

# Basics of Statistics :

Understanding your Data

Seton Office of Research Administration  
September 14, 2009

Christine A. Jesser, ScD  
Senior Epidemiologist

Seton Family of Hospitals



# Overview

- Types of data
- Descriptive methods for categorical data
- Descriptive methods for continuous data
- Data distribution

# Types of Data

An investigator asks,

“I have collected data from 200 people.

All the data are between zero and four. How should I summarize the data?”

# Types of Data

- Not enough information yet
- Statisticians encounter many different types of numeric data
- Need to know:
  - what the data represent
  - where they came from
  - what the investigator wants to learn

# Types of Data

- Do the numbers represent
  - Arbitrary codes for unordered categories?
  - Ordered categories?
  - Count data?
  - Measurements?

# Types of Data

- Nominal data
  - Values represent unordered categories
  - Informally called categorical data
  - Example (simplest case-2 categories)

0 -- male

1 -- female

# Types of Data

- Nominal data

- Example 2: Race

- 0 -- did not answer

- 1 -- black

- 2 -- white

- 3 -- Asian

- 4 -- other

# Types of Data

- Ordinal Data

- Values represent ordered categories
- Example – toxicity grades

0 -- none

1 -- mild

2 -- moderate

3 -- severe

4 -- life-threatening

# Types of Data

- Note that there is a natural ordering for the ordinal data
- No natural ordering for nominal data

# Types of Data

- Discrete data
  - Both ordering and magnitude are important
  - Numbers represent measurable quantities
  - Can take on only specified values that differ by fixed amounts
  - Often count data (e.g., number of hospitalizations)

# Types of Data

- Discrete data
- Example – number of risk factors
  - 0 -- no risk factors
  - 1 -- one risk factor
  - 2 -- two risk factors
  - 3 -- three risk factors
  - 4 -- four risk factors

# Types of Data

- Continuous data
  - Represents measurable quantities
  - Not restricted to specified values
  - Fractional values possible
  - Addition/subtraction can be applied
  - Limiting factor is degree of precision

# Types of Data

- Continuous data examples:
  - Temperature
  - Time
  - Weight
  - Cholesterol level
  - Concentration of fluoride in drinking water

# Types of Data

- Continuous data example
  - Concentration level of fluoride (ppm)
  - Values may get arbitrarily large, but in our sample we observe values ranging from 0.07 to 3.63
  - Values represent a measured quantity

# Types of Data

- Continuous data
  - Interval scales
    - Spacing between values meaningful
    - Zero value is arbitrary
  - Ratio scales
    - Spacing between values meaningful
    - Zero value is meaningful

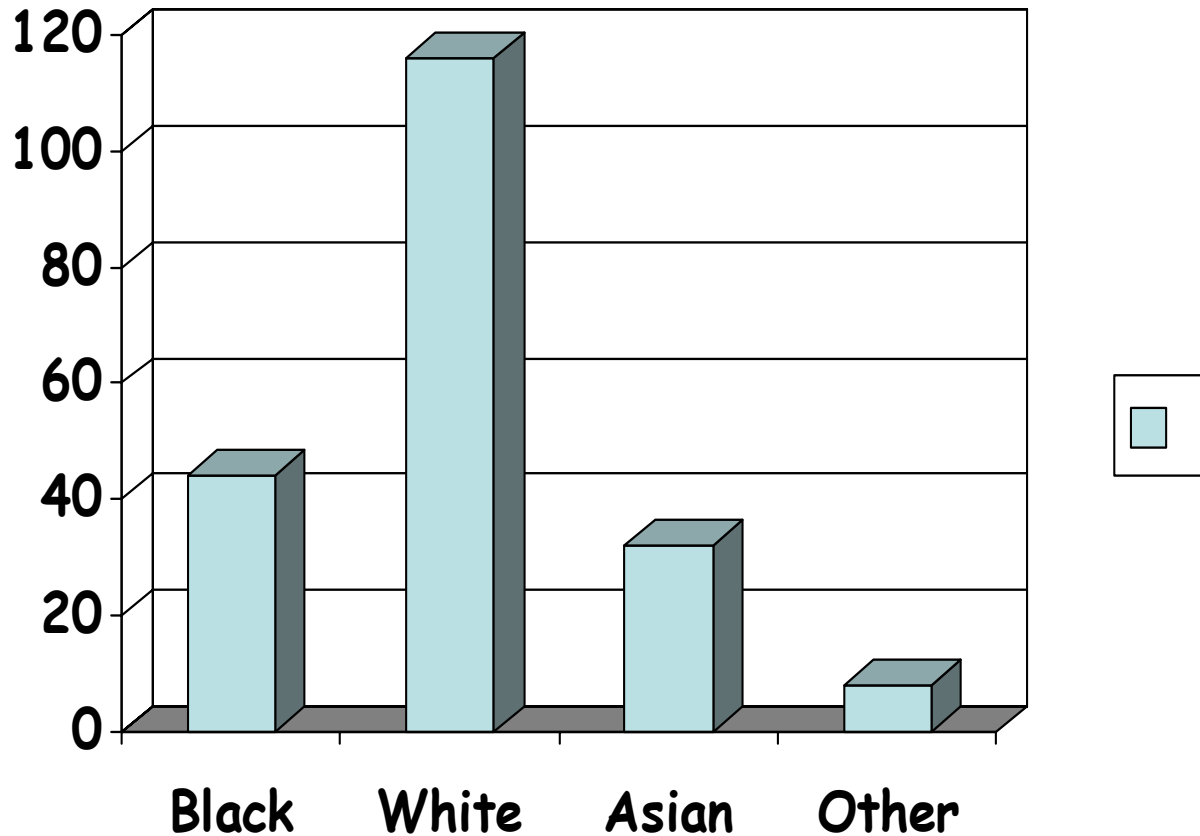
# Types of Data

- Continuous data examples:
  - Temperature – interval scale
  - Time – ratio scale
  - Weight – ratio scale
  - Cholesterol level – ratio scale
  - Concentration of fluoride in drinking water – ratio scale

# Nominal Data

<u>Race</u>	<u>N</u>
Black	44
White	116
Asian	32
Other	8

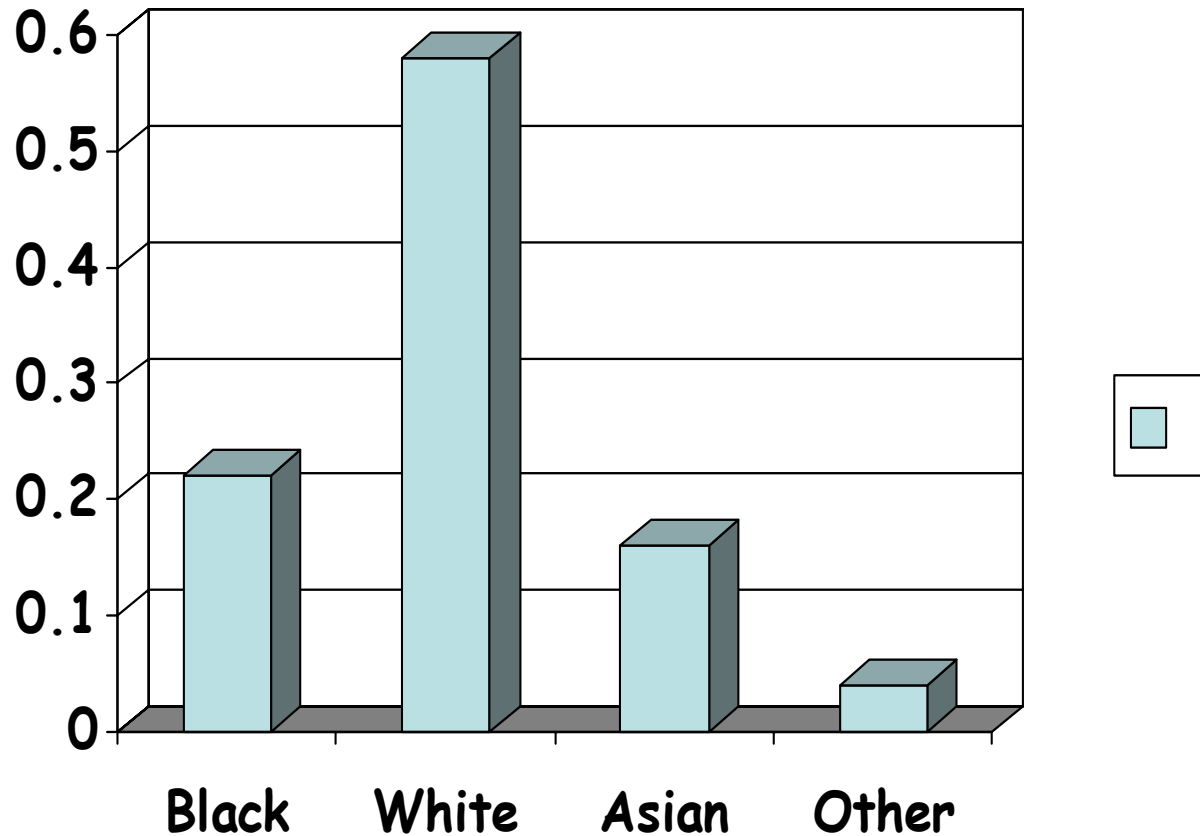
# Histogram



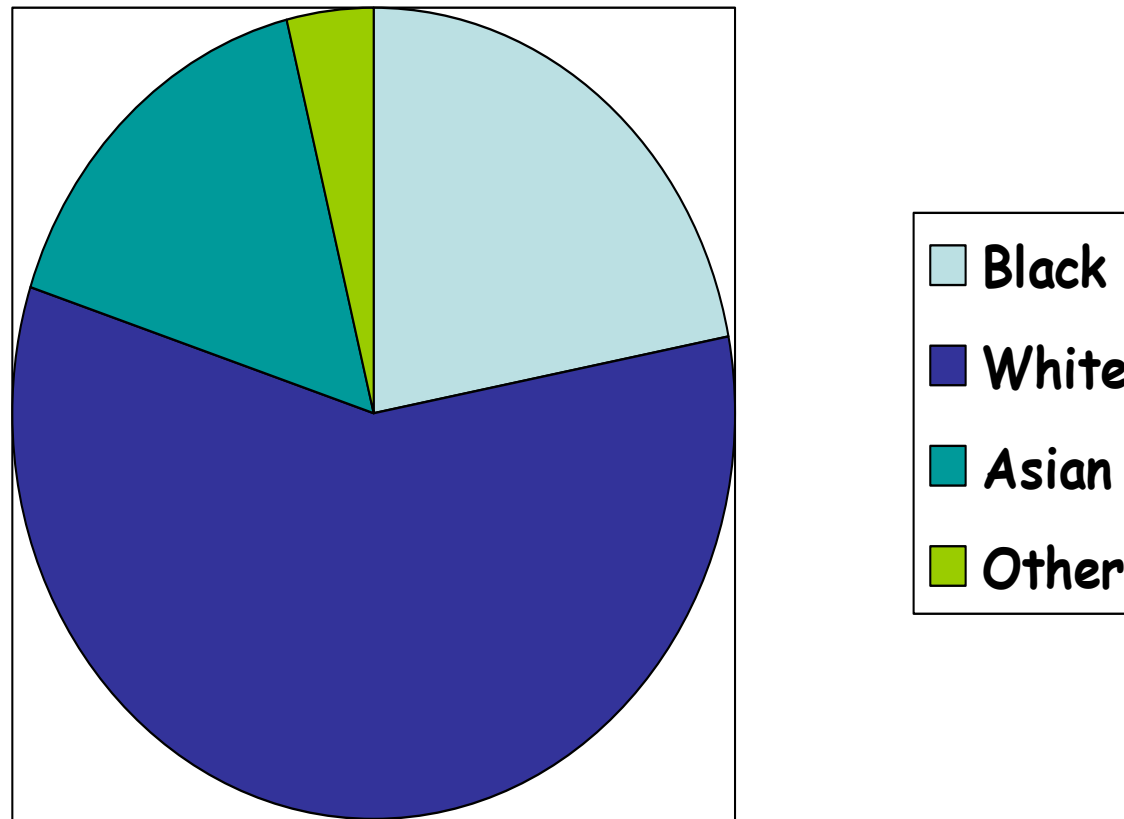
# Nominal Data

<u>Race</u>	<u>N</u>	Relative <u>frequency</u>
Black	44	.22
White	116	.58
Asian	32	.16
Other	8	.04
Total	200	1.00

# Relative Frequency Diagram



# Pie Chart



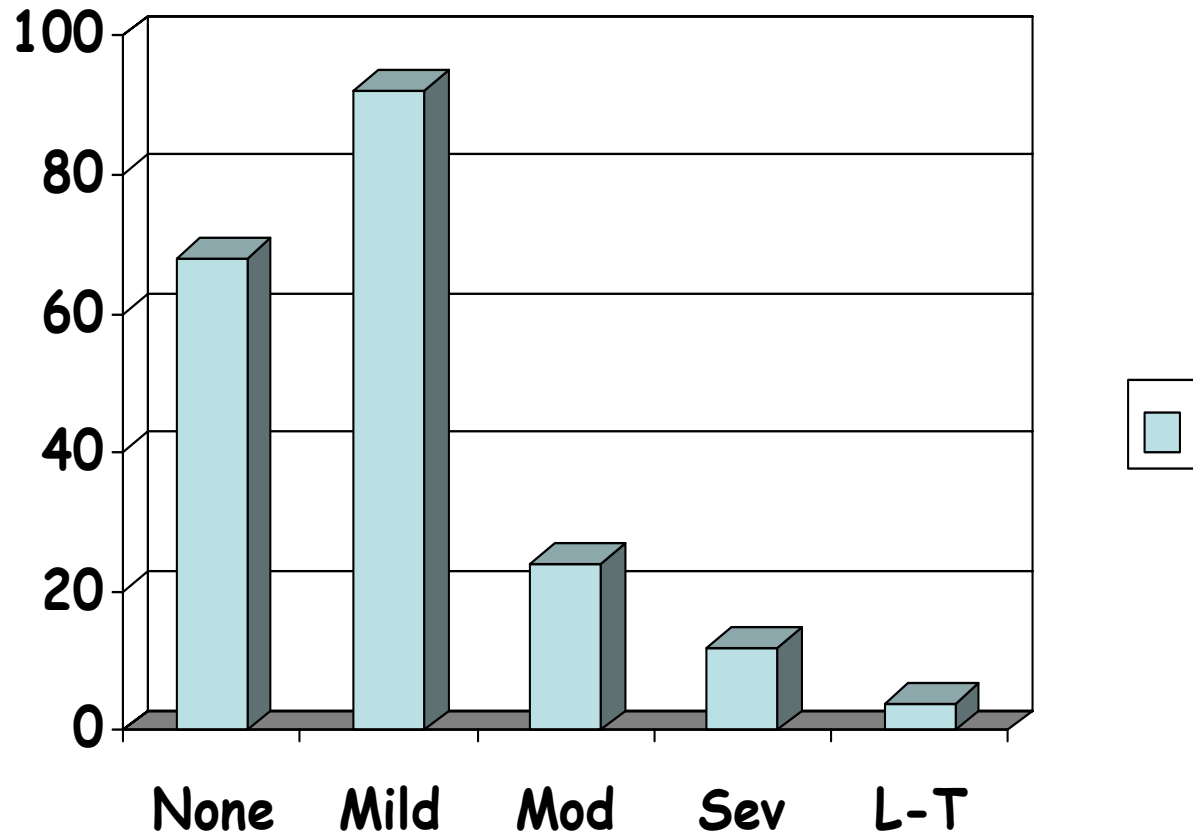
# Data Presentation

- Nominal data
  - Limited options for presenting data
    - Counts in each category
    - Histograms
    - Relative frequencies
    - Relative frequency diagram

# Ordinal Data

<u>Toxicity</u>	<u>N</u>
None	68
Mild	92
Moderate	24
Severe	12
Life-threatening	4

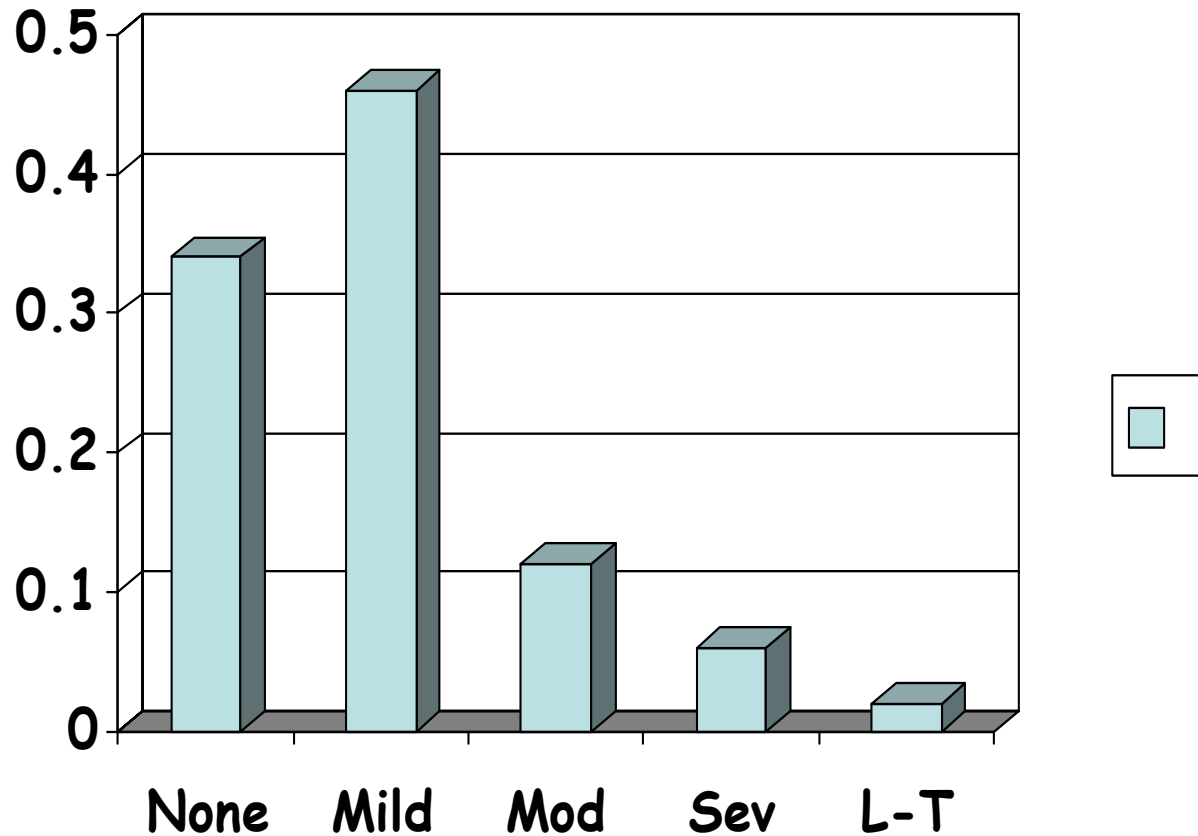
# Histogram



# Ordinal Data

<u>Toxicity</u>	<u>N</u>	<u>Relative_Freq</u>
None	68	.34
Mild	92	.46
Moderate	24	.12
Severe	12	.06
Life-threatening	4	.02

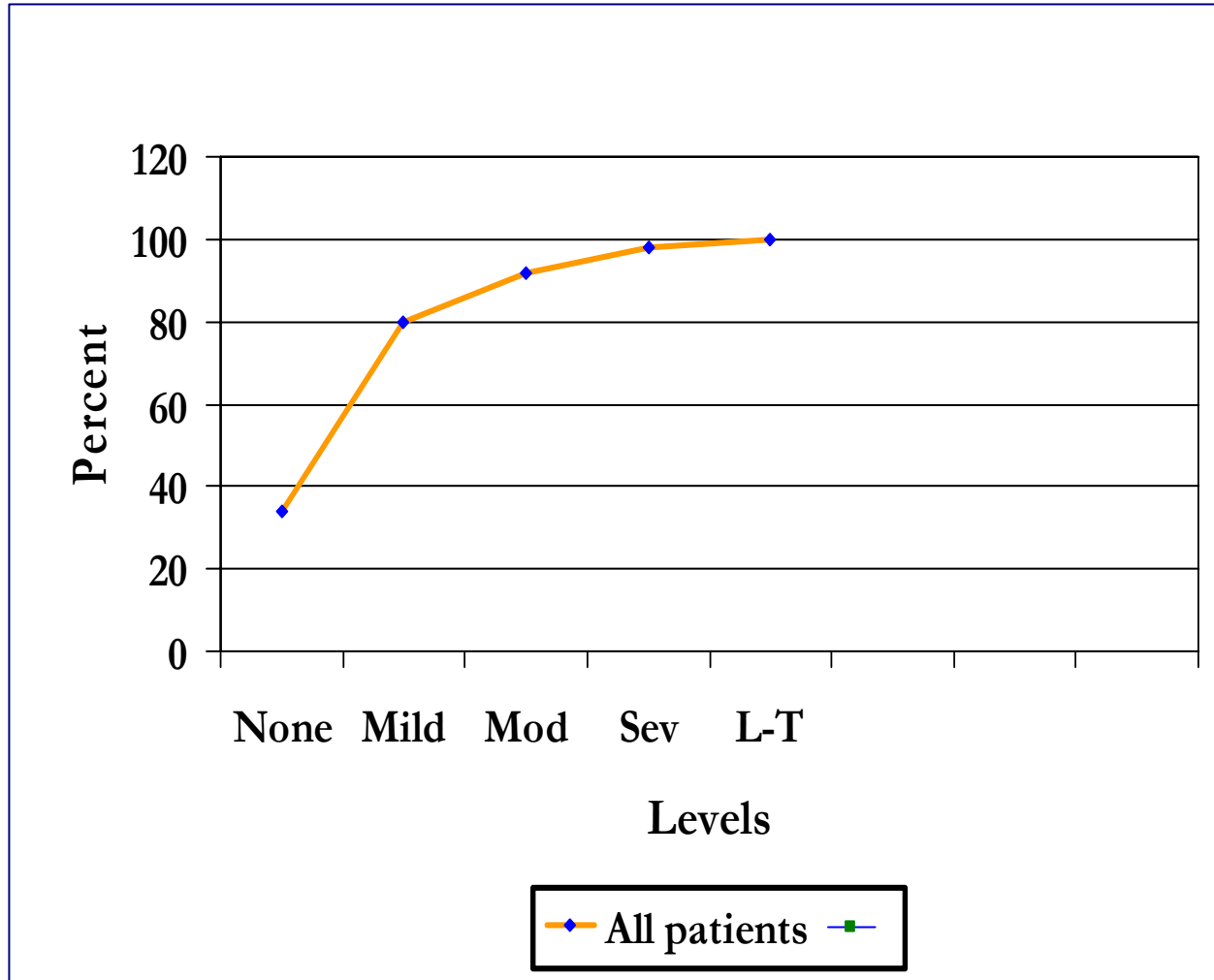
# Relative Frequency Diagram



# Ordinal Data

<u>Toxicity</u>	<u>N</u>	Relative <u>Freq</u>	<u>Cumulative</u> <u>Freq</u>
None	68	.34	.34
Mild	92	.46	.80
Moderate	24	.12	.92
Severe	12	.06	.98
Life-threatening	4	.02	1.00

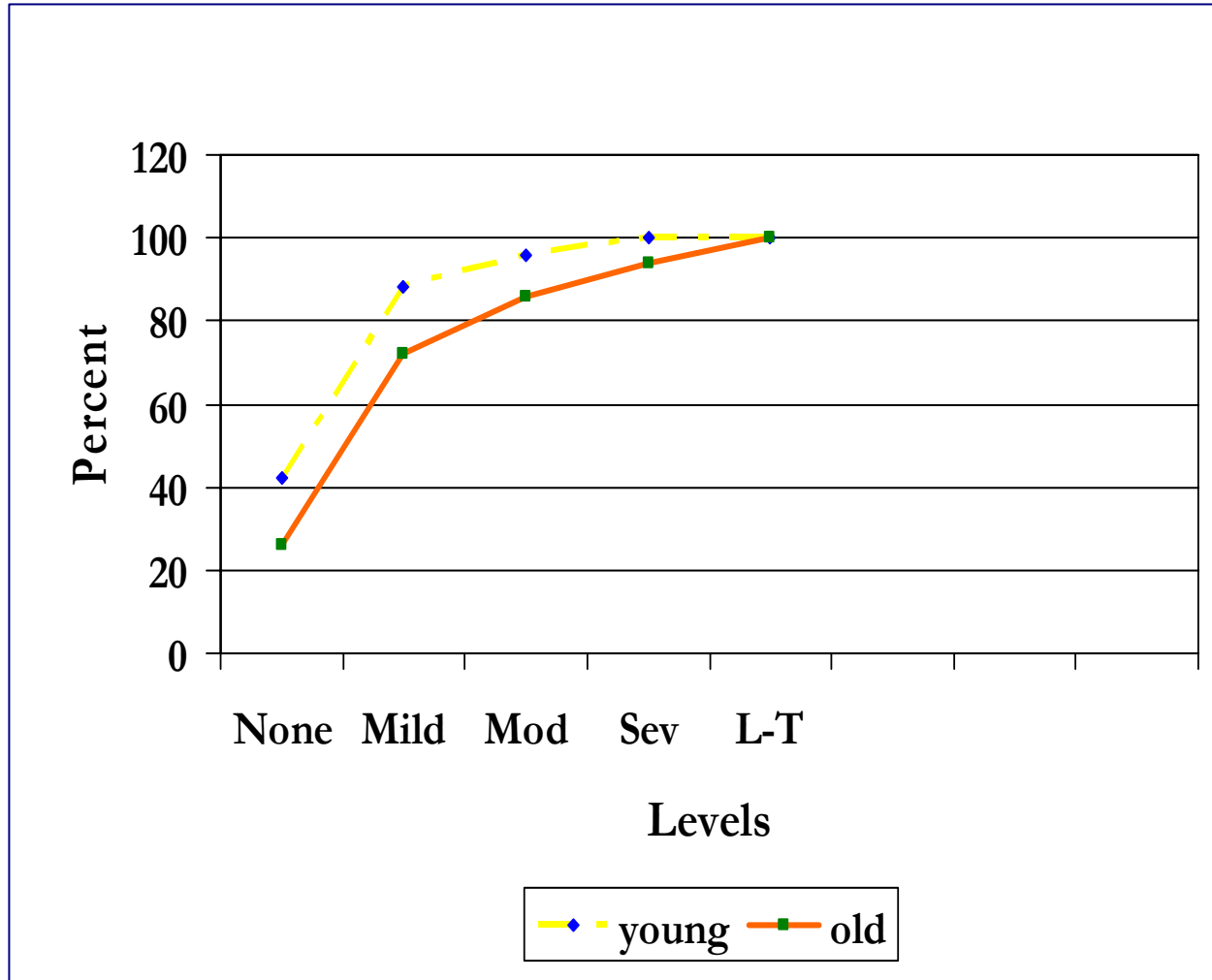
# Cumulative Frequency Polygon



## Ordinal Data

- Since the data are ordered, we can also use percentiles
- Percentiles divide a distribution into equal or ordered parts
- The ***median*** (50<sup>th</sup> percentile) divides a population into 2 equal-size parts (more on medians in a couple minutes)
- The median toxicity is mild

# Cumulative Freq by Age



# Data Presentation

- Ordinal data
  - All the tools for nominal data
  - Additional options for presenting data
    - Cumulative frequencies
    - Cumulative frequency polygons
    - Percentiles

# Discrete Data

Number of <u>Risk Factors</u>	<u>N</u>
0	62
1	86
2	32
3	14
4	6

## Discrete Data

- The values of the observations now represent actual numeric values
- We can calculate a mean (simple arithmetic average)
- Mean = sum of observation / N

## Discrete Data

- In our example,

$$\begin{aligned}\text{Mean} &= [(0 \times 62) + (1 \times 86) + (2 \times 32) \\ &\quad + (3 \times 14) + (4 \times 6)] / 200 \\ &= [0 + 86 + 64 + 42 + 24] / 200 \\ &= 216 / 200 \\ &= 1.08 \text{ risk factors}\end{aligned}$$

# Data Presentation

- Discrete data
  - All the tools for ordinal data
  - Additional options for presenting data
    - means

# Continuous Data

- Data on fluoride levels (1<sup>st</sup> 15 values)

0.079	0.146	0.112
0.071	1.335	0.072
0.224	0.553	0.071
0.159	0.415	0.119
2.467	0.288	0.154

# Data Presentation

- Continuous data
  - Wide variety of options
    - numerical
    - graphical
  - Important to consider both central tendency and dispersion
  - Examining the distribution is important

## Measures of Central Tendency

- We can calculate the sample mean as before, i.e., add all the observations together and divide by  $N$ , the number of observations
  - Mean = 0.697

# Measures of Central Tendency

- To calculate the median, we order observations from smallest to largest.
- If  $N$  is odd, define  $j = (N + 1) / 2$ . The  $j$ th ordered observation is the median
- If  $N$  is even,  $j = N / 2$ . The median is the average of ordered observations  $j$  and  $j + 1$ .
- In the example, the 100<sup>th</sup> and 101<sup>st</sup> ordered observations are 0.367 and 0.373 so the median is 0.370

## Measures of Central Tendency

- Suppose we have 9 observations:  
2.60, 2.75, 2.89, 4.05, 2.25,  
2.68, 3.00, 4.02, 2.85
- The mean is  $27.09 / 9 = 3.01$
- The median is 2.85

## Measures of Central Tendency

- Suppose the 8<sup>th</sup> observation were 40.02 instead of 4.02  
2.60, 2.75, 2.89, 4.05, 2.25,  
2.68, 3.00, 40.02, 2.8
- The mean is now  $63.09 / 9 = 7.01$
- The median is 2.85

# Measures of Dispersion

- It is also important to something about the variability of the data
- Common measures include:
  - Range: difference between the largest and smallest observations
  - Interquartile range: difference between the 25<sup>th</sup> and 75<sup>th</sup> percentiles
  - Standard deviation: square root of the variance

## Standard Deviation

	Individual Deviations	Individual Deviations- Squared
x	x - mean	(x - mean) <sup>2</sup>
2.60	-0.41	0.1681
2.75	-0.26	0.0676
2.89	-0.12	0.0144
4.05	1.04	1.0816
2.25	-0.76	0.5776
2.68	-0.33	0.1089
3.00	-0.01	0.0001
4.02	1.01	1.0201
2.85	-0.16	0.0256
total	0.00	3.0640

# Standard Deviation

- Variance is the sum of the squared deviations divided by (N-1)
  - Variance =  $3.0640 / 8 = 0.383$
- Standard deviation is the square root of the variance
  - s.d. = 0.619

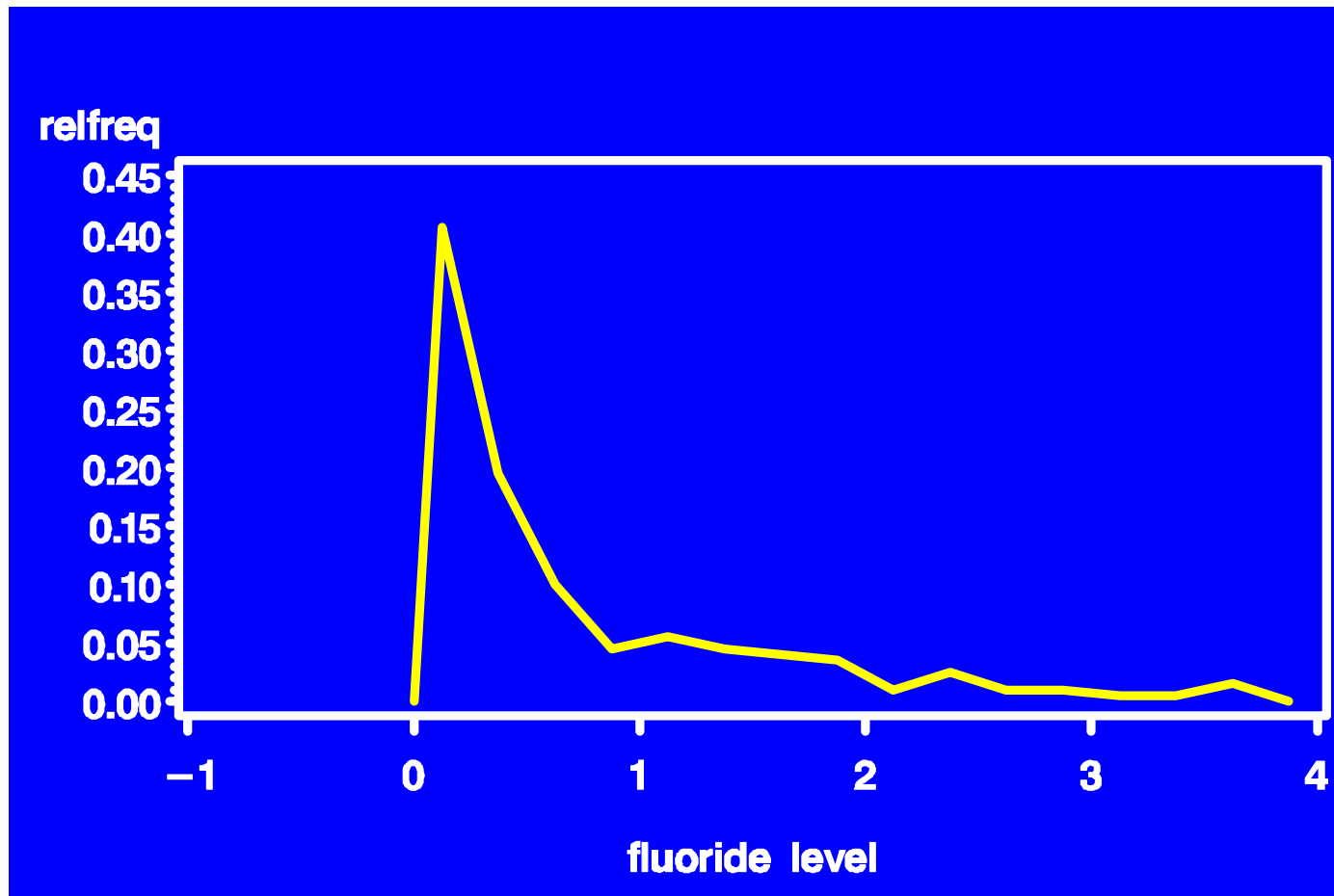
# Measures of Dispersion

- Getting back to the fluoride data
  - Range =  $3.627 - 0.070 = 3.557$
  - IQR =  $1.054 - 0.135 = 0.919$
  - s.d. =  $0.797$

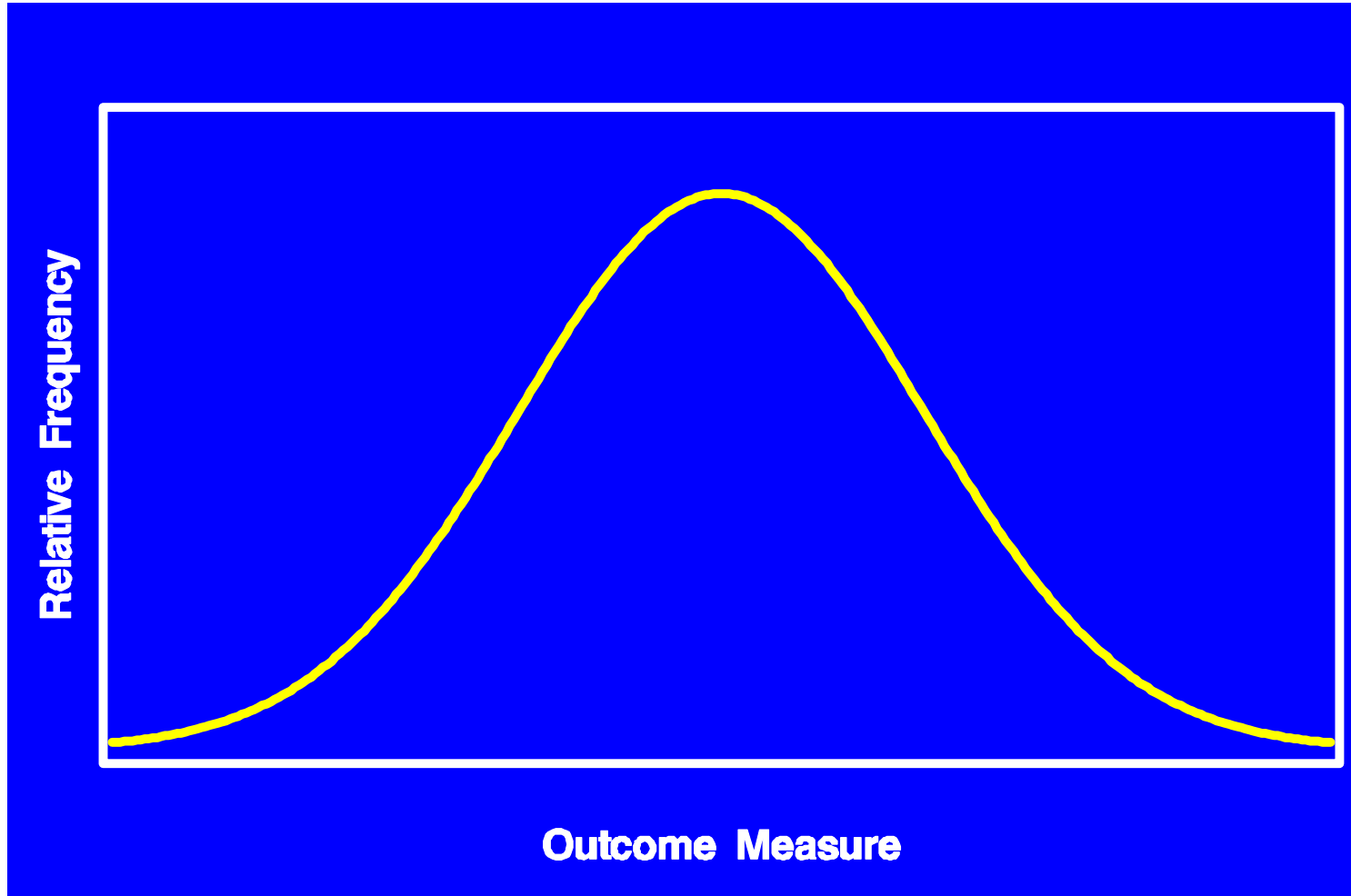
# Continuous Data

<u>Fluoride (ppm)</u>	N
0 – 0.24	81
0.25 – 0.49	39
0.5 – 0.74	20
0.75 – 0.99	9
1.0 – 1.24	11
1.25 – 1.49	9
1.5 – 1.99	15
2.0 – 2.99	11
3.0 – 3.99	5

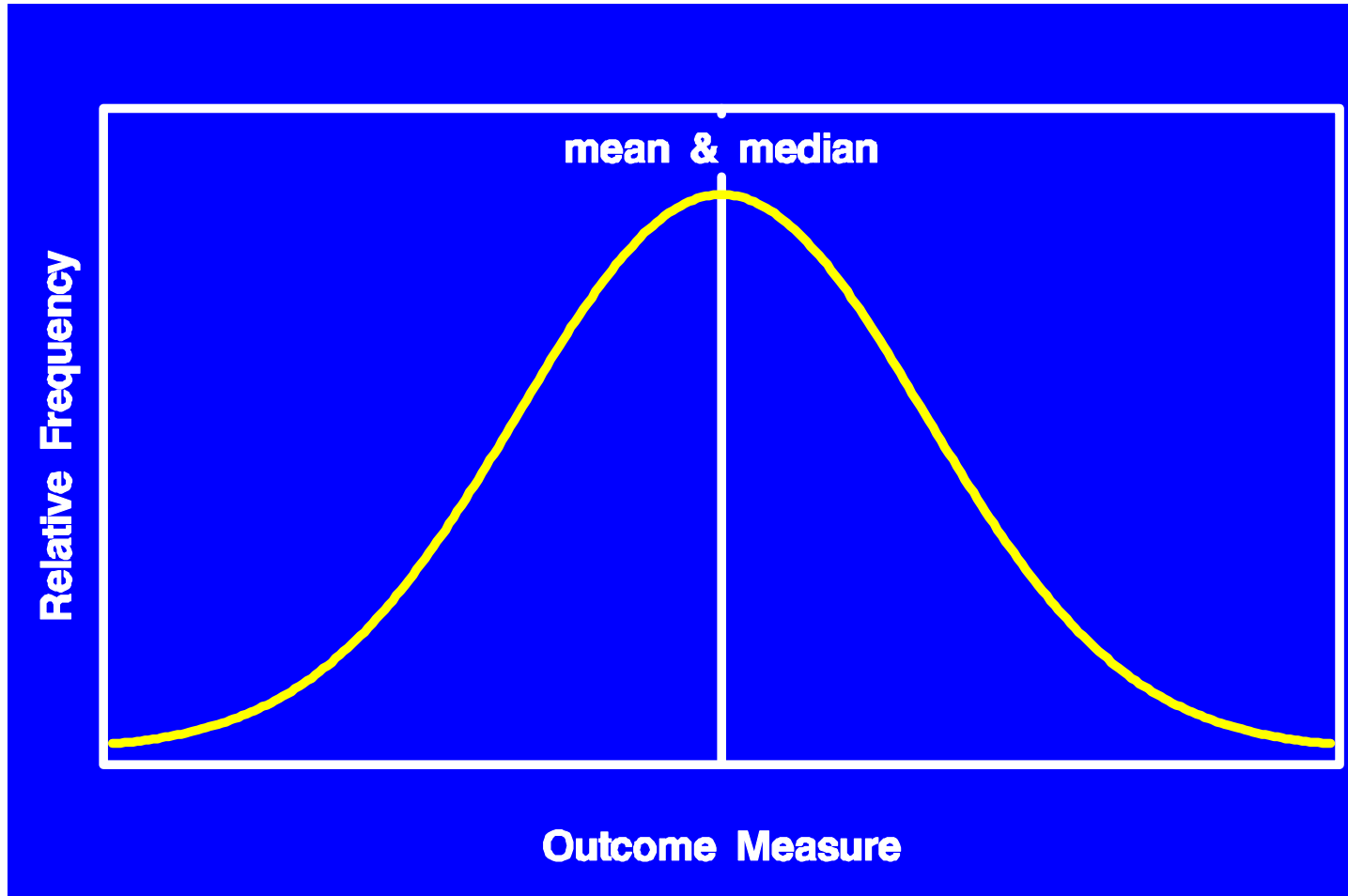
# Density Function



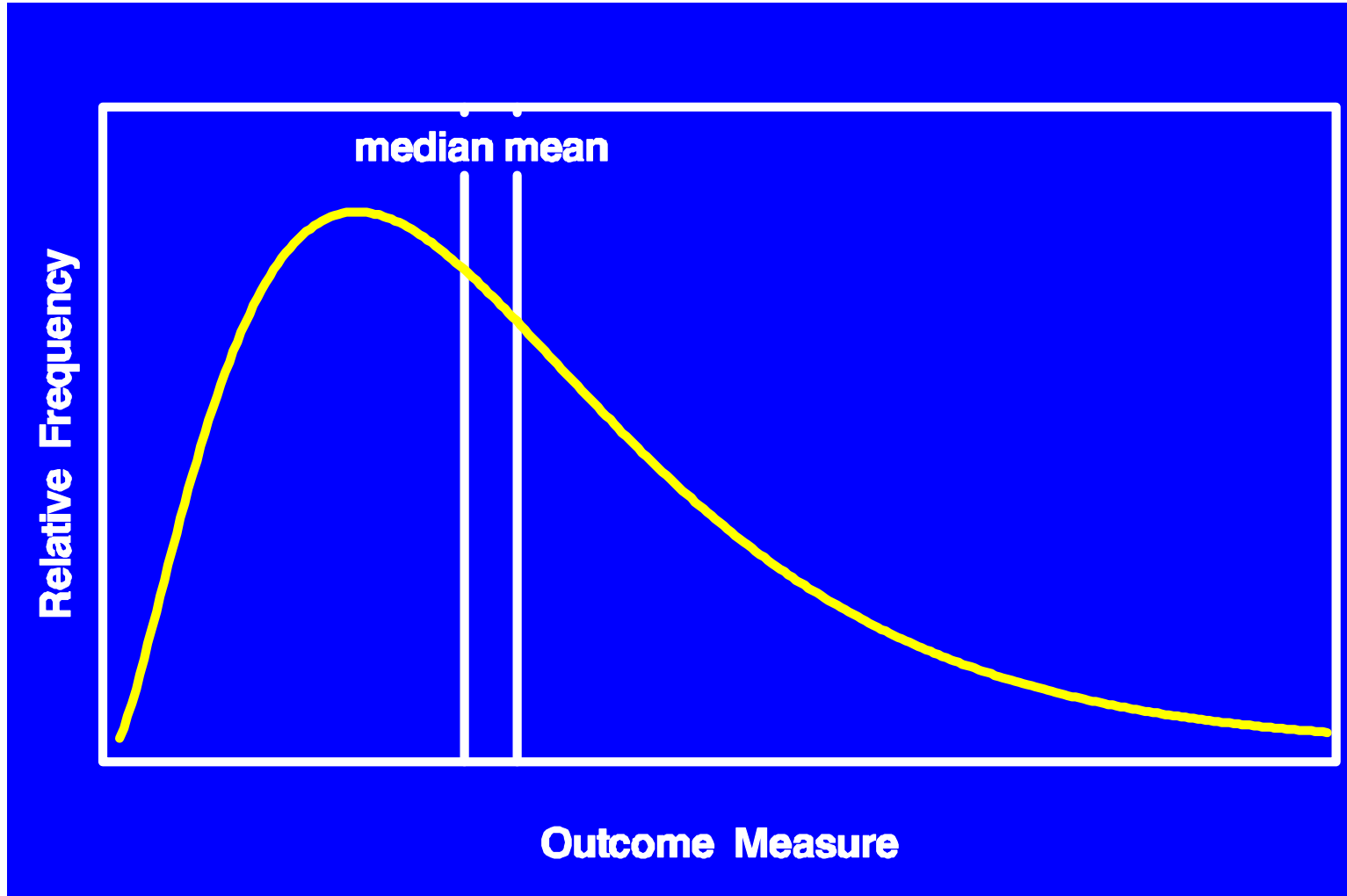
# Symmetric Density



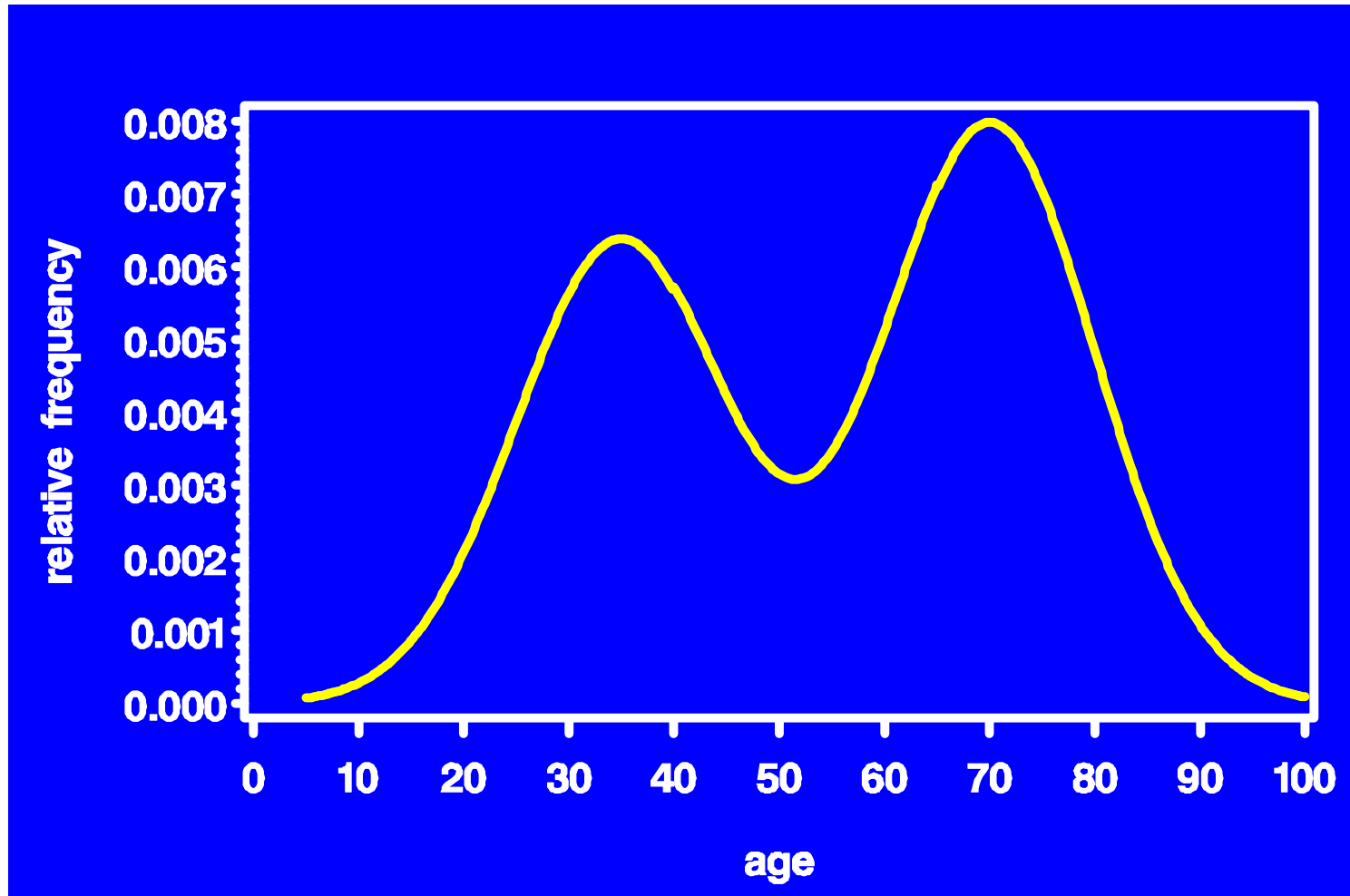
# Symmetric Density



# Skewed Distribution



# Bimodal Distribution



# Original Distribution

