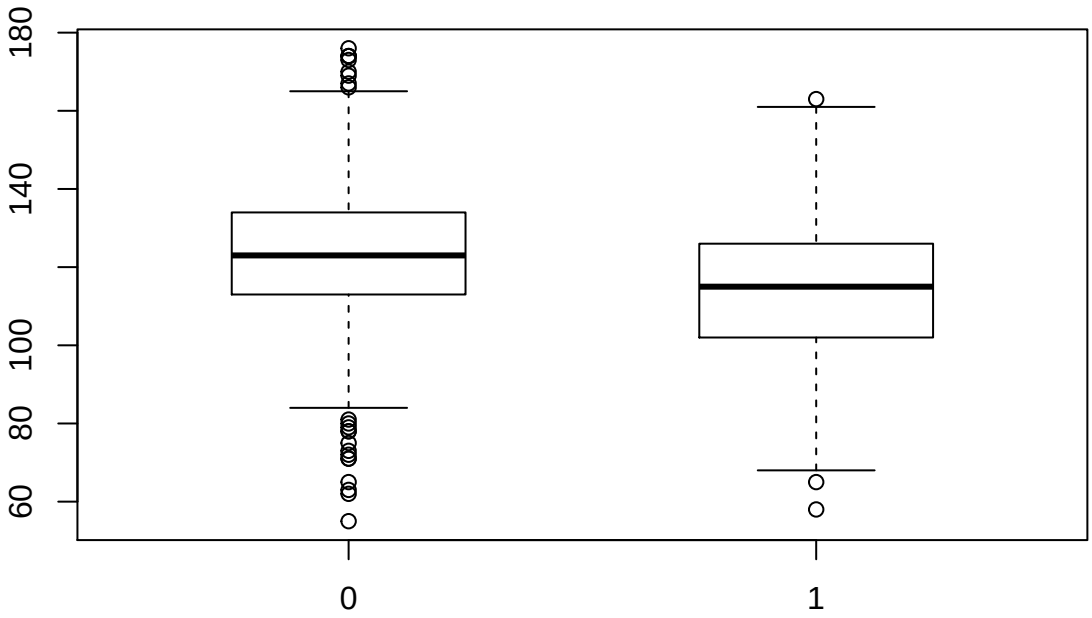
Analysis

Ziao JU

June 4, 2015

PART I:

```
baby = read.table("https://www.stat.berkeley.edu/~statlabs/data/babiesI.data", header = TRUE)
baby = baby[baby$smoke != 9, ]
baby$smoke = as.factor(baby$smoke)
baby$smoke = droplevels(baby$smoke)
boxplot(bwt ~ smoke, data = baby, boxwex = 0.5)
```



From the box plots above, we can see that the two distributions seem to have different means, but approximately same spread. To confirm this, let's check the standard deviation of each distribution

```
smoke = baby[baby$smoke == 1, ]
nosmoke = baby[baby$smoke == 0, ]
sd(smoke$bwt)
```

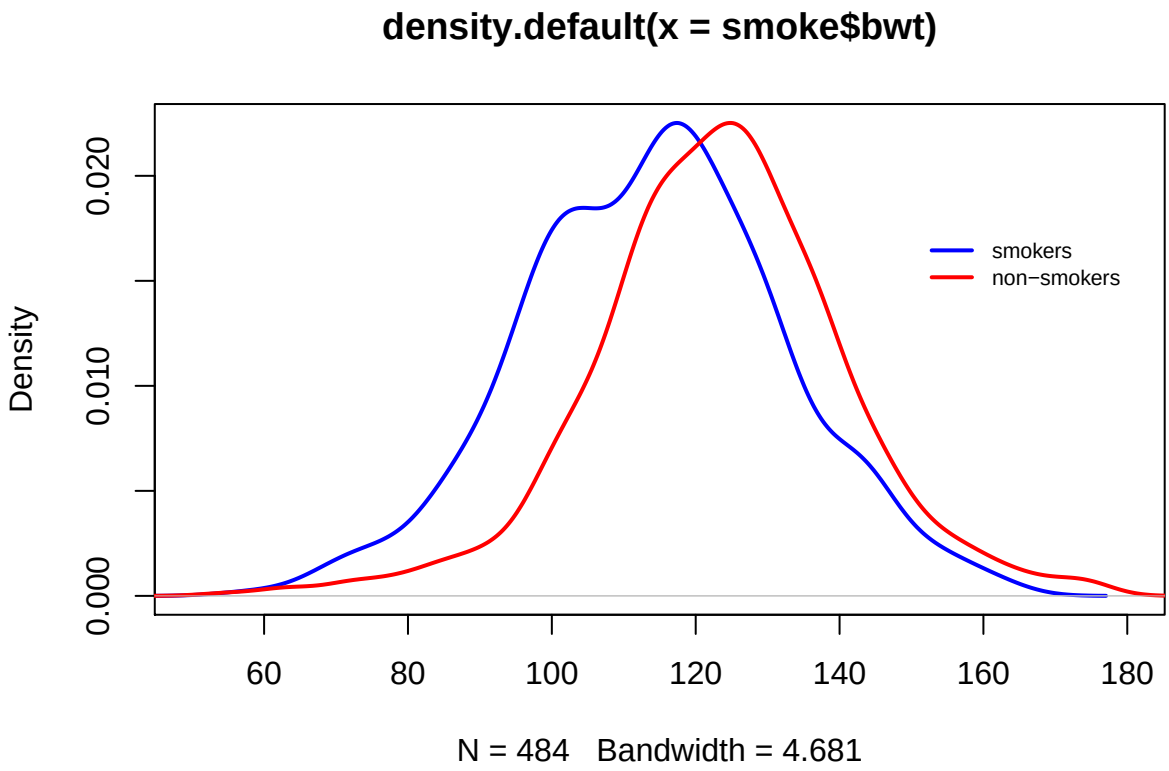
```
## [1] 18.09895
```

```
sd(nosmoke$bwt)
```

```
## [1] 17.39869
```

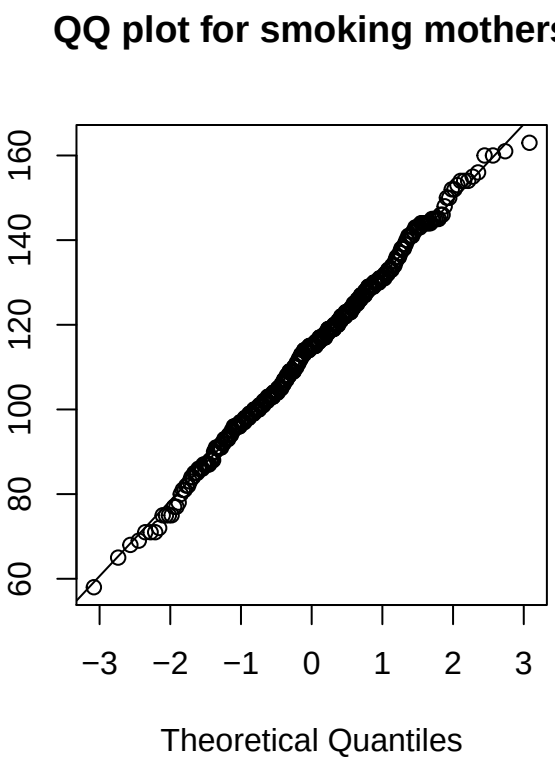
Indeed, the two distributions have very close standard deviations. Next, let's look at the density curve of each distribution:

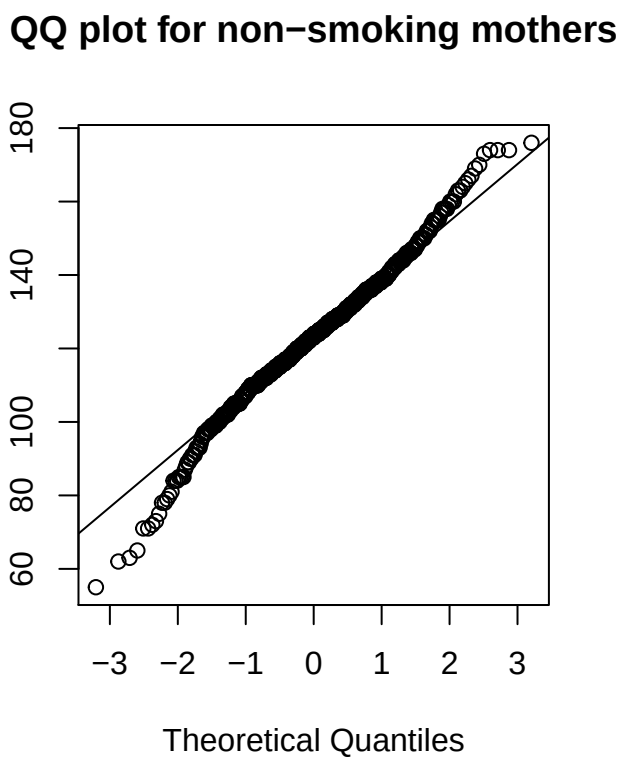
```
plot(density(smoke$bwt), lwd = 2, col = "blue", xlim = c(50, 180))
par(new=TRUE)
plot(density(nosmoke$bwt), lwd = 2, col = "red", xlim = c(50, 180),
     axes = FALSE, xlab = NA, ylab = NA, main = NA)

legend(150, 0.02, lwd = c(2,2), cex = 0.7,
       legend = c("smokers", "non-smokers"),
       col = c("blue", "red"), bty = "n")
```



Next, let's assess the normality/skewness of each distribution individuality.

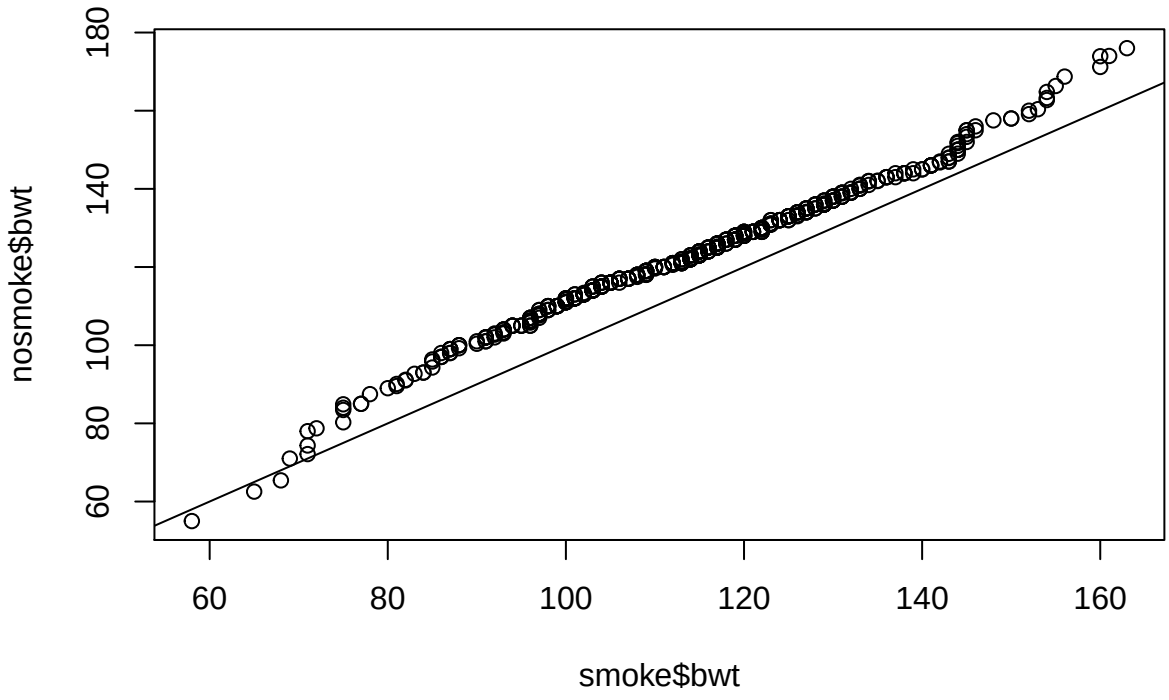
```
par(mfrow = c(1,2), mar = c(5,2,5,2))
qqnorm(smoke$bwt, main = "QQ plot for smoking mothers")
qqline(smoke$bwt)
qqnorm(nosmoke$bwt, main = "QQ plot for non-smoking mothers")
qqline(nosmoke$bwt)
```





It seems that the distribution of birth weights of babies from smoking mothers has a closer resemblance of normal distribution. Next, let's compare the two distributions together.

```
par(mar=c(5.1, 4.1, 4.1, 2.1))
par(mfrow=c(1,1))
qqplot(smoke$bwt, nosmoke$bwt)
abline(0,1)
```



From the graph, the qq plot is roughly a straight line with slope of 1; this suggests that the two distributions are approximately equal. The qq plot is above $y = x$ line, indicating that the y-intercept is above 0, because the average birth weight of babies born by non smoking moms is higher than that born by smoking moms.

To test whether the difference in average birth weights of babies is significant across two groups, let's performa a one-way ANOVE test:

```
summary(aov(bwt ~ smoke, data = baby))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## smoke          1  23400   23400   74.87 <2e-16 ***
## Residuals   1224 382529     313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ouput, we can see that the p-value is very small and hence we have enough evidence to reject the null and support that the two distributions indeed have different means.

PART II: regression analysis

Now, let's add more variables to our analysis by including gestation days, mothers' age (in years), mothers' height (in inches), and mothers' pregnancy weight (in pounds)

```
baby2 = read.table("https://www.stat.berkeley.edu/~statlabs/data/babies.data", header = TRUE)
baby2 = baby2[baby2$smoke != 9, ]
```

Let's first look at the correlation matrix to see which variables are more relevant to our analysis

```
cor.mat = cor(baby2)
print(cor.mat, digits = 3)
```

```
##               bwt gestation   parity      age  height  weight   smoke
## bwt         1.0000   0.06195 -0.04457  0.02548  0.1270  0.0474 -0.2401
## gestation   0.0619   1.00000 -0.00877 -0.00335  0.0662  0.0495 -0.0322
## parity     -0.0446  -0.00877  1.00000 -0.29788 -0.0159 -0.0696 -0.0128
## age         0.0255  -0.00335 -0.29788  1.00000  0.0485  0.0633 -0.0588
## height      0.1270   0.06617 -0.01591  0.04854  1.0000  0.6013  0.0545
## weight      0.0474   0.04948 -0.06959  0.06327  0.6013  1.0000  0.0379
## smoke      -0.2401  -0.03219 -0.01282 -0.05882  0.0545  0.0379  1.0000
```

By looking at the correlation matrix, we can first rule out mother's age and parity. There is high correlation between mother's height and mother's pregnancy weight, 0.4852; hence it might be a good idea to just keep one to avoid multicollinearity. Since mother's height has a higher correlation with birth weight, we might want to keep mother's height and drop mother's pregnancy weight. Therefore, our desired regression model is

$$bwt_i = \beta_0 + \beta_1 gestation_i + \beta_2 height_i + \beta_3 smoke_i + e_i$$

Before we run the regression analysis, let's first remove the unknown data:

```r
baby2 = baby2[baby2$gestation != 999, ]
baby2 = baby2[baby2$height != 99, ]
baby2 = baby2[baby2$weight != 999, ]
summary(lm(bwt ~ gestation + height + smoke, data = baby2))
```

```
##
## Call:
## lm(formula = bwt ~ gestation + height + smoke, data = baby2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.369 -10.341  -0.445   9.824  53.225
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -82.92467   13.86691  -5.980 2.96e-09 ***
## gestation     0.43620    0.02914  14.971  < 2e-16 ***
## height        1.31112    0.18435   7.112 1.98e-12 ***
## smoke        -8.50928    0.95274  -8.931  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.91 on 1171 degrees of freedom
## Multiple R-squared:  0.2481, Adjusted R-squared:  0.2462
## F-statistic: 128.8 on 3 and 1171 DF,  p-value: < 2.2e-16
```

Interpretation: The estimated coefficient on the *smoke* variable is about -8.5. It refers to the difference in baby's weight between a mother who smokes and a mother who doesn't smoke. The conclusion we can make here is that, keeping mother's pregnancy weight and gestation days constant, a mother who smokes is expected to have a baby 8.5 ounces lighter than a mother who doesn't smoke.