I. Introduction

THIS report investigates the concerns raised by the client regarding the services provided by Respond2Emergencies. In particular, the client requests analysis on the following three issues:

1) The effectiveness *and* consistency of the off-site training service, *We Come to You*, as compared to the standard on-site training.
2) The effectiveness of the training between the state and the 5 surrounding states.
3) The claim made by Respond2Emergencies on the *We Come to You* service that an increase in time spent by off-site instructors increases client preparedness, contradictory to feedback from training sessions.

The methods applied in this analysis are divided into two distinct statistical tests: parametric and non-parametric. In general, nonparametric methods require minimal assumptions about the form of the distributions of the populations of interest. Parametric methods require that the form of the population distribution be completely specified except for a finite number of parameters.

The additional concern of the client regarding the application of the t-test over non-parametric tests is answered throughout the report; however, as the client expresses urgency on the matter, the basic assumptions of the t-test are enumerated as followed [1][2]:

1) The observations are independent or drawn randomly from an infinite sample.
2) The population is normally distributed.
3) The variances of the scores from each group are equivalent.
4) There are no outliers.

If these assumptions are not satisfied for the provided data then the t-test is not applicable. Analysis is performed in *Sec. III*.

II. Summary of Data

General analysis is performed in order to uncover general trends in the data. In doing so, proper hypotheses are prescribed to the tests.

A. Raw Data

Data regarding issue (1) is replicated in full as well as the first few observations for both (2) and (3). [1]

(1) On and Off Site Training: A total of 14 observations are recorded by the training teams.

[1](1), (2), and (3) correspond to the data sets outlined in I.

Location	1	2	3	4
On-site	85.34	84.12	86.53	85.52
Off-site	81.48	81.26	79.71	81.46
Location	5	6	7	8
On-site	84.61	86.94	84.05	86.29
Off-site	79.20	81.65	-	-

TABLE I
2014 Emergency Preparedness Test Scores

(2) The State and Surrounding States: Let '0' denote 'Our State' and the subsequent numbers denote the surrounding states. The first four entries for each state have been listed. Furthermore, the entries have been simplified to two decimal places.

0	1	2	3	4	5
85.24	83.51	84.87	83.59	76.13	83.41
86.22	86.03	84.83	83.67	74.69	83.51
83.98	84.92	85.06	84.79	76.09	83.90
84.70	86.91	86.08	83.48	74.61	83.92
...	...	...	...	...	...

TABLE II
Trans-state Emergency preparedness test scores

(3) Training Duration: A total of 10 observations are recorded per time interval.

<30min	≈60min	≈90min	≈120min	≈180min
83.37	81.21	81.50	82.58	83.71
83.98	81.41	80.23	80.14	80.37
82.98	83.41	82.62	80.79	83.08
80.99	81.87	83.52	80.86	82.86
...	...	...	...	...

TABLE III
Instructor teaching duration and test scores

B. Quantile-Quantile Plots

Quantile-Quantile (Q-Q) plots are utilized gather evidence pertaining to the normality of the data. If the data points sufficiently maps to the indicated line, then there is evidence to assume that the data is normally distributed. (*Fig. 1-3*)

C. Histograms

Histograms are also utilized to gather evidence pertaining to the normality of the data. If both the Q-Q plot and histogram of the data suggest normality, then normality can be assumed and allows for the use of proper tests. (*Fig. 4-6*)

D. Boxplots

Boxplots are utilized to observe, but not determine, a difference in variances and means. The length of the 'whiskers' determines the variance. (*Fig. 7-9*)

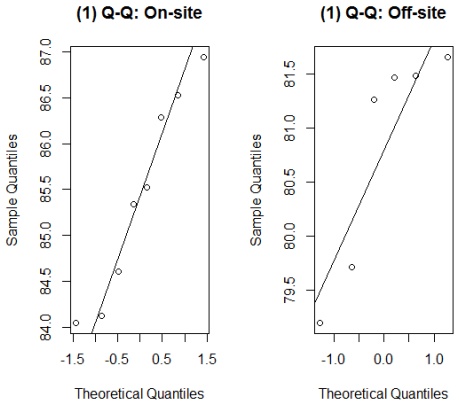

Fig. 1. Q-Q plots of the on-site and off-site training scores (1).

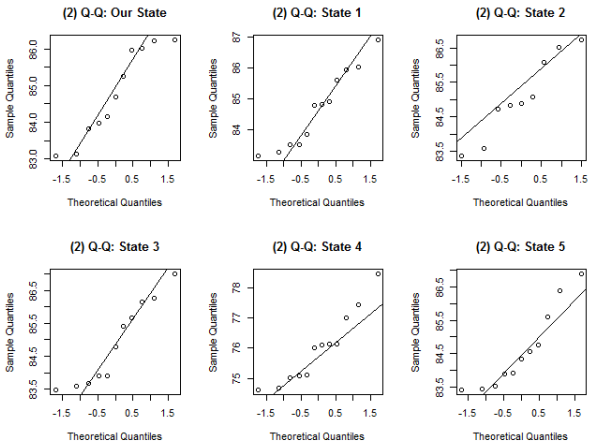

Fig. 2. Q-Q plots of the scores from the state and surrounding states (2).

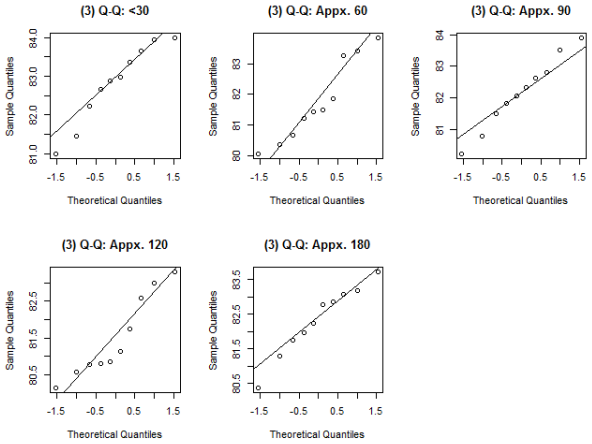

Fig. 3. Q-Q plots of the scores by varying training duration (3).

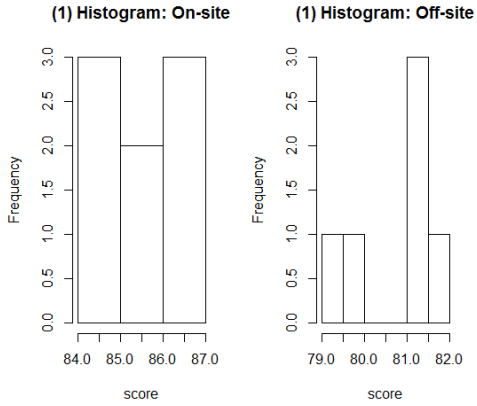

Fig. 4. Histogram of (1).

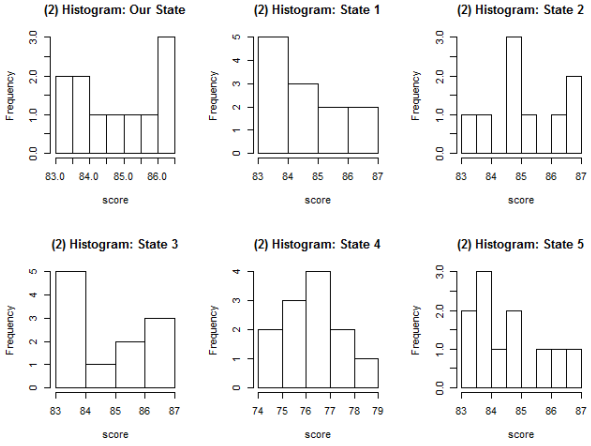

Fig. 5. Histogram of (2).

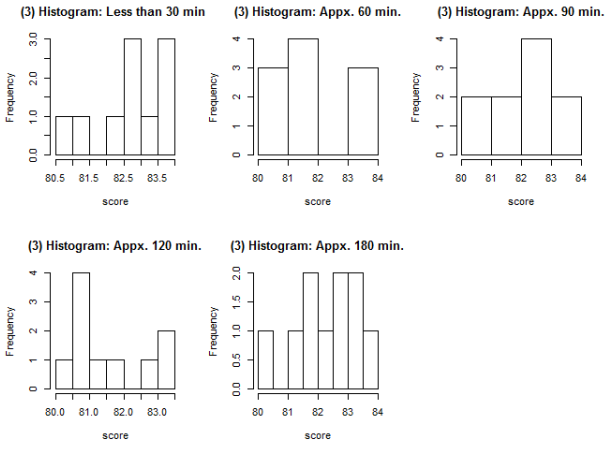

Fig. 6. Histogram of (3).

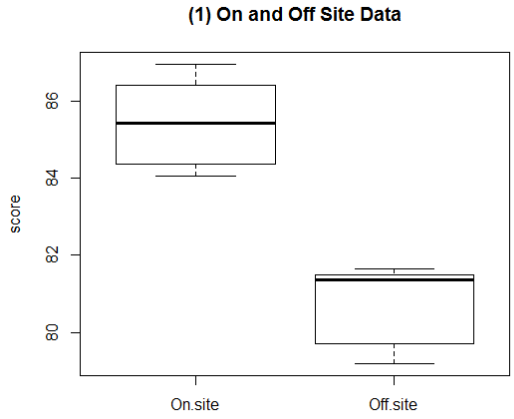

Fig. 7. Boxplot of (1).

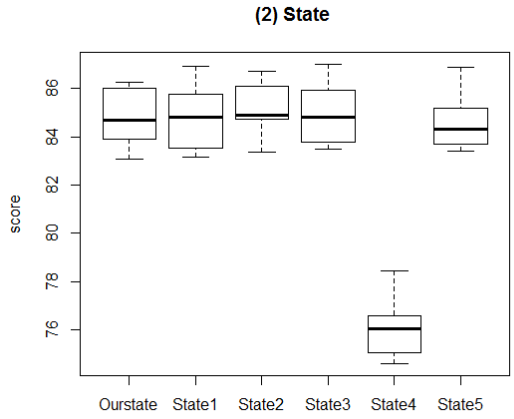

Fig. 8. Boxplot of (2).

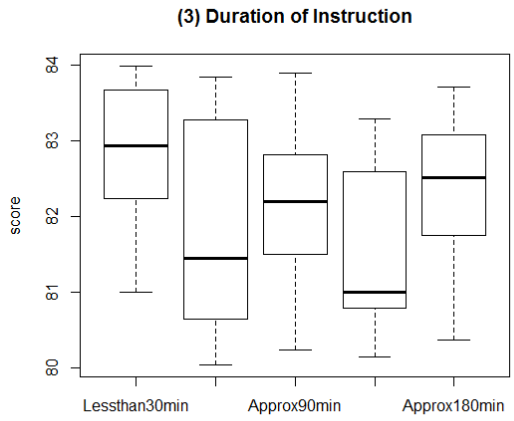

Fig. 9. Boxplot of (3). The intervals on the x-axis are: $< 30, \approx 60, \approx 90, \approx 120,$ and ≈ 180 minutes.

III. Methodology

As it was briefly discussed in *Sec. I*, certain assumptions are needed to be met to be able to apply the t-test to a set of data. One of which is the normality assumption. For (1), *fig 1* and *fig. 4* indicate no form of normality. This is compounded by the fact that the sample size is too small to generate accurate plots of the distribution. The same logic applies to (2) with *fig. 2* and *fig. 5*. Although it may be argued that the plot "(2) Historgram: State 4" of *fig. 5* indicates a relatively normally distributed data set, the corresponding Q-Q plot, "(2) Q-Q State 4" of *fig. 3* contradicts the normality assumption. The same logic is applied to (3). None of the data sets satisfy the normality assumption; therefore, no parametric tests that assume normality can be applied to these sets.

Iterated from *Sec. I*, nonparametric methods require minimal assumptions about the form of the distributions of the populations of interest. Parametric methods require that the form of the population distribution be completely specified except for a finite number of parameters.

The following subsections detail the procedures and assumptions made on (1), (2), and (3).

A. (1) On and Off Site Training

Perform a permutation test on deviances for an unknown location parameter. A permutation test is a parametric test which can be applied without assuming normality. The assumptions for the permutation test are:

1) The observations are independent.
2) The observations are exchangeable.

The first assumption is satisfied since the two groups did not interact with each other and the scores are independent. The second assumption is satisfied since the individual observations are identically distributed under the null hypothesis.

The test for an unknown location parameter is applied instead of a known location parameter because the population means of the two groups are unknown.

This test will either accept or reject if there is a significant difference in the variances between the two groups. This test alone will answer the client's question if there is a significant difference in the consistency between the training types; however, there is no indication on either method's effectiveness.

To determine which method is more effective, either the Wilcoxon Rank-Sum test or permutation test [5] is applied. The assumptions for either of these tests are:

1) The observations are independent.
2) The observations are exchangeable.
3) The variances between the groups are equivalent.

The first two assumptions are already met, and the third assumption may be satisfied by the previous test. If the previous test fails to show equality in variances, the Wilcoxon Rank-Sum and permutation test cannot be used and there is no statistically significant conclusion; however, this is not likely since by *fig. 7* the variances between the group appear to be equivalent. If the previous test does show equality of variances between the groups, then either of the two tests on medians is performed on the data to conclude if there is a difference in the medians of the scores. In particular, *fig. 7* suggests that the scores for the off-site training would be less than the on-site training. So the alternative hypothesis is formulated such that the scores for the off-site training is less than the scores of the on-site training. Furthermore, the power of each test is analyzed in another section. [3]

The tests are then applied to conclude if there exists a difference in medians rather than means. A test on medians has the advantage that medians are more protected when there are no assumptions about the distribution of the data. It takes into consideration of outliers better than the tests on means. In addition to this is that the difference in medians is better suited for non-normal distributions or heavy-tailed distributions, which can be inferred from *fig. 4-6*.

B. (2) The State and Surrounding States

The procedures for (2) are similar to (1), except that multiple groups are now taken into consideration. Parametric tests must be applied because the normality assumption is not met. In particular, perform a permutation F-test by randomly sampling different permutations. A sampling method is used instead of an exact method because of large number of possible permutations of the data. Furthermore, this test is used over a one-way analysis of variance F-test since the experimental units are not drawn from a larger population, rather from the the units that are available. Since the observations were taken from the available states and counties with respect to proximity, the permutation F-test is used. The assumptions for a permutation F-test are:

1) The observations are independent.
2) The observations are exchangeable.
3) The variances between the groups are equivalent.

The third assumption is satisfied by the client's claim that training for each respective state is consistent over the years; however, to be absolutely certain pairwise permutation tests on deviances with unknown means much like (1) are also performed. Pairwise tests are performed for each of the 15 possible permutations since the transitive property does not apply. If even one pair of groups have different variances then the third assumption is not satisfied. By observation from *fig. 8*, it seems that the variances are equivalent but observations alone does not provide concrete evidence to state a statistically significant conclusion.

The permutation F-test will determine if there exists any difference between the distributions of the groups. If there does not exist a difference, then accept the null hypothesis that the training provided by the state and surrounding states are equally effective with one another. If there does exist a difference, then individual two-sample tests for medians as in (1) are performed between the groups to determine which groups are different. Like in (1), there are two viable tests. The permutation F-test is analogous to the permutation test (for two samples) and the Kruskal-Wallis test is analogous to the Wilcoxon Rank-Sum test (for two samples). Both are performed as each have their own strengths and weaknesses.

C. (3) Training Duration

The procedure for this set of data is identical to (2). However, there are a few notable differences. It is more difficult to discern if the variances between the groups are equal from *fig. 9* as compared to *fig. 8.*; however, tests for equality of variances are carried out between each pairs of the groups. If there is a single pair of data which have different variances, then the permutation F-test cannot be carried out and there is no conclusion. If the variances are the same, perform the individual two-sample tests for medians to determine which, if any, of the medians are different from one another.

IV. Results

A. (1) On and Off Site Training

Here are the results of the tests outlined in *Sec. III.A*:

Test Type	Statistic	Value
Permutation Test on Deviances	p-value	0.7133
Permutation Test on Medians	p-value	0.0031
	Power	-
Wilcoxon Rank-Sum Test	W (Mann-W.)	48
	p-value	0.0007
	Power	-

TABLE IV
Results of (1)

B. (2) The State and Surrounding States

Here are the results of the tests outlined in *Sec. III.B*:

Test Type	Statistic	Value
Permutation Test on Deviances	min(p-value)	0.1830
Permutation F-Test	p-value	$\approx$0
	Power	-
Kruskal-Wallis Test	χ^2	29.858
	p-value	$\approx$0
	Power	-
Permutation Test on Medians	p-value	*Table VI*
	Bonf. Adj.	

TABLE V
Results of (2)

The rows and columns for the following table indicate the groups which were tested against each other.

	0	1	2	3	4	5
0	-	13.35	11.55	15.00	0.00**	12.60
1	-	-	11.85	15.00	0.00**	9.00
2	-	-	-	11.85	0.00**	3.60
3	-	-	-	-	0.00**	11.10
4	-	-	-	-	-	0.00**
5	-	-	-	-	-	-

TABLE VI
p-values Bonferroni Adjusted

** Statistically significant.

C. (3) Training Duration

Here are the results of the tests outlined in *Sec. III.C*:

Test Type	Statistic	Value
Permutation Test on Deviances	min(p-value)	0.0895
Permutation F-Test	p-value	0.1022
	Power	-
Kruskal-Wallis Test	χ^2	7.5586
	p-value	0.1092
	Power	-

TABLE VII
Results of (3)

V. Discussion

The tests performed on the data sets were **all conducted at a significance level of 5%**, and all tests produced conclusive results; however, it is inconclusive as to which method is the most powerful without making assumptions about the form of the underlying distributions.

A. (1) On and Off Site Training

The permutation test on deviances for an unknown location parameter is performed on (1). The calculated p-value for the null hypothesis that all the variances of the groups are equivalent is 0.7133. This indicates that the null hypothesis is not rejected and that the variance in scores between the on-site and off-site groups are the same. Furthermore, this also implies that the training between the two group are consistent with one another.

The permutation test on medians is performed on (1). The calculated p-value for the null hypothesis that the medians are the same is 0.0031. This indicates that the null hypothesis is rejected in favor of the alternative hypothesis that the off-site scores are lower than the on-site scores.

The Wilcoxon Rank-Sum test is also performed on (1). The calculated p-value for the same null hypothesis as the previous test is 0.0070 (with a Mann-Whitney test statistic of 48). This indicates that the null hypothesis is rejected in favor of the alternative hypothesis that the off-site scores are lower than the on-site scores.

Both tests conclude that the on-site training produced statistically higher scores than the off-site training; however, the power of the tests may be different. According to Higgins, the permutation test is more powerful than the Wilcoxon Rank-Sum test for heavy-tailed distributions than for light-tailed distributions. Based on the provided data in *fig. 4* the distribution is more heavy-tailed than light tailed. By this logic, the permutation test is preferred. If more data is collected to suggest that the underlying distribution is light-tailed then Wilcoxon Rank-Sum test is preferred.

B. (2) The State and Surrounding States

The permutation test on deviances for an unknown location parameter is performed on (2). The calculated p-value for the null hypothesis that all the variances of the groups are equivalent is 0.1830. This indicates that the null hypothesis is not rejected and that the variance in scores between the states are the same. Furthermore, this also implies that the training between the six groups are consistent.

Then the permutation F-test is performed. The calculated p-value of the null hypothesis that the distributions are identical is approximately 0. The Kruskal-Wallis test is also performed and also results in a similar p-value of approximately 0 (with a chi-squared statistic of 29.858). Either of these tests indicate that the null hypothesis is rejected in favor of the alternative hypothesis that there is a difference in distributions between at least one of the groups.

There are a total of 15 pairwise groups and the permutation test for medians is performed between them. Their values are indicated in *table VI* accounting for the Bonferroni adjustment. The table indicates that the fourth state has a different median from any of the other states. Since the Bonferroni adjustment is extremely conservative in calculating a p-value an additional, less conservative, test is performed on the data between the second and fifth states. This particular pair was chosen because, excluding the fourth state, the pairwise permutation test indicates the lowest p-value out of all possible pairs. If a less conservative test does not indicate a significant difference for this pair, then all other pairs will not exhibit a significant difference by the same test. By performing Tukey's HSD for unequal sample sizes (Tukey-Kramer), it is determined that even with a less conservative adjustment that this pairwise group do not have different medians ($p = 0.39$).

As in (1), the arguments for power are analogous between 2-groups and more than 2-groups. By *fig. 5*, the data is heavy-tailed therefore the permutation F-test is more powerful than the Kruskal-Wallis test [3]. If additional evidence is provided that suggests that the underlying distribution is light-tailed then the Kruskal-Wallis test is preferred.

C. (3) Training Duration

The permutation test on deviances for an unknown location parameter is performed on (3). The calculated p-value for the null hypothesis that all the variances of the groups are equivalent is 0.0895. This indicates that the null hypothesis is not rejected and that the variance in scores between the states are the same. Furthermore, this also implies that the training between the two group are consistent.

Then the permutation F-test is performed. The calculated p-value of the null hypothesis that the distributions are identical is 0.1022. The Kruskal-Wallis test is also performed and also results in a similar p-value of 0.1092 (with a chi-squared statistic of 7.5586). Either of these tests accept the null hypothesis that there is no difference in medians between the groups. Unlike (2), since there is no difference in medians there is no reason to perform permutation tests on medians for the pairwise groups.

By *fig. 6*, the data is heavy tailed and the permutation F-test is more powerful than the Kruskal-Wallis test. If additional evidence is provided that suggests that the underlying distribution is light-tailed then the Kruskal-Wallis test is preferred.

The data is partitioned into discrete steps of 30 minutes. One major concern regarding (3) is that the interval for $\approx$ 150 minutes of training time is not present.

D. Power

The power of a test indicates the probability of not committing a type II error, that is, the probability of not accepting a significant difference in medians that is false. The higher the power, the more reliable the test is for applications. The greatest issue facing power calculations is that an assumption about the form of the underlying distribution has to be made. For light-tailed tests, the Wilcoxon Rank-Sum and Kruskal-Wallis tests are preferred over the permutation and permutation F-tests, respectively. With an increased sample size a stronger claim can be made for the underlying distribution and a more powerful test can be applied; however, in light of this both tests for (1), (2), and (3) reach the same conclusion. The only claim provided by the client on the distribution of the data is that it is not Cauchy distributed, a heavy-tailed distribution.

VI. Conclusion

The hypothesized tests were valid for the data sets and concrete conclusions are drawn. For (1): the training methods provided by Respond2Emergencies are consistent between the on-site and off-sites; however, those who were trained on-site performed better than those who were trained off-site. For (2): the training methods provided by the emergency response services between the state and the five surrounding states are all equally consistent. All of the states performed equally well, with the exception of the fourth state which under-performed. For (3): the claim made by Respond2Emergencies on the *We Come to You* service that an increase in time spent by off-site instructors increases preparedness is false. There was no significant difference in scores regardless of the time spent by the instructors. Furthermore, the tests may be optimized by increasing the sample size through proper recording of scores at the county and state levels. In doing so, the tests become more powerful and the probability of committing an error decreases.

Code: Written in R

```
# Set the working directory to the folder with all the data.
setwd("C:/Users/Tommy/Desktop/3480")

# Import/write all the data.
data.1 = data.frame(
  "On-site" = c(85.34, 84.12, 86.53, 85.52, 84.61, 86.94, 84.05, 86.29),
  "Off-site"= c(81.48, 81.26, 79.71, 81.46, 79.20, 81.65, NA   , NA)
)
data.2 = read.table('state.csv', header = T, sep = ",")
data.3 = read.table('internal.txt', header = T)

# Q-Q plots
par(mfrow = c(1,2))
qqnorm(data.1$On.site, main = "(1) Q-Q: On-site"); qqline(data.1$On.site)
qqnorm(data.1$Off.site, main = "(1) Q-Q: Off-site"); qqline(data.1$Off.site)

par(mfrow = c(2,3))
qqnorm(data.2$Ourstate, main = "(2) Q-Q: Our State"); qqline(data.2$Ourstate)
qqnorm(data.2$State1, main = "(2) Q-Q: State 1"); qqline(data.2$State1)
qqnorm(data.2$State2, main = "(2) Q-Q: State 2"); qqline(data.2$State2)
qqnorm(data.2$State3, main = "(2) Q-Q: State 3"); qqline(data.2$State3)
qqnorm(data.2$State4, main = "(2) Q-Q: State 4"); qqline(data.2$State4)
qqnorm(data.2$State5, main = "(2) Q-Q: State 5"); qqline(data.2$State5)

par(mfrow = c(2,3))
qqnorm(data.3$Lessthan30min, main = "(3) Q-Q: <30"); qqline(data.3$Lessthan30min)
qqnorm(data.3$Approx60min, main = "(3) Q-Q: Appx. 60"); qqline(data.3$Approx60min)
qqnorm(data.3$Approx90min, main = "(3) Q-Q: Appx. 90"); qqline(data.3$Approx90min)
qqnorm(data.3$Approx120min, main = "(3) Q-Q: Appx. 120"); qqline(data.3$Approx120min)
qqnorm(data.3$Approx180min, main = "(3) Q-Q: Appx. 180"); qqline(data.3$Approx180min)

# Histograms
par(mfrow = c(1,2))
hist(data.1$On.site, xlab = "score", main = "(1) Histogram: On-site")
hist(data.1$Off.site, xlab = "score", main = "(1) Histogram: Off-site")

par(mfrow = c(2,3))
hist(data.2$Ourstate, xlab = "score", main = "(2) Histogram: Our State")
hist(data.2$State1, xlab = "score", main = "(2) Histogram: State 1")
hist(data.2$State2, xlab = "score", main = "(2) Histogram: State 2")
hist(data.2$State3, xlab = "score", main = "(2) Histogram: State 3")
hist(data.2$State4, xlab = "score", main = "(2) Histogram: State 4")
hist(data.2$State5, xlab = "score", main = "(2) Histogram: State 5")

par(mfrow = c(2,3))
hist(data.3$Lessthan30min, xlab = "score", main = "(3) Histogram: Less than 30 min.")
hist(data.3$Approx60min, xlab = "score", main = "(3) Histogram: Appx. 60 min.")
hist(data.3$Approx90min, xlab = "score", main = "(3) Histogram: Appx. 90 min.")
hist(data.3$Approx120min, xlab = "score", main = "(3) Histogram: Appx. 120 min.")
hist(data.3$Approx180min, xlab = "score", main = "(3) Histogram: Appx. 180 min.")

# Boxplots
par(mfrow = c(1,1))
boxplot(data.1, main = "(1) On and Off Site Data", ylab = "score")

par(mfrow = c(1,1))
boxplot(data.2, main = "(2) State ", ylab = "score")

par(mfrow = c(1,1))
boxplot(data.3, main = "(3) Duration of Instruction", ylab = "score")

# Test of deviances
library(combinat)

test.deviance = function(trt1,trt2){
dev1 = trt1 - median(trt1);dev2 <- trt2 - median(trt2)
all = c(dev1, dev2)
index = seq(along=all)
indexIntrt1 = combn(index, 5)
RMD = NULL
for(i in 1:dim(indexIntrt1)[2]){
  RMD[i] = mean(abs(all[indexIntrt1[, i]]))/mean(abs(all[-indexIntrt1[, i]]))
}
p_value = sum(RMD >= RMD[1])/length(RMD)
p_value
}

# (1)
attach(data.1)
# Permutation Test on Deviances
test.deviance(On.site,Off.site[!is.na(Off.site)])

# Permutation Test for Medians
mediantst <- function(x, y, nreps=100) {

  d.obs <- abs(median(x) - median(y))

  nx <- length(x)
  ny <- length(y)
  tail.prob <- 0
  for(i in 1:nreps) {
    xy <- sample(c(x,y))           # permute combined list
    x <- xy[1:nx]                  # first nx are assigned to x
    y <- xy[seq(nx+1,nx+ny)]       # next ny to y
    d.sim <- abs(median(x) - median(y))
    if(d.sim >= d.obs)             # increment tail prob
      tail.prob <- tail.prob + 1
  }
  tail.prob <- tail.prob / nreps
  return(tail.prob)
}
# Modify the above function for the one-sided test:
mediantst.2 <- function(x, y, nreps=100) {
  d.obs <- median(x) - median(y)
  nx <- length(x)
  ny <- length(y)
  tail.prob <- 0
  for(i in 1:nreps) {
    xy <- sample(c(x,y))           # permute combined list
    x <- xy[1:nx]                  # first nx are assigned to x
    y <- xy[seq(nx+1,nx+ny)]       # next ny to y
    d.sim <- median(x) - median(y)
    if(d.sim >= d.obs)             # increment tail prob
      tail.prob <- tail.prob + 1
  }
  tail.prob <- tail.prob / nreps
  return(tail.prob)
}
mediantst.2(On.site,Off.site[!is.na(Off.site)], 10000)

# Wilcoxon Rank-Sum Test
wilcox.test(On.site,Off.site[!is.na(Off.site)])
# Power
pval = replicate(1000, wilcox.test(On.site,Off.site[!is.na(Off.site)])$p.value)
summary(pval)

# (2)
# Permutation Test on Deviances
states = list(0,0,0,0,0,0)
for (i in 1:6){
  states[[i]] = data.2[[i]][!is.na(data.2[i])]
}
p.values.list = NULL
for (i in 1:6){
  for(j in 1:6){
  p.values.list = c(p.values.list, test.deviance(states[[i]],states[[j]]))
}
}
min(p.values.list)<=0.05

# Permutation F-test
data.2.stack = stack(data.2)
runtime.anova = lm(data.2.stack$values~data.2.stack$ind)
teststat.obs = summary(runtime.anova)$fstatistic[1]
teststat = rep(NA, 10000)
for (i in 1:10000){
  ratingSHUFFLE = sample(data.2.stack$ind)
  SHUFFLE.anova = lm(data.2.stack$values~ratingSHUFFLE)
  teststat[i] = summary(SHUFFLE.anova)$fstatistic[1]
}
sum(teststat>=teststat.obs)/10000

# Kruskal-Wallis Test
kruskal.test(data.2.stack$values~data.2.stack$ind)

# Permutation Test for Medians
perms = NULL
for (i in 1:6){
  for (j in 1:6){
    perms = c(perms,
      (mediantst(data.2[[i]][!is.na(data.2[i])],data.2[[j]][!is.na(data.2[j])])))
    )
}
}
perms.frame = data.frame(d1 = c(perms[1:6]),
                         d2 = c(perms[7:12]),
                         d3 = c(perms[13:18]),
                         d4 = c(perms[19:24]),
                         d5 = c(perms[25:30]),
                         d6 = c(perms[31:36])
                         )
for (i in 1:6){
  for (j in 1:i){
  perms.frame[i,j] = NA
}}
perms.frame.bonf = perms.frame*15

library(DTK)
fac = gl.unequal(2, c(9,11))
TK.test(
  x = c(data.2$State2[1:9],data.2$State5[1:11]),
  f = fac,
  a = 0.05)

# (3)
# Permutation Test on Deviances
states = list(0,0,0,0,0)
for (i in 1:5){
  states[[i]] = data.3[[i]][!is.na(data.3[i])]
}
p.values.list = NULL
for (i in 1:5){
  for(j in 1:5){
    p.values.list = c(p.values.list, test.deviance(states[[i]],states[[j]]))
  }
}
min(p.values.list)<=0.05

# Permutation F-test
data.3.stack = stack(data.3)
runtime.anova = lm(data.3.stack$values~data.3.stack$ind)
teststat.obs = summary(runtime.anova)$fstatistic[1]
teststat = rep(NA, 10000)
for (i in 1:10000){
  ratingSHUFFLE = sample(data.3.stack$ind)
  SHUFFLE.anova = lm(data.3.stack$values~ratingSHUFFLE)
  teststat[i] = summary(SHUFFLE.anova)$fstatistic[1]
}
sum(teststat>=teststat.obs)/10000

# Kruskal-Wallis Test
kruskal.test(data.3.stack$values~data.3.stack$ind)
```

References

[1] J. J. Higgins, *Introduction to Modern Nonparametric Statistics*, Brooks/Cole, Belmont, CA, 2004.

[2] R. C. Blair, J. J. Higgins, *A Comparison of the Power of Wilcoxon's Rank-Sum Statistic to That of Student's t Statistic under Various Nonnormal Distributions*, American Educational Research Association, Journal of Educational Statistics, Vol. 5, No. 4 (Winter, 1980), pp. 309-335, DOI: 10.2307/1164905.

[3] M. Mahoney, R. Magel, *Estimation of the Power of the Kruskal-Wallis Test*, Biometrical Journal, Vol. 38, No. 5 (1996), pp. 613-630.

[4] M. K. Lau, *DTK: Dunnett-Tukey-Kramer Pairwise Multiple Comparison Test Adjusted for Unequal Variances and Unequal Sample Sizes.*

[5] A. R. Rogers, *Permutation Tests*, University of Utah. URL: http://content.csbs.utah.edu/∼rogers/datanal/labprj/permtst/index.html