Income Structure and Expenditure Structure

Chenhao Lin

June 2022

1 Introduction

The income structures and consumption structures for different neighborhoods play a key role in the city's development. The yearly income and expenditure structure can reveal things about the current status of our society. For example, the income structure of each neighborhood shows the difference between the rich and the poor in the neighborhoods. The structure of income and expenditure can also show people's consumption level. When these structures are similar, that means the city or the neighborhood is in overall development. Therefore, investigating these structures are important and we can learn whether or not the city develops stably.
In this paper, we will specifically focus on the income and expenditure structure of three neighborhoods around the Seattle area: University District, Downtown Seattle, and Capitol Hill. The reason we choose them is that they contain different resources. University District contains University of Washington, so it has education resources. Downtown Seattle contains lots of companies, like Amazon, so it has employment resources. Capitol Hill has many bars, eateries, and clubs, so it has entertainment resources.

2 Data Simulation

We set both yearly income and yearly expenditure measured in ten thousand dollars.

2.1 Setup for Yearly Income

2.1.1 Model Selection

Since there are a few households have much higher income than the others, the density of yearly income should be right-skewed. Besides, the yearly income cannot be a negative value. Therefore, to simulate the distribution, we set this follows the absolute value of T-distribution with degrees of freedom k, centered at a small value s, and scaled by s.

$$\text{Income} = |s \cdot T_k + m|$$

Then, the probability of income can be calculated by:

$$f(\text{Income}) = \frac{1}{s}[f_{T_k}(\frac{\text{Income} - \mu}{s}) + f_{T_k}(\frac{-\text{Income} - \mu}{s})]$$

To simulate the overall yearly income of these three areas, we mixed the sample in a proportion of p, which is the ratio of population in these three areas.

Then, the probability of mixed income can be calculated by:

$$f(\text{Mixed Income}) = \sum_{i=1}^{3} p_i \cdot f_i(\text{Income})$$

where i is the index of area.

2.1.2 Data Selection

The real-world data is obtained from Statistical Atlas, which is provided by **US Census Bureau**.

Neighborhood	Median Income (10k$)	Number of Population
University District	4.73	27006
Downtown Seattle	4.41	7990
Capitol Hill	8.08	29343

We set the parameters as:

Neighborhood	m	s	k	p
University District	5	3	3	0.42
Downtown Seattle	4	4	4	0.13
Capitol Hill	8	7	5	0.45

2.2 Setup for Yearly Expenditure

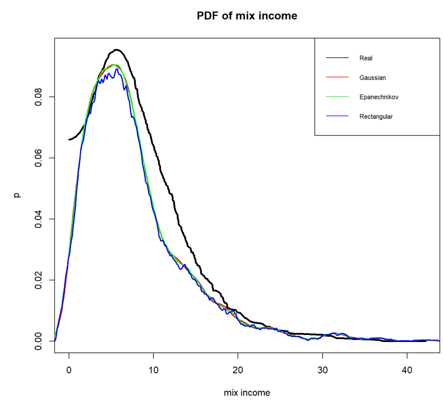
The yearly expenditure is calculated by:

$$\text{Mixed Expenditure} = |2.5 + 0.7(\text{Mixed Income} - \mu_{\text{Mixed Income}}) + T_4|$$

where T_4 is a standard T random variable on 4 degrees of freedom.

2.3 Kernel Density Estimation

Since the probability density function of mixed expenditure is unknown, to observe its distribution, we use kernel density estimation.
We use three most common kernels: Gaussian kernel, Epanechnikov kernel, and rectangular kernel. The kernels are estimated based on a sample of mixed income which has a size of 1000, and the optimal bandwidth is selected by R, which equals to 0.9324615.



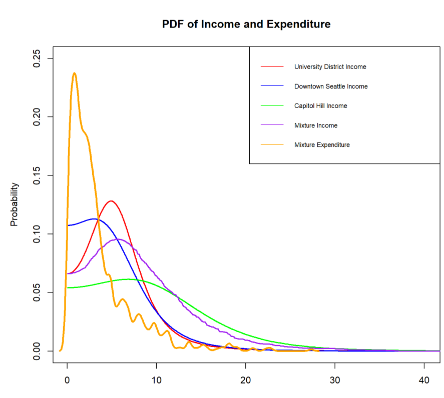
To find out what kernel has more accurate estimation, we compare the bias and mean square error (MSE) for these three kernels. To calculate them, we focus on the first value of mixed income.

Kernel	Bias	MSE
Gaussian	-0.007223027	5.217212e-05
Epanechnikov	-0.007309737	5.343226e-05
Rectangular	-0.007309737	5.343226e-05

The Gaussian kernel has the smallest bias and MSE, which means its estimate is more accurate than the other two kernels.

2.4 Probability Density Function

The following graph for probability density function is generated by a sample which has a size of 1000.
Since Gaussian kernel has the most accurate result when estimates the mixed income, we choose it to estimate the probability distribution function for expenditure. The optimal bandwidth selected by R is 0.2853636.



3 Analysis on Distribution

We will perform statistical tests to examine our hypothesis on different sample sizes at a significance level of $\alpha = 0.05$. And we want to find out the optimal sample size that makes the lower 95% confidence interval for the power of the tests above the threshold we chosen (0.8).

3.1 Preparation and Thoughts

3.1.1 Hypothesis Testing

$$H_0 : \text{Selected ones follows same distribution}$$
$$H_a : \text{Selected ones follows different distribution}$$

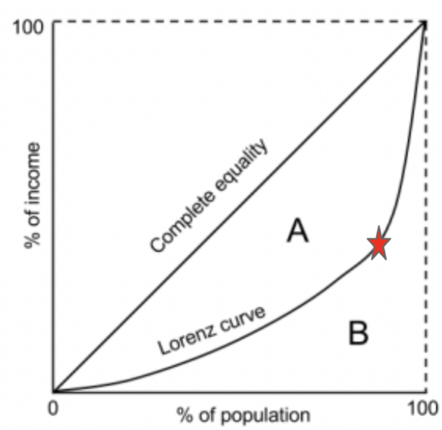
3.1.2 Test Selection

We choose to perform three tests.

1. `KS Test`
2. `Permutation Test on Interquartile Range`
3. `Permutation Test on Median`

The reason why we choose these tests is that they do not need us to make any assumptions about the distribution of our data. Hence, they are great non-parametric approaches for our purpose.

3.1.3 Selection for Permutation Test



In Economics, the Lorenz curve (Shown above) is used to represent the distribution of income, it shows the proportion of overall income or wealth assumed by the bottom x% of the people. As we can see in the left, there's a spike at the red star, where a small percentage of population shares a large percentage of income. And this, will inevitably create outliers in our income data. Hence, when it comes to choosing the parameters to test on for our permutation test, we should choose population parameters that is less likely to be affected by outliers, that is, **median** and **interquartile range**.

3.1.4 Monte Carlo and Permutation

To have more accurate results, we will use the Monte Carlo Method by simulating 10000 times for KS test. We will simulates 100 times for these two permutation test, with each time do 1000 times of permutation.

3.1.5 Performance of Monte Carlo and Permutation

We use **Power** as a way of measurement of the performance of our Monte Carlo and permutation tests. To compare these tests, we set a **Power threshold** to be 0.8.
The reason of doing so can be explained by:

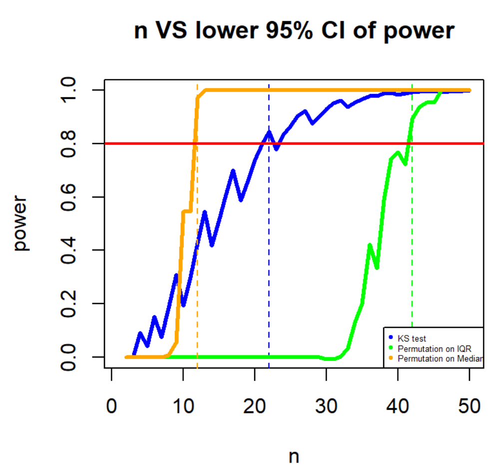
$$\text{power} = \Pr(\text{Reject } H_0 | H_a \text{ is True})$$
$$\text{power} = 1 - \beta(\text{Maximum Type II Error Rate})$$

It's not hard to see that choosing a low Statistical Power will introduce higher risk of committing Type II errors, e.g. a false negative. Hence it's optimal for us to choose a reasonable high Power, so we are less likely to commit a Type II error.
Since $\alpha = 0.05$, α and β have inverse relationship, we should balance things out at the cost of computational complexity. Since each increase in power at $\alpha = 0.05$ will drastically increase the sample size and the runtime, we finalized our power threshold to be 0.8.

3.2 Test Result

3.2.1 Distribution of Mixed Income and Expenditure

To calculate the optimal sample size, we start calculating power from a size of 2 to a size of 50. Each time the sample size will increase by 1.

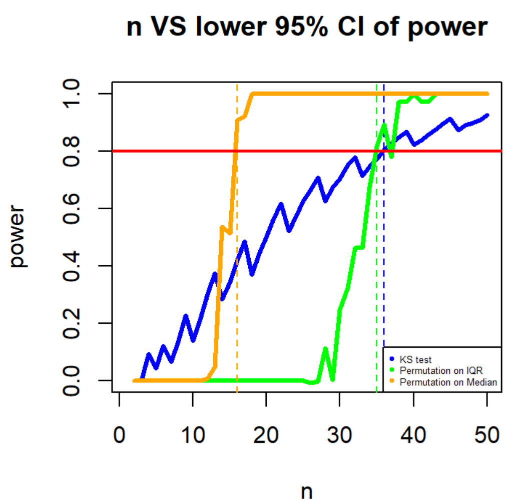


The optimal sample size is:

Test	Optimal Sample Size
KS Test	22
Permutation Test on Interquartile Range	42
Permutation Test on Median	12

3.2.2 Distribution of Downtown Seattle and Capitol Hill

Same setup with **3.2.1**.

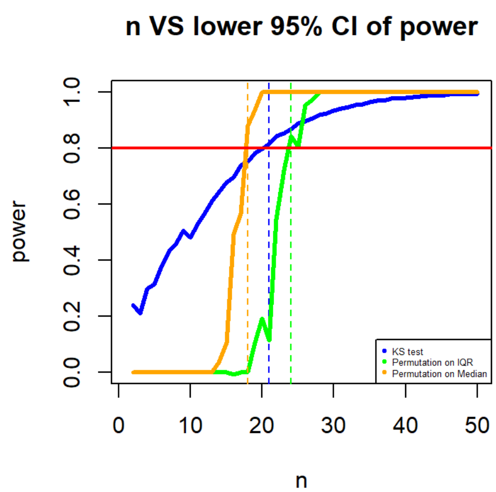


The optimal sample size is:

Test	Optimal Sample Size
KS Test	36
Permutation Test on Interquartile Range	35
Permutation Test on Median	16

3.2.3 Distribution of University District and Capitol Hill

Same setup with **3.2.1**.

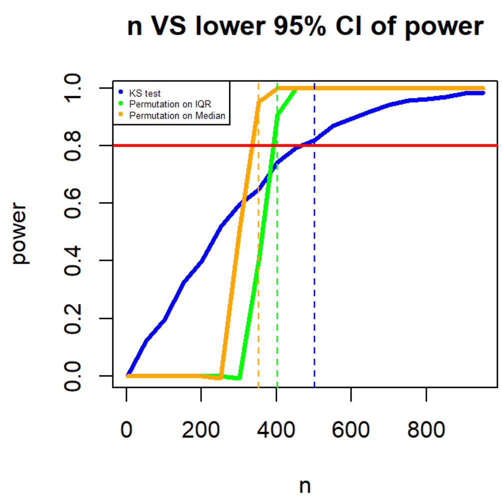


The optimal sample size is:

Test	Optimal Sample Size
KS Test	21
Permutation Test on Interquartile Range	24
Permutation Test on Median	18

3.2.4 Distribution of University District and Downtown Seattle

Since the income of University District and income of Downtown Seattle have similar distribution, to calculate the optimal sample size, we start calculating power from a size of 2 and end at a larger size (1000). To reduce the run-time, each time the sample size will increase by 50.

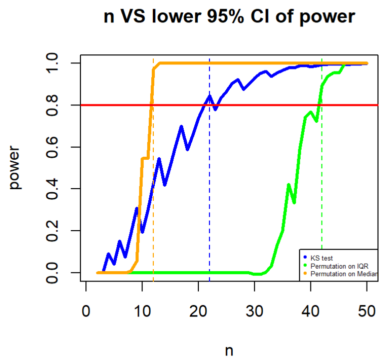


The optimal sample size is:

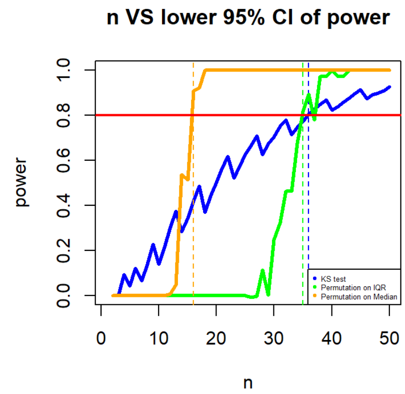
Test	Optimal Sample Size
KS Test	502
Permutation Test on Interquartile Range	402
Permutation Test on Median	352

3.3 Summary in Statistics

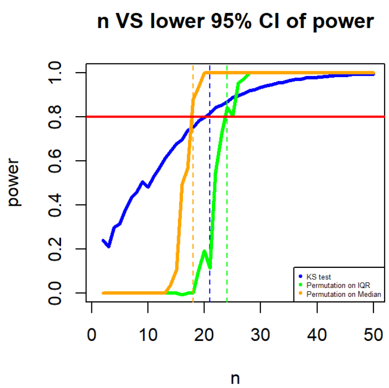
The optimal sample size to have a power of at least 0.8 is small. Even though the result in University District and Downtown Seattle is much larger than the others, it is only about 1% of the total households in these two areas. Therefore, there is enough evidence to reject the null hypothesis that they have same distribution.



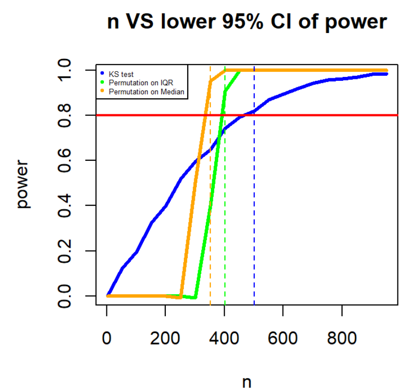
(a) Mixed Income and Expenditure



(b) Downtown Seattle and Capitol Hill



(c) University District and Capitol Hill



(d) University District and Downtown Seattle

Compare these four tests, the permutation test on median is always better than the permutation test on interquartile range since the optimal sample size is smaller. For the KS test, the lower bound of 95 percent confidence interval for the power increases slower than the other two. But when the sample size is small, KS test has fewer zeros and works better than the other two.

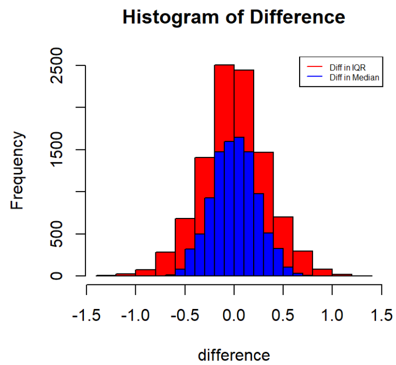
3.4 Discussion in Economics

Statistically, we rejected the null hypothesis in favor of the alternative hypothesis. This indicates that the income and expenditure hold different distributions. According to Mike Maciag [3], larger cities create more income inequalities because they tend to attract more individuals at opposite ends of the city's economic spectrum. From our analysis, we observed Seattle follows Maciag's view of more significant income inequality in larger cities.

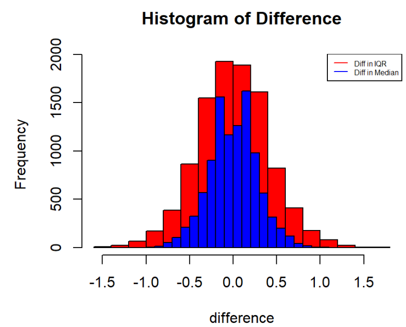
Though a household's income and expenditure are directly related: an increase in income leads to an increase in demand for service and goods, they don't have the same distribution due to the variability among individuals' spending habit [2]. All households would spend money on essential needs such as rent, food, insurance, and transportation and it follows the rule that the rich would spend more on essential goods for higher quality of living while the poor is the opposite [4].The rest of the money one holds depend on one's consumption habit. For example, some people would spend these money on normal and luxury goods such as clothes, accessories, and cars while some people would save it for later use [1]. The variability among individuals' spending habit affects the expenditure structure overall. Thus, economically, the income and expenditure have different distributions.

3.5 Reason Behind Conclusion

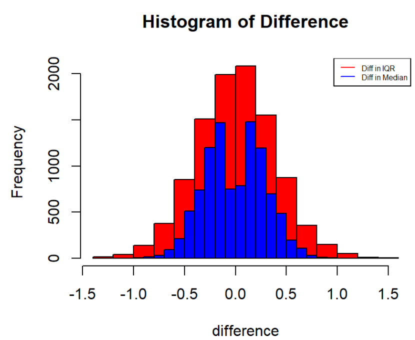
To find out the reason why permutation test on median is always better than the permutation test on interquartile range, we calculate the permutation distribution of difference in median and interquartile range. We set the sample size be 1000 and do 10000 times of permutation.



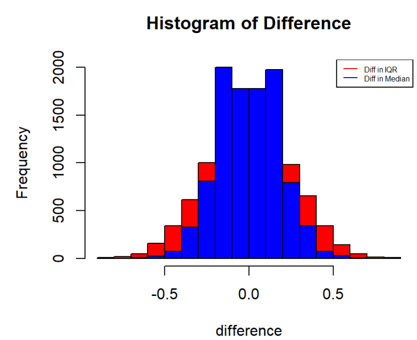
(e) Mixed Income and Expenditure



(f) Downtown Seattle and Capitol Hill



(g) University District and Capitol Hill



(h) University District and Downtown Seattle

The following is the variance for the permutation result.

Difference	Variance of Interquartile Range	Variance of Median
Mixed Income and Expenditure	0.1105686	0.05272357
Downtown Seattle and Capitol Hill	0.1654201	0.07471841
University District and Capitol Hill	0.1464419	0.0849756
University District and Downtown Seattle	0.063329	0.03240265

As we can see from above, the variance of median is lower than the variance of interquartile range. Hence, performing permutation test on median is always better than that on interquartile range.

4 Appendix

4.1 Literature Cited

References

[1] Household expenditures and income. The Pew Charitable Trusts. (n.d.). Retrieved June 4, 2022, from https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2016/03/household-expenditures-and-income

[2] Horowitz, J. M., Igielnik, R., & Kochhar, R. (2020, August 17). 1. trends in income and wealth inequality. Pew Research Center's Social & Demographic Trends Project. Retrieved June 4, 2022, from https://www.pewresearch.org/social-trends/2020/01/09/trends-in-income-and-wealth-inequality/

[3] Maciag, M. (2021, April 21). Income inequality climbing in many large cities. Governing. Retrieved June 4, 2022, from https://www.governing.com/archive/gov-income-inequality-big-cities.html#::text=Large%20metros%20have%20higher%20rates%20of%20income%20inequality,only%20shot%20at%20affordable%20housing%20within%20a%20region

[4] Team, T. I. (2021, November 11). What are the consequences of income effect? Investopedia. Retrieved June 4, 2022, from https://www.investopedia.com/ask/answers/041415/what-effect-does-income-effect-have-my-business.asp