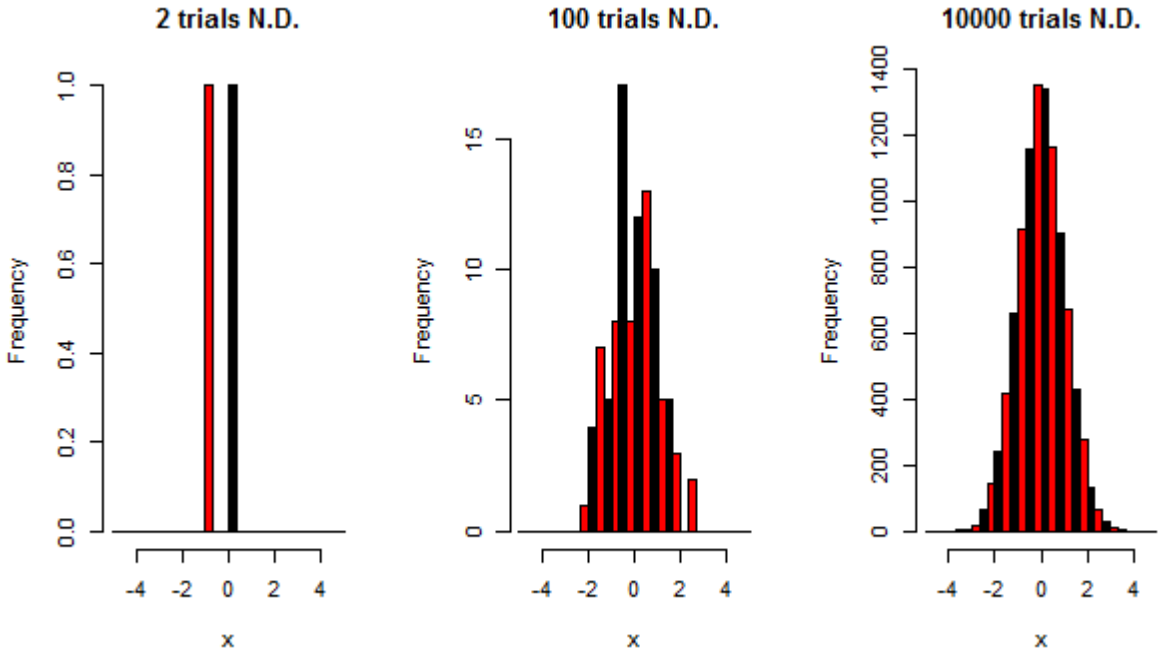
1. Show (with the use of histograms) the distribution of sample values generated using normal distribution. Is there a difference in the sample's distribution if we generate 2 numbers, 100 numbers or numbers? What is the shape of the distribution? Use the $\mu = 0$ and $\sigma = 1$.

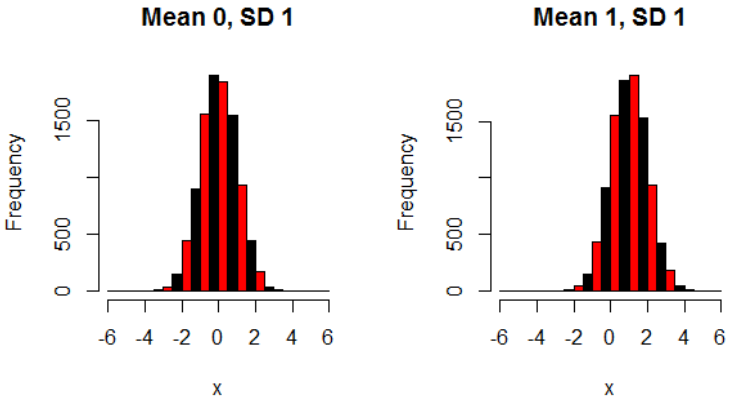



```
> br = seq(-5,5,1/3)
> colors = c("red", "black")
> mean = 0
> sd = 1
>
> par(mfrow=c(1,3))
> x <- rnorm(2,mean,sd)
> hist(x, breaks = br, col = colours, main = "2 trials N.D.")
>
> x <- rnorm(100,mean,sd)
> hist(x, breaks = br, col = colours, main = "100 trials N.D.")
>
> x <- rnorm(10000,mean,sd)
> hist(x, breaks = br, col = colours, main = "10000 trials N.D.")
```

Again, when we have two trials it is hard to discern what the distribution will be like. With 100 trials we start to see the bell-shaped pattern but there are still some buckets that differ wildly from expectation such as the tallest line being -1.33 to -1.67 and no results from 2.00 to 2.33. Finally for 10,000 trials we are finally seeing a very-good approximation. The two tallest buckets are in the middle at the mean and as we move away from the mean each bucket is smalller than the one that precedes it.

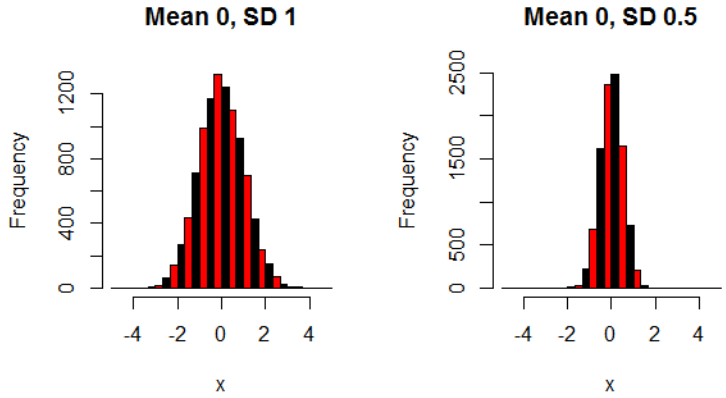
2. How do the results change if you change the parameters μ and σ?

```
> par(mfrow=c(1,2))
> br = seq(-6,6,0.5)
>
> x <- rnorm(10000,mean,sd)
> hist(x, breaks = br, col = colours,
main = "Mean 0, SD 1")
>
> mean = 1
>
> x <- rnorm(10000,mean,sd)
> hist(x, breaks = br, col = colours,
main = "Mean 1, SD 1")
```



The two histograms look very similar but the two graphs peak about their means. Since they have two different means, 0 and 1 respectively, they peak about different points.

```
> br = seq(-5,5,1/3)
> colors = c("red", "black")
> mean = 0
> sd = 1
> par(mfrow=c(1,2))
>
> x <- rnorm(10000,mean,sd)
> hist(x, breaks = br, col = colours,
main = "Mean 0, SD 1")
>
> sd = 0.5
>
> x <- rnorm(10000,mean,sd)
> hist(x, breaks = br, col = colours,
main = "Mean 0, SD 0.5")
```



When we change only the standard deviation we see that both histograms peak about the same point, the mean 0, the one with the larger standard deviation, on the left, looks fatter but note the y-axis on the right has larger numbers. Remember that the area under the curve must equal 1. Therefore if one distribution has taller around the mean then the other must be taller around the tails.

3. Suppose that X has normal distribution for which the mean is 1 and the variance is 4. Find the value of each of the following probabilities:

NOTE: As shown in part c, P(X = x) = 0 therefore we can use equivalently P(X < x) and P(X <= x)

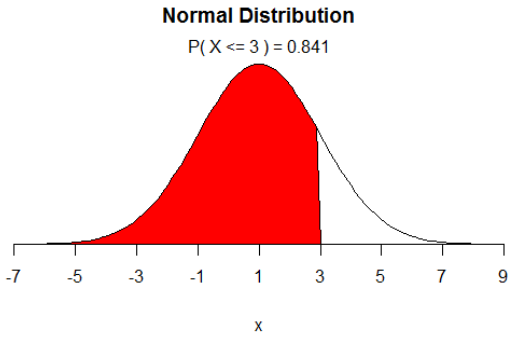
```
> shadebetween <- function(lower, upper, mean=1, sd=2)
+ {
+
+   par(mfrow=c(1,3))
+
+   shadebelow(upper)
+   shadebelow(lower)
+
+   lb=lower; ub=upper
+
+   x <- seq(-4,4,length=100)*sd + mean
+   hx <- dnorm(x,mean,sd)
+
+   plot(x, hx, type="n", ylab="",
+        main="Normal Distribution", axes=FALSE)
+
+   i <- x >= lb & x <= ub
+   lines(x, hx)
+   polygon(c(lb,x[i],ub), c(0,hx[i],0), col="red")
+
+   area <- pnorm(ub, mean, sd) - pnorm(lb, mean, sd)
+   result <- paste("P(",lb," < X <",ub,") =",
+                   signif(area, digits=3))
+   mtext(result,3)
+   axis(1, at=seq(mean-4*sd, mean+4*sd, sd), pos=0)
+   #modified from http://www.statmethods.net/advgraphs/probability.html
+ }
```

The function shadebetween colours in the appropriate area under a normal curve. The functions shadebelow and above are similar but take one number with shadebelow setting the lower bound to mu–4sd, and shadeabove setting the upper bound to mu+4sd.

```
> mu = 1
> sd = 2
```

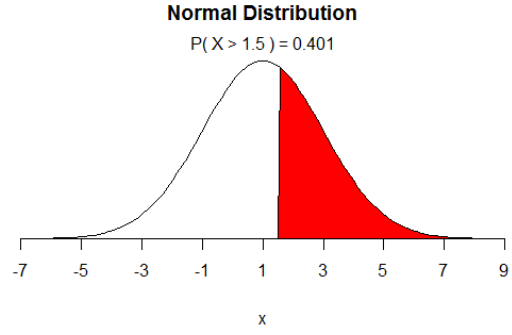
(a) $P(X \leq 3)$

```
> shadebelow(3)
> pnorm(3,mu,sd)
[1] 0.8413447
> pnorm((3-mu)/sd)
[1] 0.8413447
```



(b) $P(X > 1.5)$

```
> shadeabove(1.5)
> 1-pnorm(1.5,mu,sd)
[1] 0.4012937
> pnorm(1.5,mu,sd, lower.tail = F)
[1] 0.4012937
> pnorm((1.5-mu)/sd, lower.tail = F)
[1] 0.4012937
```



(c) P(X = 1)

I saw this addressed in the forum but there are two ways to get an answer of 0. One is that for continuous distributions the probability between a and b is given by f '(x) evaluated from b to a or f '(b) – f '(a). P(X = 1) is equivalent to P(1 < X < 1) or f ' (1) – f '(1) which must equal zero.

Alternatively we can use 1 – (area above the point + area below the point)

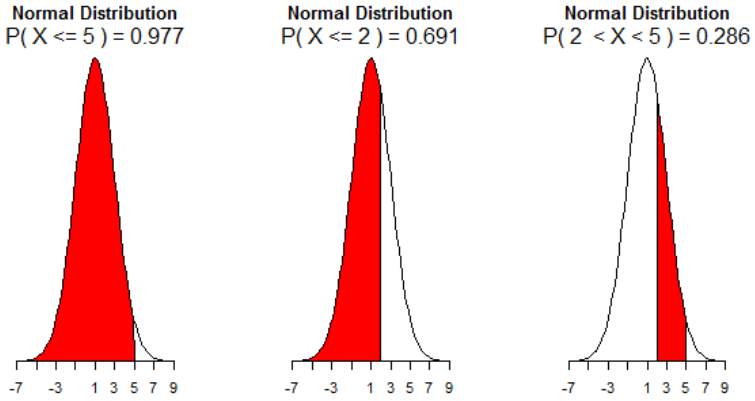
```
> 1-(pbinom(1.999999,10,0.5) + (1 - pbinom(2.000001,10,0.5)))
[1] 0.04394531
> 1-(pnorm(1.999999,mu,sd) + (1 - pnorm(2.000001,mu,sd)))
[1] 1.079819e-07
```

```
First of all the answer using pbinom is correct because it is dicrete
> dbinom(2,10,0.5)
[1] 0.04394531

We cannot get 0 for continuous because we cannot use infinite precision.
However the value is equal to width*height – height being dnorm – which should
not be equal but is because of r's lack of precision:
> 0.000002*dnorm(2,mu,sd)
[1] 1.079819e-07
```

(d) P(2 < X < 5)

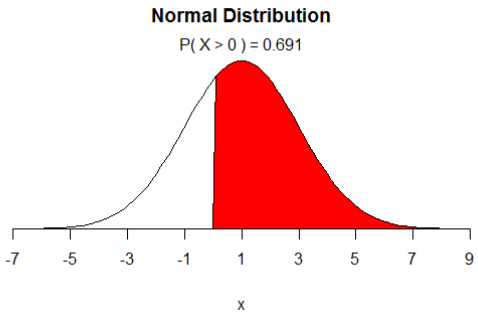
```
> shadebetween(2,5)
> pnorm(5, mu, sd)
[1] 0.9772499
> pnorm(2, mu, sd)
[1] 0.6914625
> pnorm(5, mu, sd) - pnorm(2,
mu, sd)
[1] 0.2857874
```



(e) P(X ≥ 0)

Thankfully from our earlier result that P(X=x) = 0 we can reformulate this as P(X > 0)

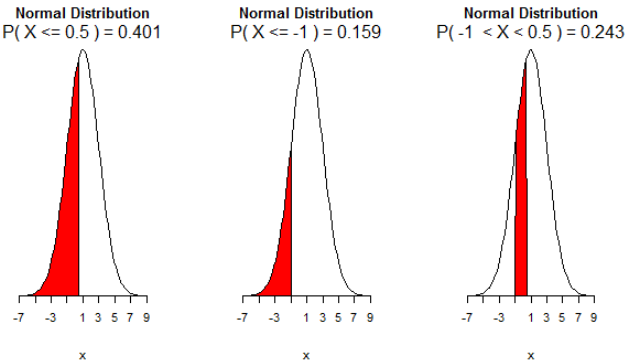
```
> shadeabove(0)
> pnorm(0,mu,sd, lower.tail = F)
[1] 0.6914625
```



(f) P(−1 < X < 0.5)

Similarly here we can use P(-1 < X <= 0.5)

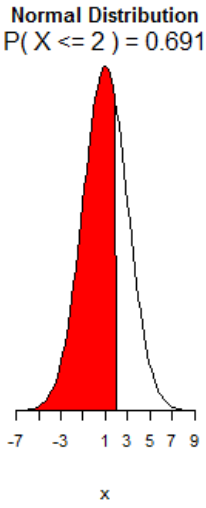
```
> shadebetween(-1,0.5)
> pnorm(0.5,mu,sd)
[1] 0.4012937
> pnorm(-1,mu,sd)
[1] 0.1586553
> pnorm(0.5,mu,sd) -pnorm(-1,mu,sd)
[1] 0.2426384
```

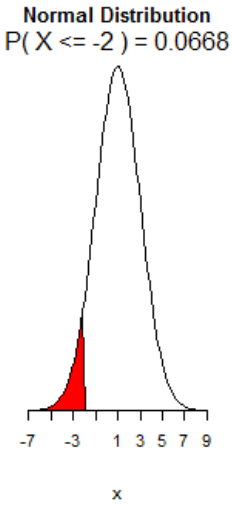


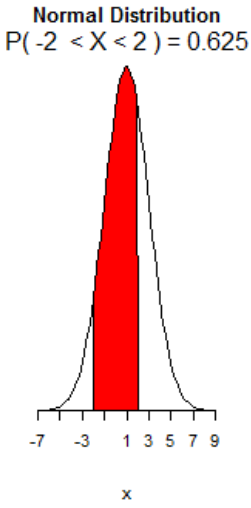
(g) P(|X| ≤ 2)

equivalent to P(-2 <= X <= 2)

```
> shadebetween(-2,2)
```







(h) P(1 ≤ −2X + 3 ≤ 8)

i)

1 <= -2x+3

-2 <= -2x

1 >= x

x <= 1

ii)

-2x+3 <= 8

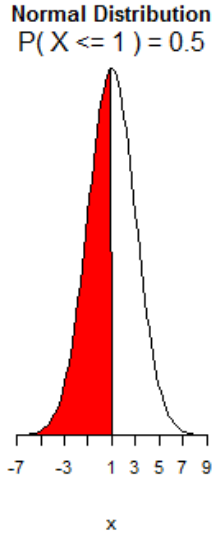
-2x <= 5

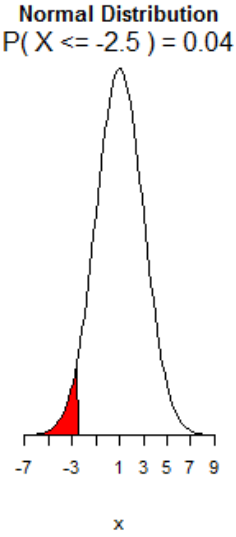
x >= -2.5

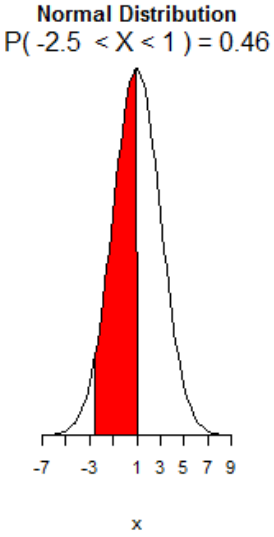
-2.5 <= x

-2.5 <= x <=1

```
> shadebetween(-2.5,1)
```







4. Suppose that the measured voltage in a certain electric circuit has the normal distribution with mean 120 and standard deviation 2. If three independent measurements of the voltage are made, what is the probability that all three measurements will lie between 116 and 118?

```
> in_range_measurement = pnorm(118, 120, 2) - pnorm(116, 120, 2)
> in_range_measurement
[1] 0.1359051
> in_range_measurement**3
[1] 0.002510195
```

We cube it because the measurements are independent.

5. Show that standard normal distribution is symmetric. Is the normal distribution symmetric as well?

Normal Probability Density Function

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

To show the distribution is symmetric we need to show that values that are equidistant from mu are equal i.e:

f(mu + a) = f(mu – a)

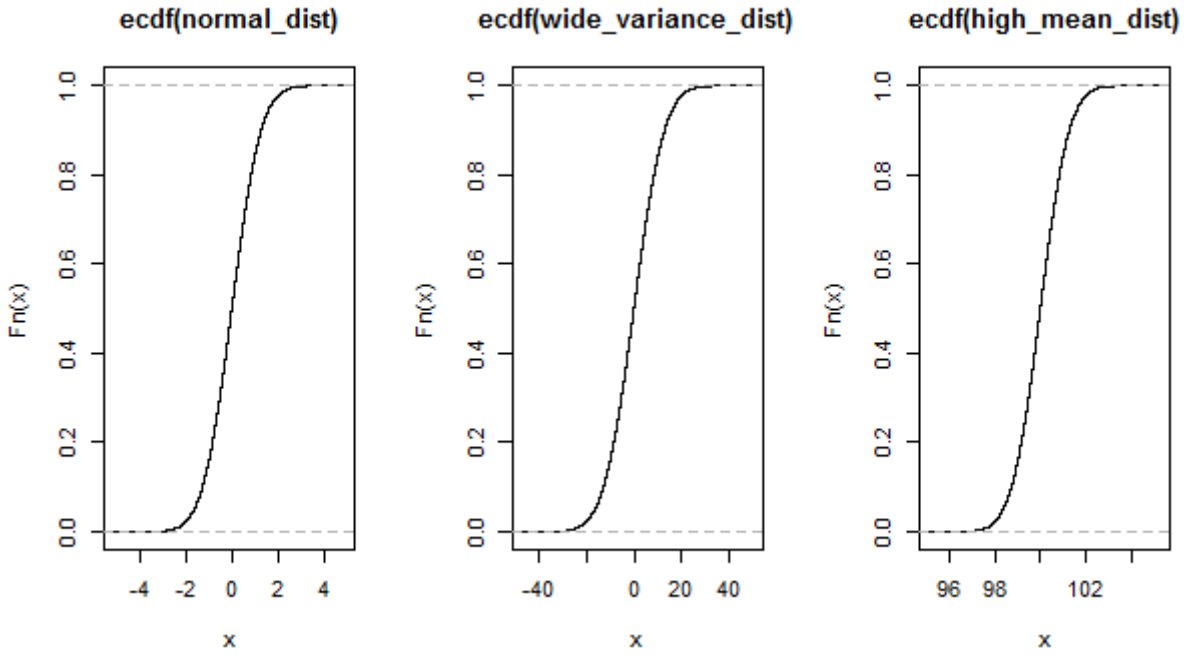
Fortunately the only place x appears is in -(x-mu)^2 so we can ignore the rest of the rest as they are constants if we are using the same distribution i.e equal mean and variance.

f(mu + a) =. - ((mu+a) – mu)^2 = - (a)^2 = -a^2
f(mu – a) = - ((mu-a) – mu)^2 = - (-a)^2 = -a^2

Happily this is true for all means and variances – except 0 variance because of division by zero but you don't need a distribution then - so all normal distributions are symmetric.

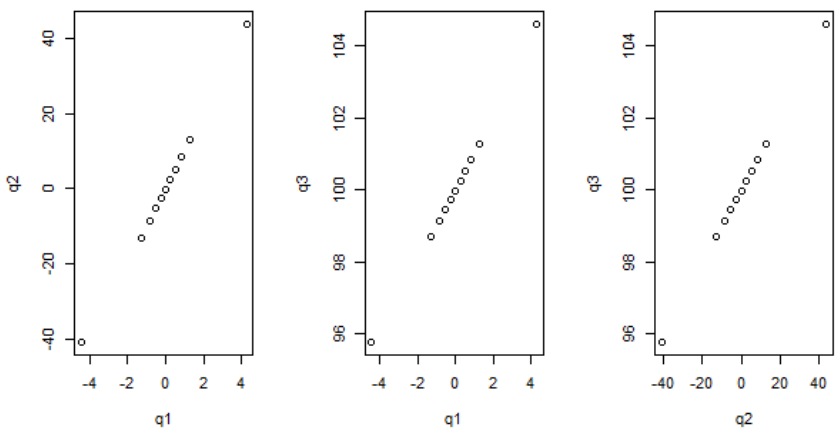
6. How does the CDF function look like for the normal distribution? How does it change when you change mean and standard deviation?

```
> mean1 = 0
> mean2 = 100
> sd1 = 1
> sd2 = 10
>
> normal_dist = rnorm(100000, mean1, sd1)
> wide_variance_dist = rnorm(100000, mean1, sd2)
> high_mean_dist = rnorm(100000, mean2, sd1)
>
> plot(ecdf(normal_dist))
> plot(ecdf(wide_variance_dist))
> plot(ecdf(high_mean_dist))
```



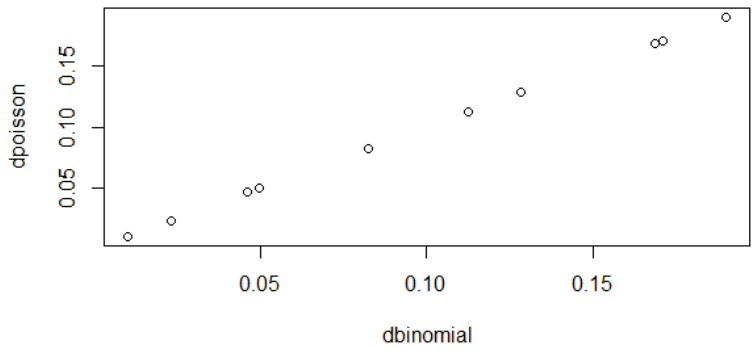

```
> qqplot(dbinomial, dpoisson)
> breaks = seq(0,1,0.1)
>
> q1 = quantile(normal_dist, breaks)
> q2 = quantile(wide_variance_dist, breaks)
> q3 = quantile(high_mean_dist, breaks)
>
> plot(q1,q2)
> plot(q1,q3)
> plot(q2,q3)
```

And we can see all three plotted along the diagonal.



7. A store owner believes that customers arrive at his store at a rate of 4.5 customers per hour on average. He wants to find the distribution of the actual number X of customers who will arrive during a particular one-hour period later in the day. He models customer arrivals in different time periods as independent of each other. As a first approximation, he divides the one-hour period into 3600 seconds and thinks of the arrival rate as being 4.5/3600 = 0.00125 per second. He then says that during each second either 0 or 1 customers will arrive, and the probability of an arrival during any single second is 0.00125. He then tries to use the binomial distribution with parameters n = 3600 and p = 0.00125 for the distribution of the number of customers who arrive during the one-hour period later in the day. Based on this example, show how Poisson distribution can be used to approximate binomial distribution.

```
> p = 4.5/3600
> n = 3600
> par(mfrow=c(1,1))
>
> dbinomial <- dbinom(1:10, n, p)
> dpoisson = dpois(1:10, 4.5)
>
> plot(dbinomial,dpoisson)
> cor(dbinomial,poisson)
[1] 0.9999998
```



The quantiles plot along thediagonal and the correlation is nearly perfect.