Midterm

Ryan Honea

10/12/2017

Problem One

If $X_1, X_2, ..., X_n$ be an iid random sample from Uniform(0,1). Let M_n be the sample median. Compute $E[M_n]$ when n is odd.

Solution: Given that $X_1 ... X_n \sim U(0,1)$, then we know that the density of the order statistic is

$$f(X_{(i)}) = n\binom{n-1}{i-1}F(x)^{i-1}(1-F(x))^{n-i}f(x) = n\binom{n-1}{i-1}x^{i-1}(1-x)^{n-i}$$

By definition of the median when n is odd, $M_n = X_{(n+1)/2}$ and so we have

$$\begin{aligned}E[M_n] &= E[X_{((n+1)/2)}]\\ &= \frac{n!}{(\frac{n+1}{2}-1)!(n-\frac{n+1}{2})!}\int_0^1 x^{(n+1)/2}(1-x)^{n-(n+1)/2}dx\\ &= \frac{\Gamma(n+1)}{\Gamma(\frac{n+1}{2})\Gamma(n-\frac{n+1}{2}+1)}\int_0^1 x^{(n+1)/2}(1-x)^{n-(n+1)/2}dx\end{aligned}$$

From this point, we notice that the integral is similar to the beta distribution and engineer it so that it integrates to 1 by making it into the beta distribution where $\alpha = \frac{n+1}{2}+1$ and $\beta = n - \frac{n+1}{2}+1$

$$\begin{aligned}E[M_n] &= E[X_{((n+1)/2)}]\\ &= \frac{\Gamma(n+1)}{\Gamma(\frac{n+1}{2})\Gamma(n-\frac{n+1}{2}+1)}\int_0^1 x^{(n+1)/2}(1-x)^{n-(n+1)/2}dx\\ &= \frac{\Gamma(n+1)}{\Gamma(\frac{n+1}{2})\Gamma(n-\frac{n+1}{2}+1)}\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{(n+1)/2}(1-x)^{n-(n+1)/2}dx\\ &= \frac{\Gamma(n+1)}{\Gamma(\frac{n+1}{2})\Gamma(n-\frac{n+1}{2}+1)}\frac{\Gamma(\frac{n+1}{2}+1)\Gamma(n-\frac{n+1}{2}+1)}{\Gamma(n+2)}\int_0^1 \frac{\Gamma(n+2)}{\Gamma(\frac{n+1}{2}+1)\Gamma(n-\frac{n+1}{2}+1)}x^{(n+1)/2}(1-x)^{n-(n+1)/2}dx\\ &= \frac{\Gamma(n+1)}{\Gamma(\frac{n+1}{2})\Gamma(n-\frac{n+1}{2}+1)}\frac{\Gamma(\frac{n+1}{2}+1)\Gamma(n-\frac{n+1}{2}+1)}{\Gamma(n+2)} = \frac{\Gamma(n+1)\Gamma(\frac{n+1}{2}+1)}{\Gamma(\frac{n+1}{2})\Gamma(n+2)}\\ &= \frac{n!(\frac{n+1}{2})!}{(\frac{n+1}{2}-1)!(n+1)!} = \frac{\frac{n+1}{2}}{n+1} = \frac{1}{2}\end{aligned}$$

Problem Two

Let $X_1, X_2, ..., X_n$ be a random sample from population with pdf $f_x(x) = 2e^{-2x}; \quad x \geq 0$. Compute the expected value of the range. That is compute $E[X_{(n)} - X_{(1)}]$.

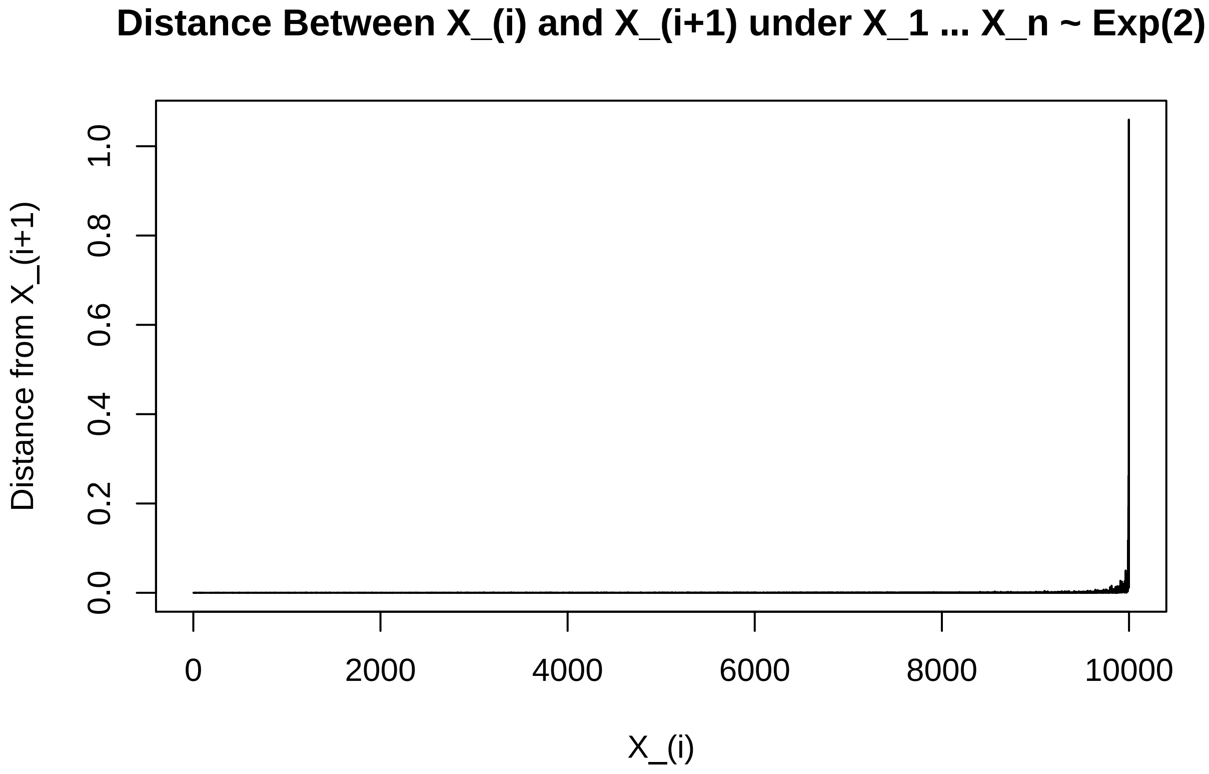
Solution: Given that $f(x_{(k)}) = n\binom{n-1}{i-1}F(x)^{i-1}\left[1-F(x)\right]^{n-i}f(x)$, and $X_1...X_n \sim Exp(\lambda = 2)$

$$\begin{aligned} f(x_{(1)}) &= n\binom{n-1}{0}\left[e^{-2x}\right]^{n-1}2e^{-2x} \\ &= 2ne^{-2x}e^{-2xn+2x} \\ &= 2ne^{-2xn} \end{aligned}$$

$$\begin{aligned} f(x_{(n)}) &= n\binom{n-1}{n-1}\left[1-e^{-2x}\right]^{n-1}2e^{-2x} \\ &= 2ne^{-2x}\left[1-e^{-2x}\right]^{n-1} \end{aligned}$$

There isn't much that can be done with the distribution of the max order statistic in the method above, so we use the memoryless property of the exponential distribution to find the expected maximum. By the memoryless property, the distance between $X_{(1)}$ and $X_{(2)}$ is the same as the distance until the first of $n-1$ observations takes on a value, or $E[X_{(2)} - X_{(1)}] = \frac{1}{(n-1)2}$, and then similarly, the distance between $X_{(3)}$ and $X_{(2)}$ is the same as the distance until the first of $n-2$ observations takes on a value, or $E[X_{(3)} - X_{(2)}] = \frac{1}{(n-2)2}$. This property is demonstrated in the graph below.

```r
sim <- rexp(10000,2)
sim <- sim[order(sim[])]
distance <- c()
for (i in 1:9999) {
  distance[i] <- sim[i+1] - sim[i]
}
plot(seq(1,9999), distance,
     main = "Distance Between X_(i) and X_(i+1) under X_1 ... X_n ~ Exp(2)",
     xlab = "X_(i)", ylab = "Distance from X_(i+1)", type = "l")
```



By extension of the above, $E[X_{(i+1)} - X_{(i)}] = \frac{1}{2(n-i)}$ and thus we can find $E[X_{(n)}]$ as the sum of the distances from $E[X_{(0)}] = 0$ to $E[X_{(n-1)}]$

$$E[X_{(1)}] = \int_0^\infty 2xne^{-2xn} \qquad\qquad E[X_{(n)}] = \sum_{i=0}^{n-1} \frac{1}{2(n-i)}$$

$$= E[x|\lambda = 2n] \qquad\qquad = \sum_{i=0}^{n} \frac{1}{2i}$$

$$= \frac{1}{2n}$$

which gives us

$$E[X_{(n)} - X_{(1)}] = \sum_{i=0}^{n} \frac{1}{2i} - \frac{1}{2n} = \sum_{i=0}^{n-1} \frac{1}{2i}$$

Problem Three

If $X_{(r)}$ be the r^{th} order statistic of a random sample of size n from a cdf F_X. Compute $P(X_{(r)} \leq t)$.

Solution: Note that $X_{(r)}$ is less than or equal to t if and only if r or more of the n observations are less than t. Then, if we set R to be the number of successes, we can write $P(X_{(r)} \leq t)$ as

$$P(X_{(r)} \leq t) = P(R = r) + P(R = r + 1) + ... + P(R = n) = \sum_{k=r}^{n} P(R = k)$$

By expressing this as the probability of successes, $P(R = k)$ is distributed binomial, and so we have

$$P(X_{(r)} \leq t) = \sum_{k=r}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k}$$

Problem Four

Show that $D_n^- = \max\{\max_{1\leq i\leq n}[F_0(X_{(i)}) - \frac{i-1}{n}], 0\}$

Solution: We know by defition that $D_n^- = \sup\limits_x [F_n(x) - F_0(x)]$ and $F_n(x) = \frac{i}{n}$.

$$\begin{aligned} D_n^- &= \sup_x \left[F_0(x) - \frac{i}{n}\right] \\ &= \max_{0\leq i\leq n}\left[\sup_{X_{(i)}\leq x<X_{(i+1)}}\left[F_0(x) - \frac{i}{n}\right]\right] \end{aligned}$$

The suprememum of the difference is whatever value maximized $F_0(x)$, that is $\sup\limits_{X_{(i)}\leq x<X_{(i+1)}}\left[F_0(x) - \frac{i}{n}\right] = \sup\limits_{X_{(i)}\leq x<X_{(i+1)}}[F_0(x)] - \frac{i}{n}$

$$\begin{aligned} D_n^- &= \max_{0\leq i\leq n}\left[\sup_{X_{(i)}\leq x<X_{(i+1)}}[F_0(x)] - \frac{i}{n}\right] \\ &= \max_{0\leq i\leq n}\left[F_0(X_{(i+1)}) - \frac{i}{n}\right] \\ &= \max_{0\leq i\leq n}\left[F_0(X_{(i)}) - \frac{i-1}{n}\right] \\ &= \max\left\{\max_{1\leq i\leq n}\left[F_0(X_{(i)}) - \frac{i-1}{n}\right], 0\right\} \end{aligned}$$

Note that $\left[F_0(X_{(i+1)}) - \frac{i}{n}\right]$ is rewritten in terms of i and $i-1$ so as to make it similar to D_n^+.

Thus we have our desired result.

Problem Five

In a psychological experiment, the research question of interest is whether a rat learned its way through a maze during 64 trials. Suppose the time-ordered observations on number of correct choices by the rat on each trial are as follows:

0	1	2	1	1	2	3	2
2	2	1	1	3	2	1	2
1	2	2	1	1	2	2	1
4	3	1	2	2	1	2	2
2	2	3	2	2	3	4	3
2	3	3	2	3	3	2	3
3	2	3	4	3	3	4	2
3	3	4	3	4	4	4	4

Test these data for randomness, using the dichotomizing criterion that 0, 1, or 2 correct choices indicate no learning, while 3 or 4 correct indicate some learning.

Solution: First, we enter the data.

```
rat.learning <- c(0,1,2,1,1,2,3,2,2,2,1,1,3,2,1,2,
                  1,2,2,1,1,2,2,1,4,3,1,2,2,1,2,2,
                  2,2,3,2,2,3,4,3,2,3,3,2,3,3,2,3,
                  3,2,3,4,3,3,4,2,3,3,4,3,4,4,4,4)
```

So, we have the following hypotheses:

$$H_0 : \text{The data is randomly sampled.}$$
$$H_1 : \text{The data is not randomly sampled.}$$

We can use the Wald-Wolfowitz Test in R to determine whether this sampling is random or not by setting a threshold of 2.5 so that 0, 1, and 2 are considered below and 3 and 4 are considered above for purposes of the test usage. And now we run the test.

```
runs.test(rat.learning, threshold = 2.5, pvalue = "exact")
```

```
##
##  Runs Test
##
## data:  rat.learning
## statistic = -3.1576, runs = 20, n1 = 27, n2 = 37, n = 64, p-value
## = 0.002298
## alternative hypothesis: nonrandomness
```

```
runs.test(rat.learning, threshold = 2.5)
```

```
##
##  Runs Test
##
## data:  rat.learning
## statistic = -3.1576, runs = 20, n1 = 27, n2 = 37, n = 64, p-value
## = 0.001591
## alternative hypothesis: nonrandomness
```

In the first function usage, an exact pvalue is calculated and in the second, a normal approximation is used. With this in mind, we can reject the null hypothesis at the 99% confidence level, and conclude that the sampling is likely not random.

Problem Six

A group of four coins is tossed 160 times, and the following data are otained:

Number of Heads	0	1	2	3	4
Frequency	16	48	55	33	8

Do you think the four coins are balanced?

Solution: If the four coins are balanced, then $X \sim Bin(4, .5)$. That is,

$$\begin{aligned} H_0 &: X \sim Bin(4, .5) \\ H_a &: X \nsim Bin(4, .5) \end{aligned}$$

We can test this with the Chi-square Goodness of Fit test which is done below.

```
observed <- c(16, 48, 55, 33, 8)
expected <- dbinom(seq(0,4), 4, .5)
chisq.test(x = observed, p = expected)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 7.2417, df = 4, p-value = 0.1237
```

With an exact p-value of 0.1237, at a confidence level of 95% we fail to reject the null hypothesis that the coins are balanced and so do not have enough evidence to suggest that the coins are not balanced.

Problem Seven

In a vibration study, a random sample of 15 airplane components were subjected to severe vibrations until they showed structural failures. The data given are failures in minutes.

1.6   10.3   3.5   13.5   18.4   7.7   24.3   10.7   8.4   4.9   7.9   12.0   16.2   6.8   14.7

(a) Test the null hypothesis that these observations can be regarded as sample from exponential population with density function $f(x) = \frac{1}{10}e^{-\frac{x}{10}}$

(b) Plot the empirical cdf of the data and also plot the empirical cdf of the density under H_0.

Solution to (a): We are specifically testing

$$\begin{aligned} H_0 &: X \sim Exp(\lambda = 1/10) \\ H_1 &: X \not\sim Exp(\lambda = 1/10) \end{aligned}$$

Because we are testing whether or not is is distributed exponentially which is continuous, the Kolgomorov-Smirnoff test is applicable and is computed below.

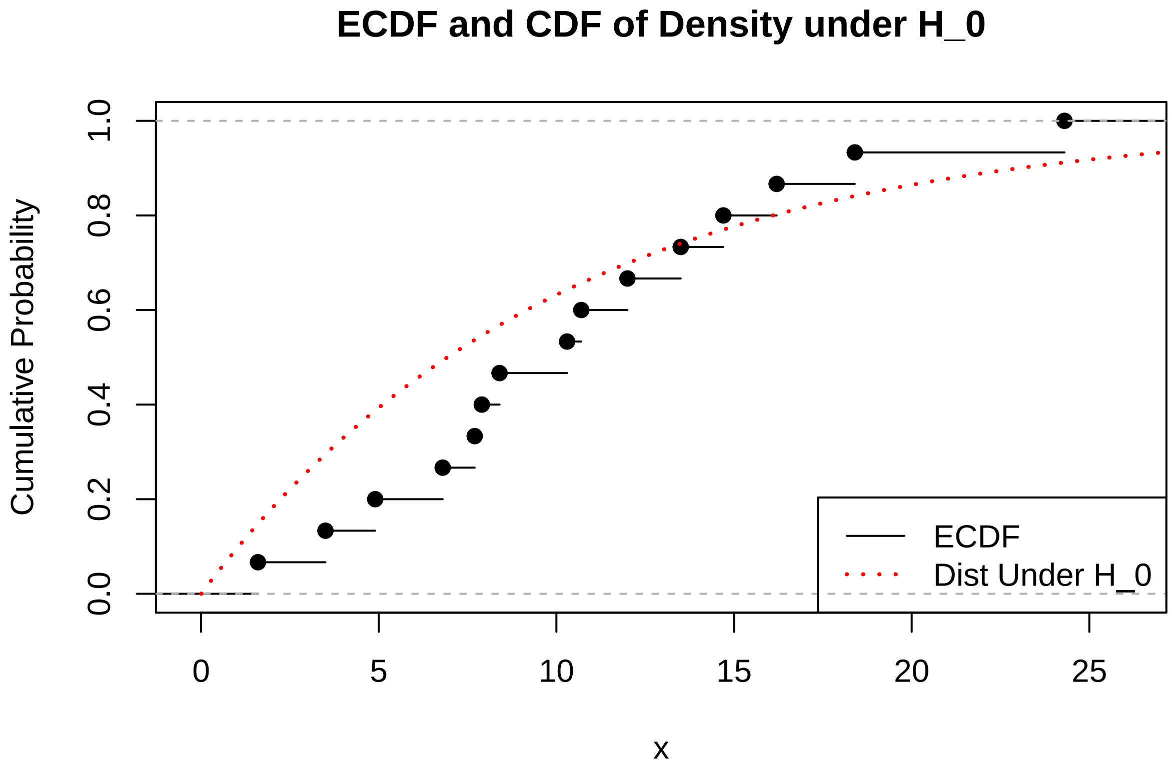
```
observed <- c(1.6, 10.3, 3.5, 13.5, 18.4, 7.7, 24.3, 10.7,
              8.4, 4.9, 7.9, 12.0, 16.2, 6.8, 14.7)
ks.test(observed, "pexp", rate = .1, exact = TRUE)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  observed
## D = 0.29338, p-value = 0.1224
## alternative hypothesis: two-sided
```

With an exact p-value of 0.1224, we fail to reject the null hypothesis, and therefore do not enough have enough evidence to suggest that the sample is not distributed exponentially with parameter $\lambda = 1/10$.

Solution to (b):

```
plot(ecdf(observed), main = "ECDF and CDF of Density under H_0",
     ylab = "Cumulative Probability")
exponential.cum <- function(x, lambda) {
  return(1 - exp(-lambda*x))
}
lines(seq(0,30,by = .05),
      exponential.cum(seq(0,30,by = .05), .1),
      lty = 3, col = "red", lwd = 2)
legend("bottomright", c("ECDF", "Dist Under H_0"), lty = c(1,3),
       col = c("black", "red"), lwd = c(1,2))
```


ECDF and CDF of Density under H_0
Cumulative Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
x
ECDF
Dist Under H_0

Problem Eight

According to test theory, scores on a certain IQ test are normally distributed. This test was given to 18 girls of similar age and their scores were

114	81	87	114	114	87	111	89	93
108	99	93	100	95	93	95	106	108

Test the null hypothesis that these scores are normally distributed with

(a) Unspecified mean and unspecified variance.
(b) Mean 100 and variance 100
(c) Plot the QQ plot for the data set and also comment on the QQ plot.

Solution (a): First, we enter the data.

```
IQ <- c(114,81,87,114,114,87,111,89,93,
        108,99,93,100,95,93,95,106,108)
```

And in this case our hypotheses are:

$$H_0 : \text{Sample IQ is Normal}$$
$$H_a : \text{Sample IQ is not Normal}$$

We can use the Anderson Darling test to determine if these are normally distributed with unknown parameters.

```
require(nortest)
```

```
## Loading required package: nortest
```

```
nortest::ad.test(IQ)
```

```
##
##  Anderson-Darling normality test
##
## data:  IQ
## A = 0.47032, p-value = 0.2169
```

With p-value of 0.2169, we fail to reject the null hypothesis and so do not have enough evidence to suggest that the sample IQ is not normal.

Solution (b):

$$H_0 : IQ \sim N(100, 10)$$
$$H_a : IQ \nsim N(100, 10)$$

We have two options to use in the case of this problem. We can use `goftests` variation of the `ad.test()` function, or we can use the `ks.test()`. I will use both for sake of rigor. Note though that the Anderson-Darling test here is preferred due to the presence of ties.

```
require(goftest)
```

```
## Loading required package: goftest
```

```
##
## Attaching package: 'goftest'
## The following objects are masked from 'package:nortest':
##
##     ad.test, cvm.test
```

```
goftest::ad.test(IQ, null = "pnorm", mean = 100, sd = 10)
```

```
##
##  Anderson-Darling test of goodness-of-fit
##  Null hypothesis: Normal distribution
##  with parameters mean = 100, sd = 10
##
## data:  IQ
## An = 0.60835, p-value = 0.6378
```

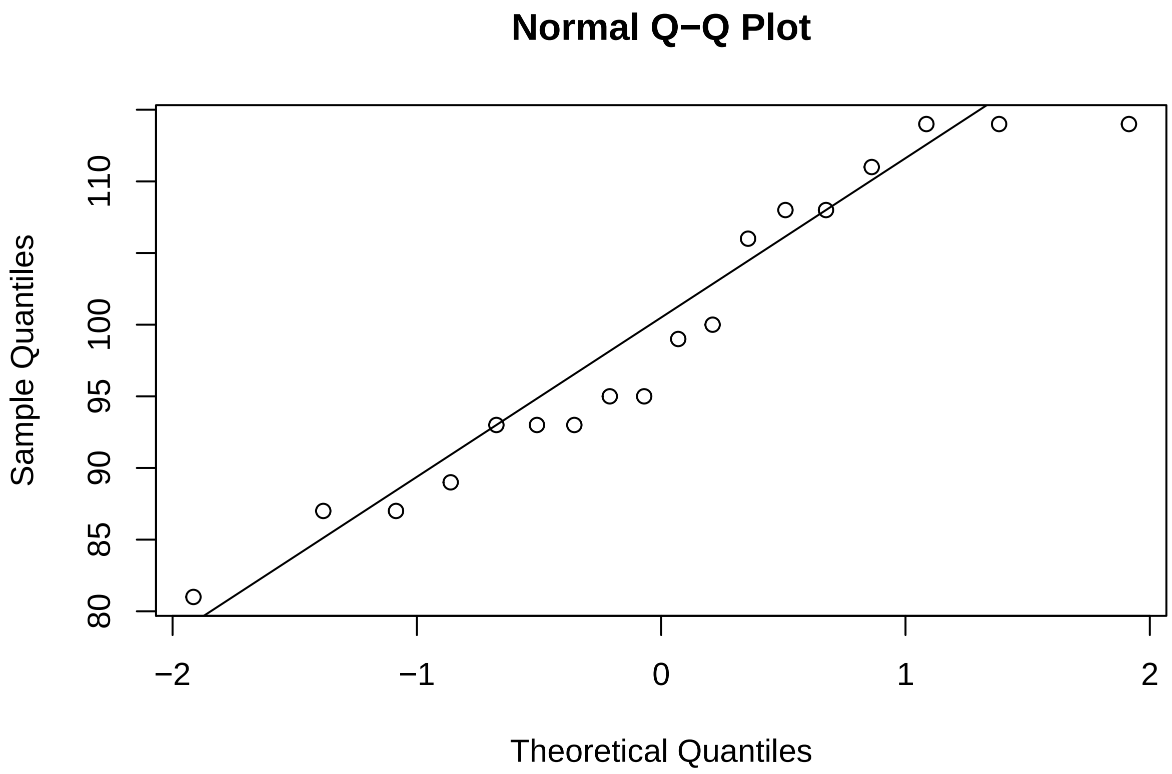
```
ks.test(IQ, "pnorm", mean = 100, sd = 10)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  IQ
## D = 0.19146, p-value = 0.5243
## alternative hypothesis: two-sided
```

In both cases, we fail to reject the null hypothesis and can reasonably assume that our sample IQs are normally distributed with $\mu = 100$ and $\sigma^2 = 100$.

Solution (c): We use the QQ Norm function because our CDF under H_0 is normal, and we showed earlier that we can indeed assume our sample comes from a normal distribution.

```
qqnorm(IQ); qqline(IQ)
```



There's definitely some deviations from the QQ Line in the plot, but not that are far enough to suggest that anything is off or that our data is not distributed normally.

Problem Nine

Find the exact probability distribution function of the Wilcoxon Signed rank test when $n = 4$.

Solution: To find the exact probability distribution function, we list all possible permutations of $n = 4$ of the ranks.

Possibilities				r_+
1	2	3	4	10
1	2	3	-4	6
1	2	-3	4	7
1	2	-3	-4	3
1	-2	3	4	8
1	-2	3	-4	4
1	-2	-3	4	5
1	-2	-3	-4	1
-1	2	3	4	9
-1	2	3	-4	5
-1	2	-3	4	6
-1	2	-3	-4	2
-1	-2	3	4	7
-1	-2	3	-4	3
-1	-2	-3	4	4
-1	-2	-3	-4	0

r_+	Occurrences	$P(R = r_+)$	$P(R <= r_+)$
0	1	0.0625	0.0625
1	1	0.0625	0.125
2	1	0.0625	0.1875
3	2	0.125	0.3125
4	2	0.125	0.4375
5	2	0.125	0.5625
6	2	0.125	0.6875
7	2	0.125	0.8125
8	1	0.0625	0.875
9	1	0.0625	0.9375
10	1	0.0625	1

The left table represents all possible permutations and the right table represents the probability density function and the cumulative density function for varying ranks.

Problem Ten

Let $X_1, X_2, ..., X_n$ be a random sample. Also $F_n(x)$ is the empirical cdf of X. Show that for

$$-\infty < x < y < \infty, \quad \operatorname{Cov}(F_n(x), F_n(y)) = \frac{F_X(x)(1 - F_X(y))}{n}.$$

Where $F_X(x)$ is cdf of X.

Solution: The Covariance is defined as

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

So in this case, we have

$$Cov(F_n(x), F_n(y)) = E[F_n(x)F_n(y)] - E[F_n(x)]E[F_n(y)]$$

We find $E[F_n(x)F_n(y)]$ and $E[F_n(x)]$ and $E[F_n(y)]$ below.

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

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

$$\begin{aligned} E[F_n(x)F_n(y)] &= E[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x)\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(X_j \leq y)] \\ &= \frac{1}{n^2}E[\sum_{i=1}^{n} \mathbb{1}(X_i \leq x)\sum_{j=1}^{n} \mathbb{1}(X_j \leq y)] \end{aligned}$$

To further compute this, note that in the case of $i = j$, $E[\sum_{i=1}^{n} \mathbb{1}(X_i \leq x) \sum_{j=i}^{n} \mathbb{1}(X_i \leq y)] = F_X(x)$ due to $x < y$. For all other cases, the expected product of two independent variables is the product of the expectations of the variables, i.e. $E[X_1X_2] = E[X_1]E[X_2]$.

$$
\begin{aligned}
E[F_n(x)F_n(y)] &= \frac{1}{n^2}E[\sum_{i=1}^{n}\mathbb{1}(X_i \leq x)\sum_{j=1}^{n}\mathbb{1}(X_j \leq y)] \\
&= \frac{1}{n^2}\left[nF_X(x) + n(n-1)E[F_n(x)]E[F_n(y)]\right] \\
&= \frac{1}{n^2}\left[nF_X(x) + n(n-1)F_X(x)F_X(y)\right] \\
&= \frac{1}{n}\left[F_X(x) + nF_X(x)F_X(y) - F_X(x)F_X(y)\right] \\
Cov(F_n(x), F_n(y)) &= \frac{1}{n}\left[F_X(x) + nF_X(x)F_X(y) - F_X(x)F_X(y)\right] - F_X(x)F_X(y) \\
&= \frac{1}{n}\left[F_X(x) + nF_X(x)F_X(y) - F_X(x)F_X(y)\right] - \frac{1}{n}nF_X(x)F_X(y) \\
&= \frac{1}{n}\left[F_X(x) - F_X(x)F_X(y)\right] \\
&= \frac{F_X(x)(1 - F_X(y))}{n}
\end{aligned}
$$

Problem Eleven

A sample of three girls and five boys are given instructions on how to complete a certain task. Then they are asked to perform the task over and over until they complete it correctly. The numbers of repetitions necessary for correct completion are

Girls	1	2	5		
Boys	4	8	9	10	12

Find the P value for the alternative that on the average the girls learn the task faster than the boys.

Solution: We are testing whether or not girls learn faster than boys, or specifically, if the amount of repetitions necessary for girls is on average less than the amount of repetitions necessary for boys. That is,

$$H_0 : \mu_{girls} = \mu_{boys}$$
$$H_a : \mu_{girls} < \mu_{boys}$$

This can be tested with the Mann-Whitney U (or Wilcoxon) test.

```r
girls <- c(1,2,5)
boys  <- c(4,8,9,10.12)
wilcox.test(girls,boys, alternative = "l")
```

```
##
##  Wilcoxon rank sum test
##
## data:  girls and boys
## W = 1, p-value = 0.05714
## alternative hypothesis: true location shift is less than 0
```

With a p-value of .05714, we are barely outside of the rejection region. With 95% confidence, we fail to reject the null hypothesis and so there is not enough evidence to suggest that girls learn faster than boys. That being said, if we were to lower our confidence to 94%, we would reject the null hypothesis. That being said, I will opt for the 95% confidence level and fail to reject.

Problem Twelve

A researcher is interested in learning if a new drug is better than a placebo in treating a certain disease. Because of the nature of the disease, only a limited number of patients can be found. Out of these, five are randomly assigned to the placebo and five to the new drug. Suppose that the concentration of a certain chemical in blood is measured and smaller measurements are better and the data are

Drug	3.2	2.1	2.3	1.2	1.5
Placebo	3.4	3.5	4.1	1.7	2.1

Perform a hypothesis test to see if the drug is effective. Compute exact and approximate p-values of the test.

Solution: We are testing if the concentration of a chemical in the drug group is smaller than the concentration of a chemical in the placebo group, that is

$$H_0 : \mu_{drug} = \mu_{placebo}$$
$$H_a : \mu_{drug} < \mu_{placebo}$$

Normally, we would need to define a modified Mann Whitney U, but R can do this for us.

```r
drug <- c(3.2, 2.1, 2.3, 1.2, 1.5)
placebo <- c(3.4, 3.5, 4.1, 1.7, 2.1)
wilcox.test(drug, placebo, alternative = "l")
```

```
## Warning in wilcox.test.default(drug, placebo, alternative = "l"): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  drug and placebo
## W = 5.5, p-value = 0.08661
## alternative hypothesis: true location shift is less than 0
```

```r
kruskal.test(drug, placebo)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  drug and placebo
## Kruskal-Wallis chi-squared = 4, df = 4, p-value = 0.406
```

Because of the ties, an exact p-value can't be computed, so R uses the normal approximation. That being said, with a p-value of 0.08661 and at the 95% confidence level, we fail to reject the null hypothesis and thefore do not have enough evidence to suggest that the drug is more effective than the placebo.

The Kruskal-Wallis test is also utilized which does not necessarily assume ties, so it offers an exact p-value that also leads us to reject the null hypothesis. That being said, the Kruskal-Wallis test does not follow under our alternative, because it is a two-tailed test of location. With this p-value though, we are led to believe that the locations of the samples are the same and so by extension, the drug could not be more effective than the placebo.

Problem Thirteen

Let X_1, X_2 and Y_1, Y_2 be random samples from two populations. Derive the probability mass function of R_X, the sum of the X ranks, under $H_0 : F_X(x) = F_Y(y)$. Determine whether this distribution is symmetric and if so, identify the point of symmetry. Calculate the mean and variance of R_X.

Solution: Similar to problem nine, we can find the probability mass function by checking all possible permutations.

Possibilities				R_x
X_1	X_2	Y_1	Y_2	
1	2	3	4	3
1	2	4	3	3
1	3	2	4	4
1	3	4	2	4
1	4	2	3	5
1	4	3	2	5
2	1	3	4	3
2	1	4	3	3
2	3	1	4	5
2	3	4	1	5
2	4	1	3	6
2	4	3	1	6
3	1	2	4	4
3	1	4	2	4
3	2	1	4	5
3	2	4	1	5
3	4	1	3	7
3	4	3	1	7
4	1	2	3	5
4	1	3	2	5
4	2	1	3	6
4	2	3	1	6
4	3	1	4	7
4	3	4	1	7

R_X	Occurences	$P(R = R_X)$	$P(R <= R_X)$
3	4	1/6	1/6
4	4	1/6	1/3
5	8	1/3	2/3
6	4	1/6	5/6
7	4	1/6	1

The left table is all of the permutations and the right table has the probability mass function and the cumulative mass function. This is a symmetrical distribution and the point of symmetry is at $R_x = 5$.

$$\begin{aligned}
E[R_X] &= \sum_{i=1}^{n} R_{X,i} * P(R = R_{X,i}) \\
&= 3 * \frac{1}{6} + 4 * \frac{1}{6} + 5 * \frac{1}{3} + 6 * \frac{1}{6} + 7 * \frac{1}{6} \\
&= \frac{1}{2} + \frac{2}{3} + \frac{5}{3} + 1 + \frac{7}{6} \\
&= 5 Var[R_X] \qquad &= \sum_{i=1}^{n} P(R = R_X)(R_{X,i} - E[R_X]) \\
&= \frac{1}{6}(-2)^2 + \frac{1}{6}(-1)^2 + \frac{1}{3}0 + \frac{1}{6}1^2 + \frac{1}{6}2^2 \\
&= \frac{4}{6} + \frac{1}{6} + \frac{1}{6} + \frac{4}{6} = 1\frac{2}{3}
\end{aligned}$$

Problem Fourteen

Giambra and Quilter (1989) performed a study to investigate gender and age difference in ability to sustain attention when given Mackworth's clock-test. This clock is metal with a plain white face and a black pointer that moves around the face in 100 discrete steps of 36 each. During the test period, the pointer made 23 doule jumps, defined as moving twice the normal distance or 7.2 in the same time period, at random and irregular intervals. Subjects were told that doule jumps would occur and asked to signal their recognition of occurrence by pressing a button. Scores were the numer of correct recognitions of the double jumps. The scores below are for 10 men aged 18 – 29 and 10 men aged 50 – 59. Determine whether median number of correct scores is larger for young men than for older men.

Age 18 – 29	11	13	15	15	18	19	20	21	21	22
Age 50 – 59	8	9	10	11	12	13	5	17	19	23

Solution: We are testing whether or not the younger men have have median number of scores larger than older man, that is

$$H_0 : M_{18--29} = M_{50--59} H_a \qquad : M_{18--29} > M_{50--59}$$

This can be solved very easily with the sign test.

```
require(BSDA)
```

```
## Loading required package: BSDA

## Loading required package: lattice

##
## Attaching package: 'BSDA'

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

```
young <- c(11,13,15,15,18,19,20,21,21,22)
old   <- c(8,9,10,11,12,13,5,17,19,23)
SIGN.test(young, old, alternative = "g")
```

```
##
##  Dependent-samples Sign-Test
##
## data:  young and old
## S = 9, p-value = 0.01074
## alternative hypothesis: true median difference is greater than 0
## 95 percent confidence interval:
##  2.893333      Inf
## sample estimates:
## median of x-y
##             4
##
## Achieved and Interpolated Confidence Intervals:
##
##                     Conf.Level L.E.pt U.E.pt
## Lower Achieved CI     0.9453 3.0000    Inf
## Interpolated CI       0.9500 2.8933    Inf
## Upper Achieved CI     0.9893 2.0000    Inf
```

With a p-value of .01074, we can reject the null hypothesis that the medians are equal and favor the alternative that young men have a higher median score on the clock test.

Problem Fifteen

A self-concept test was given to a random sample consisting of six normal subjects and three sujects under psychiatric care. Higher scores indicate more self-esteem. The data are as follows:

Normal	62	68	78	92	53	81
Psychiatric	54	70	79			

Find a p-value relevant to the alternative that psychiatric patients have lower-esteem than normal patients.

Solution: We are testing whether the psychiatric patients have lower scores than normal subjects, that is

$$H_0 : \mu_{normal} = \mu_{psychiatric}$$
$$H_a : \mu_{normal} > \mu_{psychiatric}$$

We can use the Mann-Whitney U test to determine a p-value.

```r
normal <- c(62,68,78,92,53,81)
psychiatric <- c(54,70,79)
wilcox.test(normal, psychiatric, alternative = "g")
```

```
##
##  Wilcoxon rank sum test
##
## data:  normal and psychiatric
## W = 10, p-value = 0.4524
## alternative hypothesis: true location shift is greater than 0
```

With a p-value of 0.4524, we fail to reject the null hypothesis that the psychiatric scores are the same as the scores of the normal subjects. That is, we do not have sufficient evidence to suggest that the psychiatric patients have lower self esteem than the normal patients.