Programming Assignment 2

INFO-5502 (Section 002): Analytic Tools, Techniques and Methods

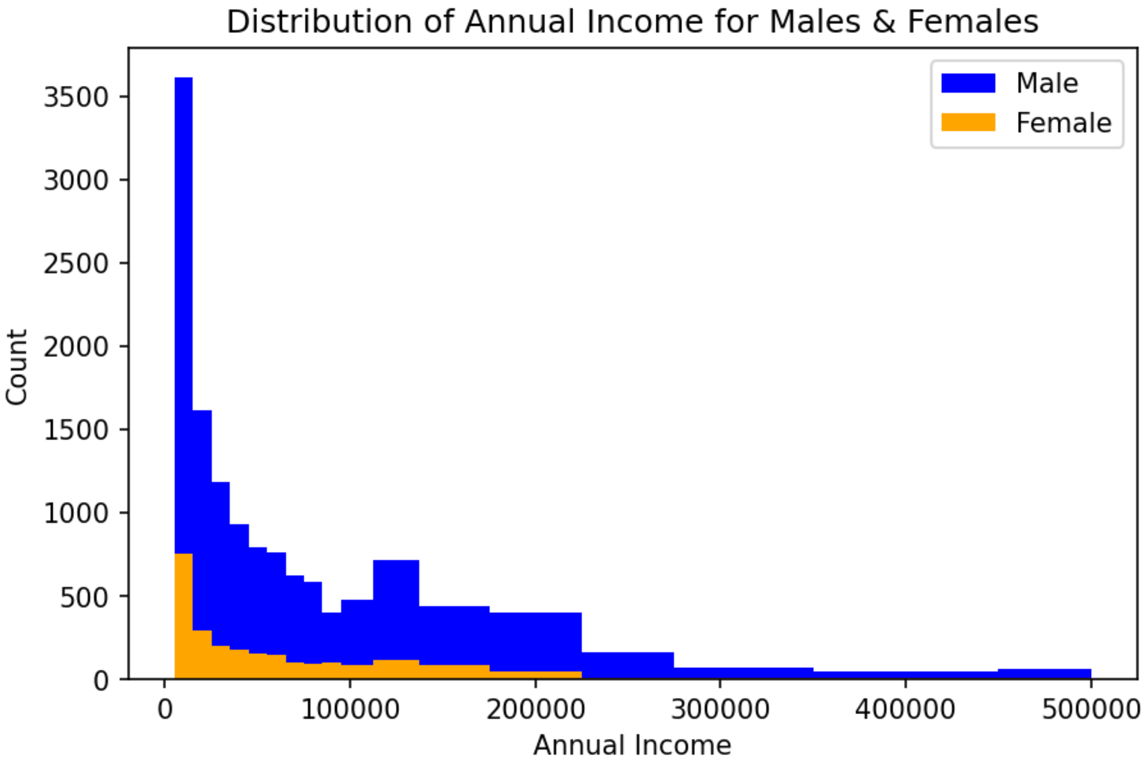
Ramandeep Harjai

February 13, 2021

1. Calculate the median income of male employees and the median income of female employee in the population. Consider the set of all employees in the datasets as the population. (1 point)

```
Population Size:  15429
Population: Median Income:  25000.0
Population: Median Income (Male):  35000.0
Population: Median Income (Female):  25000.0
```

2. Draw an overlaid graph to show the histograms of the incomes of female and male employees in the population. (You create one histogram for male, and another histogram for female, but the two histograms should be on display in the same graph with different colors. (1 point)



3. Use: random sampling, empirical distributions, sample comparisons, bootstrap, and hypothesis testing as well as A/B testing - that we discussed in the class - to analyze the income gap between female and male employees

i. Select a sample from the population. Make sure your sample include 500 employees selected from the population and consider how to ensure the sampling strategy is fair since the datasets include an overwhelming number of male employees compared to female employees (1 point).

```
sample_male = get_sample(income_male, 250,
                           with_replacement=False)

sample_female = get_sample(income_female, 250,
                             with_replacement=False)

sample = np.concatenate((sample_male, sample_female), axis=0)
```

```
Sample Size:  500
Sample: Median Income:  25000.0
Sample: Median Income (Male):  35000.0
Sample: Median Income (Female):  25000.0
```

ii. Define the test statistic, the null hypothesis and the alternative hypothesis (1 point).

Test Statistic: difference between median income for males and females

$\textbf{Test Statistic} = \text{Median Income}_{male} - \text{Median Income}_{female}$

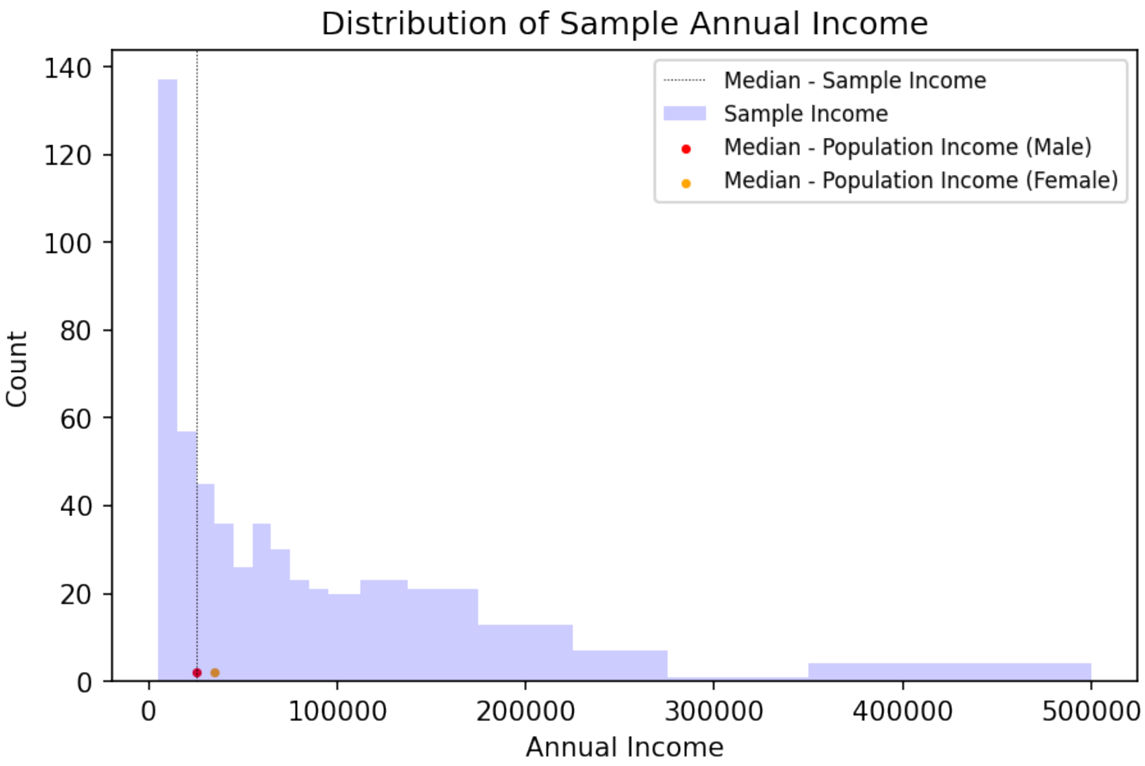
H₀ = There is no difference in median income for males and females

$\text{Median Income}_{male} - \text{Median Income}_{female} = 0$

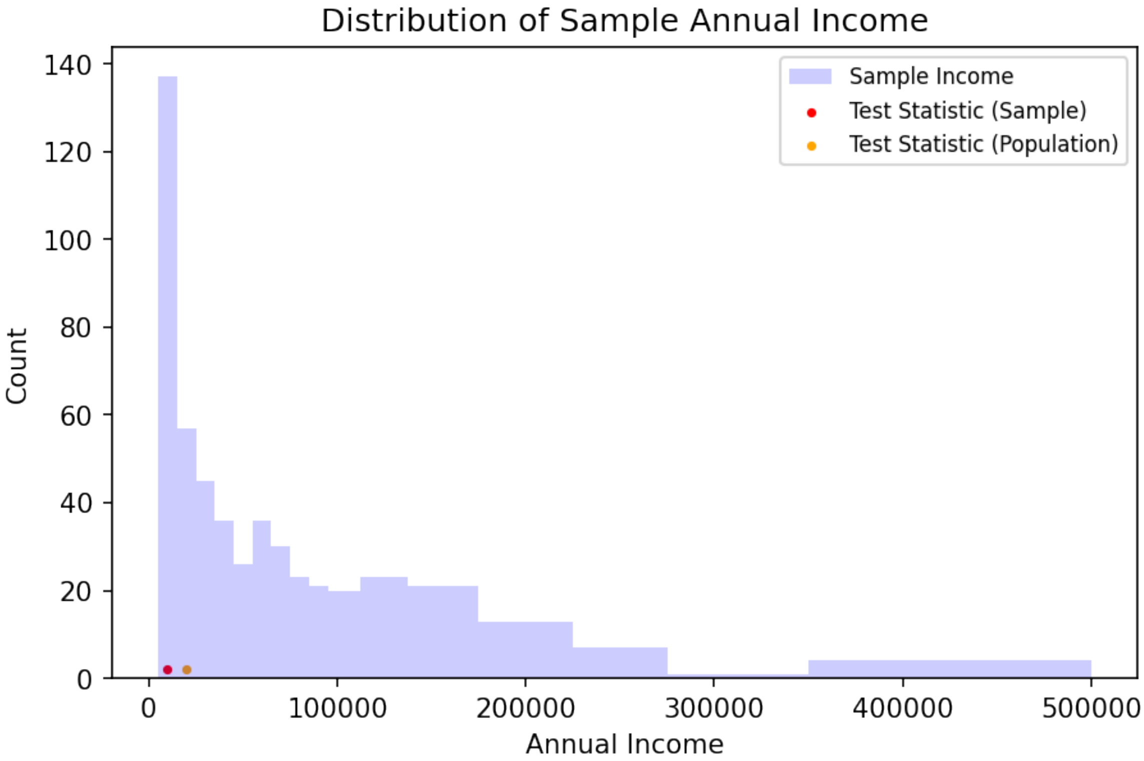
H_A = There is a difference in median income for males and females

$\text{Median Income}_{male} - \text{Median Income}_{female} \neq 0$

iii. Draw the income histogram for the sample; calculate the median income of the sample; and draw a red dot and a yellow dot for the female median income and male median income of the population respectively, in the histogram (1 point).



iv. Draw the income histogram for the sample and draw a red dot to show the corresponding test statistic of the population (e.g., the difference of the median incomes between female and male employees) in the diagram (1 point).



v. Write a procedure to use bootstrap to produce at least 5000 samples (1 point).

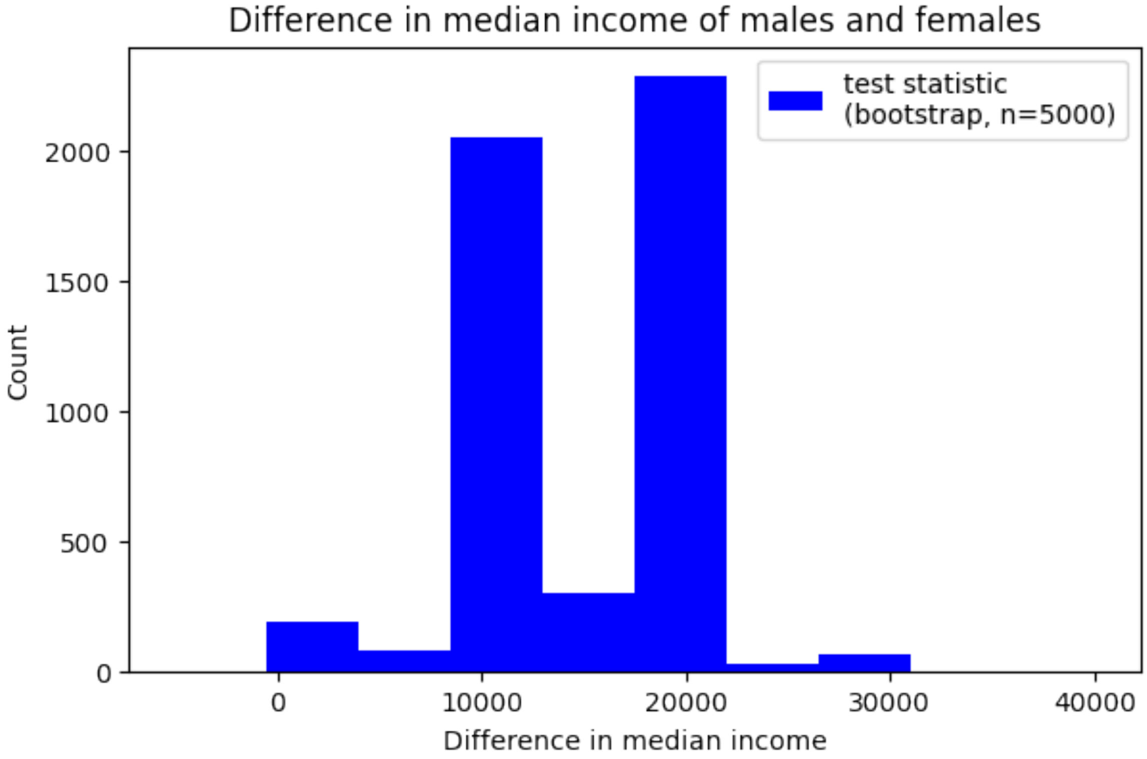
```
test_stats = []

repetitions = 5000
for i in range(repetitions):
  male_smpl = get_sample(sample_male, size=250,
                           with_replacement=True)

  female_smpl = get_sample(sample_female, size=250,
                           with_replacement=True)

  test_stats.append(get_test_stat(get_stat(male_smpl),
                                      get_stat(female_smpl)))
```

vi. Draw the histogram of the test statistic of the bootstrap samples (1 point).



vii. Define the confidence interval and P-value to validate the hypothesis you defined (2 points).

```
test_stat_arry = np.array(test_stats)

ci_low = np.percentile(test_stat_arry, 2.5, axis=0)
ci_high = np.percentile(test_stat_arry, 97.5, axis=0)

print("CI 95%: [{}, {}]".format(round(ci_low,2),
                                  round(ci_high,2)))
```

```
CI 95%: [0.0, 30000.0]
```

```
pvalue = np.mean(test_stat_arry >
                      get_test_stat(smpl_income_median_male,
                                    smpl_income_median_female))

print("pvalue: ", pvalue)
```

```
pvalue:  0.561
```

Explanation:

Sample test statistic (difference between median income for male and females) observed in the sample of 500 (including 250 Males and 250 Females) is: $35,000 - $25,000 = **$10,000**

95% Confidence Level obtained from 5,000 bootstrap samples: **[0.0, 30000.0]**. While the 95% confidence level includes the observed test statistic of $10,000; the confidence level also includes **0** (Null hypothesis value). The p-value obtained is **0.561**, which is greater than α = 0.05.

Based on the results obtained we fail to reject the null hypothesis that there is no difference between median income of male and female.

****** end of document ******