ASSIGNMENT 6

Brandon Lampe
STAT 527
Advanced Data Analysis I

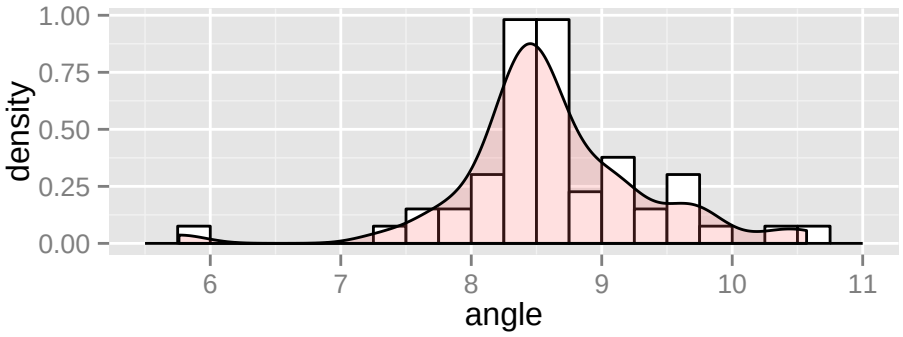
November 16, 2014

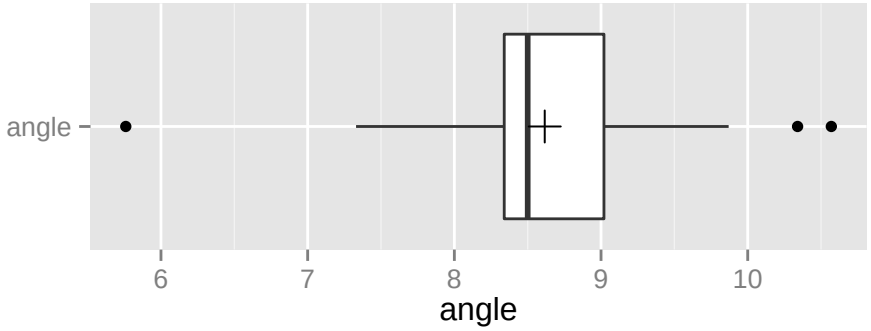
1 Parallax:

```
#  load all data for assignment and write to .csv files because website download
#  is not reliable
dir <- getwd() # current directory
# parallax <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_06_F14-1.csv")
# write.table(parallax, paste(dir, "p1.csv", sep = "/"), sep = ",", row.names = FALSE )
parallax <- read.csv(paste(dir, "p1.csv", sep = "/"))
angle <- parallax$angle;
```

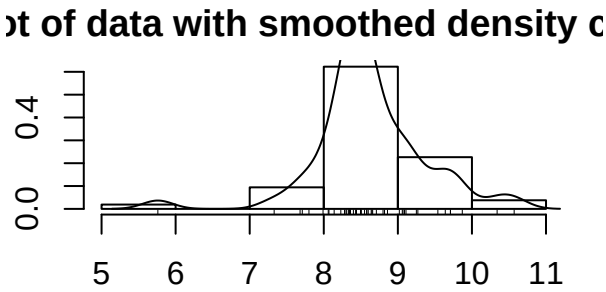
```
summary(angle)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.760   8.340   8.500   8.616   9.020  10.570
sd(angle) # standard deviation
## [1] 0.7490205
max(angle) - min(angle) #spread
## [1] 4.81
fivenum(angle)[4] - fivenum(angle)[2] #IQR
## [1] 0.68
```

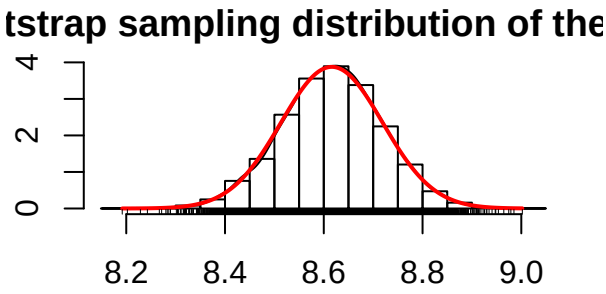
```
#create box plot
angle.hist <- ggplot(parallax, aes(x = angle))
angle.hist <- angle.hist + geom_histogram(aes(y = ..density..),
                                          binwidth = .25,color = "black", fill = "white")
angle.hist <- angle.hist + geom_density(alpha = 0.2, fill = "#FF6666")
angle.hist <- angle.hist + labs(x = "angle")

# boxplot of angle
angle.box <- ggplot(parallax, aes(x = "angle", y = angle)) # boxplot of angle
angle.box <- angle.box + geom_boxplot()
angle.box <- angle.box + coord_flip()
angle.box <- angle.box + labs(x = "", y = "angle")
angle.box <- angle.box + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 4)
# plot
grid.arrange(angle.hist, angle.box, nrow = 2)
```

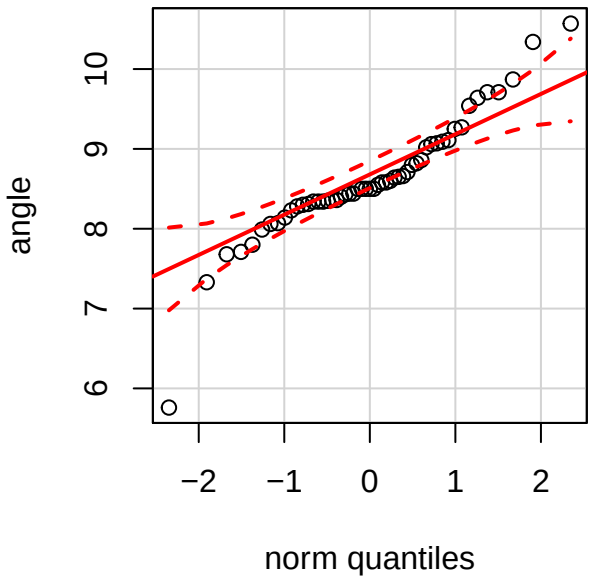





```
bs.one.samp.dist(angle) # bootstrap
qqPlot(angle)
```







1(a) (5 pts) Describe the distribution of determinations of the parallax. Be complete.

The distribution of measurements of the parallax using Short's device is unimodal and has a standard deviation that is approximately equal to the IQR. The data do not appear symmetric. Also, the data contains extreme outiers on both left and right sides and has tails of nearly equal length. The bootstrap sampling distribution of the mean parallax angle using Short's device appears symmetric and normal. The Q-Q plot shows that the data fall within the range of a normal distribution except for the extreme outliers on the left and right sides, which cause the mean to be greater than the median.

1(b) (10 pts) Perform the standard t-test on these data, at the 5Interpret the results, given the question of interest.

```
# t-test
angle.t <- t.test(angle, mu = 8.798, conf.level = 0.95)
angle.t
##
##  One Sample t-test
##
## data:  angle
## t = -1.7667, df = 52, p-value = 0.08314
## alternative hypothesis: true mean is not equal to 8.798
## 95 percent confidence interval:
##  8.409771 8.822682
## sample estimates:
## mean of x
##  8.616226
diff(angle.t$conf.int) #width of CI
## [1] 0.4129112
```

We are interested in if it is plausible that the mean of all potential measurements using Short's device agrees with the currently accepted parallax value of 8.798. In notation, $H_0 : \mu = 8.798$ against $H_A : \mu \neq 8.798$. The one sample t-test results in a p-value of 0.08, which is greater than 0.05, and a 95% confidence interval of 8.41 to 8.82, which includes 8.798. Therefore, based on these results it is plausible for the population mean of all potential measurements of the parallax using Short's device to equal 8.798, and I fail to reject the null.

1(c) (10 pts) Repeat the analysis using a suitable non-parametric method, and contrast the results with part (b). Which analysis seems most reasonable, and what are your conclusions based on that analysis, given the question of interest?

```
# sign test
sign.t <- SIGN.test(angle, md = 8.798)
##
##  One-sample Sign-Test
##
## data:  angle
## s = 17, p-value = 0.01266
## alternative hypothesis: true median is not equal to 8.798
## 95 percent confidence interval:
##  8.394796 8.651301
## sample estimates:
## median of x
##         8.5
sign.t
##                    Conf.Level L.E.pt U.E.pt
## Lower Achieved CI      0.9466 8.4000 8.6500
## Interpolated CI        0.9500 8.3948 8.6513
## Upper Achieved CI      0.9730 8.3600 8.6600
diff(sign.t[2,c(2,3)]) # width of CI
## U.E.pt
## 0.2565
#wilcoxon test
w.t <- wilcox.test(angle, mu = 8.798, conf.int = TRUE)
w.t
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  angle
## V = 465, p-value = 0.02686
```

```
## alternative hypothesis: true location is not equal to 8.798
## 95 percent confidence interval:
##  8.430072 8.774983
## sample estimates:
## (pseudo)median
##       8.574997
diff(w.t$conf.int) #width of CI
## [1] 0.3449107
```

The sign test determines if it is plausible that the median of all potential measurements using Short's device agrees with the currently accepted parallax value of 8.798. In notation, $H_0 : \eta = 8.798$ against $H_A : \eta \neq 8.798$. The sign test results in a p-value of 0.01, which is less than 0.05, and an interpolated 95% confidence interval of 8.39 to 8.65, which does not include 8.798. Therefore, based on these results it is not plausible for the population median of all potential measurements of the parallax using Short's device to equal 8.798, and I reject the null in favor of the alternative.

The Wilcoxon procedure test determines if it is plausible that the mean of all potential measurements using Short's device agrees with the currently accepted parallax value of 8.798. In notation, $H_0 : \mu = 8.798$ against $H_A : \mu \neq 8.798$. The Wilcoxon test results in a p-value of 0.03, which is less than 0.05, and an interpolated 95% confidence interval of 8.430 to 8.775, which does not include 8.798. Therefore, based on these results it is not plausible for the population mean of all potential measurements of the parallax using Short's device to equal 8.798, and I reject the null in favor of the alternative.

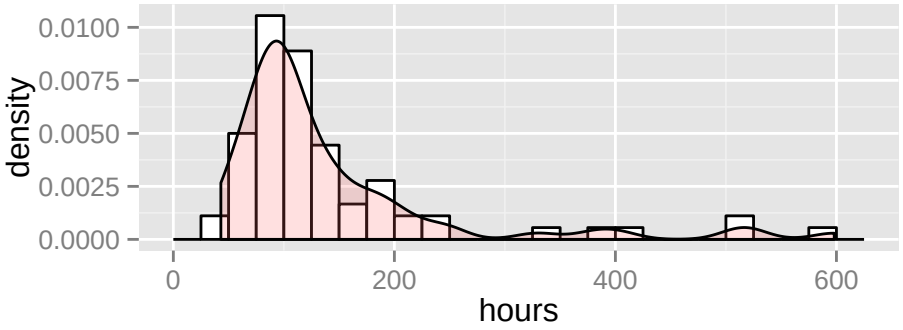
- (b): I believe the t-test is the most reasonable for this analysis because the question of interest is in regards to the population mean, which is the parameter evaluated by the t-test. The normality assumption of the t-test appears to be valid. The sign and Wilcoxon tests do not use the mean as a parameter, rather they are with respect to the median. Additionally, the median is considerably far from the mean based on the IQR. The Wilcoxon test assumes the sample distribution is symmetric, which this is not true based on the difference between the mean and median.

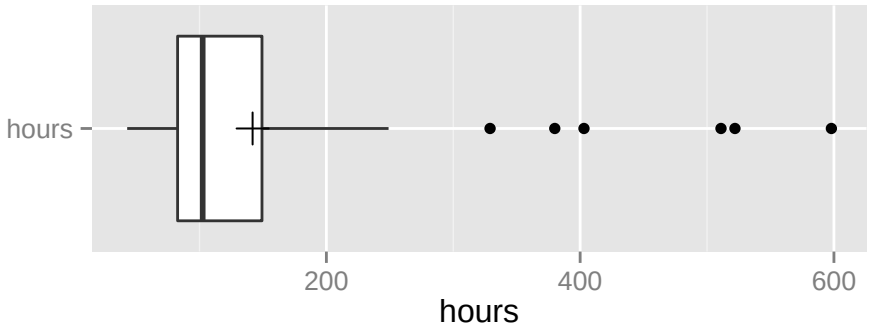
2 Guinea pigs:

```
# guinea <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_06_F14-2.csv")
# write.table(guinea, paste(dir, "p2.csv", sep = "/"), sep = ",", row.names = FALSE )
guinea <- read.csv(paste(dir, "p2.csv", sep = "/"))
hours <- guinea$hours
```

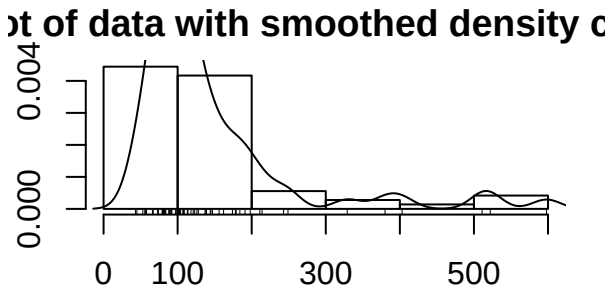
```
summary(hours)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.00   82.75  102.50  141.80  149.20  598.00
sd(hours) # standard deviation
## [1] 109.2086
max(hours) - min(hours) #spread
## [1] 555
fivenum(hours)[4] - fivenum(hours)[2] #IQR
## [1] 69
```

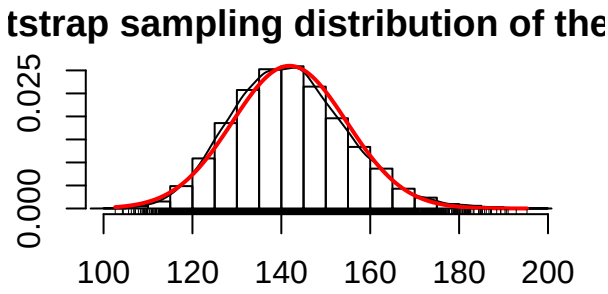
```
#create box plot
hours.hist <- ggplot(guinea, aes(x = hours))
hours.hist <- hours.hist + geom_histogram(aes( y= ..density..),
                                          binwidth = 25,color = "black", fill = "white")
hours.hist <- hours.hist + geom_density(alpha = 0.2, fill = "#FF6666")
hours.hist <- hours.hist + labs(x = "hours")

# boxplot of hours
hours.box <- ggplot(guinea, aes(x = "hours", y = hours)) # boxplot of hours
hours.box <- hours.box + geom_boxplot()
hours.box <- hours.box + coord_flip()
hours.box <- hours.box + labs(x = "", y = "hours")
hours.box <- hours.box + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 4)
# plot
grid.arrange(hours.hist, hours.box, nrow = 2)
```

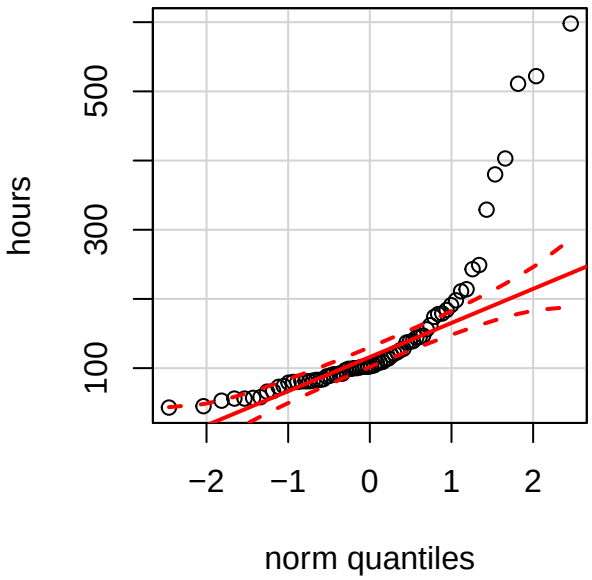





```
bs.one.samp.dist(hours) # bootstrap
qqPlot(hours)
```







2(a) (10 pts) Obtain a 95% t-CI for the mean survival time.)

```
hours.t <- t.test(hours, conf.level = 0.95)
hours.t
##
##  One Sample t-test
##
```

```
## data:  hours
## t = 11.0212, df = 71, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  116.1845 167.5100
## sample estimates:
## mean of x
##  141.8472
diff(hours.t$conf.int)
## [1] 51.32554
```

We are interested in the mean survival times in hours for guinea pigs after they were injected with a dose of tubercule bacilli. The 95% confidence interval, from the one sample t-test, indicates that the mean survival time for guinea pigs after injection ranges from 116.18 to 167.51 hours. This confidence interval has a width of 51.33 hours about the mean.

2(b) (10 pts) Repeat part (a) using a suitable nonparametric method.)

```
# sign test
sign.hr <- SIGN.test(hours)
##
##  One-sample Sign-Test
##
## data:  hours
## s = 72, p-value = 6.661e-16
## alternative hypothesis: true median is not equal to 0
## 95 percent confidence interval:
##   97.35184 120.47224
## sample estimates:
## median of x
##       102.5
diff(sign.hr[2,c(2,3)]) # width of CI
##  U.E.pt
## 23.1204
# #wilcoxon test
# w.hr <- wilcox.test(hours, conf.int = TRUE)
# w.hr
# diff(w.hr$conf.int) #width of CI
```

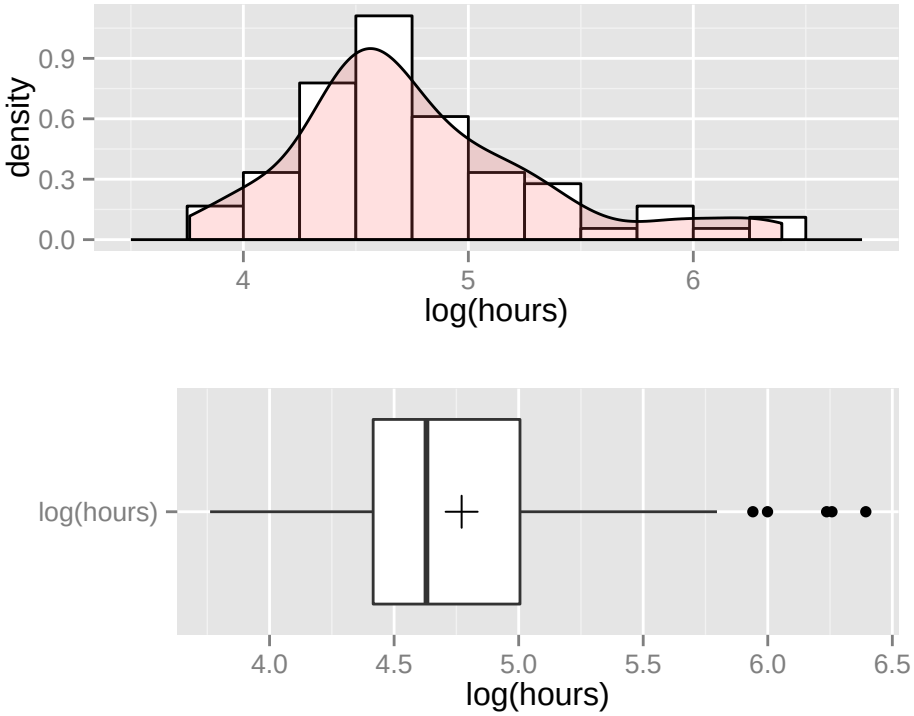
As seen in the boxplot, the data are skewed heavily to the right. The difference between the mean and median is over 50% of the IQR. From these observations and because the Wilcoxon test assumes symmetry, it is not appropriate for this analysis. The 95% confidence interval, about the median hours of survival for guinea pigs after injection, from the sign test is from 97.35 to 120.47 hours. This confidence interval has a width of 23.12 hours.

2(c) (10 pts) Take the log of survival time and find a 95% t-CI for mean log survival time.)

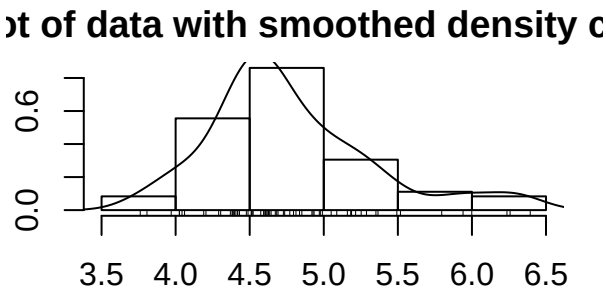
```
hours.log <- log(hours)
summary(hours.log)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.761   4.416   4.630   4.771   5.005   6.394
sd(hours.log) # standard deviation
## [1] 0.5595629
max(hours.log) - min(hours.log) #spread
## [1] 2.632391
fivenum(hours.log)[4] - fivenum(hours.log)[2] #IQR
## [1] 0.6073644
```

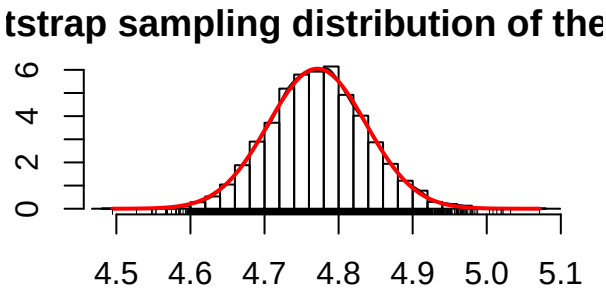
```
#create box plot
hours.log.hist <- ggplot(guinea, aes(x = hours.log))
hours.log.hist <- hours.log.hist + geom_histogram(aes(y = ..density..),
                                                  binwidth = .25, color = "black",
```

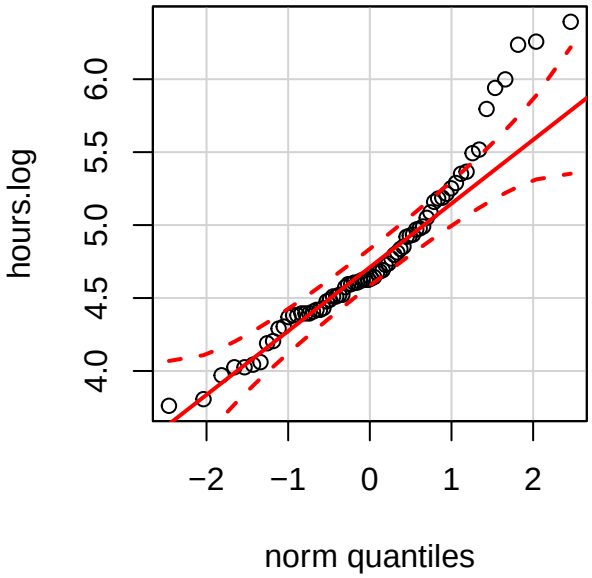
```
                                                fill = "white")
hours.log.hist <- hours.log.hist + labs(x = "log(hours)")
hours.log.hist <- hours.log.hist + geom_density(alpha = 0.2, fill = "#FF6666")

# boxplot of hours.log
hours.log.box <- ggplot(guinea, aes(x = "log(hours)", y = hours.log)) # boxplot of hours.log
hours.log.box <- hours.log.box + geom_boxplot()
hours.log.box <- hours.log.box + coord_flip()
hours.log.box <- hours.log.box + labs(x = "", y = "log(hours)")
hours.log.box <- hours.log.box + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 4)
# plot
grid.arrange(hours.log.hist, hours.log.box, nrow = 2)
```




```
bs.one.samp.dist(hours.log) # bootstrap
qqPlot(hours.log)
```








```
hours.tlog <- t.test(hours.log, conf.level = 0.95)
hours.tlog
##
##  One Sample t-test
##
## data:  hours.log
## t = 72.3491, df = 71, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.639577 4.902558
## sample estimates:
## mean of x
##  4.771067
diff(hours.tlog$conf.int)
## [1] 0.2629817
```

The 95% confidence interval, from the one sample t-test, indicates that the log transformed mean survival time for guinea pigs after injection ranges from 4.64 to 4.90 log(hours). This confidence interval has a width of 0.27 log(hours) about the mean.

2(d) (10 pts) Repeat part (c) using a suitable nonparametric method.

```
# # sign test
# sign.hr.log <- SIGN.test(hours.log)
# diff(sign.hr.log[2,c(2,3)]) # width of CI

#wilcoxon test
w.hr.log <- wilcox.test(hours.log, conf.int = TRUE)
w.hr.log
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  hours.log
## V = 2628, p-value = 1.691e-13
## alternative hypothesis: true location is not equal to 0
```

```
## 95 percent confidence interval:
##  4.599833 4.842302
## sample estimates:
## (pseudo)median
##       4.708047
diff(w.hr.log$conf.int) #width of CI
## [1] 0.2424688
```

The distribution for this test is nearly symmetric. The difference between the mean and median is approximately 20% of the IQR; therefore, because it is more powerful, the Wicoxon method will be used for this analysis.

The 95% confidence interval, about the log transformed median hours of survival for guinea pigs after injection, from the Wilcoxon test is from 4.60 to 4.84 log(hours). This confidence interval has a width of 0.24 log(hours).

2(e) (10 pts) Compare your 4 CIs, and contrast the nonparametric with the t-CIs. If they differ much, explain why they differ. Which analysis appears most appropriate? Explain.

Data Type	Analysis Method	Metric	CI
Untransformed	t-test	mean	141.85 ± 25.66
Untransformed	sign test	median	102.5 ± 11.56
Transformed	t-test	mean	4.77 ± 0.13
Transformed	Wilcoxon test	median	4.71 ± 0.12

- **Untransformed Data:** The t-test resulted in a CI with a width of 51.33 hours arond the mean of 141.85 hours, where the sign test resulted in a CI with a width of 23.12 hours around the median of 102.5 hours. The data are heavily skewed and not normal; therefore, the t-test and Wilcoxon tests are not suitable because their respective assumptions of normality and symmetry are not satisfied. The severe skewness of the data resulted in the mean being substantially larger than the median, which results in the mean not being representative of typical data. The sign test appears the most appropriate test method for evaluating the untransformed data.
- **Transformed Data:** The t-test resulted in a CI with a width of 0.26 log(hours) around the mean of 4.77 log(hours), where the Wicoxon test resulted in a CI with a width of 0.24 log(hours) around the median of 4.71 log(hours). The log transformation greatly minimized the skewness of the data, as is shown in the box plot. The boostrap distribution of the transformed data also appears normal. Based on these observations, Both the Wilcoxon and t-test assumptions are met. The width of CI from the parametric and nonparametric methods are nearly equal and either method appears appropriate for use on the transformed data.

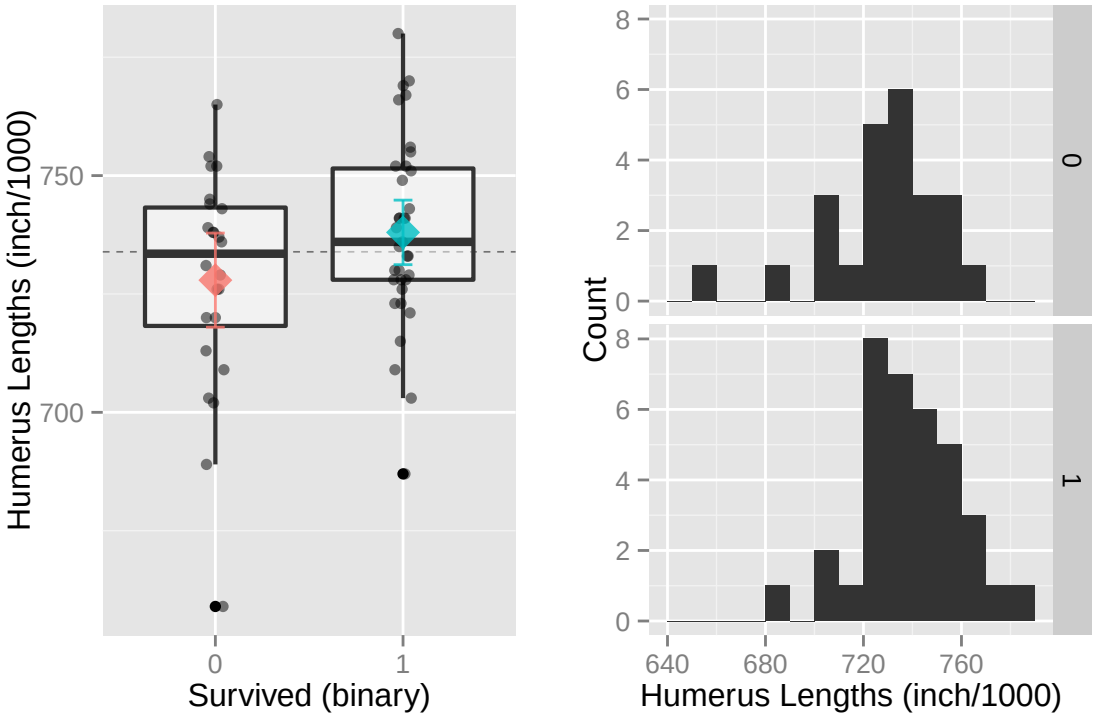
3 Humerus sparrows:

```
# sparrows <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_06_F14-3.csv")
# write.table(sparrows, paste(dir, "p3.csv", sep = "/"), sep = ",", row.names = FALSE )
sparrows <- read.csv(paste(dir, "p3.csv", sep = "/"))
sparrows$survived <- factor(sparrows$survived)
```

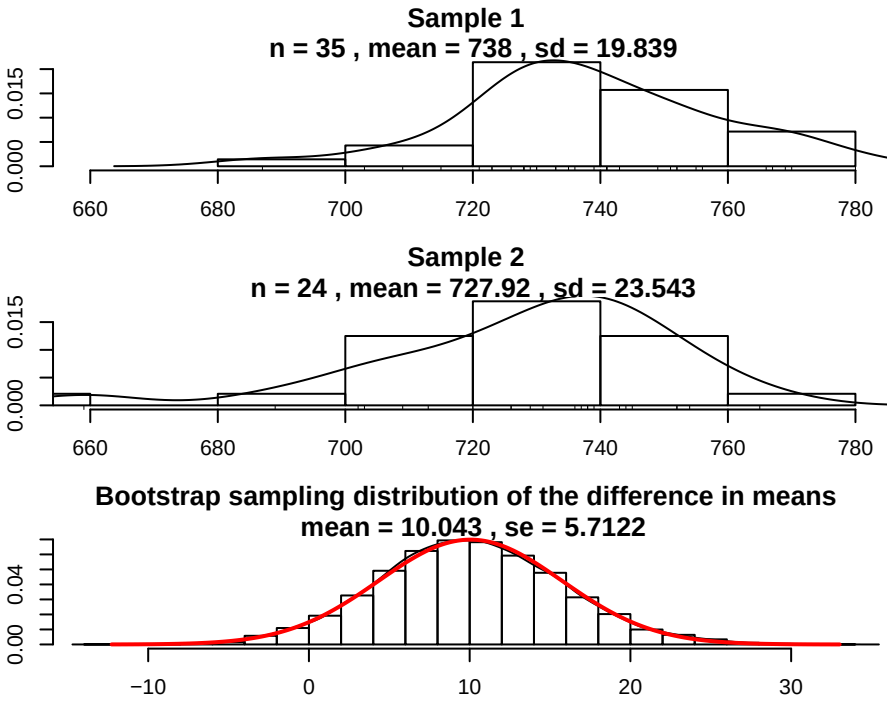
3(a) (10 pts) Make appropriate graphical displays to compare the humerus lengths in the two samples

```
#create box plot
spar.p <- ggplot(sparrows, aes(x = survived, y = humerus))
spar.p <- spar.p + geom_hline(yintercept = mean(sparrows$humerus), color = "black",
                              linetype = "dashed", size = 0.3,  alpha = 0.5)
spar.p <- spar.p + geom_boxplot(size = 0.75, alpha = 0.5) # boxplot
spar.p <- spar.p + geom_point(position = position_jitter(w = 0.05, h = 0),
                              alpha = 0.5, size = 2)
spar.p <- spar.p + stat_summary(fun.y = mean, geom = "point", shape = 18,
                                size = 6, aes(color = survived), alpha = 0.8)
spar.p <- spar.p + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
```

```
                              width = .1, aes(color = survived), alpha = 0.8)
spar.p <- spar.p + labs(y ="Humerus Lengths (inch/1000)",x = "Survived (binary)")

spar.p <- spar.p + guides(color = FALSE)
# create histograms
spar.hist <- ggplot(sparrows, aes(x = humerus)) + geom_histogram(binwidth = 10)
spar.hist <- spar.hist + facet_grid(survived ~ .)
spar.hist <- spar.hist+ labs(y = "Count", x = "Humerus Lengths (inch/1000)")
# plot boxplot and historgram
grid.arrange(spar.p, spar.hist, ncol = 2)
```




```
#boostrap of the mean
spar.1 <- sparrows[sparrows$survived == 1,1]
spar.0 <- sparrows[sparrows$survived != 1,1]

summary(spar.1)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   687.0   728.0   736.0   738.0   751.5   780.0
summary(spar.0)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   659.0   718.2   733.5   727.9   743.2   765.0
bs.two.samp.diff.dist(spar.1, spar.0)
```



3(b) **(10 pts) Test at the 5% level whether there is any difference in the population mean humerus lengths for those that perished and those that survived. Use both the t-test and an appropriate nonpara- metric procedure.**

```
# two sample t-test with equal variance
t.spar <- t.test(humerus ~ survived, data = sparrows, var.equal = TRUE)
t.spar
##
##  Two Sample t-test
##
## data:  humerus by survived
## t = -1.777, df = 57, p-value = 0.0809
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21.446053   1.279386
## sample estimates:
## mean in group 0 mean in group 1
##        727.9167        738.0000
# kruskal-Wallis test with equal variance
fit.spar <- kruskal.test(humerus ~ survived, data = sparrows)
fit.spar
##
##  Kruskal-Wallis rank sum test
##
## data:  humerus by survived
## Kruskal-Wallis chi-squared = 1.8881, df = 1, p-value = 0.1694
#WILCOXON-MANN-WHITNEY test of equal population dedians
wilcox.test(spar.1, spar.0, conf.int = TRUE)
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  spar.1 and spar.0
## W = 509, p-value = 0.1718
## alternative hypothesis: true location shift is not equal to 0
```

```
## 95 percent confidence interval:
##  -3.000012 19.000020
## sample estimates:
## difference in location
##               7.000016
```

The two sample t-test determines if it is plausible that the difference in the mean humerus length of sparrows that died and sparrows that survived is equal to zero. In notation, $H_0 : \mu_1 - \mu_2 = 0$ against $H_A : \mu_1 - \mu_2 \neq 0$. The two sample t-test resulted in a p-value of 0.08, which is greater than 0.05.

The Kruskal-Wallis tests if it is plausible that the difference in the median humerus length of sparrows that died and sparrows that survived is equal to zero. In notation, $H_0 : \eta_1 - \eta_2 = 0$ against $H_A : \eta_1 - \eta_2 \neq 0$. The Kruskal-Wallis test resulted in a p-value of 0.17, which is also greater than 0.05. Therefore, at the 5% level, I fail to reject the null hypothesis and there is no difference between the median humerus length of sparrows that survived and sparrows that died.

3(c) (10 pts) Compute and interpret a 95% CI for the difference in population mean humerus lengths for those that perished and those that survived. Repeat for an appropriate nonparametric procedure.

The 95% confidence interval from the two sample t-test for the difference in the mean humerus length ranges from -21.45 to 1.28 (inch/1000). The 95% confidence interval from the Wilcoxon-Mann-Whitney for the difference of population median humerus length resulted in a range from -3.00 to 19.00.

3(d) (10 pts) Discuss any statistical assumptions that you have made in carrying out the analyses, and whether the assumptions seem reasonable.

These methods both assumed that independent random samples from two populations were utilized. Also, the Wicoxon-Mann-Whitney test assumed that the samples came from populations having a similar distribution, e.g., similar shaped distributions. The two sample t-test relies on the assumption that the distribution of the differences is normal. Based on the plots, these assumptions appear valid.

3(e) (10 pts) Write a short summary for the problem. What analysis seems most appropriate?

The two sample t-test determines if it is plausible that the difference in the mean humerus length of sparrows that died and sparrows that survived is equal to zero. In notation, $H_0 : \mu_1 - \mu_2 = 0$ against $H_A : \mu_1 - \mu_2 \neq 0$.

The Kruskal-Wallis tests if it is plausible that the difference in the median humerus length of sparrows that died and sparrows that survived is equal to zero. In notation, $H_0 : \eta_1 - \eta_2 = 0$ against $H_A : \eta_1 - \eta_2 \neq 0$. The Kruskal-Wallis test resulted in a p-value of 0.17, which is greater than 0.05. Therefore, at the 5% level, I fail to reject the null hypothesis and there is no difference between the median humerus length of sparrows that survived and sparrows that died.

Based on the assumptions for the respective tests, both tests appear appropriate, but the t-test assumption of normality is the most rigorous assupmtion. Because the two sample t-test with equal variance has the most rigorous assumptions that are satisified, it is the most appropriate. The two sample t-test resulted in a p-value of 0.08, which is greater than 0.05.

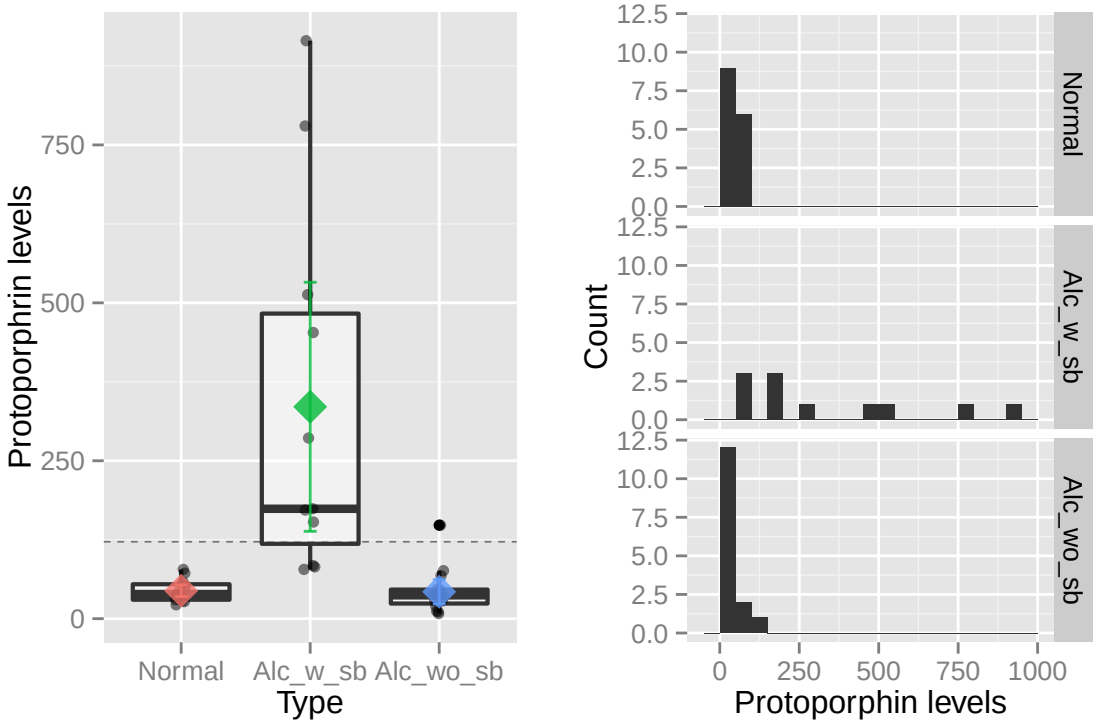
4 Protoporphin levels among alcoholics:

```
# proto <- read.table("http://statacumen.com/teach/ADA1/ADA1_HW_06_F14-4.txt")
# write.table(proto, paste(dir, "p4.csv", sep = "/"), sep = ",", row.names = FALSE )
proto <- read.csv(paste(dir, "p4.csv", sep = "/"), header = TRUE, stringsAsFactors = FALSE)
proto$row <- seq(1:nrow(proto))
proto.long <- melt(proto, id.vars = c("row"),
                   measure.vars = c("Normal", "Alc_w_sb", "Alc_wo_sb"),
                   variable.name = "type", value.name = "level", na.rm = TRUE)
```

4(a) Analysis if Untransformed Data

```
#create box plot
proto.p <- ggplot(proto.long, aes(x = type, y = level))
proto.p <- proto.p + geom_hline(yintercept = mean(proto.long$level), color = "black",
```

```
                              linetype = "dashed", size = 0.3,  alpha = 0.5)
proto.p <- proto.p + geom_boxplot(size = 0.75, alpha = 0.5) # boxplot
proto.p <- proto.p + geom_point(position = position_jitter(w = 0.05, h = 0),
                              alpha = 0.5, size = 2)
proto.p <- proto.p + stat_summary(fun.y = mean, geom = "point", shape = 18,
                                size = 6, aes(color = type), alpha = 0.8)
proto.p <- proto.p + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                                width = .1, aes(color = type), alpha = 0.8)
proto.p <- proto.p + labs(y ="Protoporphrin levels",x = "Type")
proto.p <- proto.p + guides(color = FALSE)

# create histograms
proto.hist <- ggplot(proto.long, aes(x = level)) + geom_histogram(binwidth = 50)
proto.hist <- proto.hist + facet_grid(type ~ .)
proto.hist <- proto.hist + labs(y = "Count", x = "Protoporphin levels")
# plot boxplot and historgram
grid.arrange(proto.p, proto.hist, ncol = 2)
```




```
#statistical summary of each type
by(proto.long$level, proto.long$type, summary)
## proto.long$type: Normal
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.00   30.00   39.00   44.27   54.50   78.00
## ------------------------------------------------------------
## proto.long$type: Alc_w_sb
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    78.0   118.5   174.0   335.5   483.0   915.0
## ------------------------------------------------------------
## proto.long$type: Alc_wo_sb
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   24.00   37.00   42.53   46.00  148.00
by(proto.long$level, proto.long$type, function(X) {c(IQR(X), sd(X), length(X))})
## proto.long$type: Normal
## [1] 24.50000 16.79909 15.00000
## ------------------------------------------------------------
```

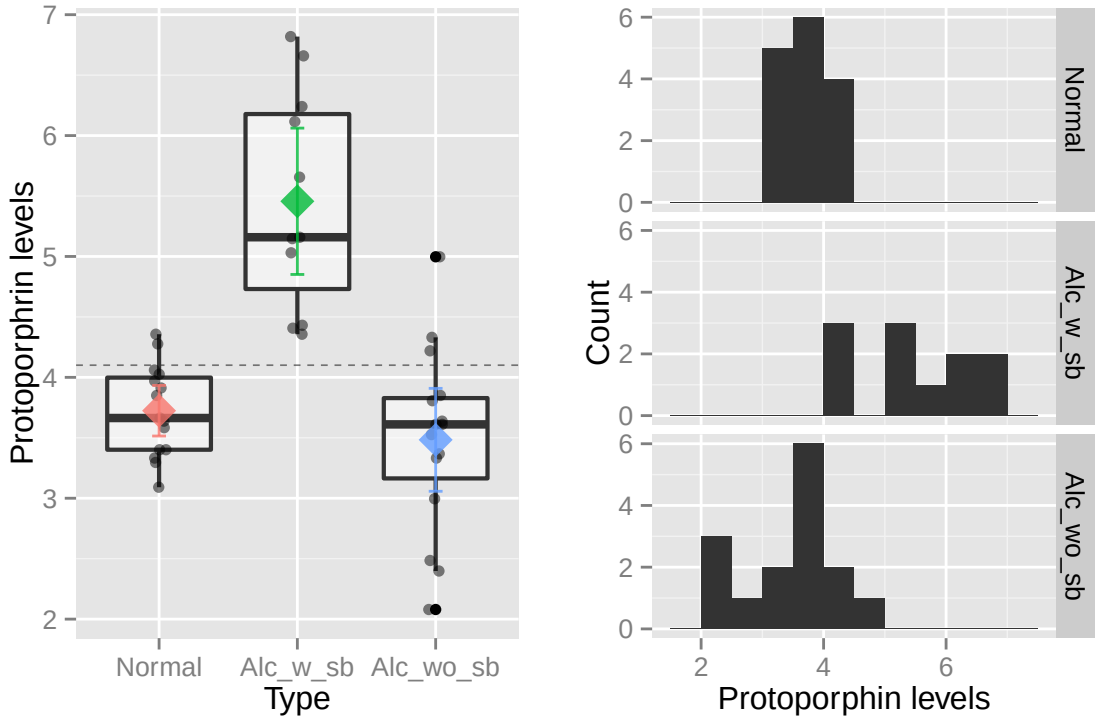
```
## proto.long$type: Alc_w_sb
## [1] 364.5000 293.4459  11.0000
## ------------------------------------------------------------
## proto.long$type: Alc_wo_sb
## [1] 22.00000 34.94254 15.00000
```

```
# kruskal-Wallis test with equal variance
fit.proto <- kruskal.test(level ~ type, data = proto.long)
fit.proto
##
##  Kruskal-Wallis rank sum test
##
## data:  level by type
## Kruskal-Wallis chi-squared = 23.1182, df = 2, p-value = 9.549e-06
#group comparisons
#WILCOXON-MANN-WHITNEY test of equal population dedians
wilcox.test(proto$Normal, proto$Alc_w_sb, conf.int = TRUE)
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  proto$Normal and proto$Alc_w_sb
## W = 0.5, p-value = 2.324e-05
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -435.00001  -61.99995
## sample estimates:
## difference in location
##                   -144
wilcox.test(proto$Normal, proto$Alc_wo_sb, conf.int = TRUE)
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  proto$Normal and proto$Alc_wo_sb
## W = 137.5, p-value = 0.3093
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -8.000014 21.000040
## sample estimates:
## difference in location
##               8.000074
wilcox.test(proto$Alc_wo_sb, proto$Alc_w_sb, conf.int = TRUE)
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  proto$Alc_wo_sb and proto$Alc_w_sb
## W = 3, p-value = 4.117e-05
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -442.00000  -71.99999
## sample estimates:
## difference in location
##                   -146
```

4(b) Analysis of Log Transformed Data

```
#create log transformed data
proto.long$log <- log(proto.long$level)
#create box plot
proto.log.p <- ggplot(proto.long, aes(x = type, y = log))
```

```
proto.log.p <- proto.log.p + geom_hline(yintercept = mean(proto.long$log), color = "black",
                                linetype = "dashed", size = 0.3,  alpha = 0.5)
proto.log.p <- proto.log.p + geom_boxplot(size = 0.75, alpha = 0.5) # boxplot
proto.log.p <- proto.log.p + geom_point(position = position_jitter(w = 0.05, h = 0),
                                alpha = 0.5, size = 2)
proto.log.p <- proto.log.p + stat_summary(fun.y = mean, geom = "point", shape = 18,
                                  size = 6, aes(color = type), alpha = 0.8)
proto.log.p <- proto.log.p + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                                  width = .1, aes(color = type), alpha = 0.8)
proto.log.p <- proto.log.p + labs(y ="Protoporphrin levels",x = "Type")
proto.log.p <- proto.log.p + guides(color = FALSE)

# create histograms
proto.log.hist <- ggplot(proto.long, aes(x = log)) + geom_histogram(binwidth = .5)
proto.log.hist <- proto.log.hist + facet_grid(type ~ .)
proto.log.hist <- proto.log.hist + labs(y = "Count", x = "Protoporphin levels")
# plot boxplot and historgram
grid.arrange(proto.log.p, proto.log.hist, ncol = 2)
```




```
#statistical summary of log transformed data
by(proto.long$log, proto.long$type, summary)
## proto.long$type: Normal
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.091   3.401   3.664   3.724   3.998   4.357
## ------------------------------------------------------------
## proto.long$type: Alc_w_sb
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.357   4.731   5.159   5.457   6.178   6.819
## ------------------------------------------------------------
## proto.long$type: Alc_wo_sb
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.079   3.164   3.611   3.483   3.828   4.997
by(proto.long$log, proto.long$type, function(X) {c(IQR(X), sd(X), length(X))})
## proto.long$type: Normal
## [1]  0.5966244  0.3778145 15.0000000
```

```
## ------------------------------------------------------------
## proto.long$type: Alc_w_sb
## [1]  1.4474566  0.9011606 11.0000000
## ------------------------------------------------------------
## proto.long$type: Alc_wo_sb
## [1]  0.6644367  0.7685407 15.0000000
```

```
# ANOVA on log transformed data
fit.proto.log <- aov(log ~ type, data = proto.long)
summary(fit.proto.log)
##             Df Sum Sq Mean Sq F value   Pr(>F)
## type         2  28.07  14.035      29 2.25e-08 ***
## Residuals   38  18.39   0.484
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# from long to wide
proto.wide <- dcast(proto.long, row ~ type, value.var = "log")

# Tukey
TukeyHSD(fit.proto.log)
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log ~ type, data = proto.long)
##
## $type
##                          diff        lwr        upr     p adj
## Alc_w_sb-Normal     1.7326582  1.0592061  2.4061103 0.0000007
## Alc_wo_sb-Normal   -0.2406837 -0.8601695  0.3788021 0.6140177
## Alc_wo_sb-Alc_w_sb -1.9733419 -2.6467940 -1.2998898 0.0000000
```

- **Graphical Summary:** Considering either the mean or the median, the typical protoporphin level increases dramatically when sideroblasts are present, where protoporphin levels in the normal and alcoholic populations appear similar. Additionally, the IQR and standard deviation in protoporhin levels is substantially greater in the population with sideroblasts than those populations without sideroblasts. The difference in the means, medians, and variance appears substantial across the three populations.

  The log transformed data have much less spread and variances that are far more consistent across the three populations. The log transformation did not alter the order of respecstive values for mean and median.

- **Kruskal-Wallis ANOVA:** Let η_0 be the median protoporphin level in a normal population and define η_1 and η_2 as the median protoporpin level in a population of alcoholics with sideroblasts and a population of alcoholics in without sideroblasts. We are interested in if it is plausible that the difference in the median protophorphin levels across three populations is equal. In notation, $H_0 : \eta_1 = \eta_2 = \eta_3$ against H_A : not H_0. The p-value from this test was $9.5 * 10^{-6}$, which is less than 0.05; therefore, I reject the null hypothesis in favor of the alternative.

- **Assumptions of Kruskal-Wallis:** These methods assumed that independent random samples from two populations were utilized and that the samples came from populations having an identical distribution, e.g., distributions with similar shape and spread.

- **Groupings from Wicoxon Rank Tests:**

  ```
  Alc_wo_sb   Normal   Alc_w_sb
  ----------------------   -------
  ```

- **ANOVA of Log Transformed Data:** Let μ_0 be the mean protoporphin level in a normal population and define μ_1 and μ_2 as the mean protoporpin level in a population of alcoholics with sideroblasts and a population of alcoholics in without sideroblasts. We are interested in if it is plausible that the difference in the mean protophorphin levels across three populations is equal. In notation, $H_0 : \mu_1 = \mu_2 = \mu_3$ against H_A : not H_0.The p-value from this test was $2.3 * 10^{-5}$, which is less than 0.05; therefore, I reject the null hypothesis in favor of the alternative.

- **Assumptions of ANOVA:** These methods assumed that independent random samples from two populations were utilized and that the samples came from populations with normal distributions having similar variance.

- **Groupings from Tukey HSD**

```
Alc_wo_sb   Normal   Alc_w_sb
-----------------   -------
```

- **Conclusions:** Based on the non parametric Kruskal-Wallis ANOVA of the untransformed data and the parametric ANOVA of the log transformed data, both the median and mean protoporphin levels are different across the three populations. The different analysis methods resulted in similar groupings where the normal and alcoholic with out sideroblasts populations had similar means and medians at the 5% levels, and the alcoholic with sideroblasts population had a different mean and median from the other two populations at the 5% level.