Lab 3: Distributions

Hao Wang

1/16/2017

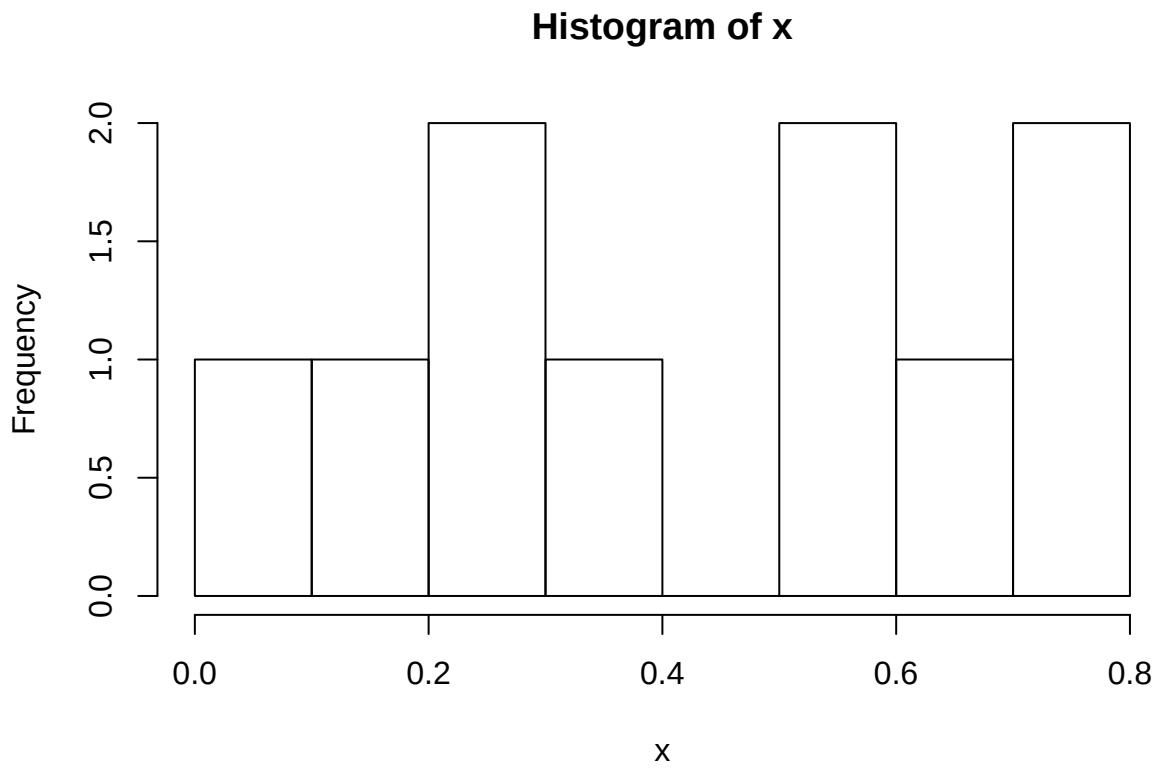
Distributions

R is great in creating random variables of different distributions. Today we'll create random variables in uniform distribution, bernouli distribution, binomial distribution, poisson distribution, normal distribution and student t-distribution.

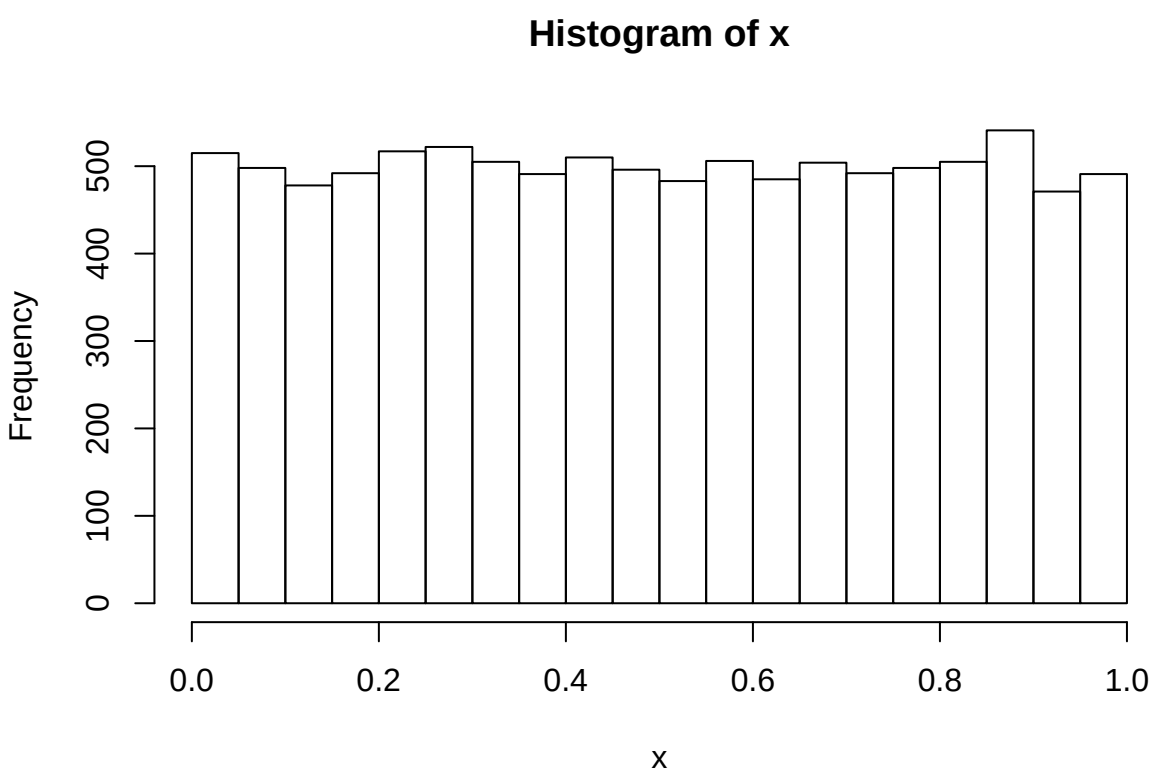
Uniform distribution

A uniform distribution, sometimes also known as a rectangular distribution, is a distribution that has constant probability. The probability density function and cumulative distribution function for a continuous uniform distribution on the interval are.

```
set.seed(1000) #set.seed identifies a fixed random number. Make the results reproducible
x <- runif(10) #create a list of 10 values of uniform distribution
hist(x)
```




```
x <- runif(10000)
hist(x)
```



- Lab practice 1: why the histograms of x are different? And whay may happen if you delete the set.seed line? What may happen if you change the number of the set.seed() line?

1. Discrete Distributions

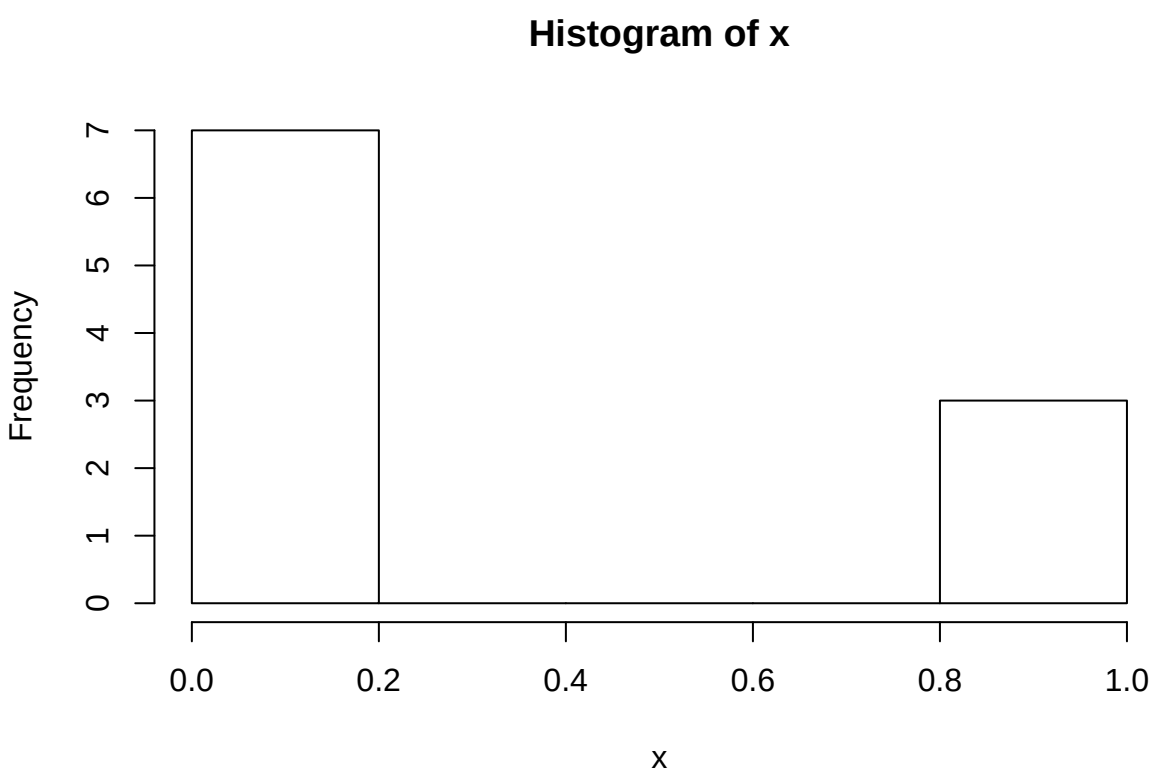
1.a Bernouli distribution

A Bernoulli trial is a chance event that can have one of two outcomes, usually called "success" or "failure."

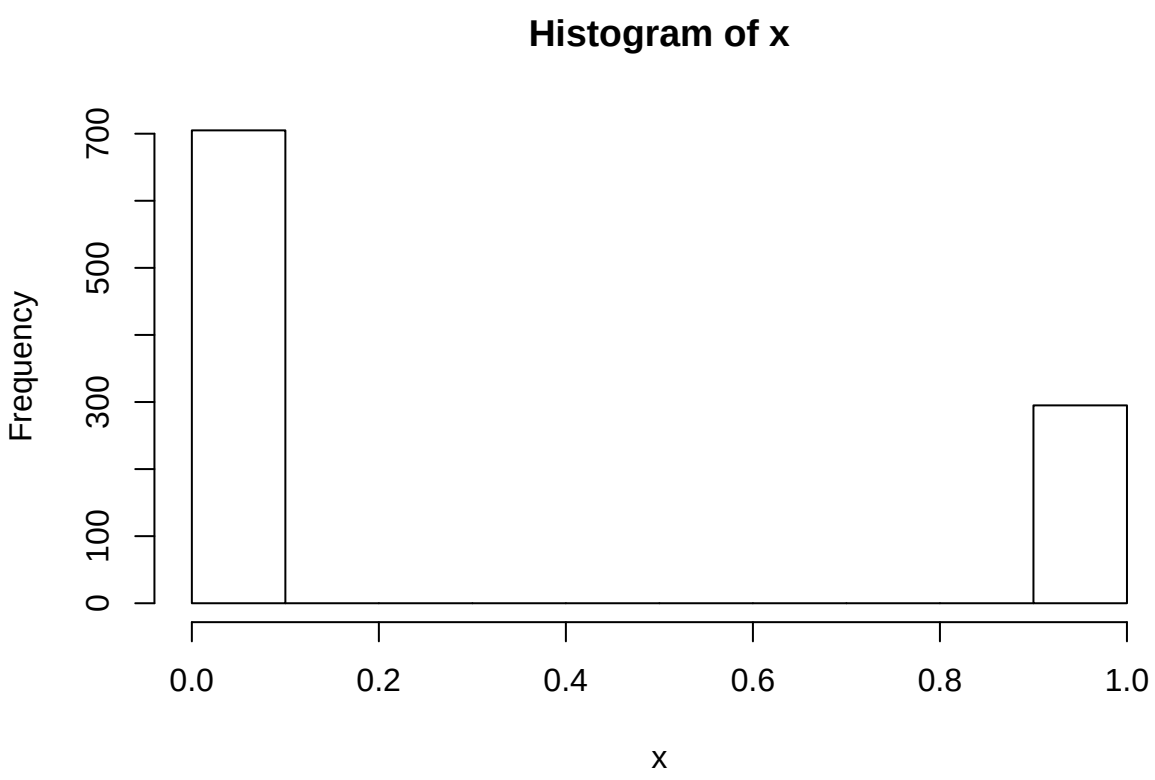
This distribution has one parameter, the unobserved probability of success, p. The probability of failure, often designated q, is the complement of p: 1-p. Examples are: tossing a coin, rain or not rain. Bernouli distribution only has two outcomes.

The rbinom() function can be used to simulate N independent bernouli/binomial random variables.

```
x <- rbinom(10, 1, 0.3)# it generates 10 random numbers, with 0.3 probability of being 1
# and 0.7 probability of being 0.
hist(x)
```




```
x <- rbinom(1000, 1, 0.3) #this time we generate 1000 random numbers
hist(x)
```



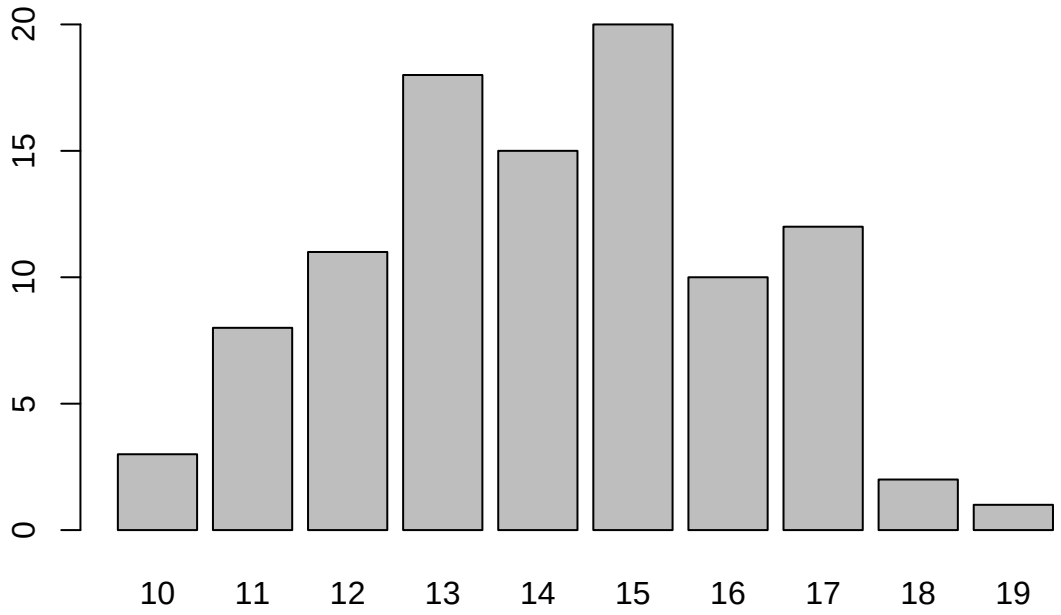
1.b Binomial distribution

In probability theory and statistics, the binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p. A success/failure experiment is also called a Bernoulli experiment or Bernoulli trial; when n = 1, the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance.

$$f(x; n, p) = Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

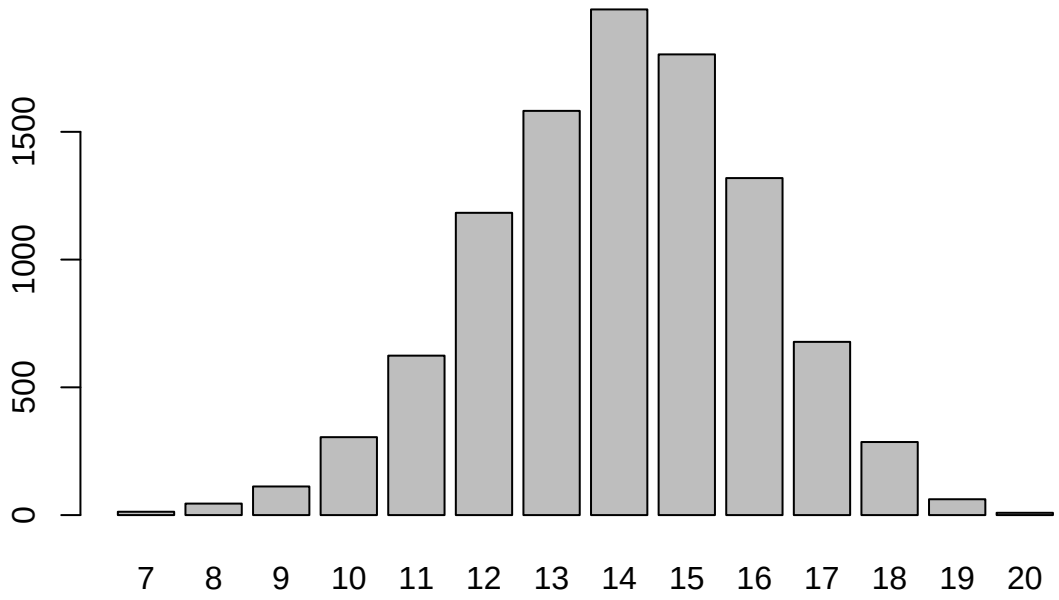
For example, we can generate 100 binomial random numbers with parameters n = 20 and p = .7. In this case x measures the number of successes in 20 bernouli trails. Taking 20 trials as one experiment, we repeat this experiment 100 times.

```
x <- rbinom(100, 20, .7) #x is the binomial random numbers
barplot(table(x))
```



We can also increase the size of n

```
x <- rbinom(10000, 20, .7) #x is the binomial random numbers
barplot(table(x))
```



1.c Poisson distribution

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume. For instance, an individual keeping track of the amount of mail they receive each day may notice that they receive an average number of 4 letters per day.

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

$$x! = x(x-1)(x-2)(x-3)...2*1$$

λ is the mean and variance of poisson distribution. We can simulate poisson distribution in R

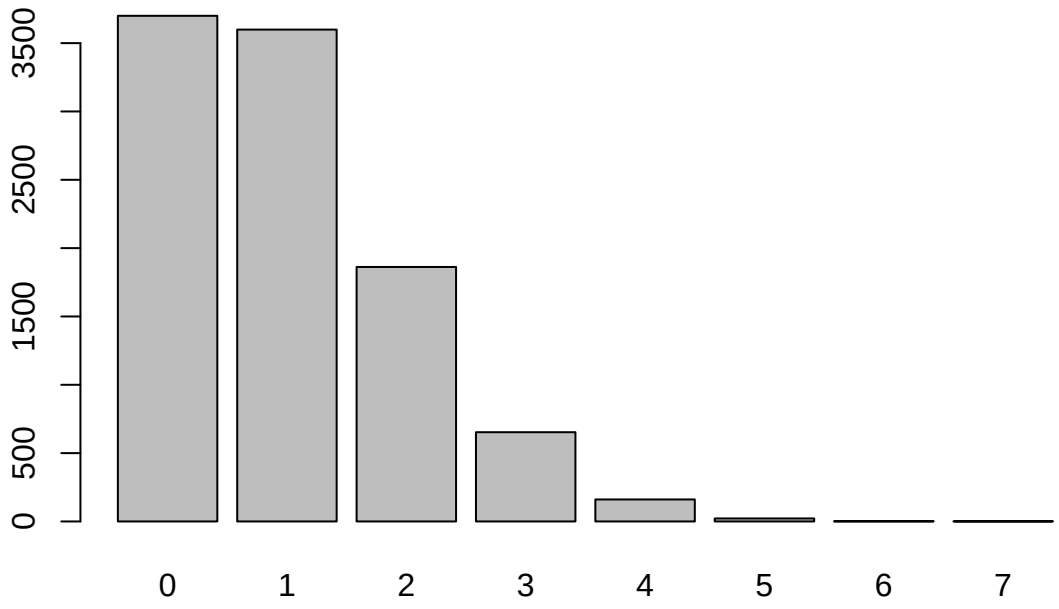
```
set.seed(1)
x <- rpois(10000, 1) #1000 poisson random numbers, lambda is 1
probability <-dpois(x, 1)
cat("The mean of x is", mean(x))
```

```
## The mean of x is 1.0055
```

```
cat("The variance of x is", var(x))
```

```
## The variance of x is 1.00617
```

```
barplot(table(x))
```



- Lab Practice 2: change the value of λ, how does the shape vary?

2. Continuous Distirbutions

2.a Normal distribution

The normal distribution is defined by the following probability density function, where μ is the population mean and σ^2 is the variance.

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

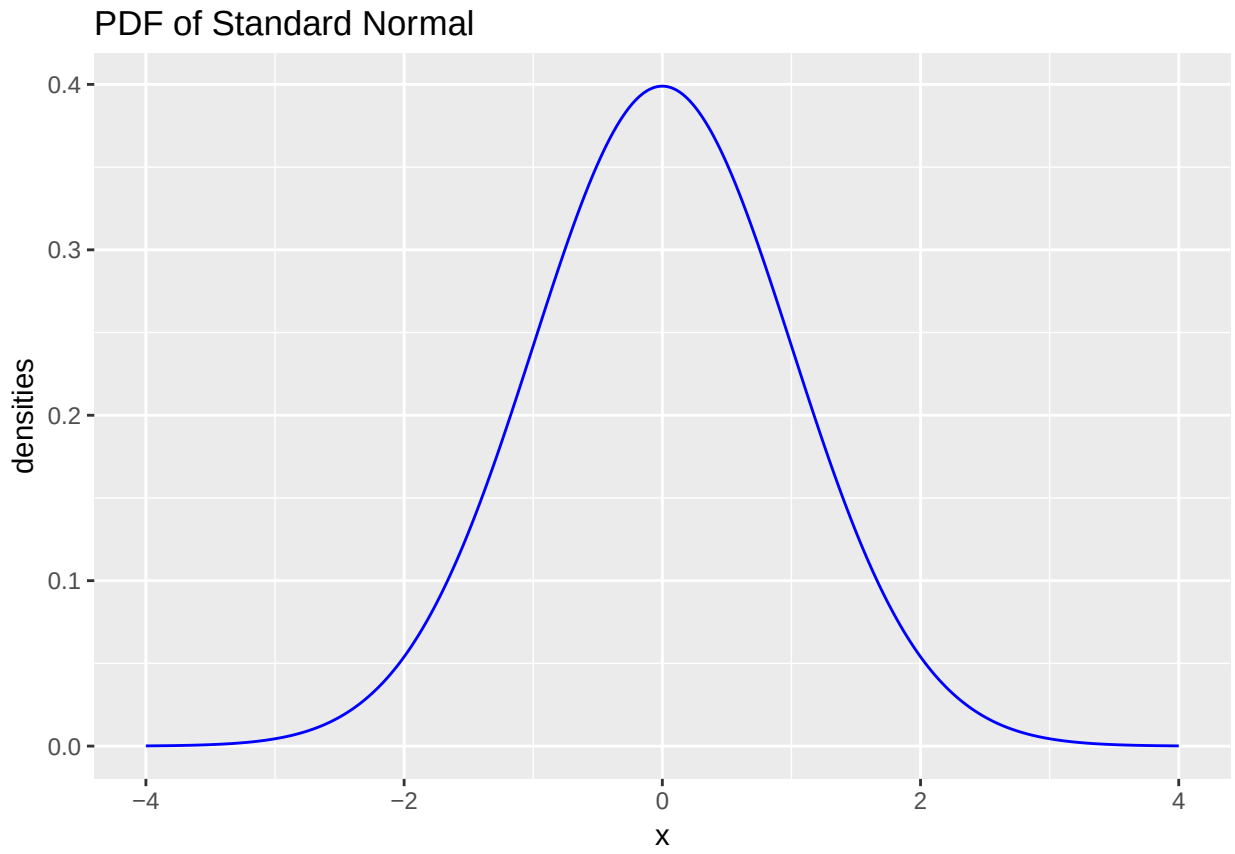
When μ is 0 and σ is 1, we call this standard normal distribution

```
require(ggplot2)
```

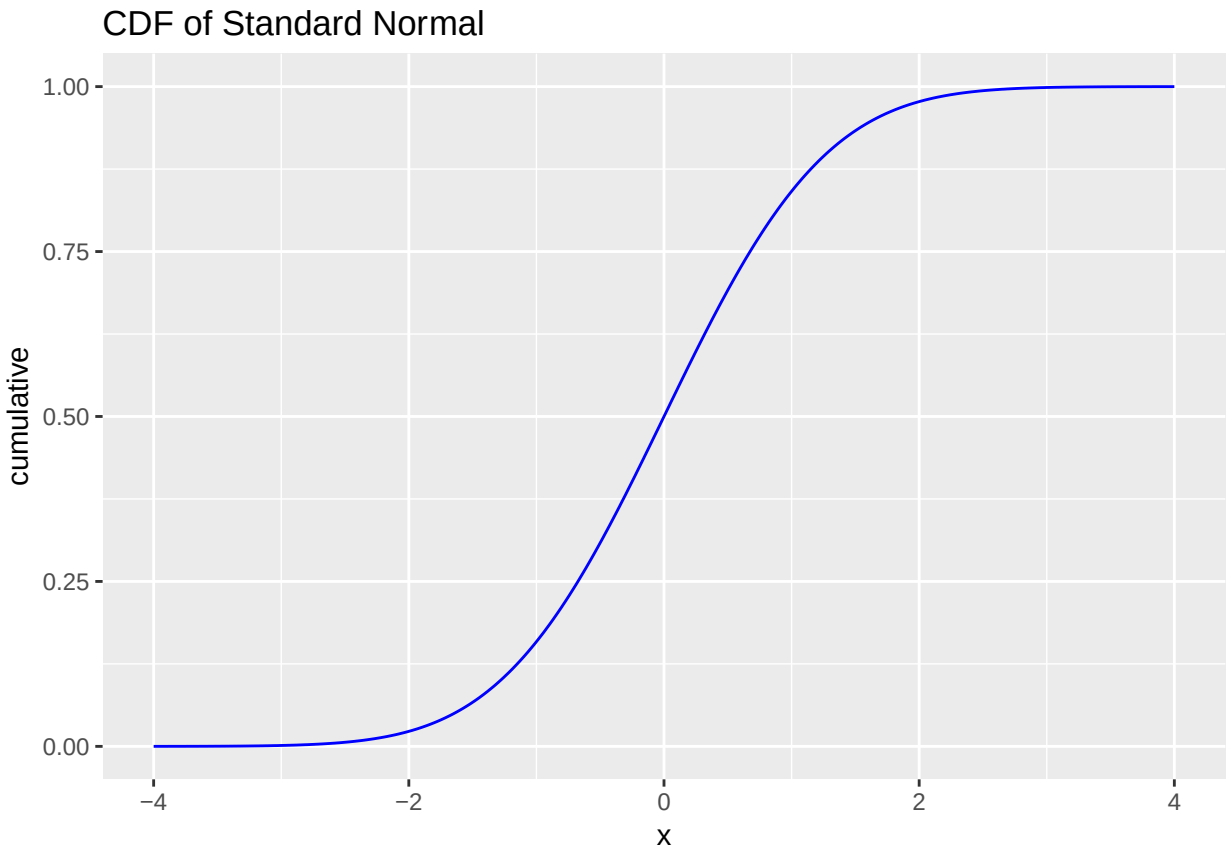
```
## Loading required package: ggplot2
```

```
set.seed(1)
x<-seq(-4,4,.01) #generates values from -4 to 4, with a step of 0.01
densities<-dnorm(x, 0, 1) #This calcualtes the pdf(probability dense function) of std normal
cumulative<-pnorm(x, 0, 1) #This calcualtes the cdf(cumulative dense function) of std normal
randomdeviates<-rnorm(1000,0,1) #1000 random normal numbers, mean is 0 and std is 1
```

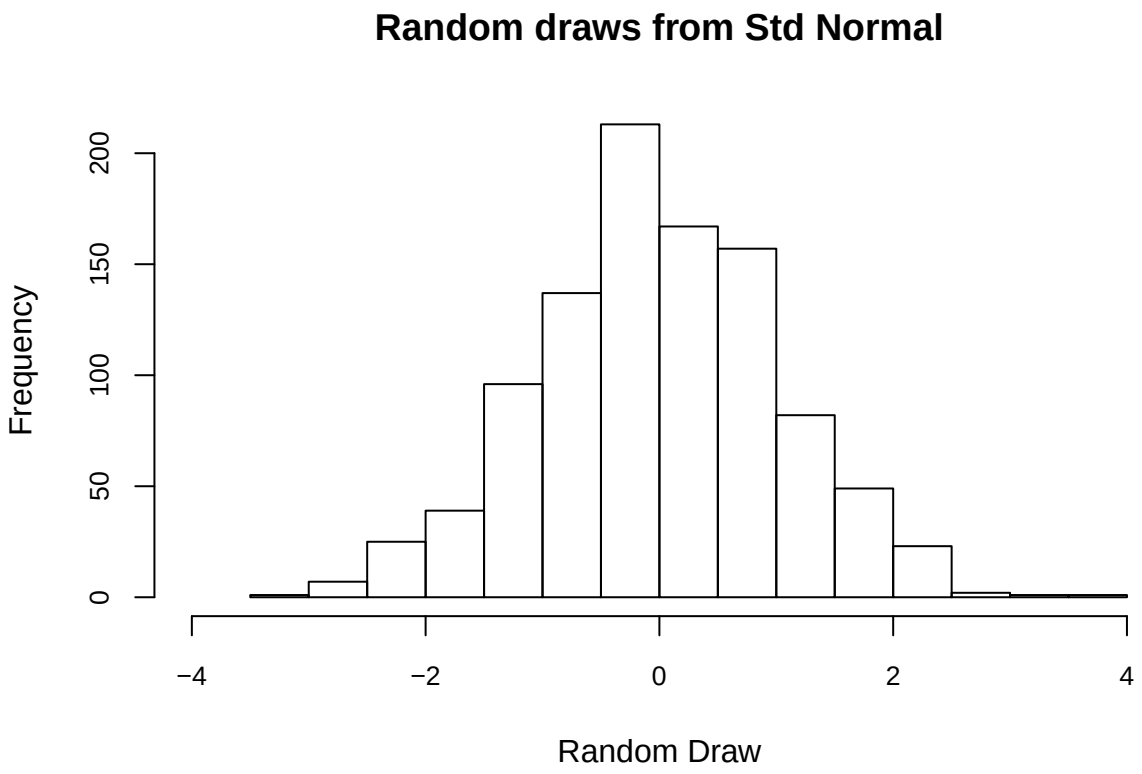
```r
plotdata <- data.frame(x,densities, cumulative)


#par(mfrow=c(1,3), mar=c(3,4,4,2))

ggplot(data = plotdata, aes(x = x, y = densities))+
      geom_line(color = "blue") +
      ggtitle("PDF of Standard Normal")
```




```r
ggplot(data = plotdata, aes(x = x, y = cumulative))+
      geom_line(color = "blue") +
      ggtitle("CDF of Standard Normal")
```




```
hist(randomdeviates, main="Random draws from Std Normal", xlab = "Random Draw", cex.axis=.8, xlim=c(-4,
```



-Lab Practice 3: change the values of densities<-dnorm(x, 0, 1), see what happens to the plots?

2.a.1 Z score

$$z = \frac{x - \mu}{\sigma}$$

z score tells us how far away the point is from mean. We can calculate the probability of a point (obs.) by z score. Besides, we can calculate the cdf and do significance test.

```
#If I want to know the probabilty of a std normal when x > 2
#We need to calculate F(2)
x2 <- pnorm(2, 0, 1)
cat("cumulative probability of x<2 is", x2)
```

```
## cumulative probability of x<2 is 0.9772499
```

```
cat("cumulative probability of x>2 is", 1- x2)
```

```
## cumulative probability of x>2 is 0.02275013
```

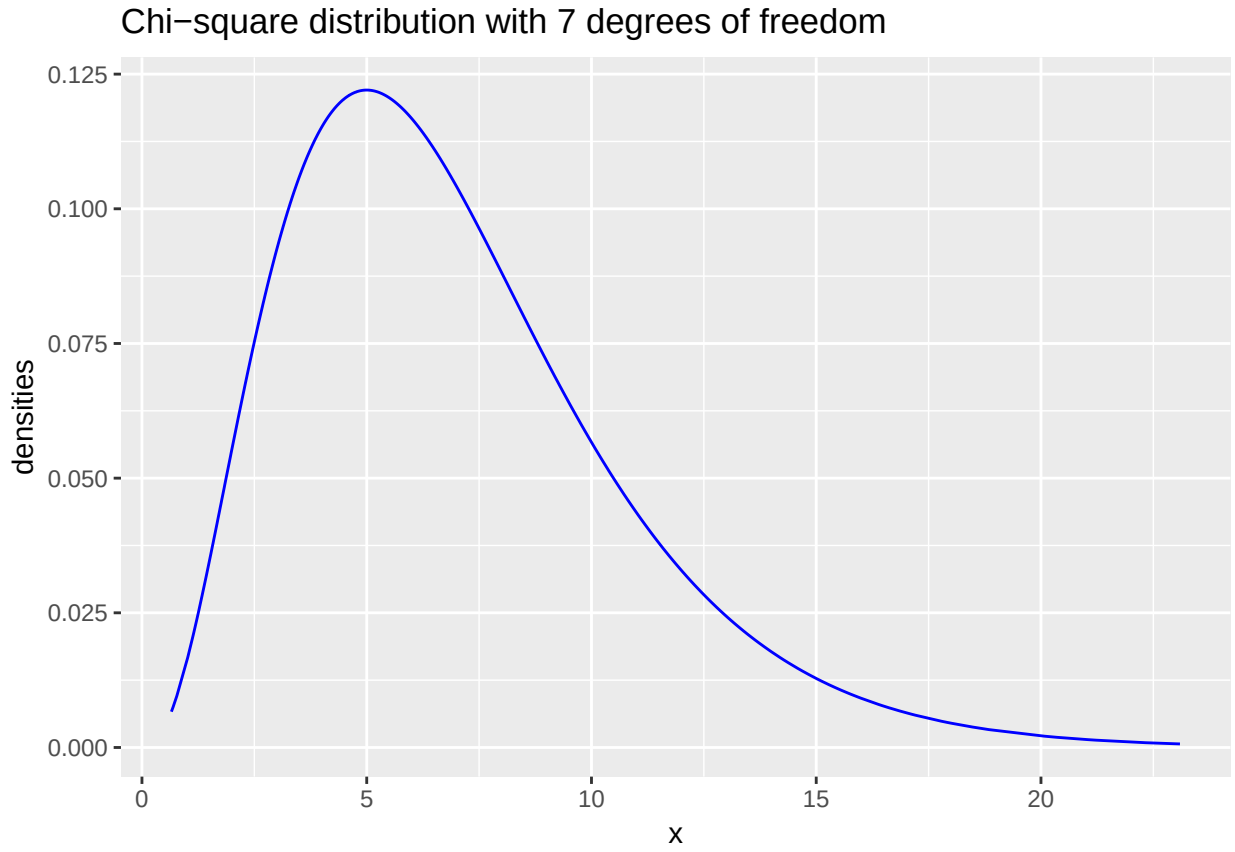
The example above is a single-tailed test. Normally we have two-tailed test. In a two-tailed test we calculate the probabilties of both tails. Normally, $\Phi(z = 1.96) = 95\%$. $\Phi(x)$ calculates the cdf of normal distribution.

2.b Chi-square distribution

f X1,X2,.,Xm 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 2m.

$$V = X_1^2 + X_2^2 + X_3^2 \sim x^2_{(m)}$$

```
x <- rchisq(1000, 7) #1000 random variables, Chi-square distribution of 7 degrees of freedom
densities <- dchisq(x, 7) #probability of x with chi-sq dist and m= 7
#We can plot its density in this way
ggplot(data.frame(x, densities),aes(x=x, y=densities)) +
  geom_line(color="blue") +
  ggtitle("Chi-square distribution with 7 degrees of freedom")
```



2.c 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.

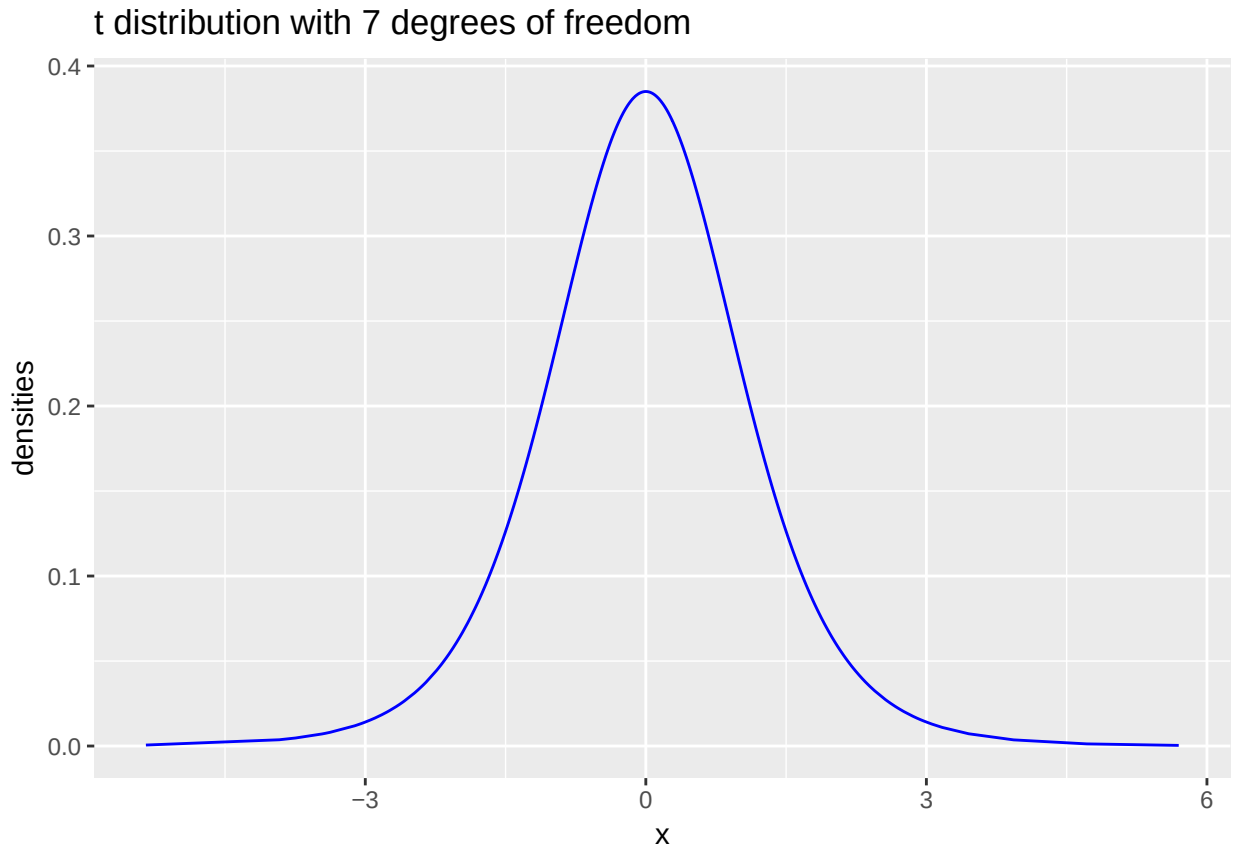
$$t = \frac{Z}{\sqrt{(V/m)}} \sim t_{(m)}$$

Alternatively, with X as a random variable (N =n), mean = μ and standard error = s_x

$$t = \frac{\bar{x} - \mu}{s_x/\sqrt{n}}$$

This is often called the test statistics.

```
x <- rt(1000, 7) #1000 random variables, t distribution of 7 degrees of freedom
densities <- dt(x, 7) #probability of x with t dist and m= 7
cumlative <- pt(x, 7) #cumulative probabilities of x
#We can plot its density in this way
ggplot(data.frame(x, densities),aes(x=x, y=densities)) +
  geom_line(color="blue") +
  ggtitle("t distribution with 7 degrees of freedom")
```



- t distribution properties :
- The mean of the distribution is equal to 0 .
- The variance is equal to m / (m - 2), where m is the degrees of freedom (see last section) and m > 2.
- The variance is always greater than 1, although it is close to 1 when there are many degrees of freedom.

- Lab Practice 4: calculate the following probabilities

1. Standard normal distribution, Pr(x > 3)
2. T distribution, df =10, Pr(x = 2)
3. T distribution, df =6, Pr (x < 1)

```
#pnorm(x, 0, 1)
#dt(x, 10)
#pt(x, 6)
```