Since We Have the Means, Let's Find the Happy Median

Frank A. Cirillo, MS
Economist
Division of Consumer Expenditure Surveys
Bureau of Labor Statistics
EEA 2/27/21

Introduction

- When looking at data, the way we present it frames our understanding of the world around us.
  - Hypothetically, if New York City has a homicide rate of 1 per 100,000 people in April 2019 and in May of 2019 it has a homicide rate of 2 per 100,000 people, it would be true to say that there was a 100% increase in the homicide rate from April to May. However, it would also be true to say that the homicide rate increased by 1 in 100,000 people from April to May 2019.
  - A headline stating, "Homicide rates double in NYC from April to May" gets a lot more clicks than "Homicide rate increases by 1 per 100,000 people from April to May"

The CE Connection

- How does this relate to BLS Consumer Expenditure Surveys (CE) data?

- When we publish annual and semi annual reports, we decide how to frame and present our data. This can affect how the outside world perceives and reports on issues pertaining to Consumer Expenditure Data.

- All data and calculations in today's presentation were made with internal CE Microdata; replicating with CE Public-use microdata files might yield slightly different results.

Medians vs Means

- Currently, the CE program only reports mean values for the data we collect and release in our tables and annual reports. These mean values include null values for consumers who don't report any purchase of that item.
- What if there was another way to present our data, what would that show us?
- Medians are another measure of centrality that have distinct advantages and disadvantages compared to means.

Outliers

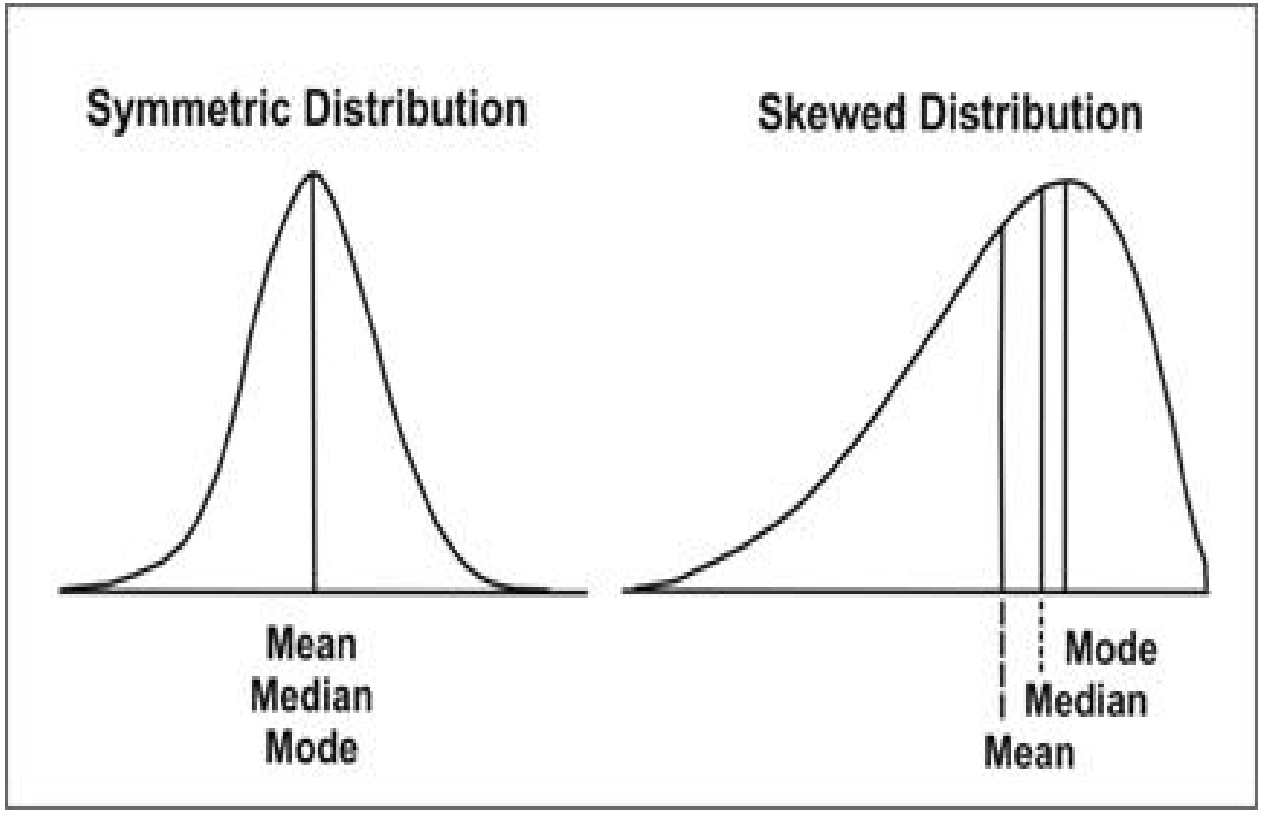
- One of the advantages of using medians over means is the lack of sensitivity to outliers.
- Means can be highly sensitive to extreme values on either end of data sets and this bias is reflected in the mean value.
- Medians on the other hand, are resistant to outliers, meaning you can get a more representative picture of the data using medians when outliers are included in your data.
- For example, if you take the numbers 0 through 10 as your sample, the mean and the median are both 5. But if you change the 10 to 1,000, the mean jumps to 95 while the median is unchanged.

Population Distribution

- Another advantage of using medians compared to means is how they are affected by skewed data sets.

- Means are pulled up or down depending on the skew of the data where as medians maintain a more stable representation of the data as shown in chart 1.

Chart 1



Percent Reporting

- When sampling a large group of respondents, many times you will have null or 0 values for certain reported categories, and the CE data are no exception.
- Currently, CE includes null values in our mean calculations. This creates a problem if we were to adapt medians as a new measure of our data.
- For any category that does not have at least 50% of respondents reporting, the median value will be 0. For this reason we will use this value (50% reporting) as a bench mark when looking for medians in CE data.
- We can also compare the median and mean values of only those consumers who reported data for certain categories to see if the reported values skew in any direction.

All 2019 categories with percent reporting >50%

Category	% reporting	Category	% reporting	Category	% reporting
Bakery products [D]	65.16	Fruits and vegetables [D]	73.03	Pensions and Social Security [I]	77.21
Cable and satellite television services [I]	56.66	Gasoline [I]	88.61	Personal care services [I]	60.17
Cellular phone service [I]	76.9	Gasoline, other fuels, and motor oil [I]	89.57	Personal insurance and pensions [I]	83.4
Cereals and bakery products [D]	70.75	Health insurance [I]	76.41	Personal taxes (contains some imputed values) [I]	81.72
Computer information services (internet) [I]	72.85	Housekeeping supplies [D]	52.12	Property taxes [I]	62.61
Dairy products [D]	67.39	Income after taxes [I]	99.89	Shelter [I]	97.82
Deductions for Social Security [I]	76.84	Income after taxes [I]	99.8	State and local income tax (imputed) [I]	62.18
Electricity (owned home) [I]	62.36	Lunch [D]	52.46	State and local income taxes [I]	62.18
Electricity [I]	91.71	Meats, poultry, fish, and eggs [D]	66.21	Telephone services [I]	87.77
Estimated market value of owned home [I]	63.69	Miscellaneous foods [D]	72.78	Utilities, fuels, and public services [I]	97.19
Estimated monthly rental value of owned home [I]	63.8	Money income before taxes [I]	99.76	Vehicle insurance [I]	73.09
Federal income tax (imputed) [I]	76.12	Net change in total liabilities [I]	57.89	Vehicle rental, leases, licenses, and other charges [I]	53.67
Federal income taxes [I]	76.12	Nonalcoholic beverages [D]	59.42	Wages and salaries [I]	75.23
Fresh fruits [D]	56.29	Other dairy products [D]	54.75	Water and other public services [I]	67.04
Fresh milk and cream [D]	50.46	Owned dwellings [I]	64.16	Water and sewerage maintenance [I]	59.79
Fresh vegetables [D]	56.36				

Ease of Use

- Mean values have a few advantages that medians do not have, mainly the ability to sum mean values of smaller groups to easily calculate the mean value of a larger group made up of the smaller groups.
- For example, adding the means of Fresh Milk, Cream, Butter, Cheese, Ice Cream and related products, other dairy products, and Miscellaneous diary products yields the mean of Diary products.
- In terms of general use, medians have no other functional downside compared to means and shouldn't be substituted in lieu of means.
- Medians should be used to complement means to show a more holistic picture of CE data. Let's look at a few examples of this.

Apples (UCC=110110)

2019 AllCUPrePub	Apples	2019 Diary Data	Apples	Apples (purchasers only)
Yrly Mean est.	$43.71	Yrly Median est.	$0	$207.48
Mean (purchasers)	$255.61	Qtly Median est.	$0 (10,682)[a]	$51.87 (1,908)[b]
Percent reporting	17.10%	Medians percent reporting		17.86% (b/a)

- To illustrate the problem of using a universal classification code (UCC) with a percent reporting under 50%, consider apples. The mean expenditure value for all consumer units (CU) on apples in 2019 was $43.71 while the median value was $0. However, the median of apple purchasers was $207.48 while the mean value of all purchasers was $255.61. (43.71/0.1710)
- What does this tell us about the distribution of apple consumption, how much apple purchasers pay, or what the average consumer spent on apples in 2019, etc.?

Cellular Phone Service (270102)

2019 AllCUPrePub	CPS	2019 Interview Data	CPS	CPS (purchasers only)
Yrly Mean est.	$1,218.08	Yrly Median est.	$760.00	$1,080.00
Mean (purchasers)	$1,583.98	Qtly Median est.	$190.00 (24,956)[a]	$270.00 (19,197)[b]
Percent reporting	76.90%	Medians percent reporting		76.92% (b/a)

- Cellular Phone Service (CPS) on the other hand did have a percent reporting above 50%. When taking the median value of CPS we get something much more representative than with apples. Our median is $760 while our mean is $1,218.08
- If you look closely at the data you will realize that the median value of CPS purchasers alone ($1,080.00) was actually significantly lower than the mean value of purchasers ($1,583.98)
- This shows that outlier values on the top end of our data or perhaps a heavy upper tail, skew our mean value in the extreme.

Fresh Milk and Cream

Fresh Milk, All types(090110) and Cream (090210)

2019 AllCUPrePub	FM&C	2019 Diary Data	FM&C	FM&C (purchasers only)
Yrly Mean est.	$140.39	Yrly Median est.	$51.48	$206.96
Mean (purchasers)	$278.22	Qtly Median est.	$12.87 (10,682)[a]	$51.74 (5,455)[b]
Percent reporting	50.46%	Medians percent reporting		51.07% (b/a)

- Unlike the prior examples, this category, Fresh Milk and Cream (FM&C) contains 2 UCC's as well as a percent reporting above 50%.
- When comparing the median value of $51.48 for all consumers to the mean value of $140.39, there is very noticeable difference. The mean of all consumers is almost 3 times larger than the median value of all consumers. However, the median value of purchasers, $206.96, is actually pretty similar to the mean value of purchasers, $278.22.
- In this case means can be a more accurate measure of centrality for FM&C, but medians can still be useful to check how much outliers are influencing the mean purchasers value. In this case, the mean purchasers value is not very different from the median purchasers value.

Income Before Taxes (980000)

2019 AllCUPrePub	Income before taxes	2019 Interview Data	Income before taxes
Yrly Mean est.	$82,852	Yrly Median est.	$45,478.60
Percent reporting	100%	Qtly Median est.	$11,369.65

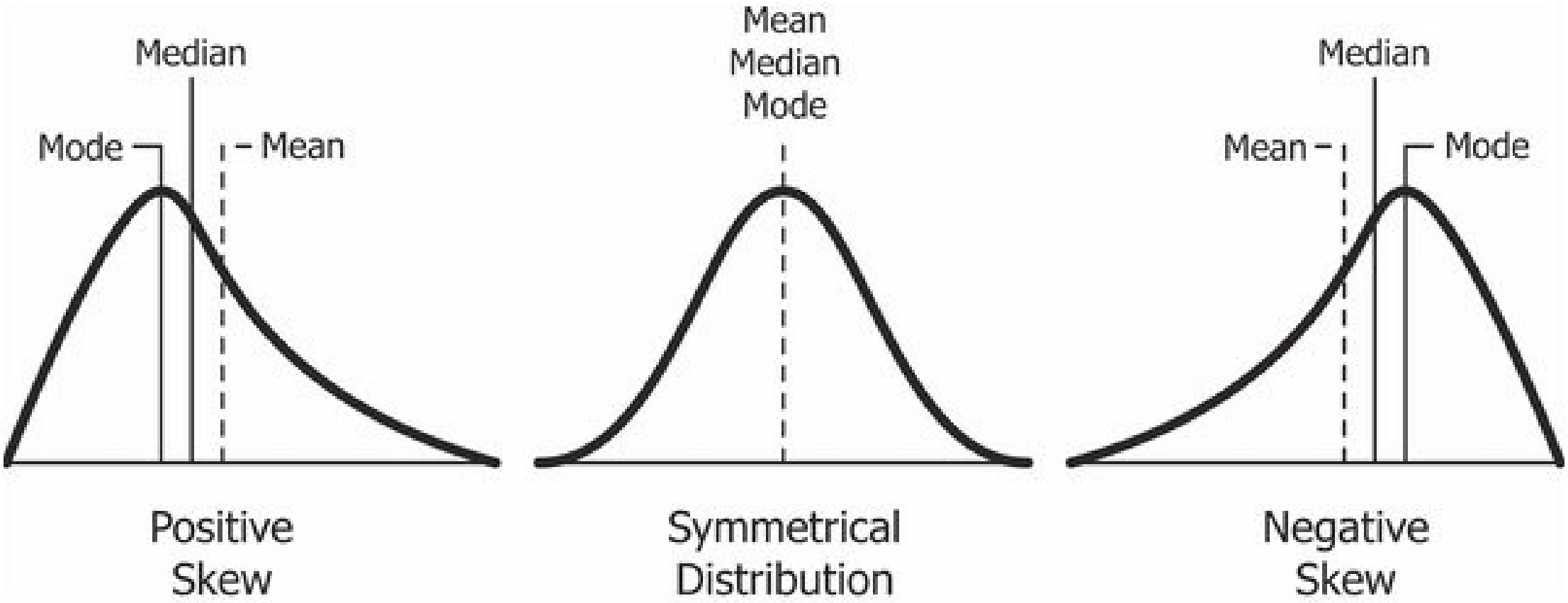
Income After Taxes (980017)

2019 AllCUPrePub	Income after taxes	2019 Interview Data	Income after taxes
Yrly Mean est.	$71,487	Yrly Median est.	$43,362.00
Percent reporting	100%	Qtly Median est.	$10,840.50

Findings

- In all examples our median values were smaller than our mean values, regardless of the percent reporting. This means our data sample has a positive skew.

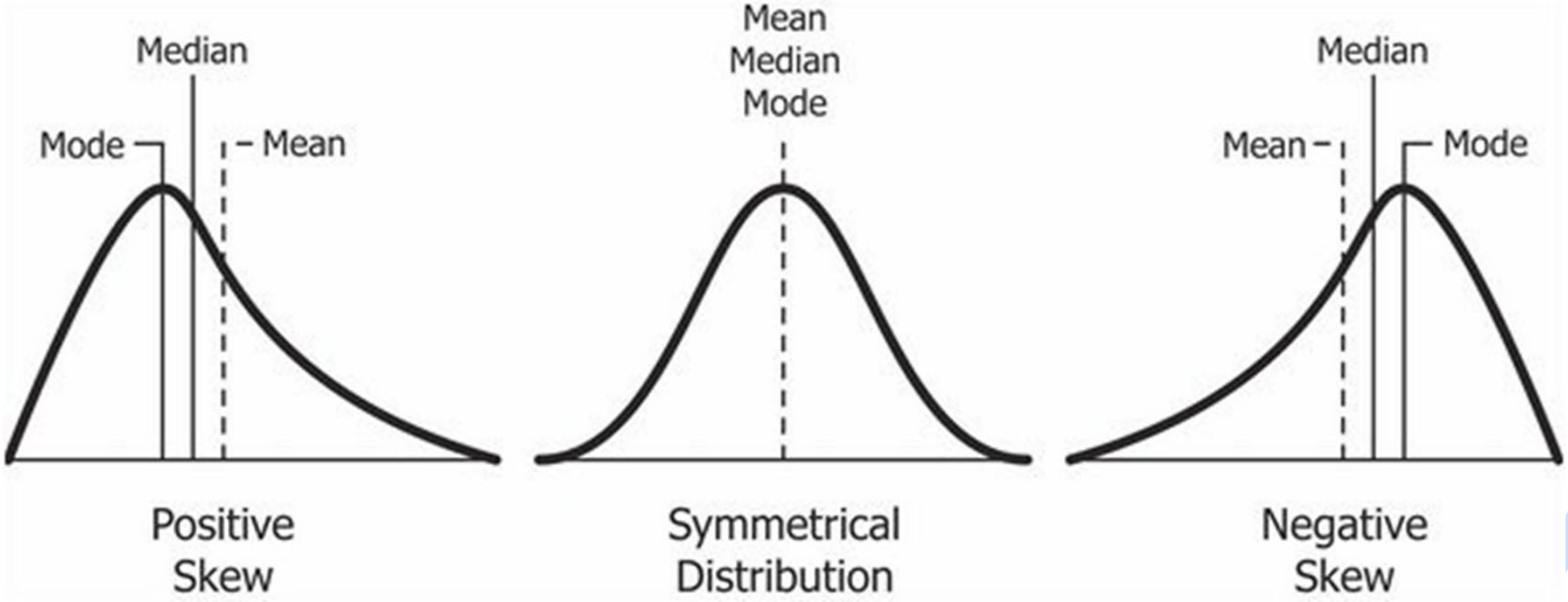
Chart 2



Findings

- Sometimes the median value of purchasers was higher or lower than the mean value of all consumer units, but the median value of purchasers was always lower than the mean value of purchasers, further highlighting the positive skew of our data samples as shown in chart 2.

Chart 2



Findings

- Since medians show lower values of consumption than means, how does this affect our reports?
- How does this shift affect the public perception of CE data and reports?
- Should certain published tables be republished with medians instead of means?
- Can we break populations up into subcategories and expect means to be an accurate reflections of these groups expenditures and other data points?

Conclusion

- How we frame our data and our findings is extremely important.
- By using new metrics (medians) along with current metrics (means) we can dive deeper into the makeup of US consumer expenditures, their distributions, skew, and resulting bias to show a more accurate representation of the data we are analyzing.
- More analysis will be needed to further compare median value changes from year to year to see how else we can frame our CE data as well as how the non normal distribution of our data affects the methodology and outcomes of CE's models.
- Hopefully you can apply this mindset to reexamine how you present your data or how you examine others in the future!

Sources

1. https:/www.cdc.gov/csels/dsepd/ss1978/lesson2/section8.html
2. https://www.bls.gov/cex/2019/research/allcuprepub.pdf
3. https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

Limitations of BLS Data

- All of the data calculations done in today's presentation were done using CE's current internal methodology including imputation, top coding, and model weighting.
- In future publications medians maybe published using different methodology than what was used in today's presentation.

Contact Information

Frank A. Cirillo, MS
Economist
Division of Consumer Expenditure Surveys
www.bls.gov/cex
202-691-6905
cirillo.frank@bls.gov