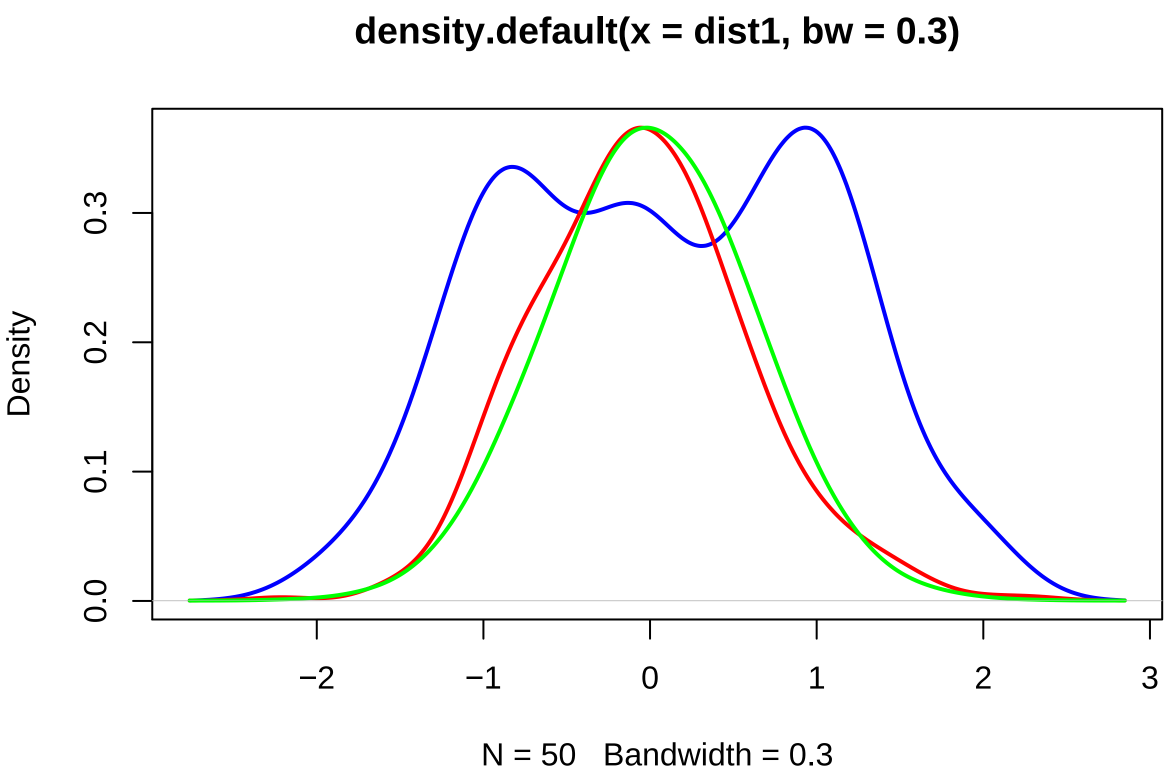
QQ Plot Tutorial

Ziao JU

June 4, 2015

PART I: the relatinoship between the sample size and normality.

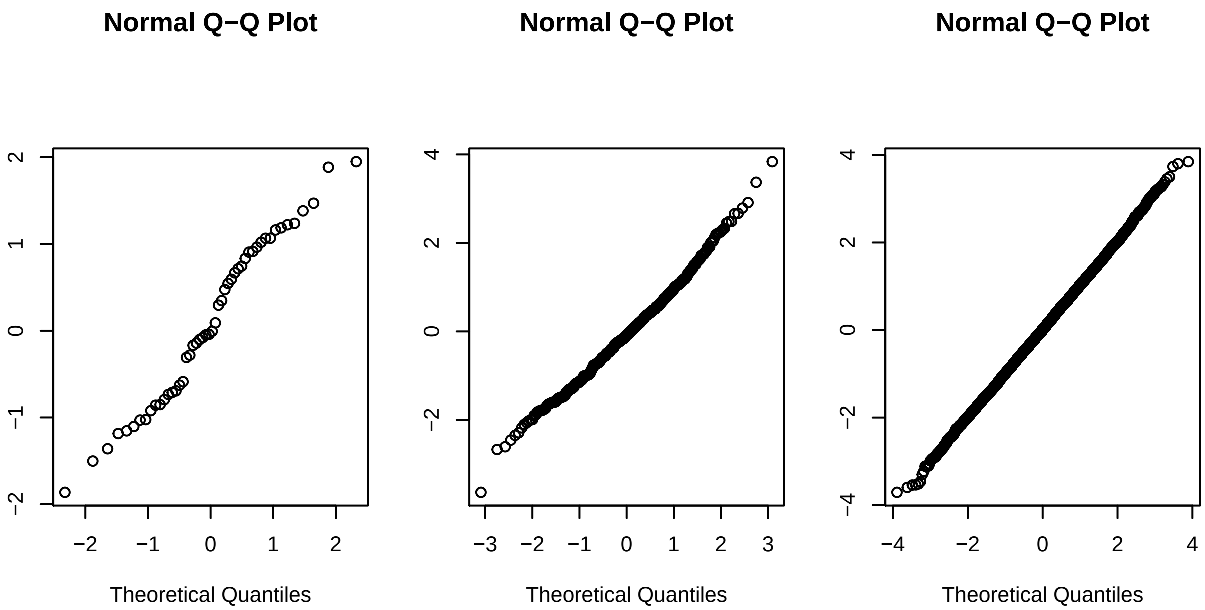
```
set.seed(133)
dist1 = rnorm(50)
dist2 = rnorm(500)
dist3 = rnorm(10000)

plot(density(dist1, bw = 0.3), lwd = 2, col = "blue")
par(new = TRUE)
plot(density(dist2, bw = 0.3), lwd = 2, col = "red",
     axes = FALSE, xlab = NA, ylab = NA, main = NA)
par(new = TRUE)
plot(density(dist3, bw = 0.3), lwd = 2, col = "green",
     axes = FALSE, xlab = NA, ylab = NA, main = NA)

legend(1, 2, lwd = c(2,2,2), cex = 0.5,
legend = c("dist 1", "dist 2", "dist 3"),
col = c("blue", "red", "green"), bty = "n")
```



As we can see from the density curves, as the sample size increases, the distribution resembles more and more closely a standard normal curve. Next let's explore how their qq norm curves

look like.

```r
par(mfrow = c(1,3), mar = c(10,2,10,2))
qqnorm(dist1)
qqnorm(dist2)
qqnorm(dist3)
```

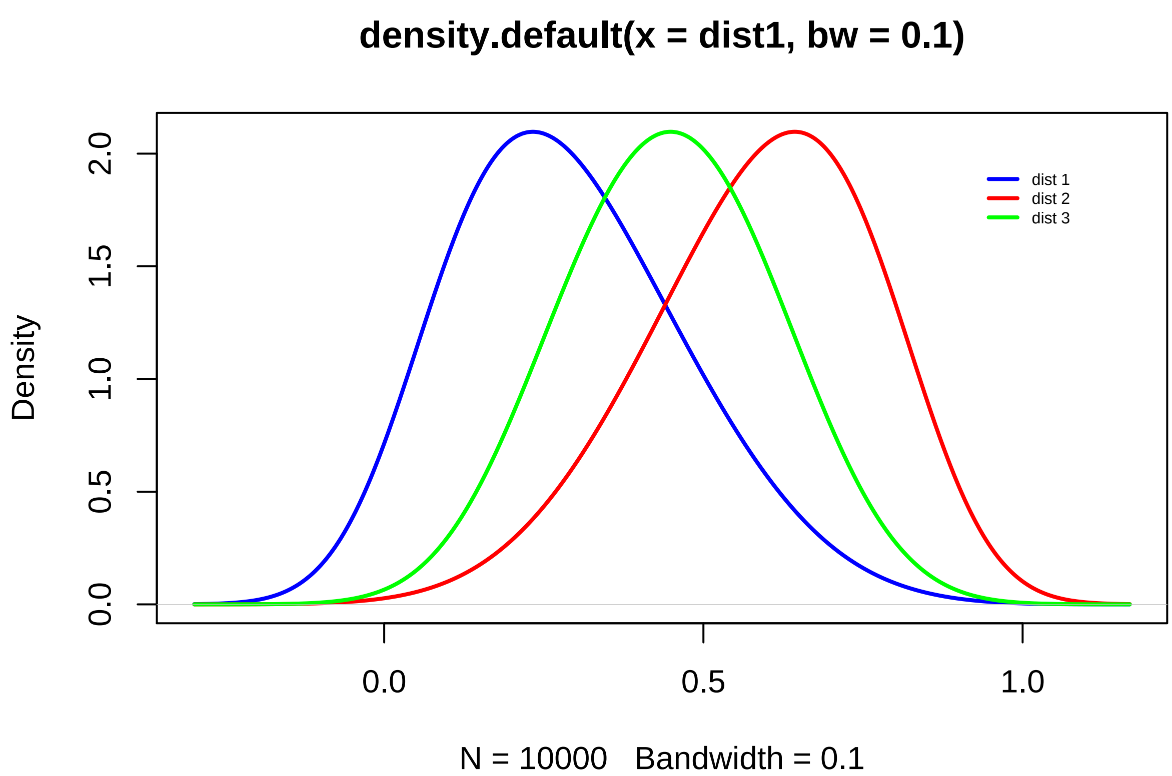


We can see that as the sample size increases, the qq plot is getting closer to a straight line passing through (0,0).

PART II: skewness

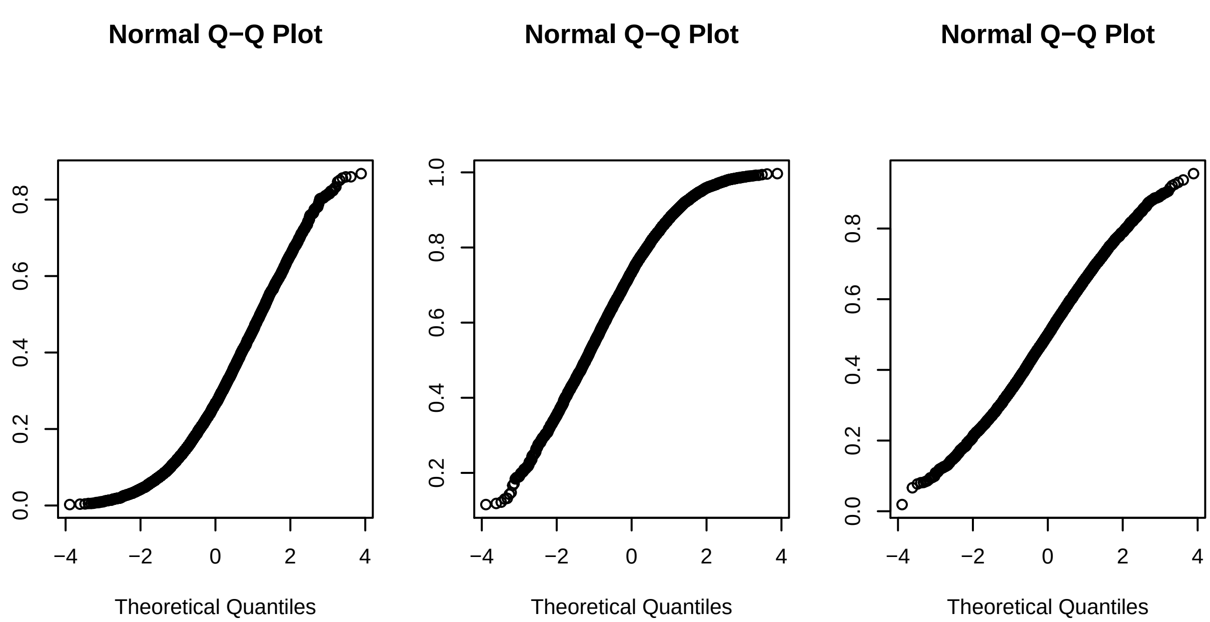
We have three distributiosn, dist 1 (blue) is right (negtively) skewed, dist 2 (red) is left (positively) skewed and dist 3 (green) is symmetrical. Now let's explore how their qq norm curves look like.

```r
set.seed(133)
dist1 = rbeta(10000, 2, 5)
dist2 = rbeta(10000, 5, 2)
dist3 = rbeta(10000, 5, 5)
plot(density(dist1, bw = 0.1), lwd = 2, col = "blue")

par(new = TRUE)
plot(density(dist2, bw = 0.1), lwd = 2, col = "red",
     axes = FALSE, xlab = NA, ylab = NA, main = NA)
par(new = TRUE)
plot(density(dist3, bw = 0.1), lwd = 2, col = "green",
     axes = FALSE, xlab = NA, ylab = NA, main = NA)

legend(1, 2, lwd = c(2,2,2), cex = 0.5,
legend = c("dist 1", "dist 2", "dist 3"),
col = c("blue", "red", "green"), bty = "n")
```



qq normal plots

```
par(mfrow = c(1,3), mar = c(10,2,10,2))
qqnorm(dist1)
qqnorm(dist2)
qqnorm(dist3)
```



Let's take a close look at the curvature of the three qq normal curves. The first plot has a upward sloping convex curve; the second has a concave curve; the last has a roughly straight line.

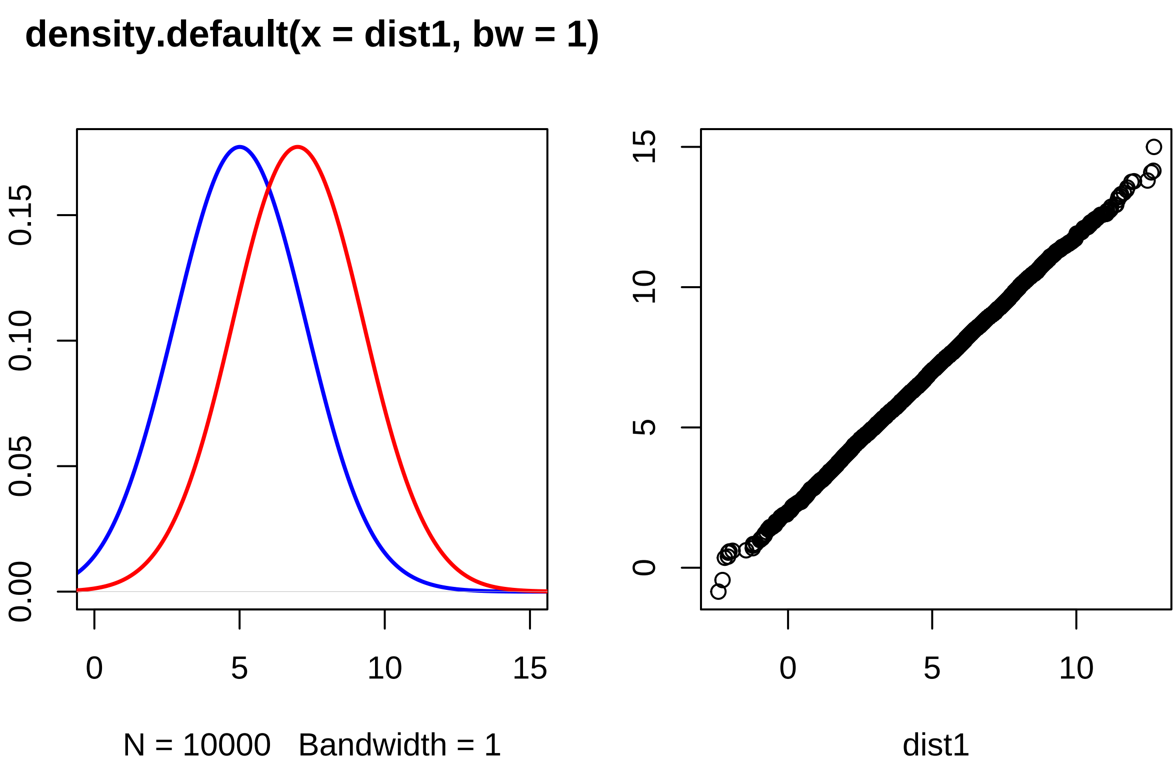
PART III: scaling and translation

Now, let's compare two normal distributions with different means or different standard deviations.

**Fact: if the two distributions are identical, then the qq plot should pass through the origin (0,0) and have a slope of 1.

Case 1: translation (same mean, different sd)

```
set.seed(133)
dist1 = rnorm(10000, mean = 5, sd = 2)
dist2 = rnorm(10000, mean = 7, sd = 2)
par(mfrow = c(1,2), mar = c(5,2,5,2))

plot(density(dist1, bw = 1), lwd = 2, col = "blue", xlim = c(0,15))
par(new = TRUE)
plot(density(dist2, bw = 1), lwd = 2, col = "red", xlim = c(0,15),
     axes = FALSE, xlab = NA, ylab = NA, main = NA)
qqplot(dist1, dist2)
```

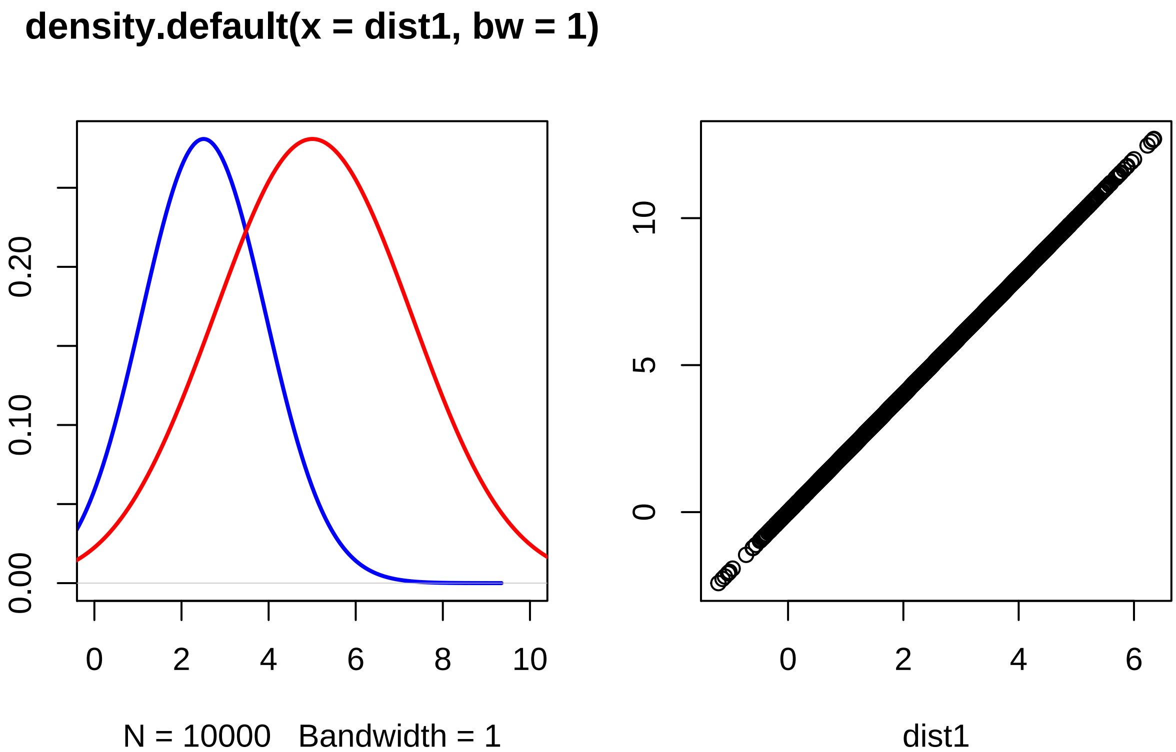


Notice that the qq plot still have slope of 1, but the y-intercept is now (0,2).

Case 2: scaling (ratio of means = ratio of sd)

```
set.seed(133)
dist1 = rnorm(10000, mean = 2.5, sd = 1)
dist2 = 2 * dist1
par(mfrow = c(1,2), mar = c(5,2,5,2))

plot(density(dist1, bw = 1), lwd = 2, col = "blue", xlim = c(0,10))
```

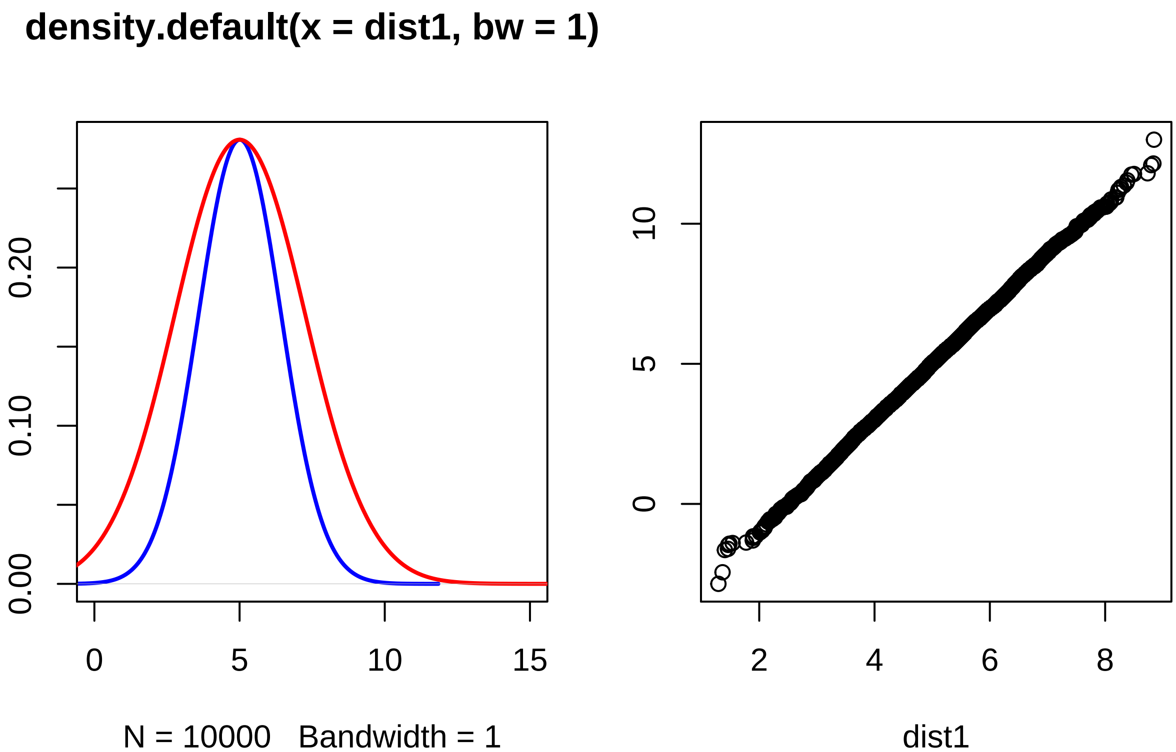
```r
par(new = TRUE)
plot(density(dist2, bw = 1), lwd = 2, col = "red", xlim = c(0,10),
     axes = FALSE, xlab = NA, ylab = NA, main = NA)
qqplot(dist1, dist2)
```



Notice that the qq plot still passes through (0,0), but the slope is the ratio of mean/sd, which in this case is 2.

Case 3: What if the two distributions the same mean, but different sd?

```r
set.seed(133)
dist1 = rnorm(10000, mean = 5, sd = 1)
dist2 = rnorm(10000, mean = 5, sd = 2)
par(mfrow = c(1,2), mar = c(5,2,5,2))

plot(density(dist1, bw = 1), lwd = 2, col = "blue", xlim = c(0,15))
par(new = TRUE)
plot(density(dist2, bw = 1), lwd = 2, col = "red", xlim = c(0,15),
     axes = FALSE, xlab = NA, ylab = NA, main = NA)
qqplot(dist1, dist2)
```



Notice that the qq plot has slope of 2 and intercept of (0, -5). If we change the sd of dist2 from 2 to 3, then the slope remains 2, but the y-intercept changes -5 to -10. So in general, if the two distributions have the same mean μ, but different sd, n and m respectively where $m > n$, then the slope will be $s = \frac{m}{n}$ and the y-intercept will be $-(s-1)\mu$.

```
par(mar=c(5.1, 4.1, 4.1, 2.1))
par(mfrow=c(1,1))
```