# Probability Distributions

A probability distribution describes how the values of a random variable is distributed. For example, the collection of all possible outcomes of a sequence of coin tossing is known to follow the binomial distribution. Whereas the means of sufficiently large samples of a data population are known to resemble the normal distribution. Since the characteristics of these theoretical distributions are well understood, they can be used to make statistical inferences on the entire data population as a whole.

```
> help(runif)
```

## Regarding d, p, and q

The letters d, p, and q have special meanings:

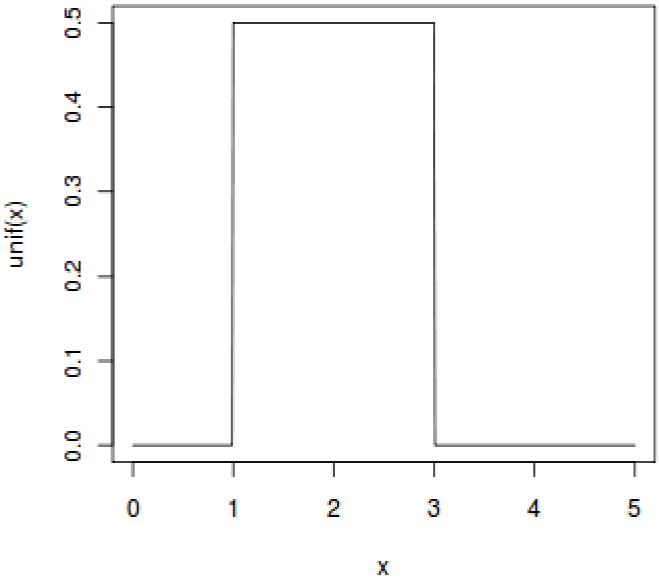
- "d" is for "density." It is used to find values of the probability density function.
- "p" is for "probability." It is used to find the probability that the random variable lies to the left of a given number.
- "q" is for "quantile." It is used to find the quantiles of a given distribution.

## Continuous Uniform Distribution

The continuous uniform distribution is the probability distribution of random number selection from the continuous interval between *a* and *b*. Its density function is defined by the following.

$$f(x)=\left\{\begin{matrix}\frac{1}{b-a}\\ 0\end{matrix}\right\}\left\{\begin{matrix}when\,a\leqslant x\leqslant b\\ when\,x<a.or.x<b\end{matrix}\right\}$$

Here is a graph of the continuous uniform distribution with *a = 1, b = 3.*



We could have used **dunif**, which will produce density values for the uniform distribution.

```
> curve (dunif(x , min = 1 , max = 5) , from=0, to=6, n=1000)
```

```
> x = seq(1,5,length=200)
> y = dunif(x,min=1,max=5)
> polygon(c(1,x,5),c(0,y,0),col="lightgray",border=NA)
```

Note that the arguments **min=1** and **max=5** provide the endpoints of the interval [1,5] on which the uniform probability density function is defined.

**Using punif**

Suppose that we would like to find the probability that the random variable X is less than or equal to 2.

To calculate this probability, we would shade the region under the density function to the left of and including 2, then calculate its area.

```
> curve (dunif(x , min = 1 , max = 5) , from=0, to=6, n=1000)

> x = seq(1,2,length=100)
> y = dunif(x,min=1,max=5)
> polygon(c(1,x,2),c(0,y,0),col="lightgray",border=NA)
```

Note that the width of the shaded area is 1, the height is 1/4, so the area of the shaded region is 1/4.

Thus, the probability that x ≤ 2 is 1/4.

We can use the **punif** command to compute the probability that x ≤ 2.

```
> punif(2,min=1,max=5)
[1] 0.25
```

If now, we can compute the probability that x ≥ 2.

```
> punif(2,min=1,max=5,lower=FALSE)
[1] 0.75
```

Let's look at another example. Suppose that we wanted to find the probability that x lies between 2 and 4. We could draw the uniform distribution, then shade the area under the curve between 2 and 4.

```
> curve (dunif(x , min = 1 , max = 5) , from=0, to=6, n=1000)

> x = seq(2,4,length=100)
> y = dunif(x,min=1,max=5)
> polygon(c(2,x,4),c(0,y,0),col="lightgray",border=NA)
```

Note that the width of the shaded area is 2, the height is 1/4, so the area is 1/2. That is, the probability that 2 < x < 4 is 1/2.

We can use **punif** to arrive at the same conclusion. To find the area between 2 and 4, we must subtract the area to the left of 2 from the area to the left of 4.

```
> punif(4,min=1,max=5)-punif(2,min=1,max=5)
[1] 0.5
```

Finally, if we need to find the area to the right of a given number, simply subtract the are to the left of the given number from the total area; i.e., subtract the area from the left of a given number from 1. So, the following calculation will find the probability that x > 4.

```
> 1-punif(4,min=1,max=5)
[1] 0.25
```

```
> punif(4,min=1,max=5,lower=FALSE)
[1] 0.25
```

## Using qunif

Thus, to find the 25th percentile for the uniform distribution on the interval [1,5], we execute the following code.

```
> qunif(0.25,min=1,max=5)
[1] 2
```

## Problem

Select ten random numbers between one and three. Compute the probability that  x ≤ 2.

## Solution

We apply the generation function *runif* of the uniform distribution to generate ten random numbers between one and three.

```
> runif(10, min=1, max=3)
[1] 1.6121 1.2028 1.9306 2.4233 1.6874 1.1502 2.7068
[8] 1.4455 2.4122 2.2171
```

We can use the **punif** command to compute the probability that x ≤ 2.

```
> punif(2,min=1,max=5)
[1] 0.25
```

Now compute the probability that x ≥ 2.

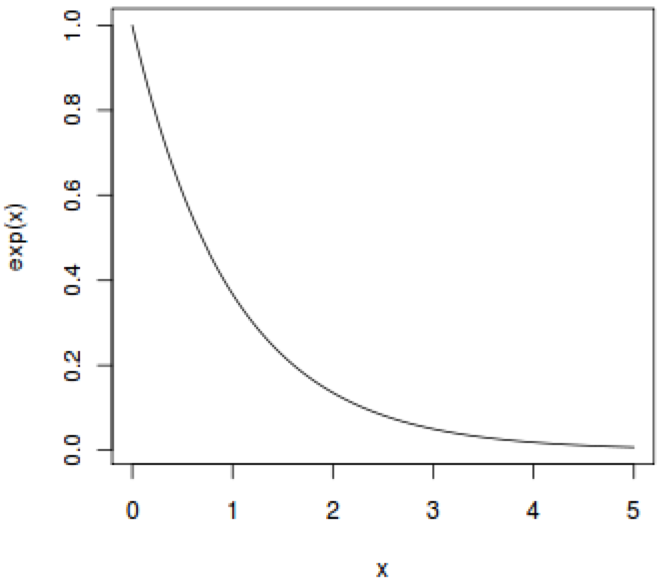
```
> punif(2,min=1,max=5,lower=F)
[1] 0.75
```

## Exponential Distribution

The exponential distribution describes the arrival time of a randomly recurring independent event sequence. If *μ* is the mean waiting time for the next event recurrence, its probability density function is:

$$f(x) = \begin{Bmatrix} \frac{1}{\mu} e^{-x/\mu} \\ 0 \end{Bmatrix} \begin{Bmatrix} when\, x \geqslant 0 \\ when\, x < 0 \end{Bmatrix}$$

Here is a graph of the exponential distribution with *μ = 1*.



The exponential probability density function is defined on the interval [0, ∞]. It has the following definition.

The exponential distribution some time change to have mean μ = 1/λ and standard deviation σ = 1/λ.

Suppose that we set λ = 1. Then the mean of the distribution should be μ = 1 and the standard deviation should be σ = 1 as well. We will be more efficient using the **dexp** command.

```
> curve (dexp(x , rate=1) , from=0, to=4, n=1000)
```

Note that the argument **rate** expects us to respond with the value of λ.

The exponential probability density function is shown on the interval [0,4]. However, remember that the full domain is on [0,∞), so we've shown only part of the full picture. In determining on what domain to draw the function, we extended the interval three standard deviations to the right of the mean.

Unlike the normal and uniform distributions, the exponential distribution is not symmetric about its mean.

### Using pexp

Suppose that we want to find the probability that x ≤ 1. We would shade the area under the exponential probability density function to the left of 1.

```
> curve (dexp(x , rate=1) , from=0, to=4, n=1000)

> x=seq(0,1,length=200)
> y=dexp(x,rate=1)
> polygon(c(0,x,1),c(0,y,0),col="lightgray")
```

Now, just as we did with the uniform distribution above, we will use the **pexp** command to compute the area of the shaded region.

```
> pexp(1,rate=1)
[1] 0.6321206
```

You may find it surprising that the answer was not 50%! However, the mean at x = 1 is not the median! The graph of the exponential is "skewed to the right" and the extreme outliers at the right strongly influence the mean, pushing it to the right of the median.

As a second example, suppose that we want to find the probability that x lies between 1 and 2.

```
> curve (dexp(x , rate=1) , from=0, to=4, n=1000)

> x=seq(1,2,length=200)
> y=dexp(x,rate=1)
> polygon(c(1,x,2),c(0,y,0),col="lightgray")
```

Again, to find the shaded area, we must subtract the area to the left of x = 1 from the area to the left of x = 2.

```
> pexp(2,rate=1)-pexp(1,rate=1)
[1] 0.2325442
```

Thus, the probability of selecting a number between 1 and 2 from this distribution is approximately 0.2325442.

Finally, if we need to find the area to the right of a given number, simply subtract the area to the left of the given number from the total area. For example, to find the probability that x > 3, subtract the probability that x ≤ 3 from 1.

```
> 1-pexp(3,rate=1)
[1] 0.04978707
```

The command **qexp** will find quantiles for the exponential distribution in the same way as we saw the **qunif** find quantiles for the uniform distribution. Thus, to find the 50th percentile for the exponential distribution on the interval, we execute the following code.

```
> qexp(0.50,rate=1)
[1] 0.6931472
```

This result is in keeping with the fact that the distribution is skewed badly to the right. The outliers at the right end greatly influence the mean, pushing it to the right. With the mean at x = 1, the current result for the median makes good sense. Fifty percent of the data lies to the left of x = 0.6931472 and fifty percent of the data lies to the right of x = 0.6931472.

**Problem**
Suppose the mean checkout time of a supermarket cashier is three minutes. Find the probability of a customer checkout being completed by the cashier in less than two minutes.

**Solution**
The checkout processing rate is equals to one divided by the mean checkout completion time. Hence the processing rate is 1/3 checkouts per minute. We then apply the function *pexp* of the exponential distribution with rate=1/3.

```
> pexp(2, rate=1/3)
[1] 0.48658
```

**Answer**
The probability of finishing a checkout in under two minutes by the cashier is 48.7%

## Binomial Distribution

The binomial distribution is a discrete probability distribution. It describes the outcome of *n* independent trials in an experiment. Each trial is assumed to have only two outcome, labeled as success or failure. If the probability of a successful trial is *p*, then the probability of having *x* successful trials in an experiment is as follows.

$$f(x)=\binom{n}{x}p^{x}(1-p)^{(n-x)}$$

where $x = 0, 1, 2, \ldots, n$

```
> x1 = rbinom(1e4,size=12,prob=0.2)
> hist(x1)

> x2 = rbinom(1e4,size=12,prob=0.5)
> hist(x2)

> x3 = rbinom(1e4,size=12,prob=0.8)
> hist(x3)


> curve(dbinom(x, size=2000, prob=0.314),from=0, to=2000)

> curve(dbinom(x, size=2000, prob=0.314, log=TRUE),from=0, to=2000)
```

**Problem**
Suppose there are twelve multiple choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Find the probability of having four or less correct answers if a student attempts to answer every question at random.

**Solution**
Since only one out of five possible answers is correct, the probability of answering a question correctly by random is 1/5=0.2. To find the probability of having four or less correct answers by random attempt, we apply the function *pbinom* with $x = 4$, $n = 12$, $p = 0.2$.

```
> pbinom(4, size=12, prob=0.2)
[1] 0.92744
```

**Answer**
The probability of four or less questions answered correctly by random in a twelve question multiple choice quiz is 92.7%.

## Poisson Distribution

The Poisson distribution is the probability distribution of independent events occurrence in an interval.

If *λ* is the mean occurrence per interval, then the probability of having *x* occurrence within a given interval is:

$$f(x)=\frac{\lambda^{x} e^{-\lambda}}{x!}$$

where $x = 0, 1, 2, 3, ...$

```
> x1 = rpois(1e4,lambda=1.)
> hist(x1)

> x2 = rpois(1e4,lambda=10.)
> hist(x2)

> x3 = rpois(1e4,lambda=100.)
> hist(x3)
```

**Problem**
If there are twelve cars crossing a bridge per minute on average, find the probability of having sixteen or more cars crossing the bridge in a particular minute.

**Solution**
We compute the upper tail probability of the Poisson distribution with the function *ppois*.

```
> ppois(16, lambda=12, lower=FALSE)   # find upper tail
[1] 0.10129
```

**Answer**
If there are twelve cars crossing a bridge per minute on average, the probability of having sixteen or more cars crossing the bridge in a particular minute is 10.1%.

## Normal Distribution

The normal distribution is defined by the following probability density function, where $\mu$ is the population mean and $\sigma^2$ is the variance.

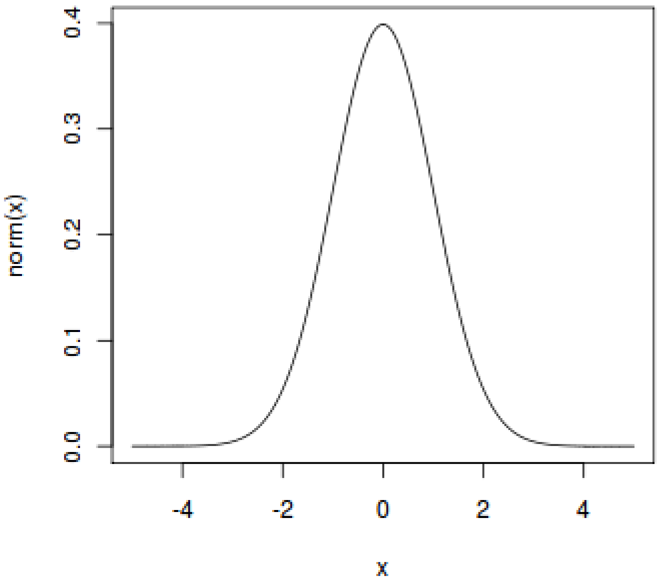
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$$

If a random variable *X* follows the normal distribution, then we write:

$$X\sim N(\mu,\sigma^2)$$

In particular, the normal distribution with $\mu = 0$ and $\sigma = 1$ is called the standard normal distribution, and is denoted as *N(0,1)*. It can be graphed as follows.

The normal distribution is important because of the Central Limit Theorem, which states that the population of all possible samples of size n from a population with mean $\mu$ and variance $\sigma^2$ approaches a normal distribution with mean $\mu$ and $\sigma^2/n$ when *n* approaches infinity.



```
> par(mfrow=c(2,2))
> curve (dnorm(x , mean = 0 , sd = 1) , from=-5, to=5, n=1000)
> curve (dnorm(x , mean = 5 , sd = 2) , from=-10, to=20, n=1000)
> curve (dnorm(x , mean =-5 , sd = 5) , from=-25, to=15, n=1000)
> curve (dnorm(x , mean = 15 , sd = 0.75) , from=5, to=25, n=1000)
```

### Problem

Assume that the test scores of a college entrance exam fits a normal distribution. Furthermore, the mean test score is 72, and the standard deviation is 15.2. What is the percentage of students scoring 84 or more in the exam?

### Solution

We apply the function *pnorm* of the normal distribution with mean 72 and standard deviation 15.2. Since we are looking for the percentage of students scoring higher than 84, we are interested in the upper tail of the normal distribution.

```
> pnorm(84, mean=72, sd=15.2, lower.tail=FALSE)
[1] 0.21492
```

### Answer

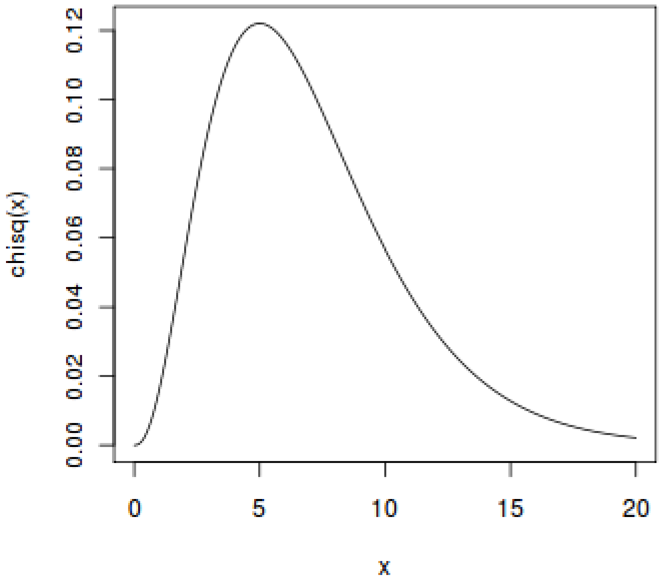
The percentage of students scoring 84 or more in the college entrance exam is 21.5%.

## Chi-squared Distribution

If $X_1, X_2, \ldots, X_m$ are m independent random variables having the standard normal distribution, then the following quantity follows a Chi-Squared distribution with m degrees of freedom. Its mean is *m*, and its variance is 2*m*.

$$V = X_1^2 + X_2^2 + \cdots + X_m^2 \sim X_{(m)}^2$$

Here is a graph of the Chi-Squared distribution 7 degrees of freedom.



**Problem**
Find the 95th percentile of the Chi-Squared distribution with 7 degrees of freedom.

**Solution**
We apply the quantile function *qchisq* of the Chi-Squared distribution against the decimal values 0.95.

```
> qchisq(.95, df=7)       # 7 degrees of freedom
[1] 14.067
```

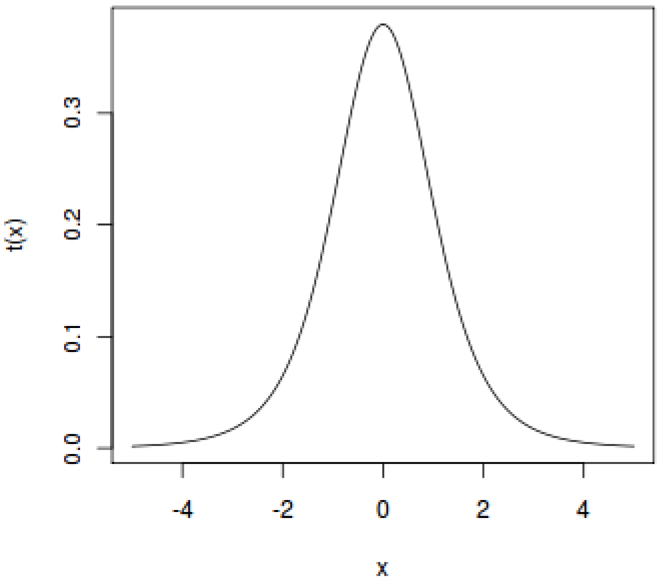
**Answer**
The 95th percentile of the Chi-Squared distribution with 7 degrees of freedom is 14.067.

## Student t Distribution

Assume that a random variable *Z* has the standard normal distribution, and another random variable *V* has the Chi-Squared distribution with m degrees of freedom. Assume further that *Z* and *V* are independent, then the following quantity follows a Student t distribution with *m* degrees of freedom.

$$t = \frac{Z}{\sqrt{V/m}} \sim t_{(m)}$$

Here is a graph of the Student t distribution with 5 degrees of freedom.



**Problem**
Find the 2.5th and 97.5th percentiles of the Student t distribution with 5 degrees of freedom.

**Solution**
We apply the quantile function qt of the Student *t* distribution against the decimal values 0.025 and 0.975.

```
> qt(c(.025, .975), df=5)   # 5 degrees of freedom
[1] -2.5706  2.5706
```

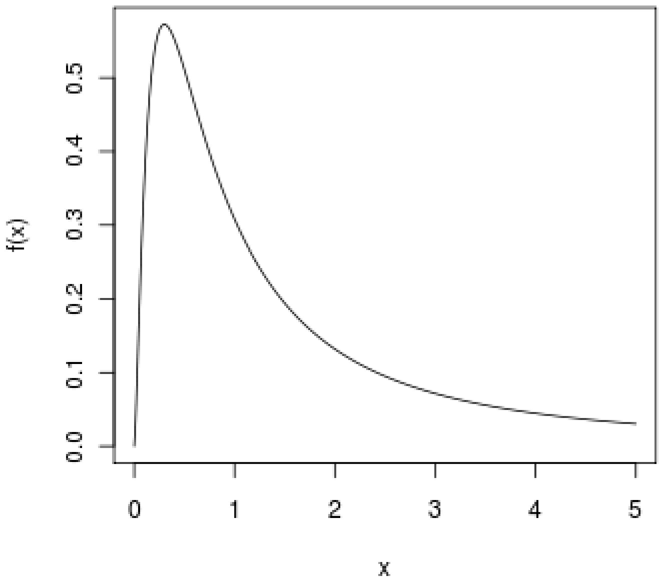
**Answer**
The 2.5th and 97.5th percentiles of the Student *t* distribution with 5 degrees of freedom are -2.5706 and 2.5706 respectively.

## F Distribution

If $V_1$ and $V_2$ are two independent random variables having the Chi-Squared distribution with $m_1$ and $m_2$ degrees of freedom respectively, then the following quantity follows an *F* distribution with $m_1$ numerator degrees of freedom and $m_2$ denominator degrees of freedom, i.e., ($m_1$,$m_2$) degrees of freedom.

$$F = \frac{V_1/m_1}{V_2/m_2} \sim F_{(m_1, m_2)}$$

Here is a graph of the *F* distribution with (5, 2) degrees of freedom.



**Problem**
Find the 95th percentile of the *F* distribution with (5, 2) degrees of freedom.

**Solution**
We apply the quantile function *qf* of the *F* distribution against the decimal value 0.95.

```
> qf(.95, df1=5, df2=2)
[1] 19.296
```

**Answer**
The 95th percentile of the *F* distribution with (5, 2) degrees of freedom is 19.296.