Chapter 4: Continuous Random Variables and Probability Distributions

Curtis Miller

2018-06-13

Introduction

Continuous probability models are the other major class of probability models. In addition to extending our probabilistic framework to continuous phenomena (namely, measurements), the Normal[1] distribution is both a continuous distribution and arguably the most important distribution in statistics and probability theory, due to its role in the central limit theorem. Many of the concepts we covered for discrete random variables carry over to the continuous case, including pmfs (although they become density functions rather than mass functions), cdfs, and expectations. In fact, the continuous case may be slightly easier than the discrete case since $\mathbb{P}(X = c) = 0$ for all $c \in \mathbb{R}$ and $\mathbb{P}(X < x) = \mathbb{P}(X \leq x)$.

[1] Another name for the Normal distribution is the Gaussian distribution, named after the great mathematician Carl Friedrich Gauss. No one is sure where the name "Normal" came from, but some theorize that the distribution attracted so much attention authors began to refer to it as the "typical" distribution, although most natural phenomena doesn't follow a Normal distribution. Thus I capitalize the word "Normal" to refer to a particular distribution but as a reminder that the distribution doesn't automatically describe a phenomenon.

Section 1: Probability Density Functions

The analogue to the probability mass function seen for discrete random variables is the **probability density function (pdf)**. The pdf is a non-negative function $f(x)$ such that, for any two numbers a and b with $a \leq b$

In order for f to be a valid pdf we must also have

Example 1

Confirm that the function

$$f(x; a, b) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

is a valid pdf. Then, plot the pdf. A random variable U following this distribution is said to follow the uniform distribution, denoted

by $U \sim \text{UNIF}(a, b)$.

Example 2

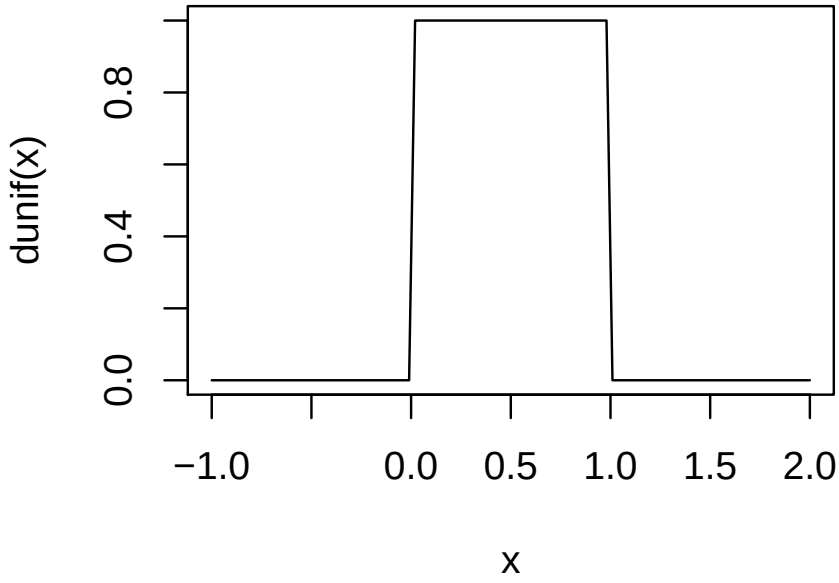
Confirm that the function

$$f(x; \mu) = \begin{cases} \frac{1}{\mu} e^{-\frac{1}{\mu}x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

is a valid pdf. Then, plot the pdf. A random variable X following this distribution is said to follow the exponential distribution, denoted by $X \sim \text{EXP}(\mu)$[2].

[2] This notation is *not* standard and depends ultimately on who is writing the document. It turns out that μ is the mean of the exponential random variable when specified this way, but an alternative specification uses the rate $\lambda = \frac{1}{\mu}$. While the rate is often easier to work with mathematically, statisticians usually are interested in the mean. As a result, probabilists usually specify exponential random variables using the rate and write $X \sim \text{EXP}(\lambda)$ while statisticians prefer to specify exponential random variables using the mean. I do the latter as this is a statistics course, but be aware of the controversy.

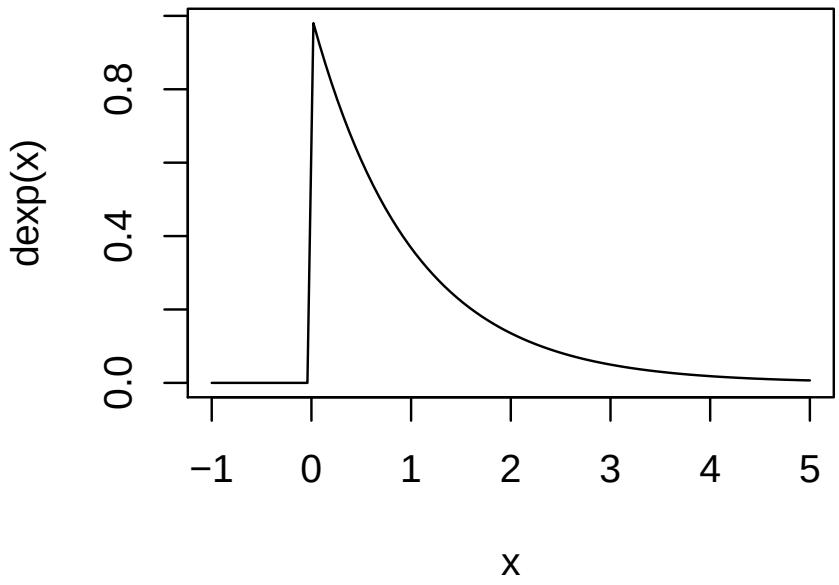
```r
# UNIF(0,1)
curve(dunif, -1, 2)      # Plot the pdf
```




```r
integrate(dunif, -1, 2)  # Integrage (numerically) the pdf to see it is one
```

```
## 1 with absolute error < 1.1e-15
```

```r
# EXP(1)
curve(dexp, -1, 5)
```




```r
integrate(dexp, 0, Inf)
```

```
## 1 with absolute error < 5.7e-05
```

Example 3

Accidents along a certain stretch of road are presumed to occur a distance of X miles from the nearest city center, where $X \sim \text{UNIF}(100, 150)$. Compute

1. $\mathbb{P}(110 \leq X \leq 130)$

2. $\mathbb{P}(127 < X \leq 144)$

3. $\mathbb{P}(X > 148)$

```r
integrate(dunif, 110, 130, min = 100, max = 150)  # 1
```

```
## 0.4 with absolute error < 4.4e-15
```

```r
integrate(dunif, 127, 144, min = 100, max = 150)  # 2
```

```
## 0.34 with absolute error < 3.8e-15
```

```r
integrate(dunif, 148, Inf, min = 100, max = 150)  # 3
```

```
## 0.03999993 with absolute error < 0.00011
```

Example 4

The time (in minutes) taken by a worker at the Tuition and Financial Aid office of a certain university to service a student follows an exponential distribution with $T \sim \text{Exp}(10)$. Compute the following:

1. $\mathbb{P}(T < 20)$

2. $\mathbb{P}(6 < T < 9)$

3. $\mathbb{P}(T \geq 22)$

```
integrate(dexp, -Inf, 20, rate = 1/10)  # 1
```

```
## 0.8646644 with absolute error < 3.8e-05
```

```
integrate(dexp, 6, 9, rate = 1/10)      # 2
```

```
## 0.142242 with absolute error < 1.6e-15
```

```
integrate(dexp, 22, Inf, rate = 1/10)   # 3
```

```
## 0.1108032 with absolute error < 1.3e-05
```

Section 2: Cumulative Distribution Functions and Expected Values

The cdf of a continuous random variable is

Thanks to the fundamental theorem of calculus we have the following relationship between the pdf and cdf of a random variable:

Rules for using the cdf to compute the probability of a continuous random variable taking values in an interval are given below.

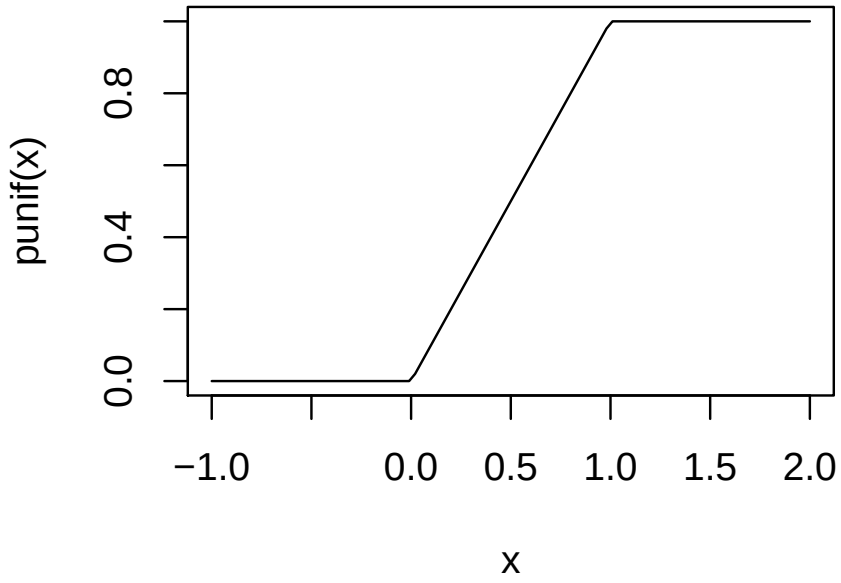
Example 5

Compute the cdf of $X \sim \text{UNIF}(a, b)$ and plot it.

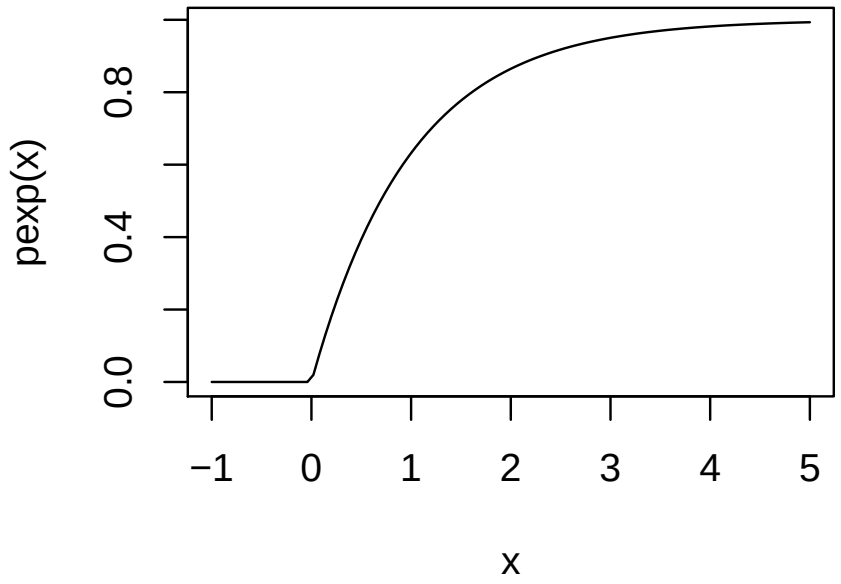
Example 6

Compute the cdf of $X \sim \text{EXP}(\mu)$ and plot it.

```r
curve(punif, -1, 2)  # CDF of UNIF(0, 1)
```




```r
curve(pexp, -1, 5)   # CDF of EXP(1)
```



Example 7

Answer the questions posed in Example 3 and Example 4 but using the cdf of the respective random variables.

```r
# Example 3
punif(130, min = 100, max = 150) - punif(110, min = 100, max = 150)  # 1
```

```
## [1] 0.4
```

```r
punif(144, min = 100, max = 150) - punif(127, min = 100, max = 150)  # 2
```

```
## [1] 0.34
```

```r
1 - punif(148, min = 100, max = 150)                                 # 3
```

```
## [1] 0.04
```

```r
# Example 4
pexp(20, rate = 1/10)                        # 1
```

```
## [1] 0.8646647
```

```r
pexp(9, rate = 1/10) - pexp(6, rate = 1/10)  # 2
```

```
## [1] 0.142242
```

```r
1 - pexp(22, rate = 1/10)                    # 3
```

```
## [1] 0.1108032
```

The $100p$**th percentile** (also referred to as quantiles) of a distribution is the number $\eta(p)$ such that $F(\eta(p)) = p$. If F can be inverted over its support, we can use F^{-1} to find percentiles.

A particularly interesting percentile is the 50th percentile, otherwise known as the **median**, $\tilde{\mu}$.

Example 8

Find percentile functions for the uniform and exponential distributions. Then find $\eta(0.5)$.

```r
# Example for UNIF(0,1) and EXP(1)

qunif(0.5)

## [1] 0.5

qexp(0.5)

## [1] 0.6931472
```

Below are formulas for $\mathbb{E}[X]$, $\mathbb{E}[h(X)]$, and $\text{Var}(X)$ in the continuous case.

The shortcut formula for the variance in the discrete case also holds in the continuous case.

Proposition 1.

$$\text{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$$

Example 9

Compute $\mathbb{E}[X]$ and $\text{Var}(X)$ for uniform and exponential random variables.

```
(mu1 <- integrate(function(x) {x * dunif(x, 0, 1)}, -1, 2))  # Mean of UNIF(0,1)

## 0.5 with absolute error < 5.6e-16

integrate(function(x) {(x - mu1$value)^2 * dunif(x, 0, 1)},  # Var of UNIF(0,1)
          -1, 2)

## 0.08333333 with absolute error < 8.6e-05

(mu2 <- integrate(function(x) {x * dexp(x)}, 0, Inf))         # Mean of EXP(1)

## 1 with absolute error < 6.4e-06

integrate(function(x) {(x - mu2$value)^2 * dexp(x)}, 0, Inf)  # Var of EXP(1)

## 1 with absolute error < 5.8e-05
```

Section 3: The Normal Distribution

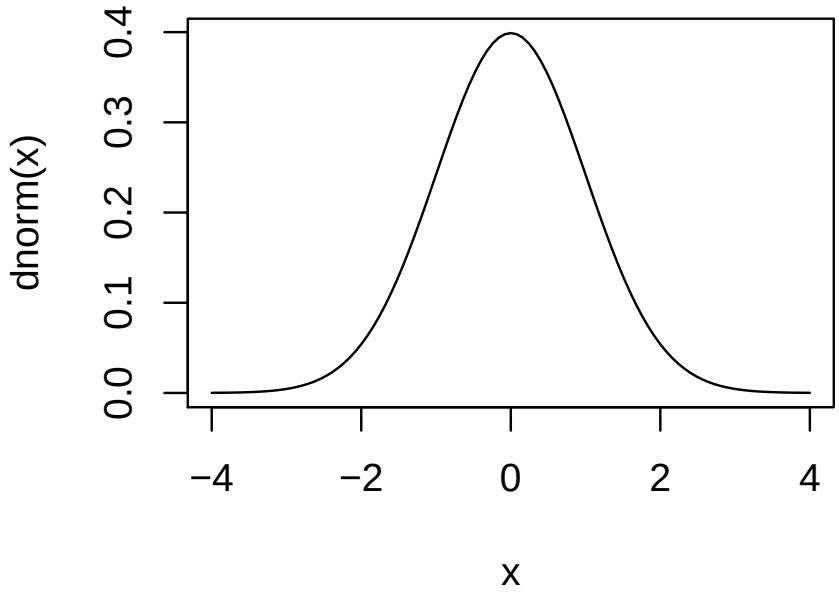
We say that a random variable X follows the **Normal distribution**[3], or $X \sim N(\mu, \sigma)$[4], if it has the pdf:

Below is a sketch of the density curve for the Normal distribution:

[3] Of all the probability distributions, the Normal distribution is arguably the most important. It plays a prominent role in one of the key theorems of probability, the central limit theorem, and as a result many random variables start to resemble Normally distributed random variables under certain conditions; we will see examples in this section. It is a well-behaved distribution; while any real number could be generated by the Normal distribution, it is effectively supported on the interval $[\mu - 3\sigma, \mu + 3\sigma]$. It naturally describes phenomena we would say results from an error process. That said, not everything is Normally distributed. Stock price movements, for example, are modeled with the Normal distribution yet we see fluctuations that would never be seen in billions of years if the Normal distribution were actually the appropriate distribution.

[4] Frequently the Normal distribution is specified with σ^2 instead of σ. In this class we use σ, but be aware that in academic settings it may be more common to see the Normal distribution using σ^2 instead. This is because the math is generally easier when using σ^2 and the notation extends well to multivariate or even functional cases.

```r
curve(dnorm, -4, 4)  # Plot of the density curve for N(0,1)
```



$\mathbb{E}[X]$, $\text{Var}(X)$, and $\text{SD}(X)$ are given below.

One property of the Normal distribution is the **68-95-99.7 rule**:

If $Z \sim N(0,1)$, we say that Z follows the **standard Normal distribution**. This distribution is useful since we can relate $X \sim N(\mu, \sigma)$ to the standard Normal distribution, and vice versa:

Let $\Phi(z) = \mathbb{P}(Z \leq z)$ be the cdf of the standard Normal distribution. Then if $F(x) = \mathbb{P}(X \leq x)$, we have the following relationship between F and Φ:

This means that we only need to worry about tabulating values for $\Phi(z)$[5] for working with any Normal distribution, as done in Table A.3.

Example 10

Compute the following:

1. $\mathbb{P}(Z \leq 0)$

2. $\mathbb{P}(Z \leq 1.23)$

[5] Notice that

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$$

As mentioned, there is no closed form solution to this integral, but that is not a problem. Numerical methods can easily compute these quantities and they can then be tabulated. On a more general note, we encounter integrals without closed form solutions all the time, yet the functions they represent can still be very well-behaved, so there is no problem leaving the integral in the expression of the quantity; we know the integral exists, we can evaluate it numerically, and we can even talk about its properties. Not every integral needs to be like the ones seen in the calculus sequence of classes.

```r
pnorm(0)     # 1
```

```
## [1] 0.5
```

```r
pnorm(1.23)  # 2
```

```
## [1] 0.8906514
```

3. $\mathbb{P}\left(-1.97 \leq Z \leq 2.1\right)$

4. $\mathbb{P}\left(Z \geq 1.8\right)$

5. $\mathbb{P}\left(Z > 5.2\right)$

```r
pnorm(2.1) - pnorm(-1.97)       # 3
```

```
## [1] 0.9577164
```

```r
1 - pnorm(1.8)                  # 4
```

```
## [1] 0.03593032
```

```r
pnorm(5.2, lower.tail = FALSE)  # 5
```

```
## [1] 9.964426e-08
```

Example 11

IQ scores are said to be Normally distributed with mean 100 and standard deviation 15. Let Q be a randomly selected individual's IQ score. Compute the following:

1. $\mathbb{P}(85 \leq Q \leq 115)$

2. $\mathbb{P}(Q > 90)$

3. The International Society for Philosophical Enquiry requires potential members to have an IQ of at least 135 in order to join the society. Based on this, what proportion of the population is eligible for membership?

```r
pnorm(115, mean = 100, sd = 15) - pnorm(85, mean = 100, sd = 15)  # 1
```

```
## [1] 0.6826895
```

```r
pnorm(90, mean = 100, sd = 15, lower.tail = FALSE)                # 2
```

```
## [1] 0.7475075
```

```r
pnorm(135, mean = 100, sd = 15, lower.tail = FALSE)               # 3
```

```
## [1] 0.009815329
```

Here the notation z_α is used to mean $\Phi(z_\alpha) = 1 - \alpha$. We can relate this back to general $\eta(p)$, defined for an arbitrary Normally distributed random variable.

$z_{1-\alpha}$ can be found using Table A.3 using a reverse lookup.

Example 12

1. What is $z_{0.5}$?

2. What is $z_{0.05}$?

3. What are the first and third quartiles of the standard Normal distribution?

```r
qnorm(0.5)                      # 1
```

```
## [1] 0
```

```r
qnorm(0.05, lower.tail = FALSE)  # 2
```

```
## [1] 1.644854
```

```r
qnorm(0.25)                     # 3a
```

```
## [1] -0.6744898
```

```r
qnorm(0.75)                     # 3b
```

```
## [1] 0.6744898
```

Example 13

Using the description of the random variable Q from Example 11, answer the following questions.

1. Mensa International require individuals have an IQ score that would place them in the top 2% of the population. What is the minimum IQ score needed to be a member of Mensa?

2. The part of the population with the lowest 5% of IQ scores is considered to be intellectually disabled. What is the highest IQ score needed to be in this group?

```r
qnorm(0.02, mean = 100, sd = 15, lower.tail = FALSE)  # 1
```

```
## [1] 130.8062
```

```r
qnorm(0.05, mean = 100, sd = 15)                      # 2
```

```
## [1] 75.3272
```

Due to the symmetry of the Normal distribution we have the following useful identities for Φ:

As mentioned before, Φ can be used to approximate the cdf of other random variables. Below is a particular example for binomial random variables when n is large[6]:

[6] A rule of thumb is that if $np \geq 10$ and $n(1-p) \geq 10$, it is safe to use this approximation.

Example 14

A manufacture will reject a batch of widgets if, in a sample of 100 randomly selected widgets from the batch, 15 or more are defective.

If 12% of the widgets in the batch are defective, what is the probability of rejecting the batch? (Use the Normal approximation to answer this question.)

```
1 - pnorm((15 + 0.5 - (.12 * 100))/sqrt(.12 * .88 * 100))
```

```
## [1] 0.1407288
```

The approximation works for Poisson random variables too, when λ is large; choose $\mu = \lambda$ and $\sigma = \sqrt{\lambda}$ for the approximation.[7]

[7] Many of the distributions we see can be related to the Normal distribution in some way.

Example 15

Suppose $X \sim \text{POI}(100)$. Estimate $\mathbb{P}(X \leq 110)$.

```r
pnorm(110 + 0.5, mean = 100, sd = sqrt(100))
```

```
## [1] 0.8531409
```

```r
ppois(110, 100)  # For comparison
```

```
## [1] 0.8528627
```

Section 4: The Exponential and Gamma Distributions

We have investigated the properties of the exponential distribution already; below we recall what we have seen:

Exponential random variables can be used to model waiting times, particularly when a process is **memoryless**; that is, the time remaining until the process terminates is independent of how long the process has currently taken.

Proposition 2 (Memoryless property). *Let $T \sim EXP(\mu)$. Then*

$$\mathbb{P}(T \geq t + t_0 | T \geq t_0) = \mathbb{P}(T \geq t)$$

Exponential random variablest play an important role in Poisson processes. The time between subsequent jumps of a Poisson process with parameter α follow an exponential distribution with mean $\mu = \frac{1}{\alpha}$.

Example 16

Your daughter's team score on average 10 points per game. You model the points scored by her team in a game with a Poisson pro-

cess, and $t = 1$ is a whole game.

1. Based on this, what is the expected time between points score by your daughter's team?

2. Suppose that by the start of the second half your daughter's team has scored 3 points. Given this, what is the expected time when your daughter's team score is 4 points?

```
(mu3 <- integrate(function(x) {x * dexp(x, rate = 10)}, 0, Inf))  # 1
```

```
## 0.1 with absolute error < 4.9e-05
```

```
0.5 + mu3$value
```

```
## [1] 0.6
```

The **gamma function**, $\Gamma(\alpha)$, is given below:

The gamma function has interesting properties, including

(Based on this we can say that the gamma function is the continuous analogue to $n!$.)

The **(lower) incomplete gamma function**, $\gamma(\alpha, x)$, is given below:

This yields the obvious asymptotic relationship between $\gamma(\alpha, x)$ and $\Gamma(\alpha)$:

The following are the pdf and cdf of the **gamma distribution** with parameters α and β (we write $X \sim \text{GAMMA}(\alpha, \beta)$ to say X follows such a distribution):

If $\beta = 1$ then we refer to GAMMA$(\alpha, 1)$ as the **standard gamma distribution**. Table A.4 gives values of the cdf of the standard gamma distribution for particular α and x.

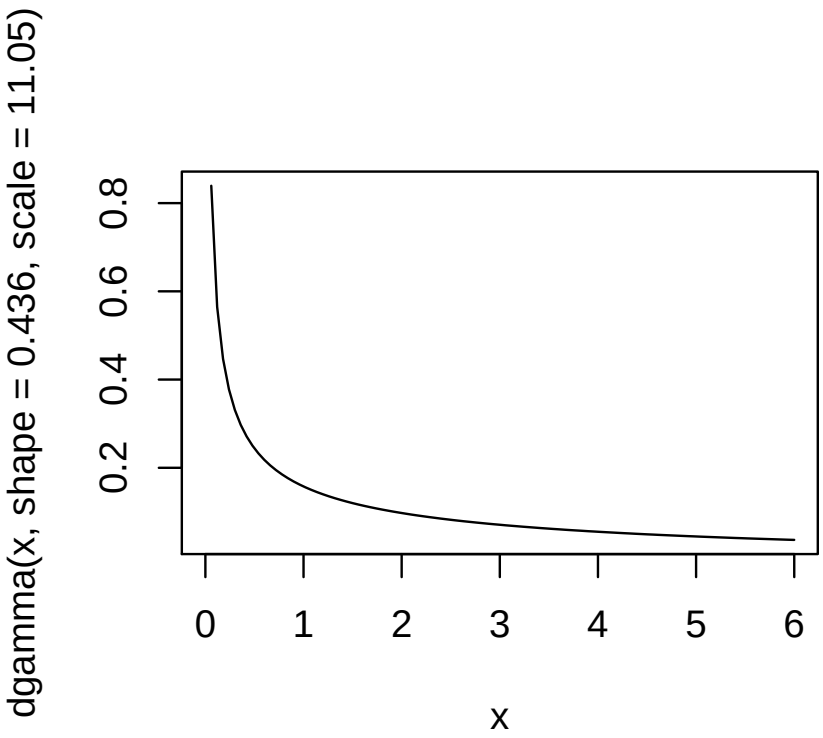
Standard gamma distributions can be used to compute probabilities involving non-standard gamma distributions in the following way:

The mean and variance of gamma-distributed random variables is given below:

Example 17

In a paper by Husak et al. [2007] the amount of rain (in mm) in Istanbull is fitted to a gamma distribution and the author estimated that the distribution of the amount of rain in April is $R \sim \text{GAMMA}(0.436, 11.05)$. Based on this, compute the mean and standard deviation of April rainfall.

```r
curve(dgamma(x, shape = 0.436, scale = 11.05), 0, 6)
```




```r
(mur <- integrate(function(x) {x * dgamma(x, shape = 0.436, scale = 11.05)},
                  0, Inf))
```

```
## 4.8178 with absolute error < 0.00029
```

```r
(varr <- integrate(function(x) {(x - mur$value)^2 * dgamma(x,
                                                           shape = 0.436, scale = 11.05)},
                   0, Inf))
```

```
## 53.23669 with absolute error < 0.0041
```

```r
sqrt(varr$value)
```

```
## [1] 7.296348
```

```r
# The probability the random variable is greater than 1
pgamma(1, shape = 0.436, scale = 11.05, lower.tail = FALSE)
```

```
## [1] 0.6145785
```

Let X_t be a Poisson process with rate parameter α. Let T_k be the time until the process is equal to k; that is, T_k is the smallest t such that $X_t = k$, so $X_{T_k} = k$. The distribution of T_k is known.

Example 18

You model the points your daughter's soccer team scores in a single game with a Poisson process with rate parameter $\alpha = 10$, with $t = 1$ representing a single game.

1. What is the mean and standard deviation of the time until her team scores 5 points?

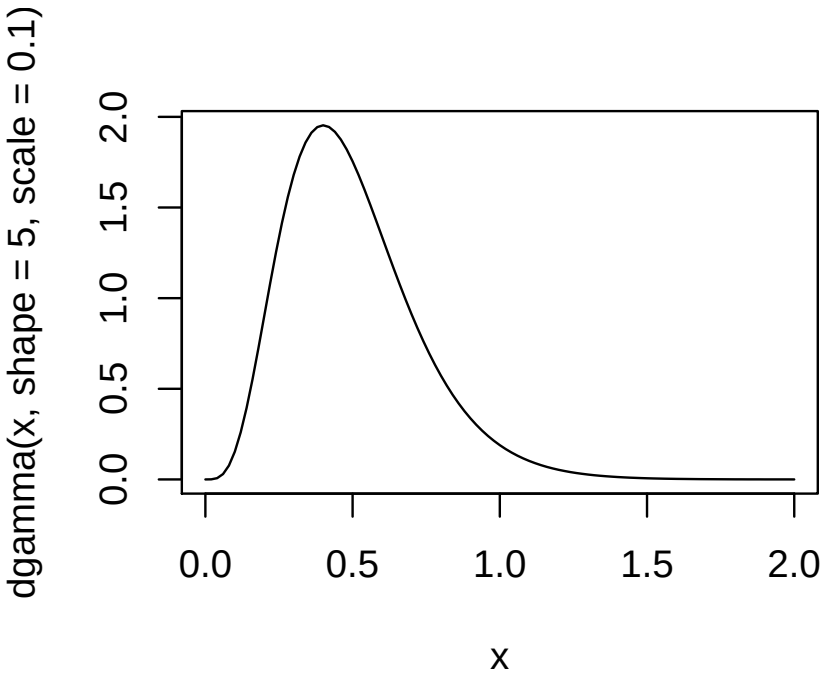
2. What is the probability that the time until her team scores 5 points is before half time ($t = 0.5$)?[8]

[8] Consider computing $\mathbb{P}(T_k > t)$, where t is a natural number. This is the event that the time when the process reaches k is greater than t; in other words, at time t, the process is less than k at time t, or $X_t < k$. Thus $\mathbb{P}(T_k > t) = \mathbb{P}(X_t < k)$. The left-hand side of this equality is an integral and the right-hand side is the sum $\sum_{x=0}^{k-1} \frac{e^{-\alpha t}(\alpha t)^x}{x!}$, so we have

$$\int_t^\infty \frac{\alpha^k x^{k-1} e^{-\alpha x}}{(k-1)!} dx = \sum_{x=0}^{k-1} \frac{e^{-\alpha t}(\alpha t)^x}{x!}$$

This identity could have been found with an inductive argument and integration by parts, but we have a probabilistic argument that explains why the identity holds, which is more illuminating.

```
curve(dgamma(x, shape = 5, scale = 0.1), 0, 2)
```




```
(mus <- integrate(function(x) {x * dgamma(x, shape = 5, scale = 0.1)},
                  0, Inf))
```

```
## 0.5 with absolute error < 3.5e-07
```

```
(vars <- integrate(function(x) {(x - mus$value)^2 * dgamma(x,
                                                    shape = 5, scale = 0.1)},
                0, Inf))
```

```
## 0.05 with absolute error < 2.7e-05
```

```
sqrt(vars$value)
```

```
## [1] 0.2236068
```

```
# The probability the random variable is greater than 1
pgamma(0.5, shape = 5, scale = 0.1)
```

```
## [1] 0.5595067
```

Notice that there is a relationship between the gamma distribution and the exponential distribution:

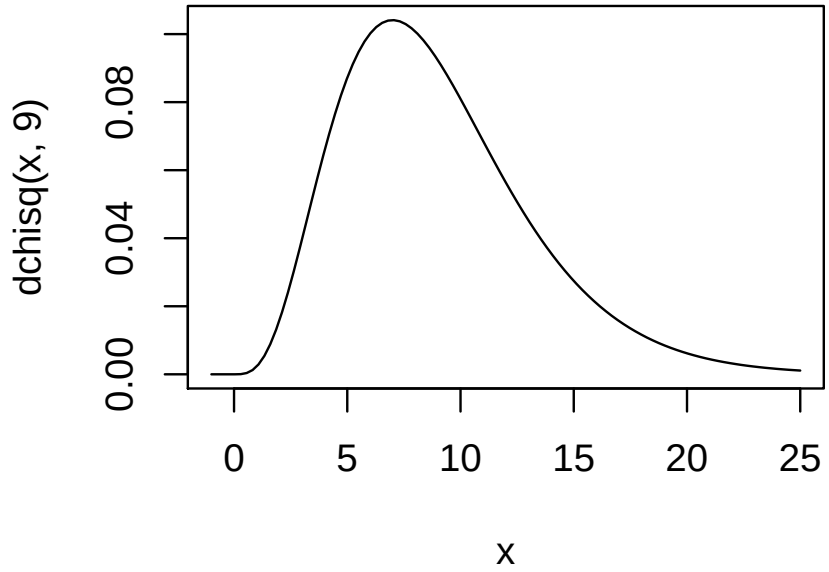
In this sense the exponential family of distributions is a subset of the gamma family of distributions.

The **chi-square distribution** is another distribution that belongs to the gamma family of distributions; we write $X \sim \chi^2(\nu)$ to indicate a chi-square distributed random variable. In particular, $X \sim \chi^2(\nu) \iff X \sim \text{GAMMA}(\nu/2, 2)$. This distribution is important in statistics for describing the sampling distribution of certain statistics. Values of the cdf of the chi-square distribution are given in Table A.7.

Example 19

Suppose $S^2 \sim \chi^2(9)$. Compute $\mathbb{E}\left[S^2\right]$, $\text{Var}\left(S^2\right)$, and $\mathbb{P}\left(S^2 > 3.325\right)$.

```
curve(dchisq(x, 9), -1, 25)
```




```
(mus2 <- integrate(function(x) {x * dchisq(x, 9)}, 0, Inf))
```

```
## 9 with absolute error < 7.6e-06
```

```
integrate(function(x) {(x - mus2$value)^2 * dchisq(x, 9)}, 0, Inf)
```

```
## 18 with absolute error < 0.00012
```

```
pchisq(3.325, 9, lower.tail = FALSE)
```

```
## [1] 0.9500055
```

Section 5: Other Continuous Distributions

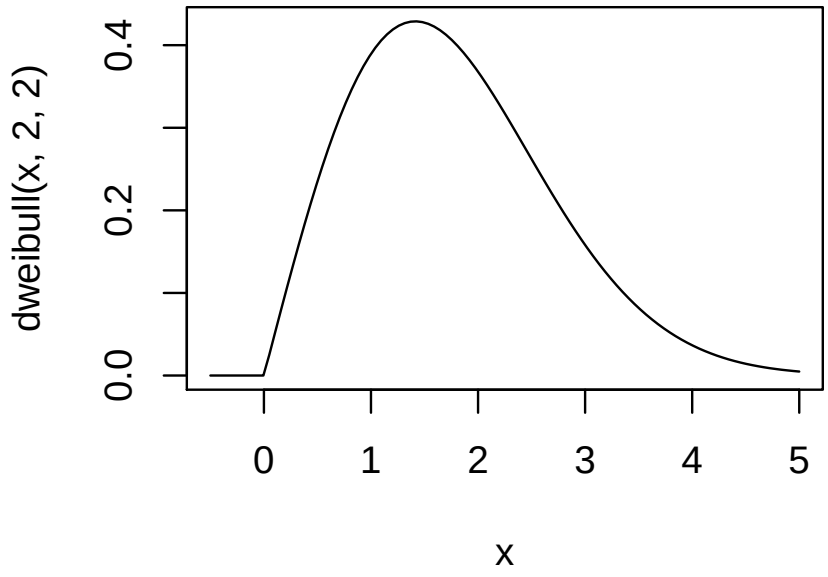
We say that X follows the **Weibull distribution** with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$, or $X \sim \text{WEI}(\alpha, \beta)$[9], if the pdf of X is

[9] Sometimes $X \sim \text{WEI}(\alpha, \beta, \gamma)$ is seen, which means that $X - \gamma \sim \text{WEI}(\alpha, \beta)$; that is, X is a shifted version of the usual Weibull distribution.

The mean, variance, and cdf of the Weibull distribution are given below

If $\alpha = 1$ the Weibull distribution is an exponential distribution.
Below is a sketch of the pdf of the Weibull distribution

```
curve(dweibull(x, 2, 2), -0.5, 5)
```



Example 20

Wind speed (in meters per second) at the site of a wind turbine is believed to follow a Weibull distribution with $\alpha = 2$ and $\beta = 8$. Compute the mean and median wind speeds and the standard deviation of wind speed.

The turbine will not turn if wind speed is below two meters per second. Compute the probability this occurs.

```
(muwind <- integrate(function(x) {x * dweibull(x, 2, 8)}, 0, Inf))
```

```
## 7.089815 with absolute error < 2.8e-06
```

```
(varwind <- integrate(function(x) {(x - muwind$value)^2 * dweibull(x, 2, 8)},
                      0, Inf))
```

```
## 13.73452 with absolute error < 6.6e-05
```

```
sqrt(varwind$value)
```

```
## [1] 3.706011
```

```
qweibull(0.5, 2, 8)  # Median
```

```
## [1] 6.660437
```

```
pweibull(2, 2, 8)
```

```
## [1] 0.06058694
```

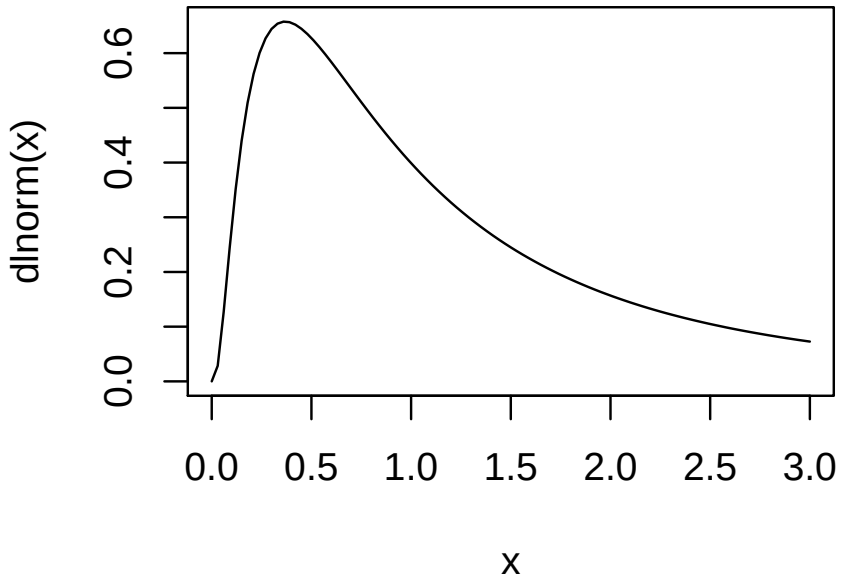
X is said to follow a **lognormal distribution**, denoted $X \sim LN(\mu, \sigma)$, if $\ln(X)$ follows a Normal distribution, or $\ln(X) \sim N(\mu, \sigma)$. X has pdf

We can express the cdf of X in terms of Φ like so:

μ and σ^2 are *not* the mean and variance of X. Instead we have

Below is a sketch of the pdf of X:

```
curve(dlnorm, 0, 3)
```



Example 21

The current price of the stock with ticker symbol CGM is \$26.18. The quants believe the the price of the stock in a year is $Y = 26.18X$, where $X \sim LN(0.1, 0.2)$. Based on this information, find l and u such that $\mathbb{P}(l \leq Y \leq u) = 0.95$ and $\mathbb{P}(Y \leq l) = 0.025$.

```
(lprime <- qlnorm(0.025, 0.1, 0.2))

## [1] 0.7467739

(uprime <- qlnorm(0.975, 0.1, 0.2))

## [1] 1.635572

(l <- lprime * 26.18)  # lower bound

## [1] 19.55054

(u <- uprime * 26.18)  # upper bound

## [1] 42.81928
```

X follows the **beta distribution**, denoted $X \sim \text{BETA}(\alpha, \beta, A, B)$[10], if X has the pdf

[10] It is also common to see $X \sim \text{BETA}(\alpha, \beta)$, which refers to the standard beta distribution.

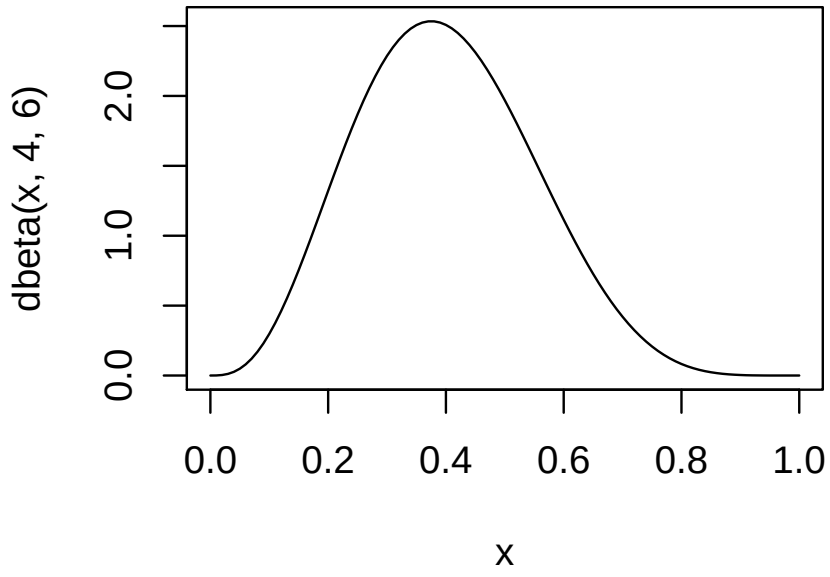
If $A = 0$ and $B = 1$, then X is said to have the **standard beta distribution**.

The mean and variance of X are given below:

The beta distribution can assume a large number of shapes depending on its shape parameters. But it has compact support, assigning positive probabilities only to regions between A and B.

Below is a sketch of what a beta distribution can look like.

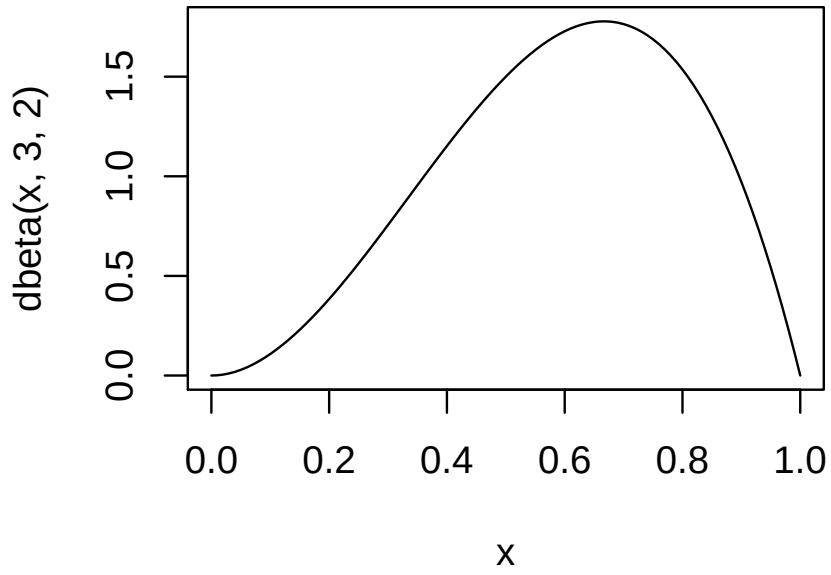
```r
curve(dbeta(x, 4, 6))
```



Example 22

Suppose $X \sim \text{BETA}(3, 2)$. Write down the pdf of X and compute $\mathbb{E}[X]$, $\text{Var}(X)$, and $\mathbb{P}(1/4 \leq X \leq 3/4)$.

```
curve(dbeta(x, 3, 2))
```




```
(mux <- integrate(function (x) {x * dbeta(x, 3, 2)}, 0, 1))
```

```
## 0.6 with absolute error < 6.7e-15
```

```
(varx <- integrate(function (x) {(x - mux$value)^2 * dbeta(x, 3, 2)}, 0, 1))
```

```
## 0.04 with absolute error < 4.4e-16
```

```
pbeta(.75, 3 ,2) - pbeta(.25, 3, 2)
```

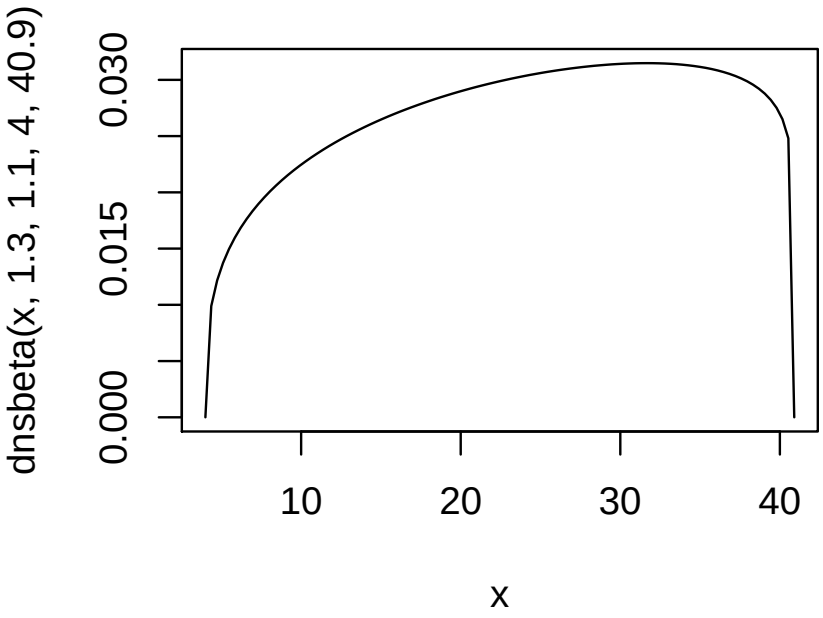
```
## [1] 0.6875
```

Example 23

In a paper by Maltamo et al. [2007], the basal diameter (in cm) of pine trees was fitted to a beta distribution. The paper suggests that, if B is the diameter of a pine tree, then $B \sim \text{BETA}(1.3, 1.1, 4.0, 40.9)$. What, then, is the mean diameter of the pine trees? What about the standard deviation?

```
suppressPackageStartupMessages(library(extraDistr))  # Package with more dist's

curve(dnsbeta(x, 1.3, 1.1, 4.0, 40.9), 4.0, 40.9)
```




```
(mudiam <- integrate(function(x) {x * dnsbeta(x, 1.3, 1.1, 4.0, 40.9)},
                     4.0, 40.9))

## 23.9875 with absolute error < 1.4e-06

(vardiam <- integrate(function(x) {(x - mudiam$value)^2 * dnsbeta(x, 1.3, 1.1,
                                                                  4.0, 40.9)},
                      4.0, 40.9))

## 99.42312 with absolute error < 0.00012

sqrt(vardiam$value)

## [1] 9.971114
```

Section 6: Probability Plots

Probability plots are a visual method used to check whether a dataset could plausibly have been drawn from a particular distribution. In essence, we compare the observed sample percentiles with the percentiles of a dataset if it had come from a chosen distribution. If the relationship between the observed and the theoretical distributions is linear, the distributional assumption seems reasonable. If

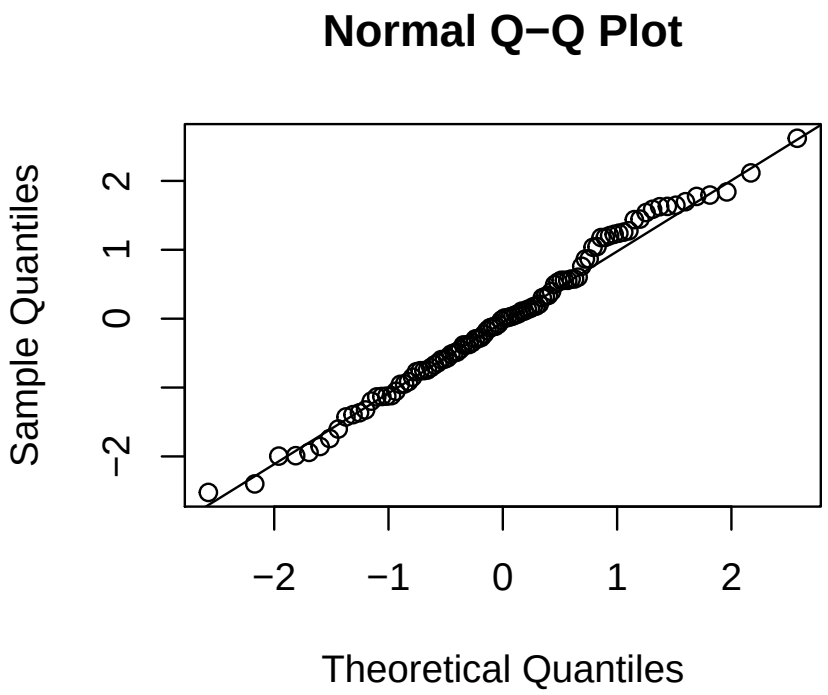
there is a nonlinear relationship, the distribution chosen is not likely a good model for the data.

While we can often argue that a certain data generating process produces a particular probability distribution in the discrete case, fitting data to distributions is more difficult in the continuous case; we can't make arguments like we could in the discrete case. Thus we turn to probability plots or statistical tests.

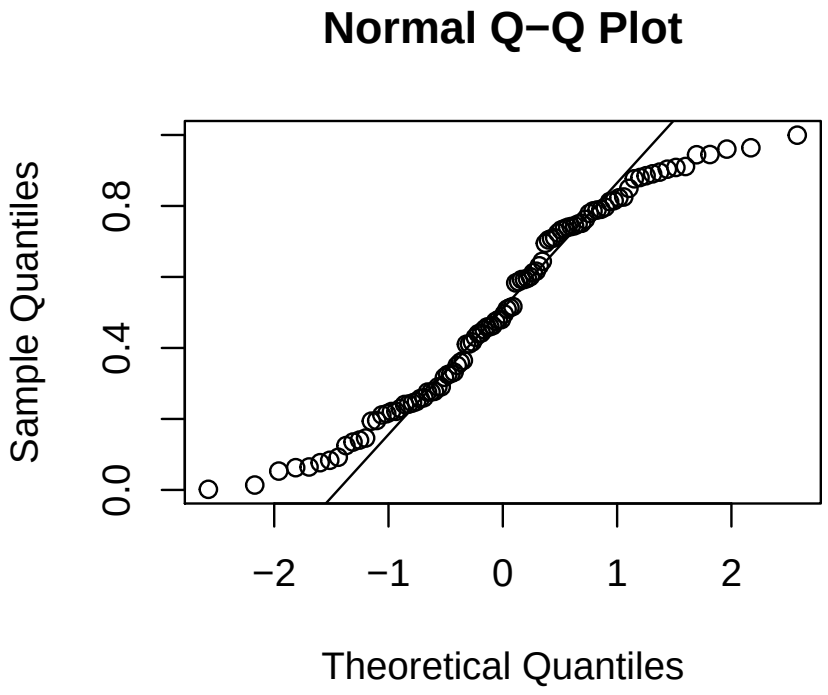
Below is an example of a probability plot.

```r
dat1 <- rnorm(100)
dat2 <- runif(100)

# Probability plot checking for Normal distributions
qqnorm(dat1); qqline(dat1)
```




```r
qqnorm(dat2); qqline(dat2)
```



Suppose we have a sample $x_1, \ldots, x_n$, and $r_1, \ldots, r_n$ is the ordered sample (with $r_1 \leq r_2 \leq \ldots \leq r_n$). We call r_i the $[100(i-.5)/n]$**th sample percentile**.

To construct a probability plot, we do the following:

1. For $i = 1, \ldots, n$, we find the $[100(i-.5)/n]$th percentile of the *theoretical* distribution; we call these the theoretical percentiles, referred to as $\eta\left(\frac{i-0.5}{n}\right)$.
2. For $i = 1, \ldots, n$, plot the point $\left(\eta\left(\frac{i-0.5}{n}\right), r_i\right)$ on a Cartesian grid; the x-axis is the theoretical percentiles and the y-axis is the observed percentiles.

If the theoretical distribution is a Normal distribution, we call the probability plot a **Normal probability plot**.

We then decide if the relationship between the theoretical and observed percentiles appears linear. If yes, then the distribution is a good fit. Otherwise, it's a bad fit.

Example 24

Consider the following dataset:

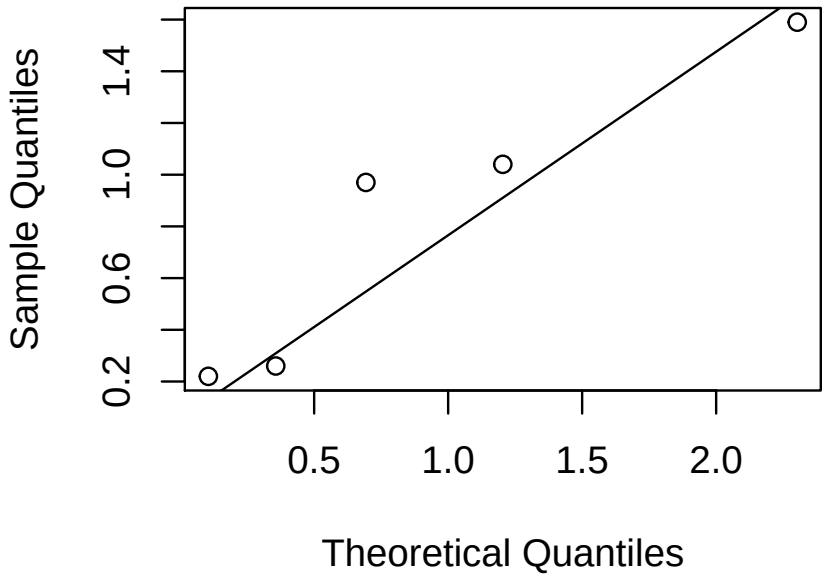
i	1	2	3	4	5
r_i	0.22	0.26	0.97	1.04	1.59

Create a probability plot to determine if it's plausible the data came from a EXP(1) distribution.

```
x <- c(0.22, 0.26, 0.97, 1.04, 1.59)
(theo <- qexp((1:5 - 0.5)/5))

## [1] 0.1053605 0.3566749 0.6931472 1.2039728
## [5] 2.3025851

qqplot(theo, x, xlab = "Theoretical Quantiles", ylab = "Sample Quantiles")
qqline(x, distribution = qexp)
```



As described we can check whether a dataset was generated by a *particular* distribution (in the last example, it was $\text{EXP}(1)$), but we usually want to know whether a dataset was generated by *a member of a family* of distributions (for example, $\text{EXP}(\mu)$). Fortunately there are tricks we can use to do the latter task.

We call θ_1 a **location** parameter and θ_2 a **scale** parameter if the cdf $F(x;\theta_1,\theta_2)$ depends on $\frac{x-\theta_1}{\theta_2}$[11]. Below are examples of parameters that are either (or are *neither*) location or scale parameters.[12]

[11] Intuitively, θ_1 shifts the pdf left or right rigidly, while θ_2 stretches or compresses the pdf.

[12] Notice that the mean is *not* always a location parameter. For the exponential distribution, the mean is a scale parameter.

If the theoretical distribution involes location and scale parameters, we estimate them; call the estimates $\hat{\theta}_1$ and $\hat{\theta}_2$. Instead of plotting using r_i, we use $\frac{r_i - \hat{\theta}_1}{\hat{\theta}_2}$, and use the *standard* theoretical distribution where $\theta_1 = 0$ and $\theta_2 = 1$[13].

Example 25

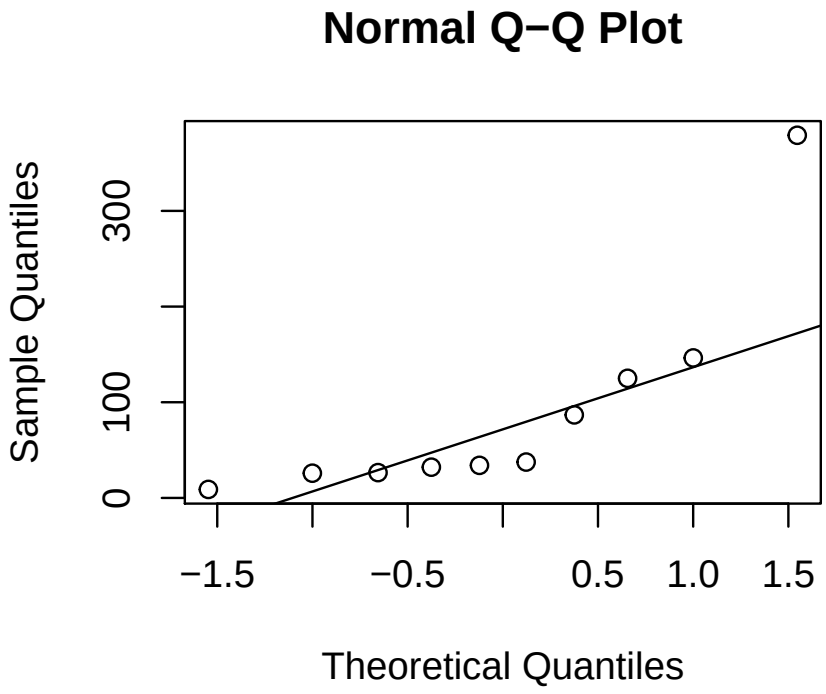
Construct a probability plot to check if the following dataset was plausibly generated by a Normal distribution.

i	1	2	3	4	5	6	7
r_i	8.89	25.86	26.47	32.16	34.07	37.49	86.80

i	8	9	10
r_i	125.02	146.36	379.06

[13] What if we need a parameter that is neither a location nor scale parameter? One trick would be to transform the data in an appropriate way. For example, if we think $X_i \sim LN(\mu, \sigma)$, neither μ nor σ are location or scale parameters, but we can create a probability plot for $\ln(X_i)$ instead and see if the new, transformed dataset is Normally distributed, as it should be if our hypothesis is correct; in this case, μ and σ can now be treated as location and scale parameters, respectively. This trick would not work if we wanted to check if $X_i \sim \text{BETA}(\alpha, \beta)$ since no transformation will turn α and β into location/scale parameters. In that case we may be forced to estimate α and β from the data, assuming that our hypothesis is true; in this example, call the estimates $\hat{\alpha}$ and $\hat{\beta}$. Then we would construct a probability plot to see if the data came from the distribution $\text{BETA}(\hat{\alpha}, \hat{\beta})$.

```
y <- c(8.89, 25.86, 26.47, 32.16, 34.07, 37.49, 86.80, 125.02, 146.36, 379.06)
qqnorm(y)
qqline(y)
```



References

Gregory J. Husak, Joel Michaelsen, and Chris Funk. Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. *International Journal of Climatology*, 27(7):935–944, 2007. URL `http://chg.ucsb.edu/publications/pdfs/2006_Husaketal_GammaDistribution.pdf`.

Matti Maltamo, Janna Puumalainen, and Risto Pivinen. Comparison of beta and Weibull functions for modelling basal area diameter distribution in stands of *pinus sylvestris* and *picea abies*. *Scandinavian Journal of Forest Research*, 10(1-4):284–295, 2007. URL `http://dx.doi.org/10.1080/02827589509382895`.