# R and Inferential Techniques (MOL8

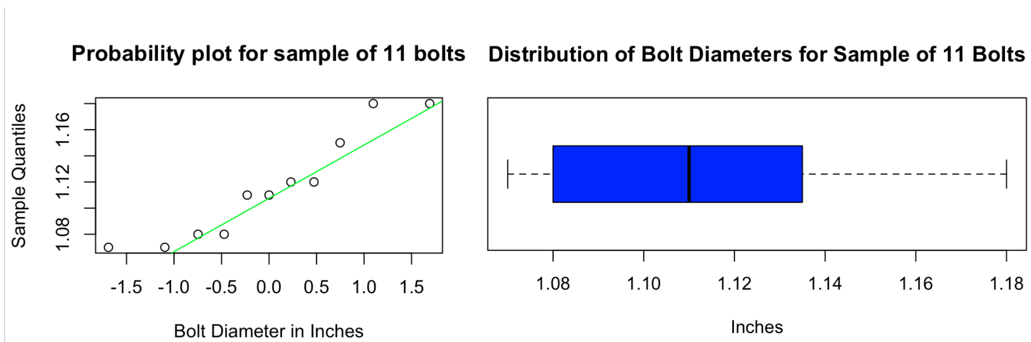
1. A bolt diameter is measured in inches. Here is a simple random sample of 11 bolt diameter measurements:

   1.11 1.07 1.11 1.07 1.12 1.08 1.08 1.15 1.18 1.18 1.12

   a. Make a probability plot and a boxplot and see if there are any major deviations from Normality and if there are any outliers. Is it reasonable to use the t procedures? Why or why not? (Copy and paste the **graphs** that result with your determination on whether it is reasonable to use the t procedures. Make sure you have meaningful titles and labels for your graph.)

**Copy and paste your code (2 pt) and resultant graphs (2 pt)**

```
> bolts <- c(1.11,1.07, 1.11, 1.07, 1.12, 1.08, 1.08, 1.15, 1.18, 1.18, 1.12)
> qqnorm(bolts, main="Probability plot for sample of 11 bolts", xlab="Bolt Diameter in Inches")
> qqline(bolts, col=”green”)
```




**State your conclusion: (1 pt) :**

The probability plot is not bad. It looks reasonably linear. This indicates that it is reasonable to assume the sample is normally distributed. We can reasonably infer it came from a population that is normally distributed. Based on the boxplot, it is reasonable to infer that there are no outliers. It is reasonable to use the t procedures. The data was from a simple random sample, knowing there are tons of millions of bolts in the world, this sample is definitely less than 5% of the population of bolts. The sampling distribution of the sample mean is approximately normal since we can assume the sample came from a normal distribution because the probability plot indicated that there were no major deviations from normality, therefore, we could reasonably assume the sample came from a population that is normally distributed. The boxplot showed no outliers so that mean is not affected by outliers.

b. Determine the 90% confidence interval for the mean diameter of these bolts.

**Copy and paste your code (2 pt) , results (1 pt) , and state the 90% confidence interval rounding to 2 decimal places (2 pt)**

90% confidence interval: (1.09, 1.14). The mean of diameter of a bolt is between 1.09 and 1.14 inches

```
> t.test(bolts, conf.level=.90)

        One Sample t-test

data:  bolts
t = 91.71, df = 10, p-value
= 5.815e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 1.093410 1.137499
sample estimates:
mean of x
 1.115455
```

c. Does the data give convincing evidence that the mean diameter of these bolts is less than 1.15 at a 0.05 significance level (doing a t test)?

**Copy and paste your code (1 pt) , results (1 pt) , and conclusion (1 pt)**

```
> bolts <- c(1.11,1.07, 1.11, 1.07, 1.12, 1.08, 1.08,
1.15, 1.18, 1.18, 1.12)
> t.test(bolts, mu=1.15, alternative="less")

        One Sample t-test

data:  bolts
t = -2.8403, df = 10, p-value = 0.00877
alternative hypothesis: true mean is less than 1.15
95 percent confidence interval:
     -Inf 1.137499
sample estimates:
mean of x
 1.115455
```

The p-value is .008 which is less than the .05 significance level thus we reject the null hypothesis that the true mean diameter of the bolts is 1.15. Based on our sample, there is sufficient evidence to support the claim the mean diameter of bolts is less than 1.15 at the .05 significance level.

2. Suppose that a random sample of *n* = 5 was selected from the vineyard properties for sale in Sonoma County,California, in each of three years. The following data are consistent with summary information on price per acre for disease-resistant grape vineyards in Sonoma County. Carry out an ANOVA (also test to see if assumptions are met – do analysis on residues for normality and for outliers) to determine whether there is evidence to support the claim that the mean price per acre for vineyard land in Sonoma County was not the same for at least one of the three years considered. Test at the 0.05 level. If there is a difference, which year(s) are different?

**1996:** 30000 34000 36000 38000 40000
**1997:** 32000 35000 37000 38000 40000
**1998:** 40000 41000 43000 44000 48000

**Copy and paste your code (2pts), results(2pts), graphs (2pts),, and conclusion(1pts):**

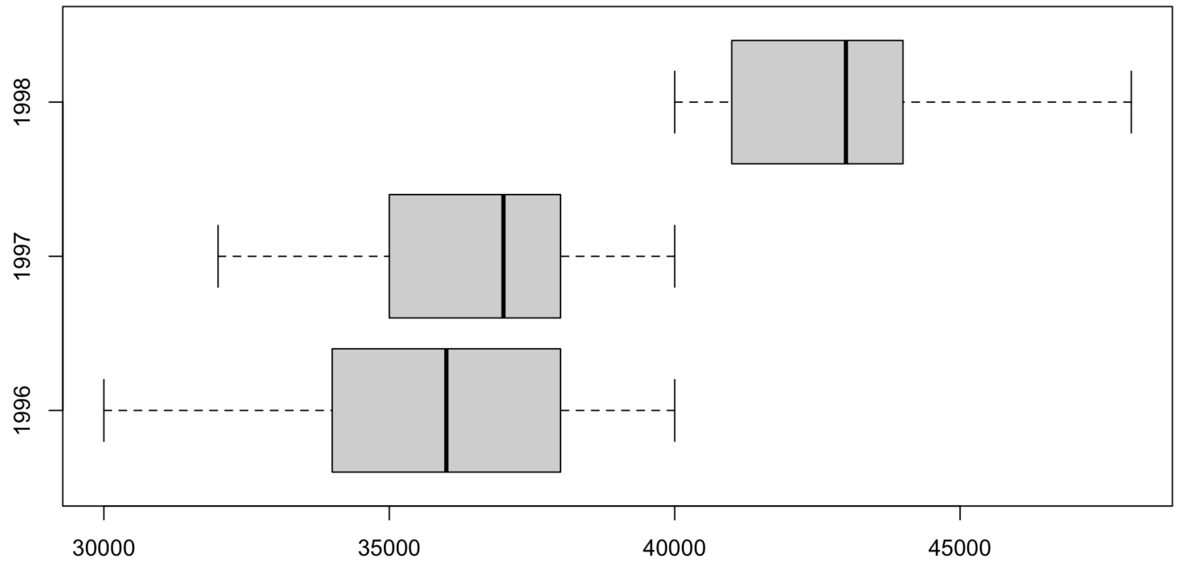
```
year1996 <- c(30000, 34000, 36000, 38000, 40000)
year1997 <- c(32000, 35000, 37000, 38000, 40000)
year1998 <- c(40000, 41000, 43000, 44000, 48000)
boxplot(year1996,year1997,year1998,horizontal = TRUE, names=c("1996","1997","1998"))
price_per_acre = c(30000, 34000, 36000, 38000, 40000, 32000, 35000, 37000, 38000, 40000,
40000, 41000, 43000, 44000, 48000)
year=c(rep("1996",5),rep("1997",5), rep("1998",5))
vineyard=data.frame(price_per_acre,year)
mean_price_per_acre <- mean(price_per_acre)
resid <- price_per_acre-mean_price_per_acre
qqnorm(resid, main="Probability Plot of Residuals for Price per Acre for Vineyard Land in
Sonoma County")
qqline(resid)
boxplot(resid, main="Distribution of Residuals for Price per Acre for Vineyard Land", horizontal =
TRUE)
sd(year1996)
sd(year1997)
sd(year1998)
results = aov(price_per_acre ~ year, data=vineyard)
summary(results)
summary(results)
TukeyHSD(results, conf.level=.95)
```

Based on the probability plot, there does not seem to be any deviations from normality. Based on the boxplot, there are no outliers. So, it is reasonable to assume normality with no outliers. Therefore, the assumption of normal population is met. The largest standard deviation is less than 2 times the smallest so the assumption of equal standard deviations is met.
Based on the sample, there is enough evidence to support the claim that there is at least one of the year's price per acre is different. (p-value = .00694). Based on the Tukey analysis, we see there is no difference between 1997 and 1996, but 1998 is different from both 1996 and 1997.

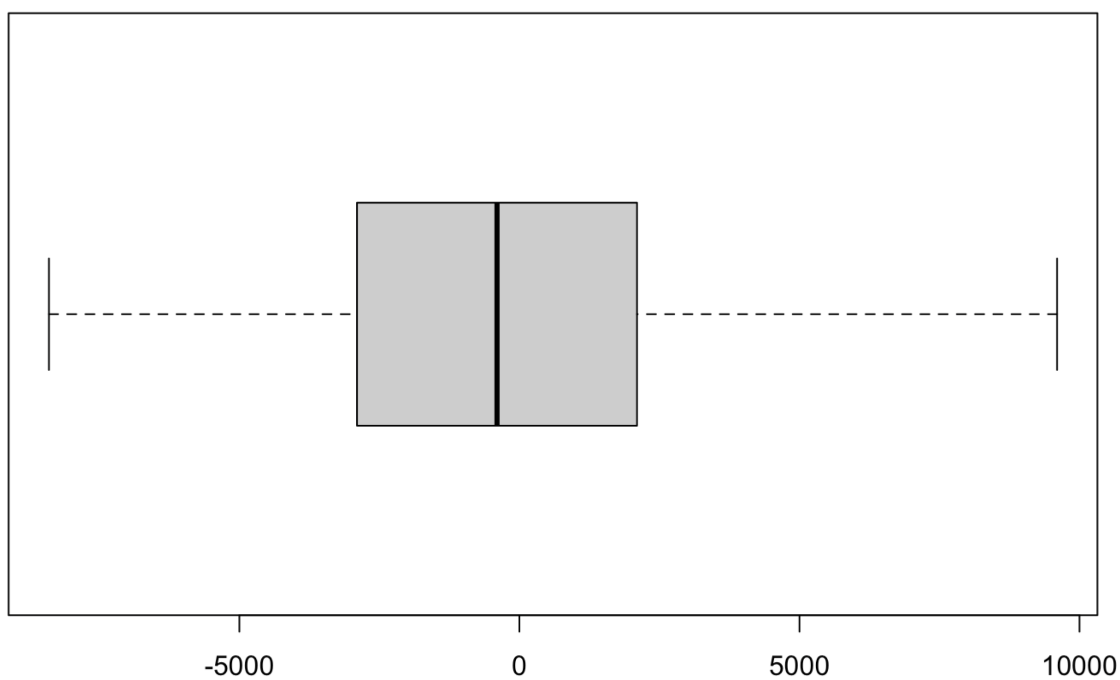
```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = price_per_acre ~ year, data = vineyard)

$year
          diff       lwr       upr     p adj
1997-1996  800 -4863.584  6463.584 0.9251845
1998-1996 7600  1936.416 13263.584 0.0097871
1998-1997 6800  1136.416 12463.584 0.0192560
```

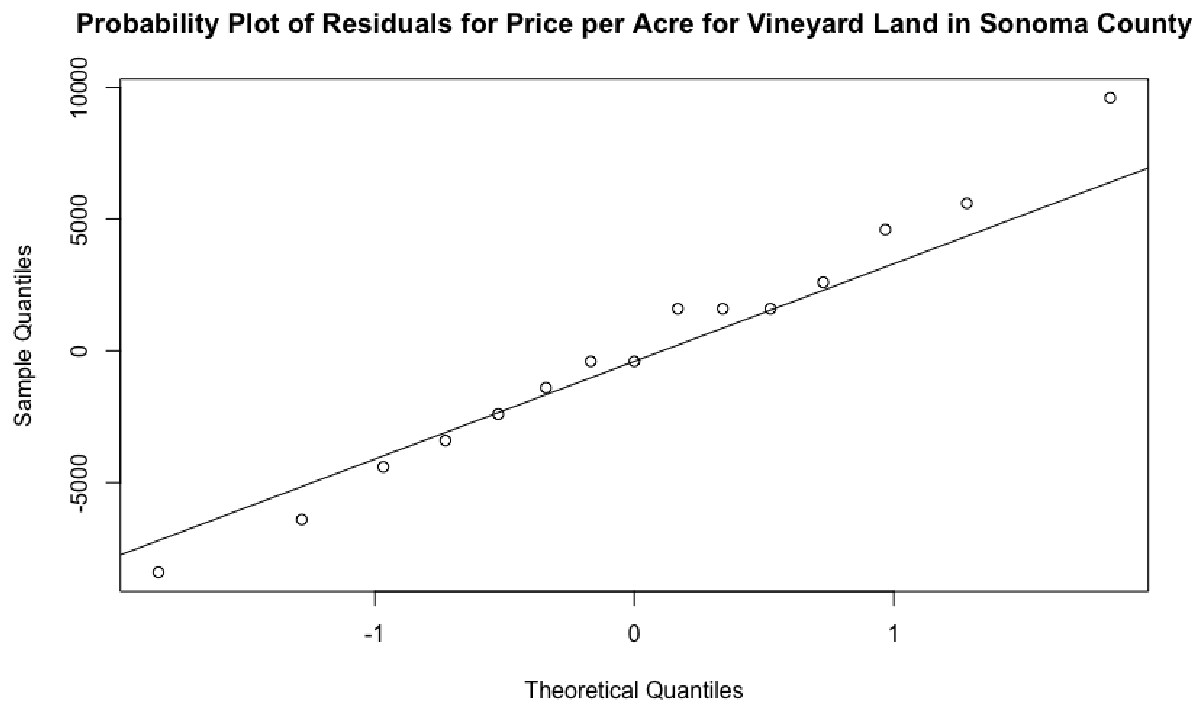
```
            Df    Sum Sq  Mean Sq F value  Pr(>F)   
year         2 174400000 87200000    7.74 0.00694 **
Residuals   12 135200000 11266667
```

```
> sd(year1996)
[1] 3847.077
> sd(year1997)
[1] 3049.59
> sd(year1998)
[1] 3114.482
```




## Distribution of Residuals for Price per Acre for Vineyard Land




Probability Plot of Residuals for Price per Acre for Vineyard Land in Sonoma County
Sample Quantiles
10000
5000
0
-5000
-1
0
1
Theoretical Quantiles