

Let x_1, x_2, \dots, x_n be a sample from some population with mean μ and standard deviation σ .

The sample mean is a statistic based on the sample. Others are the sample median, the sample standard deviation etc. Each of these is random, and has a **sampling distribution**. Usually this reflects the parent population in some way.

For example, the sample mean, \bar{x} , has mean μ and standard deviation σ/\sqrt{n} . That is it is centered around the population mean, but has a smaller spread. Furthermore, the central limit theorem tells us that the sampling distribution is approximately normal. This means we can answer questions about probabilities using the normal distribution provided n is large enough.

This project investigates sampling distributions. To do so, we will take a sample from the distribution of the statistic. (There are two populations, the parent population and the population for the statistic!) Based on the sample from the statistic we will assess normality etc.

There is a convenient function to find the sample from the sampling distribution. Download it with the command:

```
> source("http://www.math.csi.cuny.edu/st/R/sim.R")
```

The function is called `sim()` and it is used with the following arguments

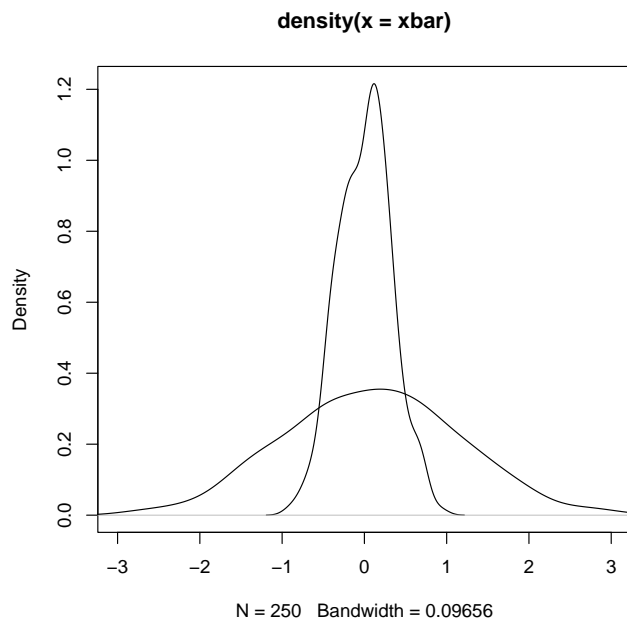
```
sim(m = number of simulations,  
    n = size of sample,  
    family = what family,  
    statistic = name of statistic)
```

The default values are $m = 200$, $n = 10$, the normal family and the statistic is the mean.

Let's see how it works,

First, suppose the population x_1, x_2, \dots, x_n is normal with mean 0 and variance 1. Let $n = 9$. The central limit theorem says that the distribution of \bar{x} should be normal with mean 0 and standard deviation $1/\sqrt{n}$. Let see.

```
> x = sim(n = 1)  
> xbar = sim(n = 9)  
> plot(density(xbar), xlim = c(-3, 3))  
> lines(density(x))
```



These commands produced the graphic. To explain, the first command `x=sim(n=1)` finds \bar{x} for $n = 1$. This is a fancy way of describing the parent distribution. The second line is the sample mean for $n = 9$. The `plot()` command draws the density of `xbar`. We do this one first, as otherwise the figure won't have a big enough y axis. However, this one won't have a large enough x axis by default so we make it wider with `xlim=c(-3,3)`.

Now, look at the graphs. Both should be normal looking, and centered at 0, but the standard deviations should be $1/3$ and 1, so noticeably different.



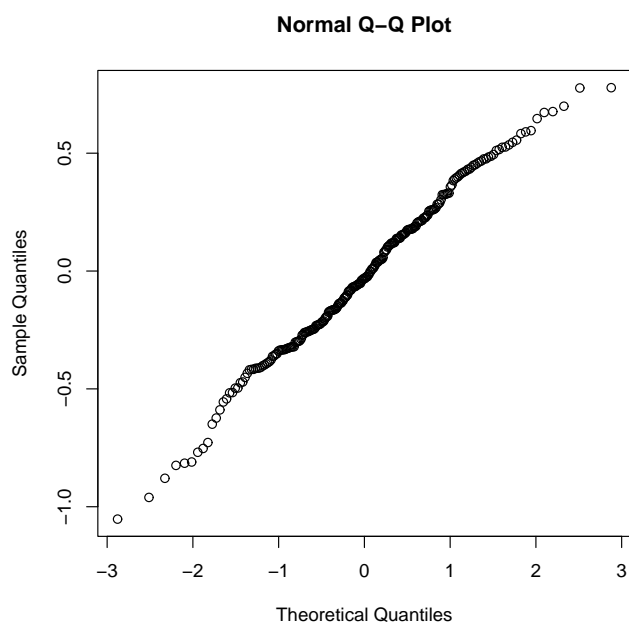
Question 1: Repeat the above. Do you get the same figure?



Question 2: The value of `xbar` are 200 random numbers. Does the empirical rule apply to them? That is, what percent are less than $1/3$ in absolute value? (hint: `sum(abs(xbar) < 1/3)`.)

The density plot is okay for looking for normality, but a quantile plot is better. These are made with the `qqnorm()` function. To see that the data is `xbar` is approximately normal we can look at the quantile plot

```
> qqnorm(xbar)
```




The quantile plot should be roughly a straight line.

The `sim()` function can be used with different families and different statistics.


For example, to simulate the median do these commands


```
> xmed = sim(n = 9, statistic = "median")
```

 Question 3: Make side by side boxplots of `xmed` and `xbar` (both for $n = 9$). Do they have the same center? Do they have the same spread?

To change the parent population of the sample can also be done. The exponential distribution has a mean and standard deviation given by the reciprocal of its rate. This is specified with the argument `rate=`. For example, this produces a sample of \bar{x} with $n = 9$ from an exponential distribution with mean 10.


```
> x = sim(n = 1, family = "exp", rate = 1/10)
> xbar = sim(n = 9, family = "exp", rate = 1/10)
```

 Question 4: Repeat the making of the density plots. Do things look normal now?

 Question 5: Look for normality of both `x` and `xbar` using a quantile plot. Are they straight lines?

Now repeat with $n = 50$.

```
> xbar.50 = sim(n = 50, family = "exp", rate = 1/10)
```

 Question 6: Make a quantile plot of `xbar.50`. Are things normal now. Find the mean and standard deviation of `xbar.50`. How do they related to the original mean and standard deviation of 10?



Question 7: For the exponential family with rate $1/10$, the median is about 6.9. For $n = 50$ can you describe the sampling distribution of the sample median? What is its shape? What is the center? Guess the spread?



Question 8: Let the parent population be normal with mean 0 and variance 1. For $n = 2, 9$ and 25 look at the sampling distribution of s , the sample standard deviation (`statistic="sd"`). Is it normal? What is the mean? How does it relate to n ? Is the *variance* related to n ?



Question 9: For $n = 1, 9, 25, 100$ store results of a simulation of \bar{x} . Make a boxplot of all four samples at once. Do this for the normal and the exponential. Is there a difference?

There are many different families in R each requiring different parameters. Here is a table of some common one

family	parameters	comments
norm	mean, sd	The normal distribution
exp	rate	Exponential. The rate is $1/\mu$, $\sigma = \mu$. This is symmetric, but not long tailed
t	df	the “t” distribution. Long tailed for small values of df , as this gets big, gets normal. Mean is 0, σ goes to 1 as df gets big.
chisq	df	The chi-square distribution. A positive distribution. Starts of skewed, but as df gets bigger, it gets normal looking. Mean and variance related to df
lnorm	meanlog, meansd	The log-normal distribution. Take a normal r.v. X and this is distribution of e^X . (That is, take a log and it is normal). Very long tailed and skewed.

Table 1: families in R



Question 10: The Central limit theorem says that as n gets large the distribution becomes normal. The size of n varies with the shape of the parent distribution. See how big n is for the families **t** with **df**=50 (symmetric, short tailed), **t** with **df**=5 (symmetric, long tailed), **exp** with the default (**rate**=1) (skewed, short tailed), **lnorm** with the default (**meanlog**=0, **meansd**=1) (skewed, long tail.)