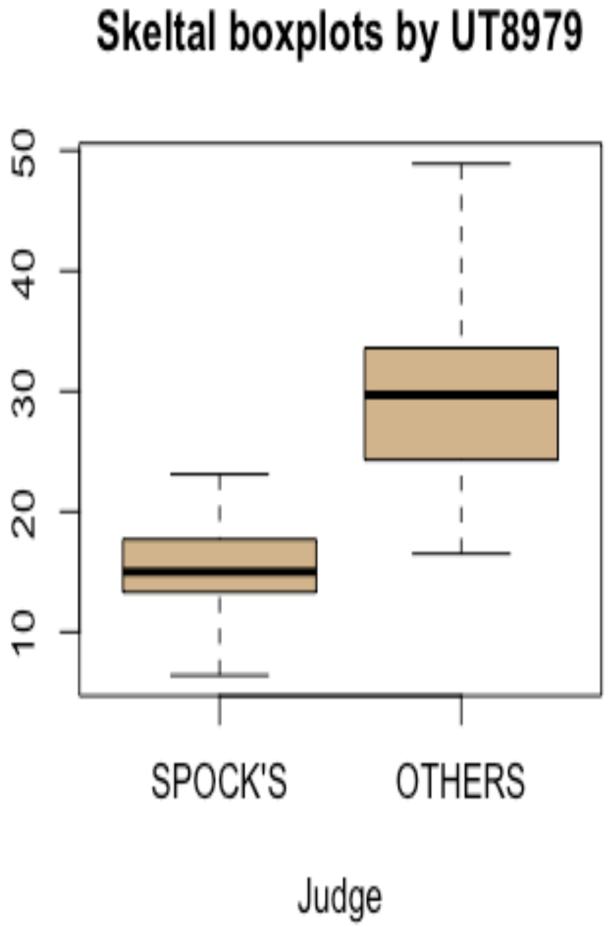
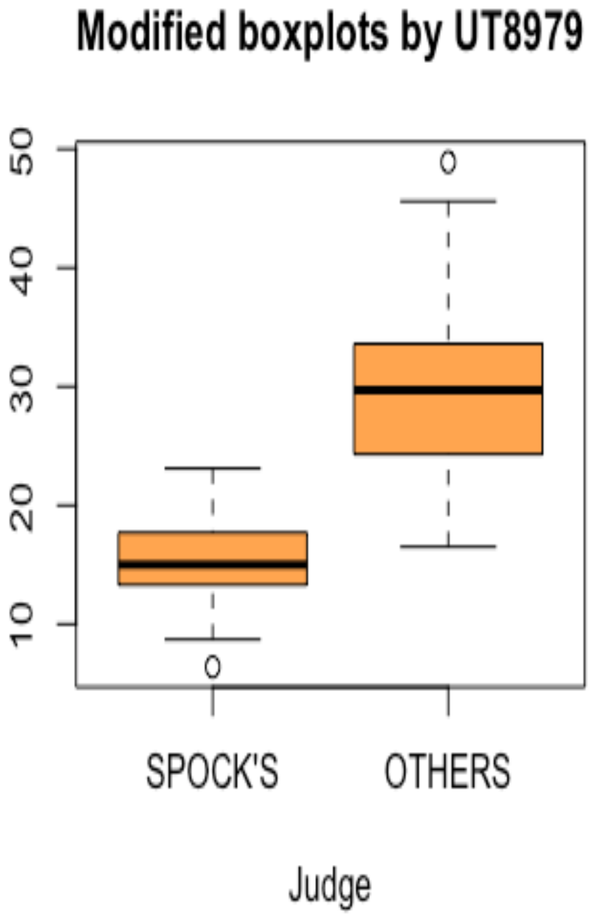
STA 303 Assignment 1

Yongwen Tan, 1002158979

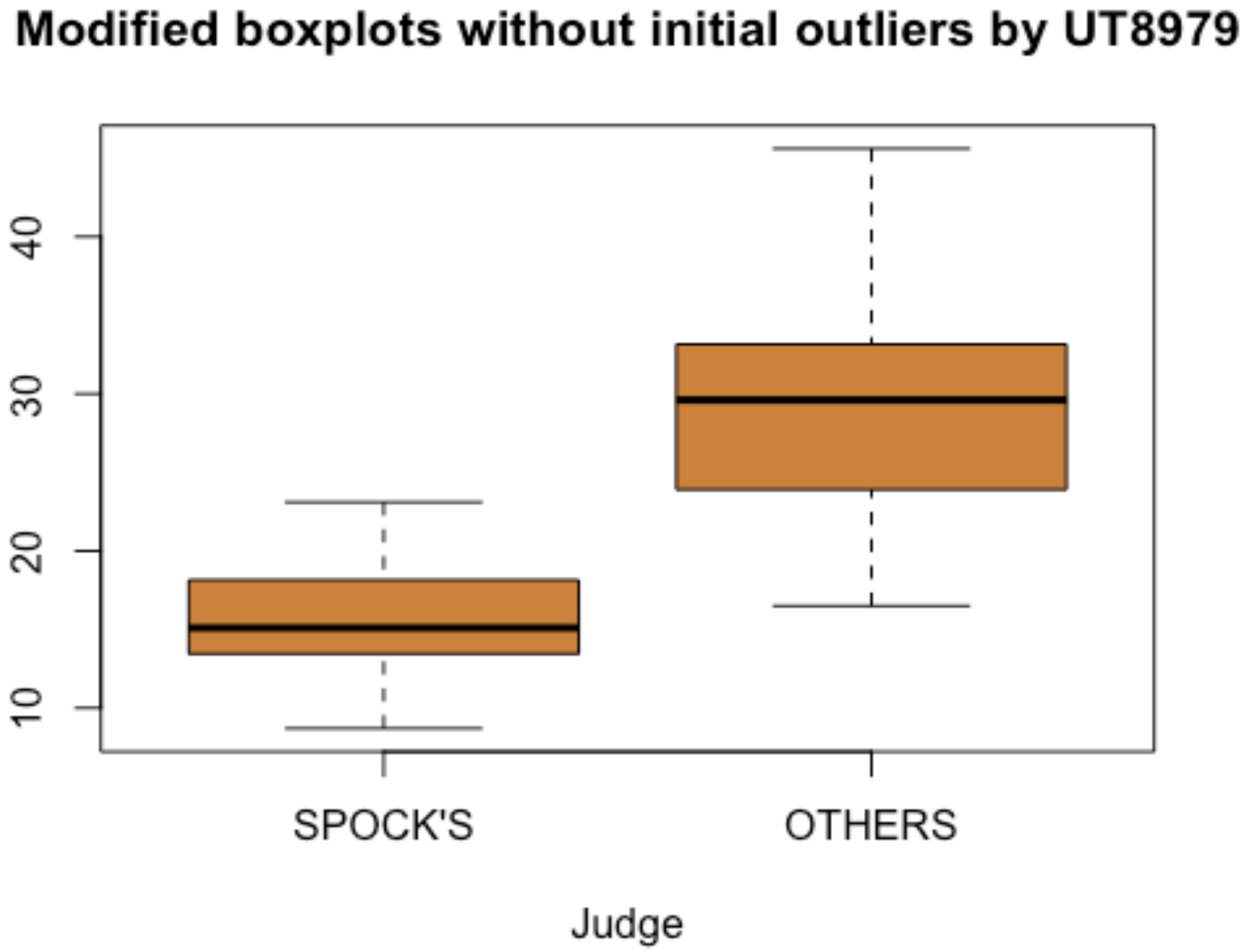
1. (a)





```
> boxplot(SPOCKS,OTHERS, range=0,xlab="Judge",names = c("SPOCK'S","OTHERS"),main="Skeltal boxplots by UT8979", col="tan")
```

```
> boxplot(SPOCKS,OTHERS, outline = TRUE,xlab="Judge",names = c("SPOCK'S","OTHERS"),main="Modified boxplots by UT8979",col="tan1")
```



```
boxplot(SPOCKS1,OTHERS1,xlab="Judge",names = c("SPOCK'S","OTHERS"),main="Modified boxplots without initial outliers by UT8979",col="tan3")
```

(b) Skeltal boxplots:

Judge	Spocks	Others
the first quartile	13.30	24.30
the second quartile	15.00	29.70
the third quartile	17.70	33.60
the end points of the two whiskers (max,min)	23.10 6.40	48.90 16.50
the extreme (outlier) points	N/A	N/A

Modified boxplots:

Judge	Spocks	Others
the first quartile	13.30	24.30
the second quartile	15.00	29.70
the third quartile	17.70	33.60
the end points of the two whiskers (max,min)	23.10 8.70	45.60 16.50
the extreme (outlier) points	17.70+1.5*(17.70-13.30)=24.3 13.30-1.5*(17.70-13.30)=6.69 6.4<6.69	33.60+1.5*(33.60-24.30)=47.55 24.30-1.5*(33.60-24.30)=10.35 48.9>47.55

Modified boxplots without initial outliers:

Judge	Spocks	Others
the first quartile	13.53	24.10
the second quartile	15.10	29.60
the third quartile	17.93	32.92
the end points of the two whiskers (max,min)	23.10 8.70	45.60 16.50
the extreme (outlier) points	N/A	N/A

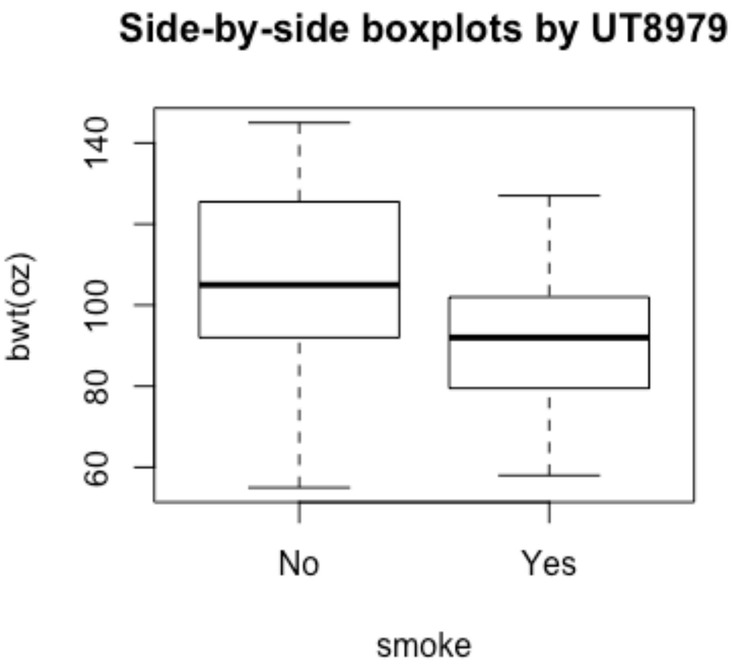
(c) (5 marks) Compare the three pairs of box plots. Which pair best represents the data and why?

I think the modified boxplot is the best represents of the data. Because this plot can clearly show if there is any outliers. This can help us analysis the data faster and easier. Also, when the data is normal distributed, the other two plots look the same.

2.(a)

Categorical variables	Number of levels, level names
Smoke	2: Yes/No
Id	99:1,2…99

(b)i. Side-by-side boxplots



ii. Null and Alternative Hypotheses:

$$H_0: \mu_{smoke} - \mu_{nonsmoke} = 0$$

$$H_\alpha: \mu_{smoke} - \mu_{nonsmoke} \neq 0$$

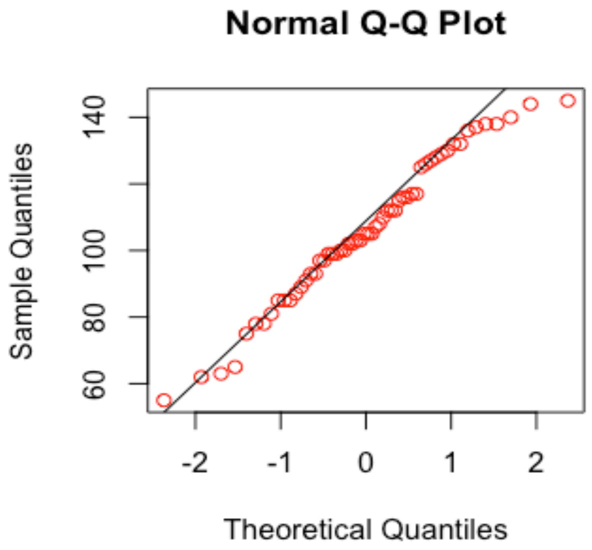
iii. A test statistic and it's distribution:

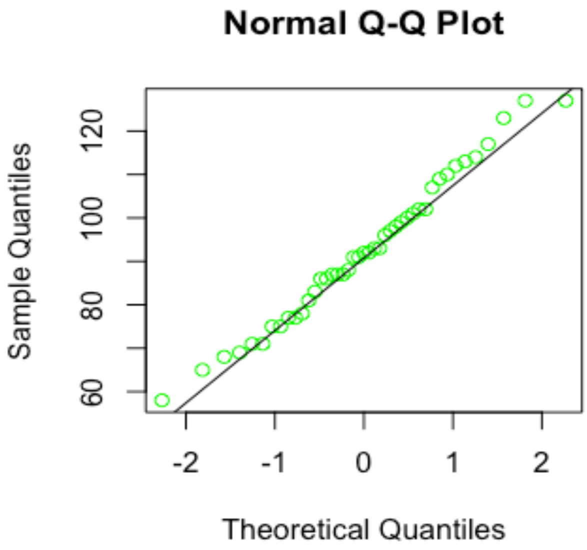
$$3.30 \sim t_{97}$$

iv. Test assumptions:

Two samples are assumed to have equal variances. Two samples should be independent. They are from the normal population

v. Test diagnostics (checking model assumptions)





The P-values in F test are 0.07511 which is larger than 0.05. Then this satisfies the assumption that two samples have the same variances.

The Normal Q-Q Plots show that two samples are normal although there are some outliers.

vi. P-value

p-value = 0.001343 < 0.05

vii. Results (brief discussion and conclusion)

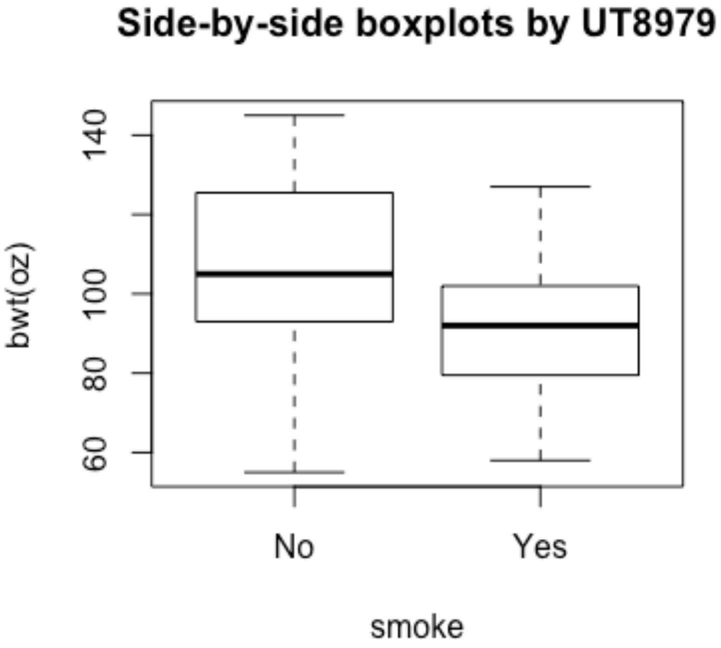
Since the p-value is less than 0.05, we need to reject to the null hypotheses. As the result, the babies born to the non-smoke mother are heavier than the babies born to the smoke mothers due to the sample means are not equal.

(c)

i). One-way ANOVA with 2 groups

ii). Simple liner regression approach with 1 dummy variable

(d) **i. Side-by-side boxplots**



ii. Null and Alternative Hypotheses:

$$H_0: \mu_{smoke} - \mu_{nonsmoke} = 0$$

$$H_\alpha: \mu_{smoke} - \mu_{nonsmoke} \neq 0$$

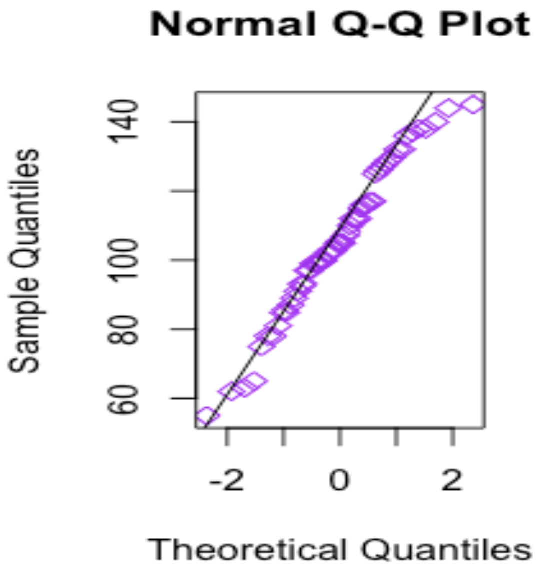
iii. A test statistic and it's distribution

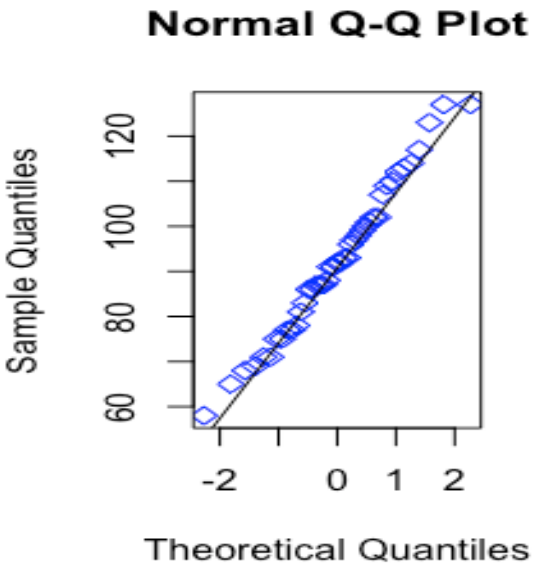
$3.38 \sim t_{96}$

iv. Test assumptions

Two samples are assumed to have equal variances. Two samples should be independent. They are from the normal population

v. Test diagnostics (checking model assumptions)





The P-values in F test are 0.07499 which is larger than 0.05. We failed to reject Null Hypotheses. Then this satisfies the assumption that two samples still have the same variances.

The Normal Q-Q Plots show that two samples are normal although the outliers still exist.

vi. P-value

p-value = 0.07499 < 0.05

vii. Results (brief discussion and conclusion)

The p-value is still less than 0.05 so that the babies born to the mothers who are nonsmokers are heavier than the babies born to the mothers who are smokers.

d)

```
subset2<-data2[-79,]
```

See the rcode in the appendix

(e) (5 marks) Compare your results of part (b) and part (d). Do you think that the observation removed was influential?

The point(79th) was removed is not an outlier. Compare to the result of part (b) and part (c), there is no significant change. Thus, the point is not influential.

Appendix

```
> #Q1(a)
> yongwendata=read.csv("~/Desktop/juries.csv")
> attach(yongwendata)
The following objects are masked from yongwensubset1:

    JUDGE, PERCENT

The following objects are masked from yongwendata (pos = 5):

    JUDGE, PERCENT

…
The following objects are masked from data (pos = 104):

    JUDGE, PERCENT


> SPOCKS=PERCENT[JUDGE=="SPOCKS"]
> OTHERS=PERCENT[JUDGE!="SPOCKS"]
>
> boxplot(SPOCKS,OTHERS, range=0,xlab="Judge",names = c("SPOCK'S","OTHERS"),main="Skeltal boxplots by UT8979", col="tan")
> boxplot(SPOCKS,OTHERS, outline = TRUE,xlab="Judge",names = c("SPOCK'S","OTHERS"),main="Modified boxplots by UT8979",col="tan1")
> yongwensubset0=yongwendata[-1,]
> yongwensubset1=subset0[-13,]
> attach(yongwensubset1)

The following objects are masked from yongwendata (pos = 3):

    JUDGE, PERCENT

……

The following objects are masked from data (pos = 105):

    JUDGE, PERCENT



> SPOCKS1=PERCENT[JUDGE=="SPOCKS"]

> OTHERS1=PERCENT[JUDGE!="SPOCKS"]

> boxplot(SPOCKS1,OTHERS1,xlab="Judge",names = c("SPOCK'S","OTHERS"),main="Modified boxplots
without initial outliers by UT8979",col="tan3")

>

>

>

>

>

> #(b)

> spocks_summary=summary(SPOCKS)

> spocks_summary

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

   6.40   13.30   15.00   14.62   17.70   23.10
```

```
> others_summary=summary(OTHERS)
> others_summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  16.50   24.30   29.70   29.49   33.60   48.90 
>
>
> IQR_spocks=spocks_summary[5]-spocks_summary[2]
> upper_spocks=spocks_summary[5]+1.5*IQR_spocks
> upper_spocks
3rd Qu. 
   24.3 
> lower_spocks=spocks_summary[2]-1.5*IQR_spocks
> lower_spocks
 1st Qu. 
6.699999 
> SPOCKS[SPOCKS>upper_spocks]
numeric(0)
> SPOCKS[SPOCKS<lower_spocks]
[1] 6.4
> IQR_others=others_summary[5]-others_summary[2]
> upper_others=others_summary[5]+1.5*IQR_others

> upper_others

3rd Qu.

   47.55

> lower_others=others_summary[2]-1.5*IQR_others

> lower_others

1st Qu.

   10.35

>

> OTHERS[OTHERS>upper_others]

[1] 48.9

> OTHERS[OTHERS<lower_others]

numeric(0)

>

>

>

>

> summary(SPOCKS1)

    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

    8.70   13.53   15.10   15.65   17.93   23.10
```

```
> summary(OTHERS1)

   Min. 1st Qu.    Median      Mean 3rd Qu.      Max.

  16.50     24.10     29.60     28.95     32.92     45.60

>

>

>

>

>

> #Q2

> #b)i

> yongwendata2=read.csv("~/Desktop/bbw99.csv",header = TRUE)

> attach(yongwendata2)

The following objects are masked from yongwendata2 (pos = 5):

    bwt, id, smoke

...

The following objects are masked from data2 (pos = 100):

    bwt, id, smoke



> boxplot(bwt~smoke,
data=yongwendata2,xlab="smoke",ylab="bwt(oz)",names=c("No","Yes"),main="Side-by-side boxplots by
UT8979")
```

```
>

> var.test(bwt~smoke)


	F test to compare two variances


data:  bwt by smoke

F = 1.7012, num df = 55, denom df = 42, p-value = 0.07511

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

 0.946692 2.989462

sample estimates:

ratio of variances

          1.701211


> t.test(bwt~smoke,var.equal=T)


	Two Sample t-test


data:  bwt by smoke

t = 3.3024, df = 97, p-value = 0.001343
```

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  5.367026 21.534967
sample estimates:
mean in group 0 mean in group 1
      105.89286        92.44186


>
> par(mfrow=c(1,2))
> qqnorm(bwt[smoke=="0"],col="red")
> qqline(bwt[smoke=="0"])
> qqnorm(bwt[smoke=="1"],col="green")
> qqline(bwt[smoke=="1"])
>
>
> #Q2 c)
> yongwensubset2=yongwendata2[-79,]
> boxplot(bwt~smoke,data=yongwensubset2, xlab="smoke",ylab="bwt(oz)",names=c("No","Yes"),
+         main="Side-by-side boxplots by UT8979")
> var.test(bwt~smoke, data=yongwensubset2)
```

```
	F test to compare two variances

data:  bwt by smoke
F = 1.7042, num df = 54, denom df = 42, p-value = 0.07499
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9467707 3.0037644
sample estimates:
ratio of variances
          1.704208

> t.test(bwt~smoke,var.equal=T, data=yongwensubset2)

	Two Sample t-test

data:  bwt by smoke
t = 3.3841, df = 96, p-value = 0.001035
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  5.718259 21.943474
sample estimates:
mean in group 0 mean in group 1
      106.27273        92.44186


>
>
> bwt1=yongwensubset2$bwt
> smoke1=yongwensubset2$smoke
> par(mfrow=c(1,2))
> qqnorm(bwt1[smoke1=="0"],col="purple",pch=5)
> qqline(bwt1[smoke1=="0"])
> qqnorm(bwt1[smoke1=="1"],col="blue",pch=5)
> qqline(bwt1[smoke1=="1"])
>
```