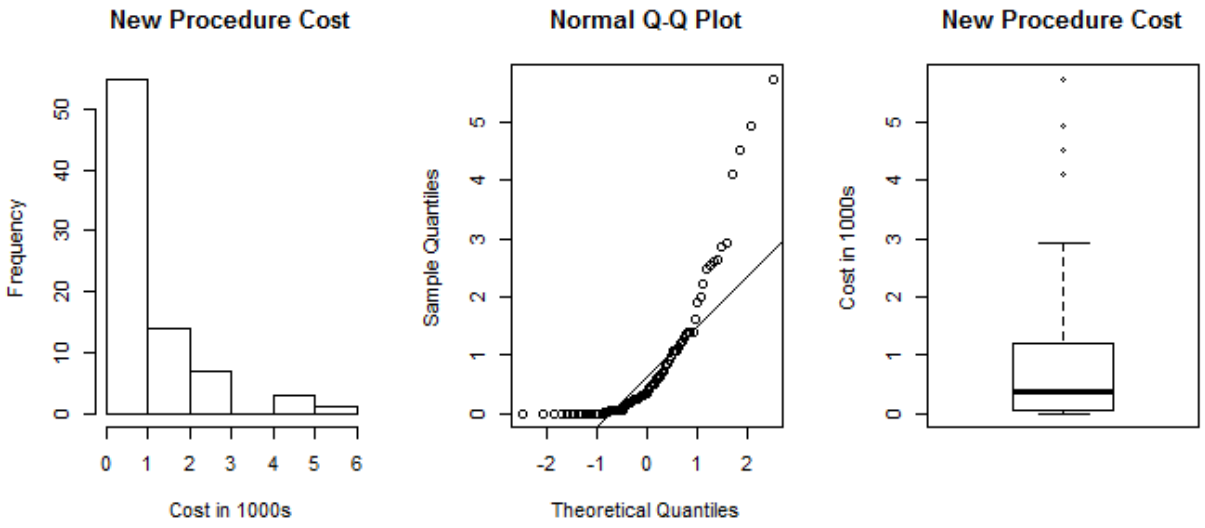
BIOS 6611 Homework 7 *Answer Key*

Due Monday, October 29, 2018 by midnight to Canvas Assignment Basket

1. Recall the data in Homework 3 on total hospital costs per patient for either of two procedures (Standard (=1) and New (=2)) over a one-month period at one hospital. Use R to read in the **ProcedureCost.csv** data, as before, and carry out the following:

   i. We will use bootstrap sampling to examine the sampling distribution of mean costs for the "New" procedure. For the observed data:

      a. Plot the observed data (e.g., histogram, normal quantile plot, boxplot, etc.)




```
df <- read.csv("~/ProcedureCost.csv")
df

### Part 1.i - bootstrap of new procedure mean
## 1.i.a - plots of observed data

par(mfrow=c(1,3)) #create plotting area for 3 figures in one row

hist(df$Cost[df$Procedure==2], main='New Procedure Cost',
xlab='Cost in 1000s') #Histogram
qqnorm(df$Cost[df$Procedure==2]);
qqline(df$Cost[df$Procedure==2]) #Normal Q-Q Plot
boxplot(df$Cost[df$Procedure==2], main='New Procedure Cost',
ylab='Cost in 1000s') #Box Plot
```

b. Describe the shape of the distribution (bell-shaped, symmetric, skewed, etc.)
From the plots of observed cost of the new procedure in (a), the distribution of cost is heavily right skewed with most observations in the 0-3 range and a few in the 4-6 range.

c. Provide summary statistics (mean, standard deviation)
The mean cost in $1000's is 0.881625 with a standard deviation of 1.210242 (or we can calculate mean cost in dollars of $881.63 with a standard deviation of $1210.24.

```
## 1.i.c - provide summary statistics for observed cost
m <- aggregate(Cost ~ Procedure, data = df, FUN = mean)[,2]
#calculate mean for each procedure
sd <- aggregate(Cost ~ Procedure, data = df, FUN = sd)[,2]
#calculate standard deviation for each procedure

m[2] #mean for new procedures
[1] 0.881625

sd[2] #sd for new procedures
[1] 1.210242
```

For the bootstrap sampling distribution:

Bootstrap Code:

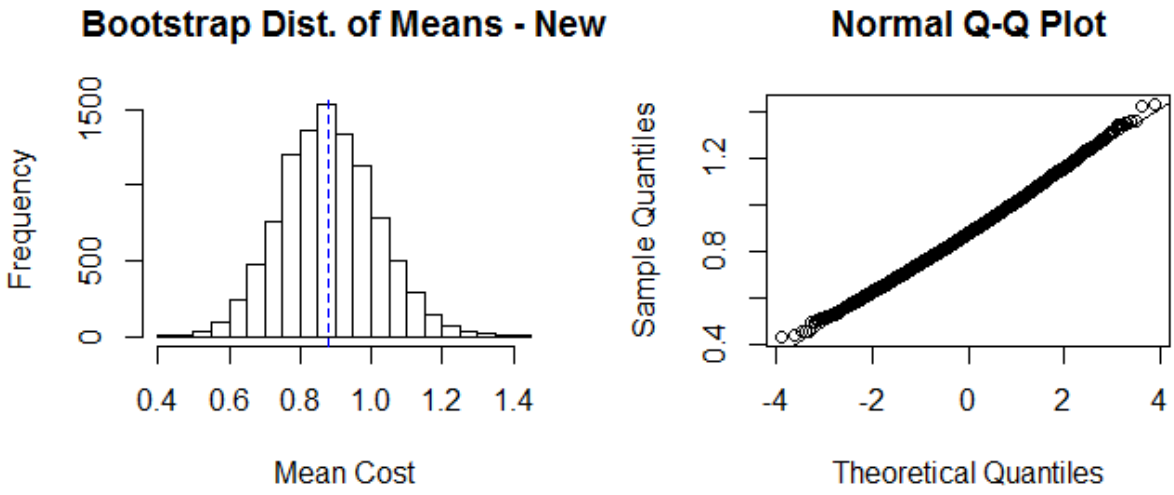
```
## 1.i.d/e/f/g - Bootstrap sampling
nN <- length(df$Procedure[df$Procedure==2]) #identify sample size
with procedure 2

B <- 10^4 #set number of bootstrap iterations

boot.mean <- numeric(B) #initialize vector to store bootstrap
mean estimates in
set.seed(515) #set seed for reproducibility

for (i in 1:B){
    xB <- sample(df$Cost[df$Procedure==2], nN, replace = TRUE)
    boot.mean[i] <- mean(xB)
}
```

d. Provide plots




```
## 1.i.d - bootstrap plots

par(mfrow=c(1,2)) #create plotting area for 2 figures in one row

hist(boot.mean, main='Bootstrap Dist. of Means - New', xlab='Mean
Cost') #histogram of mean estimates from bootstrap
abline(v = m[2], col = "blue", lty = 2) #observed mean from
procedure cost data
qqnorm(boot.mean); qqline(boot.mean) #Q-Q plot
```

e. Describe the shape and spread

The bootstrap sampling distribution of mean cost for the new procedure with samples of size 80 is symmetric and approximately normal based on the Q-Q plot points closely following the diagonal line.

f. Estimate the bootstrap mean, standard error of the mean, and bias

The mean of the bootstrap distribution is 0.8810119 ($881.01) with a standard error of 0.1337116 ($133.71). The bias is -0.00061 ($0.61), so it is not very biased.

```
## 1.i.f - mean, SE, and bias of bootstrap distribution

mean(boot.mean) # bootstrap mean
[1] 0.8810119

mean(boot.mean)-mean(df$Cost[df$Procedure==2]) # bias for New
procedure
[1] -0.000613125

sd(boot.mean) # bootstrap SE
[1] 0.1337116
```

g. Obtain the 95% normal percentile and the 95% bootstrap percentile confidence intervals and interpret the results. Comment on the coverage of the normal percentile confidence interval and the potential accuracy of the bootstrap percentile confidence interval.

The 95% normal percentile CI is (0.6189, 1.1431) and the 95% bootstrap percentile CI is (0.6319, 1.1534). For the 95% normal percentile CI, we are 95% confident that the true mean lies in this interval, assuming the central limit theorem applies. For the 95% bootstrap percentile CI, we are 95% confident that the true mean is in this interval. Additionally, because it is estimated from our data directly, 95% of the bootstrap means fall in this interval.

Based on our estimates of coverage, the 95% normal percentile estimates are too low for both the lower and upper bounds since the lower bound has coverage of 2.01% and the upper bound has coverage of 2.98% instead of the desired 2.5%, suggesting the CLT may be inaccurate. The accuracy of our bootstrap percentile can be estimated by the ratio of the bias/SE, which is -0.0046. Since this does not exceed ±0.10 we should have good accuracy.

```
## 1.i.g - 95% normal percentile and 95% bootstrap percentile
confidence intervals

# Obtain Normal percentile 95% CI and estimate of coverage
LLN <- mean(boot.mean)-1.96*sd(boot.mean) # Lower limit of 95%
Normal CI
LLN
[1] 0.6189372

ULN <- mean(boot.mean)+1.96*sd(boot.mean) # Upper limit of 95%
Normal CI
ULN
[1] 1.143087

sum(boot.mean > ULN)/B # Coverage of CI at upper end
[1] 0.0298

sum(boot.mean < LLN)/B # Coverage of CI at lower end
[1] 0.0201

# Obtain bootstrap percentile 95% CI and estimate of accuracy
quantile(boot.mean, c(0.025, 0.975))
     2.5%     97.5%
0.6318656 1.1533781

(mean(boot.mean)-mean(df$Cost[df$Procedure==2])) / sd(boot.mean)
#bias/bootstrap SE for potential accuracy of bootstrap CI, values
exceeding +/-0.10 indicate worse accuracy
[1] -0.004585429
```

ii. We will use bootstrap sampling to estimate the ratio of mean costs between the two procedures: New/Standard, from the original data. Obtain a bootstrap sampling distribution of the ratio of mean costs and:

Bootstrap Code:

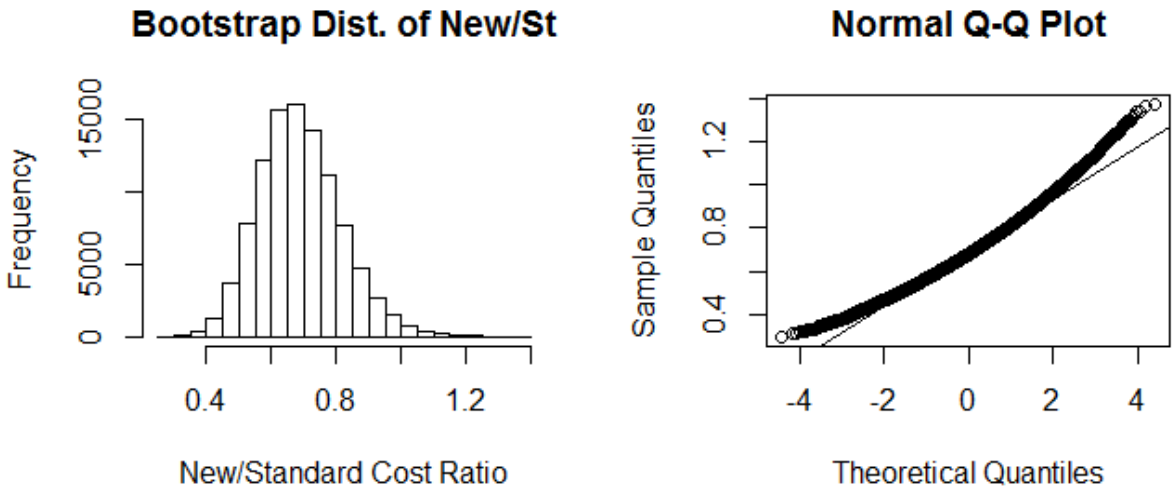
```
### Part 1.ii - Sampling distribution of Ratio of Means New to
Standard
## 1.ii.a/b - Bootstrap sampling
B <- 10^5 #set number of bootstraps
cost.ratio.mean <- numeric(B) #initialize vector to store results in

nS <- length(df$Procedure[df$Procedure==1]) #determine sample size
of standard procedure
nS

set.seed(515) #set seed for reproducibility

for (i in 1:B){
    Standard.boot <- sample(df$Cost[df$Procedure==1], nS, replace=T)
    New.boot <- sample(df$Cost[df$Procedure==2], nN, replace = TRUE)
    cost.ratio.mean[i] <- mean(New.boot)/mean(Standard.boot)
}
```

a. Provide plots and describe the shape, mean, standard error, and bias of the bootstrap sampling distribution for the ratio of mean costs.



The shape of the bootstrap distribution for ratio of mean cost for new to mean cost of standard procedure is slightly positively skewed, where new has a sample size of 80 and standard has a sample size of 120. It deviates from normality based on the Q-Q plot points curving away from the diagonal line at the extremes. The bootstrap mean ratio is 0.6883 with standard error of 0.1256. The bias is 0.0066.

```
## 1.ii.a - bootstrap plots, calculation of mean, SE, bias

par(mfrow=c(1,2)) #create plotting area for 2 figures in one row

hist(cost.ratio.mean, main='Bootstrap Dist. of New/St',
xlab='New/Standard Cost Ratio')
qqnorm(cost.ratio.mean); qqline(cost.ratio.mean)

mean(cost.ratio.mean) # bootstrap mean
[1] 0.6882697

mean(cost.ratio.mean)-
(mean(df$Cost[df$Procedure==2])/mean(df$Cost[df$Procedure==1])) #
bias for ratio
[1] 0.006557001

sd(cost.ratio.mean) # bootstrap SE
[1] 0.1255846
```

b. Obtain the 95% normal percentile and the 95% bootstrap percentile confidence intervals and interpret the results. Comment on the coverage of the normal percentile confidence interval and the potential accuracy of the bootstrap percentile confidence interval.

The 95% normal percentile CI is (0.442, 0.934) and the 95% bootstrap percentile CI is (0.469, 0.961). For the 95% normal percentile CI, we are 95% confident that the true ratio of mean cost lies in this interval, assuming the central limit theorem applies. For the 95% bootstrap percentile CI, we are 95% confident that the true ratio of mean cost is in this interval. Additionally, because it is estimated from our data directly, 95% of the bootstrap ratios of mean cost fall in this interval.

Based on our estimates of coverage, the 95% normal percentile estimates are too low at the lower and upper bounds since the lower bound has coverage of 1.2% and the upper bound has coverage of 3.6% instead of the desired 2.5%, suggesting that relying on the CLT may be inaccurate. The accuracy of our bootstrap percentile can be estimated by the ratio of the bias/SE, which is 0.052. Since this does not exceed ±0.10 we should have good accuracy.

It could also be noted that our 95% bootstrap percentile CI doesn't include 1, so we can also conclude that the mean cost of the new procedure is lower than the mean cost of the standard procedure.

```
## 1.ii.b - 95% normal percentile and 95% bootstrap percentile
confidence intervals

# Obtain Normal percentile 95% CI and estimate of coverage
LL <- mean(cost.ratio.mean)-1.96* sd(cost.ratio.mean) # Lower
limit of 95% Normal CI
LL
[1] 0.442124

UL <- mean(cost.ratio.mean)+1.96* sd(cost.ratio.mean) # Upper
limit of 95% Normal CI
UL
[1] 0.9344155

sum(cost.ratio.mean < LL)/B # Coverage of CI at lower end
[1] 0.0121

sum(cost.ratio.mean > UL)/B # Coverage of CI at upper end
[1] 0.03601

# Obtain bootstrap percentile 95% CI and estimate of accuracy
quantile(cost.ratio.mean, c(0.025, 0.975))
     2.5%     97.5%
0.4689186 0.9606854

( mean(cost.ratio.mean)-
(mean(df$Cost[df$Procedure==2])/mean(df$Cost[df$Procedure==1])) )
/ sd(cost.ratio.mean) # bootstrap CI accuracy
[1] 0.05221183
```

2. Suppose we have separately analyzed the effects of 10 single nucleotide polymorphisms (SNPs; https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism) comparing people with type I diabetes vs. controls. The p-values from these separate analyses are given below.

Effects of 10 SNPs on Type I Diabetes			
SNP	p-value	SNP	p-value
1	0.040	6	0.620
2	0.100	7	0.001
3	0.400	8	0.010
4	0.550	9	0.800
5	0.340	10	0.005

Use the FDR method to correct for multiple testing using an FDR = 0.05. After correction, which SNPs show statistically significant effects?

Before adjusting, with α=0.05, we would identify SNP 1, 7, 8, and 10 to be significantly different between our groups of Type I Diabetes and Controls. After FDR adjustment, SNPs 7, 8, and 10 are still significant.

Based on our lecture in class, there are 3 potential ways to solve this problem: using R, using SAS, or applying the algorithm by hand. Each are described below:

Problem 2 Using R:

```
# create vector of p-values

pvec2 <- c(0.04,0.1,0.4,0.55,0.34,0.62,0.001,0.01,0.8,0.005)
fdr.vals <- p.adjust( pvec2, method='fdr') #calculate FDR adjusted p-
values

# create matrix to summarize SNP and FDR value to identify which are still
significant
matrix( c( 1:10, round(fdr.vals, 3)), nrow=10, byrow=F, dimnames=list(
1:10, c('SNP','FDR') ) )  #7, 8, and 10 are sig still
```

```
   SNP   FDR
1    1 0.100
2    2 0.200
3    3 0.571
4    4 0.688
5    5 0.567
6    6 0.689
7    7 0.010
8    8 0.033
9    9 0.800
10  10 0.025
```

Problem 2 Using SAS:

```
* Read in the raw p-values;
DATA one;
INPUT SNP RAW_P; * Need to call the p-value RAW_P for MULTTEST;
CARDS;
1    .04
2    .10
3    .40
4    .55
5    .34
6    .62
7    .001
8    .01
9    .80
10         .005
;
RUN;

* Use MULTTEST to apply the FDR method;
PROC MULTTEST INPVALUES=one FDR;
RUN;
```

The Multtest Procedure

P-Value Adjustment Information	
P-Value Adjustment	False Discovery Rate

p-Values		
Test	**Raw**	**False Discovery Rate**
1	0.0400	0.1000
2	0.1000	0.2000
3	0.4000	0.5714
4	0.5500	0.6875
5	0.3400	0.5667
6	0.6200	0.6889
7	0.0010	0.0100
8	0.0100	0.0333
9	0.8000	0.8000
10	0.0050	0.0250

Problem 2 By Hand:

SNP	p-value	Rank	q (=kp/Rank)	FDR (=MIN(q for rank or higher)
7	0.001	1	0.01	0.01
10	0.005	2	0.025	0.025
8	0.01	3	0.033333333	0.03333
1	0.04	4	0.1	0.1
2	0.1	5	0.2	0.2
5	0.34	6	0.566666667	0.56667
3	0.4	7	0.571428571	0.57143
4	0.55	8	0.6875	0.6875
6	0.62	9	0.688888889	0.68889
9	0.8	10	0.8	0.8

3. Twenty-two young asthmatic volunteers were studied to assess the short-term effects of sulfur dioxide (SO_2) exposure under various conditions. The baseline data in the table (Table 12.30 from Rosner) were presented regarding the relationship of bronchial reactivity to SO_2 (cm H_2O/s) stratified by lung function (as defined by forced expiratory volume / forced vital capacity [FEV_1/FVC]) at screening.

Lung-Function Group		
Group A FEV_1/FVC $\leq$ 74%	Group B FEV_1/FVC 75-84%	Group C FEV_1/FVC $\geq$ 85%
20.8	7.5	9.2
4.1	7.5	2.0
30.0	11.9	2.5
24.7	4.5	6.1
13.8	3.1	7.5
	8.0	
	4.7	
	28.1	
	10.3	
	10.0	
	5.1	
	2.2	

Using SAS or R:

i. Assume that the variances across the groups are equal and test the hypothesis that there is an overall mean difference in bronchial reactivity among the three lung-function groups.

The overall F ratio for the test of equal mean reactivity to SO_2 over the three lung function groups is significant (p=0.0181), so at least one group has a mean reactivity that is different from the others.

Problem 3.i Using R:

```
# Create data set from table

lung <- data.frame( group=c( rep('A',5), rep('B',12), rep('C',5) ),
    react=c(20.8,4.1,30,24.7,13.8,
7.5,7.5,11.9,4.5,3.1,8,4.7,28.1,10.3,10,5.1,2.2, 9.2,2,2.5,6.1,7.5)
)

## 3.i - ANOVA with equal variances assumed
aov.lung <- aov( react ~ group, data=lung)
anova(aov.lung)

Analysis of Variance Table

Response: react
          Df Sum Sq Mean Sq F value  Pr(>F)
group      2 503.55 251.774  4.9893 0.01813 *
Residuals 19 958.80  50.463
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Problem 3.i Using SAS:

```
* Attach labels to the lung function groups;
PROC FORMAT;
VALUE Group 1='<=74%' 2 = '75-84%' 3 = '>=85';
RUN;

* Read in the raw data;
DATA lung;
INPUT group react;
CARDS;
1 20.8
2 7.5
3 9.2
1 4.1
2 7.5
3 2.0
1 30.0
2 11.9
3 2.5
1 24.7
2 4.5
3 6.1
1 13.8
2 3.1
3 7.5
2 8.0
2 4.7
2 28.1
2 10.3
2 10.0
2 5.1
2 2.2
;
RUN;
```

```
* 3.i-ii - Perform ANOVA assuming equal variances and generate
Tukey's HSD post-hoc test results;
PROC ANOVA DATA=lung;
CLASS group;
MODEL react = group;
MEANS group / TUKEY;
FORMAT group group.; * tells SAS to produce output with group labels
we defined;
RUN;
```

The SAS System
The ANOVA Procedure

Dependent Variable: react

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	503.548409	251.774205	4.99	0.0181
Error	19	958.802500	50.463289		
Corrected Total	21	1462.350909			

ii. If justified, compare the means of each pair of groups using the Tukey HSD method and summarize the results. Otherwise note why it isn't justified.

We found p=0.0181 in part (i), so post-hoc comparisons are justified. The results of post-hoc testing suggest that the lowest lung function group (A) differs from the other two:

≤74%	*75-84%*	*≥85%*
[redacted]		
	[redacted]	

Problem 3.ii Using R:

```
## 3.ii - Tukey's HSD post-hoc testing
TukeyHSD(aov.lung)

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

Fit: aov(formula = react ~ group, data = lung)

$`group`
       diff       lwr        upr     p adj
B-A -10.105 -19.71110 -0.4988964 0.0382469
C-A -13.220 -24.63375 -1.8062481 0.0217454
C-B  -3.115 -12.72110  6.4911036 0.6932026
```

Problem 3.ii Using SAS:

```
* 3.iii - Perform ANOVA without assumption of equal variances and
identify is post-hoc testing is justified;
PROC GLM DATA=lung;
CLASS group;
MODEL react = group;
MEANS group /WELCH;
FORMAT group group.; * tells SAS to produce output with group labels
we defined;
RUN;
```

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for react

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	19
Error Mean Square	50.46329
Critical Value of Studentized Range	3.59274

Comparisons significant at the 0.05 level are indicated by *.**

group Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
<=74% - 75-84%	10.105	0.499	19.711	***
<=74% - >=85	13.220	1.806	24.634	***
75-84% - <=74%	-10.105	-19.711	-0.499	***
75-84% - >=85	3.115	-6.491	12.721	
>=85 - <=74%	-13.220	-24.634	-1.806	***
>=85 - 75-84%	-3.115	-12.721	6.491	

iii. **EXTRA CREDIT:** Carry out part (i) assuming that the variances across the groups are not equal. If justified, describe a way to compare the means of each pair of groups, but do not carry out any further analysis.

At the 5% level of significance, the Welch's F-test is not significant (p=0.0585). Post-hoc tests are not justified with this result.

Problem 3.iii Using R:

```
## 3.iii - ANOVA without equal variance assumption
oneway.test( react ~ group, data=lung, var.equal=FALSE)

   One-way analysis of means (not assuming equal variances)

data:  react and group
F = 3.9682, num df = 2.0000, denom df = 8.9319, p-value = 0.05845
```

Problem 3.iii Using SAS:

Welch's ANOVA for react			
Source	**DF**	**F Value**	**Pr > F**
group	2.0000	3.97	0.0585
Error	8.9319		