y) \le w, v(x, y) \le z\}$,

$$F_{U,V}(w, z) = \int_A f_{X,Y}(x, y)\, dx\, dy.$$

The problem is to turn this into a more tractable form.

Fortunately, there are well-known results about changing variables within a multiple integral that provide the answer. We state without proof a theorem for a transformation T satisfying the following conditions. Let C and D be subsets of R^2. Suppose that T given by $T(x, y) = (u(x, y), v(x, y))$ maps C one–one onto D, with inverse T^{-1} given by $T^{-1}(u, v) = (x(u, v), y(u, v))$, which maps D one–one onto C.

We define the so-called Jacobian J as

$$J(u, v) = \begin{vmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial y}{\partial u}\frac{\partial x}{\partial v},$$

where the derivatives are required to exist and be continuous in D. Then we have the following result.

Theorem 7.1. *Let X and Y have density $f(x, y)$, which is zero outside C. Then $U = u(X, Y)$ and $V = v(X, Y)$ have joint density*

$$f_{U.V}(u, v) = f_{X,Y}(x(u, v), y(u, v))|J(u, v)| \text{ for } (u, v) \in D. \qquad (7.21)$$

It follows from (7.21):

Functions (one RV): If continuous random variables X and Y are such that $Y = g(X)$ for some function $g(\cdot)$ that is differentiable and strictly increasing, then

$$f_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}[g^{-1}(y)], \qquad (7.22)$$

where $g^{-1}(\cdot)$ is the inverse function of $g(\cdot)$ (see Example 11 about inverse function).

In general, we can write $f_Y(y) = \frac{d}{dy}\int_{\{x: g(x) \le y\}} f_X(x)dx$.

Example 7.2. Suppose in a problem involving the random variable T above (interarrival time, see section 7.4), what we really want to measure is its logarithm: $S = \ln T$.

This is also a random variable; it is negative if $T < 1$, zero if $T = 1$, and positive if $T > 1$. What are its probabilities? We may be interested in $P(a \le S \le b)$, but it is clear that we need only find $P(S \le x)$, namely the distribution function $F_S(\cdot)$ of S. The function $\ln x$ is monotone and its inverse is e^x so that $S \le x \Leftrightarrow \ln T \le x \Leftrightarrow T \le e^x$.

Hence by (7.11) $F_S(x) = P\{S \le x\} = P\{T \le e^x\} = 1 - e^{-\lambda e^x}$.

The density function f_S is obtained by differentiating:
$f_S(x) = F_S'(x) = \lambda e^x e^{-\lambda e^x} = \lambda e^{x-\lambda e^x}$.

Consider among the examples some wildly used PDFs and their numerical characteristics.

7.7. Examples

Example 1. Let X be uniformly distributed on (0, 1) with density $f(x) = 1$ if $0 < x < 1$; $f(x) = 0$ otherwise.

If $Y = -\lambda^{-1} \ln X$, where $\lambda > 0$, what is the density of Y?

Solution. First, we seek the distribution of Y: $F_Y(y) = P(-\lambda^{-1} \ln X \le y) = P(\ln X \ge -\lambda y) = P(X \ge e^{-\lambda y}) = 1 - e^{-\lambda y}$ for $y \ge 0$; $F_Y(y) = 0$ otherwise. Hence, the derivative exists except at $y = 0$, and $f_Y(y) = \lambda e^{-\lambda y}$ if $y > 0$; $f_Y(y) = 0$ if $y < 0$. This is the exponential density with parameter λ.

Example 2. Let X have the **standard normal distribution** with density

$$f(x) = (2\pi)^{-1/2}\exp(-x^2/2). \qquad (7.23)$$

Find the density of $Y = \sigma X + \mu$ for given constants μ and $\sigma \ne 0$. Also, find the density of $Z = X^2$.

Solution. $P(\sigma X + \mu \le y) = P(\sigma X \le y - \mu) =$

$$\begin{cases} P\left(X \le \dfrac{y-\mu}{\sigma}\right), \text{if } \sigma > 0; \\ P(X \ge (y-\mu)/\sigma), \text{if } \sigma < 0 \end{cases} = \begin{cases} F_X\left(\dfrac{y-\mu}{\sigma}\right), \text{if } \sigma > 0; \\ 1 - F_X((y-\mu)/\sigma), \text{if } \sigma < 0. \end{cases}$$

Hence, differentiating with respect to y,

$$f_Y(y) = f_X(y-\mu)/\sigma)/|\sigma| = (2\pi\sigma^2)^{-1/2}\exp(-(y-\mu)^2/(2\sigma^2)). \qquad (7.24)$$

Normal densities play the important role in probability theory and statistics due to central limit theorem (see part 8).

Some care is required if transformation function is not one–one as in the second case:

$$P(X^2 \le z) = P(X \le \sqrt{z}) - P(X \le -\sqrt{z}) = F_X(\sqrt{z}) - F_X(-\sqrt{z}).$$

Differentiating now gives

$$f_Z(z) = 1/(2\sqrt{z}) f_X(\sqrt{z}) + 1/(2\sqrt{z}) f_X(-\sqrt{z}) = (1/\sqrt{2\pi z})\exp(-z/2). \qquad (7.25)$$

Remark 7.1. The density given by (7.24) is known as the **normal density** with parameters μ and σ^2, sometimes denoted by N(μ, σ^2), see Figure 7.5. The standard normal density is N(0, 1) (see (7.23)).

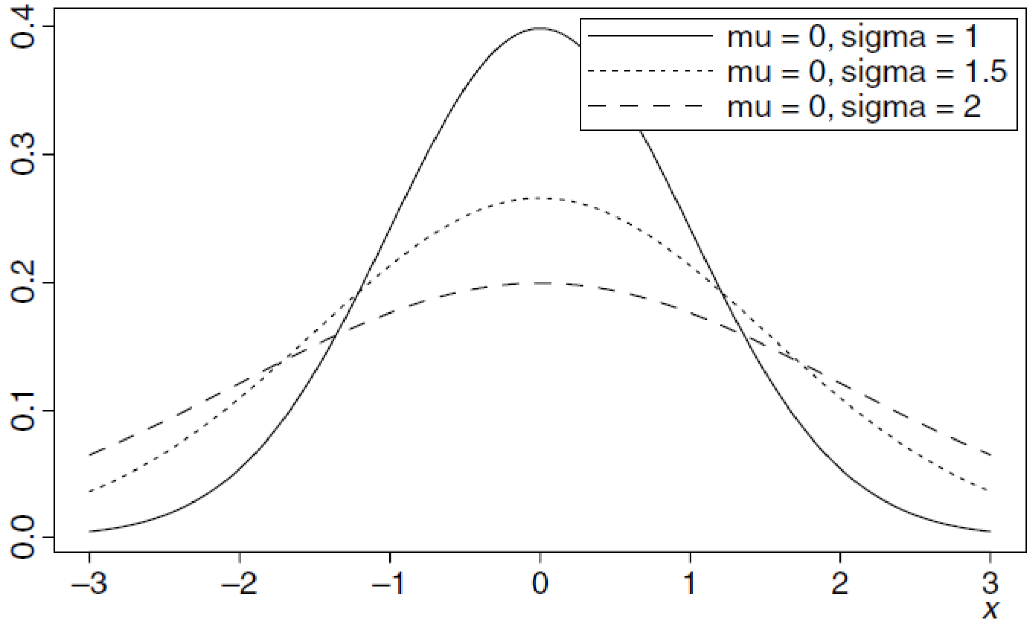


Figure 7.5. Normal density with parameters μ (mu) and σ^2 (sigma squared)

The density given by (7.25) is the **gamma density** with parameters 1/2 and 1/2. This is known as the **chi-squared density** with parameter 1, sometimes denoted by $\chi^2{}_1$.

Example 3. Let X be uniformly distributed on [−1, 1]. Find the density of $Y = X^r$ for nonnegative integers r.

Solution. First, note that X has distribution function $F(x) = (1 + x)/2$ for $-1 \le x \le 1$.

Now, if r is odd, then the function $g(x) = x^r$ maps the interval [−1, 1] onto itself in one–one correspondence. Hence, routinely:

$P(Y \le y) = P(X^r \le y) = P(X \le y^{1/r}) = (1 + y^{1/r})/2$ for $-1 \le y \le 1$, and Y has density $f(y) = y^{1/r-1}/(2r), -1 \le y \le 1$.

If r is even, then $g(x) = x^r$ takes values in [0, 1] for $x \in [-1, 1]$. Therefore, $P(Y \le y) = P(0 \le X^r \le y) = P(-y^{1/r} \le X \le y^{1/r}) = y^{1/r}$ for $0 \le y \le 1$. Hence, Y has density $f(y) = y^{1/r-1}/r, 0 \le y \le 1$.

Finally, if $r = 0$, then $X^r = 1$, $F_Y(y)$ is not continuous (having a jump from 0 to 1 at $y = 1$) and so Y does not have a density in this case. Obviously, Y is discrete, with $P(Y = 1) = 1$.

Example 4 (Expectation of uniform density). Let X be uniformly distributed on (a, b). Then

$$E(X) = \int_a^b \frac{x}{b-a}\,dx = (b + a)/2.$$

Example 5 (Expectation of normal density). Let X have the N(μ, σ^2) density. Then

$$E(X) = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} \text{y}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)dy =$$
$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \frac{y-\mu}{\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)dy +$$
$$+\frac{\mu}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)dy =$$
$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \text{y}\exp(-y^2)\,dy + \frac{\mu}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp(-y^2)\,dy.$$

on making the substitution $u = (v - \mu)/\sigma$ in both integrands. The first integrand is an odd function, so the integral over R^1 is zero. The second term is μ (why?). Hence, $E(X) = \mu$.

Prove that $\text{var}(X) = \sigma^2$.

Expectation may be infinite, as the next two examples show.

*Example 6 (**Pareto density**).* Let X have density

$$f(x) = (\alpha - 1)x^{-\alpha} \text{ for } x \geq 1 \text{ and } \alpha > 1.$$

Then if $\alpha \leq 2$, the expected value of X is infinite because

$$E(X) = \lim_{n\to\infty} \int_1^n x(\alpha - 1)x^{-\alpha}\text{d}x = (\alpha - 1)\lim_{n\to\infty} \int_1^n x^{1-\alpha}\text{d}x,$$

which diverges to ∞ for $\alpha - 1 \leq 1$. However, for $\alpha > 2$, $E(X) = (\alpha - 1)/(\alpha - 2)$.

*Example 7 (**Cauchy density**).* Let X have density

$$f(x) = 1/\pi(1 + x^2)^{-1}, -\infty < x < \infty.$$

Because $\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{xdx}{1+x^2}$ diverges X does not have an expected value.

The various moments of a random variable with a density are defined just as they were for discrete random variables: **initial moments** $\mu_k = E(X^k)$, and **central moments** $\sigma_k = E[(X - E(X))^k]$.

Example 8 (Initial moments of normal density). Let X have the density $N(0, \sigma^2)$. Find μ_k for all k.

Solution. If k is odd, then $x^k \exp(-x^2/(2\sigma^2))$ is an odd function. Hence, $\mu_k = 0$ if k is odd. If $k = 2n$, then integrating by parts gives

$$\mu_{2n} = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} y^{2n} \exp\left(-\frac{y^2}{2\sigma^2}\right)dy =$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}}\left(\left(-\sigma^2 y^{2n-1} \exp\left(-\frac{y^2}{2\sigma^2}\right)\right)\Bigg|_{-\infty}^{\infty} +\right.$$
$$\left.+\int_{-\infty}^{\infty}(2n-1)\sigma^2\, y^{2n-2} \exp\left(-\frac{y^2}{2\sigma^2}\right)dy\right) = (2n-1)\sigma^2\mu_{2n-2} = \frac{\sigma^{2n}\,(2n)!}{2^n n!}$$

on iterating and observing that $\mu_0 = 1$.

Hence, in particular, $\mu_2 = \sigma^2$.

Example 9 (Normal densities). Let X and Y be independent with common density $f(x) = k \exp(-x^2/2)$ for all x.

(a) Show that $k = (2\pi)^{-1/2}$.

(b) Show that $X^2 + Y^2$ and $\text{arc tan}(Y/X)$ are independent random variables.

Solution. Because X and Y are independent, they have joint density $f(x, y) = k^2 \exp(-(x^2 + y^2)/2)$.

Make the change of variables to polar coordinates, so that the random variables $R = (X^2 + Y^2)^{1/2}$ and $\Theta = \arctan(Y/X)$ have joint density $f(r, \theta) = k^2 r \exp(-r^2/2)$ for $0 \le r < \infty$, $0 < \theta \le 2\pi$.

Hence, R has density $f_R(r) = r \exp(-r^2/2)$ for $0 \le r < \infty$, and Θ has density $f_\Theta(\theta) = k^2$ for $0 < \theta \le 2\pi$. It follows immediately that

(a) $k^2 = (2\pi)^{-1}$.

(b) $f(r, \theta) = f_R(r)\, f_\Theta(\theta)$, so that Θ and R are independent. Hence, Θ and R^2 are independent.

Example 10 (*Uniform distribution*). Let $Q = (X, Y)$ has the uniform density over the unit circular disc C, namely,

$$f(x, y) = \pi^{-1} \text{ for } (x, y) \in C = \{(x, y): x^2 + y^2 \le 1\}.$$

It seems more natural to use polar rather than Cartesian coordinates in this case. These are given by $r = (x^2 + y^2)^{1/2}$ and $\theta = \arctan(y/x)$, with inverse $x = r\cos\theta$ and $y = r\sin\theta$. They map $C = \{x, y: x^2 + y^2 \le 1\}$ one–one onto $D = \{r, \theta: 0 \le r \le 1, 0 < \theta \le 2\pi\}$. In this case, $J(r, \theta) = r\cos^2\theta + r\sin^2\theta = r$.

Hence, the random variables $R = r(X, Y)$ and $\Theta = \theta(X, Y)$ have joint density given by $f_{R,\Theta}(r, \theta) = r/\pi$ for $0 \le r \le 1$, $0 < \theta \le 2\pi$. Notice that $f(r, \theta)$ is not uniform, as was $f(x, y)$.

(a) Are X and Y independent?

(b) Find $f_X(x)$ and $f_Y(y)$.

(c) If $X = R\cos\Theta$, and $Y = R\sin\Theta$, are R and Θ independent?

Solution. (a) The set $\{x, y: x \le -1/\sqrt{2}, y \le -1/\sqrt{2}\}$ lies outside C, so $F(-1/\sqrt{2}, -1/\sqrt{2}) = 0$. However, the intersection of the set $\{x: x \le -1/\sqrt{2}\}$ with C has nonzero area, so $F_X(-1/\sqrt{2})F_Y(-1/\sqrt{2}) > 0$. Therefore, X and Y are not independent.

(b) $f_X(x) = \int_{-1}^{1} f(x, y)\, dy = 1/\pi\, ((1 - x^2)^{1/2} + (1 - x^2)^{1/2})$.
Likewise, $f_Y(y) = 2/\pi(1 - y^2)^{1/2}$.

(c) R and Θ have joint density $f_{R,\Theta}(r, \theta) = r/\pi$, for $0 \le r < 1$, $0 < \theta \le 2\pi$.

Hence, $f_\Theta(\theta) = \int_0^1 f(r, \theta) dr = 1/(2\pi)$; $0 < \theta \le 2\pi$, and $f_R(r) = \int_0^{2\pi} f(r, \theta) dr = 2r$; $0 \le r \le 1$. Hence, $f(r, \theta) = f_\Theta(\theta)\, f_R(r)$, and so R and Θ are independent.

Example 11 (Inverse functions). Let X have distribution function $F(x)$, where $F(x)$ is continuous and strictly increasing. Let $g(x)$ be a function satisfying $F(g) = x$. Because $F(x)$ is continuous and strictly increasing, this defines $g(x)$ uniquely for every x in $(0, 1)$.

The function $g(\cdot)$ is called the **inverse function** of $F(\cdot)$ and is often denoted by $g(x) = F^{-1}(x)$. Clearly, F is the inverse function of g, that is $g(F(x)) = F(g(x)) = x$, and $g(x)$ is an increasing function.

(a) Use this function to show that $Y = F(X)$ is uniformly distributed on (0, 1).

(b) Show that if U is uniform on (0, 1), then $Z = F^{-1}(U)$ has distribution $F(\cdot)$.

Solution. (a) As usual, we seek the distribution function

$P(Y \le y) = P(F(X) \le y) = P(g(F(X)) \le g(y)) = P(X \le g(y)) = F(g(y)) = y$.

(b) Again, $P(F^{-1}(U) \le z) = P(F(F^{-1}(U)) \le F(z)) = P(U \le F(z)) = F(z)$.

The fact is used in simulation of RVs (see 7.13).

7.8. Sums, products, and quotients of random variables

The most important is the sum of two random variables.

Theorem 7.2. *Let X and Y have joint density f (x, y). If Z = X + Y, then*

$$f_Z(z) = \int_{-\infty}^{\infty} f(u, z - u)du, \tag{7.26}$$

and if X and Y are independent, then

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(u) f_Y(z - u)du. \tag{7.27}$$

Proof. First notice that the result (7.27) follows immediately from (7.26) when X and Y are independent. Turning to the proof of (7.26), we give two methods of solution.

I. Let A be the region in which $u + v \le z$. Then $P(Z \le z) = \int_{(u,v)\in A} f(u,v)du\,dv = \int_{-\infty}^{\infty}\int_{-\infty}^{z-u} f(u,v)du\,dv = \int_{-\infty}^{\infty}\int_{-\infty}^{z} f(u, w - u)\,dwdu$ on setting $v = w - u$. Now differentiating with respect to z gives $f_Z(z) = \int_{-\infty}^{\infty} f(u, z - u)\,du$.

II. This time we use the change of variable technique. Consider the transformation $z = x + y$ and $u = x$, with inverse $x = u$ and $y = z - u$. Here $J = 1$.

This satisfies the conditions of Theorem 7.1, and so $U = u(X, Y)$ and $Z = z(X, Y)$ have joint density $f(u, z - u)$. We require the marginal density of Z, which is of course just (7.26).

Example 7.3. Let X and Y have the **bivariate normal distribution**,

$$f(x, y) = (2\pi\sigma\tau)^{-1}(1 - \rho^2)^{-1/2} \exp(-(x^2/\sigma^2 - 2\rho xy/(\sigma\tau) + y^2/\tau^2)/(2(1 - \rho^2))).$$

Find the density of $aX + bY$ for constants a and b.

Remark 7.2. Parameter $\rho \in [-1, 1]$ can be identified with the correlation coefficient corr (X, Y). More precisely, var$X = \sigma^2$, var$Y = \tau^2$, and Cov $(X, Y) = \rho\sigma\tau$. Note that X and Y are independent if and only if $\rho = 0$. The fact is proved in Example 7.9.

Solution. The joint density of $U = aX$ and $V = bY$ is $g(u, v) = 1/(ab)\times f(u/a, v/b)$. Hence, the density of $Z = U + V = aX + bY$ is

$f_Z(z) = 1/(ab)\int_{-\infty}^{\infty} f(u/a, (z - u)/b)\,du$. Rearranging the exponent in the integrand we have, after a little manipulation,

$-1/(2(1 - \rho^2))[u^2/(a^2\sigma^2) - 2\rho u(z - u)/(ab\sigma\tau) + (z - u)^2/(b^2\tau^2)] =$

$-1/(2(1 - \rho^2))[\alpha(u - \beta/\alpha z)^2 + z^2/\alpha - (1 - \rho^2)/(a^2b^2\sigma^2\tau^2)]$, where

$\alpha = 1/(a^2\sigma^2) + 2\rho/(ab\sigma\tau) + 1/(b^2\tau^2)$, and $\beta = \rho/(ab\sigma\tau) + 1/(b^2\tau^2)$.

Setting $u = v + (\beta/\alpha)\, z$ in the integrand, we evaluate

$$\int_{-\infty}^{\infty} \exp\left(-\alpha v^2/(2(1-\rho^2))\right) dv = (2\pi(1-\rho^2)/\alpha)^{1/2}.$$

Hence, after a little more manipulation, we find that

$$f_Z(z) = 1/(2\pi\xi^2)^{1/2} \exp(-z^2/(2\xi^2)),$$

where $\xi^2 = a^2\sigma^2 + 2\rho ab\sigma\tau + b^2\tau^2$. That is to say, Z is $N(0, \xi^2)$.

One important special case arises when $\rho = 0$, and X and Y are therefore independent.

So we have proved the following result.

Theorem 7.3. *Let X and Y be independent normal random variables having the densities* $N(0, \sigma^2)$ *and* $N(0, \tau^2)$. *Then the sum $Z = aX + bY$ has the density* $N(0, a^2\sigma^2 + b^2\tau^2)$.

Next we turn to products and quotients.

Theorem 7.4. *Let X and Y have joint density $f(x, y)$.*
Then the density of $Z = XY$ is

$$f_Z(z) = \int_{-\infty}^{\infty} 1/|u|\, f(u, z/u)\, du$$

and the density of $W = X/Y$ is

$$f_W(w) = \int_{-\infty}^{\infty} |u|\, f(uw, u)\, du.$$

Proof. We use Theorem 7.1 again. Consider the transformation $u = x$ and $z = xy$, with inverse $x = u$ and $y = z/u$. Here, $J(u, z) = u^{-1}$. This satisfies the conditions of Theorem 7.1 and so $U = X$ and $Z = XY$ have joint density $f(u, z) = 1/|u|\, f(u, z/u)$.

The result of the theorem follows immediately as it is the marginal density of Z obtained from $f(u, z)$.

What about the quotient $W = X/Y$? First, let $V = 1/Y$. Then, by definition,

$F_{X,V}(x, v) = P(X \le x, V \le v) = P(X \le x, Y \ge 1/v) = \int_{-\infty}^{x}\int_{1/v}^{\infty} f(s,t)\, ds dt.$

Hence, on differentiating, the joint density of X and Y^{-1} is given by $f_{X,V}(x, v) = 1/v^2 f(x, 1/v)$. Now $W = XV$, so by the first part of the theorem, $f_W(w) = \int_{-\infty}^{\infty} 1/|u|\, u^2/w^2 f(u, u/w)\, du = \int_{-\infty}^{\infty} |v|\, f(vw, v)\, dv$ on setting $u = vw$ in the integrand. Alternatively, of course, you can obtain this by using Theorem 7.1 directly via the transformation $w = x/y$ and $u = y$.

Here are some illustrative examples.

Example 7.4. Let X and Y be independent with respective density functions $f_X(x) = xe^{-x^2/2}$ for $x > 0$ and $f_Y(y) = \pi^{-1}(1 - y^2)^{-1/2}$ for $|y| < 1$. Show that XY has a normal distribution.

Solution. When X and Y are independent, we have $f(x, y) = f_X(x)\, f_Y(y)$, and $f(z) = \int_{-\infty}^{\infty} 1/|u|\, f_X(u)\, f_Y(z/u) du = \int_{u>z}^{\infty} 1/|u|\, ue^{-u^2/2}\, \pi^{-1}(1 - z^2/u^2)^{-1/2} du = \pi^{-1}\int_{z}^{\infty} e^{-u^2/2}(1 - z^2/u^2)^{-1/2}\, du.$

Now we make the substitution $u^2 = z^2 + v^2$ to find that

$f(z) = 1/\pi e^{-z^2/2}\int_0^{\infty} e^{-u^2/2}\, du = (2\pi)^{-1/2} e^{-z^2/2}$, which is the $N(0, 1)$ density.

Example 7.5. Let X and Y have density $f(x, y) = e^{-x-y}$ for $x > 0, y > 0$.
Show that $U = X/(X + Y)$ has the uniform density on (0, 1).

Solution. To use Theorem 7.4, we need to know the joint density of X and $V = X + Y$. A trivial application of Theorem 7.1 shows that X and V have density $f(x, v) = e^{-v}$ for $0 < x < v < \infty$. Hence, by Theorem 7.4,

$$f(u) = \int_0^{\infty} ue^{-u} \quad du \text{ for } 0 < uv < v; f(u) = 1, \text{ for } 0 < u < 1.$$

Alternatively, we may use Theorem 7.1 directly by considering the transformation $u = x/(x + y)$, $v = x + y$, with $x = uv$, $y = v(1 - u)$ and $|J| = v$. Hence, $U = X/(X + Y)$ and $V = X + Y$ have density $f(u, v) = ve^{-v}$, for $v > 0$ and $0 < u < 1$.

The marginal density of U is 1, as required.

7.9. Moment generating function

The moment-generating function (MGF) of a random variable is in some cases an alternative definition of its probability distribution. It provides the basis of an alternative route to analytical results without working directly with probability distributions. So we can consider MGF as a technical tool (useful workhorse) for receiving some mathematical results.

Definition 7.1. The **moment generating function** $M_X(t)$ (or $M(t)$) is given by

$$M(t) = M_X(t) = E(e^{tX}).$$

If X has density f, then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tu} f(u) du,$$

if X is discrete RV then

$$M_X(t) = \sum_i \ e^{x_i} p_i.$$

MGF does not always exist. We are only interested in $M_X(t)$ for those values of t for which it is finite; this includes $t = 0$, of course.

Example 7.6 (MGF of uniform RV). Let X be uniform on $[0, a]$. Find $E(e^{tX})$. Where does it exist?

Solution. $E(e^{tX}) = \int_0^a \frac{e^{tx}}{a} dx = (e^{at} - 1)/(at)$. This exists for all t, including $t = 0$, where it takes the value 1.

Example 7.7 (MGF of normal RV). Let X be a standard normal random variable. Then $\sqrt{2\pi} M_X(t) = \int_{-\infty}^{\infty} e^{tx - x^2/2} dx = \int_{-\infty}^{\infty} e^{t^2/2 - (x-t)^2/2} dx =$ $= e^{t^2/2} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx = e^{t^2/2}\sqrt{2\pi}$.

So $M_X(t) = e^{t^2/2}$.

Now by properties of MGF (see Exercise 14) if $Y = \sigma X + \mu$ (which is $N(\mu, \sigma^2)$), then $M_Y(t) = e^{\mu t} e^{t^2\sigma^2/2}$.

Why is $M_X(t)$ called the moment generating function? There are relations between the behavior of the MGF of a distribution and properties of the distribution, such as the existence of moments.

Consider the following formal expansion:

$$E(e^{tX}) = E(\sum_{k=0}^{\infty} X^k\, t^k/k!) = \sum_{k=0}^{\infty} E(X^k)t^k/k! = \sum_{k=0}^{\infty} \mu_k t^k/k!, \quad (7.28)$$

provided the interchange of expectation and summation at (7.28) is justified. The required interchange at (7.28) is permissible if $M_X(t)$ exists in an interval that includes the origin.

The individual coefficients of the t-powers in (7.28) are the initial moments of the distribution, each divided by the corresponding factorial. Quite often this is the easiest way of calculating them. Note that this is equivalent to:

$$E(X^k) = M_X(t)^{(k)}\Big|_{t=0}$$

or, in words, to get the k-th initial moment differentiate the corresponding MGF k times (with respect to t) and set t equal to zero.

We state the following inversion theorem without proof.

Theorem 7.5. *If X has moment generating function $M_X(t)$, where for some $a > 0$, $M_X(t) < \infty$ for $|t| < a$, then the distribution of X is determined uniquely. Furthermore, the expansion (7.28) holds.*

Remark 7.3. Even though a full information about the corresponding distribution is “encoded” into a MGF, its ’decoding’ (converting MGF back into a table of probabilities or density function) is somehow more involved and will not be discussed here.

MGF is especially useful in dealing with sequences of random variables; the following theorem is the basis of this assertion. We state it without proof.

Theorem 7.6 (Continuity theorem). *Let $(F_n(x);\ n \geq 1)$ be a sequence of distribution functions with corresponding MGFs $(M_n(t);\ n \geq 1)$ that exist for $|t| < b$. Suppose that as $n\to\infty$ $M_n(t) \to M(t)$ for $|t| \leq a < b$, where $M(t)$ is the MGF of the distribution $F(x)$. Then, as $n\to\infty$, $F_n(x) \to F(x)$ at each point x where $F(x)$ is continuous.*

The main application of this theorem arises when $M(t) = e^{t^2/2}$ (MGF of what distribution?) and $F(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \exp\left(-\frac{y^2}{2}\right) dy = \Phi(x)$ (so called **probability integral**) being the CDF of normal RV.

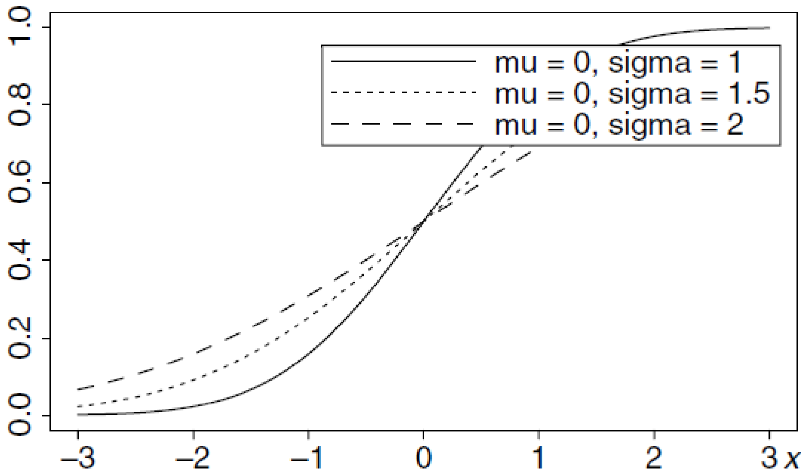


Figure 7.6. Normal (Gaussian) distribution functions with parameters μ (mu) and σ^2 (sigma squared)

Function $\Phi(x)$ is an object of paramount importance in probability theory and statistics. It is also called a **Gaussian distribution function**, named after K.F. Gauss (1777–1855), the famous German mathematician, astronomer and physicist, who made a profound impact on a number of areas of mathematics. He identified the distribution while working on the theory of errors in astronomical observations. Gaussian distribution fitted the pattern of errors much better than "double-exponential" distribution previously used by Laplace.

The values of $\Phi(x)$ have been calculated with a great accuracy for a narrow mesh of values of x and constitute a major part of the probabilistic and statistical tables. See Table A6.1 (Appendix 6).

7.10. Distributions concerning normal

A significant part of the statistics course is concerned with iid N(0,1) RVs X_1, X_2, … and their functions. The simplest functions are linear combinations, i.e. $\sum_{i=1}^{n} a_i X_i$.

It is easy to see using MGF that if $X_1, X_2, \ldots, X_n$ are independent $N(\mu_i, \sigma_i^2)$ RVs then $\sum_{i=1}^{n} a_i X_i$ is normal $N(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} (a_i \sigma_i)^2)$.

Another example is the sum of squares. Let $X_1, X_2, \ldots, X_n$ be iid N(0,1) RVs. The distribution of the sum $\sum_{i=1}^{n} X_i^2$ is called the **chi-square (**or **chi-squared) distribution** denoted by χ_n^2, the parameter n is called **degrees of freedom**. It has the PDF $f_{\chi_n^2}(\cdot)$ concentrated on the positive half-axis $(0, \infty)$:

$$f_{\chi_n^2}(x) = C x^{n/2-1} e^{-x/2}, \quad x > 0$$

with the constant of proportionality $C = \left[\Gamma(n/2) 2^{n/2}\right]^{-1}$, here

$$\Gamma(u) = \frac{1}{2^u} \int_0^\infty x^{u-1} e^{-x/2} dx.$$

A useful property of the family of χ_n^2 distributions is that it is closed under independent summation as normal distribution. That is if $Z \in \chi_n^2$ and $Y \in \chi_m^2$ independently, then $Z + Y \in \chi_{n+m}^2$.

The mean value of the χ_n^2 distribution equals n and the variance $2n$. All χ^2 PDFs are unimodal. A sample of graphs of PDF $f_{\chi_n^2}(x)$ is shown in Figure 7.7. See Table A6.2 (Appendix 6) for tabulation of chi-squared distribution.

If, as above, X_1, X_2, …, X_{n+1} is iid N(0,1) RVs then the distribution of the ratio

$$\frac{X_{n+1}}{\sqrt{\sum_{i=1}^{n} X_i^2 / n}}$$

is called the **Student's distribution**, with n degrees of freedom, or the t_n distribution for short. It has the PDF $f_{t_n}(\cdot)$ spread over the whole axis R^1 and is an even function:

$$f_{t_n}(t) = C\left(1 + t^2/n\right)^{-(n+1)/2}$$

with the proportionality constant $C = \frac{1}{\sqrt{\pi n}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)}$.

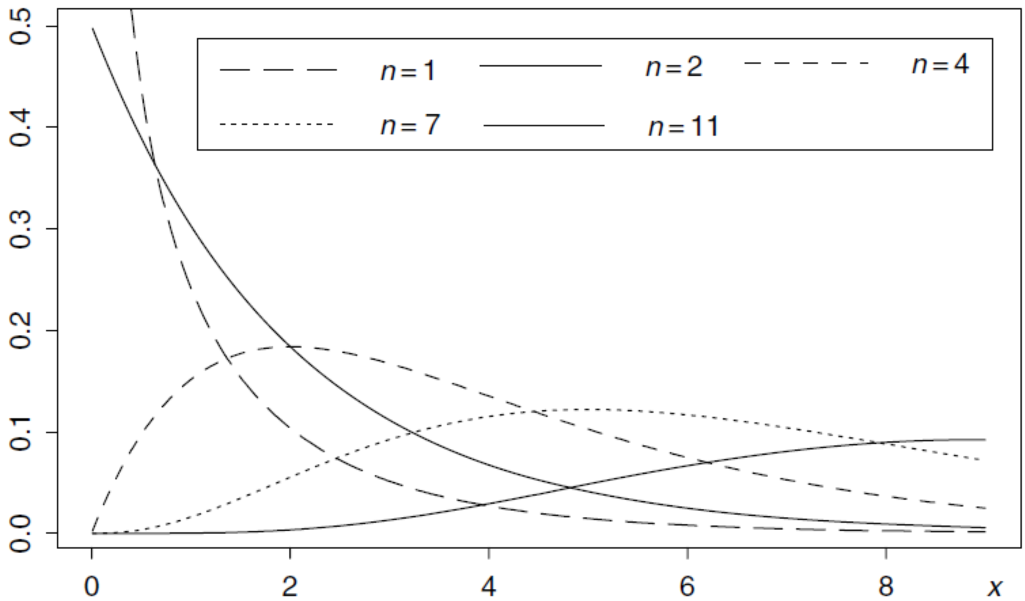


Figure 7.7. Chi-square distributions with n degrees of freedom

For $n > 1$ it has, obviously, the mean value 0. For $n > 2$, the variance is $n/(n - 2)$. All Student's PDFs are unimodal. A sample of graphs of PDF $f_{t_n}(x)$ is shown in Figure 7.8.

These PDFs resemble normal PDFs and $f_{t_n}(x)$ approaches $\varphi(x) = e^{-x^2/2}/\sqrt{2\pi}$ as $n \to \infty$ and the percentage points for t_n tend to those for N(0,1). In practice we can use normal approximation for $n \geq 30$, see Table A6.3.

However, for finite n, the "tails" of $f_{t_n}(x)$ are "thicker" than those of the normal PDF. In particular, the MGF of a t distribution does not exist (except at zero point). Note that for $n = 1$, the t_1 distribution coincides with the Cauchy distribution.

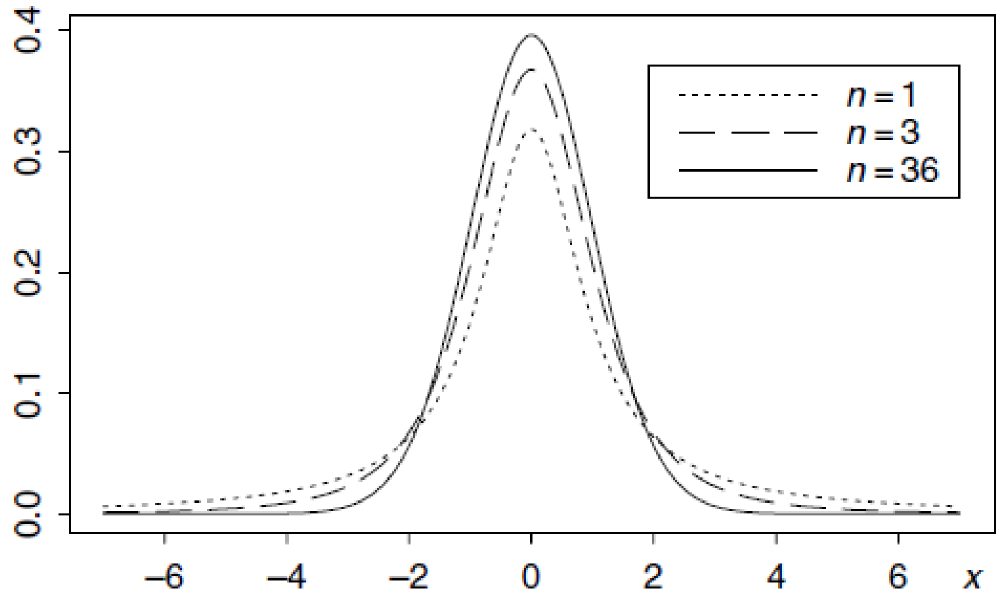


Figure 7.8. Student distributions with n degrees of freedom

7.11. Conditional density and expectation

Suppose that X and Y have joint density $f(x, y)$, and we are given the value of Y. We make the following definition.

Definition 7.2. *If X and Y have joint density f (x, y), then the* **conditional density** *of X given Y = y is given by* $f_{X|Y}(x|y) = f(x, y)/f_Y(y)$, *if* $0 < f_Y(y) < \infty$, $f_{X|Y}(x|y) = 0$ *elsewhere.*

When X and Y are discrete the definition is analogous.

We observe immediately that $f_{X|Y}(x|y)$ is indeed a density, because it is nonnegative and

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)dx = \int_{-\infty}^{\infty} f(x,y)/f_Y(y)dx = f_Y(y)/f_Y(y) = 1. \quad (7.29)$$

The corresponding **conditional distribution function** is

$$F_{X|Y}(x, y) = \int_{-\infty}^{x} f_{X|Y}(x|y)dx = P(X \le x|Y = y),$$

and we have the rule

$$P(X \in A|Y = y) = \int_A f_{X|Y}(x|y)dx.$$

Example 7.8. Let (X, Y) be the coordinates of the point Q uniformly distributed on a circular disc of unit radius. What is $f_{Y|X}(y|x)$?

Solution. Recall that for the marginal density $f_X(x) = (2/\pi)(1 - x^2)^{1/2}$. Hence, by definition,

$$f_{Y|X}(y|x) = f(x, y)/f_X(x) = 1/\pi(2/\pi)^{-1}(1 - x^2)^{-1/2} = (1 - x^2)^{-1/2}/2$$

for $|y| < (1 - x^2)^{1/2}$.

This conditional density is uniform on $(-(1 - x^2)^{1/2}, (1 - x^2)^{1/2})$.

Example 7.9. Let X and Y be independent and exponential with parameter λ. Show that the density of X conditional on $X + Y = v$ is uniform on $(0, v)$.

Solution. To use (7.29), we need to take some preliminary steps. First note that the joint density of X and Y is $f(x, y) = \lambda^2 e^{-\lambda(x+y)}$ for $x > 0, y > 0$.

Next we need the joint density of X and $X + Y$ so we consider the transformation $u = x$ and $v = x + y$, with inverse $x = u$ and $y = v - u$, so that $J = 1$.

Hence, by Theorem 7.1,

$$f_{U,V}(u, v) = \lambda^2 e^{-\lambda v} \text{ for } 0 < u < v < \infty.$$

It follows that $f_V(v) = \int_0^v \lambda^2 e^{-\lambda v}\, du = \lambda^2 v e^{-\lambda v}$ and so by definition $f_{U|V}(u|v) = f(u, v)/f_V(v) = 1/v$ for $0 < u < v$. This is the required uniform density.

This striking result is related to **the lack-of-memory property ($P(X > x + y \mid X > x) = P(X > y)$ for any x and y)** of the exponential density.

Example 7.10 (*Bivariate normal*). Let X and Y have the bivariate normal density

$$f(x, y) = (2\pi\sigma\tau)^{-1}(1 - \rho^2)^{-1/2} \exp[-(x^2/\sigma^2 - 2\rho xy/(\sigma\tau) + y^2/\tau^2)/(2(1 - \rho^2))].$$

Find the conditional density of X given $Y = y$.

Solution. $f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy =$

$$= (2\pi\sigma\tau)^{-1}(1 - \rho^2)^{-1/2}\int_{-\infty}^{\infty} \exp -\left(\left(\frac{x^2}{\sigma^2} - \frac{\rho^2 x^2}{\sigma^2} + \left(\frac{y}{\tau} - \frac{\rho x}{\sigma}\right)^2\right)/(2(1 - \rho^2))\right) dy.$$

Now setting $\frac{y}{\tau} - \frac{\rho x}{\sigma} = u$, and recalling that

$\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2(1-\rho^2)\sigma^2}\right) dy = (2\pi(1 - \rho^2))^{1/2}$ yields $f_X(x) = (2\pi\sigma^2)^{-1/2}\exp(-x^2/\sigma^2)$.

This is $N(0, \sigma^2)$ density. Interchanging the roles of x and y in the above integrals shows that $f_Y(y)$ is the $N(0, \tau^2)$ density.

Hence, $f_{X|Y}(x|y) = f(x, y)/f_Y(y) = (2\pi\sigma\tau)^{-1}(1 - \rho^2)^{-1/2}\times$
$\times\exp(-(x^2/\sigma^2 - 2\rho xy/(\sigma\tau) + y^2/\tau^2)/(2(1 - \rho^2))\ (2\pi\tau^2)^{-1/2} \exp(y^2/(2\tau^2)) =$
$= (2\pi\sigma^2(1 - \rho^2))^{-1/2} \exp(-1/(2(1 - \rho^2))(x\sigma - \rho y/\tau)^2)$.

Hence, the conditional density of X given $Y = y$ is $N(\rho\sigma y/\tau, \sigma^2(1 - \rho^2))$. Note that if $\rho = 0$, then this does not depend on y, which is to say that X is independent of Y.

Now, because $f_{X|Y}(x|y)$ is a density it may have an expected value, which naturally enough is called conditional expectation.

Definition 7.3. If $\int_{R^+} |x|\, f_{X|Y}(x|y)\, dx < \infty$, then the **conditional expectation** of X given $Y = y$ is given by

$$E(X|Y = y) = \int_R x\, f_{X|Y}(x|y) dx\,.$$

If the value of Y is left unspecified, we write $E(X|Y) = \psi(Y)$ on the understanding that when $Y = y$, $\psi(Y)$ takes the value $E(X|Y = y) = \psi(y)$. By writing $\psi(y) = E(X|Y = y)$ we emphasize the fact that the conditional expectation of X given $Y = y$ is a function of y.

Example 7.11. If X and Y are independent and exponential, then we showed that the density of X given $X + Y = v$ is uniform on $(0, v)$. Hence, $E(X|X + Y = v) = v/2$.

Actually, this is otherwise obvious because, for reasons of symmetry, $E(X|X + Y = v) = E(Y|X + Y = v)$, and trivially $E(X + Y|X + Y = v) = v$. Hence the result follows, provided it is true that for random variables X, Y and V, we have $E(X + Y|V = v) = E(X|V = v) + E(Y|V = v)$. In fact, this is true.

The conditional expectation has all properties of usual expectation. Among the most important is that

$$E(Xg(Y)|Y = y) = g(y)\psi(y) \tag{7.30}$$

for any function $g(Y)$.

It is natural to think of $E(X|Y)$ as a random variable that is a function of Y. (A more rigorous analysis can indeed justify this assumption.)

We give the following theorem without proof.

Theorem 7.7. *The expected value of* $\psi(Y)$ *is* $E(X)$; *thus,* $EX = E(E(X|Y))$.
Finally, we stress that conditional expectation is important in its own right, it should not be regarded merely as a stage on the way to calculating something else. Conditional expectation is a concept of great importance and utility.

For example, suppose that X and Y are random variables, and we want to record the value of X. Unfortunately, X is inaccessible to measurement, so we can only record the value of Y. Can this help us to make a good guess at X? First, we have to decide what a "good" guess $g(Y)$ at X is. We decide that $g_1(Y)$ is a better guess than $g_2(Y)$ if $E[(g_1(Y) - X)^2] < E[(g_2(Y) - X)^2]$.

According to this (somewhat arbitrary) rating, it turns out that the best guess at X given Y is $\psi(Y) = E(X|Y)$.

Theorem 7.8. *For any function* $g(Y)$ *of* Y, $E[(X - g(Y))^2] \geq E[(X - \psi(Y))^2]$.

Proof. Using Theorem 7.7 and (7.30), we have $E[(X - \psi)(\psi - g)] =$
$E[(\psi - g)E(X - \psi|Y)] = 0$. Hence,
$E[(X - g)^2] = E[(X - \psi + \psi - g)^2] = E[(X - \psi)^2] + E[(\psi - g)^2] \geq E[(X - \psi)^2]$.

7.12. The median, quartiles, percentiles. Skewness and kurtosis

The median of a continuous distribution is the number which will be exceeded with a 50 % probability; consequently, a smaller result is obtained with the complementary probability of 50 %, i.e. the median divides the distribution in two equally probable halves. Mathematically, we can find it as a solution of F (med) = 1/2, or (equivalently) to 1 − F (med) = 1/2 (med will be our usual notation for the median). For a symmetric distribution (uniform, normal) the mean and median must be both at the center of symmetry (yet, there is an important distinction: the mean may not always exist, the median always does).

The median of the exponential distribution is thus the solution to $e^{-\lambda \mathrm{med}} = 1/2 \Leftrightarrow \lambda \mathrm{med} = \ln 2$, yielding med = $(\ln 2)/\lambda$ ($\approx 0.6931\lambda^{-1}$) (substantially smaller that the corresponding mean λ^{-1}).

Suppose that the waiting time to catch a fish in the pond has exponential distribution. This means that if it takes, on the average, 1 hour to catch a fish, 50 % of all fishes are caught in less than 41 min. and 35 sec.

The **lower quartile** l and the **upper quartile** u are similarly defined by $F(l) = 1/4$, $F(u) = 3/4$.

Thus, the probability that X lies between l and u is 3/4−1/4 = 1/2, so the quartiles give an estimate of how spread-out the distribution is.

More generally, we define the nth **percentile** of X to be the value of x_n such that $F(x_n) = n/100$, that is, the probability that X is smaller than x_n is n %.

These values are the particular cases of so called **quantiles**: the quantile of level q is the value α_q such that $F(\alpha_q) = q$. Here and above in this section we assume that $F(\cdot)$ is strictly monotonic distribution function, otherwise the solutions of the quantiles' equations can be indeterminate.

The skewness is a measure of the asymmetry of the probability distribution. The skewness value can be positive or negative, or even undefined. Qualitatively, a negative skew indicates that the tail on the left side of the probability density function is longer than the right side and the bulk of the values (possibly including the median) lie to the right of the mean. A positive skew indicates that the tail on the right side is longer than the left side and the bulk of the values lie to the left of the mean. A zero value indicates that the values are relatively evenly distributed on both sides of the mean, typically but not necessarily implying a symmetric distribution.

If the distribution is symmetric then the mean equals median and there is zero skewness.

The skewness of a random variable X is the third standardized moment defined as

$$\mathrm{Skew}(X) = \mu_3 / \sigma^3,$$

where μ_3 is the third central moment, σ is the standard deviation.

The kurtosis is any measure of the "peakedness" of the probability distribution. In a similar way to the concept of skewness, kurtosis is a descriptor of the shape of a probability distribution.

Kurtosis is commonly defined as the fourth standardized moment of the probability distribution minus 3

$$\mathrm{Kurt}(X) = \mu_4 / \sigma^4 - 3.$$

The "minus 3" at the end of this formula is often explained as a correction to make the kurtosis of the normal distribution equal to zero.

A high kurtosis distribution has a sharper peak and longer, fatter tails, while a low kurtosis distribution has a more rounded peak and shorter, thinner tails.

For more details see http://en.wikipedia.org/wiki/Skewness and http://en.wikipedia.org/wiki/Kurtosis.

7.13. Simulation of random variables

A random variable is a mathematical concept (having no other existence) that is suggested by the outcomes of real experiments. Thus, tossing a coin leads us to define an $X(\cdot)$ such that $X(H) = 1$, $X(T) = 0$, and X is the number of heads. The coin exists, X is a concept.

A natural next step, having developed theorems about mathematical coins is to test them against reality. However, the prospect of actually tossing a large enough number of coins to check the theoretical laws is rather forbidding.

Luckily, we have machines to do large numbers of boring and trivial tasks quickly, namely, computers. These can be persuaded to produce many numbers (u_i; $i \geq 1$) that are sprinkled evenly and "randomly" over the interval (0, 1). The word randomly appears in quotations because each u_i is not really random. Because the machine was programmed to produce it, the outcome is known in advance, but such numbers behave for many practical purposes as though they were random. They are called **pseudorandom numbers**.

Now if we have a pseudorandom number u from a collection sprinkled uniformly in (0, 1), we can look to see if $u < 1/2$, in which case we call it "heads", or $u > 1/2$ in which case we call it "tails." This process is called **simulation**; we have simulated tossing a coin.

Different problems produce different random variables, but computers find it easiest to produce uniform pseudorandom numbers. We are thus forced to consider appropriate transformations of uniform random variables.

Here there are two real life examples answering the question why might we want to simulate such random variables?

Example 7.12 (Epidemic). An infection is introduced into a population. For each individual the incubation period is a random variable X, the infectious period is a random variable Y, and the number of further individuals infected is a random variable *N*, depending on *Y* and the behavior of the infected individual. What happens?

Unfortunately, exact solutions to such problems are rare and, for many diseases (e.g., the so-called “slow viruses” or prions), *X* and *Y* are measured in decades so experiments are impractical. However, if we could simulate *X* and *Y* and the infection process *N*, then we could produce one simulated realization (not a real realization) of the epidemic. With a fast computer, we could do this many times and gain a pretty accurate idea of how the epidemic would progress (if our assumptions were correct).

Example 7.13 (Toll Booths). Motorists are required to pay a fee before entering a toll road. How many toll booths should be provided to avoid substantial queues? Once again an experiment is impractical. However, simple apparatus can provide us with the rates and properties of traffic on equivalent roads. If we then simulate the workings of the booth and test it with the actual traffic flows, we should obtain reasonable estimates of the chances of congestion.

7.14. On using tables

We end this section with a few comments about using tables, not tied particularly to the normal distribution (though the examples will come from there).

Interpolation. Any table is limited in the number of entries it contains. Tabulating something with the input given to one extra decimal place would make the table ten times as bulky. Interpolation can be used to extend the range of values tabulated.

Suppose that some function *F* is tabulated with the input given to three places of decimals. It is probably true that *F* is changing at a roughly constant rate between, say, 0.28 and 0.29. So $F(0.283)$ will be about three-tenths of the way between $F(0.28)$ and $F(0.29)$.

For example, if *F* is the CDF of the normal distribution, then $F(0.28) = 0.6103$ and $F(0.29) = 0.6141$, so $F(0.283) = 0.6114$. (Three-tenths of 0.0038 is 0.0011.)

Using tables in reverse. This means, if you have a table of values of *F*, use it to find *x* such that $F(x)$ is a given value *c*. Usually, *c* won't be in the table and we have to interpolate between values x_1 and x_2, where $F(x_1)$ is just less than *c* and $F(x_2)$ is just greater.

For example, if *F* is the CDF of the normal distribution, and we want the upper quartile, then we find from tables $F(0.67)=0.7486$ and $F(0.68)=0.7517$, so the required value is about 0.6745 (since 0.0014/0.0031 = 0.45).

Nowadays we can use computers instead of tables.

7.15. Exercises

1. If X is a random variable (on a countable sample space), is it true that $X + X = 2X$, $X - X = 0$? Explain in detail.

2. Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $P(\omega_1) = P(\omega_2) = P(\omega_3) = 1/3$, and define X, Y, and Z as follows:
$X(\omega_1) = 1$, $X(\omega_2) = 2$, $X(\omega_3) = 3$;
$Y(\omega_1) = 2$, $Y(\omega_2) = 3$, $Y(\omega_3) = 1$;
$Z(\omega_1) = 3$, $Z(\omega_2) = 1$, $Z(\omega_3) = 2$.
Show that these three random variables have the same probability distribution. Find the probability distributions of $X + Y$, $Y + Z$, and $Z + X$.

3. In Ex. 2 find the probability distribution of $X + Y - Z$, $\sqrt{(X^2 + Y^2)Z}$, $Z/|X - Y|$.

4. Take Ω to be a set of five real numbers. Define a probability measure and a random variable X on it that takes the values 1, 2, 3, 4, 5 with probabilities 1/10, 1/10, 1/5, 1/5, 2/5 respectively; another random variable Y that takes the value $\sqrt{2}$, $\sqrt{3}$, π with probabilities 1/5, 3/10, 1/2. Find the probability distribution of XY. (Hint: the answer depends on your choice and is not unique.)

5. Generalize Exercise 4 by constructing Ω, P, X so that X takes the values $v_1, v_2, \ldots, v_n$ with probabilities $p_1, p_2, \ldots, p_n$ where the sequence of probabilities satisfies the normalized condition.

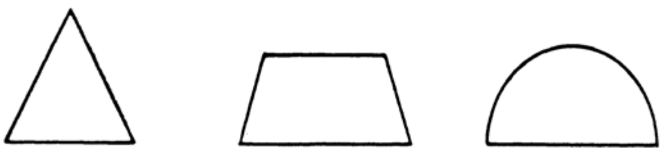
6. Let $\lambda > 0$ and define function f as follows: $f(u) = 0.5\lambda e^{-\lambda u}$ if $u \geq 0$; $0.5\lambda e^{+\lambda u}$ if $u < 0$.
This distribution is called **bilateral exponential**. If X has density f, find the density of $|X|$. (Hint: begin with the distribution function.)

7. If X is a positive random variable with density f, find the density of $+\sqrt{X}$. Apply this to the distribution of the side length of a square when its area is uniformly distributed in $[a, b]$.

8. If X has density f, find the density of (i) $aX + b$ where a and b are constants; (ii) X^2.

9. If f and g are two density functions, show that $\lambda f + \mu g$ is also a density function, where $\lambda + \mu = 1$, $\lambda \geq 0$, $\mu \geq 0$.

10. In the figure below an equilateral triangle, a trapezoid, and a semidisk are shown:



Determine numerical constants for the sides and radius to make these the graphs of density functions.

11. Suppose a target is a disk of radius 10 feet and that the probability of hitting within any concentric disk is proportional to the area of the disk. Let R

denote the distance of the bullet from the center. Find the distribution function, density function, and mean of R.

12. Agent 009 was trapped between two narrow abysmal walls. He swung his gun around in a vertical circle touching the wall, and fired a wild (random) shot. Assume that the angle his pistol makes with the horizontal is uniformly distributed between 0◦and π/2. Find the distribution of the height where the bullet landed and the mean.

13. Pick two numbers at random from [0, 1]. Let X denote the smaller, Y the larger of the two numbers so obtained. Describe the joint distribution of (X, Y), and the marginal ones. Find the distribution of $Y - X$ from the joint distribution. (Hint: draw the picture and compute areas.)

14. Prove the following properties of MGF using the properties of expectation: if X and U are independent and have MGFs $M_X(t)$ and $M_U(t)$ correspondently, then i) $Y = aX + b$ has MGF $M_Y(t) = M_X(at)\, e^{bt}$, ii) $Z = aX + bU$ has MGF $M_Z(t) = M_X(at)M_U(bt)$, where a and b are some constants.

15. Find the expectation and variance of normal distribution using MGF.

16. Find MGF of exponential distribution.

17. Proof

i) the properties of $\Phi(x)$: $\Phi(x) = 1 - \Phi(-x)\ \forall\ x \in R^1$, and $\Phi(0) = 1/2$;

ii) the equations $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y\exp\left(-\frac{y^2}{2}\right)dy = 0$, which means that the mean value of the standard Gaussian distribution is 0, and $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y^2\exp\left(-\frac{y^2}{2}\right)dy = 1$, which means that the variance of the standard Gaussian distribution is 1.

18. Take two related continuous random variables and draw what you think would be their densities. For example, you can take:

(i) distributions of income in a wealthy neighborhood and in a poor neighborhood, (ii) distributions of temperature in winter and summer in a given geographic location;

(iii) distributions of electricity consumption in two different locations at the same time of the year.

Don't forget that the "total" density is 1!