Chapter 6

Continuous Distributions

In principle it is easy to calculate probabilities such as $\mathbb{P}\{\text{Bin}(30, p) \geq 17\}$ for various values of p: one has only to sum the series

$$\binom{30}{17} p^{17}(1-p)^{13} + \binom{30}{18} p^{18}(1-p)^{12} + \ldots + (1-p)^{30}$$

With a computer (compare with the Matlab m-file BinProbs.m) such a task would not be as arduous as it used to be back in the days of hand calculation. In this Chapter I will discuss another approach, based on an exact representation of the sum as a beta integral, as a sneaky way of introducing the the concept of a continuous distribution. (Don't worry if you have never heard of the beta integral; I'll explain it all.)

For many purposes it would suffice to have to good approximation to the sum. The best known method—the normal approximation, due to de Moivre (1733)—will be described in Chapter 7. The beta integral will be starting point for one derivation of the normal approximation.

Binomial to beta

The connection between the Binomial distribution and the beta integral becomes evident when we consider a special method for simulating coin tosses. Start from a random variable U that is UNIFORMLY DISTRIBUTED on the interval $[0, 1]$. That is,

•uniformly distributed

$$\mathbb{P}\{a \leq U \leq b\} = b - a \qquad \text{for all } 0 \leq a \leq b \leq 1.$$

This distribution is denoted by Uniform$[0, 1]$. It is a different sort of distribution from the geometric or Binomial. Instead of having only a discrete range of possible values, U ranges over a continuous interval. It is said to have a CONTINUOUS DISTRIBUTION. Instead of giving probabilities for U taking on discrete values, we must specify probabilities for U to lie in various subintervals of its range. Indeed, if you put a equal to b you will find that $\mathbb{P}\{U = b\} = 0$ for each b in the interval [0,1].

•continuous distribution

To distinguish more clearly between continuous distributions and the sort of distributions we have been working with up to now, a random variable like X_n that take values in a discrete range, will be said to have a DISCRETE DISTRIBUTION.

•discrete distribution

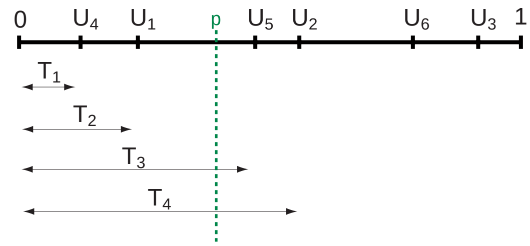
Of course, to actually simulate a Uniform$[0, 1]$ distribution on a computer one would work with a discrete approximation. For example, if numbers were specified to only 7 decimal places, one would be approximating Uniform[0,1] by a discrete distribution placing probabilities of about 10^{-7} on a fine grid of about 10^7 equi-spaced points in the interval. You might think of the Uniform$[0, 1]$ as a convenient idealization of the discrete approximation.

For a fixed n (such as $n = 30$), generate independently n random variables $U_1, \ldots, U_n$, each distributed uniformly on $[0, 1]$. Fix a p in $[0, 1]$. Then the independent events

$$\{U_1 \leq p\}, \{U_2 \leq p\}, \ldots, \{U_n \leq p\}$$

are like n independent flips of a coin that lands heads with probability p. The number, X_n, of such events that occur has a Bin(n, p) distribution.

The trick for reexpressing Binomial probabilities as integrals involves new random variables defined from the U_i. Write $T_1, \dots, T_n$ for the values of the $U_1, \dots, U_n$ rearranged into increasing order:



●order statistics

The $\{T_i\}$ are called the ORDER STATISTICS of the $\{U_i\}$, and are often written as $U_{(1)}$, $U_{(2)}, \dots, U_{(n)}$ or $U_{n:1}, U_{n:2}, \dots, U_{n:n}$. (Exercise: What is $\mathbb{P}\{U_{(1)} = U_1\}$?)

The random variables T_i, which have continuous distributions, are related to X_n, which has a discrete distribution, by the equivalence:

$$X_n \geq k \text{ if and only if } T_k \leq p.$$

That is, there are k or more of the U_i's in $[0, p]$ if and only if the kth smallest of them is in $[0, p]$. Thus

$$\mathbb{P}_p\{X_n \geq k\} = \mathbb{P}_p\{T_k \leq p\}.$$

I have added a subscript p to the probability to remind you that the definition of X_n involves p; the random variable X_n has a Bin(n, p) distribution.

Now all we have to do is find the distribution of T_k, or, more specifically, find the probability that it lies in the interval $[0, p]$.

The distribution of the order statistics from the uniform distribution

To specify the distribution of the random variable T_k we need to find the probability that it lies in a subinterval $[a, b]$, for all choices of a, b with $0 \leq a < b \leq 1$.

Start with a simpler case where the interval is very short. For $0 < t < t + \delta < 1$ and δ very small, find $\mathbb{P}\{t \leq T_k \leq t + \delta\}$.

Decompose first according to the number of points $\{U_i\}$ in $[t, t + \delta]$. If there is only one point in $[t, t + \delta]$ then we must have exactly $k - 1$ points in $[0, t)$ to get $t \leq T_k \leq t + \delta$. If there are two or more points in $[t, t + \delta]$ then it becomes more complicated to describe all the ways that we would get $t \leq T_k \leq t + \delta$. Luckily for us, the contributions from all those complicated expressions will turn out to be small enough to ignore if δ is small. Let us calculate.

$$\begin{aligned} \mathbb{P}\{t \leq T_k \leq t + \delta\} &= \mathbb{P}\{\text{exactly 1 point in } [t, t+\delta], \text{exactly } k-1 \text{ points in } [0, t)\} \\ &\quad + \mathbb{P}\{t \leq T_k \leq t + \delta,\ \text{two or more points in } [t, t+\delta]\} \end{aligned} \tag{<6.1>}$$

Let me first dispose of the second contribution on the right-hand side of <6.1>. The indicator function of the event

$$F_2 = \{t \leq T_k \leq t + \delta,\ \text{two or more points in } [t, t+\delta]\}$$

is less than the sum of indicator functions

$$\sum_{1 \leq i < j \leq n} 1\{U_i, U_j \text{ both in } [t, t+\delta]\}$$

You should check this assertion by verifying that the sum of indicators is nonnegative and that it takes a value ≥ 1 if the event F_2 occurs. Take expectations, remembering that the probability of an event is equal to the expectation of its indicator function, to deduce that

$$\mathbb{P}F_2 \le \sum_{1\le i<j\le n} \mathbb{P}\{U_i,\ U_j \text{ both in } [t, t+\delta]\}$$

By symmetry, all $\binom{n}{2}$ terms in the sum are equal to

$$\begin{aligned}&\mathbb{P}\{U_1, U_2 \text{ both in } [t, t+\delta]\}\\ &\qquad = \mathbb{P}\{t \le U_1 \le t+\delta\}\mathbb{P}\{t \le U_2 \le t+\delta\} \qquad \text{by independence}\\ &\qquad = \delta^2.\end{aligned}$$

Thus $\mathbb{P}F_2 \le \binom{n}{2}\delta^2$, which tends to zero much faster than δ as $\delta \to 0$. (The value of n stays fixed throughut the calculation.)

Next consider the first contribution on the right-hand side of <6.1>. The event

$$F_1 = \{\text{exactly 1 point in } [t, t+\delta], \text{exactly } k-1 \text{ points in } [0, t)\}$$

can be broken into disjoint pieces like

$$\{U_1, \ldots, U_{k-1} \text{ in } [0, t),\ U_k \text{ in } [t, t+\delta],\ U_{k+1}, \ldots, U_n \text{ in } (t+\delta, 1]\}.$$

Again by virtue of the independence between the $\{U_i\}$, the piece has probability

$$\mathbb{P}\{U_1 < t\}\mathbb{P}\{U_2 < t\}\ldots\mathbb{P}\{U_{k-1} < t\}\mathbb{P}\{U_k \text{ in } [t, t+\delta]\}\mathbb{P}\{U_{k+1} > t+\delta\}\ldots\mathbb{P}\{U_n > t+\delta\},$$

Invoke the defining property of the uniform distribution to factorize the probability as

$$t^{k-1}\delta(1-t-\delta)^{n-k} = t^{k-1}(1-t)^{n-k}\delta + \text{ terms of order } \delta^2 \text{ or smaller.}$$

How many such pieces are there? There are $\binom{n}{k-1}$ ways to choose the $k-1$ of the U_i's to land in $[0, t)$, and for each of these ways there are $n-k+1$ ways to choose the observation to land in $[t, t+\delta]$. The remaining observations must go in $(t+\delta, 1]$. We must add up

$$\binom{n}{k-1} \times (n-k+1) = \frac{n!}{(k-1)!(n-k)!}$$

pieces with the same probability to calculate $\mathbb{P}F_1$.

Consolidating all the small contributions from $\mathbb{P}F_1$ and $\mathbb{P}F_2$ we get

<6.2> $$\mathbb{P}\{t \le T_k \le t+\delta\} = \frac{n!}{(k-1)!(n-k)!}t^{k-1}(1-t)^{n-k}\delta + \text{ terms of order } \delta^2 \text{ or smaller.}$$

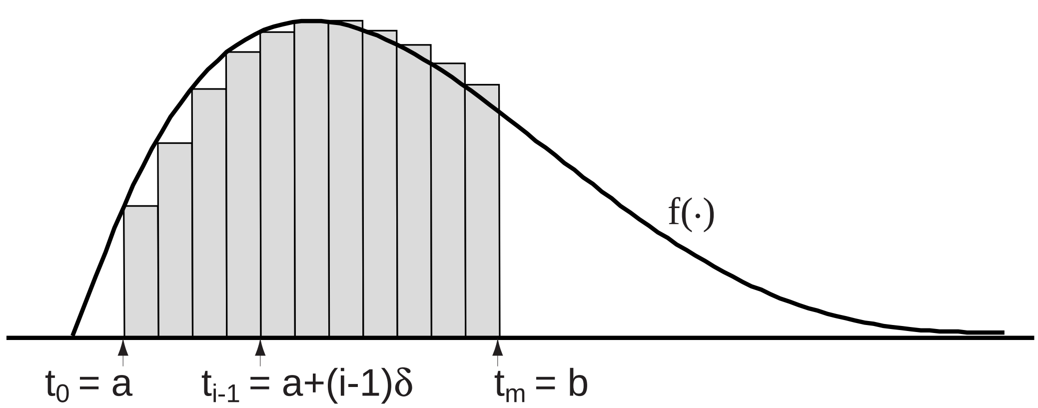
The function

<6.3> $$f(t) = \frac{n!}{(k-1)!(n-k)!}t^{k-1}(1-t)^{n-k} \qquad \text{for } 0 < t < 1$$

•density function

is called the DENSITY FUNCTION for the distribution ofT_k.

Calculate the probability that T_k lies in a longer interval $[a, b]$ by breaking the interval into many short pieces. For m a large integer, let $I_1, \ldots, I_m$ be the disjoint subintervals with lengths $\delta = (b-a)/m$ and left end points $t_{i-1} = a + (i-1)\delta$.



From <6.2>,

$$P\{T_k \in I_i\} = f(t_i)\delta + \text{ terms of order } \delta^2 \text{ or smaller}$$

Sum over the subintervals.

$$\mathbb{P}\{T_k \in [a,b]\} = \delta \sum_{i=1}^{m} f(t_i) + \text{ remainder of order } \delta \text{ or smaller.}$$

Notice how m contributions of order δ^2 (or smaller) can amount to a remainder of order at worst δ (or smaller), because m increases like $1/\delta$. (Can you make this argument rigorous?)

The sum $\delta \sum_{i=1}^{m} f(t_i)$ is an approximation to the integral of f over $[a,b]$. As δ tends to zero, the sum converges to that integral. The remainder terms tend to zero with δ. The left-hand side just sits there. In the limit we get

$$\mathbb{P}\{T_k \in [a,b]\} = \int_a^b f(t)\,dt,$$

where f denotes the density function from <6.3>.

<6.4> **Definition.** *A random variable Y is said to have a* CONTINUOUS DISTRIBUTION *with* DENSITY FUNCTION $g(\cdot)$ *if*

$$\mathbb{P}\{a \le Y <\le b\} = \int_a^b g(t)\,dt$$

for all intervals $[a,b]$. In particular (at least at points t where $g(\cdot)$ is continuous)

$$\mathbb{P}\{t \le Y \le t+\delta\} = g(t)\delta + \text{ terms of order } \delta \text{ or smaller.}$$

□

Notice that g must be non-negative, for otherwise some tiny interval would receive a negative probability. Also

$$1 = \mathbb{P}\{-\infty < Y < \infty\} = \int_{-\infty}^{\infty} g(t)\,dt.$$

In particular, for the density of the T_k distribution,

$$1 = \int_{-\infty}^{0} 0dt + \frac{n!}{(k-1)!(n-k)!}\int_0^1 t^{k-1}(1-t)^{n-k}dt + \int_1^{\infty} 0dt$$

That is,

<6.5> $$\int_0^1 t^{k-1}(1-t)^{n-k}dt = \frac{(k-1)!(n-k)!}{n!},$$

a fact that you might try to prove by direct calculation.

REMARK:I prefer to think of densities as being defined on the whole real line, with values outside the range of the random variable being handled by setting the density function equal to zero appropriately. That way my integrals always run over the whole line with the zero density killing off unwanted contributions. This convention will be useful when we consider densities that vanish outside a range depending on a parameter of the distribution; it will also help us avoid some amusing calculus blunders.

•beta function

The T_k distribution is a member of a family whose name is derived from the BETA FUNCTION, defined by

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt \qquad \text{for } \alpha > 0, \beta > 0.$$

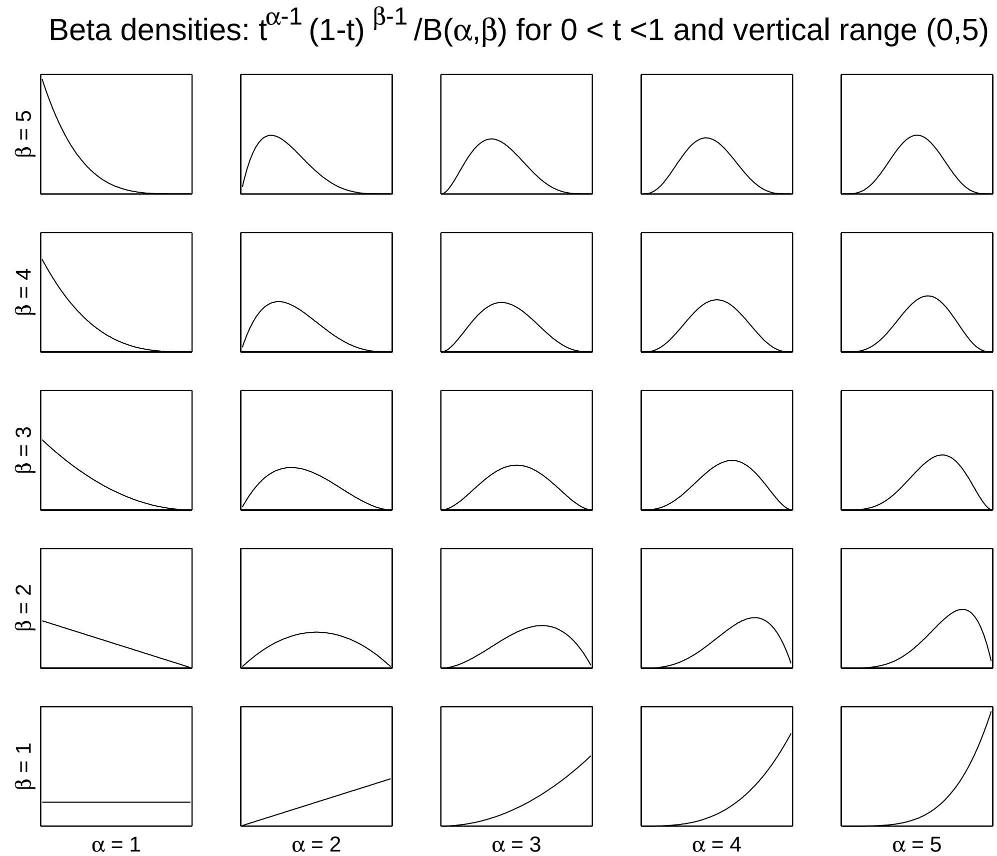
Equality <6.5> gives the value for $B(k, n-k+1)$. If we divide through by $B(\alpha,\beta)$ we get a candidate for a density function: non-negative and integrating to 1.

<6.6> **Definition.** *For* $\alpha > 0$ *and* $\beta > 0$ *the* Beta(α, β) *distribution is defined by the density function*

$$\frac{t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha, \beta)} \qquad \text{for } 0 < t < 1$$

The density is zero outside $(0, 1)$. □

For example, T_k has Beta$(k, n-k+1)$ distribution.



See the Matlab m-file drawbeta.m for the calculations used to draw all these density functions.

A Matlab digression

The function *beta* in Matlab calculates the beta function, defined for $z > 0$ and $w > 0$ by

$$\text{beta}(z, w) = \int_0^1 t^{z-1}(1-t)^{w-1}dt.$$

The function *betainc* in Matlab calculates the incomplete beta function, defined by

$$\text{betainc}(x, z, w) = \int_0^x \frac{t^{z-1}(1-t)^{w-1}}{\text{beta}(z, w)}dt \qquad \text{for } 0 \le x \le 1.$$

Expectation of a random variable with a continuous distribution

Consider a random variable X whose distribution has density function $f(\cdot)$. Let $W = g(X)$ be a new random variable defined as a function of X. How can we calculate $\mathbb{E}W$?

Construct discrete approximations to g by cutting the line into small intervals of length δ. Define

$$g'_{i,\delta} = \max\{g(x) : i\delta \le x < (i+1)\delta\}$$
$$g''_{i,\delta} = \min\{g(x) : i\delta \le x < (i+1)\delta\}$$

Near points where g is continuous, $g'_{i,\delta} \approx g(i\delta) \approx g''_{i,\delta}$. Define upper ($W'_\delta$) and lower ($W''_\delta$) approximations to $g(X)$ by putting $W'_\delta = g'_{i,\delta}$ and $W''_\delta = g''_{i,\delta}$ if $i\delta \le X < (i+1)\delta$, for $i = 0, \pm 1, \pm 2, \ldots$.

Notice that $W''_\delta \le g(X) \le W'_\delta$ always. Rule E3 for expectations (see Chapter 2) therefore gives

$$\mathbb{E}W''_\delta \le \mathbb{E}g(X) \le \mathbb{E}W'_\delta$$

Both W'_δ and W''_δ have discrete distributions, for which we can find expected values by the conditioning rule E4:

$$\begin{aligned}\mathbb{E}W'_\delta &= \sum_i \mathbb{P}\{i\delta \le X < (i+1)\delta\}\mathbb{E}(W'_\delta \mid i\delta \le X < (i+1)\delta) \\ &\approx \sum_i \delta f(i\delta) g(i\delta)\end{aligned}$$

You should recognize the last sum as an approximation to the integral $\int_{-\infty}^{\infty} g(x) f(x)\,dx$. The expectation $\mathbb{E}W''_\delta$ approximates the same integral.

As $\delta \to 0$, both approximating sums converge to the integral. The fixed value $\mathbb{E}W$ gets sandwiched between converging upper and lower bounds. In the limit we must have

<6.7> $$\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x) f(x)\,dx$$

when X has a continuous distribution with density $f(\cdot)$.

Compare with the formula for a random variable X^* taking only a discrete set of values $x_1, x_2, \ldots$:

$$\mathbb{E}g(X^*) = \sum_i g(x_i)\mathbb{P}\{X^* = x_i\}$$

In the passage from discrete to continuous distributions, discrete probabilities get replaced by densities and sums get replaced by integrals.

> You should be very careful not to confuse the formulae for expectations in the discrete and continuous cases. Think again if you find yourself integrating probabilities or summing expressions involving probability densities.