Nonparametric Statistical Methods Final

Ryan Honea

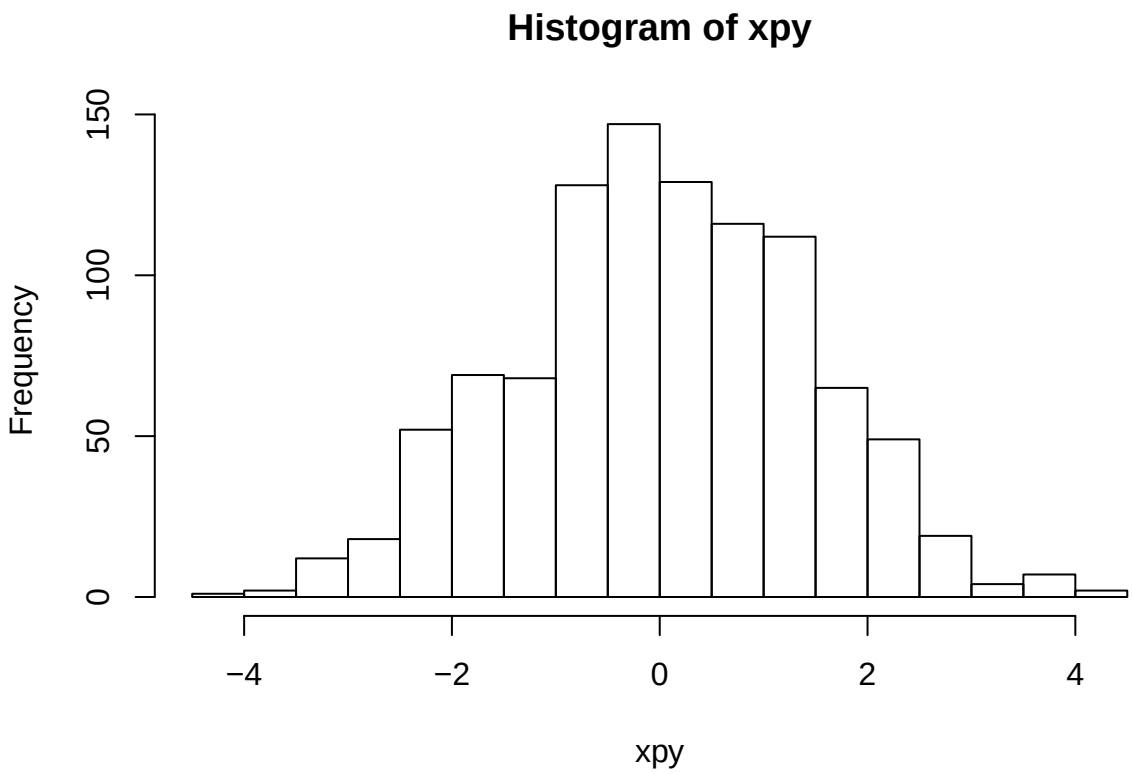
12/08/2017

Problem One

Let X and Y be independent random variables with the same density function $f(x)$. Also assume the density function $f(x)$ is symmetric about 0. Prove that $P(X+Y>0)=\frac{1}{2}$.

Solution: I will show that the mean of $E[X+Y]=0$ and that we can show from there that $P(X+Y>0)=\frac{1}{2}$. Firstly, we know that $E[X+Y]=E[X]+E[Y]$ because they are independent variables. Secondly, $E[X]=0$ because $f(x)$ is symmetric about (i.e. the mean is 0). We also know that $E[Y]=0$ for similar reasons. Therefore $E[X+Y]=E[X]+E[Y]=0$. As the mean of $X+Y$ is 0, then $P(X+Y>0)=\frac{1}{2}$. For sake of visual evidence, I run a quick simulation.

```
x <- rnorm(1000)
y <- rnorm(1000)
xpy <- x + y
hist(xpy, breaks= 30)
```




```
cat("P(x+y > 0) =",length(xpy[xpy > 0])/length(xpy))
```

```
## P(x+y > 0) = 0.503
```

Problem Two

Find the probability that the range of a random sample of size 3 from the uniform distribution is less than 0.8.

Solution: We know from Chapter 2.7.2 that the range of the standard uniform has beta distribution with parameters $n-1$ and 2, so $R\ Beta(2,2)$. We find the value with R.

```
cat("Probability:",pbeta(.8, 2, 2))
```

```
## Probability: 0.896
```

Problem Three

Find the expected value and variance of the range of a random sample of size 3 from the standard uniform distribution.

Solution: Given that $f(x_{(i)}) = \frac{n!}{(i-1)!(n-i)!}F(x)^{i-1}\left[1-F(x)\right]^{n-i}f(x)$, and $X_1, X_2, X_3 \sim U(0,1)$

$$f(x_{(3)}) = \frac{3!}{(3-1)!(3-3)!}x^{3-1}(1-x)^{3-3} \qquad f(x_{(1)}) = \frac{3!}{(1-1)!(3-1)!}x^{1-1}(1-x)^{3-1}$$
$$= 3x^2 \qquad = 3(1-x)^2$$

$$E[x_{(3)} - x_{(1)}] = E[x_{(3)}] - E[x_{(1)}]$$

$$E[x_{(3)}] = 3\int_0^1 x^3 dx \qquad E[x_{(1)}] = 3\int_0^1 x(1-x)^2 dx$$
$$= 3\left[\frac{x^4}{4}\Big|_0^1\right] \qquad = 3\left[\left(\frac{x^4}{4} - \frac{2x^3}{3} + \frac{x^2}{2}\Big|_0^1\right)\right]$$
$$= \frac{3}{4} \qquad = \frac{3}{4} - \frac{6}{3} + \frac{3}{2} = \frac{1}{4}$$

$$E[x_{(3)} - x_{(1)}] = \frac{1}{2}$$

Now that we found the expected value through the traditional means, I'll find the variance using the method from Problem Three that acknowledges that this is distributed Beta(2,2). The variance of Beta(2,2) is

$$\frac{2*2}{(2+2)^2(2+2+1)} = \frac{4}{80} = \frac{1}{20}$$

And so the expected value is 0.5 and the variance is 0.2.

Problem Four

Forty people were selected at random in the following order

$$M \quad M \quad F \quad F \quad F \quad F \quad M \quad F \quad F \quad M \quad M \quad F \quad M \quad M \quad M \quad M \quad F \quad F \quad M \quad M$$

$$F \quad M \quad F \quad F \quad M \quad M \quad M \quad M \quad F \quad F \quad M \quad F \quad M \quad M \quad F \quad F \quad M \quad M \quad M \quad F$$

Is it true that the people were selected at random?

Solution: We can enter this as a seriies of 1s and 0s and just use the Runs Test for Randomness.

So, our null hypothesis is that sequence X is random, and our alternative hypothesis is that X is not random.

```
require(randtests)
X <- c(0,0,1,1,1,1,0,1,1,0,0,1,0,0,0,0,1,1,0,0,
           1,0,1,1,0,0,0,0,1,1,0,1,0,0,1,1,0,0,0,1)
runs.test(X, threshold = 0.5)
```

```
##
##  Runs Test
##
## data:  X
## statistic = -0.25895, runs = 20, n1 = 18, n2 = 22, n = 40, p-value
## = 0.7957
## alternative hypothesis: nonrandomness
```

Based on a p-value of 0.7957, we fail to reject the null hypothesis and thus do not have sufficient evidence to assume the data is not random.

Problem Five

The three graphs in the Figure above illustrate some kind of patterns. Time is on the horizontal axis. The data vaslues are indicated by dots and the horizontal line denotes the median of the data. For each graph, use the run test to determine if the patterns are random.

In the absence of graphs, I will describe each graph by its point where 1 means it is above the median line, 0 means it is on the median line, and -1 means it is below the median line.

(a)	-1	-1	-1	1	-1	-1	-1	0	1	1	1	1	1	1	-1
(b)	-1	-1	1	-1	1	-1	-1	0	1	1	1	1	1	1	1
(c)	1	1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	-1	1

Solution: For all of these, the null hypothesis is that they are random sequences, and the alternative is that they are not random.

```
a.seq <- c(-1,-1,-1,1,-1,-1,-1,0,1,1,1,1,1,1,-1)
b.seq <- c(-1,-1,1,-1,1,-1,-1,0,1,1,1,1,1,1,1)
c.seq <- c(1,1,1,-1,-1,-1,-1,-1,1,1,1,1,-1,-1,1)
```

For **(a)**, we have

```
runs.test(a.seq, threshold = 0)
```

```
##
##  Runs Test
##
## data:  a.seq
## statistic = -1.669, runs = 5, n1 = 7, n2 = 7, n = 14, p-value =
```

```
## 0.09511
## alternative hypothesis: nonrandomness
```

It is clear that the test avoided the case where $a_{(i)} = 0$ as $n = 14$. With a p-value of 0.09511, we fail to reject the null hypothesis and thus do not have sufficient evidence to assume the data is not random.

For **(b)**, we have

```
runs.test(b.seq, threshold = 0)
```

```
##
##  Runs Test
##
## data:  b.seq
## statistic = -0.87191, runs = 6, n1 = 9, n2 = 5, n = 14, p-value =
## 0.3833
## alternative hypothesis: nonrandomness
```

It is clear that the test avoided the case where $b_{(i)} = 0$ as $n = 14$. With a p-value of 0.3833, we fail to reject the null hypothesis and thus do not have sufficient evidence to assume the data is not random.

For **(c)**, we have

```
runs.test(c.seq, threshold = 0)
```

```
##
##  Runs Test
##
## data:  c.seq
## statistic = -1.8667, runs = 5, n1 = 8, n2 = 7, n = 15, p-value =
## 0.06194
## alternative hypothesis: nonrandomness
```

With a p-value of .06194, we fail to reject the null hypothesis and thus do not have sufficient evidence to assume the data is not random.

All of these conclusions are based on a confidence level of 95%.

Problem Six

Two types of corn (golden and green-striped) carry recessive genes. When these were crosses, a first generation was obtained, which was consistently normal (neither golden nor green-strioed). When this generation was allowed to self-fertilize, four distinct types of plants were produced: normal, golden, green-striped, and golden-green-sriped. In 1200 plants, this process produced the following distribution:

Normal	Golden	Green-Striped	Golden-Green-Striped
670	230	238	62

A monk named Mendel wrote an article theorizing that in a second generation of such hybrids, the distribution of plant types should be in a 9:3:3:1 ratio. Are the above data consistent with the good monk's theory?

Solution: The frequency distribution can be summarized by the table below.

Type	Frequency	Observed Proportion $p(x)$	Expected Proportion $p_0(x)$
Normal	670	.5583	0.5625
Golden	230	.1917	0.1875
Green-Striped	238	.1983	0.1875
Golden-Green-Striped	62	.0517	.0625

So we have

$$H_0 : p(x) = p_0(x) \qquad H_a : p(x) \neq p_0(x)$$

This can be tested with the chi-squared test.

```
observed <- c(670, 230, 238, 62)
expected <- c(.5625, .1875, .1875, .0625)
chisq.test(observed, expected)
```

```
## Warning in chisq.test(observed, expected): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  observed and expected
## X-squared = 8, df = 6, p-value = 0.2381
```

With a p-value of .2381, we fail to reject the null hypothesis and thus do not have sufficient evidence to assume that $p(x) \neq p_0(x)$.

Problem Seven

Based on a sample size n, let $F_n(x)$ be the Empirical CDF of a population with CDF $F(x)$. Prove that $F_n(x)$ is an unbiased and consistent estimator of $F(x)$.

Solution: We prove that $F_n(x)$ is unbiased by showing that $E[F_n(x)] = E[F(x)]$.

$$\begin{aligned} E[F_n(x)] &= E[\sum_{i=1}^{n} \frac{\mathbb{1}(X_i \leq x)}{n}] \\ &= \frac{1}{n} \sum_{i=1}^{n} E[\mathbb{1}(X_i \leq x)] \\ &= \frac{1}{n} \sum_{i=1}^{n} P(X_i \leq x) = \frac{1}{n} n F_X(x) \\ &= F_X(x) \end{aligned}$$

To show that $F_n(x)$ is consistent, we use the law of large numbers. Because $F_n(x)$ is the sum of random variables, then $F_n(x) \xrightarrow{p} F(x)$.

Problem Eight

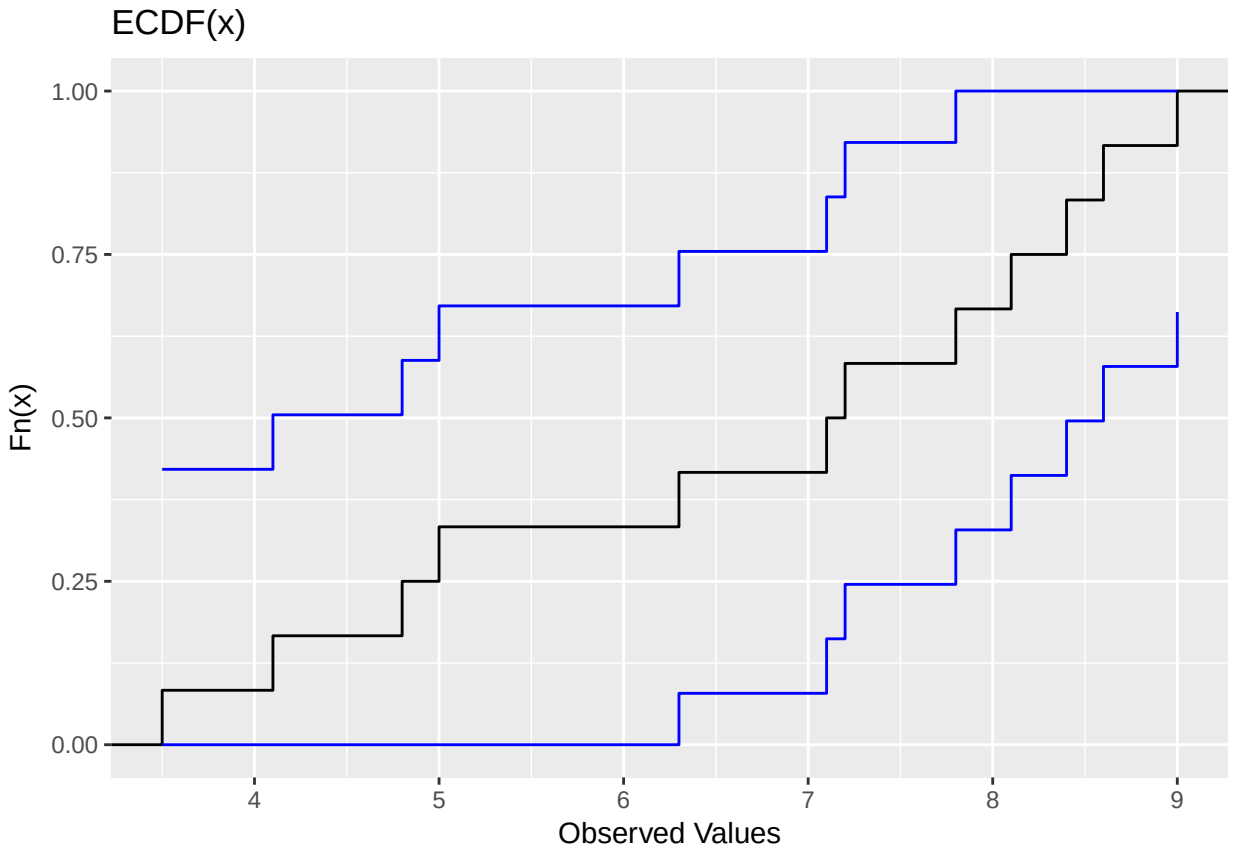
A random sample of size 12 is drawn from an unknown continuous population $F_X(x)$, with the following results

3.5 4.1 4.8 5.0 6.3 7.1 7.2 7.8 8.1 8.4 8.6 9.0

A 90% confidence band is desired for $F_X(x)$. Plot a graph of the empirical distribution function $F_n(x)$ and resulting confidence bands.

Solution: For a sample of size 12 and $\alpha = .10$, we can use a table to find the $D_{n,\alpha}$ value that can be used to create a confidence band. With this table's results, we find the value to be 0.338. So, below, we create the empirical distribution with R's commands and then find the upper and lower confidence bands.

```r
require(ggplot2)
x <- c(3.5, 4.1, 4.8, 5.0, 6.3, 7.1, 7.2, 7.8, 8.1, 8.4, 8.6, 9.0)
dist.x <- ecdf(x)
lower <- c()
upper <- c()
for (i in 1:12) {
  lower[i] <- max(dist.x(x[i]) - .338, 0)
  upper[i] <- min(dist.x(x[i]) + .338, 1)
}
pl = data.frame(x = x, y = dist.x(x))
ggplot(pl,aes(x)) +
  stat_ecdf(geom = "step") +
  geom_step(aes(y=lower), colour="blue") +
  geom_step(aes(y=upper), colour="blue") +
  labs(x = "Observed Values", y = ("Fn(x)")) +
  labs(title = "ECDF(x)")
```



The black lines represents the ECDF where the blue lines represent the confidence bands.

Problem Nine

Ten students take a test and their scores (out of 100) are as follows:

$$95 \quad 80 \quad 40 \quad 52 \quad 60 \quad 80 \quad 82 \quad 58 \quad 65 \quad 50$$

Test the null hypothesis that the cumulative distribution function of the proportion of right answers a student gets on a test is

$$F_0(x) = \begin{cases} 0, & x < 0 \\ x^2(3-2x), & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$$

Solution: Our hypotheses therefore are

$$H_0 : F(x) = F_0(x) \qquad H_a : F(x) \neq F_0(x)$$

We can test this using the Kolgomorov-Smirnov test.

```
grades <- c(.95, .80, .40, .52, .60, .82, .57, .65, .50)
dist <- function(x) {
  if (x < 0) {
    return(0)
  } else if (x > 1) {
    return(1)
  } else
  return (x^2*(3-2*x))
}
ks.test(grades,"dist")
```

```
## Warning in if (x < 0) {: the condition has length > 1 and only the first
## element will be used
```

```
## Warning in if (x > 1) {: the condition has length > 1 and only the first
## element will be used
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  grades
## D = 0.38889, p-value = 0.09783
## alternative hypothesis: two-sided
```

Based on the following p-value, we fail to reject the null hypothesis that the distributions are equal and so have no evidence to suggest that they aren't.

Problem Ten

Length, in centimeters, of 10 randomly selected pygmy sunfish are given by:

$$5.0 \quad 3.9 \quad 5.2 \quad 5.5 \quad 2.8 \quad 6.1 \quad 6.4 \quad 2.6 \quad 1.7 \quad 4.3$$

Can we conclude that the median length of pygmy sunfish differs significantly from 3.7 centimeters?

Solution: We have the following hypotheses

$$H_0 : M = 3.7\text{cm} \qquad H_a : M \neq 3.7\text{cm}$$

I'll use the sign test to test this hypothesis.

```
require(BSDA)
```

```
## Loading required package: BSDA
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
x <- c(5.0, 3.9, 5.5, 2.8, 5.1, 6.4, 2.6, 1.7, 4.3)
SIGN.test(x, md = 3.7, alternative = "two.sided")
```

```
##
##  One-sample Sign-Test
##
## data:  x
## s = 6, p-value = 0.5078
## alternative hypothesis: true median is not equal to 3.7
## 95 percent confidence interval:
##  2.615556 5.468889
## sample estimates:
## median of x
##         4.3
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level L.E.pt U.E.pt
## Lower Achieved CI     0.8203 2.8000 5.1000
## Interpolated CI       0.9500 2.6156 5.4689
## Upper Achieved CI     0.9609 2.6000 5.5000
```

With a p-value of .5078, we fail to reject the null hypothesis and thus do not have sufficient evidence to assume that $M \neq 3.7$.

Problem Eleven

Consider an independent and identically distributed continuous random sample $X_1, X_2, ..., X_n$ from a population with CDF $F_X(x)$. Let M be the median of the population. Define a variable $U(X_i)$ and Y_i as:

$$U(X_i) = \begin{cases} 1, & \text{if } X_i > M \\ 0, & \text{otherwise} \end{cases}$$

Consider another set of random variables $Y_1, Y_2, ..., Y_n$ such that:

$$Y_i = \begin{cases} 0, & \text{with probability } \frac{1}{2} \\ i, & \text{with probability } \frac{1}{2} \end{cases}$$

Show that

$$W = \sum_{i=1}^{n} U(X_i) R_i \qquad \text{and} \qquad Y = \sum_{i=1}^{n} Y_i$$

have the same probability distribution where $R_i = \text{rank}(X_i)$.

Solution: We use the probability generating functions of W and Y to show that they have the same probability distribution. Recall that the probability generating functions have the general form $PGF(X) = E[S^x]$.

$$
\begin{aligned}
PGF(W) &= E[S^W] = E[S^{\sum_{i=1}^n U(X_i)R_i}] \\
&= \prod_{i=1}^n E[S^{U(X_i)R_i}] \\
&= \prod_{i=1}^n P_{U(X_i)R_i}(S)
\end{aligned}
\qquad
\begin{aligned}
PGF(Y) &= E[S^Y] = E[S^{\sum_{i=1}^n Y_i}] \\
&= \prod_{i=1}^n E[S^{Y_i}] \\
&= \prod_{i=1}^n P_{Y_i}(S)
\end{aligned}
$$

So, we need to show that $P_{U(X_i)R_i}(S)$ and $P_{Y_i}(S)$ are equivelant.

$$
\begin{aligned}
P_{Y_i}(S) &= E[S^{Y_i}] \\
&= S^0 P(Y_i = 0) + S^i P(Y_i = i) \\
&= S^0 \frac{1}{2} + S^i \frac{1}{2} \\
&= \frac{1}{2}(1 + S^i)
\end{aligned}
\qquad
\begin{aligned}
P_{U(X_i)R_i}(S) &= E[S^{U(X_i)R_i}] \\
&= S^0 P(U_i R_i = 0) + S^{U(X_i)R_i} P(U(X_i)R_i = R_i) \\
&= \frac{1}{2} + S^{R_i} P(U(X_i) = 1) \\
&= \frac{1}{2}(1 + S^{R_i})
\end{aligned}
$$

Now, note that $\sum_{i=1}^n R_i = \sum_{i=1}^n i$ because they both span the range of $i, R_i = 1, 2, ..., n$ and so $\prod_{i=1}^n S^i = \prod_{i=1}^n S^{R_i}$.

Therefore, we have referring back to $PGF(W)$ and $PGF(Y)\ldots$

$$
\begin{aligned}
PGF(W) &= \prod_{i=1}^n \frac{1}{2}(1 + S^i) \\
&= 2^{-n} \prod_{i=1}^n (1 + S^i)
\end{aligned}
\qquad
\begin{aligned}
PGF(Y) &= \prod_{i=1}^n \frac{1}{2}(1 + S^{R_i}) \\
&= 2^{-n} \prod_{i=1}^n (1 + S^i)
\end{aligned}
$$

$$
PGF(W) = PGF(Y)
$$

Problem Twelve

In a marketing research test, 15 adult males were asked to shave one side of their face with brand A razor blade and the other side with brand B razor blade and state their preferred blade. Twelve men preferred brand A. Find the P value for the alternative that the probability of preferred brand A is greater that 0.5. [Use exact Binomial test].

Solution: We have the hypotheses

$$
H_0 : P(A) = P(B) = 0.5 \qquad H_a : P(A) > P(B)
$$

```
binom.test(12, 15, 0.5, alternative = "greater")
```

```
##
##  Exact binomial test
##
```

```
## data:  12 and 15
## number of successes = 12, number of trials = 15, p-value = 0.01758
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.5602156 1.0000000
## sample estimates:
## probability of success
##                    0.8
```

With a p-value of 0.01758, we reject the null hypothesis that the probability of preferred brand A is equal to the probability of preferred brand B. Specifically, we choose to prefer that $P(A) > 0.5$.

Problem Thirteen

Develop a hypothesis test procedure for testing a 90^{th} percentile. Also find a method to build a confidence interval.

Solution: First, we develop a test statistic utilizing the following pieces of information

$$\phi_i = \begin{cases} 1, & \text{if } x_i > \theta \\ 0, & \text{if } x_i < \theta \end{cases}, \qquad S = \sum_{i=1}^{n} \phi_i \sim \text{Bin}(n, .10)$$

And our rejection region is

$$R = \begin{cases} s > c_1, & \text{if } H_a : \theta > \theta_0 \\ s < c_2, & \text{if } H_a : \theta < \theta_0 \\ s > c_1 \cup s < c_2, & \text{if } H_a : \theta \neq \theta_0 \end{cases}$$

It's clear that S does not have symmetric taisl, and that there will be cases where the two-tailed test has significance equal to the significance of a one-tail test.

Creating a confidence interval is by finding for what values of r and s does the following hold,

$$P(x_{(r)} \leq \theta \leq x_{(s)}) \leq 1 - \alpha$$

There are obviously going to be several cases where this is true, and so the objective is to maximize $P(x_{(r)} \leq \theta \leq x_{(s)})$ subject to the condition that is is $\leq 1 - \alpha$.

Problem Fourteen

The 2000 census statistics for Alabama give the percentage changes in population between 1990 and 2000 for each of the 67 counties. These counties were divided into two mutually independent groups, Rural and Nonrural, according to population sizes of less than 25,000 in the year 2000 or not. Random samples of nine Rural and seven Nonrural counties gave the following data on percentage population change:

Rural	1.1	21.7	16.3	11.3	10.4	7.0	2.0	1.9	6.2
Nonrural	2.4	9.9	14.2	18.4	20.1	23.1	70.4		

Use Mathisan Median test and Mann-Whitney U to test the null hypothesis of equal distribution. Also mention the required assumption.

Solution: We define the distribution of the rural to be $F_R(x)$ and the distribution of the Nonrural to be $F_N(x)$. We are testing, therefore

$$H_0 : F_R(x) = F_N(x) \qquad H_a : F_R(x) \neq F_N(x)$$

We test this hypothesis in R.

```r
require(agricolae)
```

```
## Loading required package: agricolae
```

```r
rural <- c(1.1, 21.7, 16.3, 11.3, 10.4, 7.0, 2.0, 1.9, 6.2)
nonrural <- c(2.4, 9.9, 14.2, 18.4, 20.1, 23.1, 70.4)
rural <- cbind(rural, rep(1, length(rural)))
nonrural <- cbind(nonrural, rep(2, length(nonrural)))
dataset <- rbind(rural, nonrural)
Median.test(dataset[,1], dataset[,2])
```

```
##
## The Median Test for dataset[, 1] ~ dataset[, 2]
##
## Chi Square = 2.285714   DF = 1   P.Value 0.13057
## Median = 10.85
##
##   Median r Min  Max   Q25  Q75
## 1    7.0 9 1.1 21.7  2.00 11.3
## 2   18.4 7 2.4 70.4 12.05 21.6
##
## Post Hoc Analysis
##
## Groups according to probability of treatment differences and alpha level.
##
## Treatments with the same letter are not significantly different.
##
##   dataset[, 1] groups
## 2         18.4      a
## 1          7.0      a
```

```r
wilcox.test(rural, nonrural)
```

```
## Warning in wilcox.test.default(rural, nonrural): cannot compute exact p-
## value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  rural and nonrural
## W = 59.5, p-value = 0.01061
## alternative hypothesis: true location shift is not equal to 0
```

In the case of the Mathisen median test, a p-value of 0.13057 suggests that the treatments are not significant different, however in the case of the Wilcoxon (Mann-Whitney U), we obtain a p-value that suggests the location of the rural group is significantly different from the nonrural group. Just looking at the data, I'm more likely to side with the results of the Mann-Whitney U.

There are a few assumptions I need to make to do that though, namely that the dataset is symmetric and that although there are ties, the normal approximation will suffice.

Problem Fifteen

Prove that the null mean of the Wilcoxon rank-sum statistic is $m(m+n+1)/2$

Solution: We can show that for two sample sets $X_1, ..., X_n \sim F(x)$ and $Y_1, ..., Y_n \sim G(x)$, $E[W] = \frac{m(m+n+1)}{2}$.

$$W = \sum_{i}^{N} iZ_i$$

$$Z_i = \begin{cases} 1 & Z_i \in X_1, X_2, ..., X_m \\ 0 & Z_i \in Y_1, Y_2, ..., Y_n \end{cases}$$

From there, we can show $E[W]$.

$$\begin{aligned} E[W] = E[\sum_{i}^{m+n} iZ_i] &= \sum_{i}^{N} iE[Z_i] \\ &= \sum_{i}^{m+n} i\frac{m}{m+n} \\ &= \frac{m}{m+n}\frac{(m+n)(m+n+1)}{2} \\ &= \frac{m(m+n+1)}{2} \end{aligned}$$

Problem Sixteen

Show that the Wilcoxon rank-sum test and the Mann-Whitney U are equivelant.

Solution: Recall that the Mann-Whitney U Statistic is

$$U = \sum_{i}^{m}\sum_{j}^{n} D_{ij} \qquad\qquad D_{ij} = \begin{cases} 1 & \text{if } Y_j < X_i \\ 0 & \text{if } Y_j > X_i \end{cases}$$

Now, let $t_i =$ Number of X's less than or equal to X_i. Then $\sum_{i}^{n} D_{ij}$ is the number of values of j for which $Y_j < X_i$, or the rank of X_i reduced by t_i.

So, we have

$$\begin{aligned} U &= \sum_{i=1}^{m}[r(X_i) - t_i] \\ &= \sum_{i=1}^{m} r(X_i) - (t_1 ... t_m) \\ &= \sum_{i=1}^{m} iZ_i - (1 + 2 + ... + m) \\ &= W_N - \frac{m(m+1)}{2} \end{aligned}$$

so we have shown that the Wilcoxon Rank-Sum and Mann Whitney are equivelant.

Problem Seventeen

The psychology department of public universities in each of two different states accepted seven and nine applicants, respectively, for graduate study next fall. Their respective scores on the Graduate Record Examination are:

University X	1200	1220	1300	1170	1080	1110	1130		
University Y	1210	1180	1000	1010	980	1400	1430	1390	970

The sample median and mean scores of applicants to the two universities are close to being equal, so an assumption of equal location may well be justified. Use the Siegel-Tukey test to see which university has smaller variability in scores, if either.

Solution: In this case the null hypothesis is a test of scale, so we test the following

$$H_0 : \sigma_x^2 >= \sigma_y^2 \qquad H_a : \sigma_x^2 < \sigma_y^2$$

and we can test this with the Siegel-Tukey test in R.

```
require(jmuOutlier)
```

```
## Loading required package: jmuOutlier
```

```
x <- c(1200, 1220, 1300, 1170, 1080, 1110, 1130)
y <- c(1210, 1180, 1000, 1010, 980, 1400, 1430, 1390, 970)
siegel.test(x, y, alternative = "less")
```

```
## [[1]]
## [1] "   Siegel-Tukey test"
##
## $alternative
## [1] "less"
##
## $rank.x
## [1] 14 10  7 16  9 12 13
##
## $rank.y
## [1] 11 15  5  8  4  3  2  6  1
##
## $p.value
## [1] 0.01145105
```

With a p-value of 0.01, we reject the null hypothesis and have evidence to suggest that the variability of University X is smaller than that of University Y.

Problem Eighteen

Below are four sets of five measurements, each set an array of data on the smoothness of a certain type of paper, each set obtained from a different laboratory. Find an appropriate P value to test whether the median smoothness can be regarded as the same for all laboratories.

A	38.7	41.5	43.8	44.5	45.5
B	39.2	39.3	39.7	41.4	41.8
C	34.0	35.0	39.0	40.0	43.0
D	34.1	34.8	34.9	35.4	37.2

Solution: Our null hypothesis therefore is

$$H_0 : M_A = M_B = M_C = M_D \qquad H_a : M_A \neq M_B \neq M_C \neq M_D$$

. We can use Mood's Test to test this hypothesis as it can handle multiple samples. Because the R variation can only test with two groups, specialized code has to be written to test it.

```
A <- c(38.7,41.5,43.8,44.5,45.5)
B <- c(39.2,39.3,39.7,41.4,41.8)
C <- c(34.0,35.0,39.0,40.0,43.0)
D <- c(34.1,34.8,34.9,35.4,37.2)
mood.median.test <- function(w, x, y, z) {
  h <- c(w, x, y, z)
  g <- rep(1:4, c(length(w),length(x),length(y),length(z)))
  m <- median(h)
  g
  fisher.test(h < m, g)$p.value
}

mood.median.test(A, B, C, D)
```

```
## [1] 0.04874537
```

Based on the following p-value, we reject the null hypothesis and therefore have enough evidence to suggest that the medians of the data are not equal.

Problem Nineteen

A beauty contest has eight contestants. Two judges are each asked to rank the contestants in a preferential order of pulchritude.

Judge	A	B	C	D	E	F	G	H
I	2	1	3	5	4	8	7	6
II	1	2	4	5	7	6	8	3

(a) Calculate Kendal-tau coefficient.
(b) Test the null hypothesis that the judges ranked the contestants independently and find the P value.

Solution: **(a)**: We can estimate this with the function below.

```
I <- c(2,1,3,5,4,8,7,6)
II <- c(1,2,4,5,7,6,8,3)
kendalltau.coef <- function(x, y) {
  n <- length(x)
  sum <- 0
  for (i in 1:(n-1)) {
    for (j in i:n) {
      sum <- sum + (sign(x[j] - x[i])*sign(y[j]-y[i]))
    }
  }
  return(sum/choose(n,2))
}
kendalltau.coef(I, II)
```

```
## [1] 0.5
```

(b): We can test this using the kendall's rank correlation tau with R.

```
cor.test(I, II, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
```

```
##
## data:  I and II
## T = 21, p-value = 0.1087
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
## tau
## 0.5
```

Based on this p-value of 0.1087, we fail to reject the null hypothesis and so do not have evidence to suggest the judges did not grade them independently.

Problem Twenty

Suppose A and B are two small colleges, the results of the beginning physics course at each of the two schools are given in a table.

College	Pass	Fail
I	8	14
II	1	3

Do the data provide sufficient evidence to indicate that the proportions passing Physics differ for the two colleges? Use Fisher's exact test.

Solution: We test

$$H_0 : p_I = p_I I \qquad H_a : p_I \neq P_I I$$

in R with

```
pf <- matrix(c(8, 14, 1, 3),
       nrow = 2,
       dimnames = list(Test = c("Pass", "Fail"),
                       College = c("I", "II")))
fisher.test(pf)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  pf
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    0.1116273 101.0207966
## sample estimates:
## odds ratio
##   1.681372
```

With a p-value of 1, we can safely fail to reject the null hypothesis and therefore do not have enough evidence to suggest that the ratio of odds is not 1.

Problem Twenty-One

A random sample of 135 U.S. citizens were asked their opinion about the current U.S. foreign policy in Afghanistan. Forty-three reported a negative opinion and the others were positive. These 135 persons were then put on a mailing list to receive an informative newsletter about U.S. foreign policy, and then asked their opinion later. At the time, 37 were opposed and 30 of these 37 originally had a positive opinion. Find the

P value for the alternative that the probability of a change from negative to positive is greater than the corresponding probability of a change in the opposite direction.

Solution: To understand this better, I'm going to put it in a table.

	Negative	Positive
Before	43	92
After	37	98

So if thirty changed their mind from positive to negative, and there are six less, then 36 must have changed their mind from negative to positive. That is the probability of Negative to Positive $P(- \to +) = \frac{36}{43}$ and $P(+ \to -) = \frac{30}{92}$. Our hypothesis therefore is

$$H_0 : P(- \to +) \leq \frac{36}{43} \qquad H_a : P(- \to +) > \frac{36}{43}$$

We can use Fisher's exact test for this problem, but we will need to represent our data in a different way.

	After Newsletter		
Before Newsletter	**Negative**	**Positive**	**Total**
Negative	7	36	43
Positive	30	62	92
Total	37	98	135

We can then enter this into a matrix in R.

```
pf <- matrix(c(7, 30, 36, 62),
       nrow = 2,
       dimnames = list(Before = c("Negative", "Positive"),
                       After = c("Negative", "Positive")))
fisher.test(pf, alternative = "less")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  pf
## p-value = 0.03529
## alternative hypothesis: true odds ratio is less than 1
## 95 percent confidence interval:
##  0.000000 0.933575
## sample estimates:
## odds ratio
##  0.4043726
```

With a p-value of 0.03529, we can reject the null hypothesis and have evidence to suggest that the odds ratio is less than 1, i.e. $P(- \to +)$ is likely greater.