1 R Functions for Probability Distributions

2 Review of Random Variables Concepts

2.1 Variables

Before we tackle random variables, it is best to be sure we are clear about the notion of a mathematical variable. A variable is a symbol that stands for an unspecified mathematical object, like x in the expression $x^2 + 2x + 1$.

Often, it is clear from the context what kind of object the variable stands for. In this example, x can be any real number. But not all variables are numerical. We will also use vector variables and variables taking values in arbitrary sets. Thus, when being fussy, we specify the kind of mathematical objects a variable can symbolize. We do this by specifying the set of objects which are possible values of the variable. For example, we write

$$x^2 + 2x + 1 = (x + 1)^2, x \in \mathbb{R} \tag{1}$$

to show that the equality holds for any real number x, the symbol $\mathbb{R}$ indicating the set of all real numbers.

2.2 Functions

A function is a rule f that assigns to each element x of a set called the domain of the function an object $f(x)$ called the value of the function at x. Note the distinction between the function f and the value $f(x)$. There is also a distinction between a function and an expression defining the function. We say, let f be the function defined by

$$f(x) = x^2, x \in \mathbb{R} \tag{2}$$

Strictly speaking, (2) isn't a function, it's an expression defining the function f. Neither is x^2 the function, it's the value of the function at the point x. The function f is the rule that assigns to each x in the domain, which from (2) is the set $\mathbb{R}$ of all real numbers, the value $f(x) = x^2$.

2.3 Random Variables: Informal Intuition

Informally, a random variable is a variable that is random, meaning that its value is unknown, uncertain, not observed yet, or something of the sort. The probabilities with which a random variable takes its various possible values are described by a probability model.

In order to distinguish random variables from ordinary, nonrandom variables, we adopt a widely used convention of denoting random variables by capital letters, usually letters near the end of the alphabet, like X, Y, and Z.

There is a close connection between random variables and certain ordinary variables. If X is a random variable, we often use the corresponding small letter x as the ordinary variable that takes the same values.

Whether a variable corresponding to a real-world phenomenon is considered random may depend on context. In applications, we often say a variable is random before it is observed and nonrandom after it is observed and its actual value is known. Thus the same real-world phenomenon may be symbolized by X before its value is observed and by x after its value is observed.

2.4 Random Variables: Formal Definition

The formal definition of a random variable is rather different from the informal intuition. Formally, a random variable isn't a variable, it's a function.

Definition 1 *(**Random Variable**). A random variable in a probability model is a function on the sample space of a probability model.*

The capital letter convention for random variables is used here too. We usually denote random variables by capital letters like X. When considered formally a random variable X a function on the sample space S, and we can write

$$S \xrightarrow{X} T \tag{3}$$

if we like to show that X is a map from its domain S (always the sample space) to its codomain T. since X is a function, its values are denoted using the usual notation for function values $X(s)$.

3 Review of Probability Distributions Concepts

3.1 Probability Mass Functions

A probability mass function (PMF) is a function

$$S \xrightarrow{f} \mathbb{R} \tag{4}$$

whose domain S, which can be any nonempty set, is called the sample space, whose codomain is the real numbers, and which satisfies the following conditions: its values are nonnegative

$$f(x) \geq 0, x \in S \tag{5}$$

and sum to one

$$\sum_{x \in S} f(x) = 1 \tag{6}$$

3.2 Probability Density Functions

A real-valued function f defined on an interval (a, b) of the real numbers is called a probability density function (PDF) if

$$f(x) \geq 0, a < x < b \tag{7}$$

and

$$\int_a^b f(x)dx = 1 \tag{8}$$

The values $a = -\infty$ or $b = +\infty$ are allowed for endpoints of the interval.

Remark 1 *A PDF is just like a PMF except that we integrate rather than sum*

A real-valued function f defined on a region S of $\mathbb{R}^2$ is also called a PDF if

$$f(x_1, x_2) \geq 0, \quad (x_1, x_2) \in S \tag{9}$$

and

$$\iint_S f(x_1, x_2)\, dx_1 dx_2 = 1 \tag{10}$$

A real-valued function f defined on a region S of $\mathbb{R}^n$ is also called a PDF if

$$f(\mathbf{x}) \geq 0, \mathbf{x} \in S \tag{11}$$

and

$$\int_S f(\mathbf{x}) d\mathbf{x} = 1 \tag{12}$$

Here only the boldface indicates that $\mathbf{x}$ is a vector and hence we are dealing with a multiple integral (n -dimensional).

3.3 Discrete and Continuous

If X is a random variable or X is a random vector whose dis- tribution is described by a PMF, we say the distribution or the random variable or vector is discrete.

If X is a random variable or X is a random vector whose dis- tribution is described by a PDF, we say the distribution or the random variable or vector is continuous.

3.4 Quantiles

If X is a random variable and $0 < q < 1$, then a q -th quantile of X (or of the distribution of X) is any number x such that

$$\Pr(X \leq x) \geq q \text{ and } \Pr(X \geq x) \geq 1 - q \tag{13}$$

If X is a continuous random variable having distribution function F, then this simplifies to

$$F(x) = q \tag{14}$$

If X is a discrete random variable having distribution function F, then this simplifies to

$$F(y) \leq q \leq F(x), y < x \tag{15}$$

Example 1 The distribution function of the binomial $\mathrm{Bin}(3, 1/2)$ distribution is

$$F(x) = \begin{cases} 0, & x < 0 \\ 1/8, & 0 \leq x < 1 \\ 1/2, & 1 \leq x < 2 \\ 7/8, & 2 \leq x < 3 \\ 1, & 3 \leq x \end{cases} \tag{16}$$

and 0 is the unique q -th quantile for $0 < q < 1/8$ x is a $1/8$ -th quantile for $0 \leq x \leq 1$ 1 is the unique q -th quantile for $1/8 < q < 1/2$ x is a $1/2$ -th quantile for $1 \leq x \leq 2$ and so forth.

Example 2 For a continuous random variable whose support is an interval, quantiles are uniquely defined.

Suppose X is an $\mathrm{Exp}(\lambda)$ random variable, then the q -th quantile is the solution for x of

$$F(x) = 1 - e^{-\lambda x} = q \tag{17}$$

which has the solution

$$x = -\frac{1}{\lambda} \cdot \log(1 - q) \tag{18}$$

Remark 2 *Certain special quantiles have other names.*

The $1/2$ -th quantile is also called the median.

The $1/4$ -th and $3/4$ -th quantile are also called the lower quartile and upper quartile, respectively.

If $100q$ is an integer, the q -th quantile is also called the $100q$ -th percentile. (This terminology is frowned upon in serious probability theory, just like all use of percentages.)

Remark 3 *Although quantiles are not necessarily unique, it is mathematically convenient to define a function that gives for each q a q -th quantile.*

If F is a distribution function, the corresponding quantile function G is

$$G(q) = \inf\{x \in \mathbb{R} : F(x) \geq q\}, 0 < q < 1 \tag{19}$$

If the q -th quantile is unique, then it is $G(q)$. If the q -th quantile in non-unique, then $G(q)$ is the smallest q -th quantile.

3.4.1 Summary of Quantiles Continuous Random Variables

If F is the DF of a continuous random variable, then the p -th quantile of this random variable or this distribution is any x satisfying

$$F(x) = p \tag{20}$$

and the solution is unique if F is differentiable at x with $F^{\prime}(x)>0.

The solution is non-unique only if F has a flat section with value p, that is if there exist points a and b with $a < b$ such that

$$F(x) = p, a \leq x \leq b \tag{21}$$

in which case every $x \in [a, b]$ is a p -th quantile.

3.4.2 Summary of Quantiles Discrete Random Variables

If F is the DF of a discrete random variable, then the p -th quantile of this random variable or this distribution is any x satisfying

$$F(y) \leq p \leq F(x), \text{ for all } y < x \tag{22}$$

There are two cases. The solution is unique if we do not have $F(x) = p$ for any x. In that case the p -th quantile is the point x where F jumps past p, that is

$$F(y) < p < F(x), y < x \tag{23}$$

The solution is non-unique if we do have $F(x) = p$ for some x, in which case (because the DF of a discrete random variable is a step function) we must have $F(x) = p$ for all x in an interval, and every such x is a p -th quantile.

4 R Functions for Probability Distributions

Every distribution that R handles has four functions. There is a root name, for example, the root name for the normal distribution is norm. This root is prefixed by one of the letters

• p for "probability", the cumulative distribution function (c. d. f.)

• q for "quantile", the inverse c. d. f.

• d for "density", the density function (p. f. or p. d. f.)

• r for "random", a random variable having the specified distribution

For the normal distribution, these functions are `pnorm`, `qnorm`, `dnorm`, and `rnorm`. For the binomial distribution, these functions are pbinom, qbinom, dbinom, and rbinom. And so forth.

For a continuous distribution (like the normal), the most useful functions for doing problems involving probability calculations are the "p" and "q" functions (c. d. f. and inverse c. d. f.), because the density (p. d. f.) calculated by the "d" function can only be used to calculate probabilities via integrals and R doesn't do integrals.

For a discrete distribution (like the binomial), the "d" function calculates the density (p. f.), which in this case is a probability

$$f(x) = P(X = x) \tag{24}$$

and hence is useful in calculating probabilities.

R has functions to handle many probability distributions. The table below gives the names of the functions for each distribution and a link to the on-line documentation that is the authoritative reference for how the functions are used. But don't read the on-line documentation yet. First, try the examples in the sections following the table.

Distribution	Functions			
Beta	pbeta	qbeta	dbeta	rbeta
Binomial	pbinom	qbinom	dbinom	rbinom
Cauchy	pcauchy	qcauchy	dcauchy	rcauchy
Chi-Square	pchisq	qchisq	dchisq	rchisq
Exponential	pexp	qexp	dexp	rexp
F	pf	qf	df	rf
Gamma	pgamma	qgamma	dgamma	rgamma

Distribution	Functions			
Geometric	pgeom	qgeom	dgeom	rgeom
Hypergeometric	phyper	qhyper	dhyper	rhyper
Logistic	plogis	qlogis	dlogis	rlogis
Log Normal	plnorm	qlnorm	dlnorm	rlnorm
Negative Binomial	pnbinom	qnbinom	dnbinom	rnbinom
Normal	pnorm	qnorm	dnorm	rnorm
Poisson	ppois	qpois	dpois	rpois
Student t	pt	qt	dt	rt
Studentized Range	ptukey	qtukey	dtukey	rtukey
Uniform	punif	qunif	dunif	runif
Weibull	pweibull	qweibull	dweibull	rweibull
Wilcoxon Rank Sum Statistic	pwilcox	qwilcox	dwilcox	rwilcox
Wilcoxon Signed Rank Statistic	psignrank	qsignrank	dsignrank	rsignrank

4.1 The Normal Distribtion

`pnorm` is the R function that calculates the c. d. f.

$$F(x) = P(X <= x) \tag{25}$$

where X is normal.

Both of the R commands in the below cell do exactly the same thing.

Example 3 Compute $P(X < 27.4)$ when X is normal with mean 50 and standard deviation 20.

In [1]:
```
pnorm(27.4, mean=50, sd=20)
pnorm(27.4, 50, 20)
```

0.129238112240018

0.129238112240018

Example 4

What is $P(X > 19)$ when X has the $N(17.46, 375.67)$ distribution?

Remark 4 *R wants the standard deviation (s. d.) as the parameter, not the variance. We'll need to take a square root!*

In [2]:
```
1 - pnorm(19, mean=17.46, sd=sqrt(375.67))
```

0.468335635789911

4.2 Inverse of cdf

The inverse distribution function for continuous variables $F^{-1}(p)$ is the inverse of the cumulative distribution function $F(x)$ (CDF). In other words, it's simply the distribution function $F(x)$ inverted. The CDF shows the probability a random variable X is found at a value equal to or less than a certain x. Intuitively, it's how much area is under the curve at a certain point. The inversion of the CDF, the IDF, gives a value for x such that:

$$F(x) = Pr(X \leq x) = p \tag{26}$$

, where p is a given value.

`qnorm` is the R function that calculates the inverse c. d. f. F^{-1} of a normal distribution. The c. d. f. and the inverse c. d. f. are related by

$$p = F(x) = P(X \leq x) \tag{27}$$
$$x = F^{-1}(p) \tag{28}$$

So given a number p between zero and one, `qnorm` looks up the **p-th quantile** of the normal distribution. As with `pnorm`, optional arguments specify the mean and standard deviation of the distribution.

Example 5

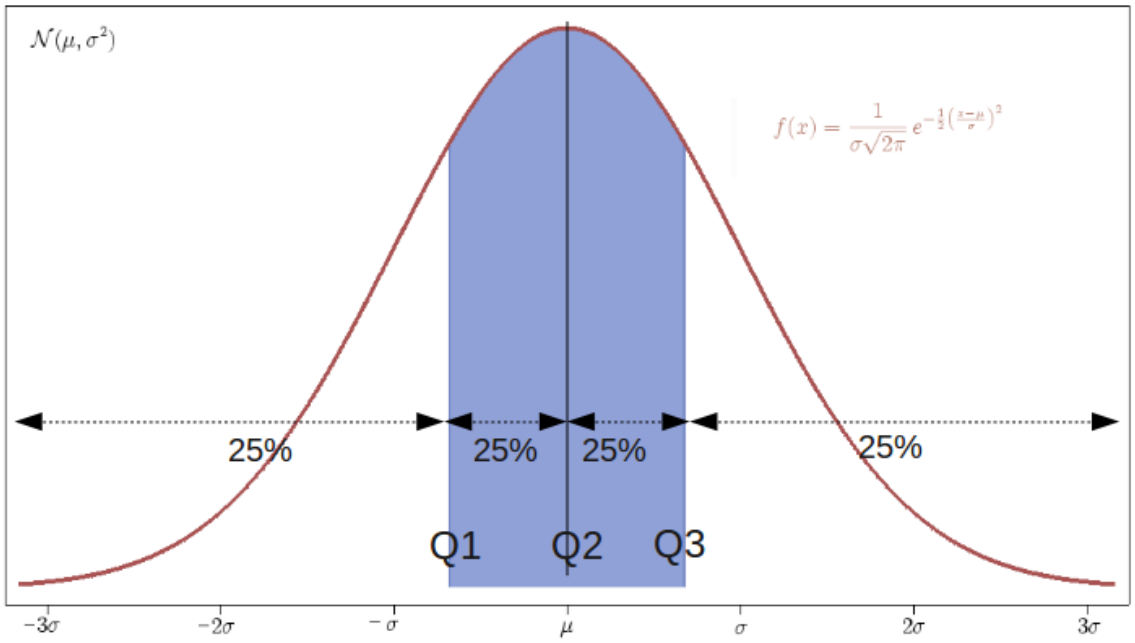
Suppose IQ scores are normally distributed with mean 100 and standard deviation 15. What is the 95th percentile of the distribution of IQ scores?

or

What is $F^{-1}(0.95)$ when X has the $N(100, 15^2)$ distribution?

Remark 5 *A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations falls. For example, the 20th percentile is the value (or score) below which 20% of the observations may be found. Equivalently, 80% of the observations are found above the 20th percentile.*

Remark 6 *In statistics and probability, quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way. There is one fewer quantile than the number of groups created. Thus quartiles are the three cut points that will divide a dataset into four equal-sized groups. Common quantiles have special names: for instance quartile, decile (creating 10 groups: see below for more). The groups created are termed halves, thirds, quarters, etc., though sometimes the terms for the quantile are used for the groups created, rather than for the cut points.*



Probability density of a normal distribution, with quantiles shown. The area below the red curve is the same in the intervals $(-\infty, Q_1)$, (Q_1, Q_2), (Q_2, Q_3), *and* $(Q_3, +\infty)$.

In [4]:
```
qnorm(0.95, mean=100, sd=15)
```

124.672804404272

4.3 Density

`dnorm` is the R function that calculates the p.d.f. f of the normal distribution. As with `pnorm` and `qnorm`, optional arguments specify the mean and standard deviation of the distribution.

There's not much need for this function in doing calculations, because you need to do integrals to use any p.d.f., and R doesn't do integrals.

In fact, there's not much use for the "d" function for any continuous distribution (discrete distributions are entirely another matter, for them the "d" functions are very useful, see the section about `dbinom`).

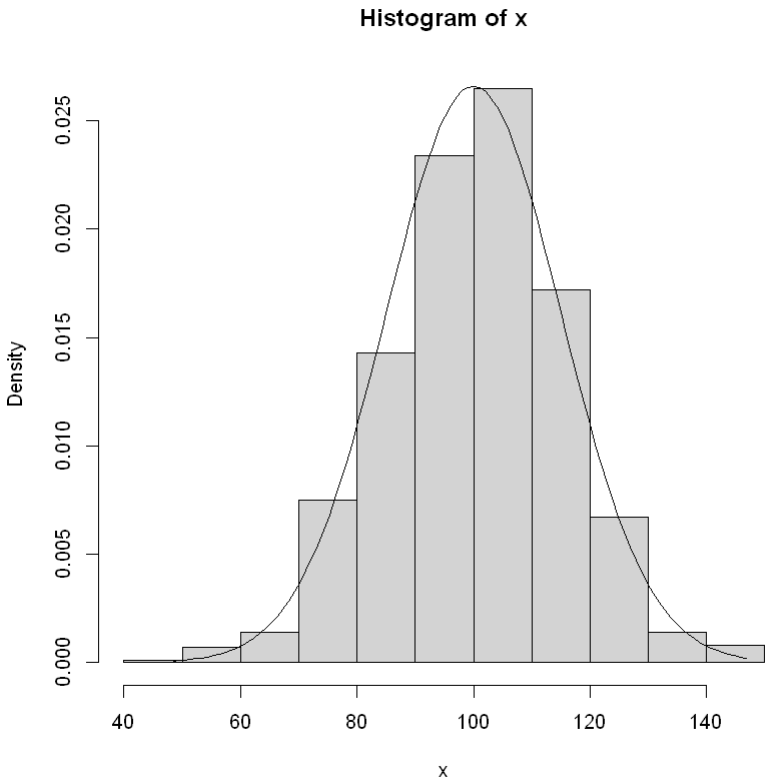
4.4 Random Variates

`rnorm` is the R function that simulates random variates having a specified normal distribution. As with `pnorm`, `qnorm`, and `dnorm`, optional arguments specify the mean and standard deviation of the distribution.

We will be using the "r" functions (such as rnorm) to generate random numbers with known distribution.

Example 6 Generate 1000 numbers from normal distribution with mean 100 and standard deviation 15

In [5]:
```
x <- rnorm(1000, mean=100, sd=15)
hist(x, probability=TRUE)
xx <- seq(min(x), max(x), length=100)
lines(xx, dnorm(xx, mean=100, sd=15))
```



This generates 1000 Independent and Identically Distributed (i.i.d). normal random numbers (first line), plots their histogram (second line), and graphs the p. d. f. of the same normal distribution (third and forth lines).

4.5 The Binomial Distribtion

`dbinom` is the R function that calculates the p.f. of the binomial distribution. Optional arguments d specify the parameters of the particular binomial distribution.

Both of the R commands in the box below do exactly the same thing.

In [11]:
```
dbinom(27, size=100, prob=0.25)
dbinom(27, 100, 0.25)
```

0.0806407548759012

0.0806407548759012

They look up $P(X = 27)$ when X is has the Bin(100, 0.25) distribution.

Example 7 What is $P(X = 1)$ when X has the $Bin(25, 0.005)$ distribution?

In [12]:
```
dbinom(1, 25, 0.005)
```

0.110831688812663

4.6 Direct Look-Up, Intervals

`pbinom` is the R function that calculates the c. d. f. of the binomial distribution. Optional arguments described on documentation (help(pdinom) specify the parameters of the particular binomial distribution.

Both of the R commands in the cell below do exactly the same thing.

Example 8 Look up $P(X <= 27)$ when X is has the $Bin(100, 0.25)$ distribution. (Note the less than or equal to sign. It's important when working with a discrete distribution!)

In [9]:
```
pbinom(27, size=100, prob=0.25)
pbinom(27, 100, 0.25)
```

0.722380513115339

0.722380513115339

Exercise 1 What is P(X <= 1) when X has the Bin(25, 0.005) distribution?

In [14]:
```
pbinom(1,25,0.0005)
```

0.999925572634752

4.7 Inverse Look-Up

`qbinom` is the R function that calculates the "inverse c. d. f." of the binomial distribution. How does it do that when the c. d. f. is a step function and hence not invertible? The on-line documentation for the binomial probability functions explains.

In [15]:
```
help(qbinom)
```

> The quantile is defined as the smallest value x such that $F(x) >= p$, where F is the distribution function.

When the p-th quantile is nonunique, there is a whole interval of values each of which is a p-th quantile. The documentation says that qbinom (and other "q" functions, for that matter) returns the smallest of these values. That is one sensible definition of an "inverse c. d. f." In the terminology of Section of the course notes, the function defined by qbinom is a right inverse of the function defined by pbinom, that is,

$q == pbinom(qbinom(q, n, p)), 0 < q < 1, 0 < p < 1, n$ a positive integer

is always true, but the analogous formula with `pnorm` and `qnorm` reversed does not necessarily hold.

Example 9 Question: What are the 10th, 20th, and so forth quantiles of the Bin(10, 1/3) distribution?

In [17]:
```
qbinom(0.1, 10, 1/3)
qbinom(0.2, 10, 1/3)
# and so forth, or all at once with
qbinom(seq(0.1, 0.9, 0.1), 10, 1/3)
#Note the nonuniqueness.
```

```
1

2

1 2 3 3 3 4 4 5 5
```

5 Learning R (online book)

http://lib.stat.cmu.edu/R/CRAN/doc/manuals/r-release/R-intro.html (http://lib.stat.cmu.edu/R/CRAN/doc/manuals/r-release/R-intro.html)

6 Statistical Analysis on the Web with R ¶

6.1 Rweb

https://rweb.webapps.cla.umn.edu/Rweb/ (https://rweb.webapps.cla.umn.edu/Rweb/)

6.2 rdrr.io

https://rdrr.io/snippets/ (https://rdrr.io/snippets/)