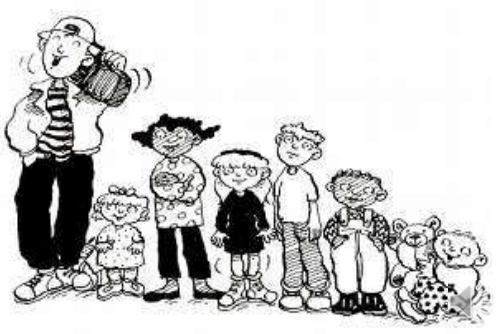
Descriptive Statistics

Frequency Distribution

- First step in describing a variable
- E.g.: Variable "child's age" is measured for 10 children: 3, 1, 2, 3, 6, 7, 6, 2, 8, 2
- Unique values are 1, 2, 3, 6, 7, 8 years old
- 1 – 1 case (10%)
  2 – 3 cases (30%)
  3 – 2 cases (20%)
  6 – 2 cases (20%)
  7 – 1 case (10%)
  8 – 1 case (10%)



Cumulative Distribution

- Only makes sense for ordinal and interval/ratio
- 1 – 1 case (10%) [10%]
  2 – 3 cases (30%) [40%]
  3 – 2 cases (20%) [60%]
  6 – 2 cases (20%) [80%]
  7 – 1 case (10%) [90%]
  8 – 1 case (10%) [100%]

3

Descriptive Statistics

- Distributions can be long and complicated
- To describe distributions concisely and to compare distributions, we use descriptive statistics

"Here's a list of 100,000 warehouses full of data. I'd like you to condense them down to one meaningful warehouse."

4

Percentage

- A percentage for each value = a statistic
- Percentages are very important statistics for <u>nominal</u> variables

WHEN STATISTICIANS ARGUE

5

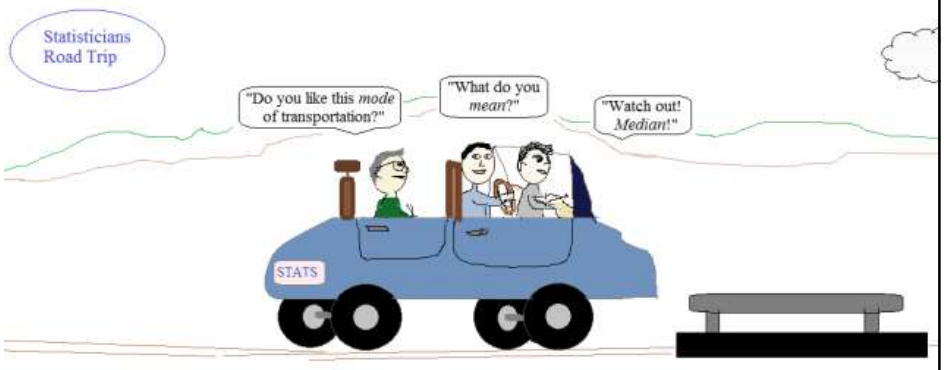
Two Key Characteristics of a Variable

<u>1. Central tendency (or average)</u> -- describes the typical value in the list

<u>2. Variability</u> -- describes spread, variation, the typical distance of numbers from the average

6

Measures of Central Tendency: Mean, Median, Mode



7

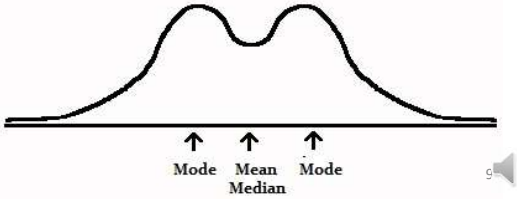
Mode: Example

- Frequency distribution for child's age variable:
- 1 – 1 case (10%)
  2 – 3 cases (30%) ← mode
  3 – 2 cases (20%)
  6 – 2 cases (20%)
  7 – 1 case (10%)
  8 – 1 case (10%)

8

Mode: When to Use

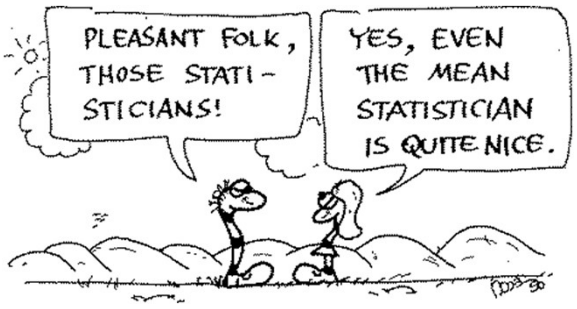
- Mode can be used with any level of measurement
- But it's not as helpful for interval/ratio, especially if more than 10-15 discrete values
- Mode for interval/ratio variables: can be useful if bimodal distribution



9

Mean

- Mean of list = sum of all numbers / how many numbers: $\bar{X} = \frac{\sum X}{n}$
- Mean can only be used with interval/ratio variables, not with nominal or ordinal



10

Example Using Frequency Distribution

- We multiply each value by its frequency:
  - 1 – 1 case: 1 x 1 = 1
  - 2 – 3 cases: 2 x 3 = 6
  - 3 – 2 cases: 3 x 2 = 6
  - 6 – 2 cases: 6 x 2 = 12
  - 7 – 1 cases: 7 x 1 = 7
  - 8 – 1 case: 8 x 1 = 8
- Add all these products and divide by total number of cases: Mean of X = (1 + 6 + 6 + 12 + 7 +8)/10=4

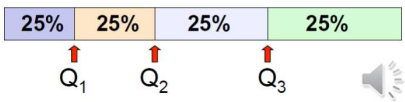
Median

- Splits the sorted list of values in half; the second quartile (Q_2), the 50th percentile
- List of children's ages, sorted:

  1, 2, 2, 2, 3 | 3, 6, 6, 7, 8. Median=3
- Odd number of cases:

  1 3 4 4 5 6 7 7 7. Median=5
- Different numbers in the middle:

  3 4 4 5 6 | 7 8 8 9 10. Median=6.5 (some use 6)



Based on Cumulative Distribution

- Median is the point where we reach or exceed 50% -- here, it's 3, we exceed 50%.

1 – 1 case (10%) [10%]
2 – 3 cases (30%) [40%]
3 – 2 cases (20%) [60%]
6 – 2 cases (20%) [80%]
7 – 1 case (10%) [90%]
8 – 1 case (10%) [100%]

Cumulative Distribution and Median: Another Example

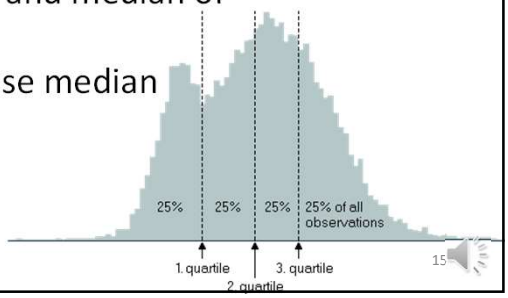
3 – 1 case (10%) [10%]
4 – 2 cases (20%) [30%]
5 – 1 case (10%) [40%]
6 – 1 case (10%) [50%]
7 – 1 case (10%) [60%]
8 – 2 cases (20%) [80%]
9 – 1 case (10%) [90%]
10 – 1 case (10%) [100%]

- Median=6 (we reached 50%)
- Compare to sorted list result: 6.5

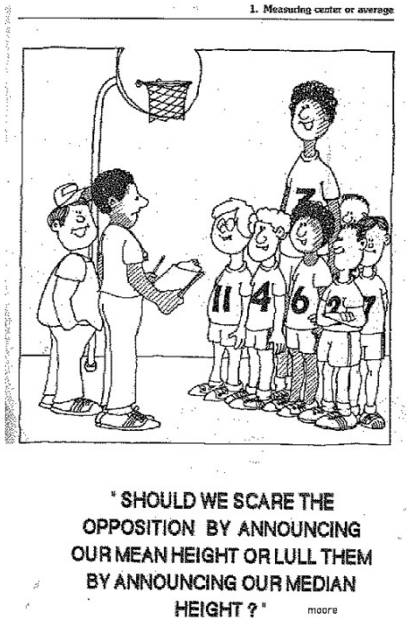
14

Median

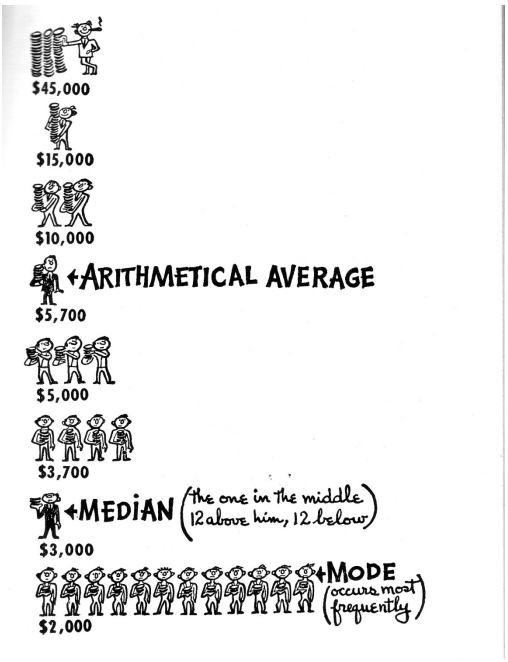
- Median can be used with ordinal or interval/ratio, <u>not</u> with nominal
- Mean vs median choice for <u>interval/ratio</u> variables:
  - Mean is very sensitive to extreme values, median isn't
  - Example: Find the mean and median of 1, 2, 3, 4, 90
  - So if you have outliers, use median



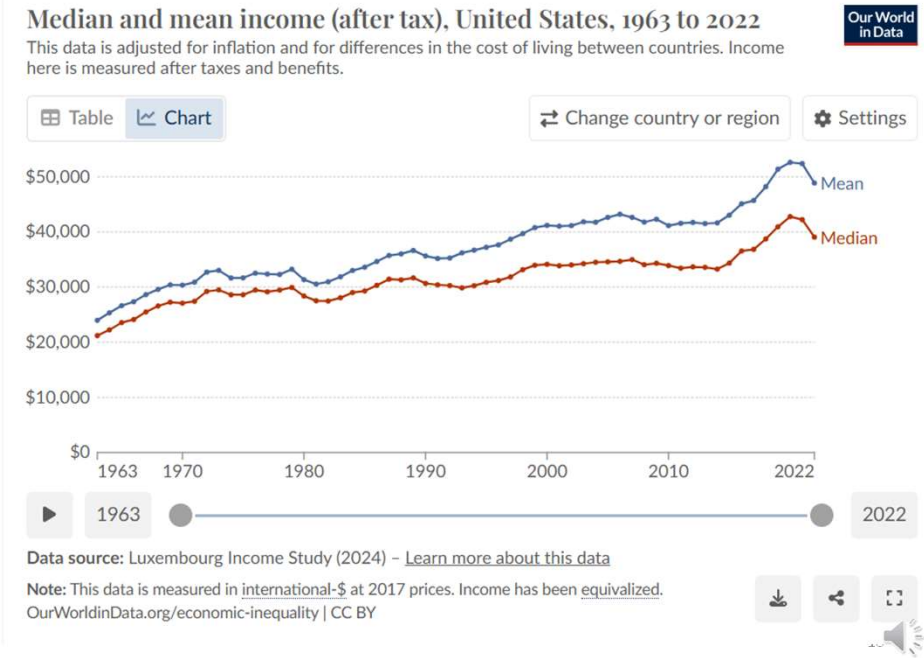
15



16


$45,000
$15,000
$10,000
+ARITHMETICAL AVERAGE
$5,700
$5,000
$3,700
+MEDIAN (the one in the middle 12 above him, 12 below)
$3,000
+MODE (occurs most frequently)
$2,000

17


Median and mean income (after tax), United States, 1963 to 2022
This data is adjusted for inflation and for differences in the cost of living between countries. Income here is measured after taxes and benefits.
Our World in Data
Table
Chart
Change country or region
Settings
$50,000
$40,000
$30,000
$20,000
$10,000
$0
Mean
Median
1963
1970
1980
1990
2000
2010
2022
1963
2022
Data source: Luxembourg Income Study (2024) – Learn more about this data
Note: This data is measured in international-$ at 2017 prices. Income has been equivalized.
OurWorldinData.org/economic-inequality | CC BY

Summary

- Nominal → mode only, NEVER mean or median
- Ordinal → median or mode; mean is sometimes used but that's not technically correct
- Interval/ratio → mean (if no outliers) or median (if outliers or skew); mode can be used but often not as useful, especially if there are many values

19 

Why Do We Need to Measure Variability?

- Consider three sets of values
  - List 1: 1, 5, 9
  - List 2: 4, 5, 6
  - List 3: 5, 5, 5
- Mean=5 but very different

20 

Variation Ratio

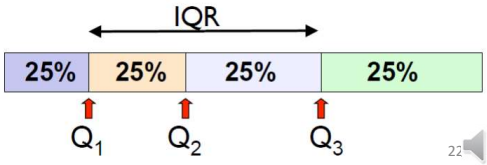
- Rarely used, only for nominal variables
- Percentage of cases NOT in modal category

21

Range and Interquartile Range

- Describes the spread of the distribution
- Used for ordinal or interval/ratio data
- To find Q1 and Q3, use the same approaches as for median (sorted list or cumulative distribution)
- E.g.: 1, 2, 2, 2, 3 | 3, 6, 6, 7, 8

  Range = 8-1=7 IQR= 6-2=4



22

Standard Deviation

- Describes the typical distance of numbers from the mean
- Not applicable to nominal level variables!
- Deviation vs Standard Deviation:
  - Deviation – distance between one number and the mean
  - Standard deviation – typical (average) distance

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

23

Variance

- Variance = squared standard deviation
- Measured in squared units → the number is not as meaningful as standard deviation
- But useful for further statistical analyses

$$s^2 = \frac{\sum(x - \bar{x})^2}{n-1}$$

24

Example Using Frequency Distribution

- List: 3, 1, 2, 3, 6, 7, 6, 2, 8, 2. Mean=4.

X	Cases	$X - \bar{X}$	$(X - \bar{X})^2$	$(X - \bar{X})^2 * \text{Cases}$
1	1	1 – 4 = -3	9	9 * 1 = 9
2	3	2 – 4 = -2	4	4 * 3 = 12
3	2	3 – 4 = -1	1	1 * 2 = 2
6	2	6 – 4 = 2	4	4 * 2 = 8
7	1	7 – 4 = 3	9	9 * 1 = 9
8	1	8 – 4 = 4	16	16 * 1 = 16
Σ				56

- Step 3: 56/(10-1)=6.22 ← variance
- Step 4: sqrt(6.22) = 2.49 ← standard deviation

Descriptive Statistics by Level of Measurement

X	Nominal	Ordinal	Interval/Ratio
Percentage	Yes	Yes	Yes (often not useful)
Mean	No	No (sometimes)	Yes
Median	No	Yes	Yes
Mode	Yes	Yes	Yes (often not useful)
Variation ratio	Yes	Yes	Yes (but not useful)
Range	No	Yes	Yes
Interquartile range	No	Yes	Yes
Standard deviation	No	No (sometimes)	Yes
Variance	No	No (sometimes)	Yes