Name: Rufus Petrie

HW4 Problem 1: El Paso Study

Part (a) Summary information:

	Sample Size	*Sample Median*	*Sample IQR*	*Outliers (if any)*	*~95 % CI for the Median*
Control	64	***53.405***	**12.935**	***13.33, 22.62, 26.08, 83.18***	**[*50.15, 58.27*]**
Past	16	***51.02***	**18.305**	***None***	**[*37.85, 61.98*]**
Current	19	**47.54**	**12.59**	**13.02, 14.86**	**[*38.09, 54.11*]**

Part (b) Rank sum test results:

```
RankSumTest[scores2, scores3,
 TwoSided → True,
 ExactDistribution → True]
```

First Sample	*Second Sample*	*Exact PValue for Two -Sided Test*
Control	Past	0.198016
Control	Current	0.00119794
Past	Current	0.271102

Part (c) Discussion:

From the results in part a, we see that the median scores for the finger-wrist tapping test seem to decrease with exposure to lead. That is, the control group, who had low lead on both occasions had the highest median. The past group, which had exposure in the past had the second highest median. The current group, which has had lead exposure recently has the lowest median. However, we should note that while the control group has a larger sample size, it also had the highest amount of outliers, which could be a topic for further discussion.

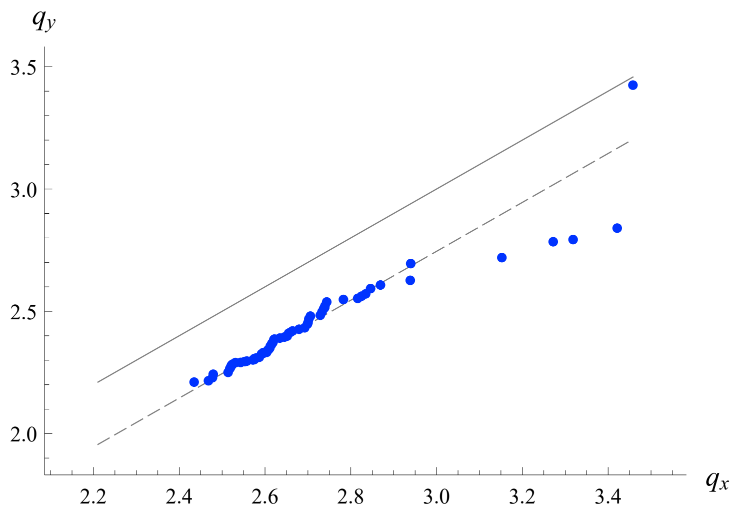
From the rank sum tests, we see that we can not reject the null hypothesis that the control and past groups have the same distribution. Likewise, we can not reject the null hypothesis that the past and current scores have the same distribution. However, we can reject the null hypothesis that the control and the current scores have the same distribution. That means it’s reasonable to conclude that the control scores and the current scores are systematically different in some way. This means that low-lead exposure may create a statistical difference in the test scores for people who have had little exposure to lead and people who have had exposure to lead recently.

HW4 Problem 2: Olympic marathon finishing times

```
RankSumTest[times1, times2,
 TwoSided → True,
 ExactDistribution → False]
```

```
Rank Sum      TwoSided    Sampling
Statistic:    P Value:    Distribution:
8512.5        0           Normal (μ=5752.5,σ=326.222)
```

```
QQPlot[times1, times2,
 MedianDifference → True]
```



```
Dashed Line:           y = x - 0.2547
HL Estimate of Δ:      0.2547
```

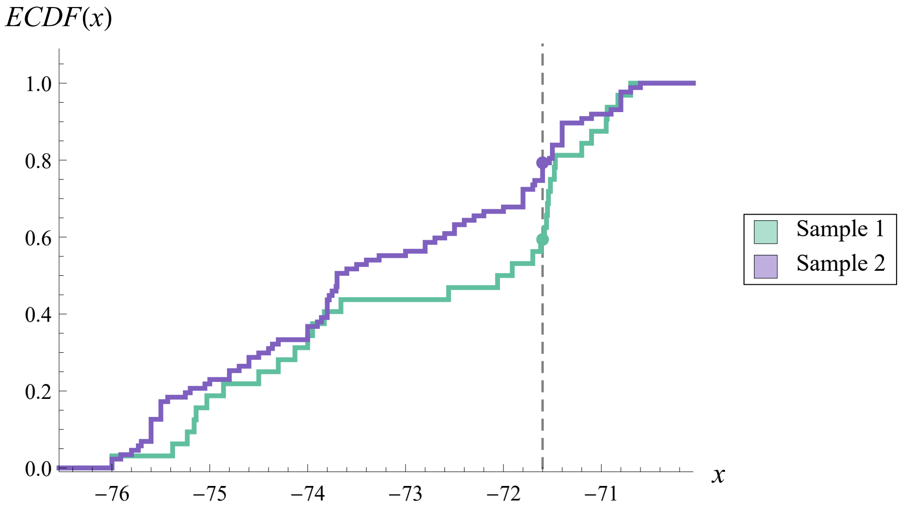
```
RankSumCI[times1, times2,
 ConfidenceLevel → 0.95,
 ExactDistribution → True]
```

```
Confidence     Confidence          Sampling
Level:         Interval:           Distribution:
0.950933       {0.2153, 0.2945}    Exact
```

Because the P-value of the rank-sum statistic is approximately 0, we can reject the null hypothesis that the finishing times for men and women have the same distribution. The Hodges-Lehman estimate of the shift parameter tells us that the distribution of the women’s finishing times is roughly the same as the men’s, but shifted by 0.2547 units. The rank-sum CI command tells us that the 95% confidence interval for the shift parameter is between 0.2153 and 0.2945. Combined, all of this data tells us that male and female olympic athletes have different finishing times for marathons. According to the Hodges-Lehman estimate, it appears that men are about fifteen minutes quicker than women. We can be 95% certain that men are somewhere between about 12 and 18 minutes quicker than women.

HW4 Problem 3: Earthquake locations

```
ECDFPlot[longitude1, longitude2,
 MaximumDifference → True]
```



```
Dashed Line:            x = -71.6
Maximum Difference:     0.199353
```

```
samples = {longitude1, longitude2};
Clear[ss]; Remove[f];
f[ss_] := SmirnovStatistic[First[ss], Last[ss]];
observed = f[samples] (* Observed Smirnov Statistic *)
```

```
0.199353
```

```
results = Table[f[RandomPartition[samples]], {1999}];
count = 1 + Length[Select[results, (# ≥ observed) &]]
estimate = N[count / 2000] (* Estimated P-value *)
```

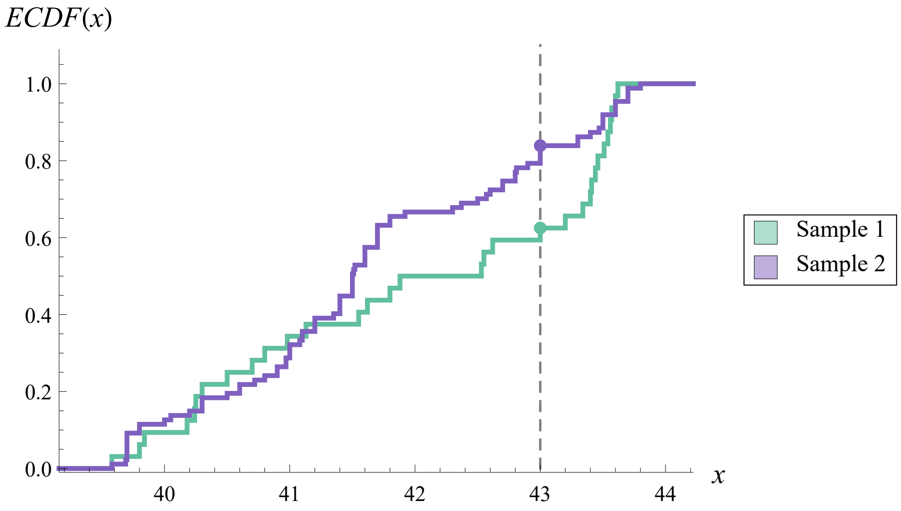
```
500
```

```
0.25
```

```
BinomialCI[count, 2000, ConfidenceLevel → 0.99]
(*P-value CI *)
```

```
{0.225432, 0.275772}
```

```
ECDFPlot[latitude1, latitude2,
 MaximumDifference → True]
```



```
Dashed Line:            x = 43.
Maximum Difference:     0.21408
```

```
samples = {latitude1, latitude2};
Clear[ss]; Remove[f];
f[ss_] := SmirnovStatistic[First[ss], Last[ss]];
observed = f[samples]
```

```
0.21408
```

```
results = Table[f[RandomPartition[samples]], {1999}];
count = 1 + Length[Select[results, (# ≥ observed) &]]
estimate = N[count / 2000]
```

```
369
```

```
0.1845
```

```
BinomialCI[count, 2000, ConfidenceLevel → 0.99]
```

```
{0.162684, 0.207838}
```

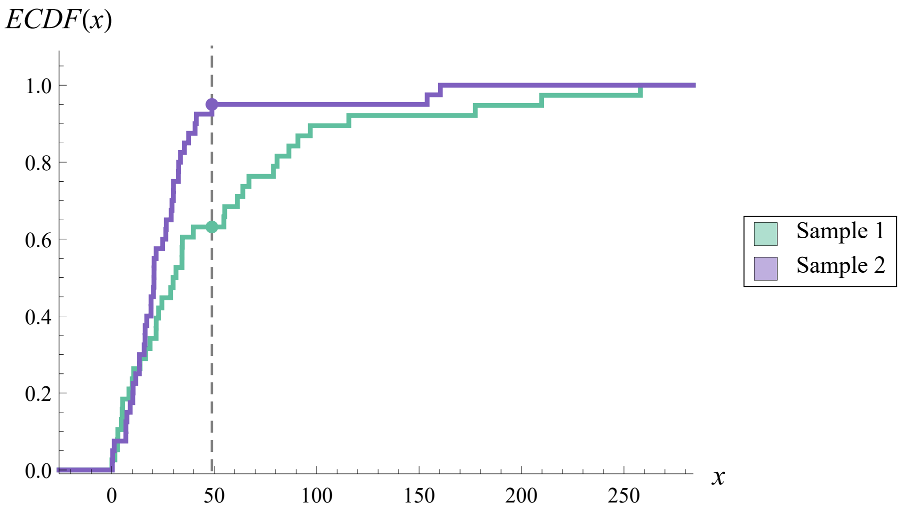
From the ECDFplots and Smirnov tests, it appears that latitude and longitude did not change significantly between the two time periods. The Smirnov test for the longitudes gave us a P-value of 0.25 with a 95% confidence interval of [0.225,0.275]. With such a high P-value, we can not reject the null hypothesis that variations in the longitude data were due to chance alone. Likewise, The Smirnov test produced a P-value of 0.18 for the latitudinal data with a 95% confidence interval of [0.162,0.207]. This also implies that we can’t reject the null hypothesis that variations in the latitudinal data were due to chance alone. While these statistical tests did not provide any compelling results, we can make some interesting observations by simply eye-balling the ECDF diagrams. Both had a relatively maximum difference. In particular, it appears that the data from sample 1 had far more longitude values below 71.6 and far more latitude values above 43. These observations may be useful for further research if it turns about that there was some variation in the location of earthquakes that did not have enough of an impact on these statistical tests to turn out a low P-value.

Problem 4: Radon exposure study

```
RankSumTest[radon1, radon2,
 TwoSided → True,
 ExactDistribution → True]
```

```
Rank Sum      TwoSided     Sampling
Statistic:    P Value:     Distribution:
1671.         0.0902649    Exact
```

```
ECDFPlot[radon1, radon2,
 MaximumDifference → True]
```



```
Dashed Line:              x = 48.87
Maximum Difference:       0.318421
```

```
samples = {radon1, radon2};
Clear[ss]; Remove[f];
f[ss_] := SmirnovStatistic[First[ss], Last[ss]];
observed = f[samples]
```

```
0.318421
```

```
results = Table[f[RandomPartition[samples]], {1999}];
count = 1 + Length[Select[results, (# ≥ observed) &]]
estimate = N[count / 2000]
```

```
65
```

```
0.0325
```

```
BinomialCI[count, 2000, ConfidenceLevel → 0.99]
```

```
{0.0231599, 0.044134}
```

```
Map[Median, {radon1, radon2}]
```

```
{30.65, 20.61}
```

```
{CI1, CI2} = {QuantileCI[radon1, 0.50, ConfidenceLevel → 0.95],
   QuantileCI[radon2, 0.50, ConfidenceLevel → 0.95]}
```

```
 Confidence     Confidence       Confidence     Confidence
{Level:         Interval:      , Level:         Interval:     }
 0.966448       {18.7, 55.13}    0.961523       {16.26, 29.02}
```

Because the rank-sum test had a P-value of 0.09, we can not reject the null hypothesis that neither of the distributions is stochastically larger than the other. However, the Smirnov test gives us a P-value of 0.0325, which indicates that we can reject the null hypothesis that the distributions are the same. We see taht this result is robust at the 5% level because the 99% confidence interval for statistic the yields [0.023,0.044]. Furthermore, we see that the maximum difference in the ECDF is 0.31, which is quite large and could indicate that something funny is happening in the data. 95% of radon2 observation fall below x=48.87, but only 63% of radon2 observations fall below that level. Furthermore, the we see that radon1 has a median of 30.65 while radon2 only has a median of 20.61. From the median confidence intervals, we see that both radon1 and radon2 have similar lower bounds, but radon1 has a significantly higher upper boundary. All of this information seems to indicate that while the levels of radon were largely similar between households with and without child cancer patients, a few outliers may have skewed some of the results of the radon1 data set upwards.