Probability Distributions in R

Contents

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 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

$f(x) = P(X = x)$

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`
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`

Warning: The parameters of these distributions may not agree with textbooks. In particular, the second parameter in the gamma distribution is the reciprocal of the second parameter in our textbook (*beta* = 1 / *lambda*).

That's a lot of distributions. Fortunately, they all work the same way. If you learn one, you've learned them all.

Of course, the discrete distributions are discrete and the continuous distributions are continuous, so there's some difference just from that aspect alone, but as far as the computer is concerned, they're all the same. We'll do a continuous example first.

The Normal Distribtion

Direct Look-Up

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

$F(x) = P(X <= x)$

where X is normal. Optional arguments described on the on-line documentation specify the parameters of the particular normal distribution.

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

```
pnorm(27.4, mean=50, sd=20)
```

```
pnorm(27.4, 50, 20)
```

They look up $P(X < 27.4)$ when X is normal with mean 50 and standard deviation 20.

Example

Question: Suppose widgit weights produced at Acme Widgit Works have weights that are normally distributed with mean 17.46 grams and variance 375.67 grams. What is the probability that a randomly chosen widgit weighs more then 19 grams?

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

Caution: R wants the s. d. as the parameter, not the variance. We'll need to take a square root!

Answer:

```
pnorm(19, mean=17.46, sd=sqrt(375.67))
```

Inverse Look-Up

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

$p = F(x)$
$x = F^{-1}(p)$

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

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

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

Answer:

```
qnorm(0.95, mean=100, sd=15)
```

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).

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 won't be using the "r" functions (such as `rnorm`) much. So here we will only give an example without full explanation.

```
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 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).

The Binomial Distribtion

Direct Look-Up, Points

`dbinom` is the R function that calculates the p. f. of the binomial distribution. Optional arguments described on the on-line documentation specify the parameters of the particular binomial distribution.

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

```
dbinom(27, size=100, prob=0.25)
```

```
dbinom(27, 100, 0.25)
```

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

Example

Question: Suppose widgits produced at Acme Widgit Works have probability 0.005 of being defective. Suppose widgits are shipped in cartons containing 25 widgits. What is the probability that a randomly chosen carton contains exactly one defective widgit?

Question Rephrased: What is $P(X = 1)$ when X has the Bin(25, 0.005) distribution?

Answer:

```
dbinom(1, 25, 0.005)
```

Direct Look-Up, Intervals

`pbinom` is the R function that calculates the c. d. f. of the binomial distribution. Optional arguments described on the on-line documentation specify the parameters of the particular binomial distribution.

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

```
pbinom(27, size=100, prob=0.25)

pbinom(27, 100, 0.25)
```

They 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!)

Example

Question: Suppose widgits produced at Acme Widgit Works have probability 0.005 of being defective. Suppose widgits are shipped in cartons containing 25 widgits. What is the probability that a randomly chosen carton contains no more than one defective widgit?

Question Rephrased: What is $P(X <= 1)$ when X has the Bin(25, 0.005) distribution?

Answer:

```
pbinom(1, 25, 0.005)
```

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.

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

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

Answer:

```
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 non-uniqueness.