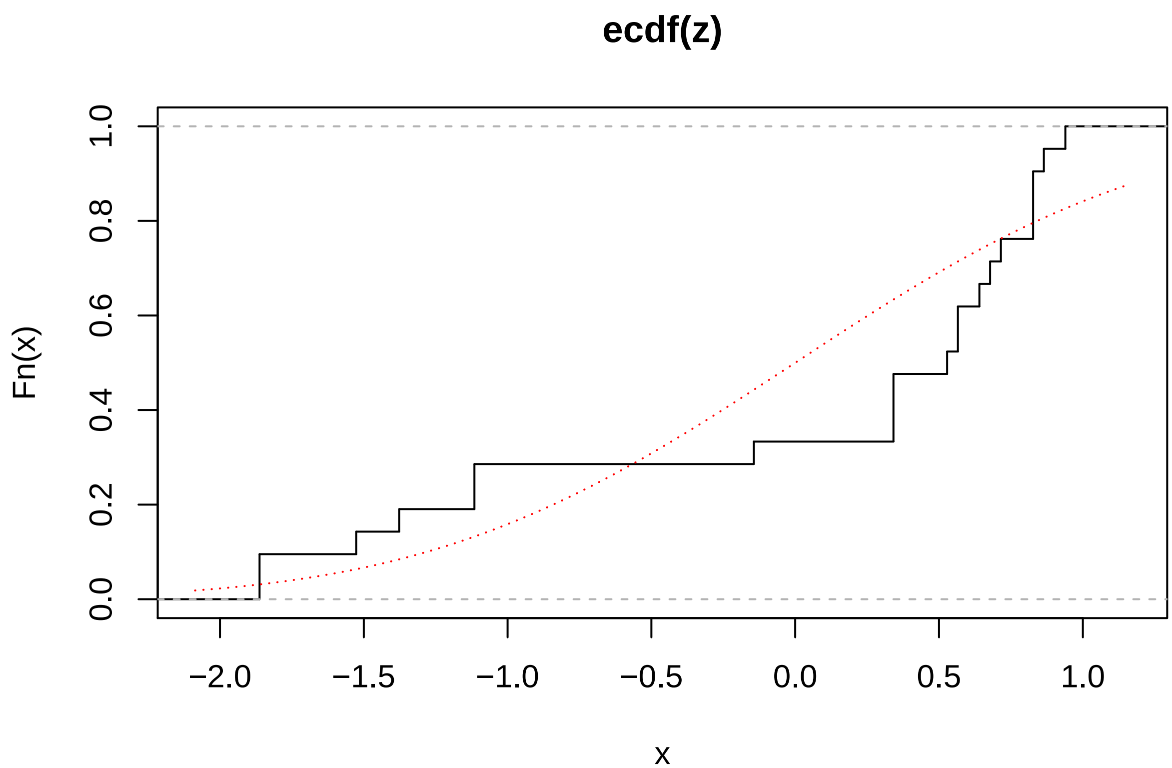
Homework 2 STAT 4500

Cody Frisby

4.1

First we take a look at the empirical CDF of our data after being normalized with the normal CDF overlayed in red:



At first glance, this does not appear to take the shape of the normal distribution CDF.

A table of our observed data and subsequent calculations for our test is displayed for reference (having removed repeated values of x):

	x	z	Pz	Sx	Diff	Diff.1
1	13	-1.8623017	0.0312803	0.0952381	-0.0639578	0.0312803
3	22	-1.5261269	0.0634891	0.1428571	-0.0793680	-0.0317490
4	26	-1.3767158	0.0843001	0.1904762	-0.1061761	-0.0585571
5	33	-1.1152465	0.1323724	0.2857143	-0.1533419	-0.0581038
7	59	-0.1440749	0.4427207	0.3333333	0.1093873	0.1570064
8	72	0.3415109	0.6336405	0.4761905	0.1574500	0.3003072
11	77	0.5282747	0.7013456	0.5238095	0.1775361	0.2251552
12	78	0.5656274	0.7141765	0.6190476	0.0951288	0.1903669
14	80	0.6403329	0.7390219	0.6666667	0.0723553	0.1199743
15	81	0.6776857	0.7510145	0.7142857	0.0367288	0.0843478
16	82	0.7150385	0.7627074	0.7619048	0.0008026	0.0484216
17	85	0.8270967	0.7959089	0.9047619	-0.1088530	0.0340041

	x	z	Pz	Sx	Diff	Diff.1
20	86	0.8644495	0.8063295	0.9523810	-0.1460515	-0.0984324
21	88	0.9391550	0.8261744	1.0000000	-0.1738256	-0.1262065

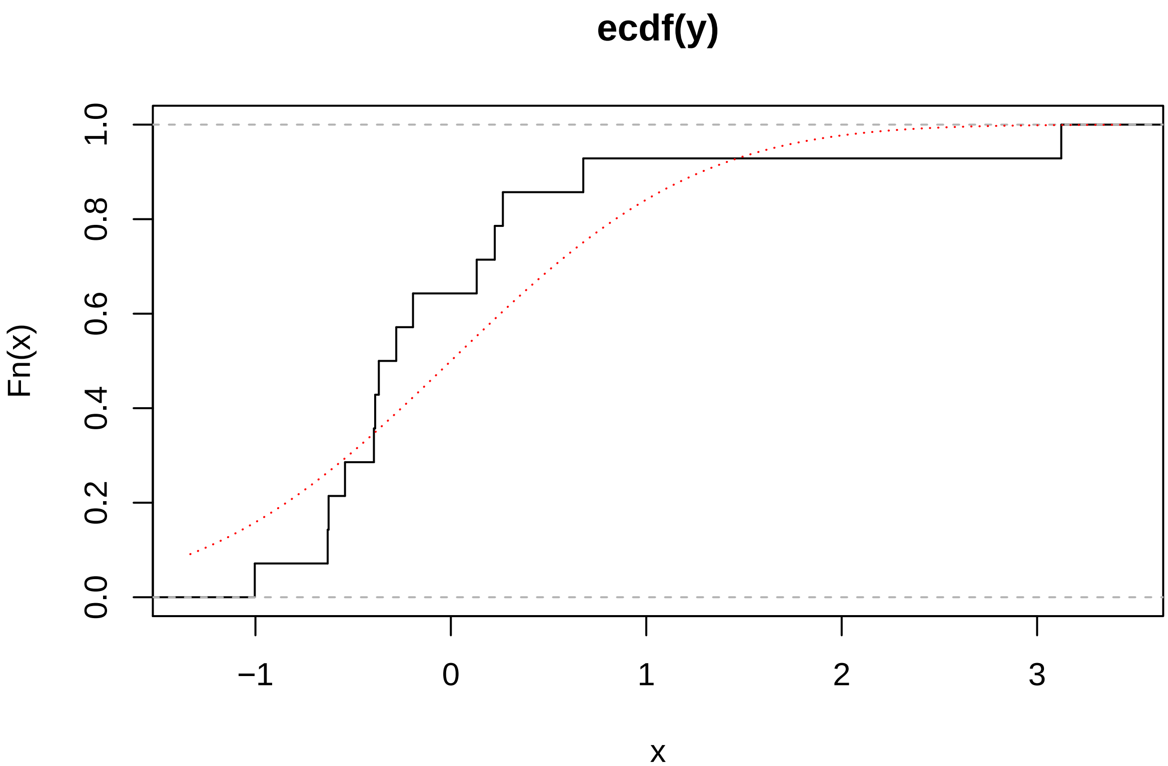
and our test statistic is the largest of the values of the differences, in this case it is 0.3003072 which under the null hypothesis is an improbable value so we reject the null hypothesis that the data is normally distributed. This same statistic, using Lilliefors' test to compute p-value, gives us $p = 2.8056044 \times 10^{-5}$.

Similarly, using the Shapiro-Wilk test for normality, the computed p-value is 3.9674968×10^{-4}, we conclude that it is **improbable** that the time at death is normally distributed. Our conclusion is the same, but our p-value is much smaller using the Shapiro-Wilk test for normality.

4.3

First, I plot the empirical CDF of the observed data with a normal CDF on top of it, after having standardized the observed values with

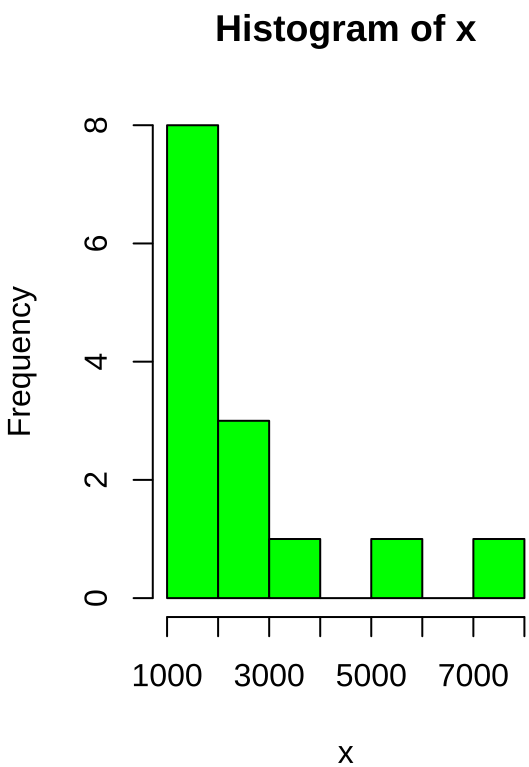
$$y_i = \frac{x_i - \bar{x}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}}$$

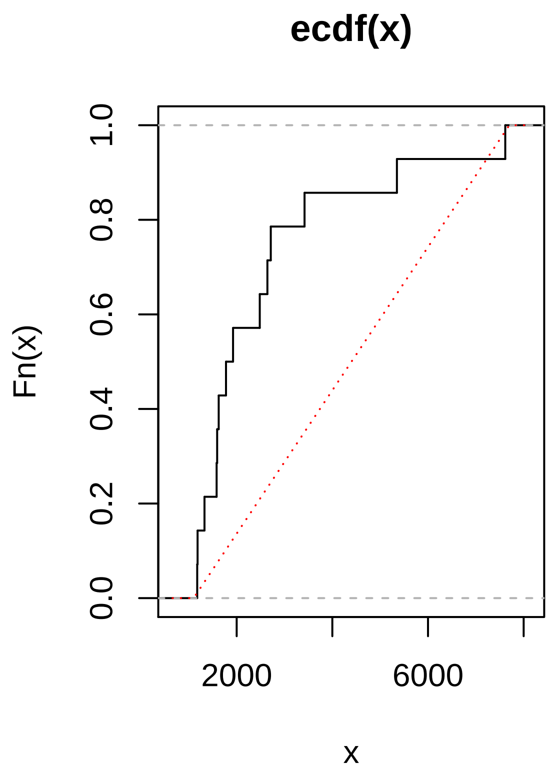


The normal CDF curve does not appear to be near "enough" to the empirical CDF. We can probably assume that the data is not normally distributed. Testing the normality hypothesis using the Shapiro-Wilk test we get a test stat of 0.7107667 and the p-value = 4.7869604×10^{-4}. We reject the null hypothesis that the data is normally distributed. The Lilliefor's test has a test statistic of 0.2522107 and a p-value of 0.0160568. We would come to the same conclusion using either test.

4.4

First, examining a historgram and a CDF plot we can start to get clues about the distribution of our data.

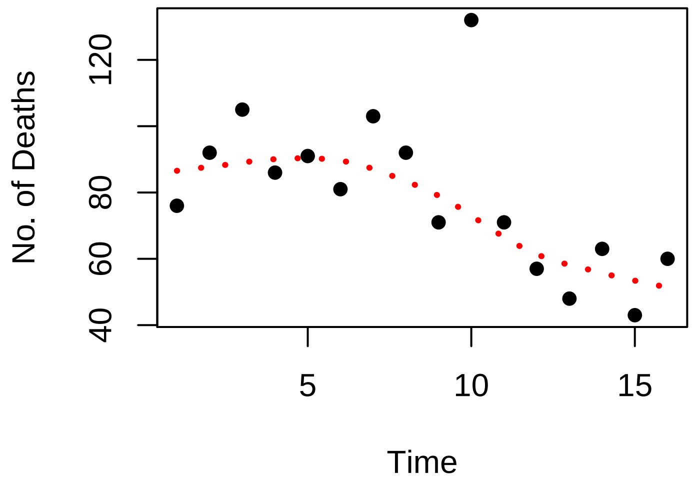




The histogram indicates a large scew to the left, meaning the data is likely to not be uniformally distirbuted. Running a test for distribution using the Kolmogorov method we find that our test statistic is 0.5413203 and the associated p-value is $p = 2.2773877 \times 10^{-4}$. There is evidence to reject that the data is from a uniform distribution, U(1100, 7700).

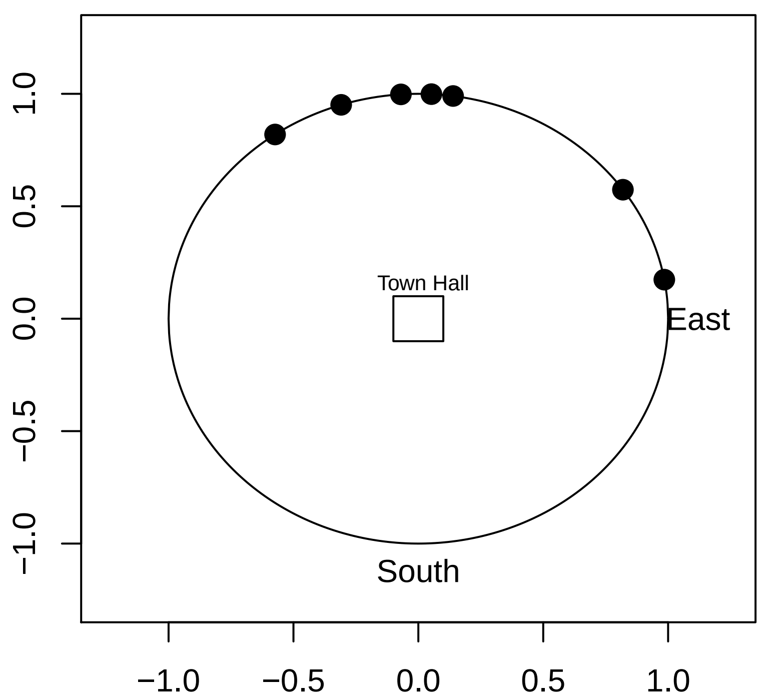
4.12

Here, we need to run a trend test. Using the Cox-Stuart test from the text, we find $p = 0.0703125$. There is weak evidence of a trend in the data. Visually, we can see this negative trend, albeit the test does show weak evidence, as does the plot. *Note: this is a two-sided p-value. If we suspected a negative or positve trend, we could run a one-sided test and our p value would be half 0.0703125.*



4.13

First I display a plot of the data where the points are the locations on the road of the observed car accidents using the unit circle.



There appears to be some clustering of car accidents. If there was no correlation to accidents and sections of road we would expect to see the points scattered randomly along the circle. Running the *Hodges-Ajne Test* where the null hypothesis is that the population of car accidents on this road are uniformly distributed along the circumfrence of a circle m = 0 since we can draw a line through the center where all of the points are on one side of.

$$m < t = \frac{n}{3} = \frac{7}{3} = 2.3333333$$

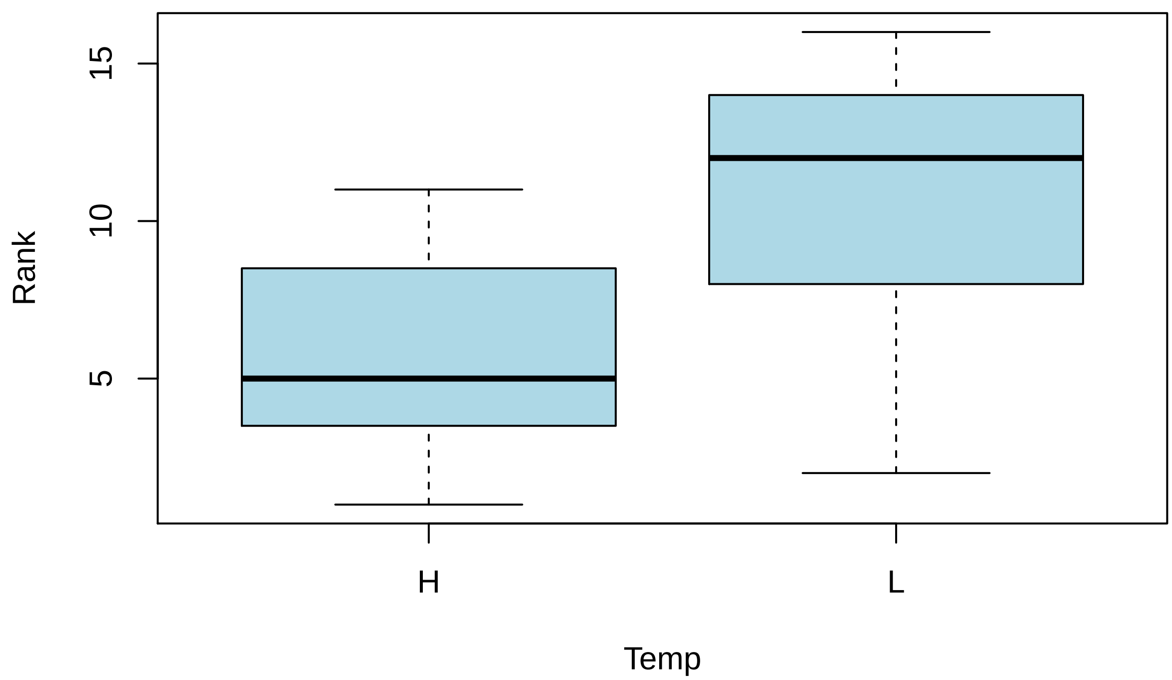
and the probability that $m = 0$ is 0.109375. Another appropriate test for this sample could be *The Range Teset* since it appears we do not have any outliers. Here $r = 2.0071286$ and the p-value is 0.0074382 using formula 4.9 in the text. We conclude the same as the *Hodges-Ajne Test*. There is evidence of clustering of car accidents.

4.17

Running a few lines of R code, finding n = 12, n1 = 6, n2 = 2 and n3 = 4 and then approximating the p-value using the asymtotic normal approach, we find the p-value is 0.0171708. This is the p-value for a two-tailed test. The book has an exact p-value for a one-sided test. If I multiply mine by 2 I get 0.0343417. This is very close to the exact p-value. At $\alpha = 0.05$ our conclusions would be the same, there is evidence of NON-randomness. The team tends to go on winning streaks and also losing streaks.

6.2

First, taking a look at a boxplt of the data for context:

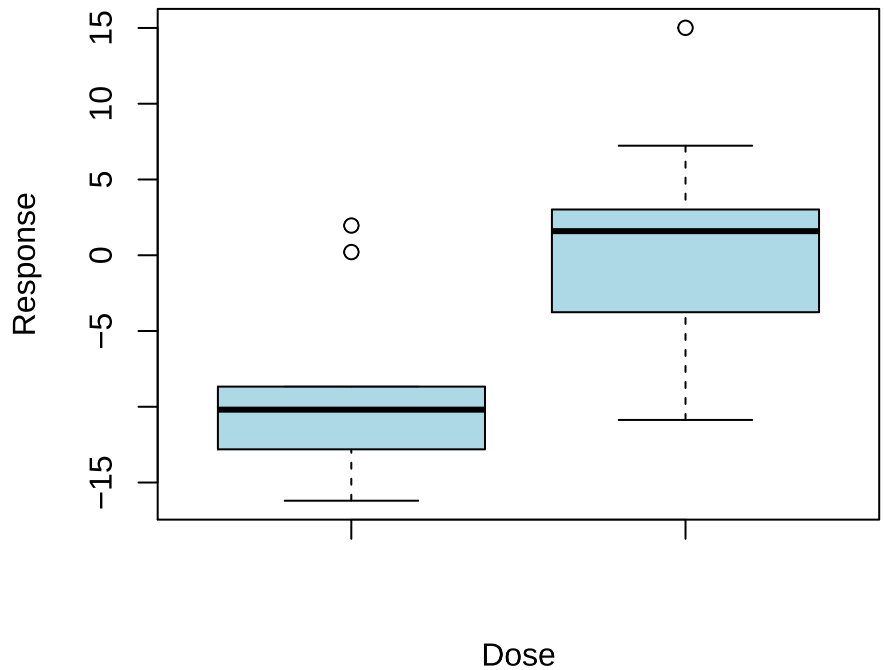


By visual inspection there appears to be a relationship between hardness and temperature, with the lower temperature creating harder specimins on average. Running a nonparametric test (WMW) we get $p = 0.0548951$. This is a "small" p-value albeit not "significant" at $\alpha = 0.05$. We would not reject at the traditional $\alpha = 0.05$ level but there is slight evidence of a relationship. Further investation is warranted.

6.4

There does not appear to be any strong evience of a difference between the two insulations. The firms' conclusions seem justified. A confidence interval contructing using the non-parametric method (WMW) contains zero [-0.2, 1.9]. Additionally, if we assume normality our conclusions do not change; 0 is also contained in the 95% confifence interval for the t test, [-0.1706533, 1.8289866].

6.10



Running a WMW test on the data and computing a confidence interval, we can see at significance level of 0.05, zero is NOT in our interval, [-17.42, -2.4]. There is evidence of a median response time difference between *dose I* and *dose II*. *Dose I* median response time appears to be lower.

6.11

Using these results

$$E(W) = \frac{n(m+n+1)}{2}$$

$$V(W) = \frac{mn(m+n+1)}{12}$$

$$Z = \frac{W_{mn} - E(W)}{\sqrt{V(W)}}$$

we compute the asymptotic z statistic $z = -2.8322445$ and the two-tailed probability of z is $p = 0.0046222$. We conclude that there is a difference between DMF scores for males and females among first year dental students.

There are quite a few ties. This could sway which test statistic is appropriate to use. However, the data does not appear to be grouped so I believe U_m is an appropriate statistic here, using mid-ranks for ties.

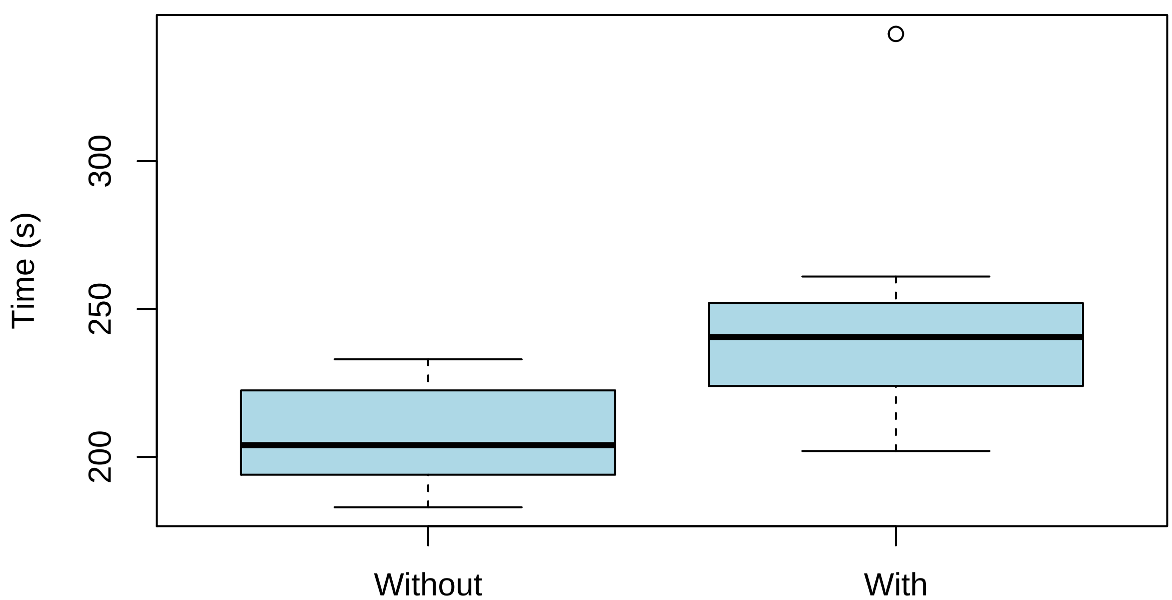
6.13

Running some R code, I rank the data according to the *Seigel-Tukey* test, after shifting the **male** values by the mean of x minus the mean of y, 58.375.

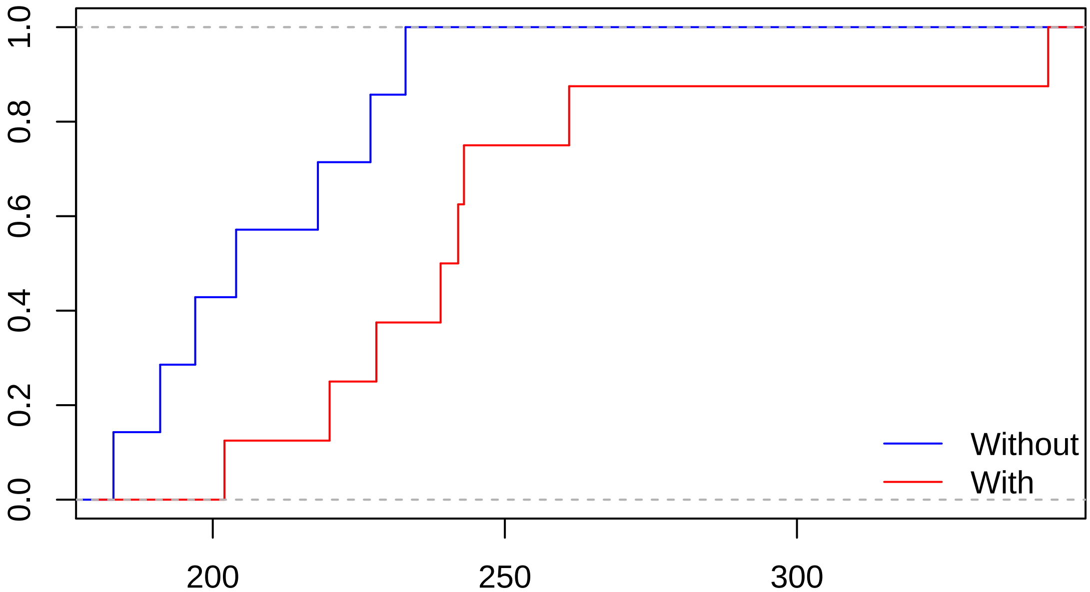
	Rank
47	1
105	4
118.625	5
126	8
141.625	9
142	12
158	13
168.625	16
171.625	17
172	20
173.625	19
197	18
209.625	15
213.625	14
220	11
225	10
230	7
238.625	6
262	3
270	2

We do NOT reject the null hypothesis that the variances are equal. The test statistic is 65 and the p-value is 0.2082877. Also $s_m = 101$.

6.15



By visual inspection, it appears there is a difference between the two samples. Plotting the CDF of both samples below we see that they are not the same. Additionally, since one sample is greater than the other at every step of the CDF, a one-sided test may be appropriate here. But, using tables provided, I report the two-tailed p-value. Having said that, there are other ways to run the test and get a p-value for a “greater” than alternative hypothesis and the results are not too different.

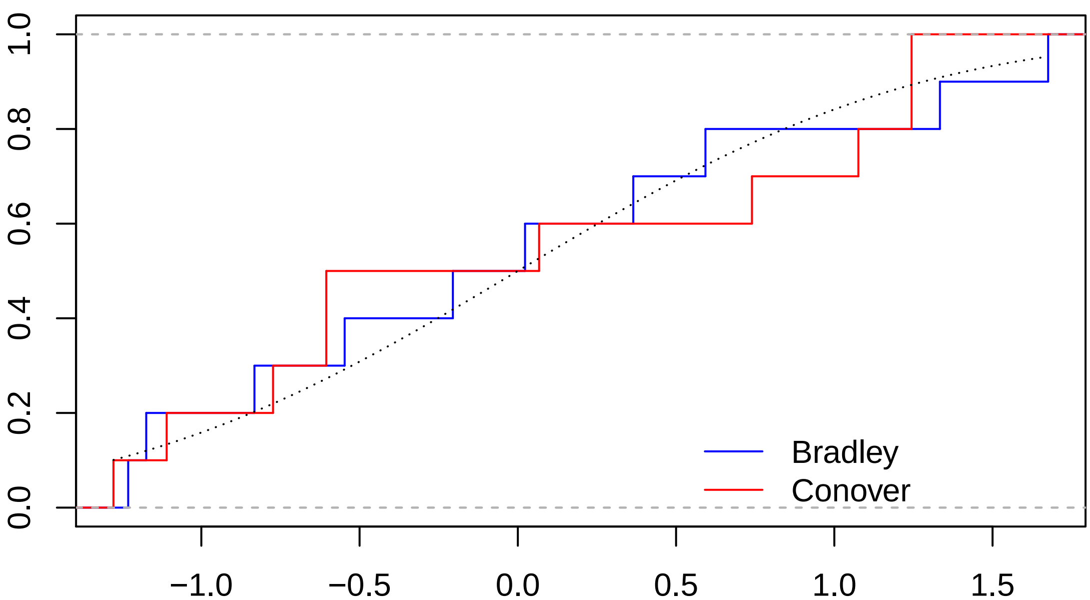


And using a Smirnov test where the psycologist believes these two samples are from different populations we find that

the probability that x lies above y is 0.0559441. We do not reject at $\alpha = 0.05$, although there is weak evidence of a difference between the two groups. Running a one tailed test we get 0.0541138.

6.16

I'd like to first plof the empirical CDF of both samples.



First off, it appears that both ecdfs follow the normal cdf very closely. This indicates that both samples are approximately normally distributed. If we can assume normality, we can run tests based on this assumption. But, for purposes of practice I run nonparametric tests on the samples.

- i) I run a wilcox test where the sum of the x observations is 120.5 and $m = 10$. There are ties among the data, which means our p value will not be exact when looking it up uisng the provided tables. However, the exact p-value and the table p-value do not differ greatly to cause any concerns. Computing a two-sided p value we get $p = 0.2474507$. We can not reject the null hypothesis, $H_0 : \eta_1 = \eta_2$.

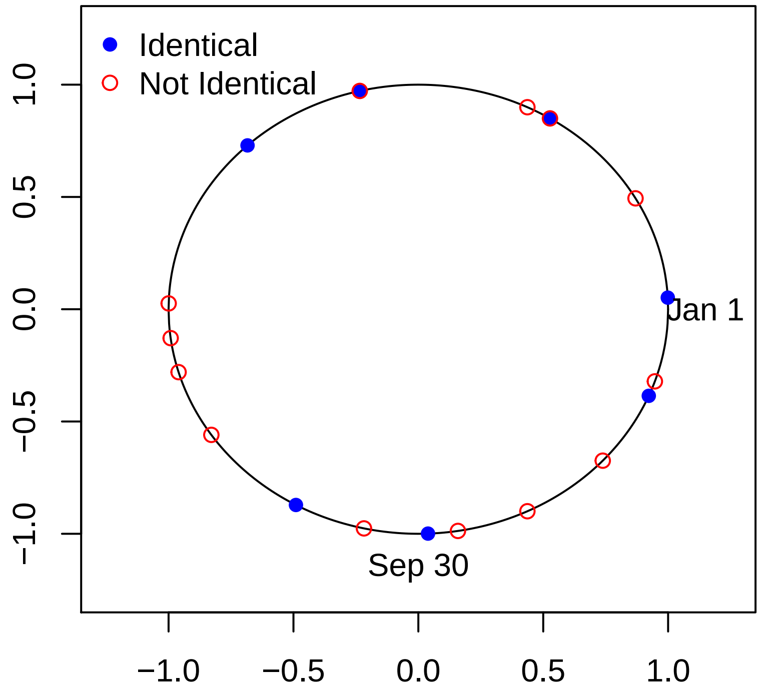
- ii) At first glance, it appears that there may be a difference in variance among the two samples. Running a non-parametric test for variance, here I use the *Siegel-Tukey* test, we find that there is evidence to reject the null hypothesis. Interestingly, both the F test and the Siegel-Tukey test both reject the null hypotheis. Using either test, we would conclude the same. **The variances are NOT equal**.

obs	Rank
13	1.0
14	4.0
17	5.0
18	8.0
20	10.5
21	14.5
25	18.5
29	19.0
31	16.5
32	12.5
35	10.0
41	7.0
45	6.0
58	3.0
64	2.0

- iii) There is not evidence to conclude that the distributions are different. Test statistic is 0.5. Looking up $mn(0.5) = 50$ on an appropriate table shows the p-value is 0.2. *Note, there are ties so this p-value is not exact.*

- iv) There is NOT evidence to conclude that data isn't normally distributed. Shapiro-Wilk, KS test, and Lilliefor's test all conclude the same.

6.17



Visually inspecting a plot of the data we can easily see that there isn't any clustering. Births appear to be occuring at random throughout the year. The number or runs around the circle is 12 (there are a few ties). The p-value is greater than 0.378. We do not reject.