Data Types and Distributions

Dr Celia Gregson
Consultant Senior Lecturer, Musculoskeletal Research Unit
University of Bristol, UK

#SAMSON19

www.theSAMSON.org

Day 1; Monday 11th

University of BRISTOL

SAMSON Sub-Saharan African MuSculOskeletal Network

By the end of this session you should be able to…

1. Distinguish between quantitative (continuous) and categorical variables
2. Know how to summarise quantitative and categorical data
3. Understand some basic properties of the normal distribution
4. Understand issues behind converting continuous variables into categorical variables

Statistical Approach to Epidemiology

- Focus of epidemiology is on populations not individuals
- Examine groups of data, rather than individual data
- Use summary statistics to characterise groups
- Formally assess whether differences between groups are meaningful

Statistical Approach - Quantitative vs. Categorical Variables

- Determines how to summarise data
  - Tables (means vs. proportions)
  - Graphs ('histogram' vs. bar graphs)
- Influences choice of "statistical test"
  (statistical inference will be covered in a later session!)
  - continuous variable → Student t-test
  - categorical variable → chi-squared test

Quantitative/continuous data

- Defined as "data in numerical quantities such as continuous measurements or counts" (Last, 1995)
- Can take on any value within a pre-specified range of values (includes any *real* number)

  0, 1, 1.232, 5.24, 12, 34.98… 100
- In practice, this definition also includes *integers* (i.e., whole numbers) and counts

  0, 1, 2, 10, 20… 100

Examples of continuous variables

- Age (years)
- Height (cm)
- Weight (kg)
- Blood pressure (mm Hg)
- Number of prescriptions (counts)

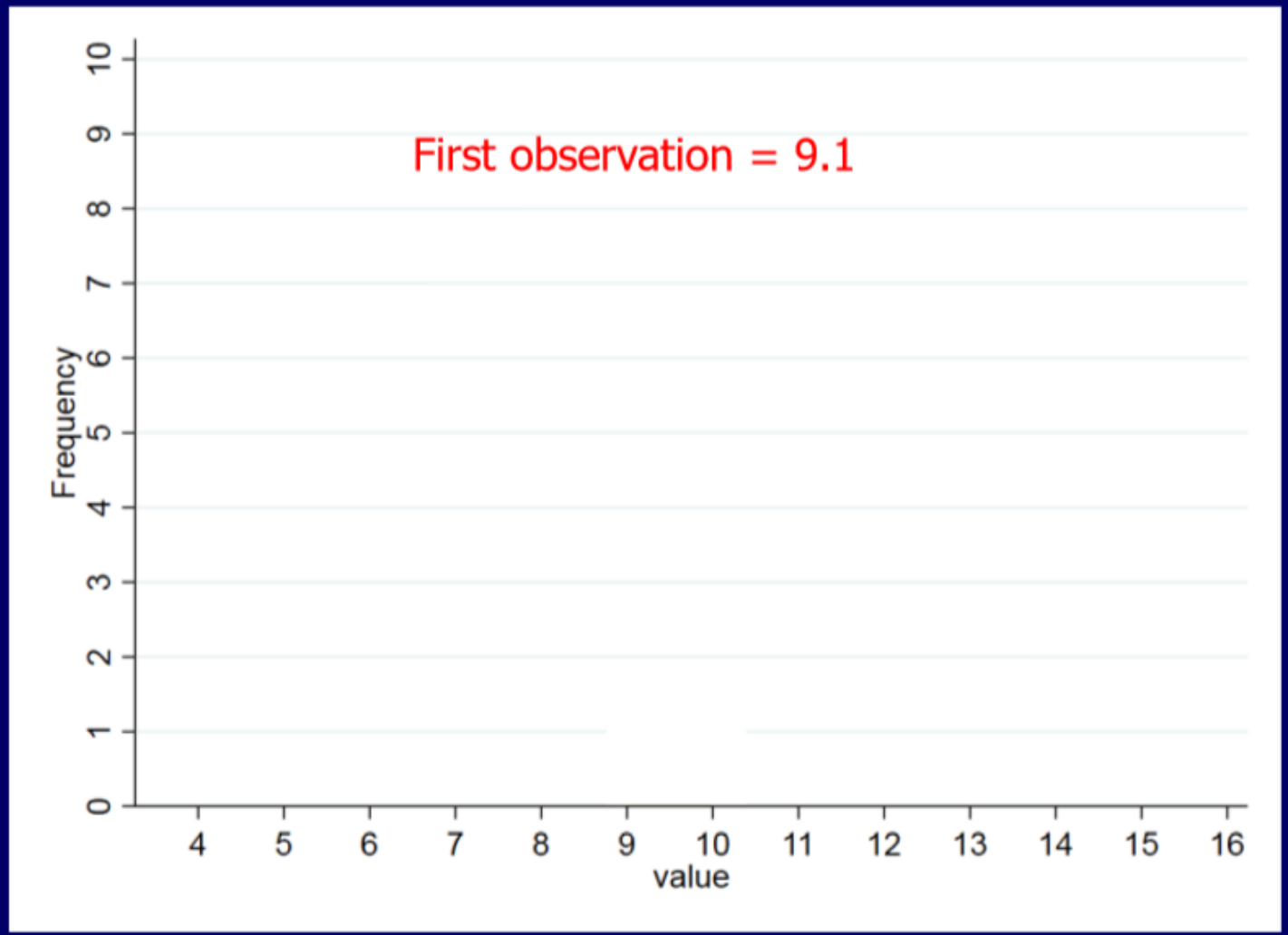
- Others?

Categorical Variables

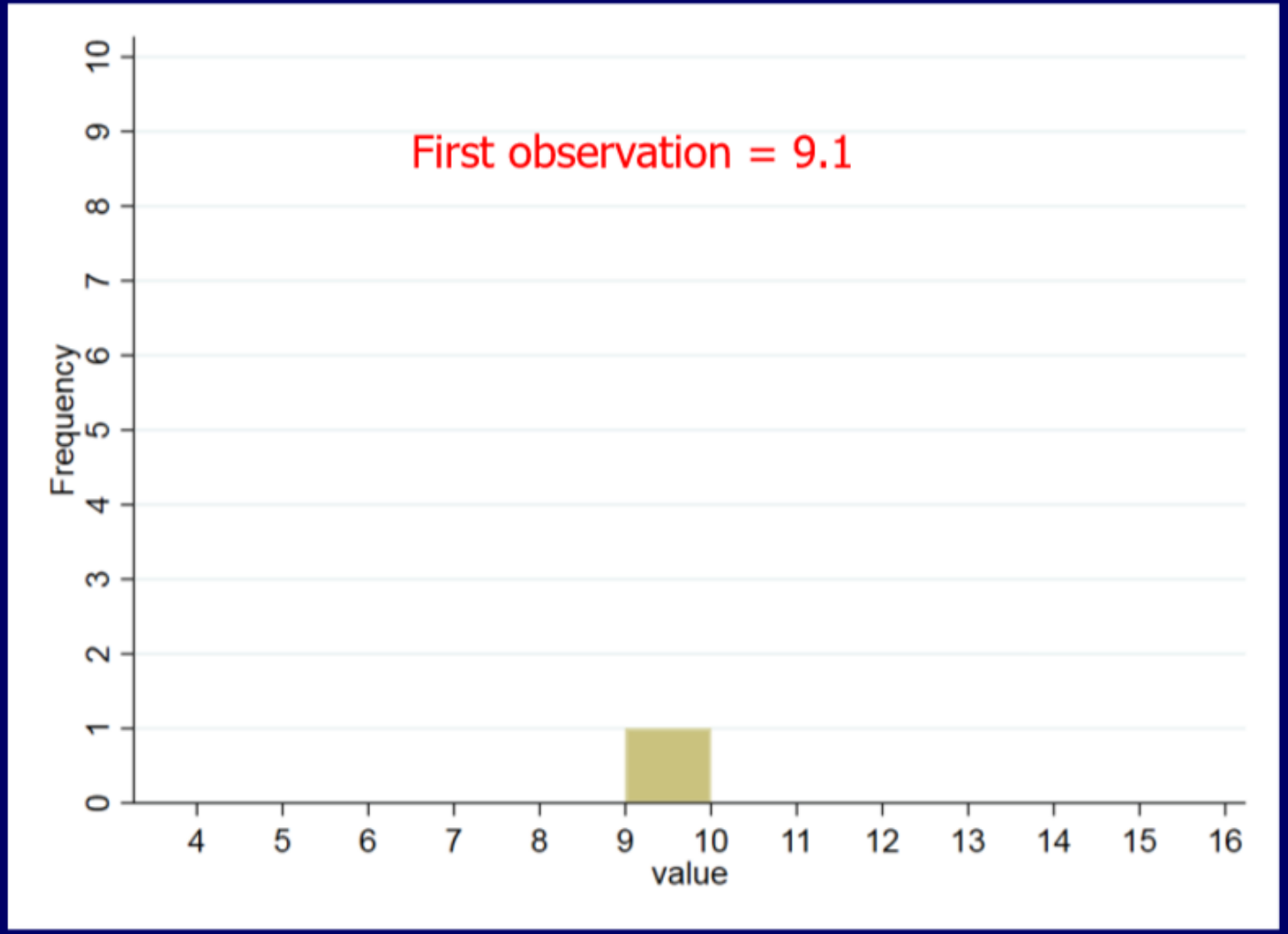
- Variables with 2+ categories (classes)
- Individual can only belong to one category
- **Nominal**
  - No intrinsic ordering of categories
  - Examples: *gender, blood group*
- **Ordinal**
  - Ordering is important
  - *Cancer staging (I-IV)*
  - *Education level*

Describing Quantitative (Continuous) Variables

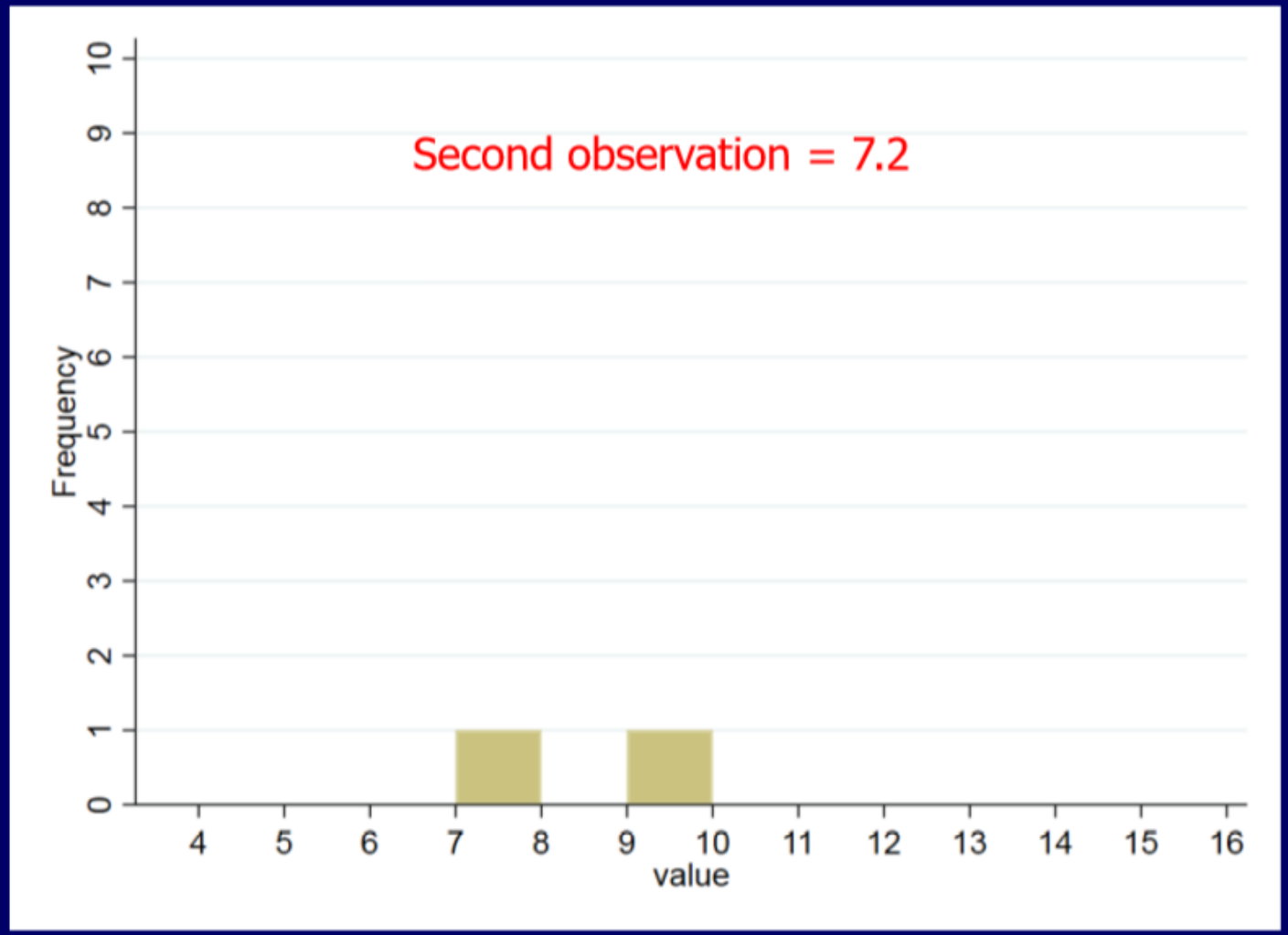
Histogram


First observation = 9.1
Frequency
0
1
2
3
4
5
6
7
8
9
10
4
5
6
7
8
9
10
11
12
13
14
15
16
value

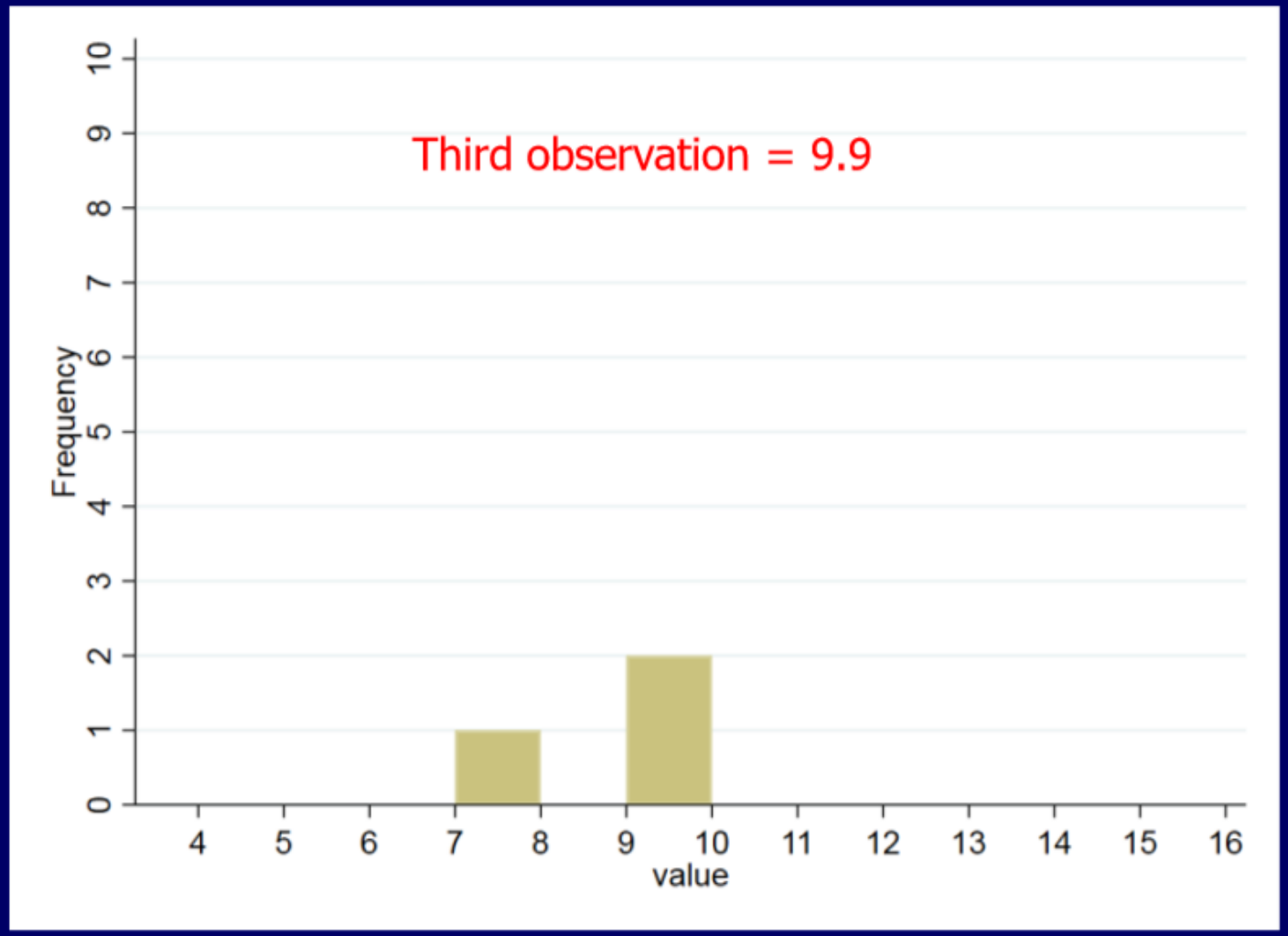
Histogram


First observation = 9.1
Frequency
0
1
2
3
4
5
6
7
8
9
10
4
5
6
7
8
9
10
11
12
13
14
15
16
value

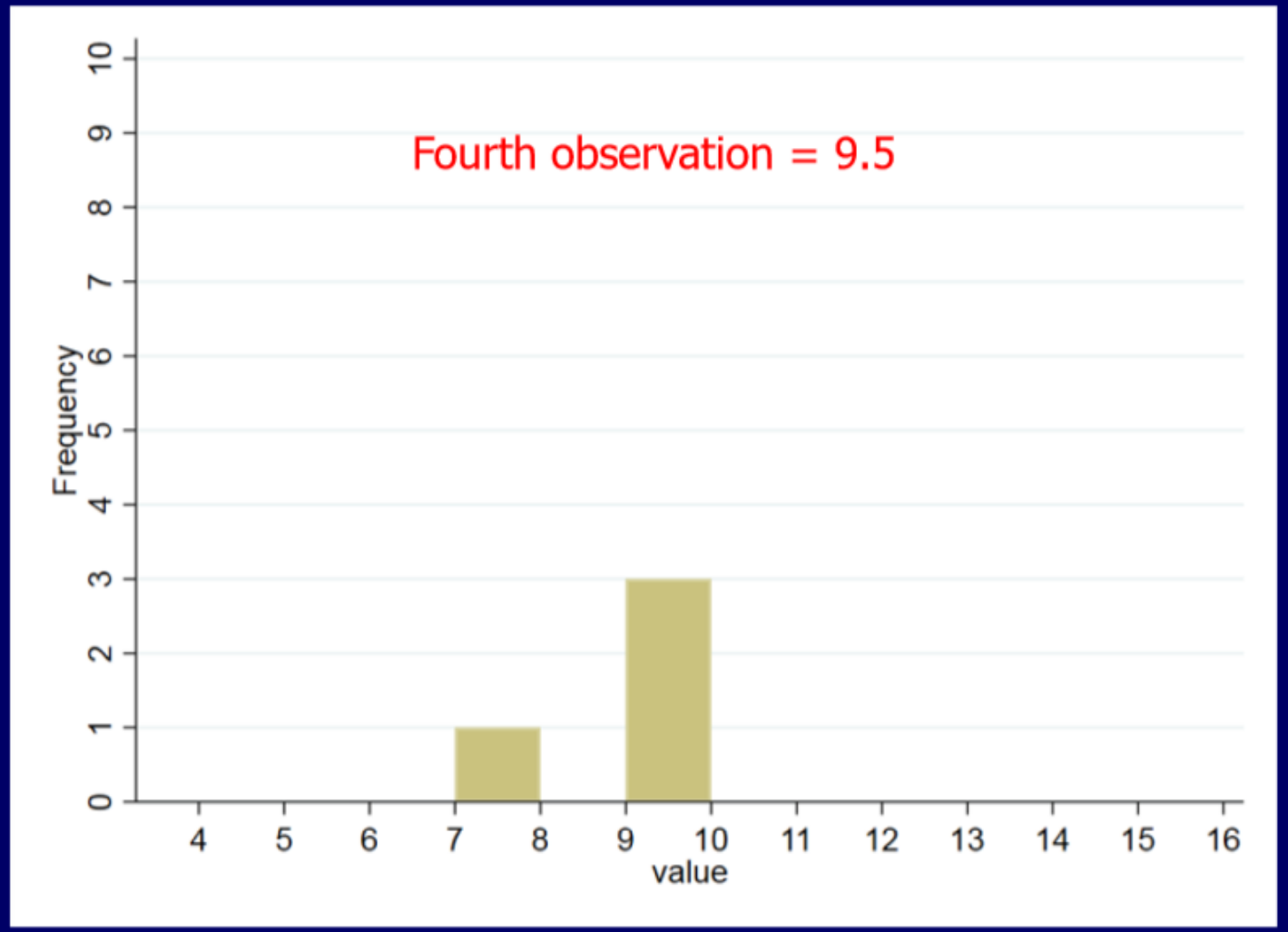
Histogram


Second observation = 7.2
Frequency
0
1
2
3
4
5
6
7
8
9
10
4
5
6
7
8
9
10
11
12
13
14
15
16
value

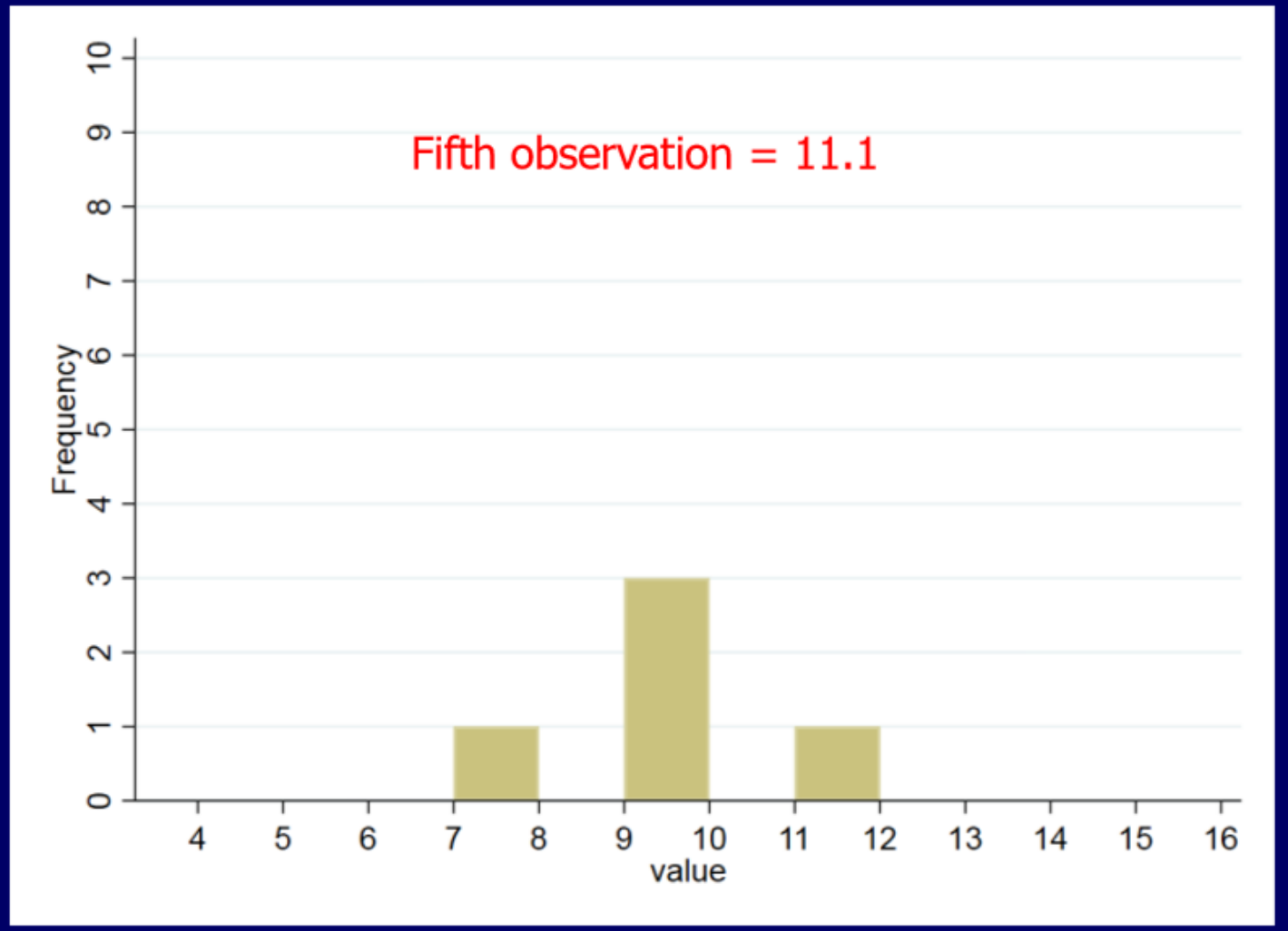
Histogram


Third observation = 9.9
Frequency
0
1
2
3
4
5
6
7
8
9
10
4
5
6
7
8
9
10
11
12
13
14
15
16
value

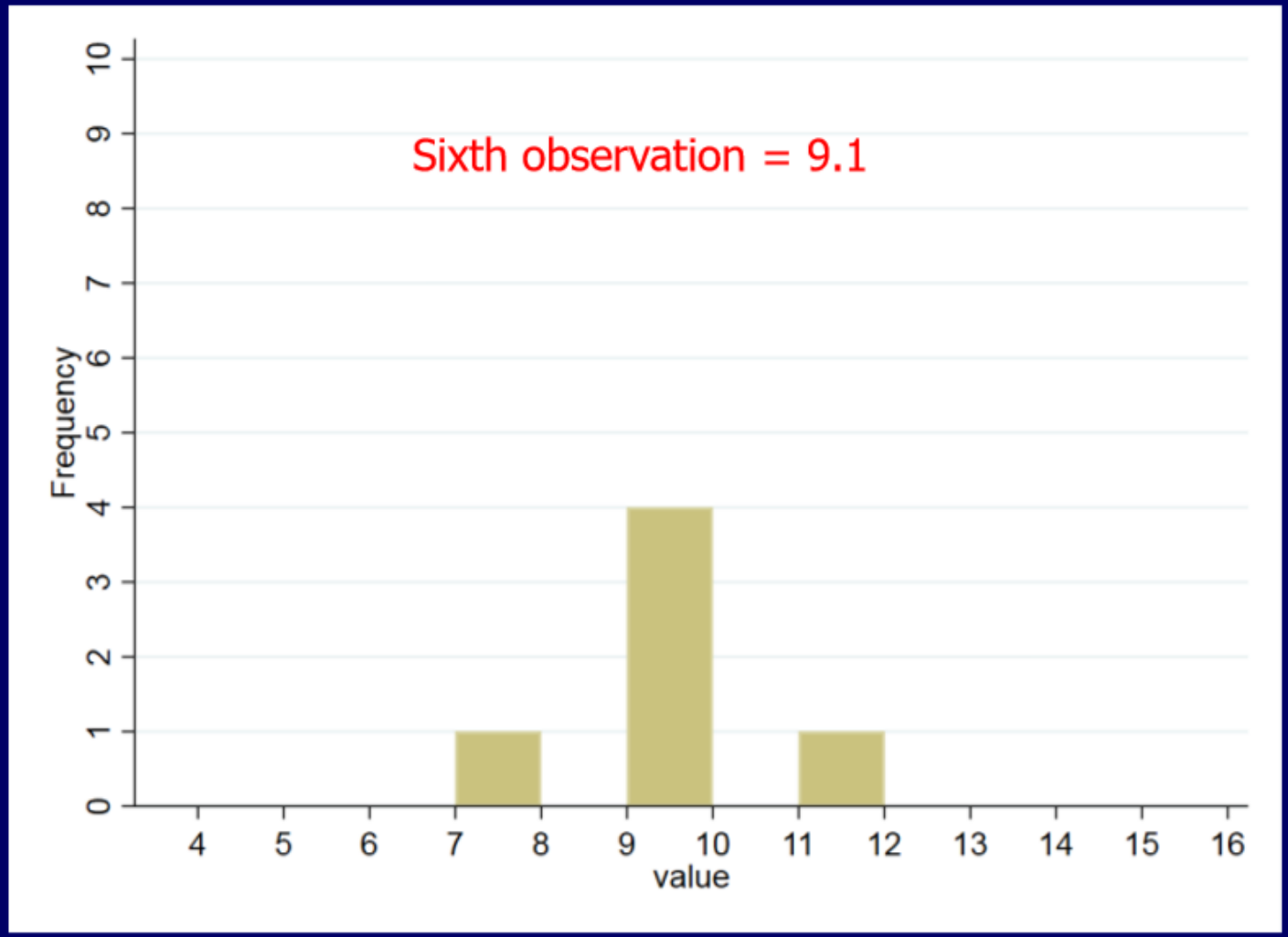
Histogram


Fourth observation = 9.5
Frequency
0
1
2
3
4
5
6
7
8
9
10
4
5
6
7
8
9
10
11
12
13
14
15
16
value

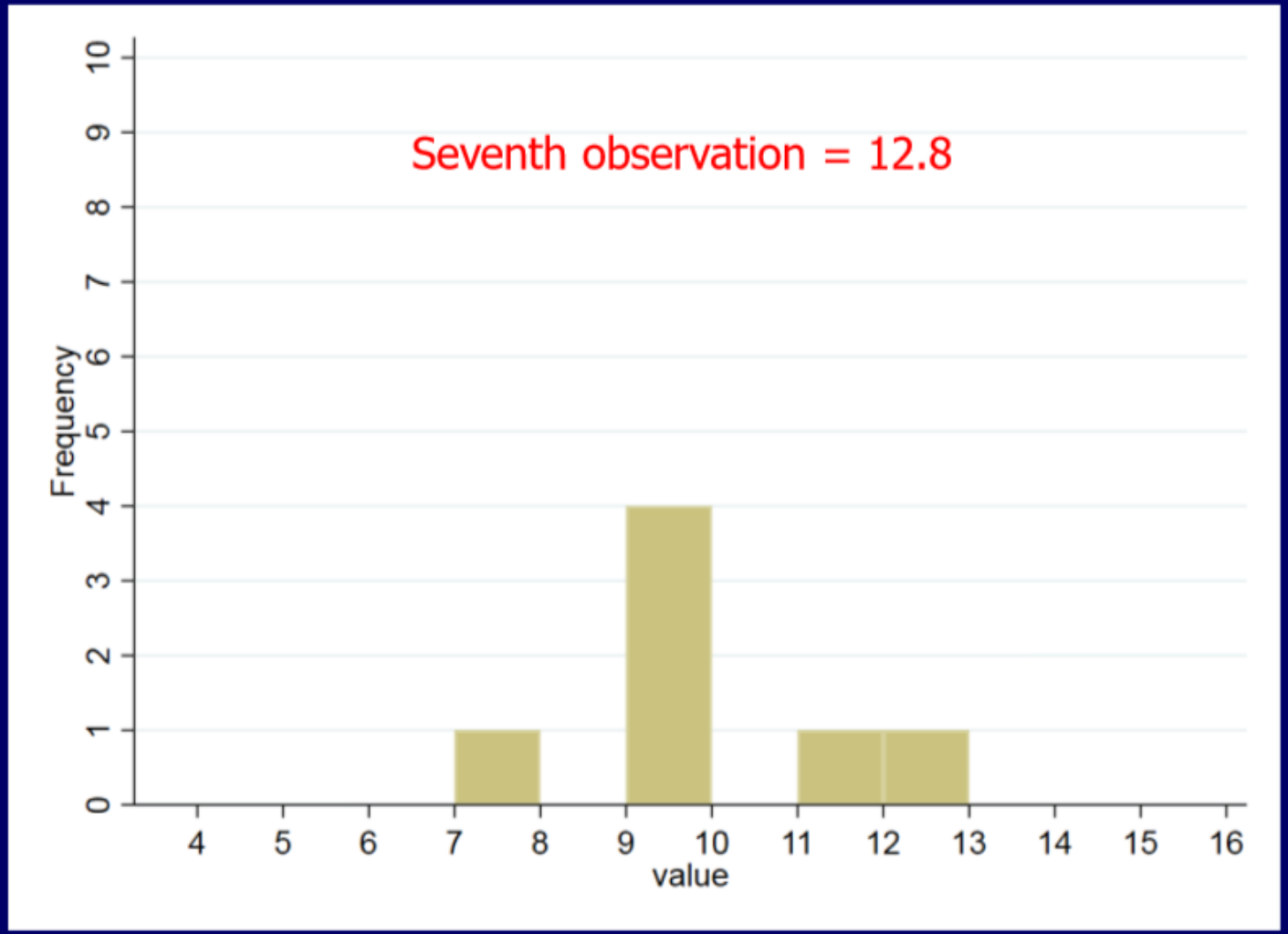
Histogram


Fifth observation = 11.1
Frequency
0
1
2
3
4
5
6
7
8
9
10
4
5
6
7
8
9
10
11
12
13
14
15
16
value

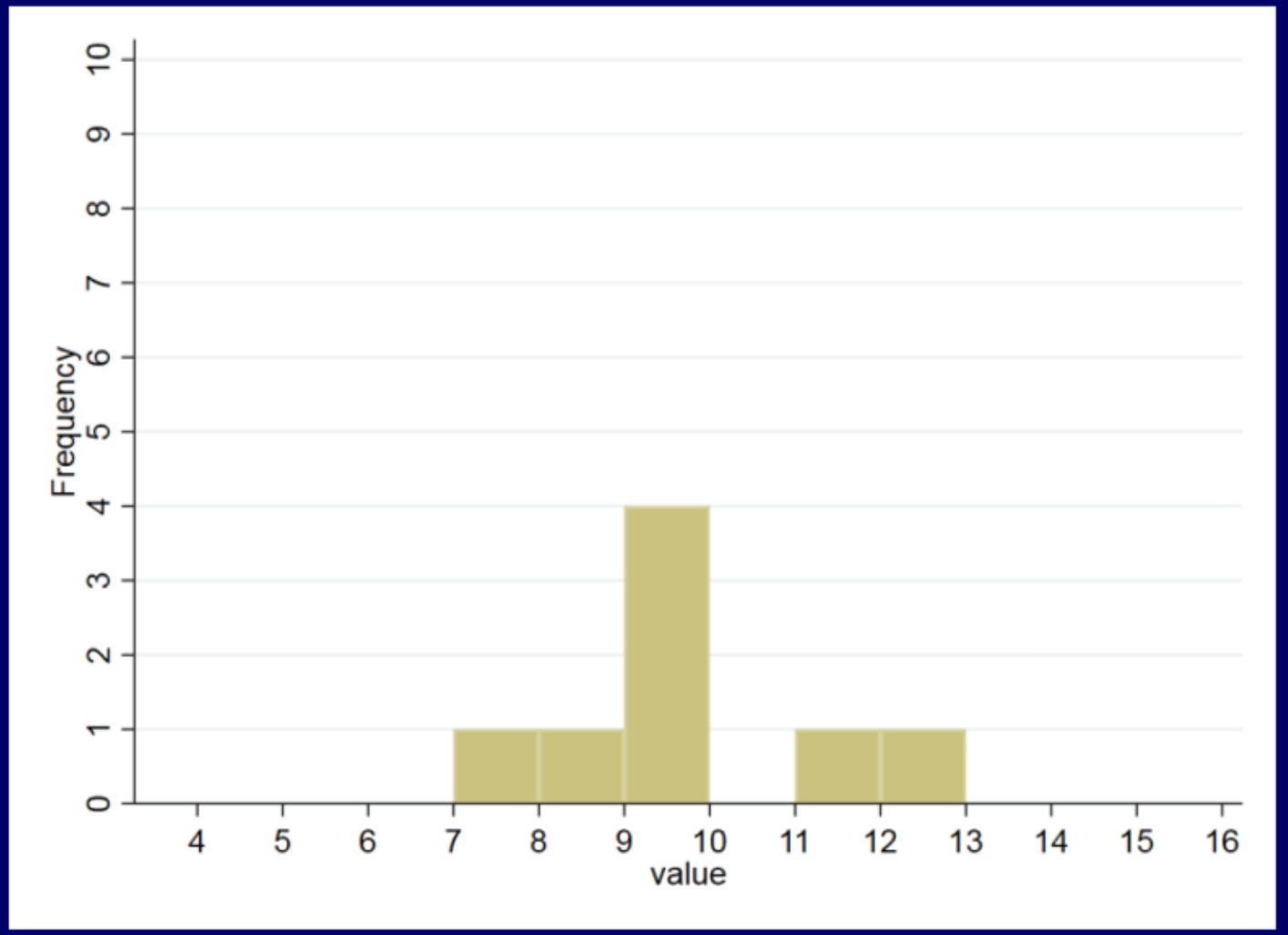
Histogram


Sixth observation = 9.1
Frequency
0
1
2
3
4
5
6
7
8
9
10
4
5
6
7
8
9
10
11
12
13
14
15
16
value

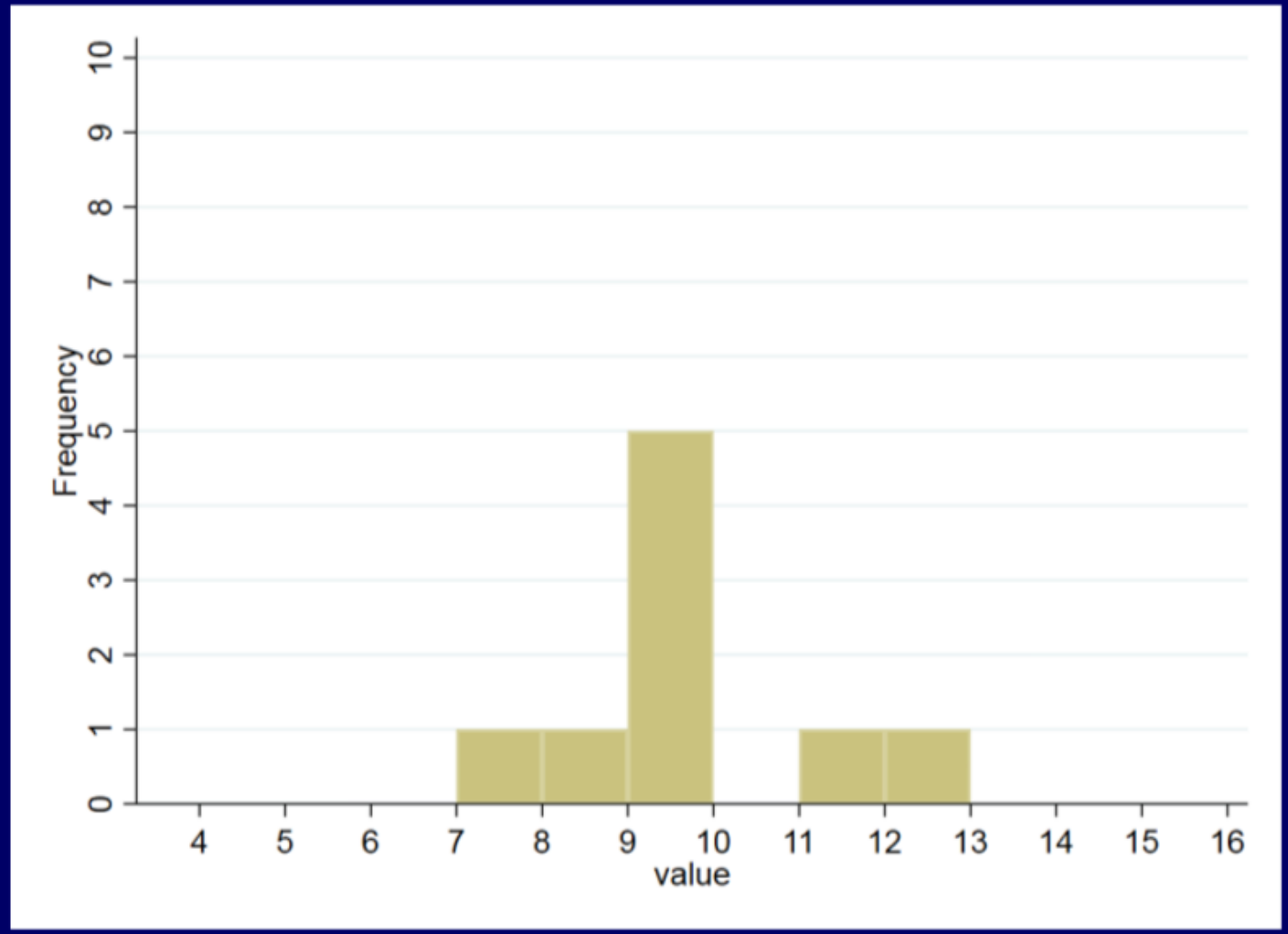
Histogram


Seventh observation = 12.8
Frequency
0
1
2
3
4
5
6
7
8
9
10
4
5
6
7
8
9
10
11
12
13
14
15
16
value

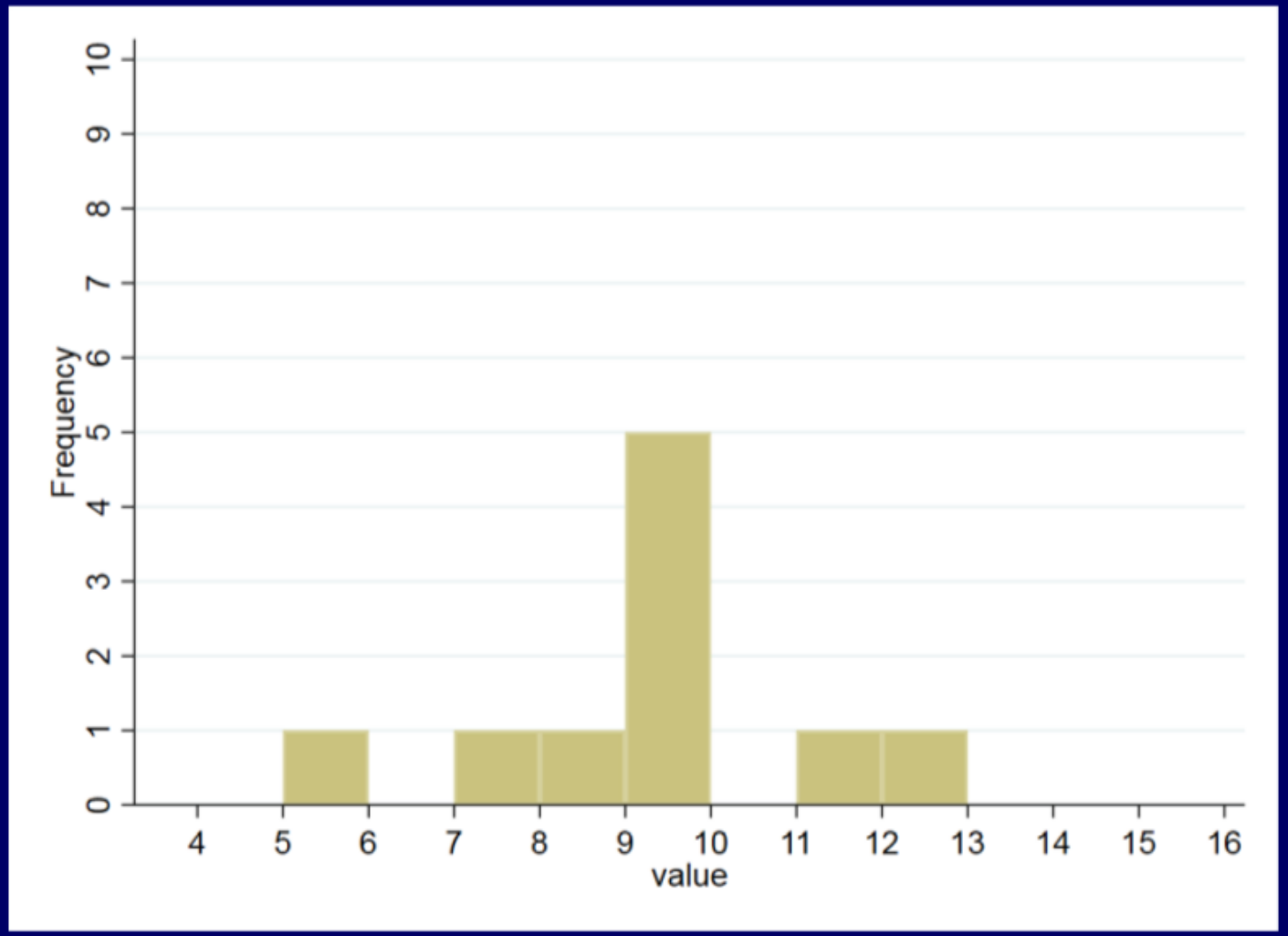
Histogram

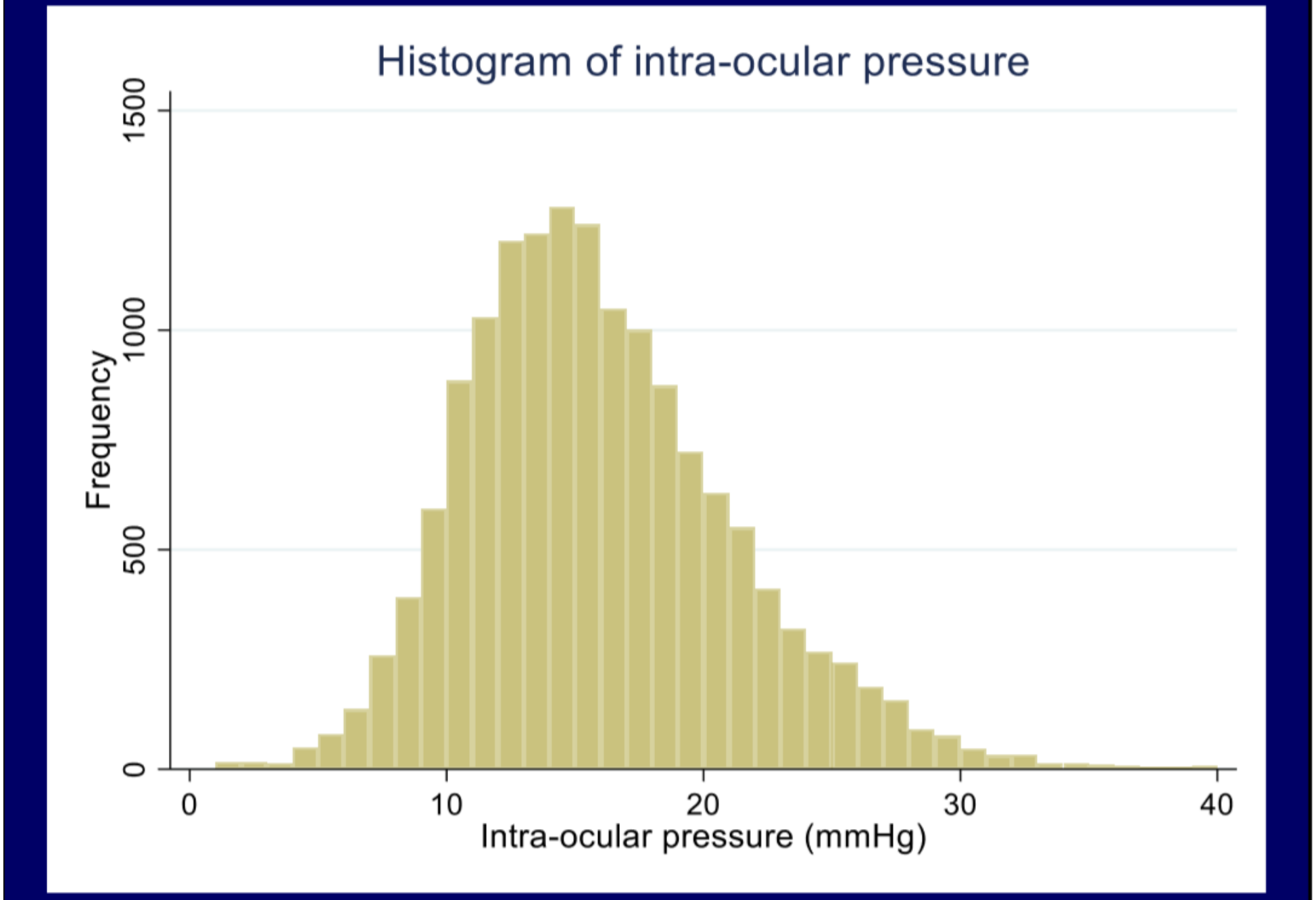


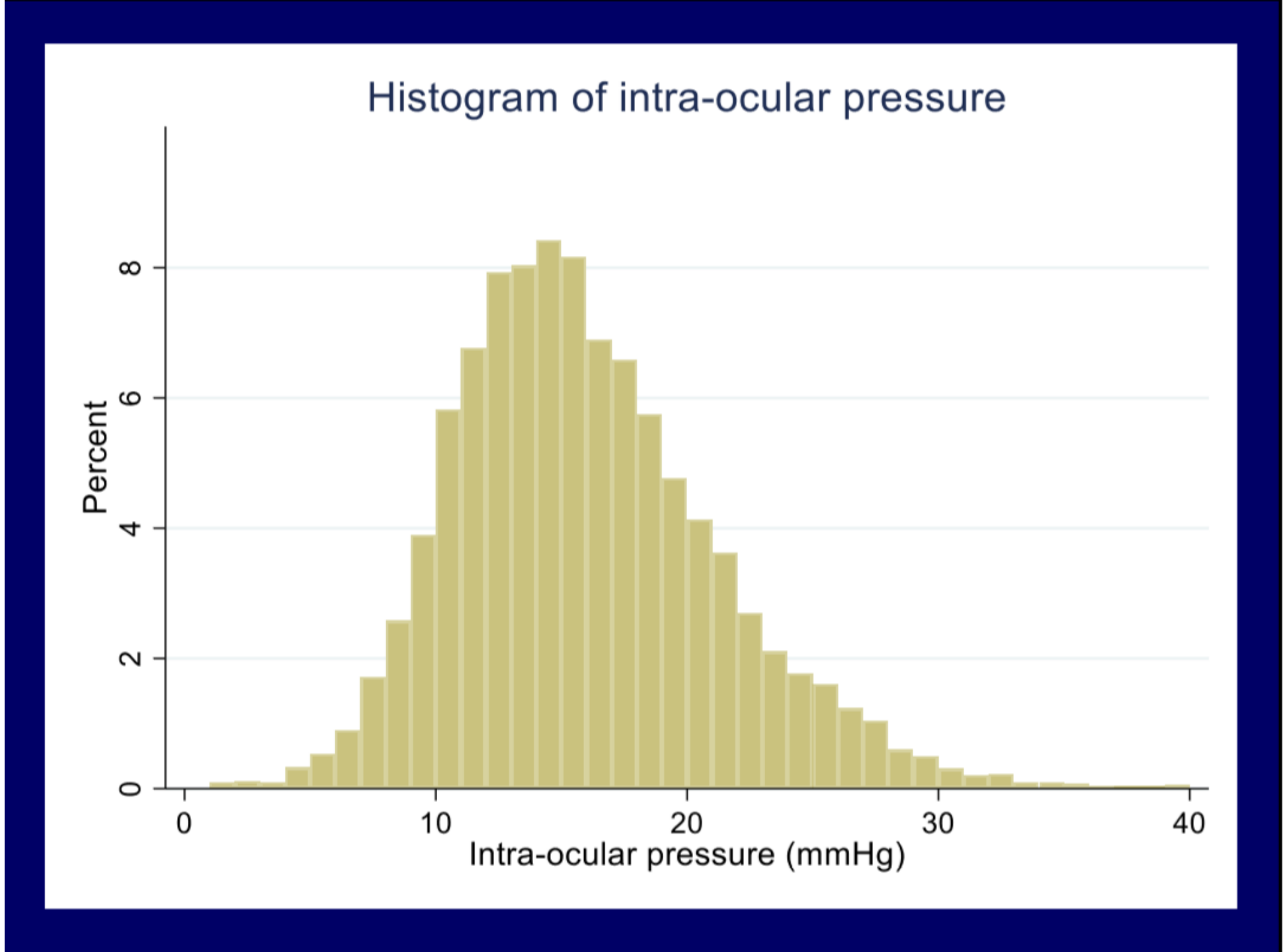
Histogram

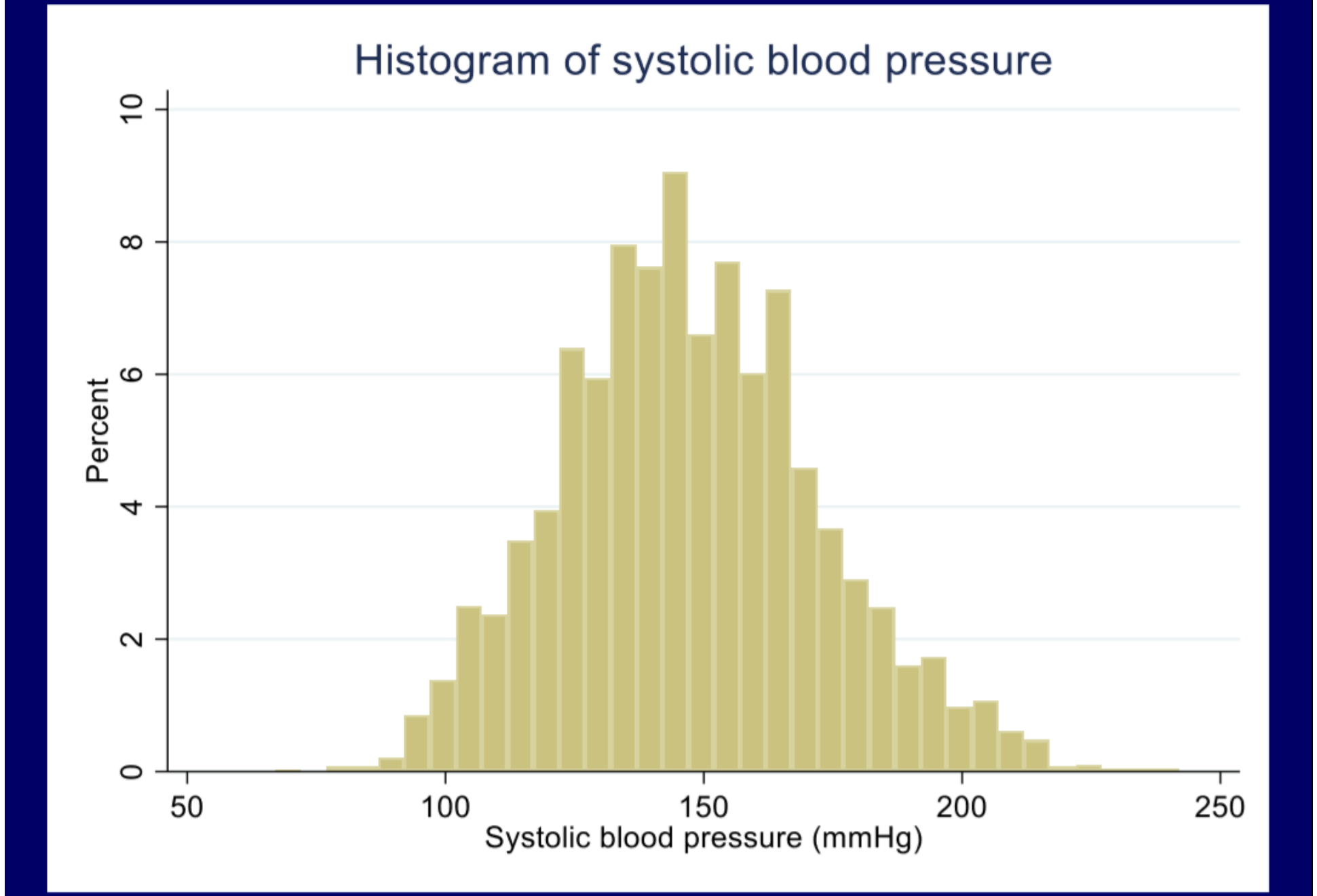


Histogram

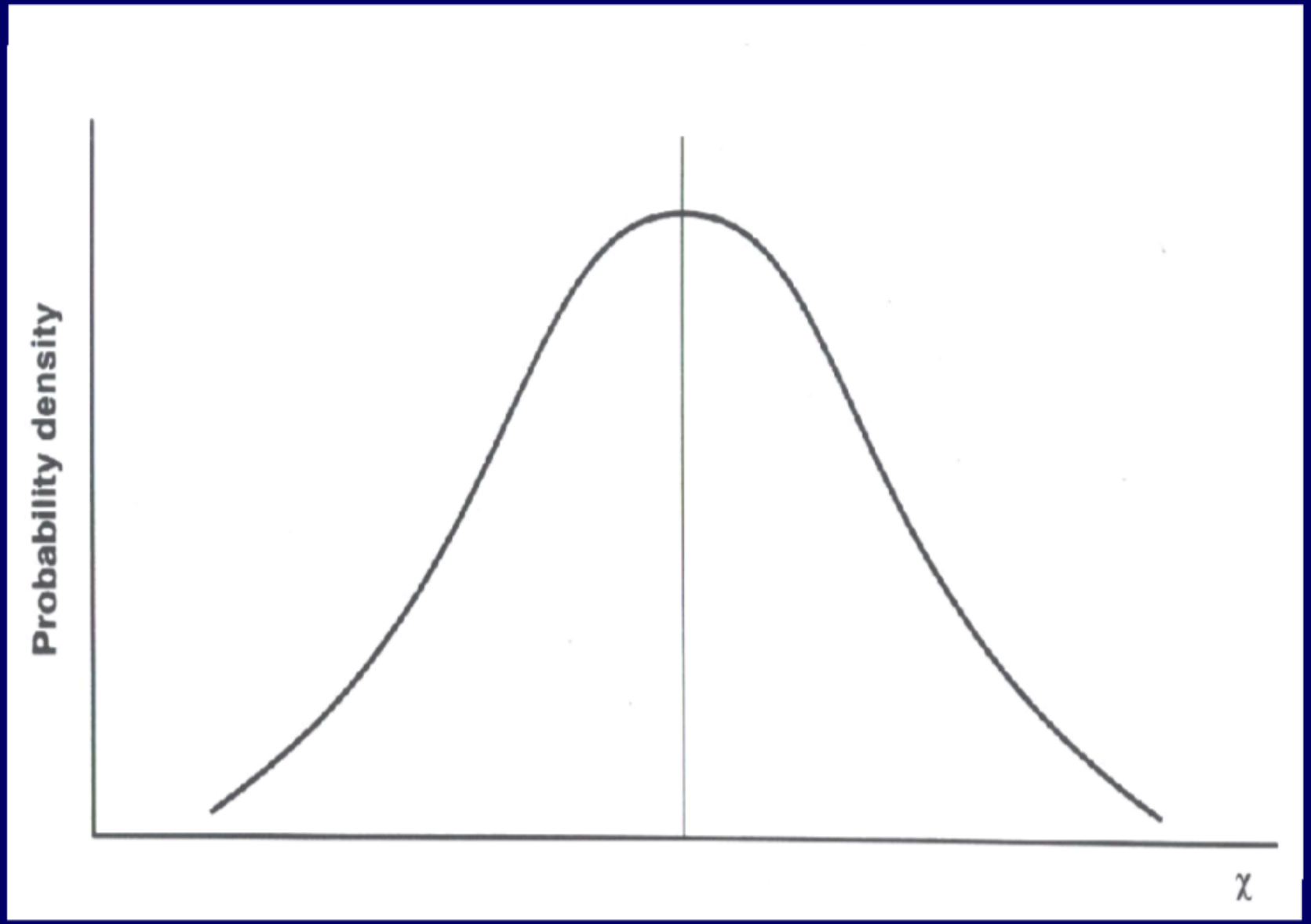


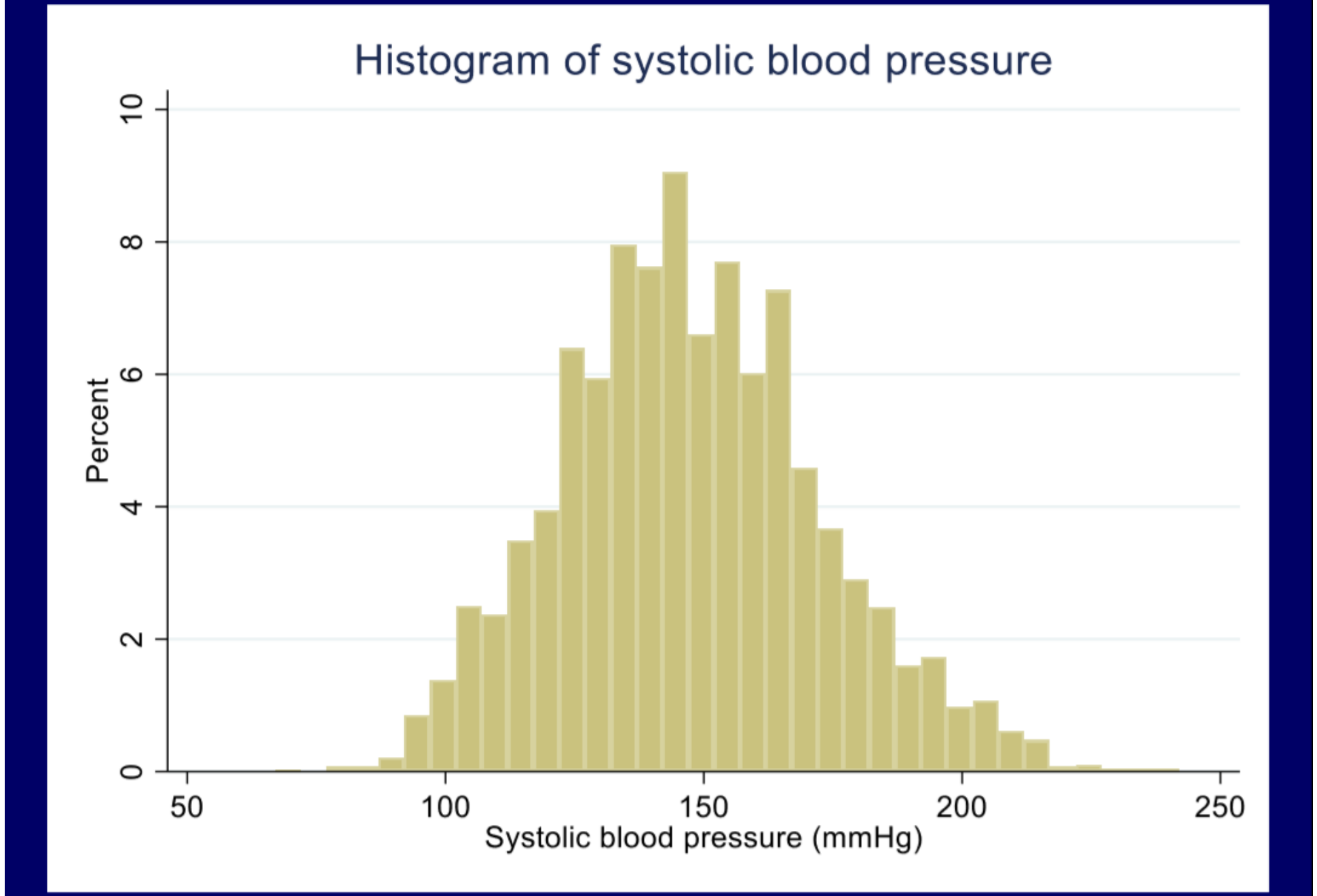

Histogram of intra-ocular pressure
Frequency
0
500
1000
1500
0
10
20
30
40
Intra-ocular pressure (mmHg)

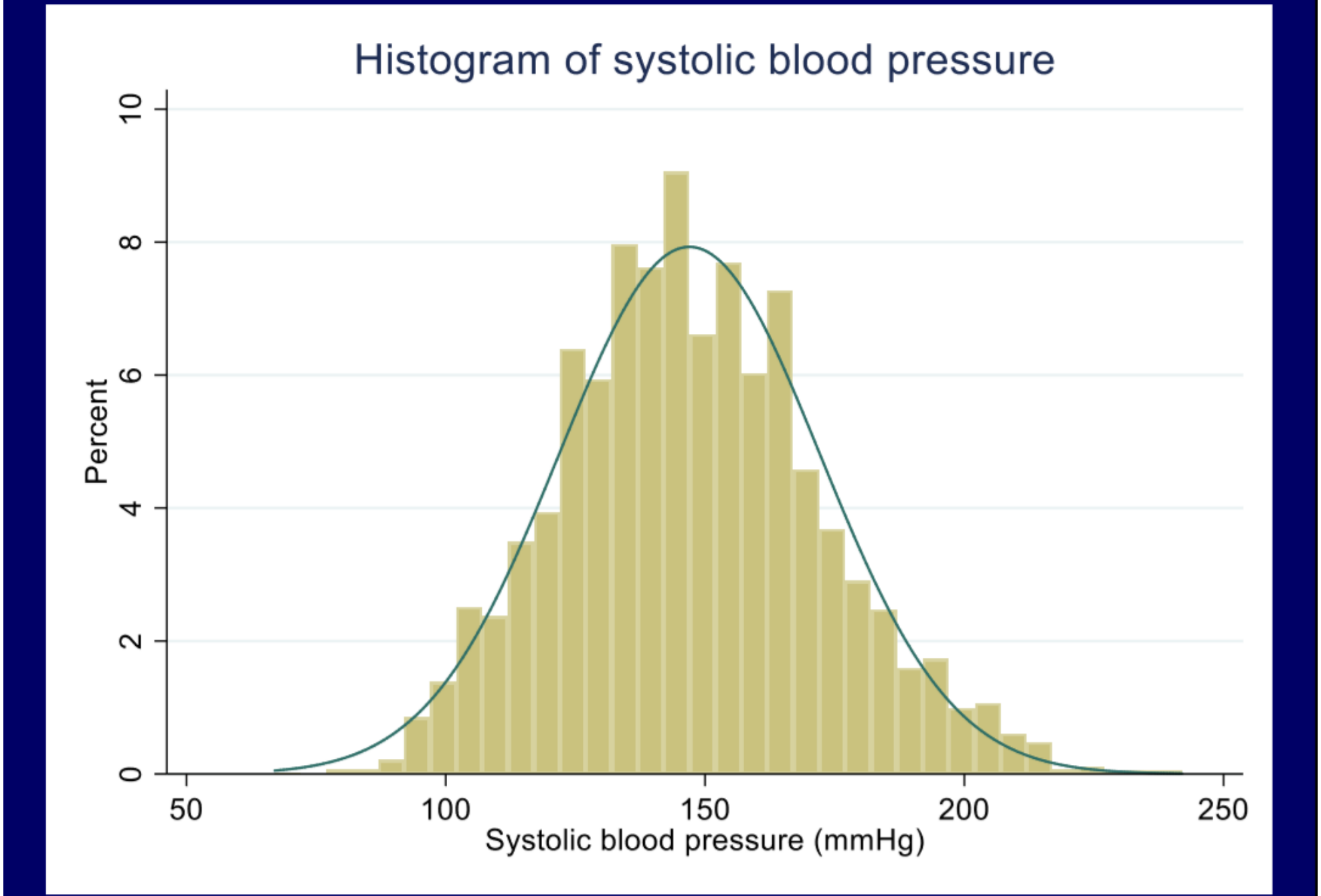

Histogram of intra-ocular pressure
Percent
0
2
4
6
8
0
10
20
30
40
Intra-ocular pressure (mmHg)


Histogram of systolic blood pressure
Percent
0
2
4
6
8
10
50
100
150
200
250
Systolic blood pressure (mmHg)

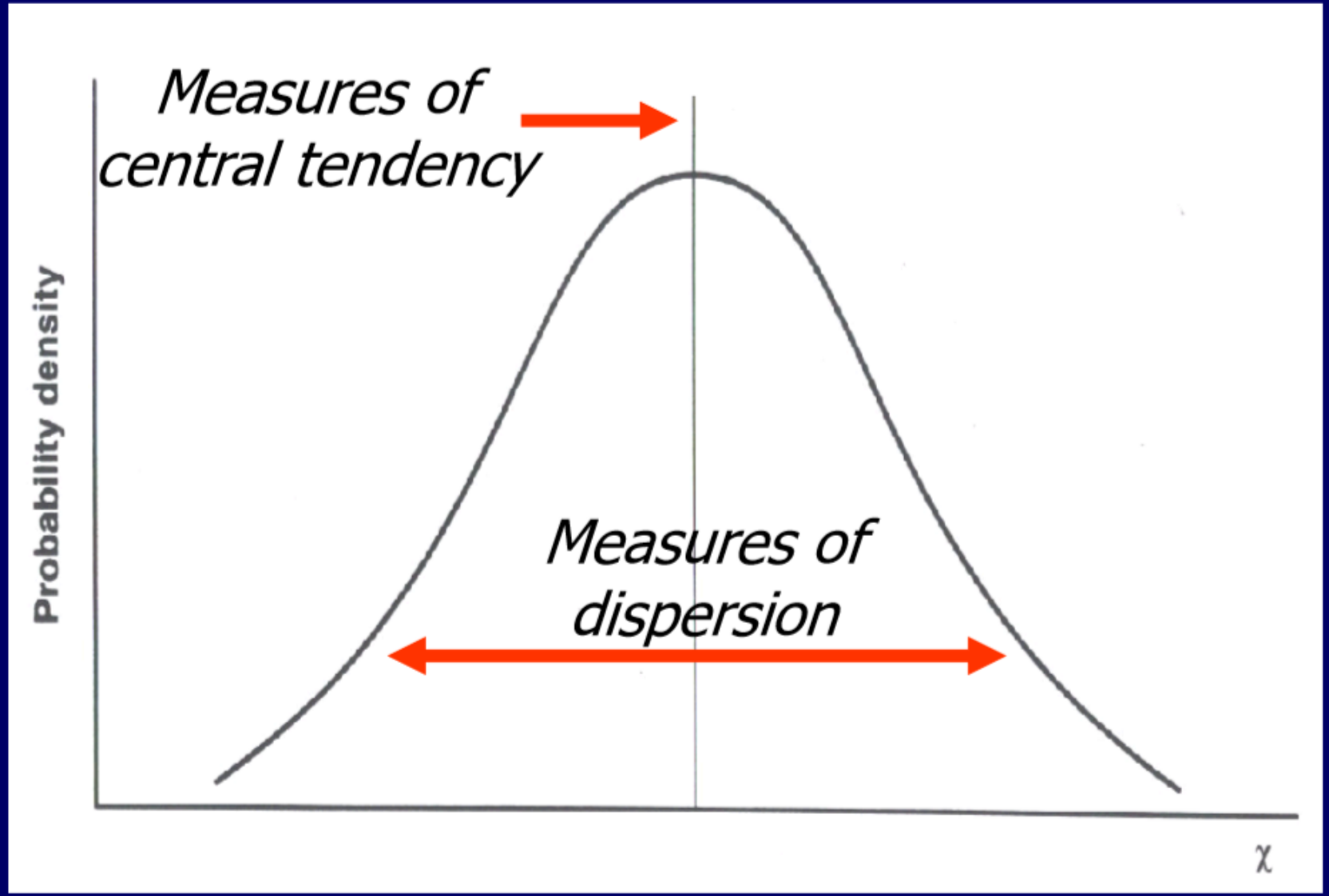
The Normal Distribution



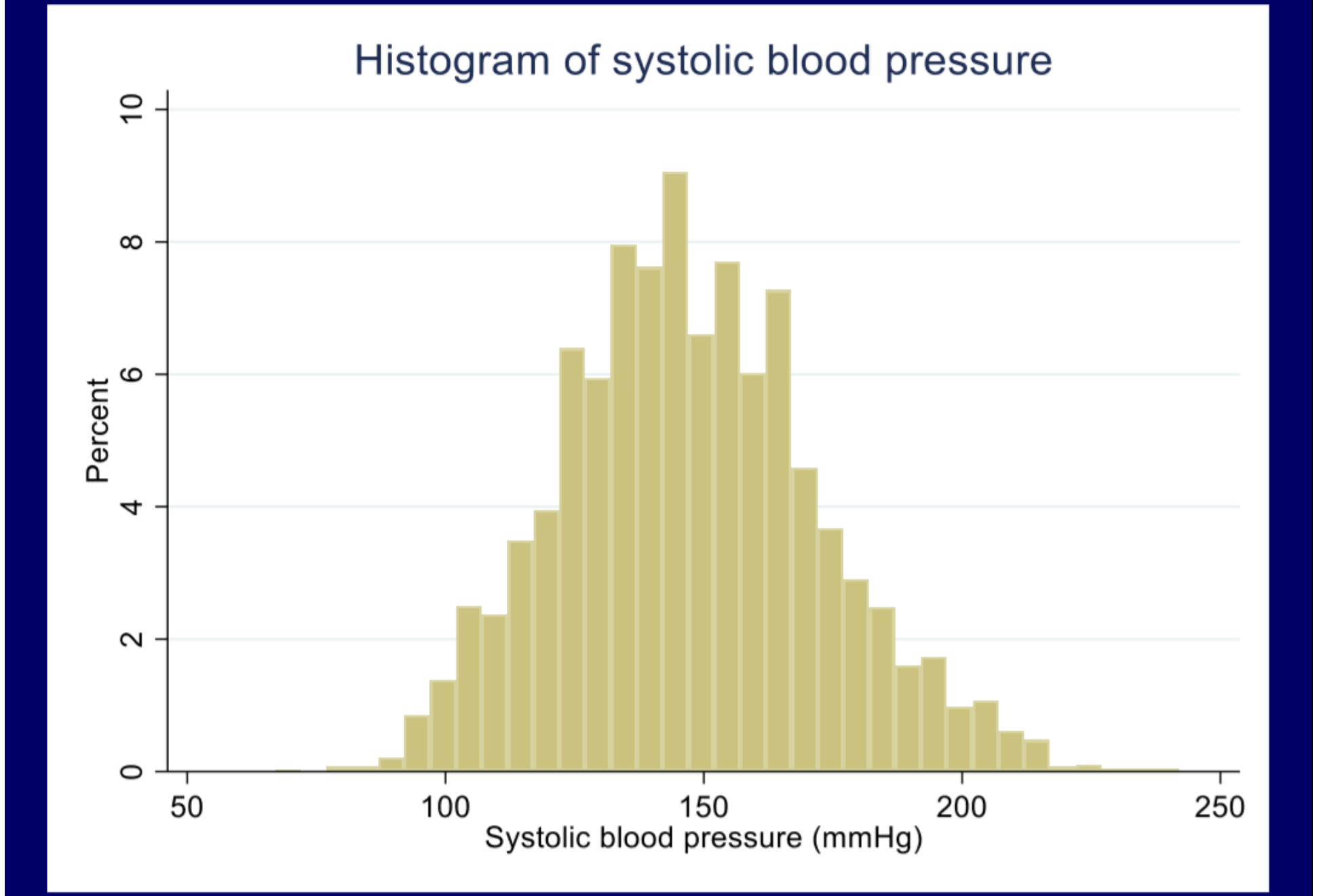

Histogram of systolic blood pressure
Percent
0
2
4
6
8
10
50
100
150
200
250
Systolic blood pressure (mmHg)


Histogram of systolic blood pressure
Percent
0
2
4
6
8
10
50
100
150
200
250
Systolic blood pressure (mmHg)

The Normal Distribution



Measures of Central Tendency (or Location)


Histogram of systolic blood pressure
Percent
0
2
4
6
8
10
50
100
150
200
250
Systolic blood pressure (mmHg)

The (Arithmetic) Mean

- What most people refer to as an "average"
- This is the sum of the observations divided by N, the number of observations.

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

The (Arithmetic) Mean

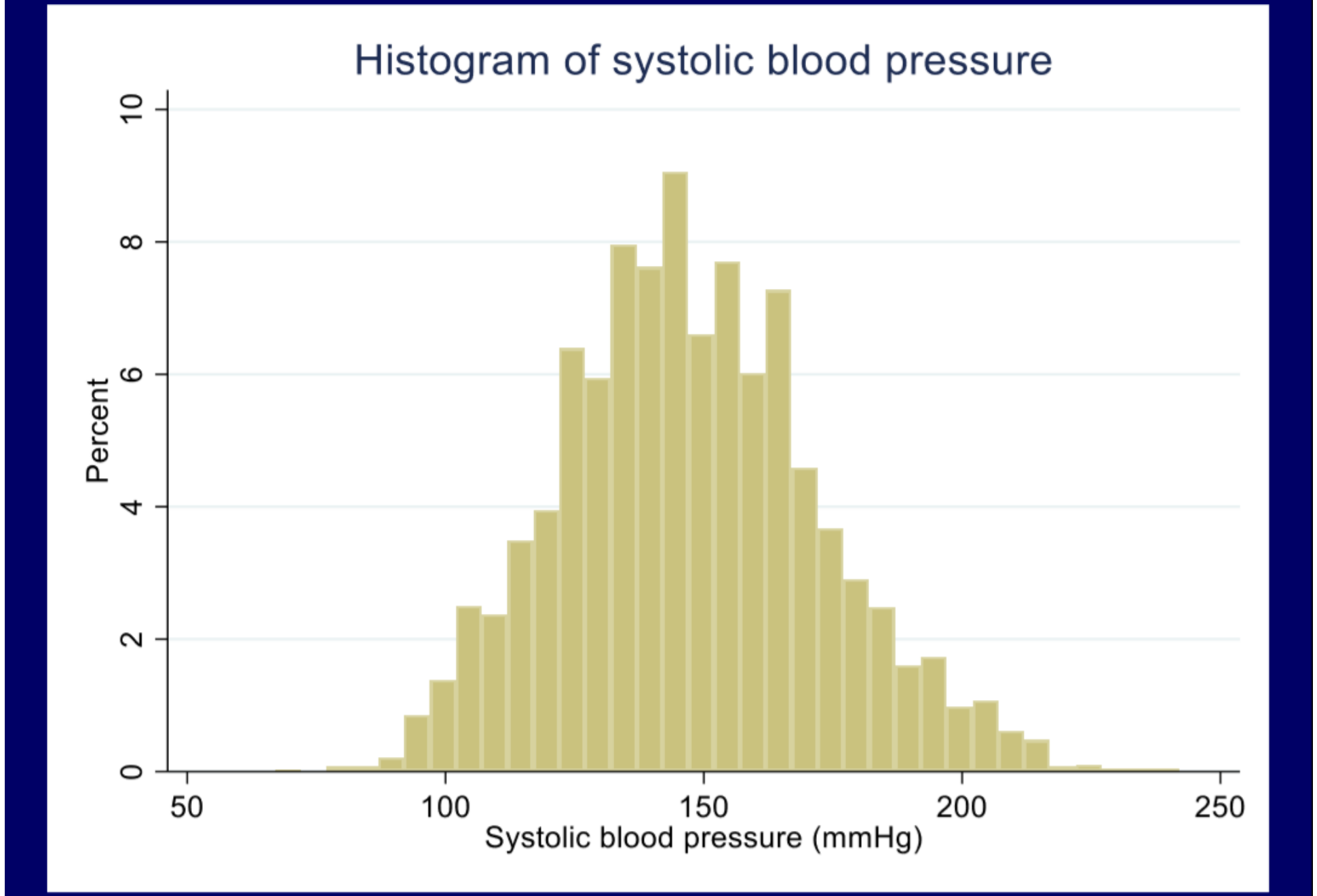
- Example: calculate the arithmetic mean of the 8 plasma volumes:

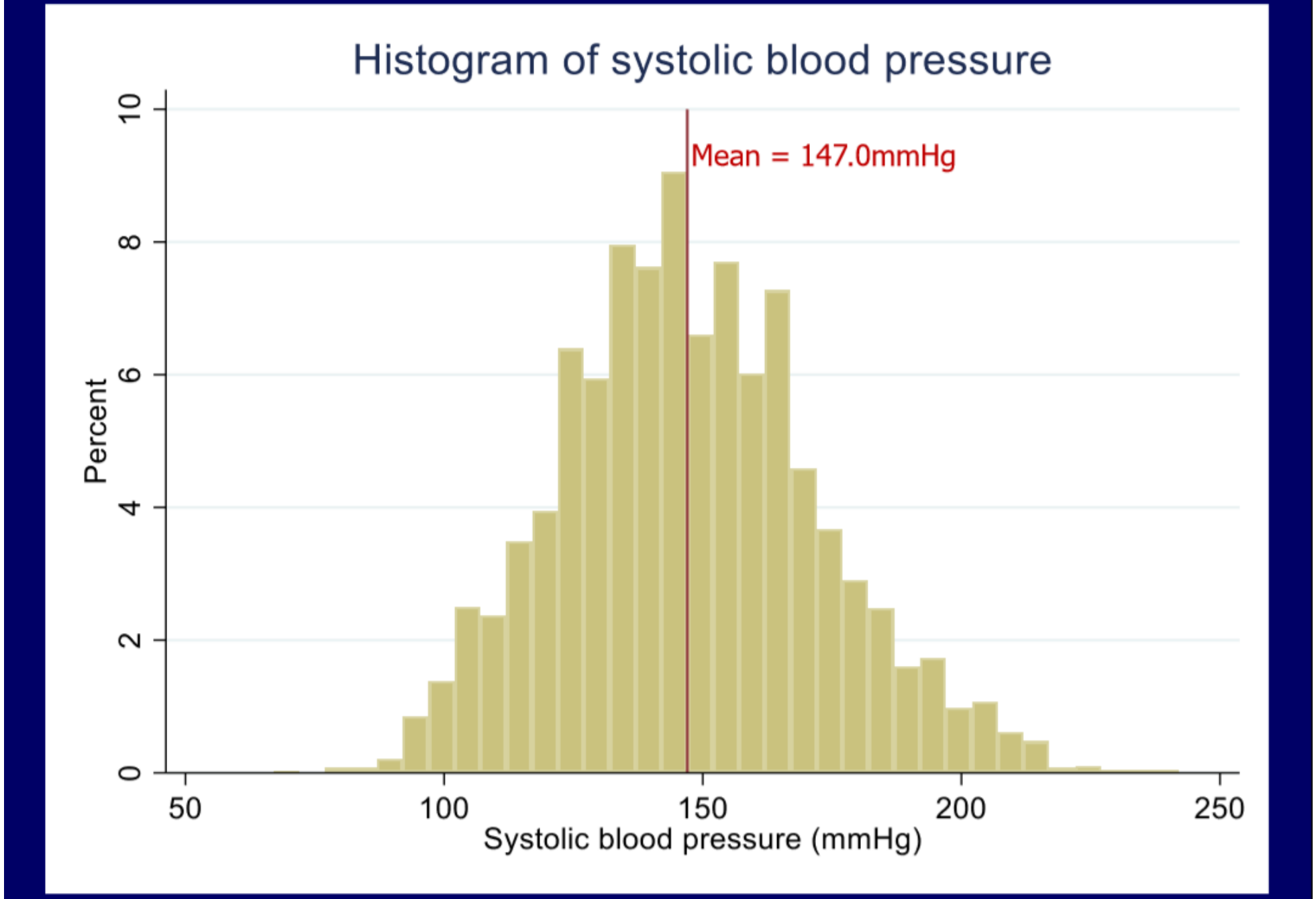
2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12

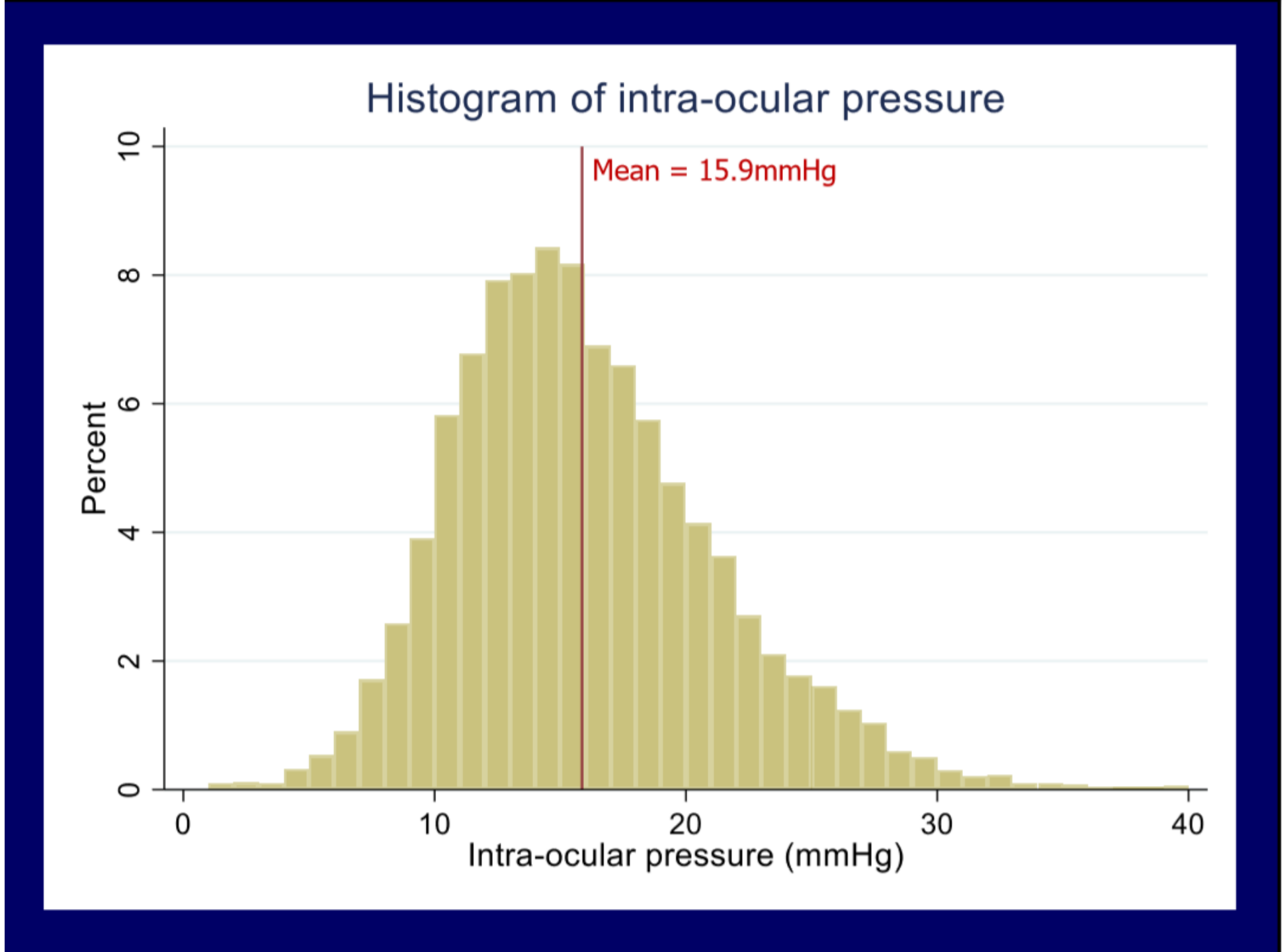
(2.75 +2.86 +3.37 +2.76 +2.62 +3.49 +3.05 +3.12) /8

= 24.02/8

= 3.0025 litres.


Histogram of systolic blood pressure
Percent
0
2
4
6
8
10
50
100
150
200
250
Systolic blood pressure (mmHg)


Histogram of systolic blood pressure
Mean = 147.0mmHg
Percent
0
2
4
6
8
10
50
100
150
200
250
Systolic blood pressure (mmHg)


Histogram of intra-ocular pressure
Mean = 15.9mmHg
Percent
0
2
4
6
8
10
0
10
20
30
40
Intra-ocular pressure (mmHg)

Outliers

- The mean is sensitive to outliers, particularly when sample size is small
- 2,2,3,4,4,5,6,6,30
- Mean of these numbers is 6.9 – which is larger than 8 out of the 9 observations
- Arithmetic mean is probably not a good summary measure of the data in this instance

The Median

- The value that divides a distribution in half.
- "50th" percentile
- Median = (n+1)/2 th value
- Or in other words, the median is the middle value of the *ordered* observations
- If even number of observations, take the mean of the two middle observations

Example: The Median

- Calculate the median of the 8 plasma volumes:

  2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12

- Ordered values:

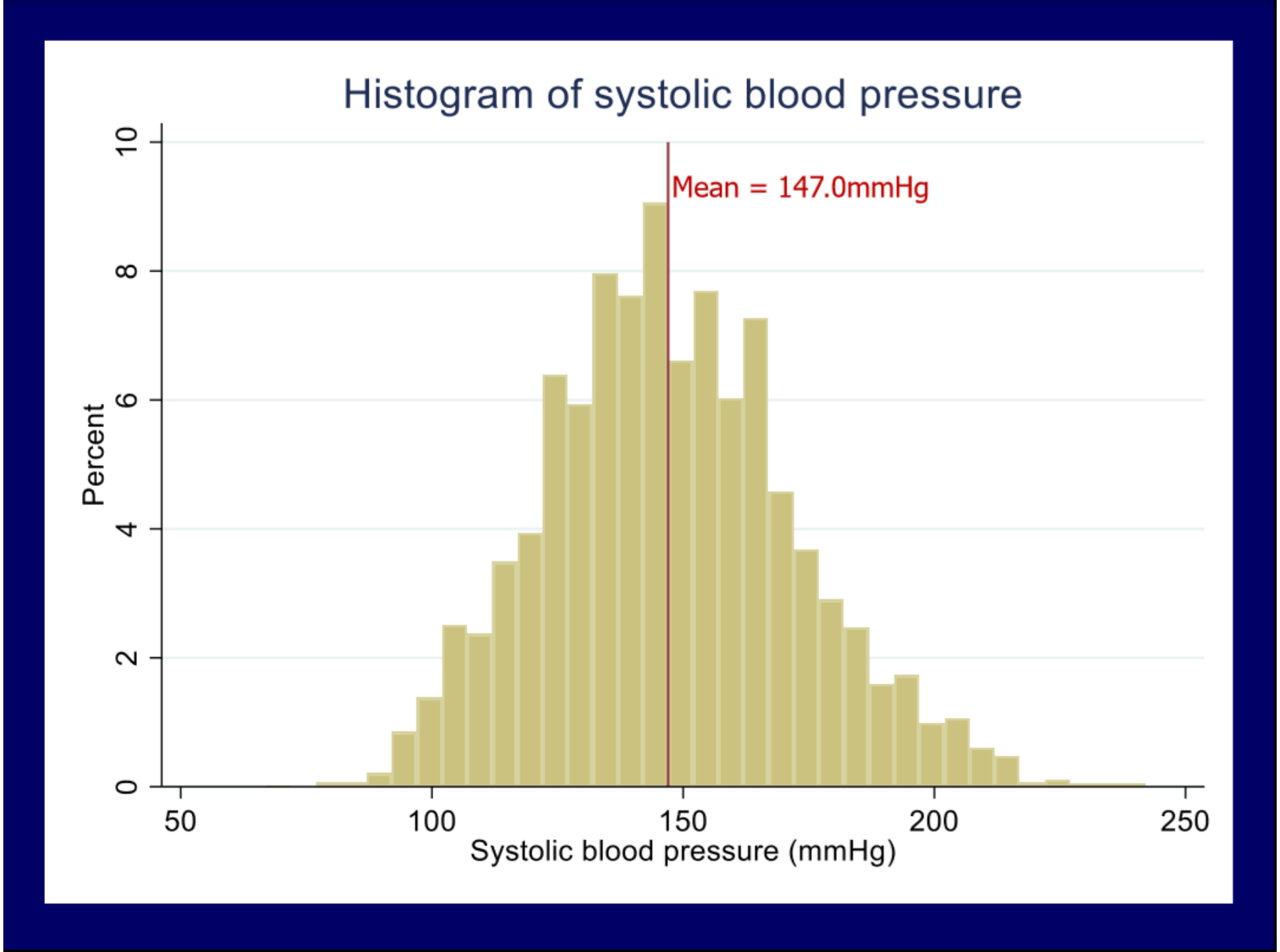
  2.62 2.75 2.76 2.86 3.05 3.12 3.37 3.49

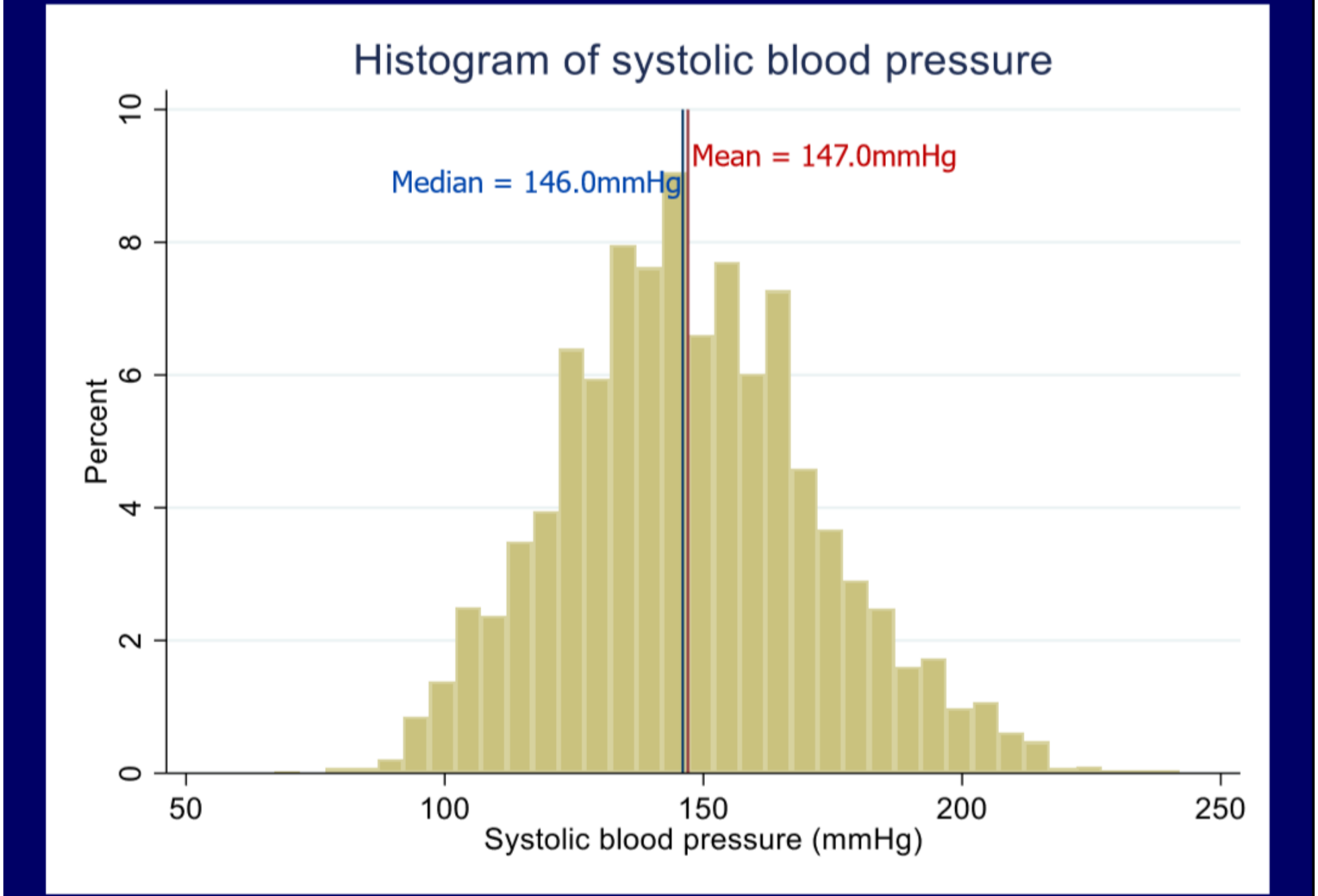
  Median = $(n+1)/2^{th}$ value = $(8+1)/2$

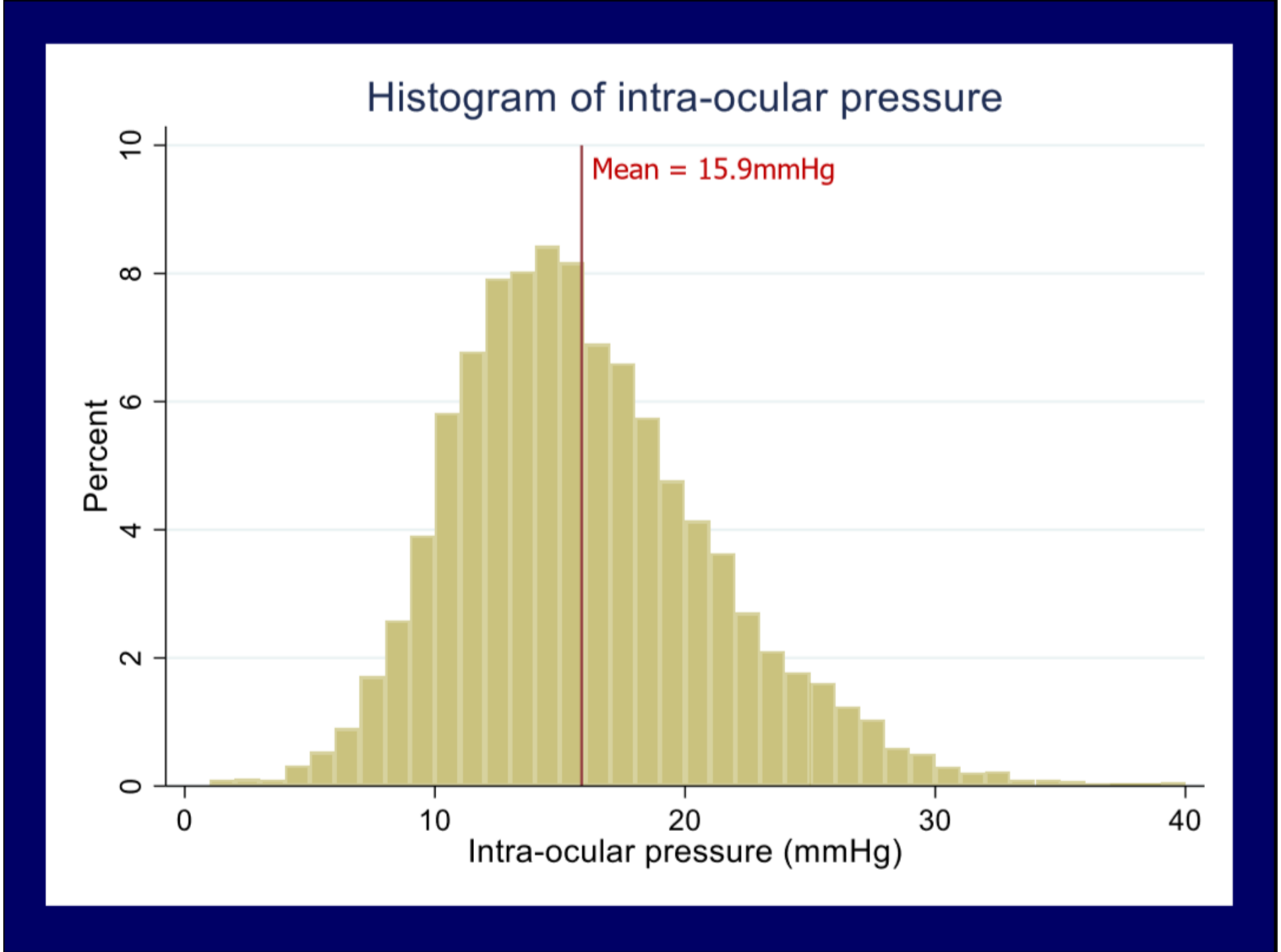
  = 4.5^{th} value

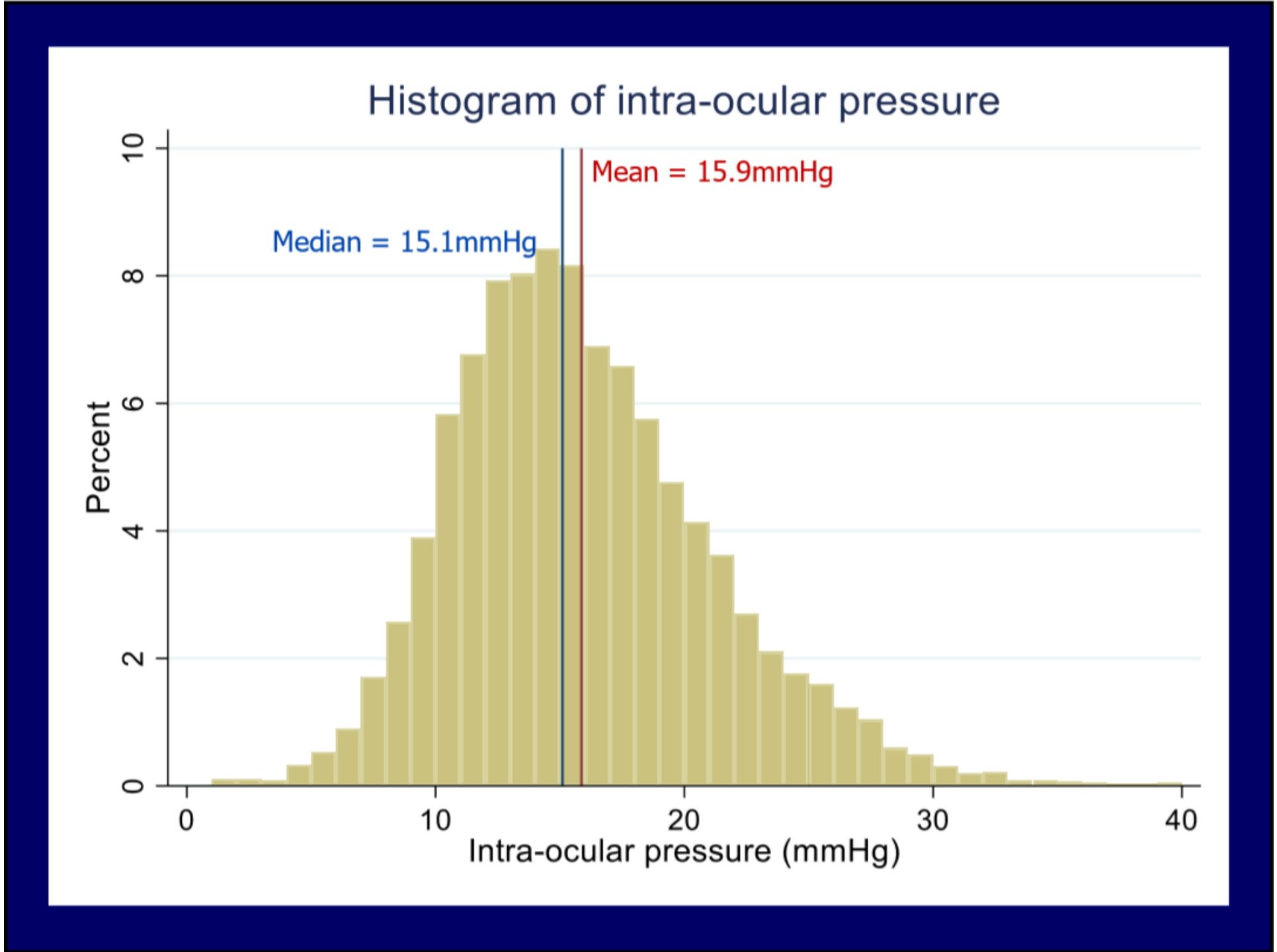
  = $(2.86 + 3.05)/ 2$

  = 2.95


Histogram of systolic blood pressure
Mean = 147.0mmHg
Percent
0
2
4
6
8
10
50
100
150
200
250
Systolic blood pressure (mmHg)


Histogram of systolic blood pressure
Median = 146.0mmHg
Mean = 147.0mmHg
Percent
0
2
4
6
8
10
50
100
150
200
250
Systolic blood pressure (mmHg)


Histogram of intra-ocular pressure
Mean = 15.9mmHg
Percent
0
2
4
6
8
10
0
10
20
30
40
Intra-ocular pressure (mmHg)


Histogram of intra-ocular pressure
Mean = 15.9mmHg
Median = 15.1mmHg
Percent
0
2
4
6
8
10
0
10
20
30
40
Intra-ocular pressure (mmHg)

The Mode

- Rarely used
- The most frequently observed value

Class Exercise

- Calculate the arithmetic mean, median and the mode of the following data

2, 2, 3, 3, 3, 5

Mean:

Median:

Mode:

Class Exercise

- Calculate the arithmetic mean, median and the mode of the following data

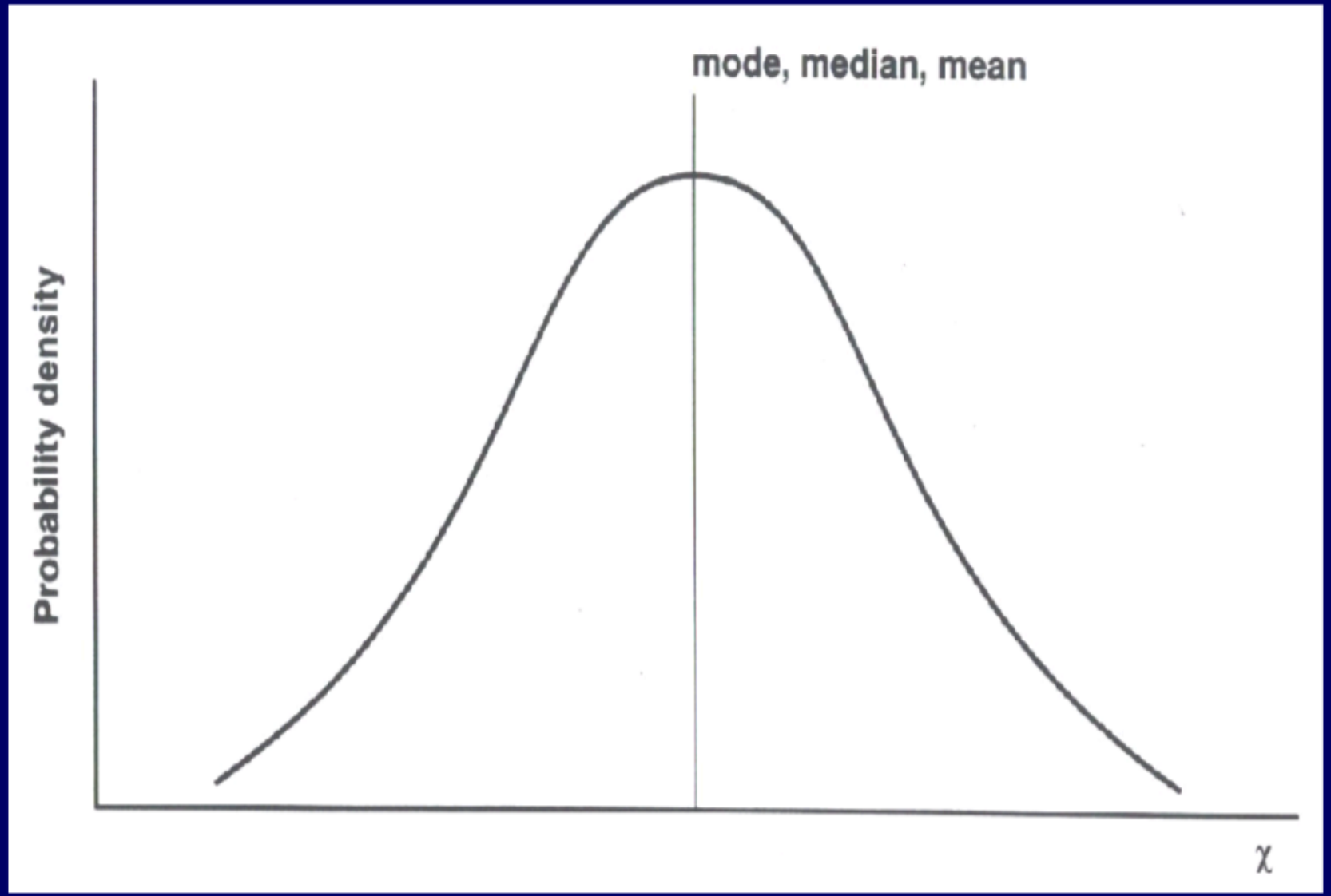
2, 2, 3, 3, 3, 5

Mean: (2+2+3+3+3+5) / 6 = 3

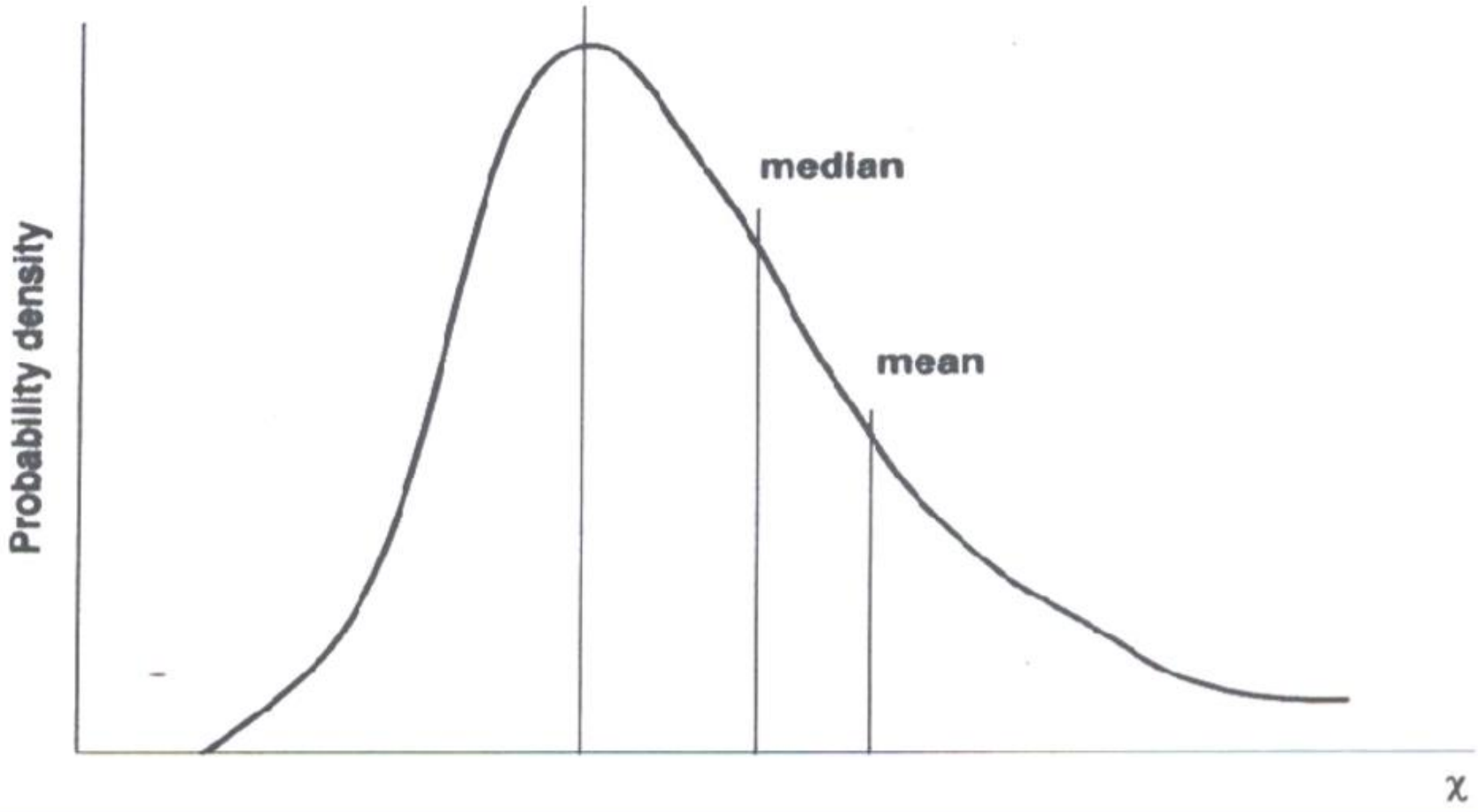
Median: 3

Mode: 3

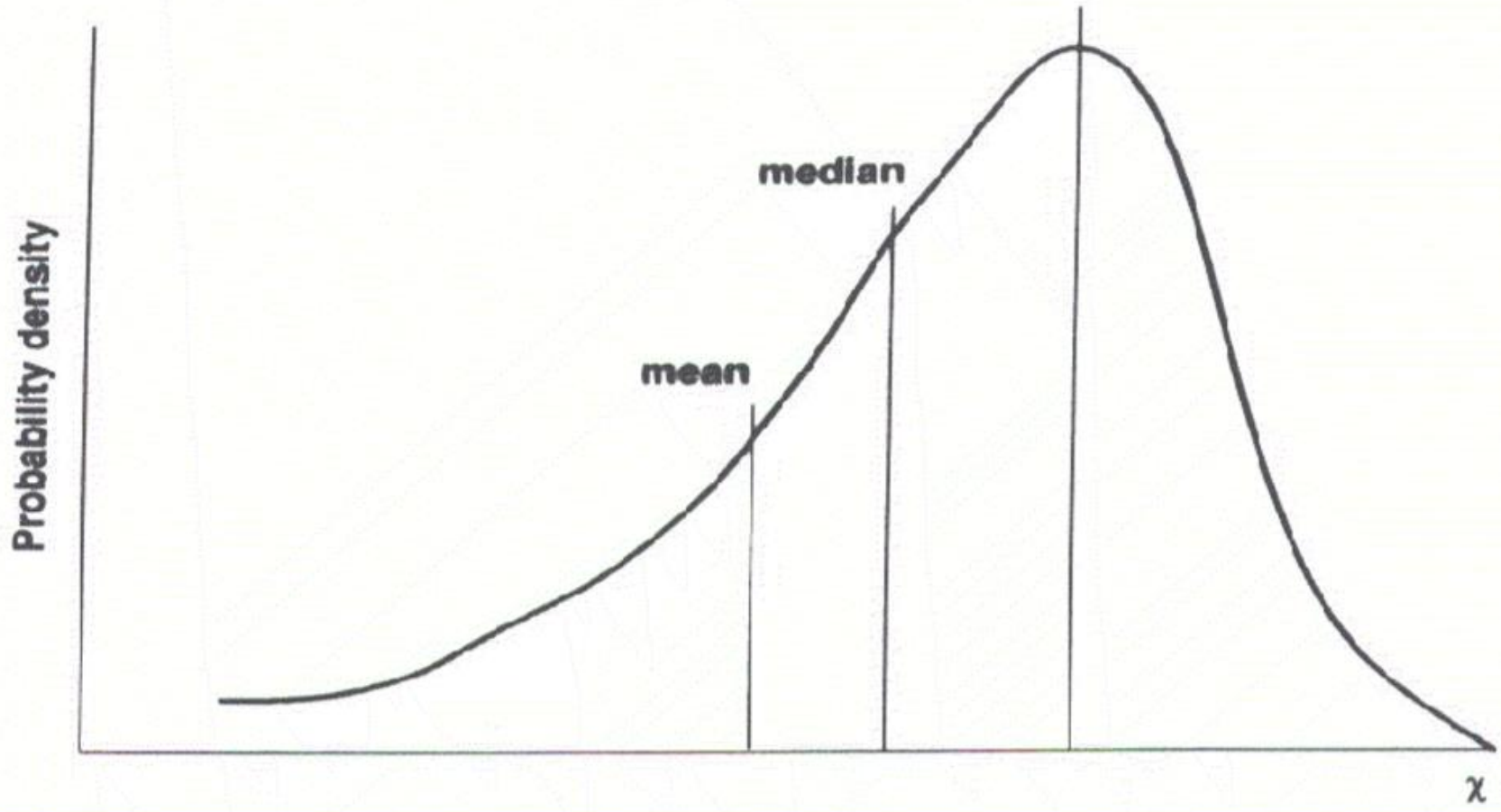
The Normal Distribution



Positively Skewed



Negatively Skewed



The Geometric Mean

- This is the "antilog" of the sum of the logged observations divided by n
- The antilog is the exponential function (e^x).
- "Transform" original data
- Commonly applied to *skewed* data

The Geometric Mean

- Example: calculate the geometric mean of the 8 plasma volumes:

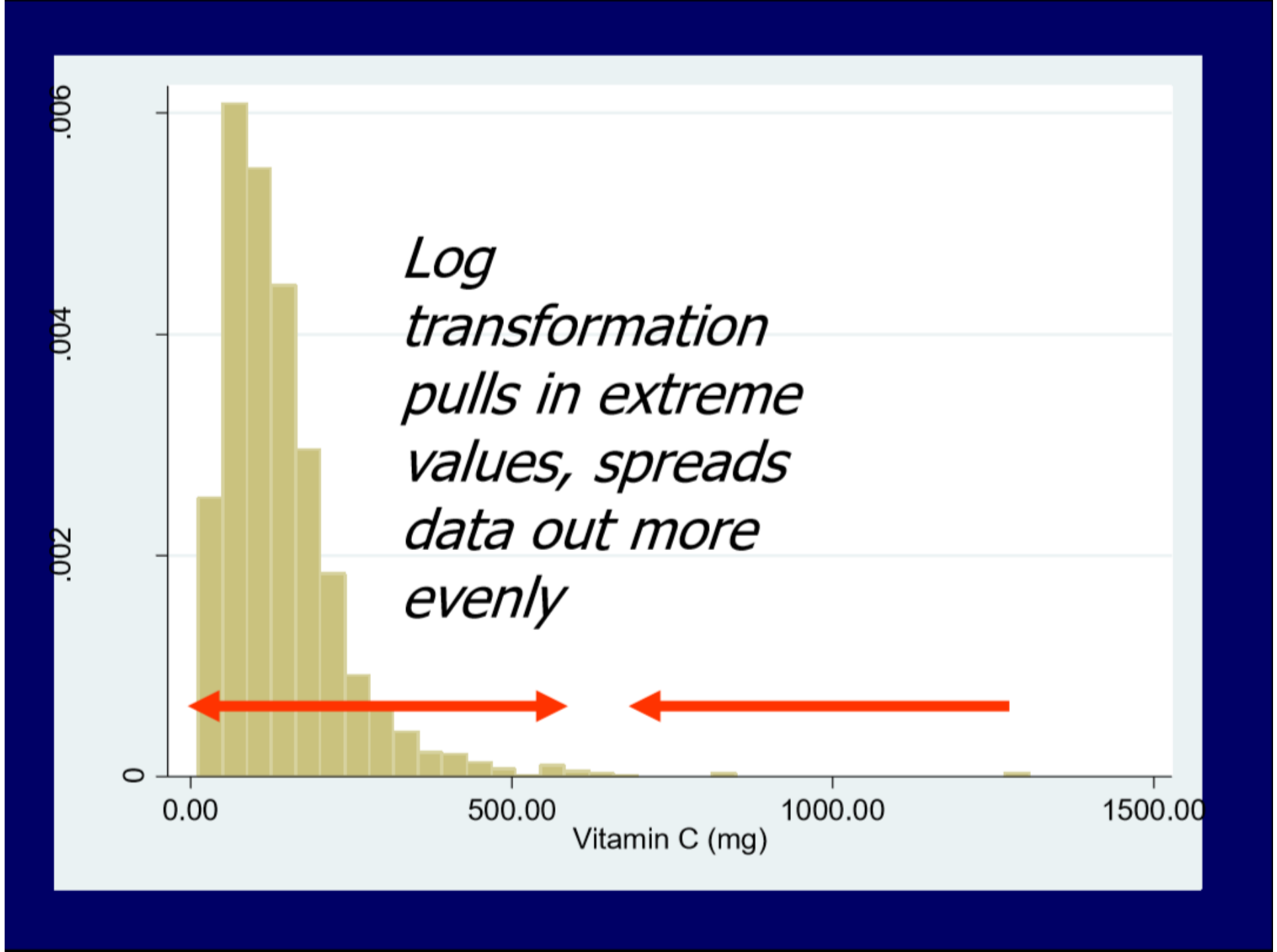
  2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12

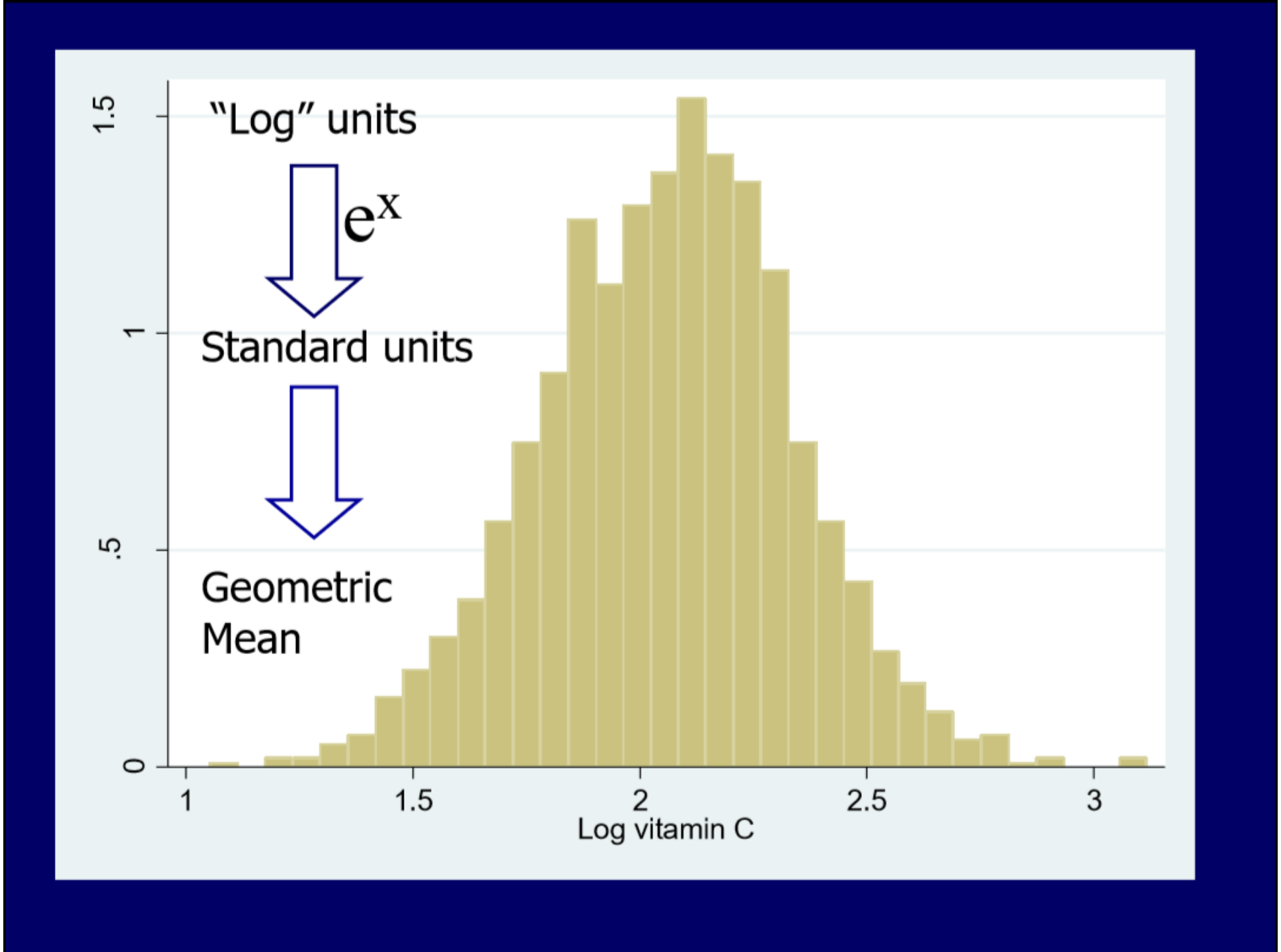
  anti-log ([ln(2.75) + ln(2.86) + … + ln(3.12)]/8)

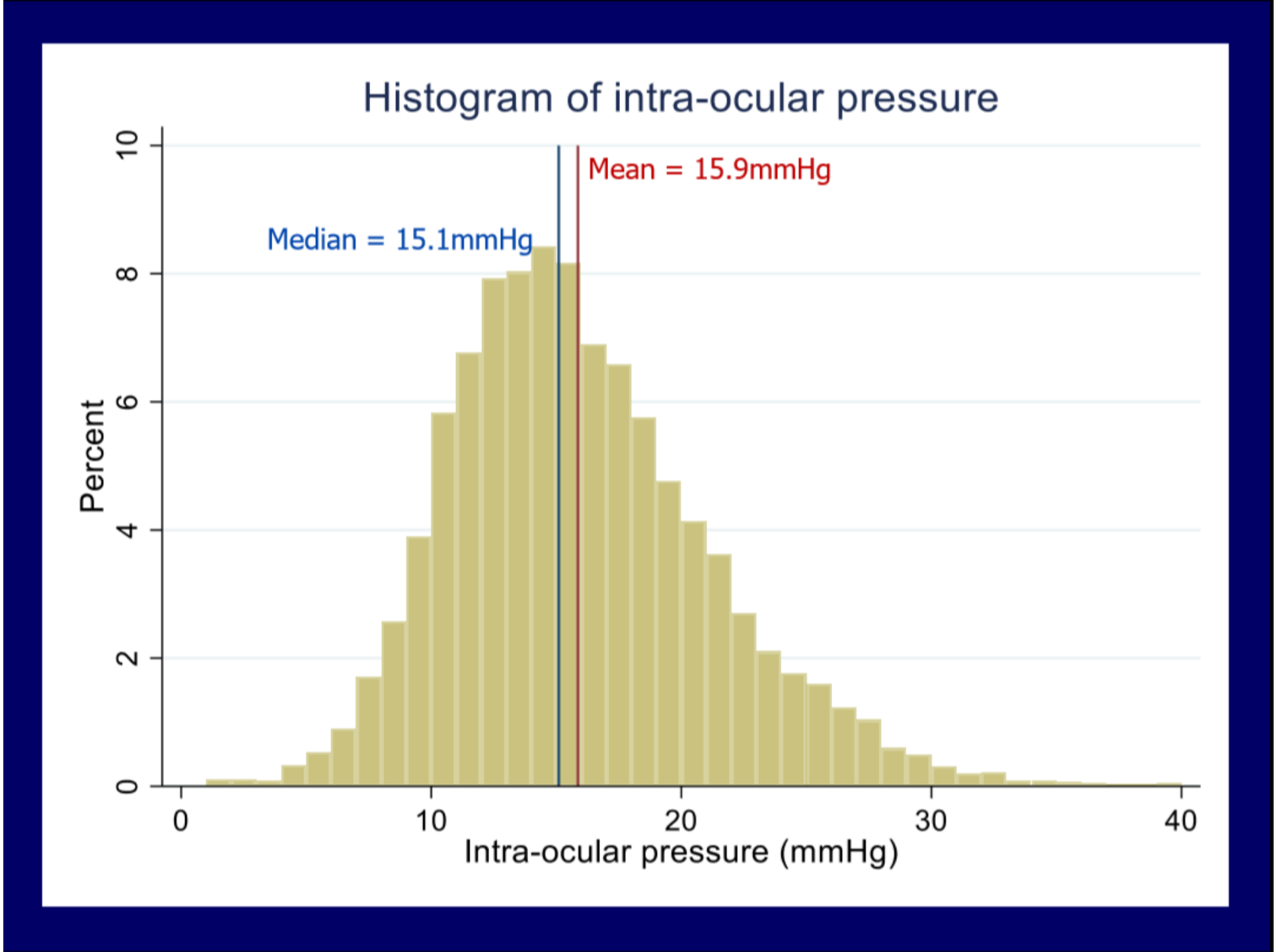
  = anti-log (8.75/8)

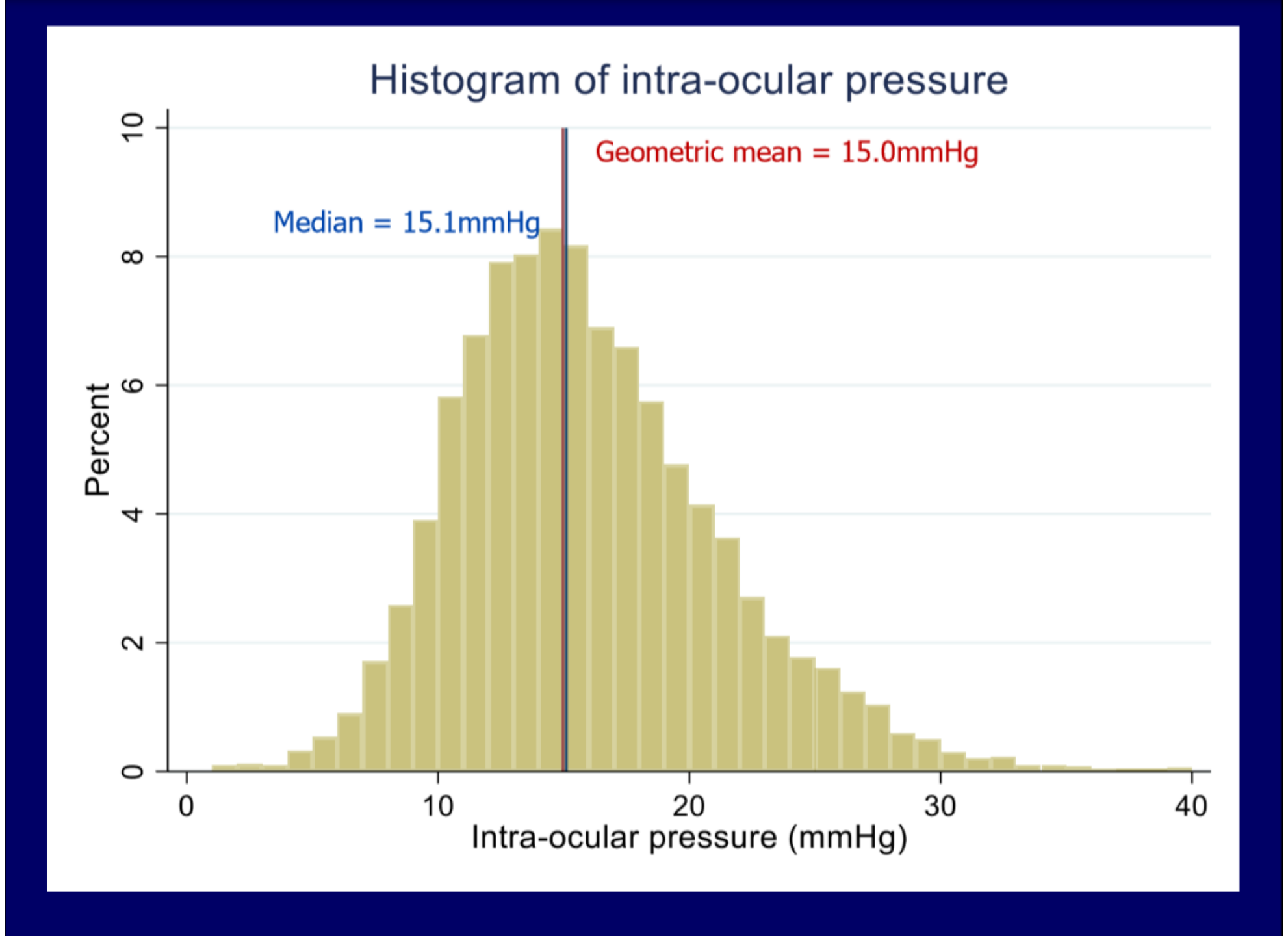
  = anti-log (1.09)

  = 2.98


.006
.004
.002
0
0.00
500.00
1000.00
1500.00
Vitamin C (mg)
Log transformation pulls in extreme values, spreads data out more evenly


"Log" units
e^x
Standard units
Geometric Mean
1.5
1
.5
0
1
1.5
2
2.5
3
Log vitamin C


Histogram of intra-ocular pressure
Mean = 15.9mmHg
Median = 15.1mmHg
Percent
0
2
4
6
8
10
0
10
20
30
40
Intra-ocular pressure (mmHg)


Histogram of intra-ocular pressure
Geometric mean = 15.0mmHg
Median = 15.1mmHg
Percent
0
2
4
6
8
10
0
10
20
30
40
Intra-ocular pressure (mmHg)

Outliers

- Take our data with one large outlying value again
- 2,2,3,4,4,5,6,6,30
- Arithmetic mean is 6.9
- Median is 4
- Geometric mean is 4.7
- Median and geometric means are less sensitive to outliers

Measures of Central Tendency

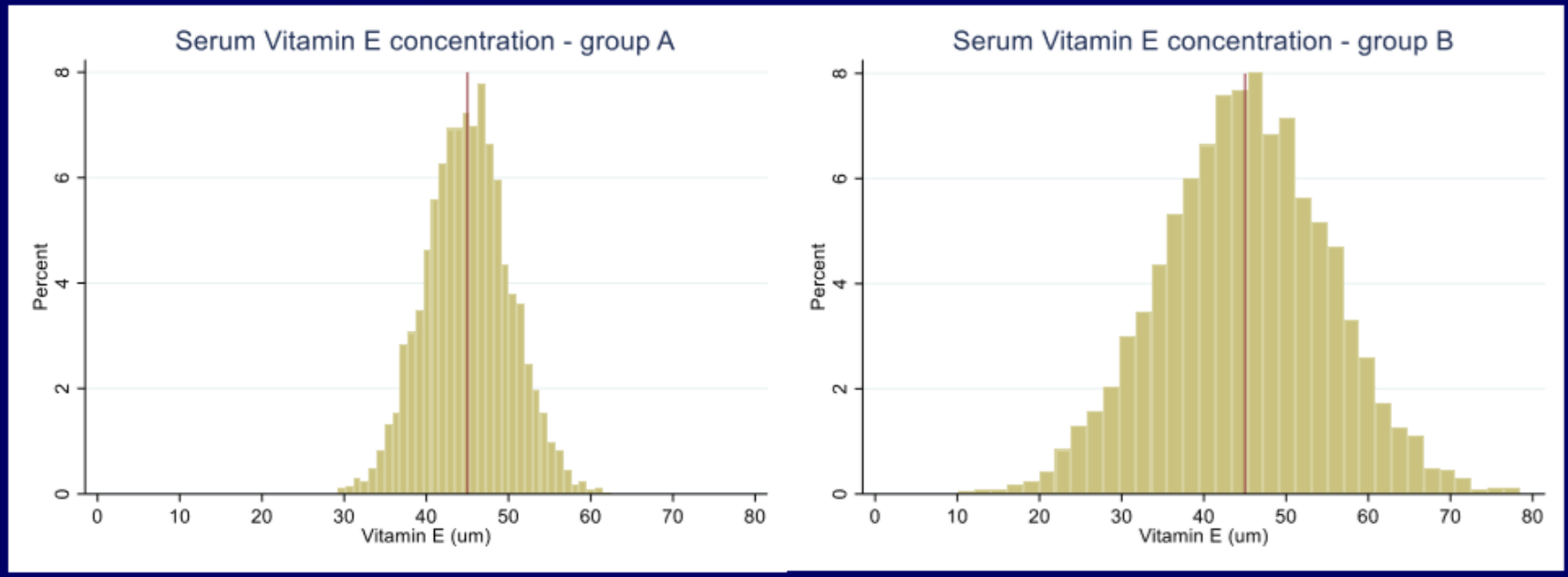
- (Arithmetic) Mean
- Geometric Mean
- Median
- Mode

Measures of Dispersion (Spread)

Measures of dispersion

- Means and medians only tell us so much
- How deep is that river…?
  - Only 25cm…

- Vitamin E concentration measured in two populations

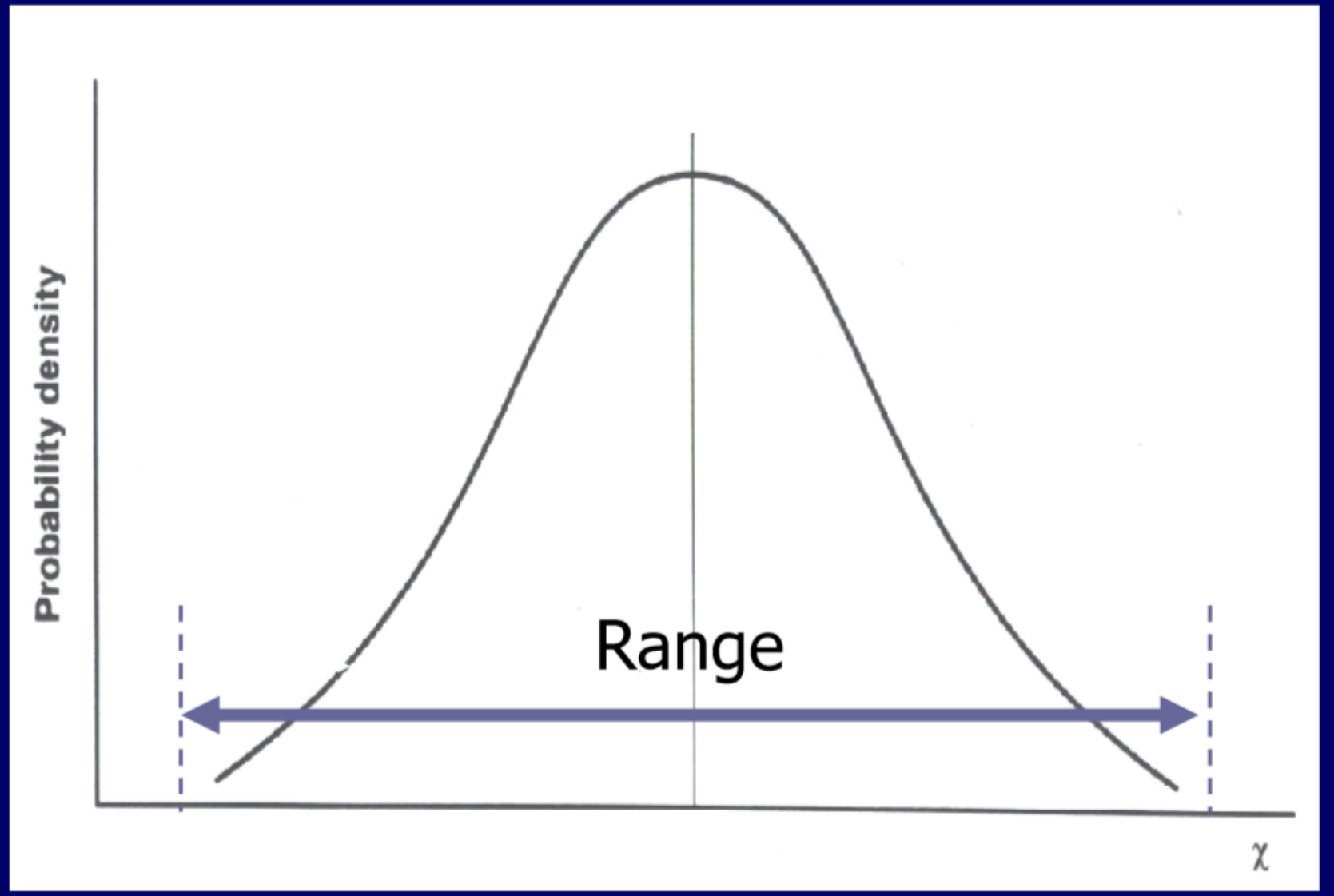


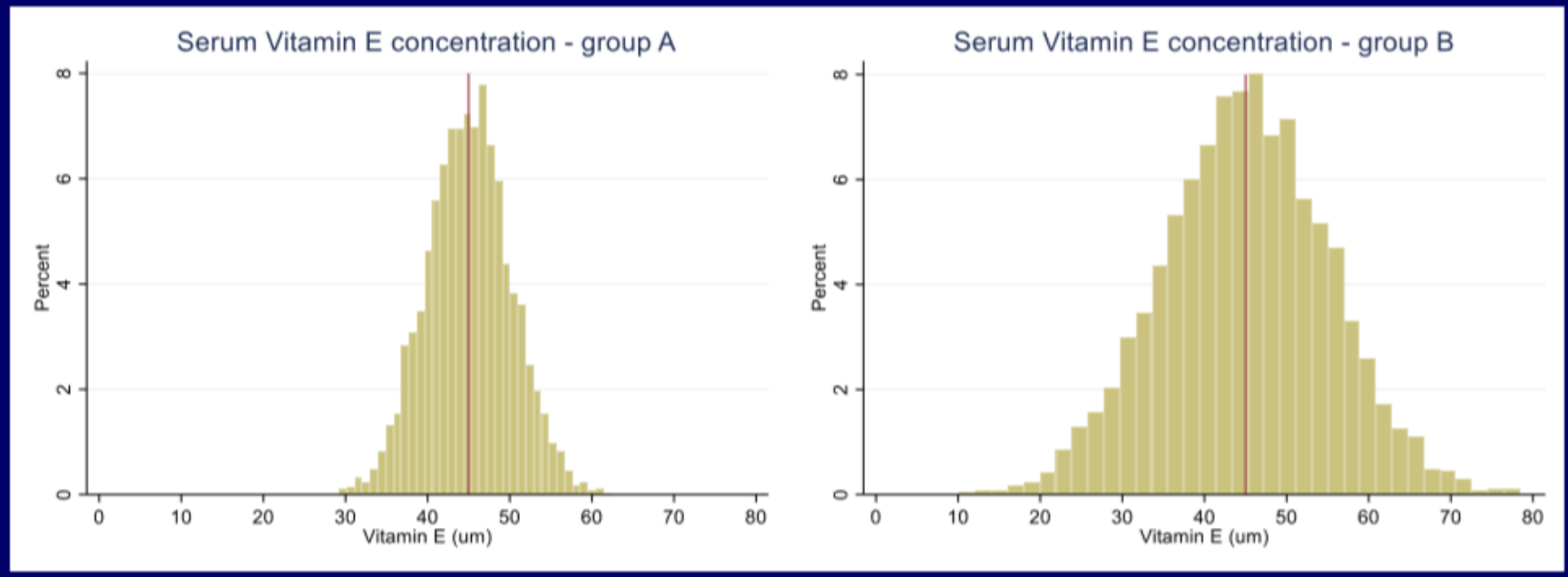
- Mean in both groups the same (45μM/l)
- But distribution of data is clearly different

Measures of Dispersion

- Range
  - Minimum to maximum value
  - Difference between minimum & maximum values
  - Highly affected by outlying values

The Normal Distribution





- Range in group A : 29-62
- Range in group B : 10-78

Measures of Dispersion

- Standard deviation (SD)
  - SD is a measure of the average spread of values about the mean
    - Small SD → most values lie very close to the mean
    - Large SD → many values lie far from the mean

Standard deviation

- How far on average observation is from the mean observation
  - Formula

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

  - Easier to explain using an example!

- Example: Calculate the standard deviation of the 8 plasma volumes:

  2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12

First calculate the mean plasma volume:

(2.75+2.86+3.37+2.76+2.62+3.49+3.05+3.12) /8

= 24.02/8

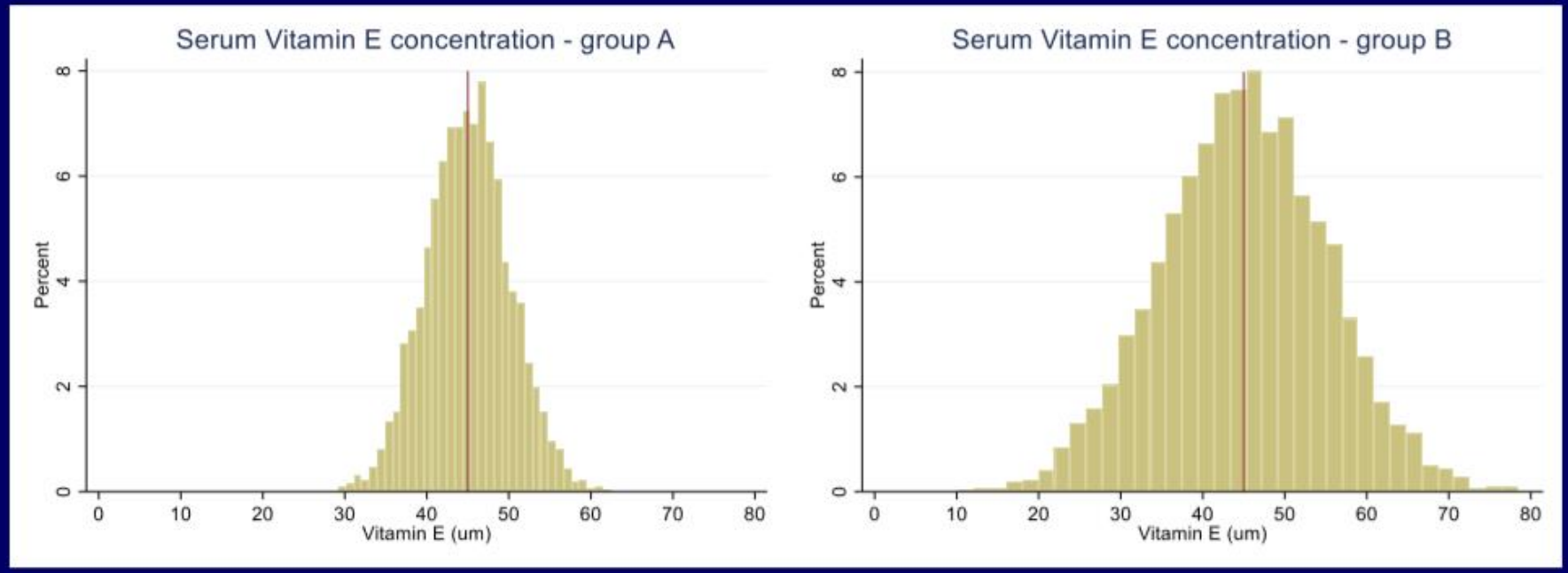
= 3.0025 litres.

Observation	Deviation from mean (3.0025)	Squared deviation
2.75	2.75-3.0025=-0.2525	0.0638
2.86	-0.1425	0.0203
3.37	0.3675	0.1351
2.76	-0.2425	0.0588
2.62	-0.3825	0.1463
3.49	0.4875	0.2377
3.05	0.0475	0.0023
3.12	0.1175	0.0138
Total	*0*	*0.6781*

Sum of squared deviations = 0.6781

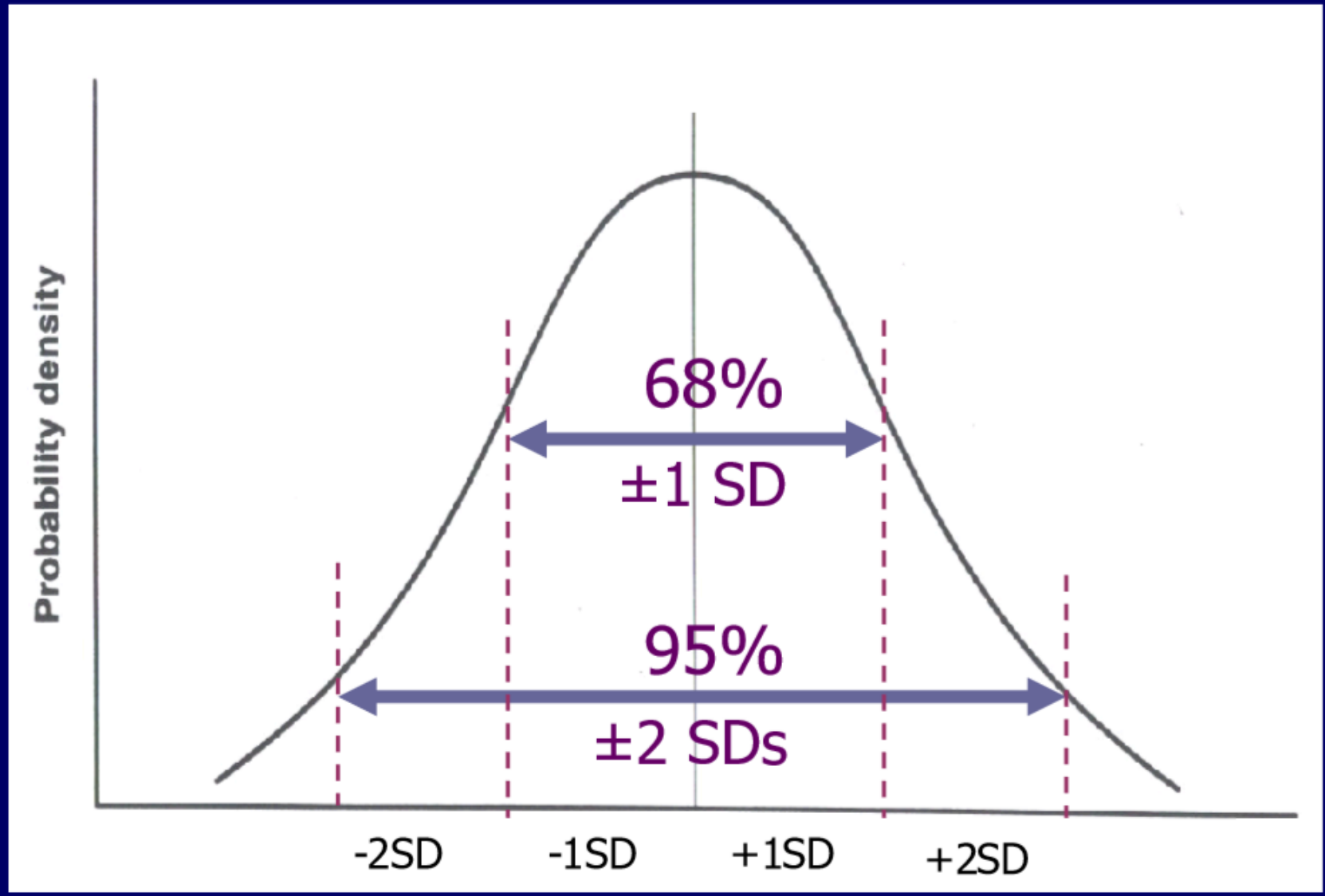
Variance = $\frac{\sum(x-\bar{x})^2}{n-1}=\frac{0.6781}{7}=0.0968$

Standard Deviation (SD) = $\sqrt{0.0968}=0.3112$



- SD in group A : 5.1
- SD in group B : 10.2

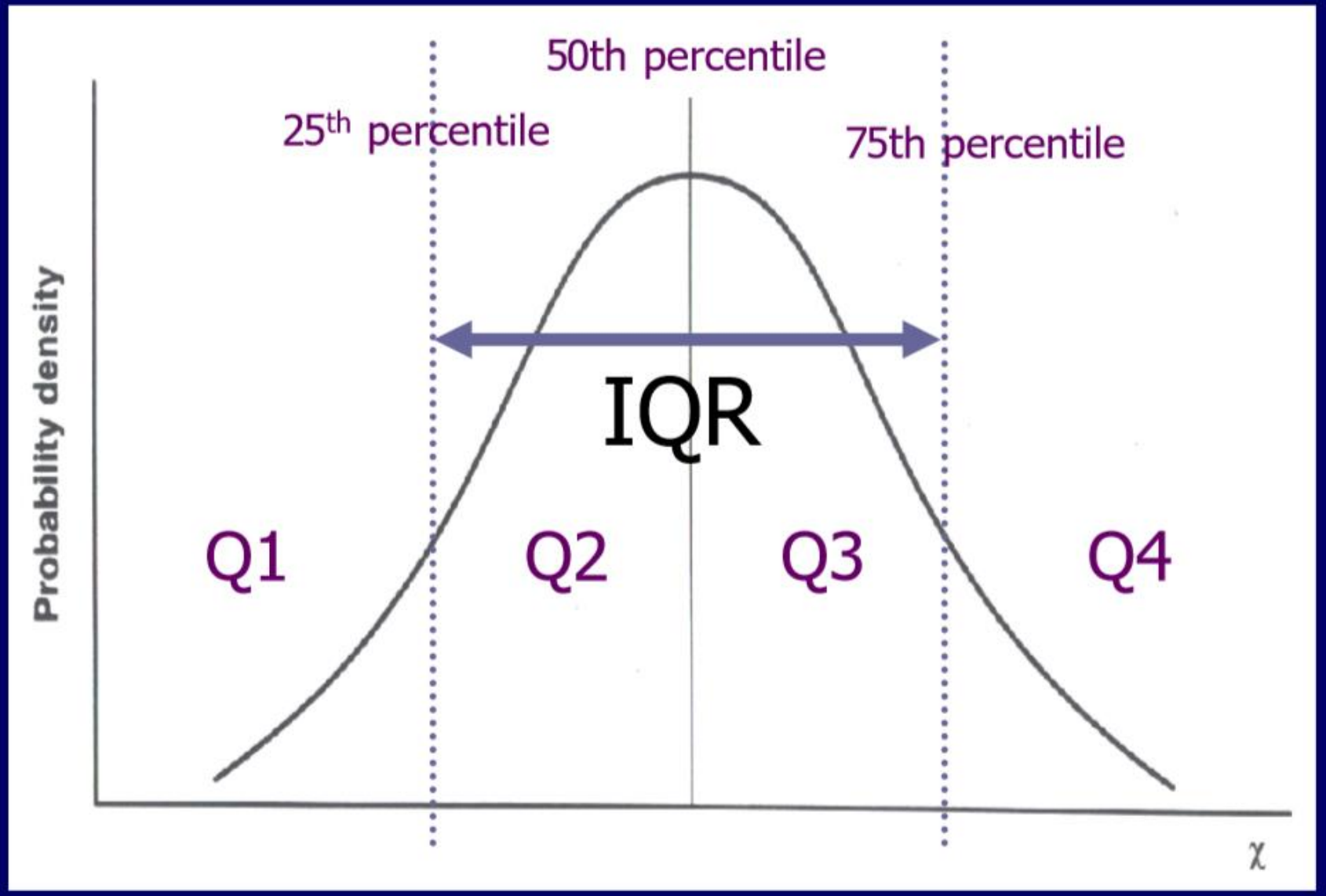
The Normal Distribution

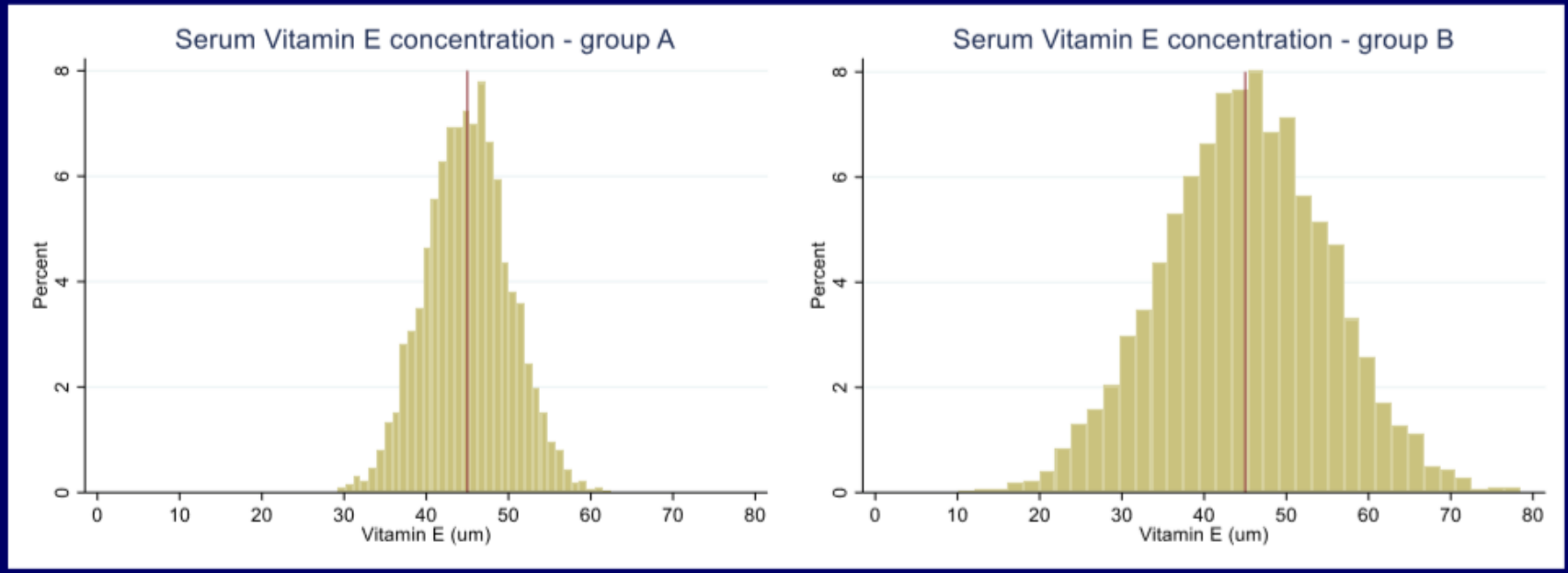


Measures of Dispersion (Spread)

- Inter-quartile range (IQR)
  - 25^{th} to 75^{th} percentile
  - (75^{th} percentile - 25^{th} percentile)

The Normal Distribution





- IQR in group A : 41.8-48.5
- IQR in group B : 38.2-52.3

Describing Categorical Variables

Categorical Variables

- Variables with 2+ categories (classes)
- Individual can only belong to one category
- **Nominal**
  - No intrinsic ordering of categories
  - Examples: *gender, blood group*
- **Ordinal**
  - Ordering is important
  - *Cancer staging (I-IV)*
  - *Education level*

Examples of Categorical Variables

- Occupation
- Pregnancy status
- Mortality
- Eye colour
- Marital status

- Examples of others?

Describing Categorical Variables

Gender	Count	Proportion
Male	381	48.2%
Female	410	51.8%
Total	791	

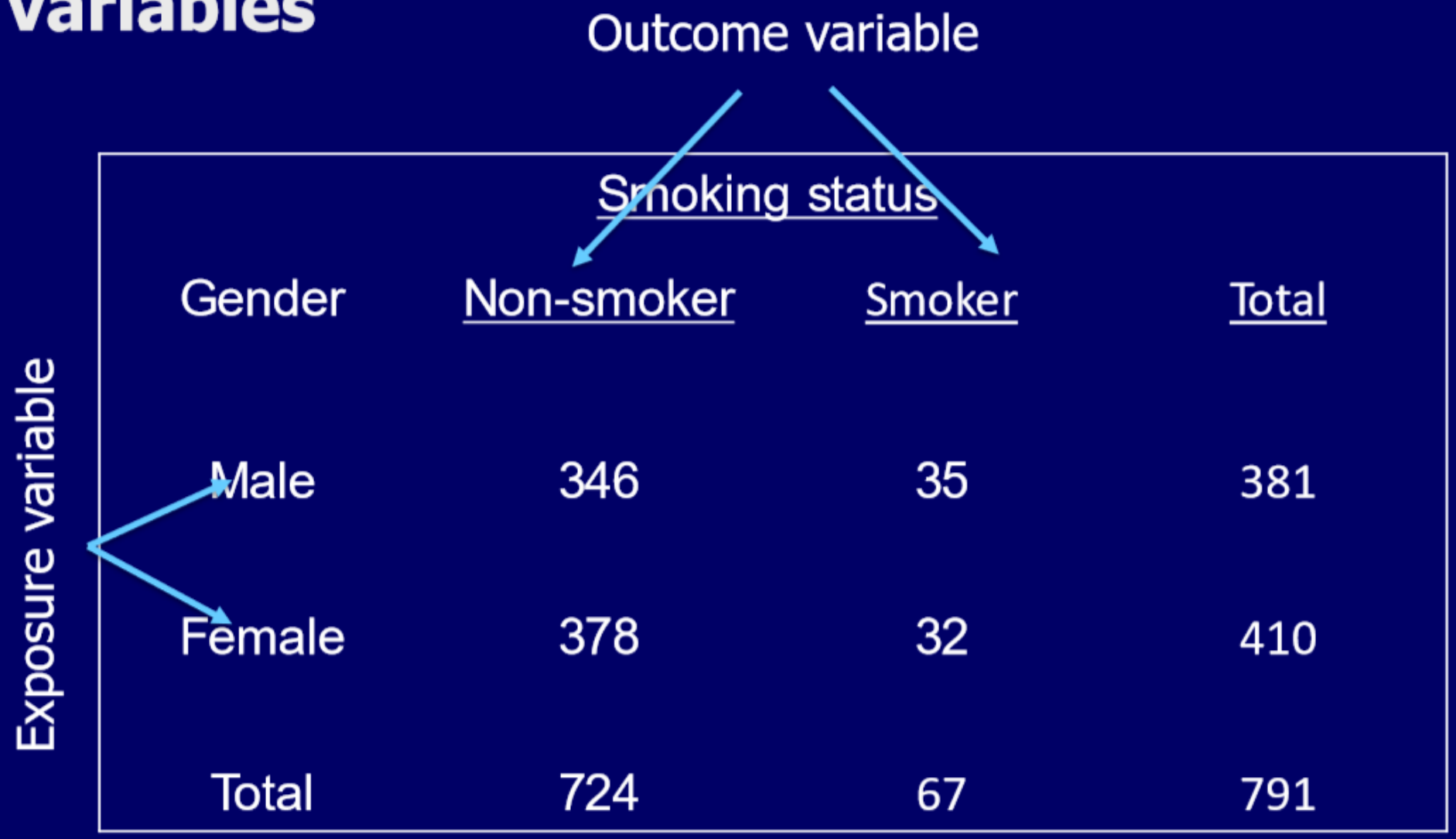
Simple tabulations are the best way to show data distribution in categorical variables

Describing Categorical Variables

Smoking status	Count	Proportion
Non-smoker	724	91.5%
Smoker	67	8.5%
Total	791	

Simple tabulations are the best way to show data distribution in categorical variables

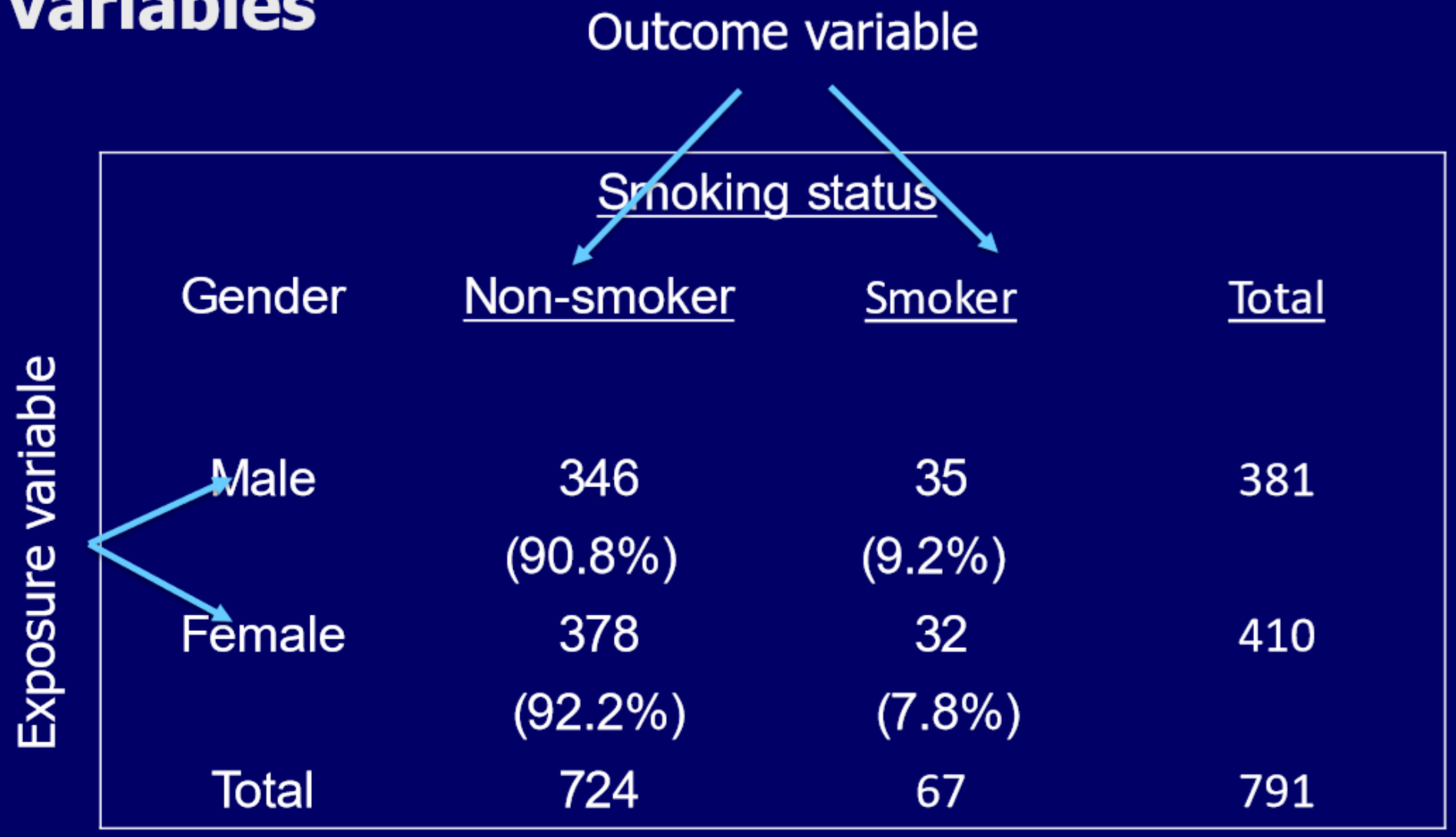
Association between two categorical variables



Gender	Non-smoker	Smoker	Total
Male	346	35	381
Female	378	32	410
Total	724	67	791

Row or column percentages?

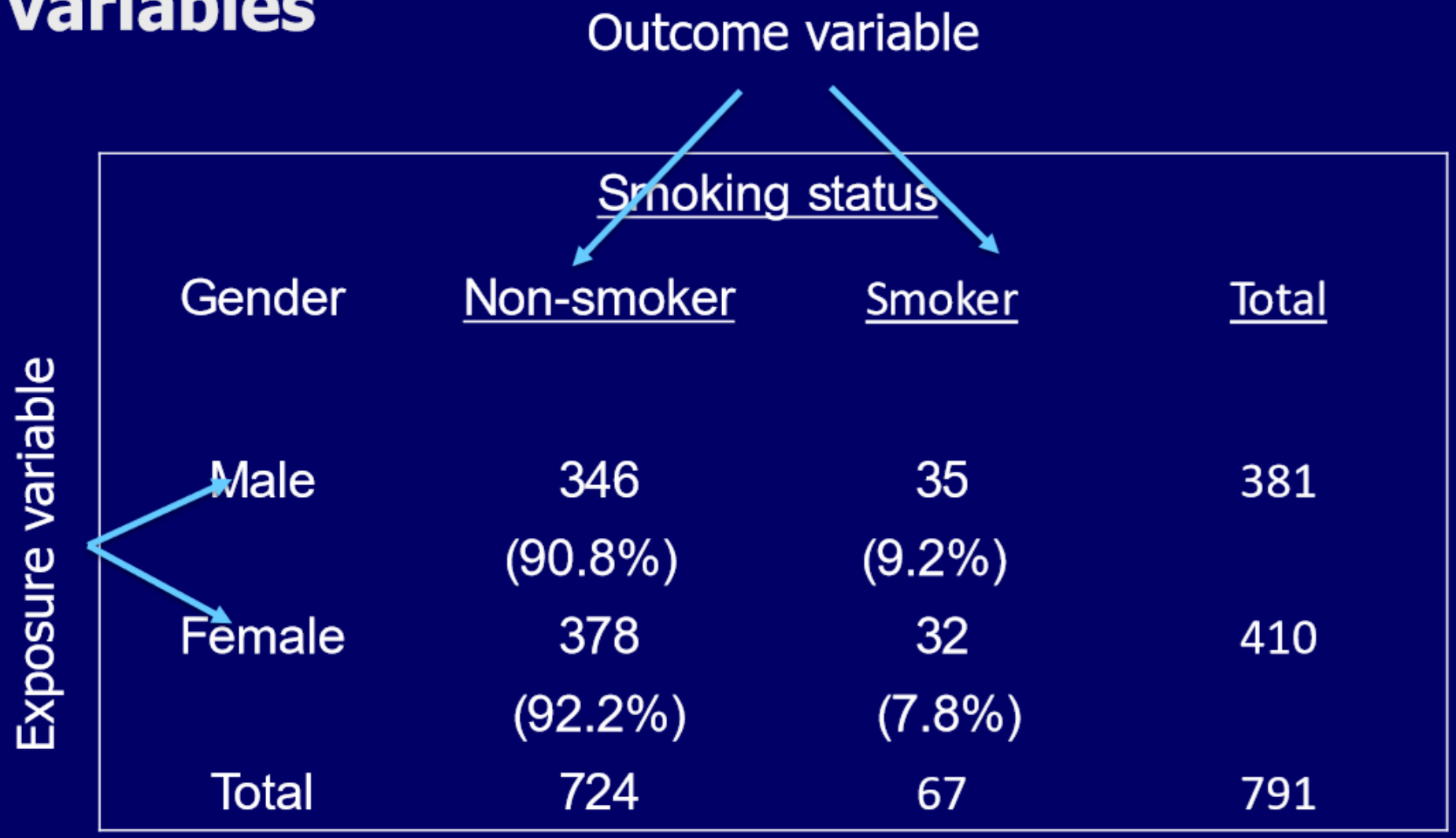
Association between two categorical variables



Gender	Non-smoker	Smoker	Total
Male	346 (90.8%)	35 (9.2%)	381
Female	378 (92.2%)	32 (7.8%)	410
Total	724	67	791

Row or column percentages?

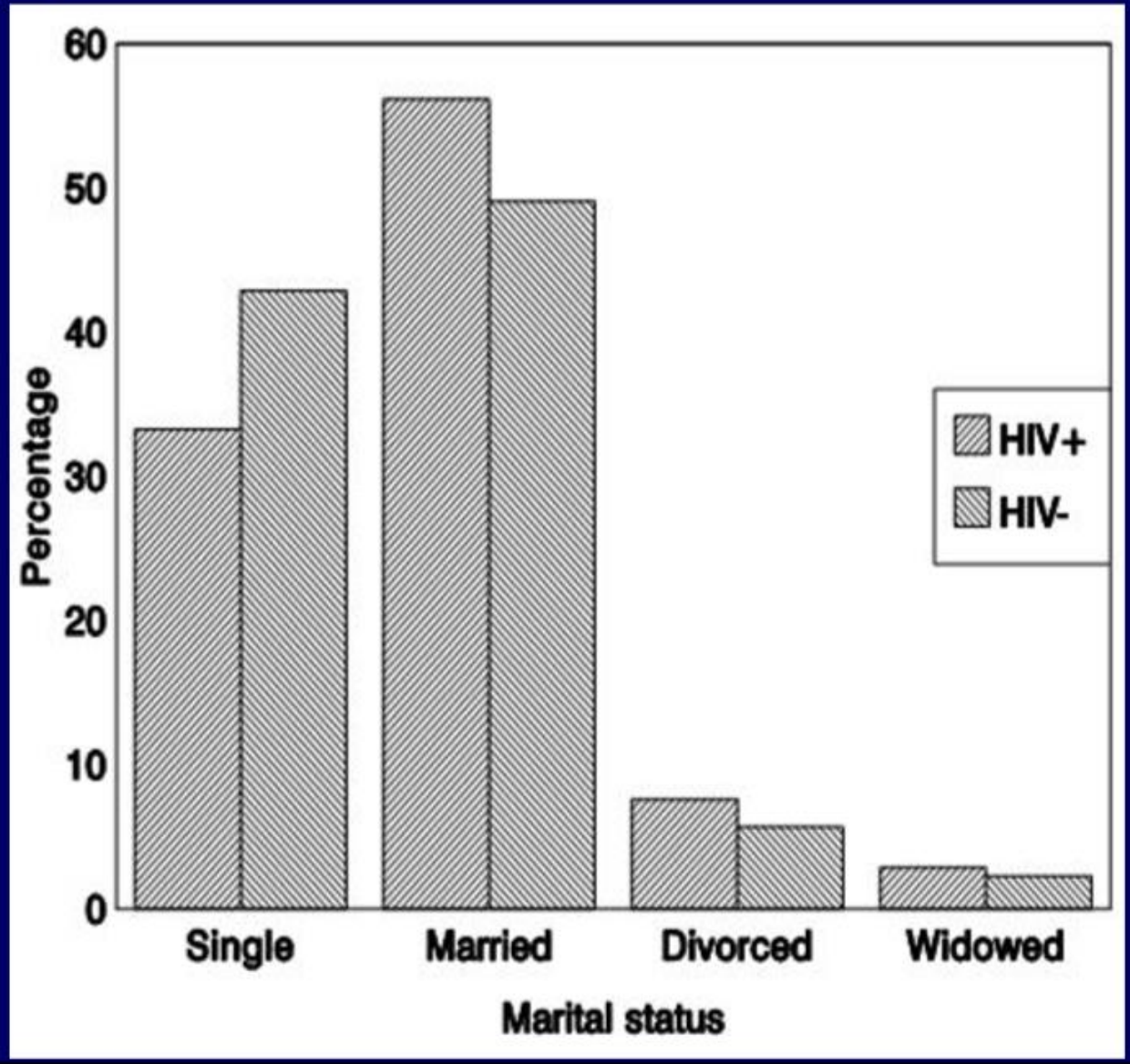
Association between two categorical variables



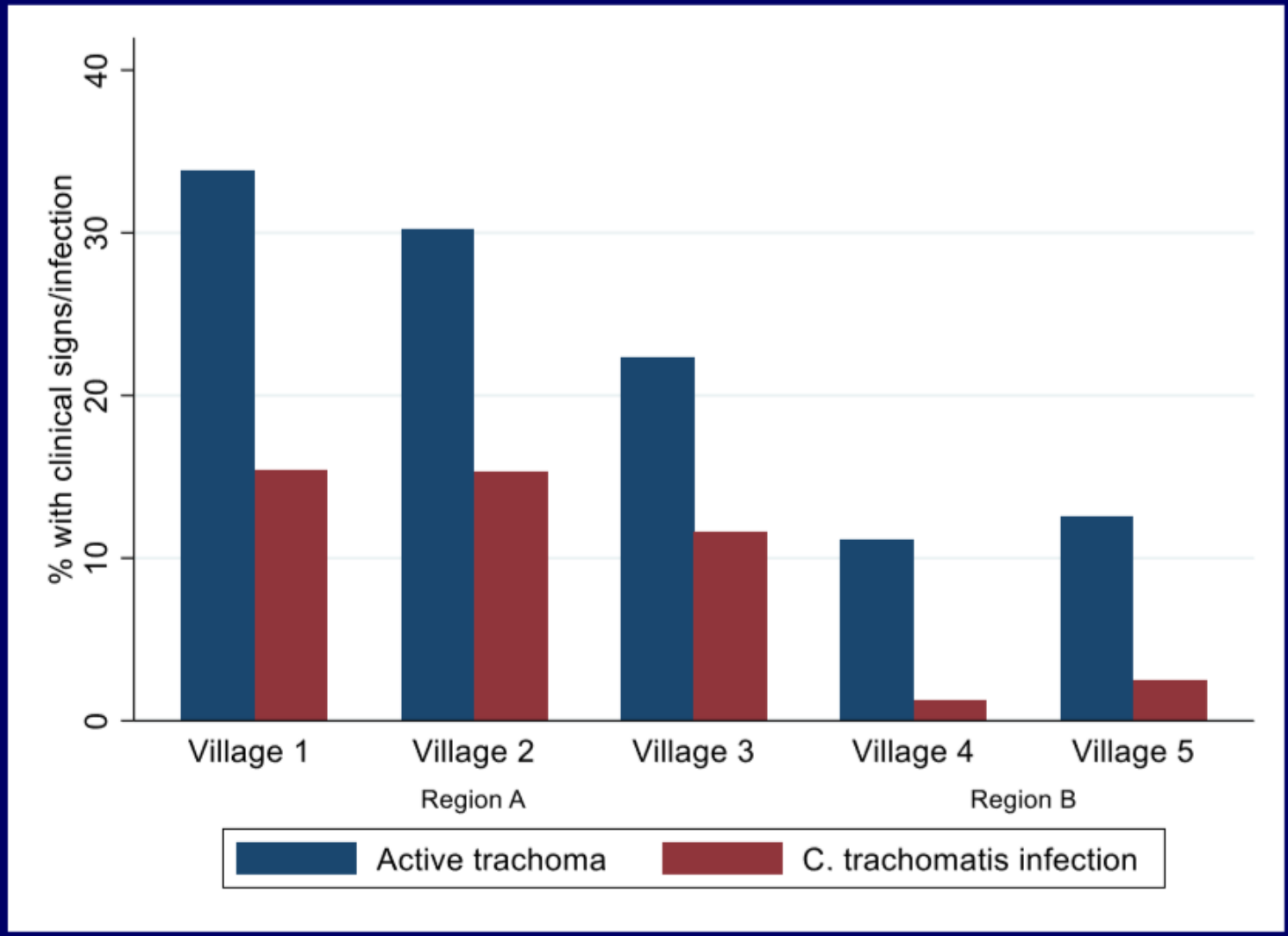
Gender	Non-smoker	Smoker	Total
Male	346 (90.8%)	35 (9.2%)	381
Female	378 (92.2%)	32 (7.8%)	410
Total	724	67	791

Think about whether you want row or column %

Bar charts to display the distribution of a categorical variable: distribution of marital status in the HIV positive and negative study participants



Bar charts to display proportions within categories: Proportion with clinical signs or trachoma and infection across 5 villages



Converting continuous variables to categorical variables

Quantitative variables converted to categorical variables…

- To form clinically important categories
- To form equally weighted categories (quartiles)

Quantitative variables converted to categorical variables…

- To form clinically important categories
- To form equally weighted categories
- Description and analysis of data are treated as a categorical variable
- "Homogeneity" assumption: values are equivalent within a given category

Example: Body Mass Index

$< 18.5\ kg/m^2$	Underweight
$18.5 – 24.9\ kg/m^2$	Normal weight
$25 – 29.9\ kg/m^2$	Overweight
$30 – 34.9\ kg/m^2$	Obese
$35+\ kg/m^2$	Morbidly obese

Example:
Age (10 year categories)

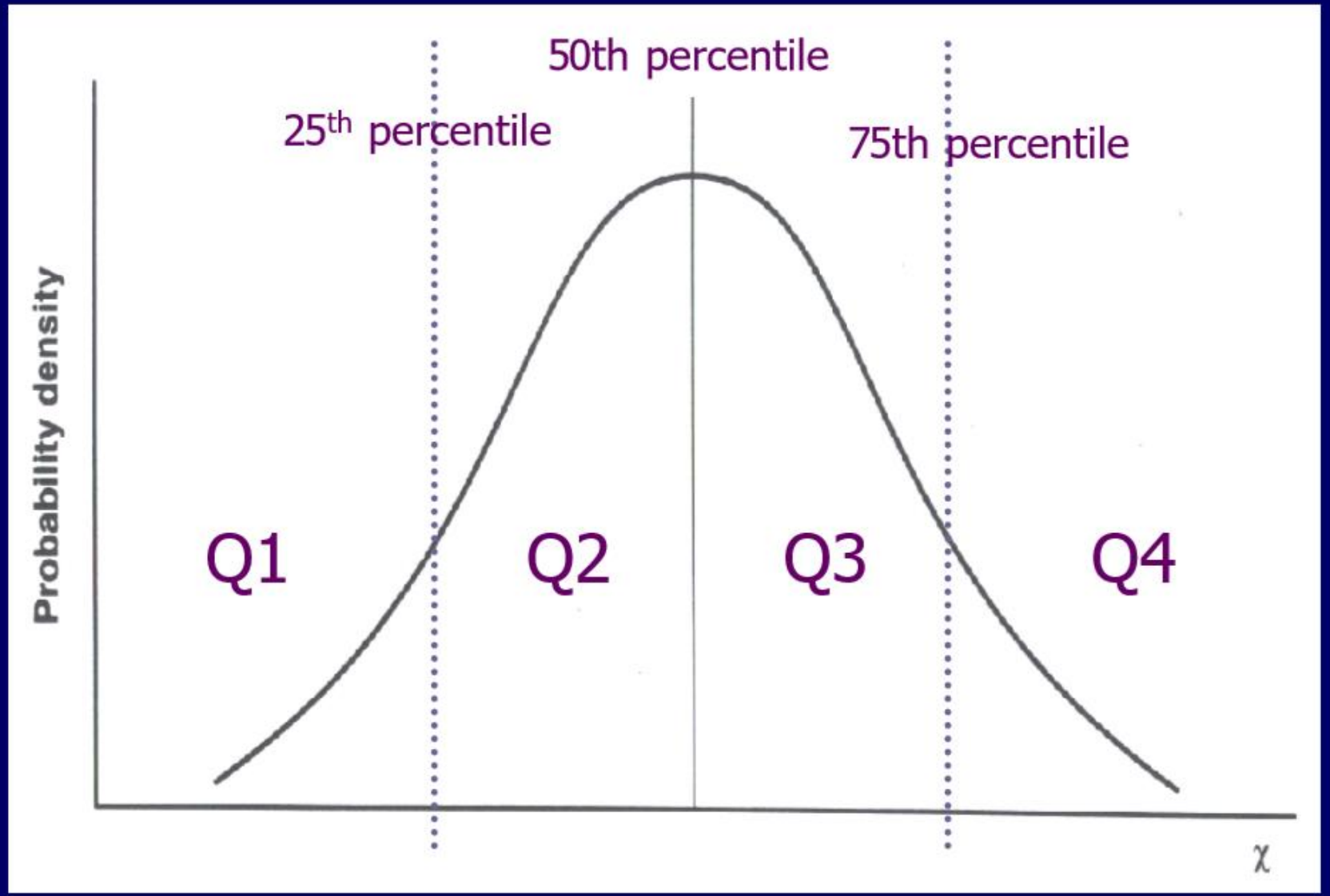
- 21 years – 30 years
- 31 years – 40 years
- 41 years – 50 years
- 51 years – 60 years
- 61 years – 70 years
- > 70 years

Example:
Age (binary categories)

- ≤ 60 years "Young" age
- > 60 years "Old" age

Any concerns?

The Normal Distribution



Example: Age (quartiles)

- 21 years – 32 years — Quartile 1
- 33 years – 55 years — Quartile 2
- 56 years – 62 years — Quartile 3
- >62 years — Quartile 4

Any concerns?

Considerations when categorising continuous variables

- Be mindful of "homogeneity" assumption
- Best to use "clinically important" categories
- Use a data-driven approach (e.g. quartiles) useful if no *a priori* information available
- If possible, use all the data! (i.e. don't categorise)

Q: Which should we use?

Arithmetic mean? Median? Geometric mean?

Arithmetic mean

Useful for statistical inference (i.e., statistical tests)
Influenced by outliers

Median

Not affected by outliers
Useful for data which are not symmetrical
Not amenable to many statistical tests

Geometric mean

Useful if data are skewed

Data and Distributions

Variable type

Categorical

- Binary
- Ordered categorical

Frequency tables (n & %)

Bar chart (or pie chart)

Quantitative

Summary statistics:

- Mean & SD
- Median & IQR
- Geometric mean
- Histogram

Summary

- The type of variable drives the choice of summary measure & graphical presentation of the data
- Quantitative & Categorical variables require different approaches to summarising data
- Be aware of distributional assumptions when using a summary measure
- Be aware of limitations when categorising a continuous variable