**Combining and Transforming Random Variables**

**Examples**

1. Toss a pair of fair dice. Let X be the result on the first die and Y the result on the second die. Then S = X + Y is the sum of the two dice and P = XY is the product.
2. Let F be the temperature of a randomly selected object in degrees Fahrenheit. Then C = 5/9(F-32) is the object’s temperature in degrees Celsius.
3. A random sample of n adult females is chosen. Let be the height (in inches) of the ith person. Then,  is the sample mean.

**General Question**: If X is the result of combining two or more random variables or of transforming a single random variable, what is the distribution of X?

**Some General Rules.**

**Recall** Var(X) **=** = E(X2) – E(X)2

1. E(X+b) = E(X) + b and Var(X+b) = Var(X).
2. E(aX) = aE(X) and Var(aX) = Var(X).
3. E(X+Y) = E(X) + E(Y).

If X and Y are random variables and the value of X doesn’t affect the probability distribution of Y, then X and Y are **independent random variables**.

**Examples**

If a pair of fair dice are tossed, X is the value of the first die, and Y is the value of the second, then X and Y are independent.

If X and Y are the height and weight of a randomly selected person, then X and Y are not independent.

**General Rules Continued**

1. If X and Y are independent, then E(XY) = E(X)E(Y).
2. If X and Y are independent, then Var(X+Y) = Var(X) + Var(Y).

**A Simulation**

> die1<-resample(1:6,10000)

> die2<-resample(1:6,10000)

> prod<-die1\*die2

> mean(die1)

[1] 3.485

> mean(die2)

[1] 3.519

> mean(prod)

[1] 12.2979

**Example** Let X and Y be independent random variables that have the uniform distribution on [0,1].

Let S = X+Y. What kind of a distribution does S have?

**Simulation**

> simx<-runif(10000,0,1)

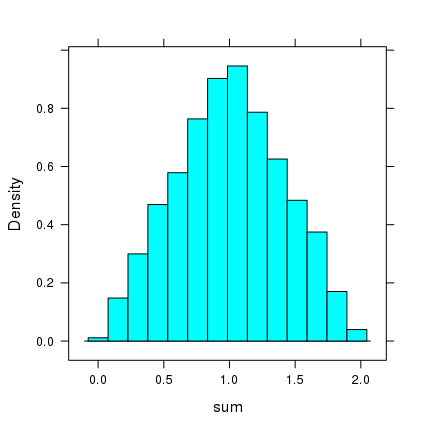
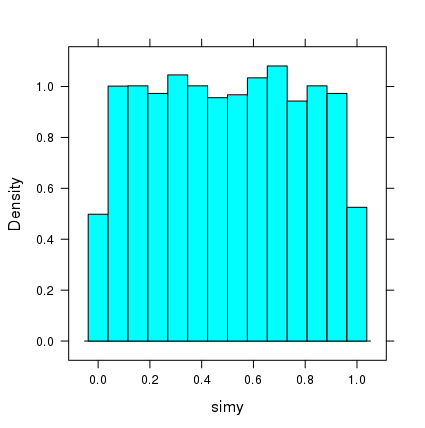
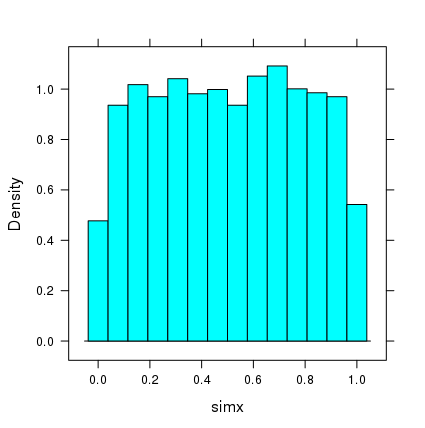
> simy<-runif(10000,0,1)

> sum<-simx+simy

> histogram(simy)

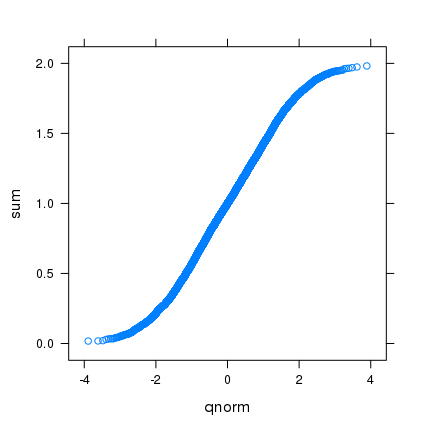
> histogram(simx)

> histogram(sum)

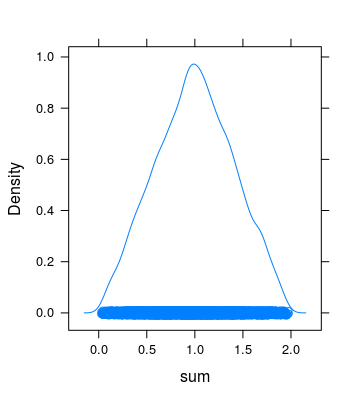


What do you guess is the density function for S? Normal?

> qqmath(sum)



> densityplot(sum)



**Linear Combinations**

If is a set of random variables, then is a **linear combination** of .

**General Rules**

6. 

7. , if  are independent.

**Normal Distributions**

**Fact 1** **Any linear combination of independent normal random variables is normal**

**Example:** If X is N(1,2), Y is N(-1,3), W is N(4,4), and X,Y,W are independent, find the distribution of

C = 2X + 3Y + W.

**Simulation.**

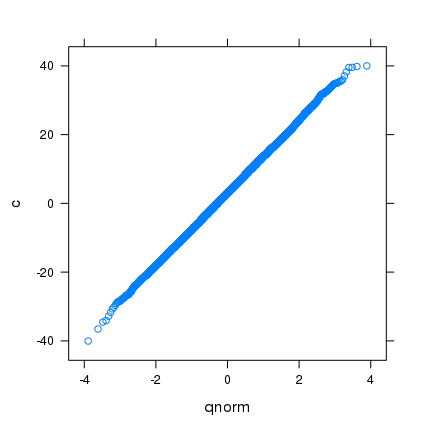
> simx<-rnorm(10000,1,2)

> simy<-rnorm(10000,-1,3)

> simw<-rnorm(10000,4,4)

> c<-2\*simx+3\*simy+simw

> qqmath(c)



> fitdistr(c,"normal")

mean sd

2.99177253 10.57286712

( 0.10572867) ( 0.07476146)

**Fact 2**

**If are independent random variables that have the same distribution, then the sum  has, approximately, a normal distribution if n is “large”. If n is “huge”, the distribution is almost exactly normal.**

Simulation

> sim1<-runif(10000,0,1)

> sim2<-runif(10000,0,1)

> sim3<-runif(10000,0,1)

> sim4<-runif(10000,0,1)

> sim5<-runif(10000,0,1)

> sim6<-runif(10000,0,1)

> sum2<-sim1+sim2

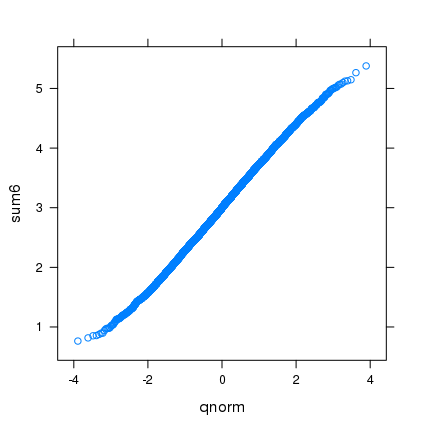
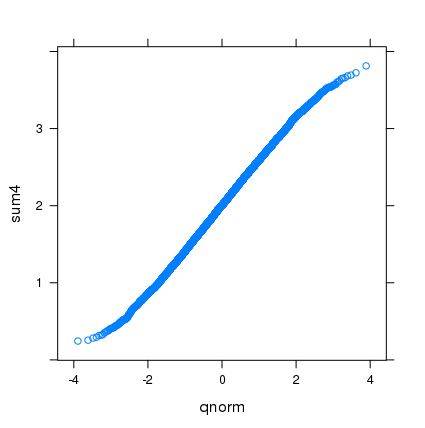
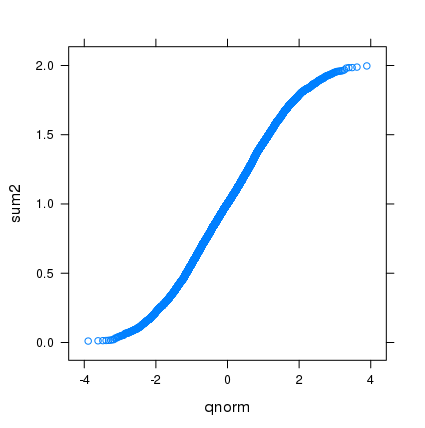
> sum4<-sim1+sim2+sim3+sim4

> sum6<-sim1+sim2+sim3+sim4+sim5+sim6

> qqmath(sum2)

> qqmath(sum4)

> qqmath(sum6)



> sim1<-rexp(10000,1)

> sim2<-rexp(10000,1)

> sim3<-rexp(10000,1)

> sim4<-rexp(10000,1)

> sim5<-rexp(10000,1)

> sim6<-rexp(10000,1)

> sim7<-rexp(10000,1)

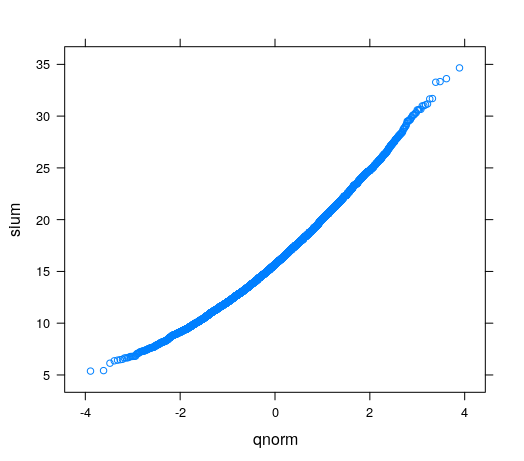
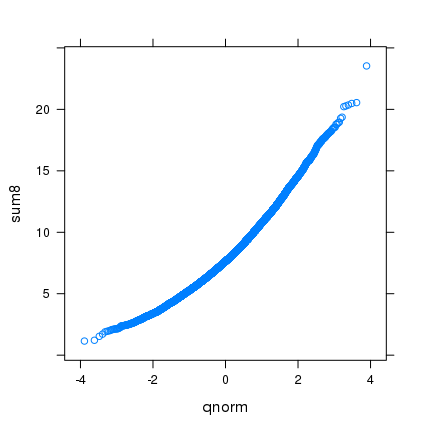
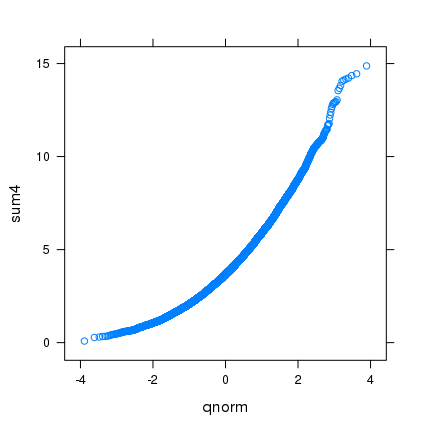
> sim8<-rexp(10000,1)

> sum4<-sim1+sim2+sim3+sim4

> sum8<-sim1+sim2+sim3+sim4+sim5+sim6+sim7+sim8

> qqmath(sum4)

> qqmath(sum8)



**Inferential Statistics (What it is all about)**

Given a population and a feature of interest, use data gathered from the population to provide a good estimate of the feature of interest.

**Example 1.** What percent of registered voters in Michigan have a favorable view of Governor Whitmer?

Population: Set of registered voters in Michigan

Feature of interest: p = the proportion of people in the population that have a favorable view of Whitmer.

The feature p = **population proportion** is a **population parameter.**

**Example 2.** What is the average household income for all households in Grand Rapids?

Population: Set of households in GR

Feature of interest:  = average household income for all households in the population.

The feature  = **population mean** is a **population parameter**.

**Basic Terminology**

**Population** The collection of “individuals” we want to know something about

**Parameter** A number that describes a feature of the population (**population parameter**)

**Sample**  A subset of the population that we have data for.

**Statistic**  A number calculated from the sample

**Example**  We have a biased coin that has an unknown probability p of producing a head.

Population All possible tosses of the coin

Parameter p = proportion of times a head occurs. (p = population proportion)

Sample Toss the coin 100 times

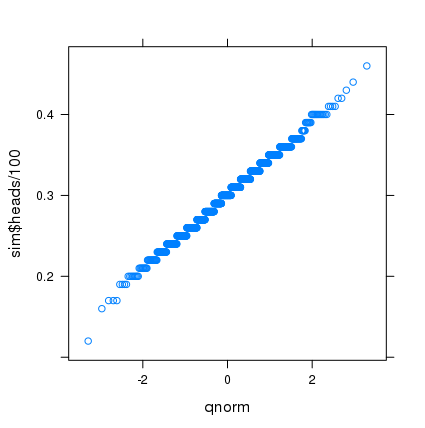
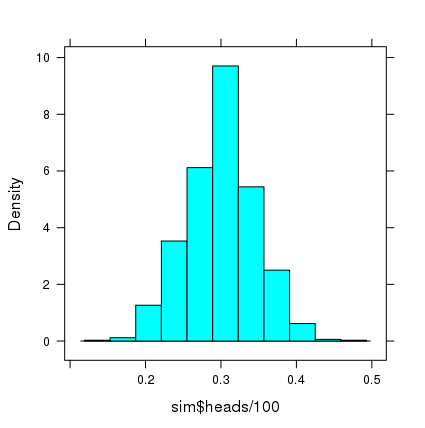
Statistic = proportion of tosses that are heads. (= sample proportion)

**Sampling distribution** The distribution of  as a random variable defined on all sets of 100 tosses.

> sim<-do(1000)\*rflip(100,p)

> histogram(sim$heads/100)

> qqmath(sim$heads)



> fitdistr(sim$heads/100,"normal")

mean sd

0.299880000 0.045727296

(0.001446024) (0.001022493)

**Exercises 10**

1. If X and Y are independent random variables, Var(X+Y) = Var(X) + Var(Y). What about Var(X-Y)? Your first guess might be that Var(X-Y) = Var(X) – Var(Y). But, this cannot be true, since if Var(Y) > Var(X) we would have Var(X-Y) < 0. Actually, If X and Y are independent, Var(X-Y) = Var(X) + Var(Y). Prove this using General Rules (2) and (5).
2. Suppose X is N(10,3), Y is N(6,2), and X and Y are independent.
3. What is the distribution of 2X – Y?
4. What is the distribution of X+Y?
5. What is P(X ≤ 4)?
6. What is P(Y ≥ 2)?
7. What is P(1≤ 2X-Y≤4)?
8. What is P(X+Y≤5)?
9. What is P(X≤Y)? (Hint: P(X≤Y) = P(X-Y≤ 0). Use the distribution of X-Y.)
10. Suppose X is a random variable with mean = 6 and sd = 2, Y is a random variable with

mean = -1 and sd = 2, and X and Y are independent.

1. What are the mean and standard deviation of 2X-Y?
2. What are the mean and standard deviation of 3X+4Y?
3. Let X and Y be independent gamma(shape = 3, rate = 4) random variables. Let S = X+Y and D = X-Y. Is a gamma distribution a good fit for S? We use a simulation with 10000 repetitions to test this.

> simx<-rgamma(10000,3,4)

> simy<-rgamma(10000,3,4)

> sum<-simx+simy

> fitdistr(sum,"gamma")

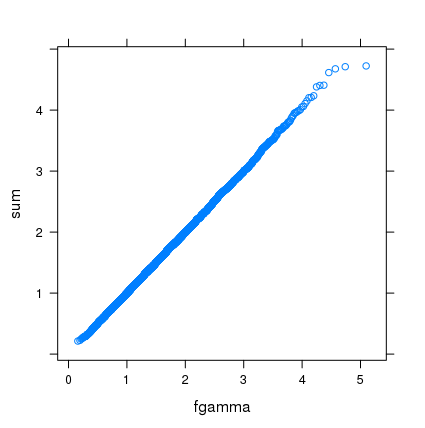
shape rate

6.06651329 4.04044750

(0.08354057) (0.05800790)

> fgamma<-function(p) qgamma(p,6.06651329,4.04044750)

> qqmath(sum,distribution=fgamma)

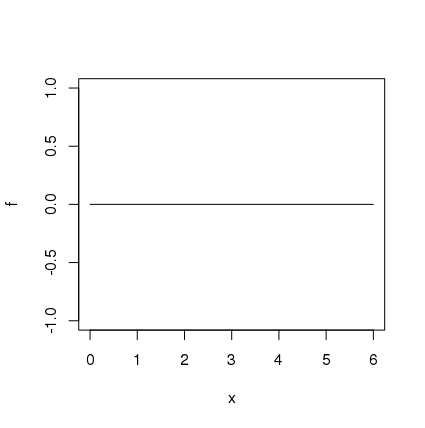


So, it looks like the sum of two independent gamma RV’s with the same shape and scale also has a gamma distribution.

1. We would not expect D to have a gamma distribution, since values for D can be negative and a gamma RV cannot be negative. Use simulations and a qq-plot to check whether it is reasonable to conclude that D has a normal distribution. Include the R commands and qq-plot.
2. Suppose X is gamma( 3,4) and W is gamma(1,5) and W is independent of X. Use simulations as in the X+Y case to check whether it is reasonable to conclude that X+W has a gamma distribution. . Include the R commands and qq-plot.
3. Another important distribution in applications is the chisquare distribution. The chisquare distribution is a special case of the gamma distribution. If n is a positive integer, then X has a chisquare distribution with n degrees of freedom if X has a gamma distribution with shape = n/2 and rate = ½; i.e., chisq(n) = gamma(n/2,1/2) . R has the chisquare distribution built in: dchisq(x,n), pchisq(x,n), qchisq(p,n), rchisq(k,n), where n = df’s. This is illustrated below.

> f<-function(x) (dchisq(x,3)-dgamma(x,3/2,1/2))

> plot(f,0,6)



Let Z be the standard normal random variable; i.e., Z is N(0,1). What kind of distribution does have? One of the following is true about Z2. It has a normal distribution or it has a chisquare distribution with 1 degree of freedom. Use a simulations and qq-plots to determine which of these two alternatives is correct. Give the two qq-plots along with your answer.