Women in the Olympics

DANIELLE LAMB

Hypothesis

Beginning in 1900, women were able to participate in the Olympics. Five events accepted women athletes that year, tennis, sailing, croquet, equestrian, and golf. “Female participation has increased steadily since then, with women accounting for more than 45 per cent of the participants at the 2016 Games in Rio,” (International Olympic Committee, 2018). The initial presence of women in the International Olympic Committee (IOC) began in 1981 when Flor Isava-Fonseca and Pirjo Haeggman were included. Currently four out of 15 IOC members are women.

In this project, the difference in events women can participate in and the number of women participants will be evaluated, comparing before women were on the IOC committee and after.

Data

	year	sports	womens_events	total_events	%_of_womens_events	women_participants	%_of_women_participants
0	1900	2	2	95	2.1	22	2.2
1	1904	1	3	91	3.3	6	0.9
2	1908	2	4	110	3.6	37	1.8
3	1912	2	5	102	4.9	48	2.0
4	1920	2	8	154	5.2	63	2.4

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27 entries, 0 to 26
Data columns (total 7 columns):
year                        27 non-null int64
sports                      27 non-null int64
womens_events               27 non-null int64
total_events                27 non-null int64
%_of_womens_events          27 non-null float64
women_participants          27 non-null int64
%_of_women_participants     27 non-null float64
dtypes: float64(2), int64(5)
memory usage: 1.6 KB
```

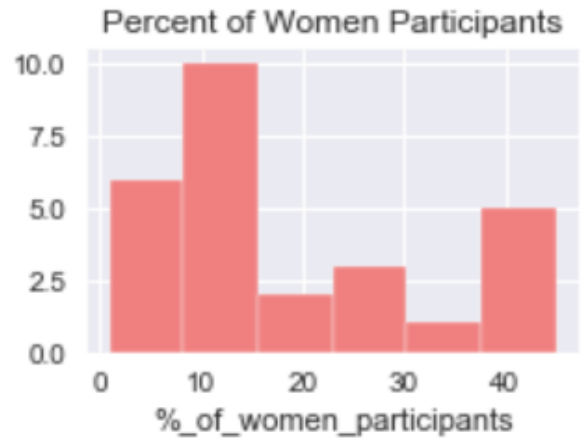
The Olympic data presents data from 1900 to 2016 which includes seven variables. All variables are either integers or floats. There is a total of 27 entries and all data is complete with zero incomplete observations.

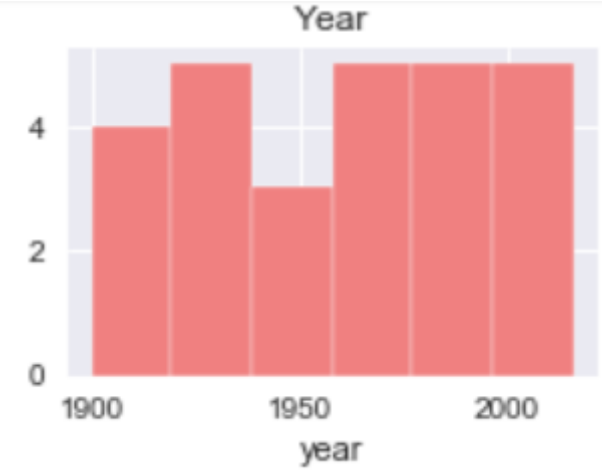
- The year in which the summer Olympics was held is represented as year.
- The number of sports that women could participate in is included in the variable sports.
- Women_events counts the number of events across all sports that women could participate in. This also includes mixed events.
- Total_events counts the total number of events during that year of the Olympics.
- The percentage of women's events each year is represented as %_of_womens_events. This variable is created from dividing women_events and total_events.
- Women_participants is the total number of women that participated in any sport or event during that year of the Olympics.
- The percentage of women participation is the final variable, %_of_women_participants.

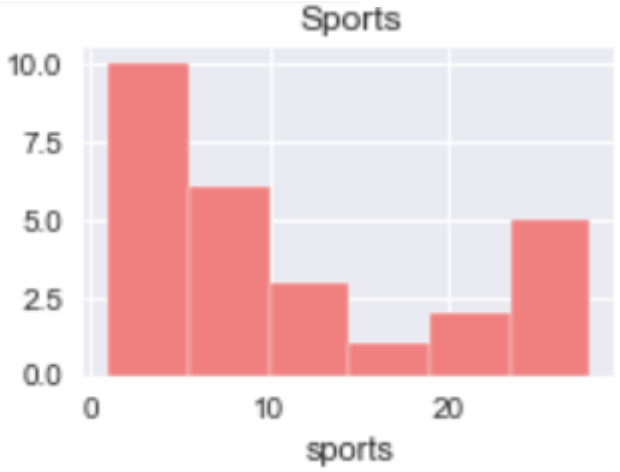
This data set was obtained from:

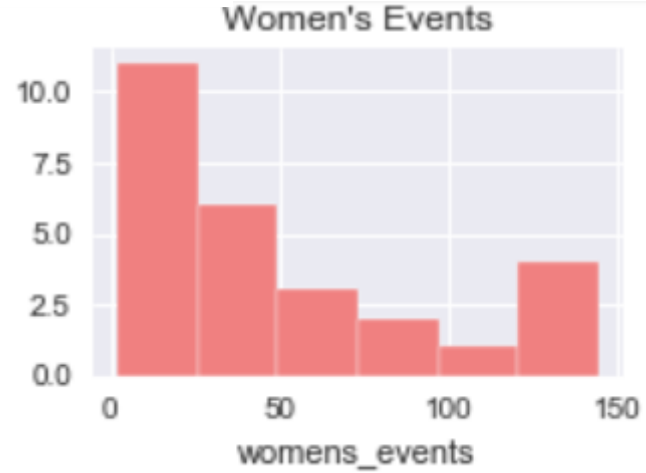
Women in the Olympic Games - dataset by sports. (2019, March 13). Retrieved from https://data.world/sports/women-in-the-olympic-games/workspace/file?filename=womens_participation_in_the_games_of_the_olympiad.csv

Histograms

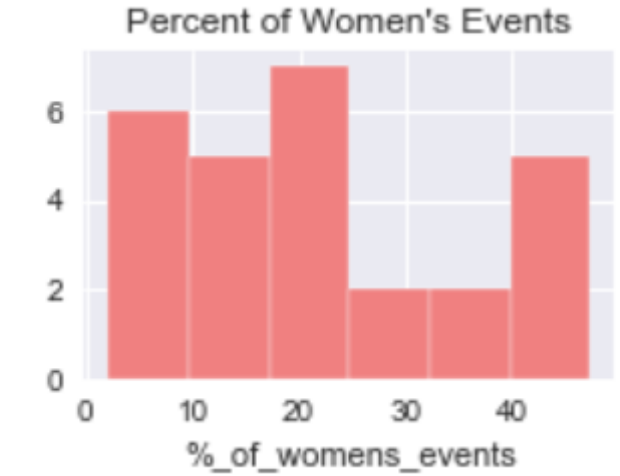

Percent of Women Participants
10.0
7.5
5.0
2.5
0.0
0
10
20
30
40
%_of_women_participants


Year
4
2
0
1900
1950
2000
year


Sports
10.0
7.5
5.0
2.5
0.0
0
10
20
sports


Women's Events
10.0
7.5
5.0
2.5
0.0
0
50
100
150
womens_events


Total Events
6
4
2
0
100
150
200
250
300
total_events


Percent of Women's Events
6
4
2
0
0
10
20
30
40
%_of_womens_events


Women Participants
15
10
5
0
0
2000
4000
women_participants

Histogram Analysis

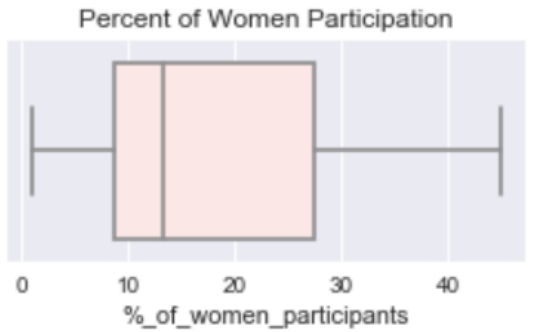
- The year histogram is almost even since the summer Olympics occurs every four years.
- The rest of the histograms present a right skewed distribution.
- There tends to be a spike in values on the far right of those same histograms.
- This could either be a sign of outliers or a rapid increase.
- This rapid increase could be as a result of the hypothetical question of women's involvement in the IOC affecting women participation.
- Moving forward, testing both normal distribution analyses and non-parametric analyses is crucial to find the best fitting methods.

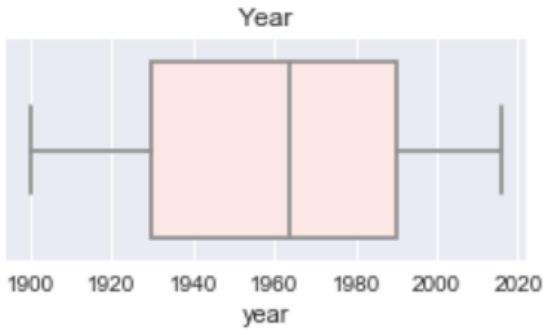
Describing data

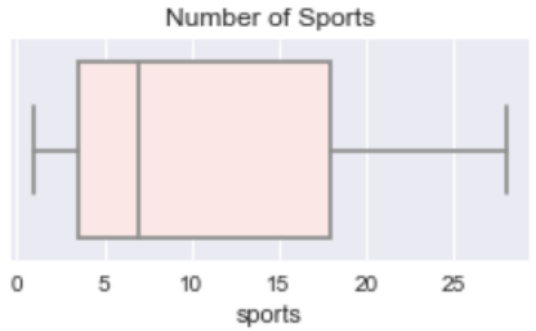
Variable	Mean	Std	Minimum	Maximum	25%	50%	75%
Year	1960.74	35.88	1900	2016	1930	1964	1990
Sport	10.85	9.14	1	28	3.5	7	18
Women Events	50.44	46.66	2	145	14	33	79
Total Events	186.92	73.36	91	306	127.5	163	247
Percent of Women Events	21.86	13.96	2.1	47.4	11.8	20.2	31.95
Women Participants	1489.78	1686.41	6	4700	206	678	2450.5
Percent of Women Participants	18.22	14.4	0.9	45	8.65	13.3	27.45

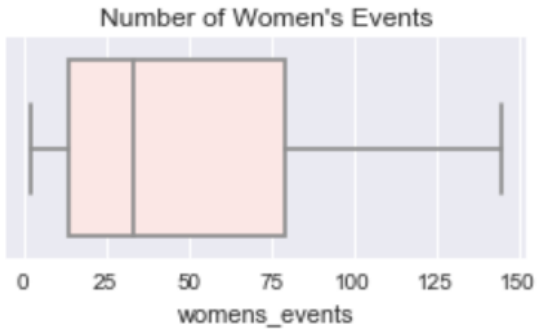
Boxplots

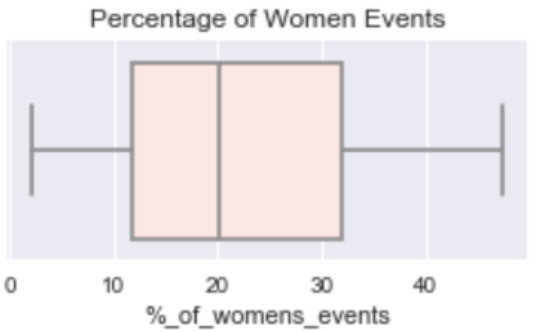
- To verify no outliers exist, based on the skewed data with concentrated right end, boxplots were made.
- No variables present any outliers based on boxplots.
- The boxplots also visualize the right skewness, as the IQR box is present on the left hand side of most of the boxplots.
- Year, total events, and percentage of women's events will be variables that are on the edge of being normal or non-parametric based on the boxplots.

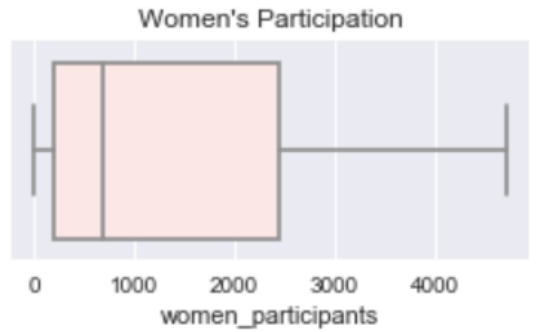






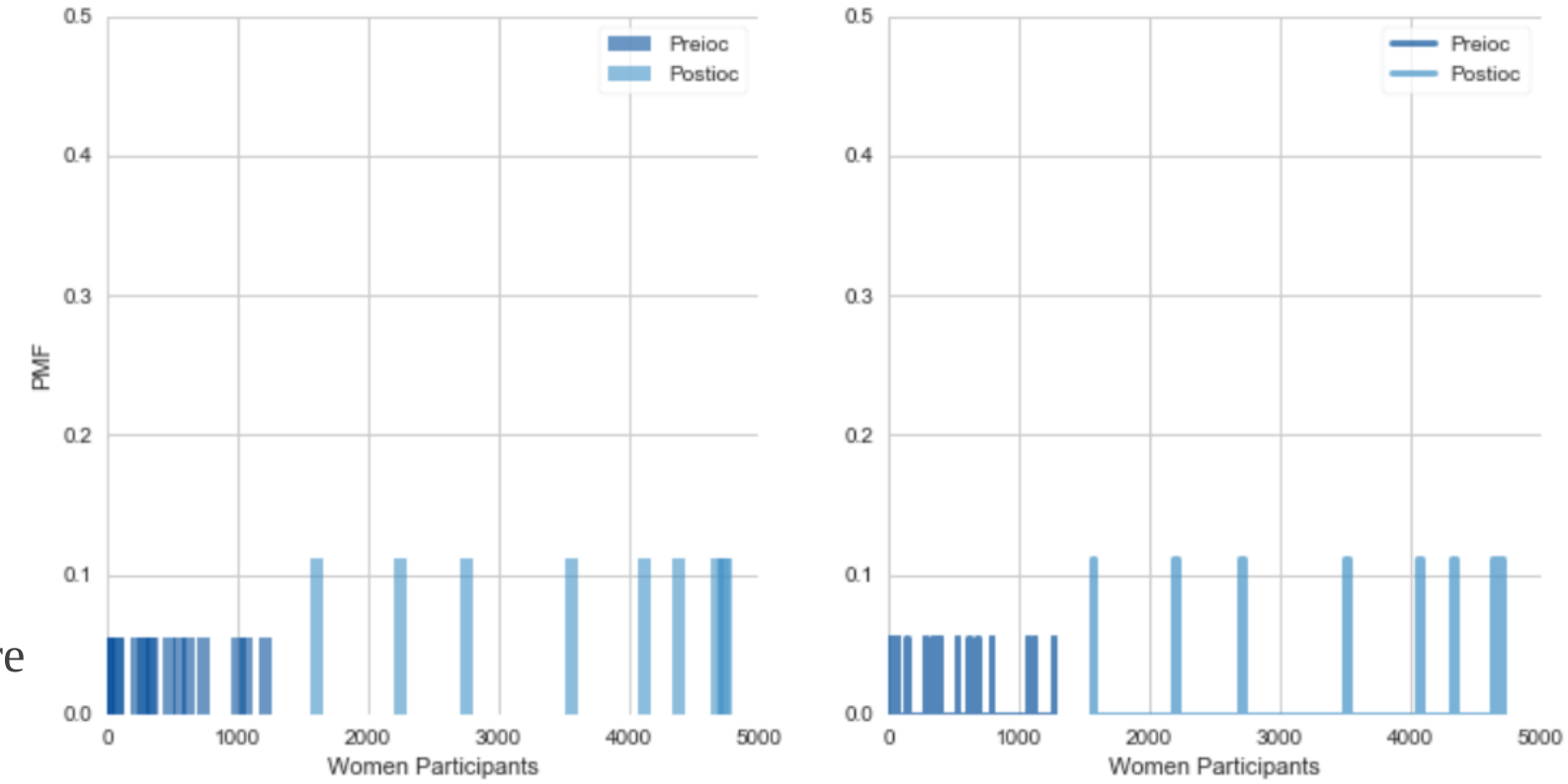






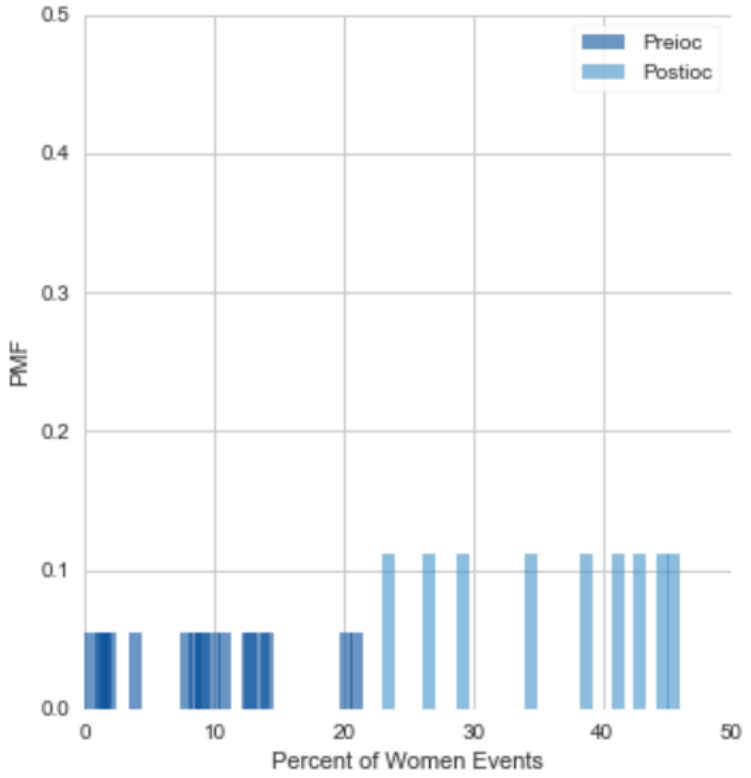
PMF One: Women participants before women in ioc (1981) vs. after

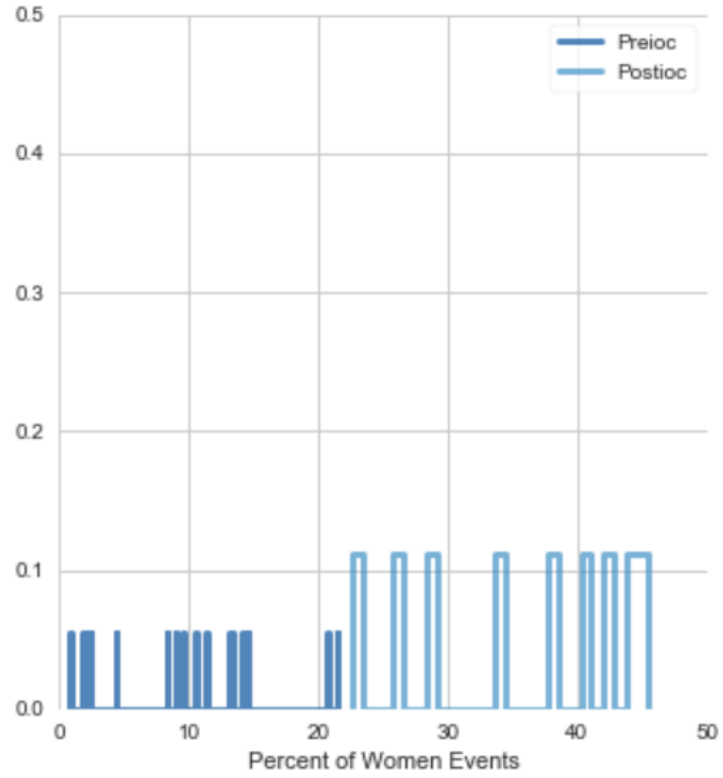
- The probability mass function is the probability of some value equaling another.
- The data chosen has unique values in that almost no variable has multiple values that are the same.
- As a result, the PMFs don’t have any pattern.
- Since the number of Preioc observations is greater than the Postioc, the Preioc PMFs are a smaller percentage.



PMF Two: Percent of women's events before women in ioc (1981) vs. after

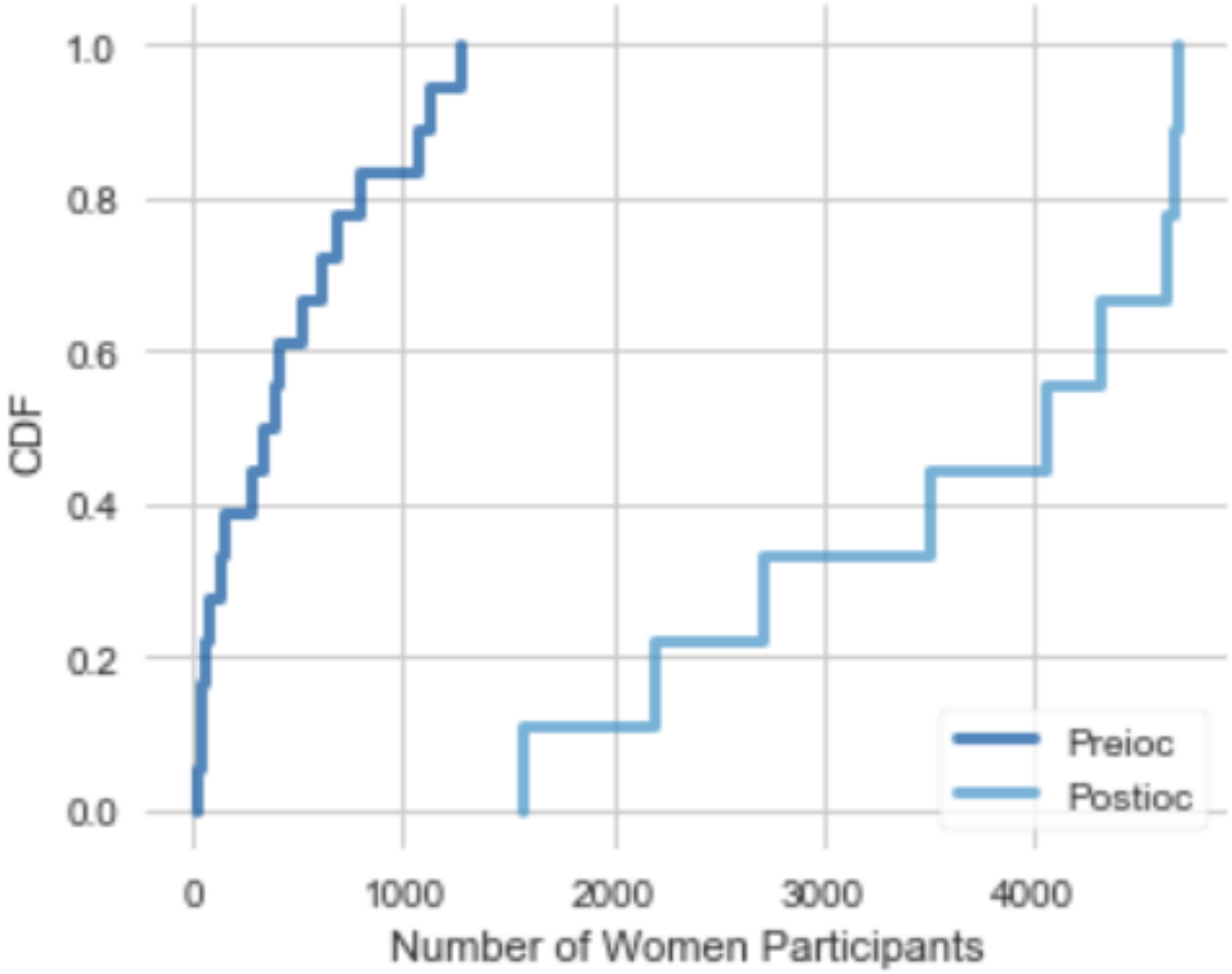
- The probability mass function is the probability of some value equaling another.
- The percent of women's events variable is similar to the woman participants in that the observations are unique.
- With both sets of PMFs, the Postioc values are higher, sitting on the right side of the x-axis.





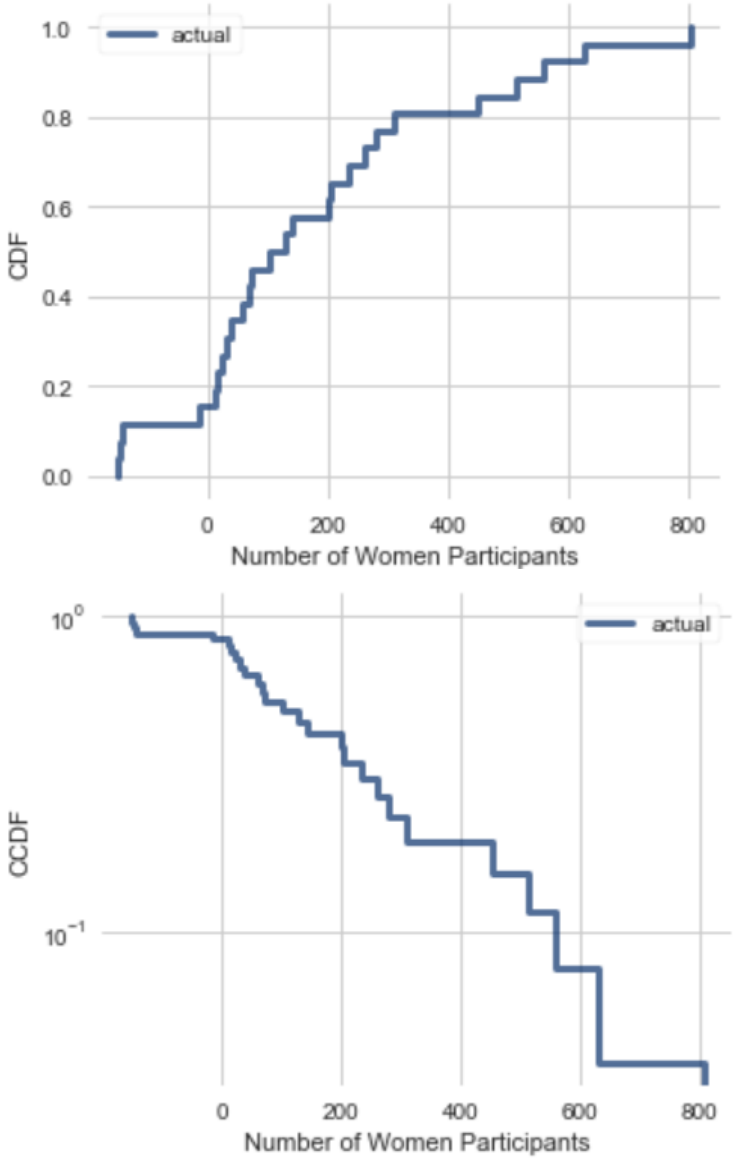
CDF: Number of women participants before women in ioc (1981) vs. after

- The Preioc cumulative distribution is very different from the Postioc CDF in that the rate at which the values rise is much greater for the Preioc.
- While the Preioc did not increase in number of participants very much (x-axis distance), the rate of increase was faster.
- The Postioc looks more like a step function, increasing at a fairly sustained rate with a rapid increase at the end.
- The behavior of these CDF's suggest there is a significant difference in women participation before and after women were included in the IOC.
- Further analyses will be performed to confirm this.



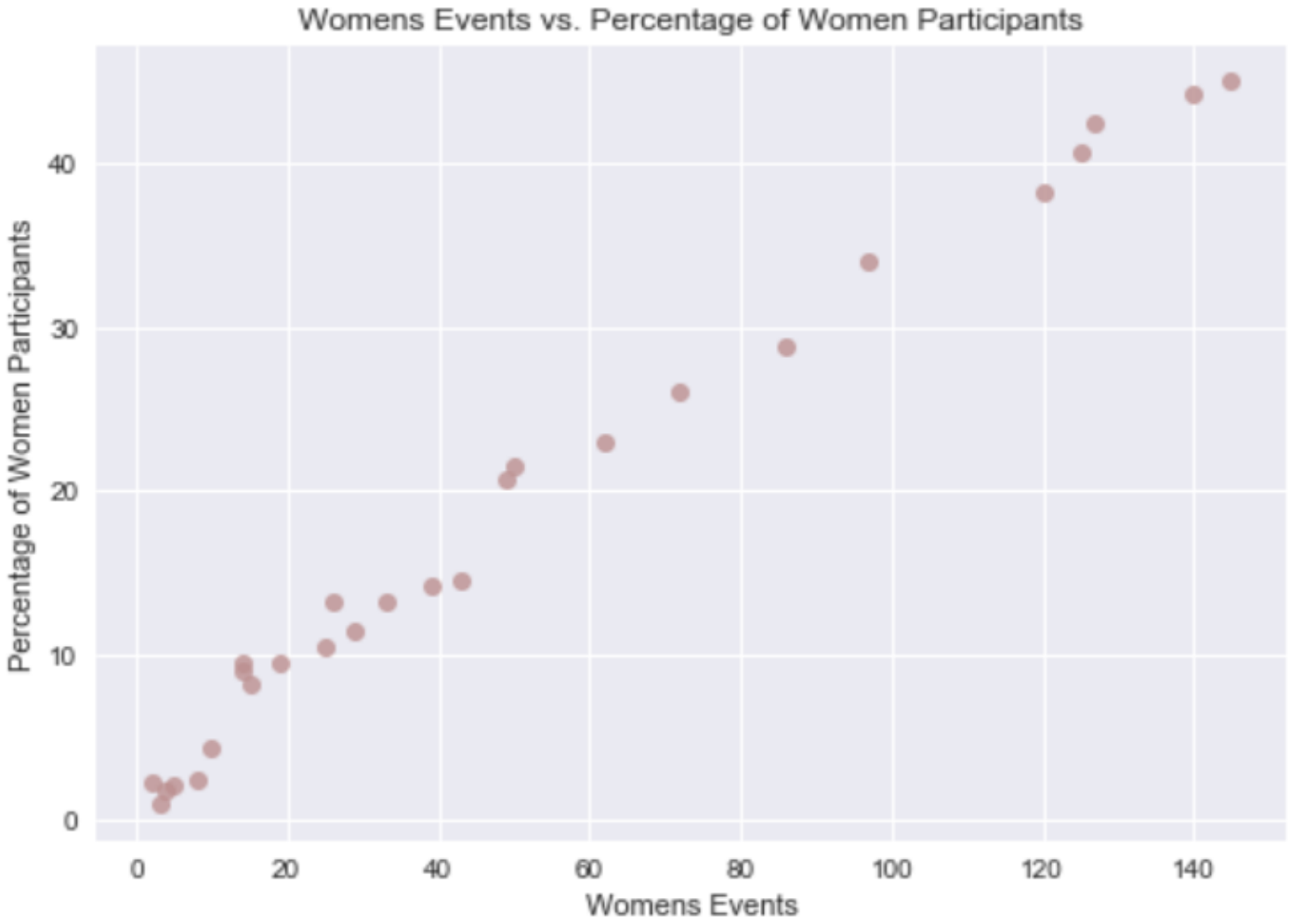
Analytical Distribution

- The analytical distribution chosen for this data was the exponential distribution.
- Alley Downey mentions that exponential distributions are often used in datasets where time is a factor, in addition to the above CDF appearing exponential.
- There is a slight curve in the top graph, indicating the model could be exponential but alone it is not enough to tell.
- The bottom graph, the CCDF is almost a straight line, with only a little deviance on the bottom right, this indicates that the exponential distribution does fit this data.



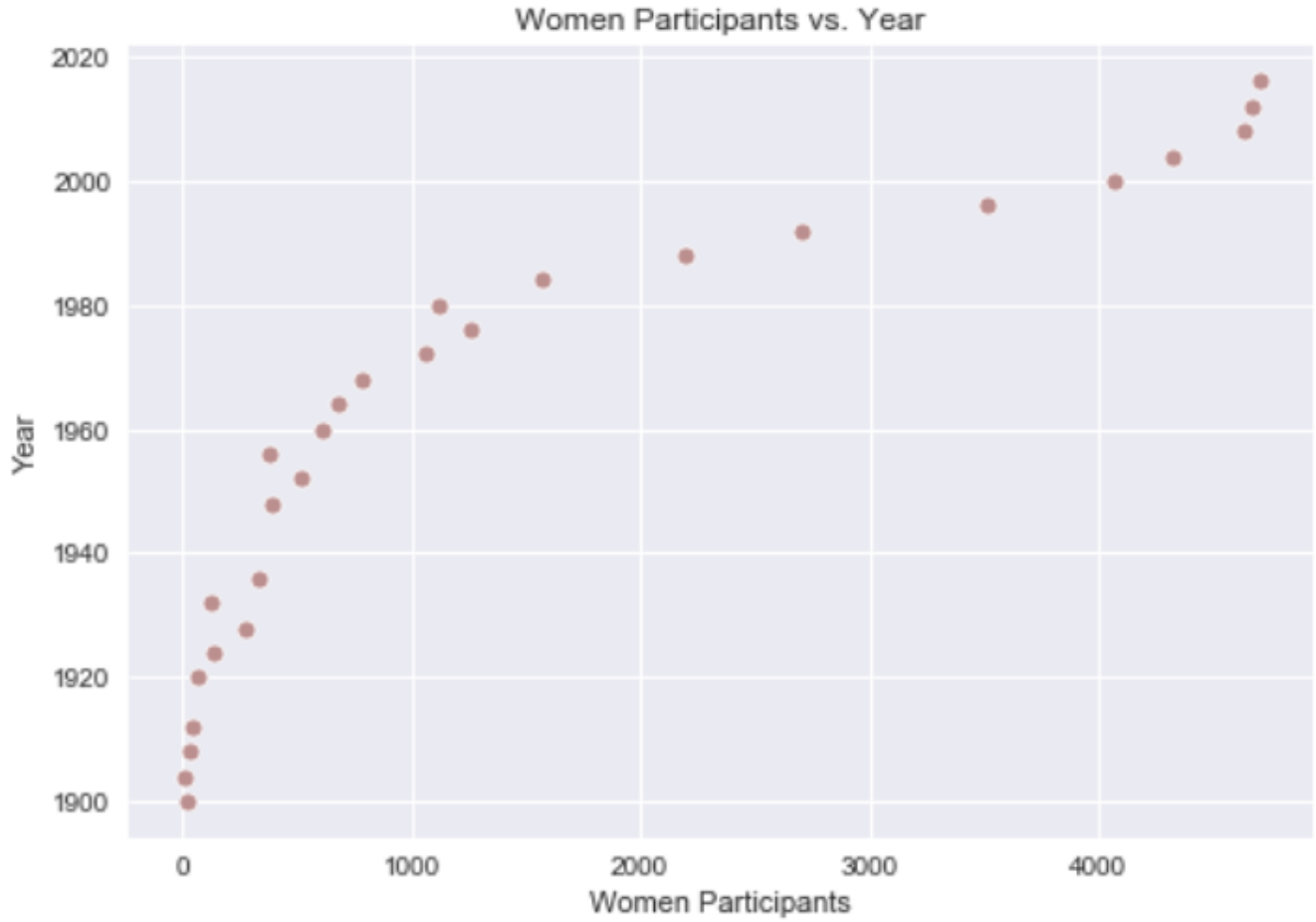
Scatterplot One: Women's events vs. percentage of women participants

- The scatterplot of women's event and percentage of women participants depicts a strong, positive relationship.
- As there are more women's events included in the summer Olympics, the more women attend, making the percentage of women greater.
- The data is more concentrated in the bottom left corner, within the 40 events and 20% women participants, then is strung out to the top right.
- The normal distribution Pearson's Correlation Coefficient resulted in 0.99, however since this data is skewed, the non-parametric Spearman's Rank correlation was also calculated and resulted in the same 0.99.
- The covariance was 641.99.
- Between the scatterplot and the numerical analyses of this data, we can conclude that women's events and percent of women participants have a strong, positive relationship



Scatterplot Two: Women participants vs. year

- The women participants per year looks similar to the CDF from the previous slides.
- Putting together the two time periods, there is a slow increase in participants over many years, then a fast increase in a smaller time period.
- This follows the hypothesis that more women have been involved in the Olympics after being involved in the IOC in 1981.
- The normal distribution Pearson's Correlation Coefficient resulted in 0.88, however since this data is skewed, the non-parametric Spearman's Rank correlation was also calculated and resulted in a higher correlation of 0.99.
- Comparing this scatterplot to the previous, this is less linear, which suggest a correlation of 0.88 being more accurate.
- The covariance was 51282.
- Between the scatterplot and the numerical analyses of this data, we can conclude that women participants and year have a strong, positive relationship



Hypothesis Test

H_0: $\mu_{Pre} = \mu_{Post}$

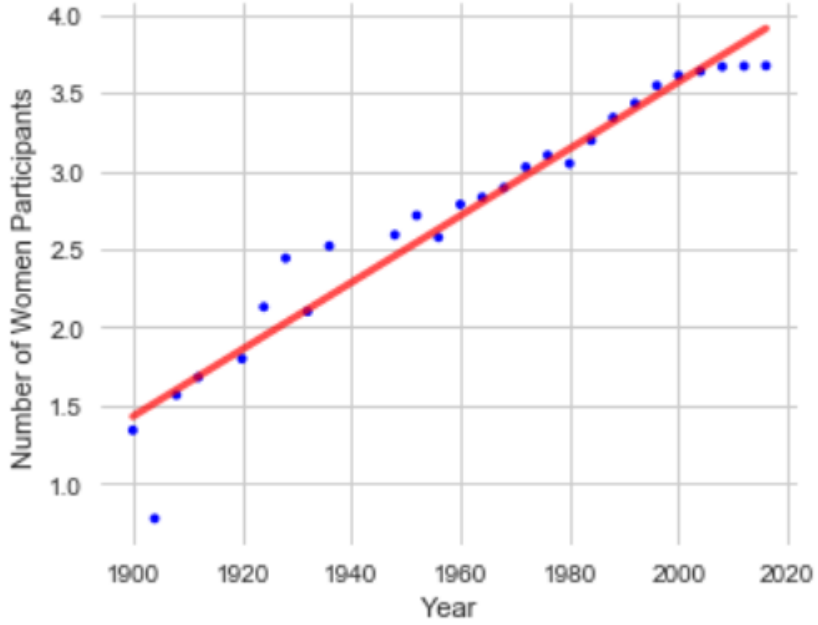
H_A: $\mu_{Pre} \neq \mu_{Post}$

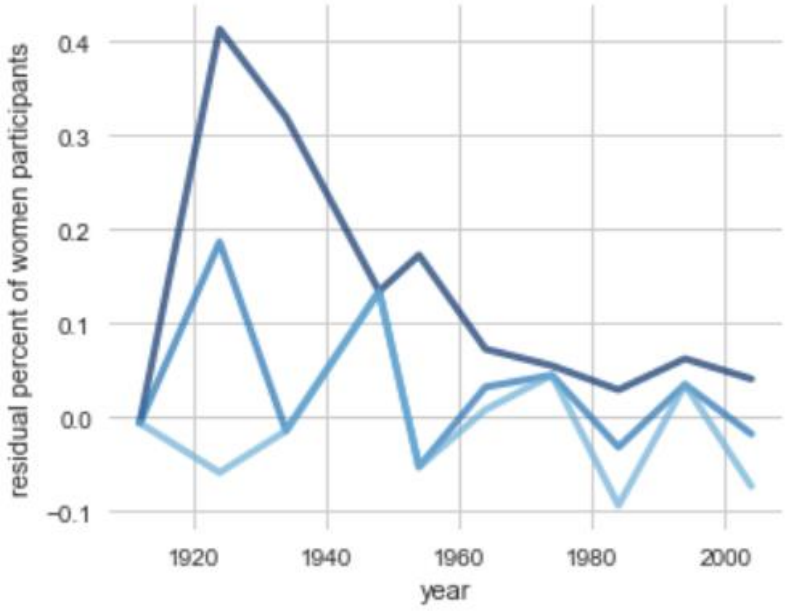
- There was some difficulty in running the Means Permute Two-Sided by itself, so the combination function RunTests() was used to obtain the correct resulting statistics.
- The null hypothesis states that the mean of the observations recorded prior to 1981 equals the mean of the observations after 1981.
- The alternative hypothesis, the question being asked, is that the mean of observations after 1981 is not equal to the mean of observations before 1981
- Since we are just looking for a difference in means, not greater or less than, we will be using the two-sided test.
- With a p-value of less than 0.0, we can conclude that there is enough evidence to reject the null hypothesis.
- In the case of the other two tests, with p-values of 1.0 and 0.985, we cannot reject the null hypothesis.

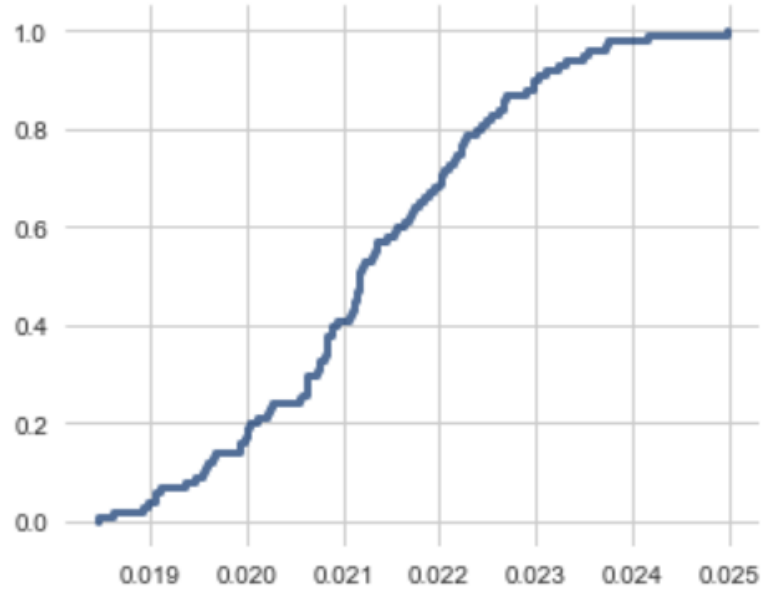
```
means permute two-sided
p-value = 0.0
actual = 3163.6666666666665
ts max = 2467.8333333333335
Writing hypothesis1.pdf
Writing hypothesis1.eps
The PostScript backend does not support transparency
The PostScript backend does not support transparency

means permute one-sided
p-value = 1.0
actual = -3163.6666666666665
ts max = 1660.8333333333335

std permute one-sided
p-value = 0.985
actual = -718.4502205794927
ts max = 1557.3885086329021
<Figure size 576x432 with 0 Axes>
```

Regression Analysis

- This section proved to be quite difficult, and I am unsure if the results are correct. A regression was run using year and number of women participants.
- Since the women participant data was skewed, the data was log-transformed before being run in the regression.
- The R-squared value for this model is 0.94, meaning, the regression model explains 94% of the data (top).
- When used in a CDF (bottom right), the p-value less than 0.000 with a standard error of 0.0013 meaning there is a small spread sample means are close to the population mean.







Citations

International Olympic Committee. (2018, October). Women in the Olympic Movement. Retrieved from https://stillmed.olympic.org/media/Document Library/OlympicOrg/Factsheets-Reference-Documents/Women-in-the-Olympic-Movement/Factsheet-Women-in-the-Olympic-Movement.pdf#_ga=2.7286004.710853385.1552501553-1697028591.1552313867

Some code used from:

Downey, A. B. (2015). *Think stats*. Sebastopol: OReilly Media.