Advanced Data Analysis 1, Stat 427/527 Homework 6 Name: Answer Key
Fall 2014 Prof. Erik B. Erhardt

Part I. (175 points) Do all calculations in R. All R code for the assignment should be included with the part of the problem it addresses (for code and output use a fixed-width font, such as Courier). Code is used to calculate result. Text is used to report and interpret results. Do not report or interpret results in the code. Also:

1. Clearly define population parameters in each problem. That is, give a verbal description of what the population mean is in the context of the problem.
2. Clearly specify hypotheses when appropriate (not every problem involves a test of hypothesis).
3. Write a coherent conclusion based on each CI or test.

(25pts) **1. Parallax:** The following determinations of the parallax of the sun (the angle spanned by the earth's radius as if it were viewed and measured from the sun's surface) were made in 1761 by noted astronomer James Short. The units are in seconds of a degree (1/360 degree).

```
8.50   8.06   8.65   9.71   8.80   7.99   8.50   8.43   8.35
8.50   8.40   8.58   7.33   8.44   8.71   8.28   8.82   8.34
8.64   8.14   8.31   9.87   9.02   9.64   9.27   7.68   8.36
8.86  10.57   8.34   9.06  10.34   8.58   5.76   9.11   8.55
9.25   8.07   7.80   8.44   8.66   9.54   9.09   8.36   7.71
8.23   8.34   9.07   8.50   9.71   8.30   8.50   8.60
```

```
parallax <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_06_F14-1.csv")
angle <- parallax$angle;
```

With a careful determination of the radius of the earth and a good average value of the parallax, the average distance of the earth to the sun can be obtained. The currently accepted value of the parallax is 8.798.
Within this framework, define

$$\mu = \text{the population mean of all potential measurements of the parallax using Short's device.}$$

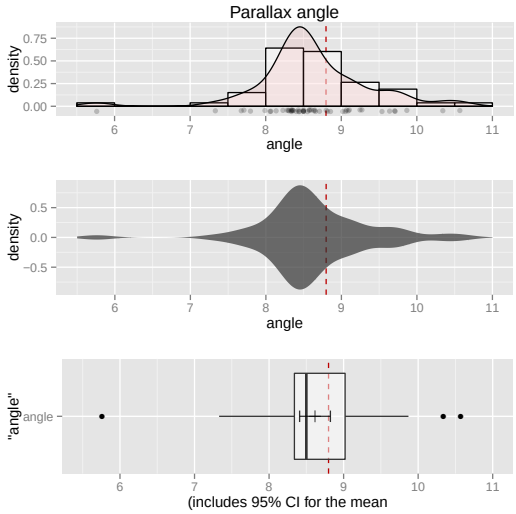
We are interested in whether μ could be the currently accepted value of the parallax of the sun, that is, we wish to test whether $\mu = 8.798$.

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

Solution: The distribution is roughly symmetric, unimodal, but heavy-tailed with outliers. The AD normality test (and others) rejects normality at the 0.05 level.

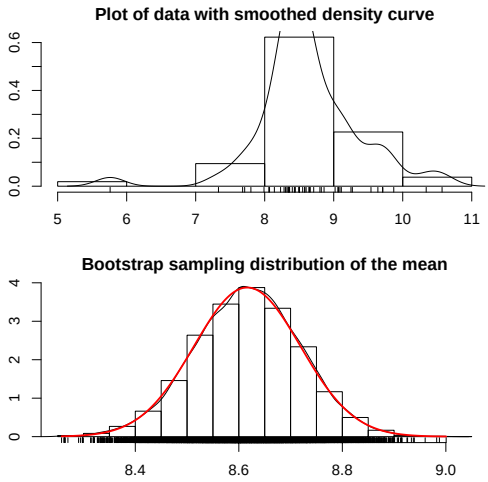
```
library(ggplot2)
  # Histogram with density instead of count on y-axis
p1 <- ggplot(parallax, aes(x = angle))
p1 <- p1 + scale_x_continuous(limits=c(5.5,11))
p1 <- p1 + geom_vline(aes(xintercept=8.798), colour="#BB0000", linetype="dashed")
p1 <- p1 + geom_histogram(aes(y=..density..)
                          , binwidth=0.5
                          , colour="black", fill="white", alpha = 0.5)
  # Overlay with transparent density plot
p1 <- p1 + geom_density(alpha=0.1, fill="#FF6666")
p1 <- p1 + geom_point(aes(y = -0.05)
                    , position = position_jitter(height = 0.01)
                    , alpha = 1/5)
p1 <- p1 + labs(title = "Parallax angle")

# violin plot
p2 <- ggplot(parallax, aes(x = angle))
p2 <- p2 + scale_x_continuous(limits=c(5.5,11))
p2 <- p2 + geom_vline(aes(xintercept=8.798), colour="#BB0000", linetype="dashed")
p2 <- p2 + geom_ribbon(aes(ymax = ..density.., ymin = -..density..)
                       , stat = "density", alpha = 0.7)

# boxplot
p3 <- ggplot(parallax, aes(x = "angle", y = angle))
p3 <- p3 + scale_y_continuous(limits=c(5.5,11))
```

```
p3 <- p3 + geom_hline(aes(yintercept=8.798), colour="#BB0000", linetype="dashed")
p3 <- p3 + geom_boxplot(alpha = 0.5)
# diamond at mean for each group
p3 <- p3 + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 3, alpha = 0.8)
# confidence limits based on normal distribution
p3 <- p3 + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                        width = .1, alpha = 0.8)
p3 <- p3 + coord_flip()
p3 <- p3 + ylab("(includes 95% CI for the mean")

library(gridExtra)
grid.arrange(p1, p2, p3, ncol=1)
```



The normality assumption of the sample mean for a one-sample test is satisfied.

```
bs.one.samp.dist(parallax$angle)
```




```
#shapiro.test(parallax$angle)
library(nortest)
ad.test(parallax$angle)
```

```
##
##  Anderson-Darling normality test
##
## data:  parallax$angle
```

```
## A = 1.52, p-value = 0.0005687
# lillie.test(parallax$angle)
#cvm.test(parallax$angle)
```

(b) (10 pts) Perform the standard t-test on these data, at the 5% level, and construct a 95% CI for μ. Interpret the results, given the question of interest.

Solution: While this is a fairly large sample and the data is symmetric, because of the outliers, the t-test may have lower power than the Wilcoxon test.
Because the p-value=0.083 > 0.05, we fail to reject H_0, concluding that we have insufficient evidence to conclude that the mean parallax is different from 8.798.
We are 95% confident that the true population mean is in the interval $(8.40977, 8.82268)$. Note the hypothesized mean is in the CI, consistent with the result of the t-test.

```
t.test(parallax$angle, mu=8.798)
##
##  One Sample t-test
##
## data:  parallax$angle
## t = -1.767, df = 52, p-value = 0.08314
## alternative hypothesis: true mean is not equal to 8.798
## 95 percent confidence interval:
##  8.410 8.823
## sample estimates:
## mean of x
##     8.616
```

(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?

Solution: Because the distribution is roughly symmetric, we use the Wilcoxon signed-rank test for the median, which is close to the mean because of symmetry. Because the p-value= 0.027 < 0.05, we reject the null hypothesis in favor of the alternative hypothesis concluding that the population median is different from 8.798.
We are 95% confident that the true population median is in the interval $(8.430, 8.770)$.
This analysis seems more appropriate because of the non-normality of the data and outliers. (Furthermore, note that the CI for the median is narrower than the CI for the mean.)

```
summary(parallax$angle)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.76    8.34    8.50    8.62    9.02   10.60
# with continuity correction in the normal approximation for the p-value
wilcox.test(parallax$angle, mu=8.798, conf.int=TRUE)
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  parallax$angle
## V = 465, p-value = 0.02686
## alternative hypothesis: true location is not equal to 8.798
## 95 percent confidence interval:
##  8.430 8.775
## sample estimates:
## (pseudo)median
##          8.575
```

(50pts) **2. Guinea pigs:** The data below are the survival times in hours of 72 guinea pigs after they were injected with a given dose of tubercule bacilli in a medical experiment. The data are from the article "Acquisition of resistance of guinea pigs injected with different doses of virulent tubercule bacilli," by T. Bjerkedal in the American Journal of Hygiene, (1960), pp. 130–148.

43	45	53	56	56	57	58	66	67	73	74	
79	80	80	81	81	81	82	83	83	84	88	
89	91	91	92	92	97	99	99	100	100	101	
102	102	102	103	104	107	108	109	113	114	118	
121	123	126	128	137	138	139	144	145	147	156	
162	174	178	179	184	191	198	211	214	243	249	
329	380	403	511	522	598						

```
guinea <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_06_F14-2.csv")
hours <- guinea$hours;
```

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

Solution: Because of the strong skewness of the sample, a t-CI does not make sense (data are not normal).
The 95% CI for survival time is $(116.184, 167.510)$.

```
summary(guinea$hours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    43.0    82.8   102.0   142.0   149.0   598.0
```

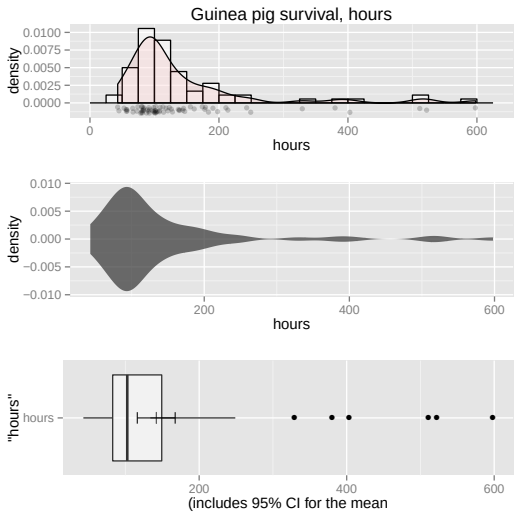
```
t.test(guinea$hours)
```

```
##
##  One Sample t-test
##
## data:  guinea$hours
## t = 11.02, df = 71, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  116.2 167.5
## sample estimates:
## mean of x
##     141.8
```

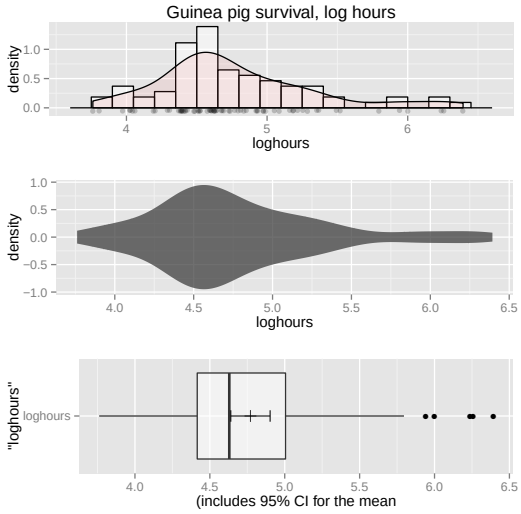
Here are plots on the hours and log(hours) scales.

```
library(ggplot2)
  # Histogram with density instead of count on y-axis
p1 <- ggplot(guinea, aes(x = hours))
p1 <- p1 + geom_histogram(aes(y=..density..)
                          , binwidth=25
                          , colour="black", fill="white", alpha = 0.5)
  # Overlay with transparent density plot
p1 <- p1 + geom_density(alpha=0.1, fill="#FF6666")
p1 <- p1 + geom_point(aes(y = -0.001)
                      , position = position_jitter(height = 0.0005)
                      , alpha = 1/5)
p1 <- p1 + labs(title = "Guinea pig survival, hours")

# violin plot
p2 <- ggplot(guinea, aes(x = hours))
p2 <- p2 + geom_ribbon(aes(ymax = ..density.., ymin = -..density..)
                       , stat = "density", alpha = 0.7)

# boxplot
p3 <- ggplot(guinea, aes(x = "hours", y = hours))
p3 <- p3 + geom_boxplot(alpha = 0.5)
# diamond at mean for each group
p3 <- p3 + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 3, alpha = 0.8)
# confidence limits based on normal distribution
p3 <- p3 + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                        width = .1, alpha = 0.8)
p3 <- p3 + coord_flip()
p3 <- p3 + ylab("(includes 95% CI for the mean")
```

```
library(gridExtra)
grid.arrange(p1, p2, p3, ncol=1)
```




```
guinea$loghours <- log(guinea$hours);
library(ggplot2)
  # Histogram with density instead of count on y-axis
p1 <- ggplot(guinea, aes(x = loghours))
p1 <- p1 + geom_histogram(aes(y=..density..)
                          , binwidth=0.15
                          , colour="black", fill="white", alpha = 0.5)
  # Overlay with transparent density plot
p1 <- p1 + geom_density(alpha=0.1, fill="#FF6666")
p1 <- p1 + geom_point(aes(y = -0.05)
                    , position = position_jitter(height = 0.01)
                    , alpha = 1/5)
p1 <- p1 + labs(title = "Guinea pig survival, log hours")

# violin plot
p2 <- ggplot(guinea, aes(x = loghours))
p2 <- p2 + geom_ribbon(aes(ymax = ..density.., ymin = -..density..)
                       , stat = "density", alpha = 0.7)

# boxplot
p3 <- ggplot(guinea, aes(x = "loghours", y = loghours))
p3 <- p3 + geom_boxplot(alpha = 0.5)
# diamond at mean for each group
p3 <- p3 + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 3, alpha = 0.8)
# confidence limits based on normal distribution
p3 <- p3 + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                        width = .1, alpha = 0.8)
p3 <- p3 + coord_flip()
p3 <- p3 + ylab("(includes 95% CI for the mean")

library(gridExtra)
grid.arrange(p1, p2, p3, ncol=1)
```



The Anderson-Darling test for normality rejects normality for the population distribution both on the hours and log(hours) scales.

```
library(nortest)
ad.test(guinea$hours)
```

```
##
##  Anderson-Darling normality test
##
## data:  guinea$hours
## A = 7.517, p-value < 2.2e-16
```

```
ad.test(guinea$loghours)
```

```
##
##  Anderson-Darling normality test
##
## data:  guinea$loghours
## A = 1.577, p-value = 0.0004277
```

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

Solution: Because the data are skewed, the sign test for the median is used.
The 95% sign test CI for the population median is $(97.4, 120.5)$.

```
library(BSDA)
```

```
## Loading required package:  e1071
## Loading required package:  class
##
## Attaching package:  'e1071'
##
## The following object is masked from 'package:Hmisc':
##
##    impute
##
##
## Attaching package:  'BSDA'
##
## The following objects are masked from 'package:car':
##
##    Vocab, Wool
##
## The following object is masked from 'package:datasets':
##
##    Orange
```

```
SIGN.test(guinea$hours)
```

```
##
##  One-sample Sign-Test
##
## data:  guinea$hours
## s = 72, p-value = 6.661e-16
## alternative hypothesis: true median is not equal to 0
## 95 percent confidence interval:
##   97.35 120.47
## sample estimates:
## median of x
##       102.5
##                   Conf.Level L.E.pt U.E.pt
## Lower Achieved CI     0.9236  99.00  118.0
## Interpolated CI       0.9500  97.35  120.5
## Upper Achieved CI     0.9556  97.00  121.0
```

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

Solution: While the distribution is more symmetric, the AD normality test rejects normality at the 0.05 level. Therefore, the t-CI of $(4.63958, 4.90256)$ may still be an unreliable summary.

```
summary(guinea$loghours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.76    4.42    4.63    4.77    5.01    6.39
```

```
t.test(guinea$loghours)
```

```
##
##  One Sample t-test
##
## data:  guinea$loghours
## t = 72.35, df = 71, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.640 4.903
## sample estimates:
## mean of x
##     4.771
```

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

Solution: The sign-rank test gives a 95% CI for the population median of $(4.578, 4.791)$, while the Wilcoxon signed-rank CI is $(4.600, 4.842)$. Because the distribution of the log(hours) is still skewed, I trust the sign-rank over the Wilcoxon, but both give similar intervals.

```
library(BSDA)
SIGN.test(guinea$loghours)
```

```
##
##  One-sample Sign-Test
##
## data:  guinea$loghours
## s = 72, p-value = 6.661e-16
## alternative hypothesis: true median is not equal to 0
## 95 percent confidence interval:
##  4.578 4.791
## sample estimates:
## median of x
##        4.63
##                   Conf.Level L.E.pt U.E.pt
## Lower Achieved CI     0.9236  4.595  4.771
## Interpolated CI       0.9500  4.578  4.791
## Upper Achieved CI     0.9556  4.575  4.796
```

```
wilcox.test(guinea$loghours, conf.int=TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
```

```
##
## data:  guinea$loghours
## V = 2628, p-value = 1.691e-13
## alternative hypothesis: true location is not equal to 0
## 95 percent confidence interval:
##  4.600 4.842
## sample estimates:
## (pseudo)median
##          4.708
```

(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.

Solution: Hours: t $(116.184, 167.510)$, sign-rank $(97.4, 120.5)$
log(hours): t $(4.63958, 4.90256)$, sign-rank $(4.578, 4.791)$, Wilcoxon $(4.600, 4.842)$
The sign-rank test gives narrower intervals for both hours and log(hours) due to the strong skewness of both distributions.
As discussed above, because of the skewness of the hours and log(hours) distributions, the sign-rank CI is most appropriate for the population median.

(50^{pts}) **3. Humerus sparrows:** In an 1898 Biology lecture at Woods Hole, Massachusetts, Hermon Bumpus reminded the audience that the process of natural selection for evolutionary change was an unproved theory. As evidence in support of natural selection, he presented measurements on house sparrows brought to his Brown University laboratory after an uncommonly severe winter storm. Some of the birds had died, and some had survived. Bumpus asked whether those that perished did so because they lacked physical characteristics enabling them to withstand the intensity of that particular instance of selective elimination.
The data we will look at are the humerus (arm bone) lengths (inch/1000) for the 24 adult male sparrows that perished and the 35 adult males that survived. You will see that the data are in two columns. The first column contains the humerus lengths for the 59 birds. The second column identifies whether the birds perished (0) or survived (1).

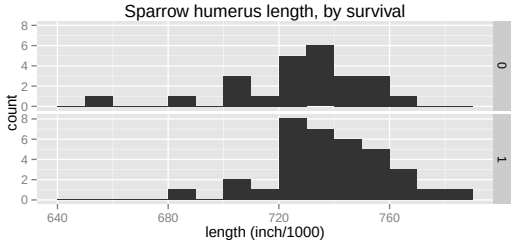
humerus	survived	humerus	survived
659	0	687	1
689	0	703	1
703	0	709	1
702	0	715	1
709	0	728	1
713	0	721	1
720	0	729	1
729	0	723	1
726	0	728	1
726	0	723	1
720	0	726	1
737	0	728	1
739	0	736	1
731	0	733	1
738	0	730	1
736	0	733	1
738	0	730	1
744	0	739	1
745	0	735	1
743	0	741	1
754	0	741	1
752	0	749	1
752	0	741	1
765	0	743	1
		741	1
		752	1
		752	1
		751	1
		756	1
		755	1
		766	1
		767	1
		769	1
		770	1
		780	1

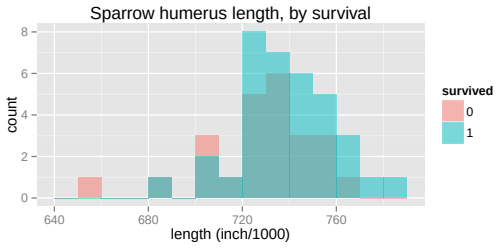
```
sparrows <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_06_F14-3.csv")
sparrows$survived <- factor(sparrows$survived)
```

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

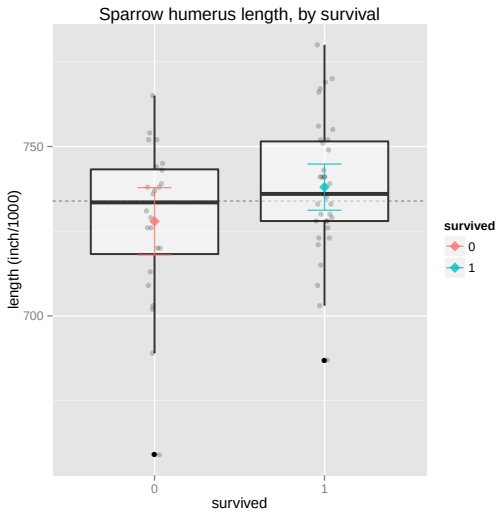
Solution: While there is some left-skewness indicated in the survived=0, neither plot is far from normal. The AD test of normality of both distribution fails to reject the null hypothesis that the distributions are normal. Furthermore, the boxplots do not indicate a relationship between the mean and variance of the distributions. Finally, tests of equal variances from the two samples fail to reject the null hypothesis of being equal (see plot), therefore we will use a pooled variance estimate. Therefore, the two-sample t-test with pooled variance is an appropriate test to compare mean humerus lengths between the sparrows that survived and those that did not.

```
# histogram using ggplot
p1 <- ggplot(sparrows, aes(x = humerus))
p1 <- p1 + geom_histogram(binwidth = 10)
p1 <- p1 + facet_grid(survived ~ .)
p1 <- p1 + labs(title = "Sparrow humerus length, by survival")
p1 <- p1 + xlab("length (inch/1000)")
#print(p1)

p2 <- ggplot(sparrows, aes(x = humerus, fill=survived))
p2 <- p2 + geom_histogram(binwidth = 10, alpha = 1/2, position="identity")
p2 <- p2 + labs(title = "Sparrow humerus length, by survival")
p2 <- p2 + xlab("length (inch/1000)")
#print(p2)

library(gridExtra)
grid.arrange(p1, p2, ncol=1)
```






```
# Plot the data using ggplot
library(ggplot2)
p <- ggplot(sparrows, aes(x = survived, y = humerus))
# plot a reference line for the global mean (assuming no groups)
p <- p + geom_hline(aes(yintercept = mean(humerus)),
                    colour = "black", linetype = "dashed", size = 0.3, alpha = 0.5)
# boxplot, size=.75 to stand out behind CI
p <- p + geom_boxplot(size = 0.75, alpha = 0.5)
# points for observed data
p <- p + geom_point(position = position_jitter(w = 0.05, h = 0), alpha = 0.2)
# diamond at mean for each group
p <- p + stat_summary(fun.y = mean, geom = "point", shape = 18, size = 4,
                      aes(colour=survived), alpha = 0.8)
# confidence limits based on normal distribution
p <- p + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                      width = .2, aes(colour=survived), alpha = 0.8)
p <- p + labs(title = "Sparrow humerus length, by survival")
p <- p + ylab("length (inch/1000)")
print(p)
```



Normality test.

```
library(nortest)
ad.test(subset(sparrows, survived==0)$humerus)
```

```
##
##  Anderson-Darling normality test
##
## data:  subset(sparrows, survived == 0)$humerus
## A = 0.5307, p-value = 0.1573
```

```
ad.test(subset(sparrows, survived==1)$humerus)
```

```
##
##  Anderson-Darling normality test
##
## data:  subset(sparrows, survived == 1)$humerus
## A = 0.307, p-value = 0.5457
```

Numerical summaries.

```
# summary of each year
by(sparrows$humerus, sparrows$survived, summary)
```

```
## sparrows$survived: 0
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     659     718     734     728     743     765
## ------------------------------------------------
## sparrows$survived: 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     687     728     736     738     752     780
```

```
# IQR and sd of each year
by(sparrows$humerus, sparrows$survived, function(X) { c(IQR(X), sd(X)) })
```

```
## sparrows$survived: 0
## [1] 25.00 23.54
## ------------------------------------------------
## sparrows$survived: 1
## [1] 23.50 19.84
```

Test for equal variance.

```
## Test equal variance
# assumes populations are normal
bartlett.test(humerus ~ survived, data = sparrows)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  humerus by survived
## Bartlett's K-squared = 0.803, df = 1, p-value = 0.3702
```

```r
# does not assume normality, requires car package
library(car)
leveneTest(humerus ~ survived, data = sparrows)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1    0.32   0.58
##       57
```

```r
# nonparametric test
fligner.test(humerus ~ survived, data = sparrows)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  humerus by survived
## Fligner-Killeen:med chi-squared = 0.0878, df = 1, p-value =
## 0.767
```

(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 nonparametric procedure.

Solution: The p-value=0.081 > 0.05, therefore we fail to reject H_0 concluding there is insufficient evidence to conclude a mean difference in humerus length between surviving and nonsurviving sparrows.
Because this is a case where the t-test is appropriate, it will also be a more powerful test than the WMW test. The WMW test gives a p-value=0.1718 > 0.05, resulting in the same conclusion for the median.

```r
t.test(humerus ~ survived, data = sparrows, var.equal=TRUE)
```

```
##
##  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:
##  -21.446   1.279
## sample estimates:
## mean in group 0 mean in group 1
##           727.9           738.0
```

```r
wilcox.test(humerus ~ survived, data = sparrows, conf.int = TRUE)
```

```
## Warning:  cannot compute exact p-value with ties
## Warning:  cannot compute exact confidence intervals with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  humerus by survived
## W = 331, p-value = 0.1718
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -19   3
## sample estimates:
## difference in location
##                     -7
```

(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.

Solution: The 95% t-CI for the difference in means is $(-21.4461, 1.2794)$, and the 95% WMW-CI for the difference in medians is $(-19.00, 3.00)$.
We are 95% confident that the true population difference in means (medians) is in the interval above.

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

Solution: As discussed in part (a), the normality assumption and equal-variance assumption is not violated for the t-test and CIs. Similarly, the symmetry required for the WMW is not violated.

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

Solution: We are interested to know whether the average humerus length of sparrows differs between those that survived and did not survive a severe winter storm. Normality assumptions are not in question for samples from both bird populations. The variances from both samples are close enough to assume they are equal and use a pooled-variance estimate. Therefore a two-sample t-test was conducted, but there was insufficient evidence to conclude that the mean humerus length differed between the two populations.

(50 pts) **4. Protoporphin levels among alcoholics:** Protoporphrin levels were determined for three groups of people — a control group of normal workers, a group of alcoholics with sideroblasts in their bone marrow, and a group of alcoholics without sideroblasts. The given data appeared in the paper "Erythocyte Coproporphyrin and Protoporphrin in Ethanol-Induced Sideroblastic Erythoporiesis" (Blood, 1974, p. 291–295).

Analyze the data, assuming you are interested in comparing the typical protoporphrin level across groups. Quantify any differences you find. Make sure to clearly define all population parameters, and assess the assumptions underlying your chosen method of analysis. Have a well-organized write-up.

Normal	Alc_w_sb	Alc_wo_sb
22	78	37
27	172	28
47	286	38
30	82	45
38	453	47
78	513	29
28	174	34
58	915	20
72	84	68
56	153	12
30	780	37
39	NA	8
53	NA	76
50	NA	148
36	NA	11

(At `statacumen.com/teach/ADA1/ADA1_HW_06_F14-4.txt`, it's up to you to read this dataset into R.)

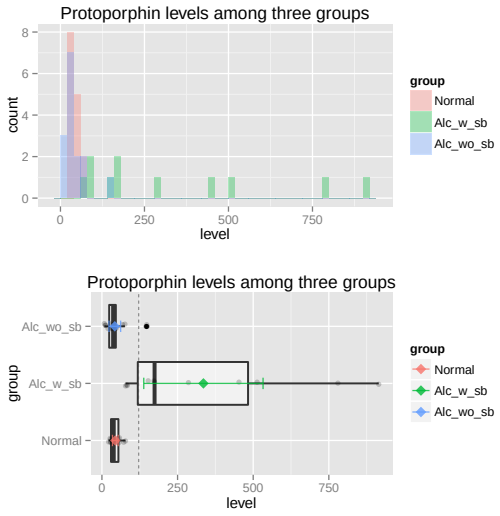
Solution:

```
# read the table in as a data.frame
proto <- read.table("http://statacumen.com/teach/ADA1/ADA1_HW_06_F14-4.txt", header=TRUE)
# reshape in long format, removing NAs
library(reshape2)
proto.long <- melt(proto,
              variable.name = "group",
              value.name = "level",
              na.rm = TRUE
             )

## Using  as id variables
proto.long$loglevel <- log(proto.long$level)
```

The three distributions are all right-skewed (so ANOVA is not appropriate), and the shapes are different (so KW ANOVA is not appropriate). Also, the boxplots show that the variance and means appear related. We'll try a log transformation of the data to improve these issues in the data to use a method we've learned about for analysis.

```
p2 <- ggplot(proto.long, aes(x = level, fill=group))
p2 <- p2 + geom_histogram(binwidth = 20, alpha = 1/3, position="identity")
p2 <- p2 + labs(title = "Protoporphin levels among three groups")
p2 <- p2 + xlab("level")
#print(p2)
```

```
# Plot the data using ggplot
library(ggplot2)
p <- ggplot(proto.long, aes(x = group, y = level))
# plot a reference line for the global mean (assuming no groups)
p <- p + geom_hline(aes(yintercept = mean(level)),
                    colour = "black", linetype = "dashed", size = 0.3, alpha = 0.5)
# boxplot, size=.75 to stand out behind CI
p <- p + geom_boxplot(size = 0.75, alpha = 0.5)
# points for observed data
p <- p + geom_point(position = position_jitter(w = 0.05, h = 0), alpha = 0.2)
# diamond at mean for each group
p <- p + stat_summary(fun.y = mean, geom = "point", shape = 18, size = 4,
                      aes(colour=group), alpha = 0.8)
# confidence limits based on normal distribution
p <- p + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                      width = .2, aes(colour=group), alpha = 0.8)
p <- p + coord_flip()
p <- p + labs(title = "Protoporphin levels among three groups")
p <- p + ylab("level")
#print(p)

library(gridExtra)
grid.arrange(p2, p, ncol=1)
```



Numerical summaries.

```
# summary of each year
by(proto.long$level, proto.long$group, summary)
```

```
## proto.long$group: Normal
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    22.0    30.0    39.0    44.3    54.5    78.0
## ------------------------------------------------------------
## proto.long$group: Alc_w_sb
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      78     118     174     335     483     915
## ------------------------------------------------------------
## proto.long$group: Alc_wo_sb
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.0    24.0    37.0    42.5    46.0   148.0
```

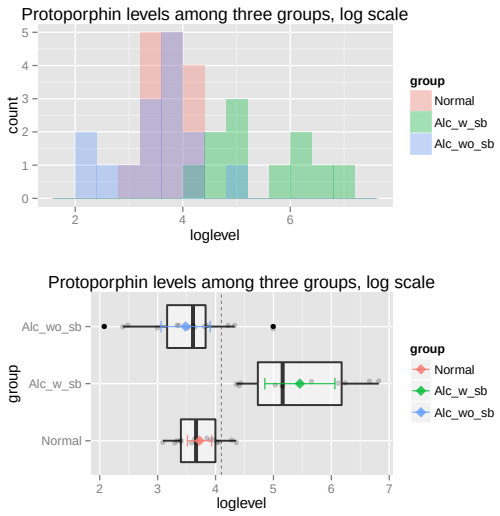
```
# IQR and sd of each year
by(proto.long$level, proto.long$group, function(X) { c(IQR(X), sd(X)) })
```

```
## proto.long$group: Normal
## [1] 24.5 16.8
## ------------------------------------------------------------
```

```
## proto.long$group: Alc_w_sb
## [1] 364.5 293.4
## ------------------------------------------------------
## proto.long$group: Alc_wo_sb
## [1] 22.00 34.94
```

The log transformation has made the data consistent with normal distributions. The standard deviations for the normal group is about half of the other two groups, but the IQR is larger for the alcoholics without sideroblasts group than the other two groups (boxplot and numerical summaries of StDev and IQR).

```
p2 <- ggplot(proto.long, aes(x = loglevel, fill=group))
p2 <- p2 + geom_histogram(binwidth = .4, alpha = 1/3, position="identity")
p2 <- p2 + labs(title = "Protoporphin levels among three groups, log scale")
p2 <- p2 + xlab("loglevel")
#print(p2)

# Plot the data using ggplot
library(ggplot2)
p <- ggplot(proto.long, aes(x = group, y = loglevel))
# plot a reference line for the global mean (assuming no groups)
p <- p + geom_hline(aes(yintercept = mean(loglevel)),
                    colour = "black", linetype = "dashed", size = 0.3, alpha = 0.5)
# boxplot, size=.75 to stand out behind CI
p <- p + geom_boxplot(size = 0.75, alpha = 0.5)
# points for observed data
p <- p + geom_point(position = position_jitter(w = 0.05, h = 0), alpha = 0.2)
# diamond at mean for each group
p <- p + stat_summary(fun.y = mean, geom = "point", shape = 18, size = 4,
                      aes(colour=group), alpha = 0.8)
# confidence limits based on normal distribution
p <- p + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                      width = .2, aes(colour=group), alpha = 0.8)
p <- p + coord_flip()
p <- p + labs(title = "Protoporphin levels among three groups, log scale")
p <- p + ylab("loglevel")
#print(p)

library(gridExtra)
grid.arrange(p2, p, ncol=1)
```



Numerical summaries.

```
# summary of each year
by(proto.long$loglevel, proto.long$group, summary)
```

```
## proto.long$group: Normal
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    3.09    3.40    3.66    3.72    4.00    4.36
## ------------------------------------------------------
## proto.long$group: Alc_w_sb
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.36    4.73    5.16    5.46    6.18    6.82
## ------------------------------------------------------
## proto.long$group: Alc_wo_sb
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.08    3.16    3.61    3.48    3.83    5.00
```

```
# IQR and sd of each year
by(proto.long$loglevel, proto.long$group, function(X) { c(IQR(X), sd(X), length(X)) })
```

```
## proto.long$group: Normal
## [1]  0.5966  0.3778 15.0000
## ------------------------------------------------------
## proto.long$group: Alc_w_sb
## [1]  1.4475  0.9012 11.0000
## ------------------------------------------------------
## proto.long$group: Alc_wo_sb
## [1]  0.6644  0.7685 15.0000
```

In spite of the differences in spread (Bartlett's test for equal variance p-value=0.012), we'll perform both an ANOVA and KW ANOVA. The ANOVA is appropriate to compare log(proto) means, and we should expect the KW ANOVA to provide a consistent result for log(proto) medians.
The result of the ANOVA below has a p-value=0.000 < 0.05, therefore we reject the null hypothesis in favor of the alternative that at least two log(proto) means are different. The result of the Bonferroni corrected pairwise two-sample t-test comparisons shows that the alcoholics with sideroblasts group is different from the other two groups, which are not different from each other.
Normality test.

```
library(nortest)
ad.test(subset(proto.long, group==unique(proto.long$group)[1])$loglevel)
```

```
##
##  Anderson-Darling normality test
##
## data:  subset(proto.long, group == unique(proto.long$group)[1])$loglevel
## A = 0.2204, p-value = 0.7962
```

```
ad.test(subset(proto.long, group==unique(proto.long$group)[2])$loglevel)
```

```
##
##  Anderson-Darling normality test
##
## data:  subset(proto.long, group == unique(proto.long$group)[2])$loglevel
## A = 0.3552, p-value = 0.3907
```

```
ad.test(subset(proto.long, group==unique(proto.long$group)[3])$loglevel)
```

```
##
##  Anderson-Darling normality test
##
## data:  subset(proto.long, group == unique(proto.long$group)[3])$loglevel
## A = 0.3141, p-value = 0.5101
```

Test for equal variance.

```
## Test equal variance
# assumes populations are normal
bartlett.test(loglevel ~ group, data = proto.long)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  loglevel by group
## Bartlett's K-squared = 8.806, df = 2, p-value = 0.01224
```

ANOVA and Bonferroni comparisons.

```
# ANOVA of rank, for illustration that this is similar to what KW is doing
fit.ll <- aov(loglevel ~ group, data = proto.long)
summary(fit.ll)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## group        2   28.1   14.04      29 2.2e-08 ***
## Residuals   38   18.4    0.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit.ll
```

```
## Call:
##    aov(formula = loglevel ~ group, data = proto.long)
##
## Terms:
##                 group Residuals
## Sum of Squares  28.07     18.39
## Deg. of Freedom     2        38
##
## Residual standard error: 0.6956
## Estimated effects may be unbalanced
```

```
# Bonferroni 95% Individual p-values
# All Pairwise Comparisons among Levels of glabella
pairwise.t.test(proto.long$loglevel, proto.long$group,
                pool.sd = TRUE, p.adjust.method = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  proto.long$loglevel and proto.long$group
##
##           Normal  Alc_w_sb
## Alc_w_sb  7.2e-07 -
## Alc_wo_sb 1       4.7e-08
##
## P value adjustment method: bonferroni
```

The KW ANOVA agrees with the ANOVA. The result of the KW ANOVA below has a p-value=0.000 $<$ 0.05, therefore we reject the null hypothesis in favor of the alternative that at least two log(proto) medians are different. The result of the Bonferroni corrected pairwise two-sample WMW comparisons shows that the alcoholics with sideroblasts group is different from the other two groups, which are not different from each other.
KW ANOVA and WMW Bonferroni comparisons.

```
# ANOVA of rank, for illustration that this is similar to what KW is doing
# KW ANOVA
fit.llk <- kruskal.test(loglevel ~ group, data = proto.long)
fit.llk
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  loglevel by group
## Kruskal-Wallis chi-squared = 23.12, df = 2, p-value =
## 9.549e-06
```

```
wilcox.test(subset(proto.long, group==unique(proto.long$group)[1])$loglevel
          , subset(proto.long, group==unique(proto.long$group)[2])$loglevel
          , conf.int = TRUE, conf.level = 0.9833)
```

```
## Warning:  cannot compute exact p-value with ties
## Warning:  cannot compute exact confidence intervals with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  subset(proto.long, group == unique(proto.long$group)[1])$loglevel and subset(proto.long, group == unique(proto.lo
## W = 0.5, p-value = 2.324e-05
## alternative hypothesis: true location shift is not equal to 0
```

```
## 98.33 percent confidence interval:
##  -2.6339 -0.8824
## sample estimates:
## difference in location
##                 -1.698
```

```
wilcox.test(subset(proto.long, group==unique(proto.long$group)[1])$loglevel
          , subset(proto.long, group==unique(proto.long$group)[3])$loglevel
          , conf.int = TRUE, conf.level = 0.9833)
```

```
## Warning:  cannot compute exact p-value with ties
## Warning:  cannot compute exact confidence intervals with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  subset(proto.long, group == unique(proto.long$group)[1])$loglevel and subset(proto.long, group == unique(proto.lo
## W = 137.5, p-value = 0.3093
## alternative hypothesis: true location shift is not equal to 0
## 98.33 percent confidence interval:
##  -0.3054  0.8110
## sample estimates:
## difference in location
##                 0.2126
```

```
wilcox.test(subset(proto.long, group==unique(proto.long$group)[2])$loglevel
          , subset(proto.long, group==unique(proto.long$group)[3])$loglevel
          , conf.int = TRUE, conf.level = 0.9833)
```

```
## Warning:  cannot compute exact p-value with ties
## Warning:  cannot compute exact confidence intervals with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  subset(proto.long, group == unique(proto.long$group)[2])$loglevel and subset(proto.long, group == unique(proto.lo
## W = 162, p-value = 4.117e-05
## alternative hypothesis: true location shift is not equal to 0
## 98.33 percent confidence interval:
##  0.928 2.951
## sample estimates:
## difference in location
##                  1.922
```

The permutation test gives the same result, without any distributional assumptions.

```
# permutation test version
library(lmPerm)
aovp.summary <- aovp(loglevel ~ group, data = proto.long)
```

```
## [1] "Settings:  unique SS "
```

```
aovp.summary
```

```
## Call:
##    aovp(formula = loglevel ~ group, data = proto.long)
##
## Terms:
##                 group Residuals
## Sum of Squares  28.07     18.39
## Deg. of Freedom     2        38
##
## Residual standard error: 0.6956
## Estimated effects may be unbalanced
```

```
summary(aovp.summary)
```

```
## Component 1 :
##             Df R Sum Sq R Mean Sq Iter Pr(Prob)
## group        2     28.1     14.04 5000   <2e-16 ***
## Residuals   38     18.4      0.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# permutation test version
library(lmPerm)

# 1 vs 2
proto.long2 <- subset(proto.long, group != unique(proto.long$group)[3])
lmp.summary <- lmp(loglevel ~ group, data = proto.long2)
```

```
## [1] "Settings:  unique SS "
```

```r
summary(lmp.summary)
```

```
##
## Call:
## lmp(formula = loglevel ~ group, data = proto.long2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0998 -0.3744 -0.0733  0.3278  1.3624
##
## Coefficients:
##        Estimate Iter Pr(Prob)
## group1   -0.866 5000   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.649 on 24 degrees of freedom
## Multiple R-Squared: 0.653,Adjusted R-squared: 0.639
## F-statistic: 45.2 on 1 and 24 DF,  p-value: 5.94e-07
```

```r
# 1 vs 3
proto.long2 <- subset(proto.long, group != unique(proto.long$group)[2])
lmp.summary <- lmp(loglevel ~ group, data = proto.long2)
```

```
## [1] "Settings:  unique SS "
```

```r
summary(lmp.summary)
```

```
##
## Call:
## lmp(formula = loglevel ~ group, data = proto.long2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4037 -0.3227  0.0847  0.3180  1.5140
##
## Coefficients:
##        Estimate Iter Pr(Prob)
## group1     0.12  200     0.34
##
## Residual standard error: 0.606 on 28 degrees of freedom
## Multiple R-Squared: 0.0406,Adjusted R-squared: 0.00633
## F-statistic: 1.18 on 1 and 28 DF,  p-value: 0.286
```

```r
# 2 vs 3
proto.long2 <- subset(proto.long, group != unique(proto.long$group)[1])
lmp.summary <- lmp(loglevel ~ group, data = proto.long2)
```

```
## [1] "Settings:  unique SS "
```

```r
summary(lmp.summary)
```

```
##
## Call:
## lmp(formula = loglevel ~ group, data = proto.long2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4037 -0.4721  0.0855  0.5863  1.5140
##
## Coefficients:
##        Estimate Iter Pr(Prob)
## group1    0.987 5000   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.826 on 24 degrees of freedom
## Multiple R-Squared: 0.601,Adjusted R-squared: 0.585
## F-statistic: 36.2 on 1 and 24 DF,  p-value: 3.28e-06
```