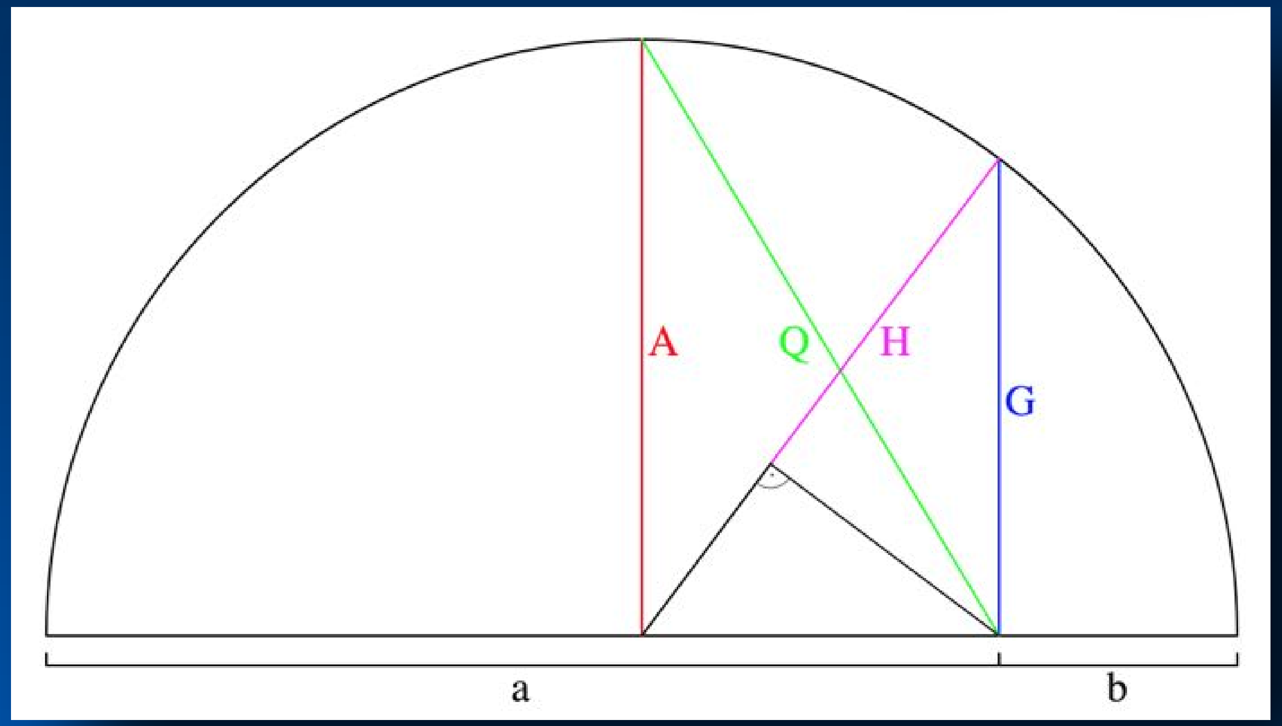
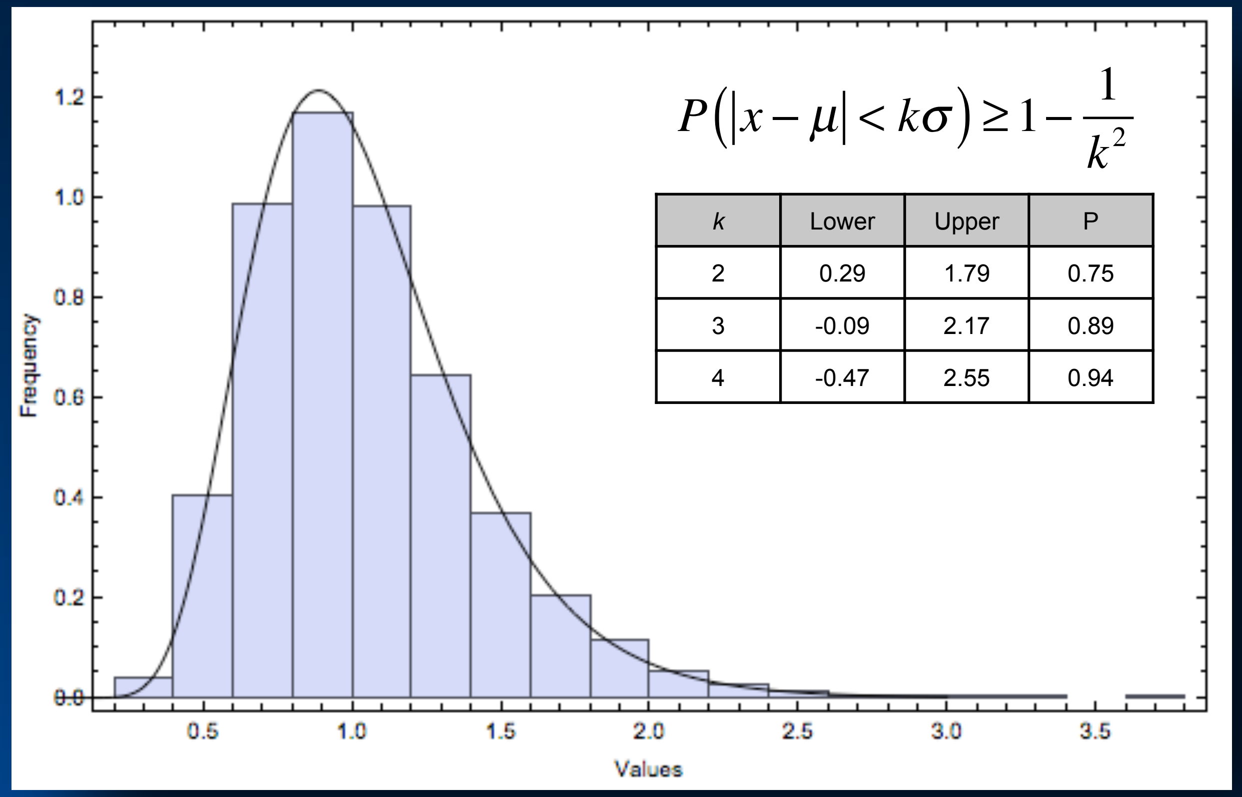
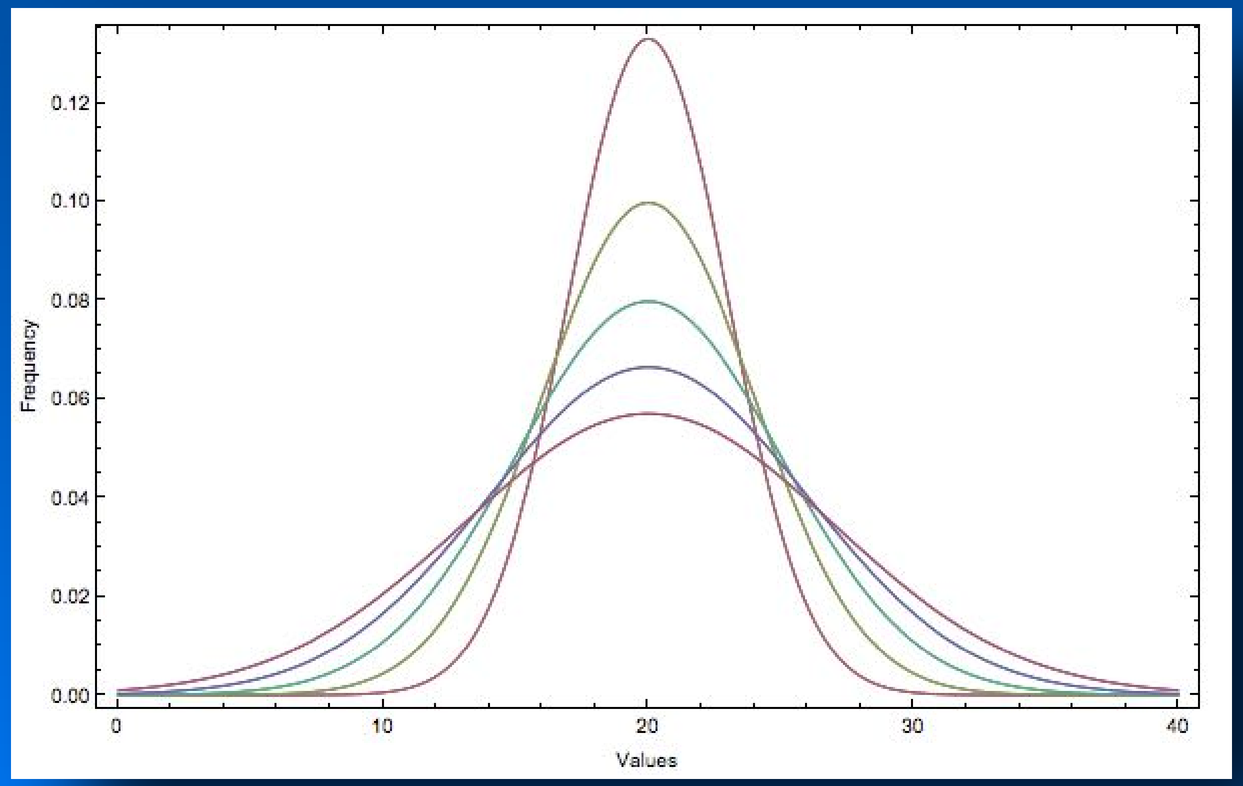
Statistics & Probability for Earth Scientists - A Review?

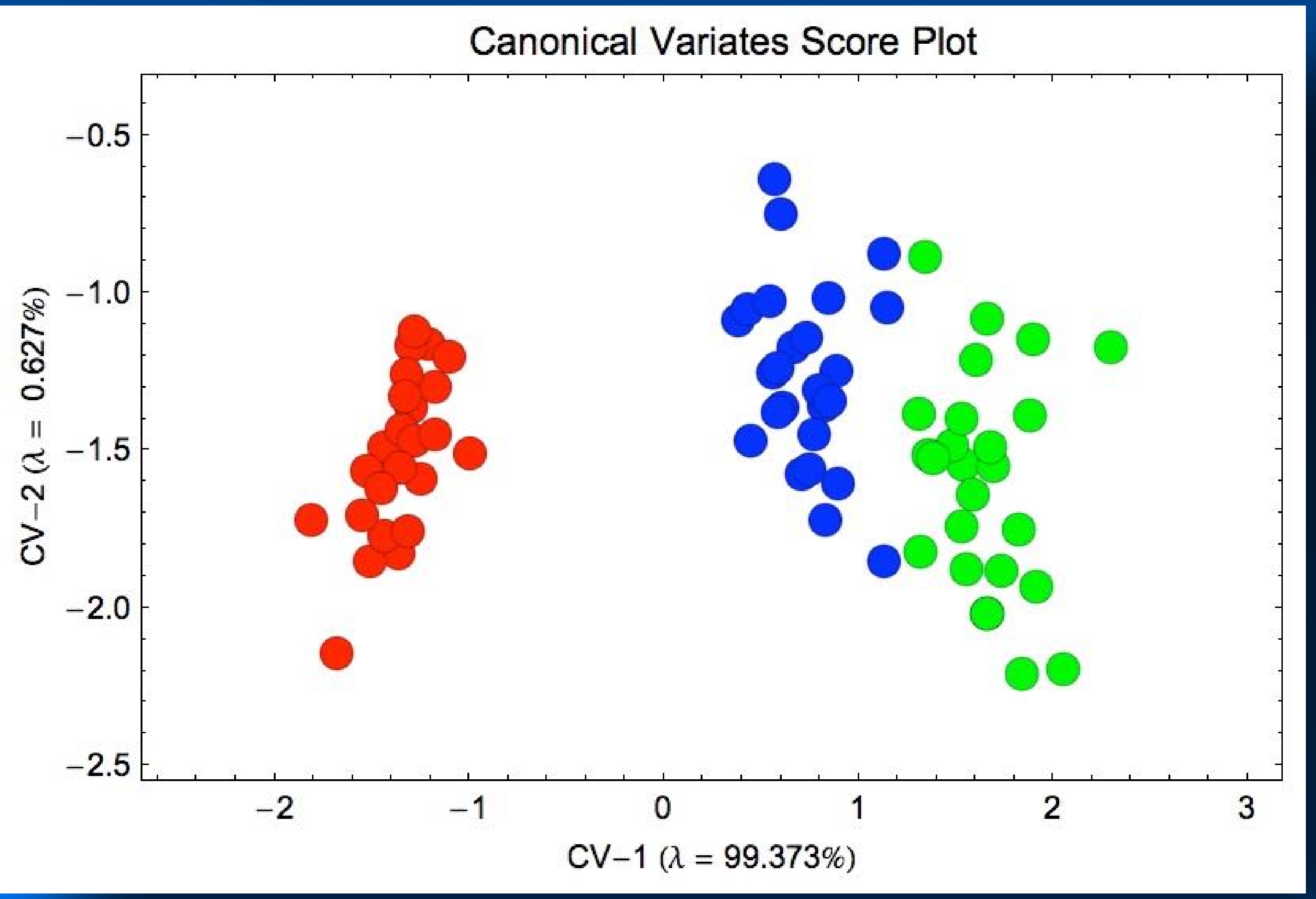
Prof. Norman MacLeod
School of Earth Sciences & Engineering, Nanjing University





k	Lower	Upper	P
2	0.29	1.79	0.75
3	-0.09	2.17	0.89
4	-0.47	2.55	0.94



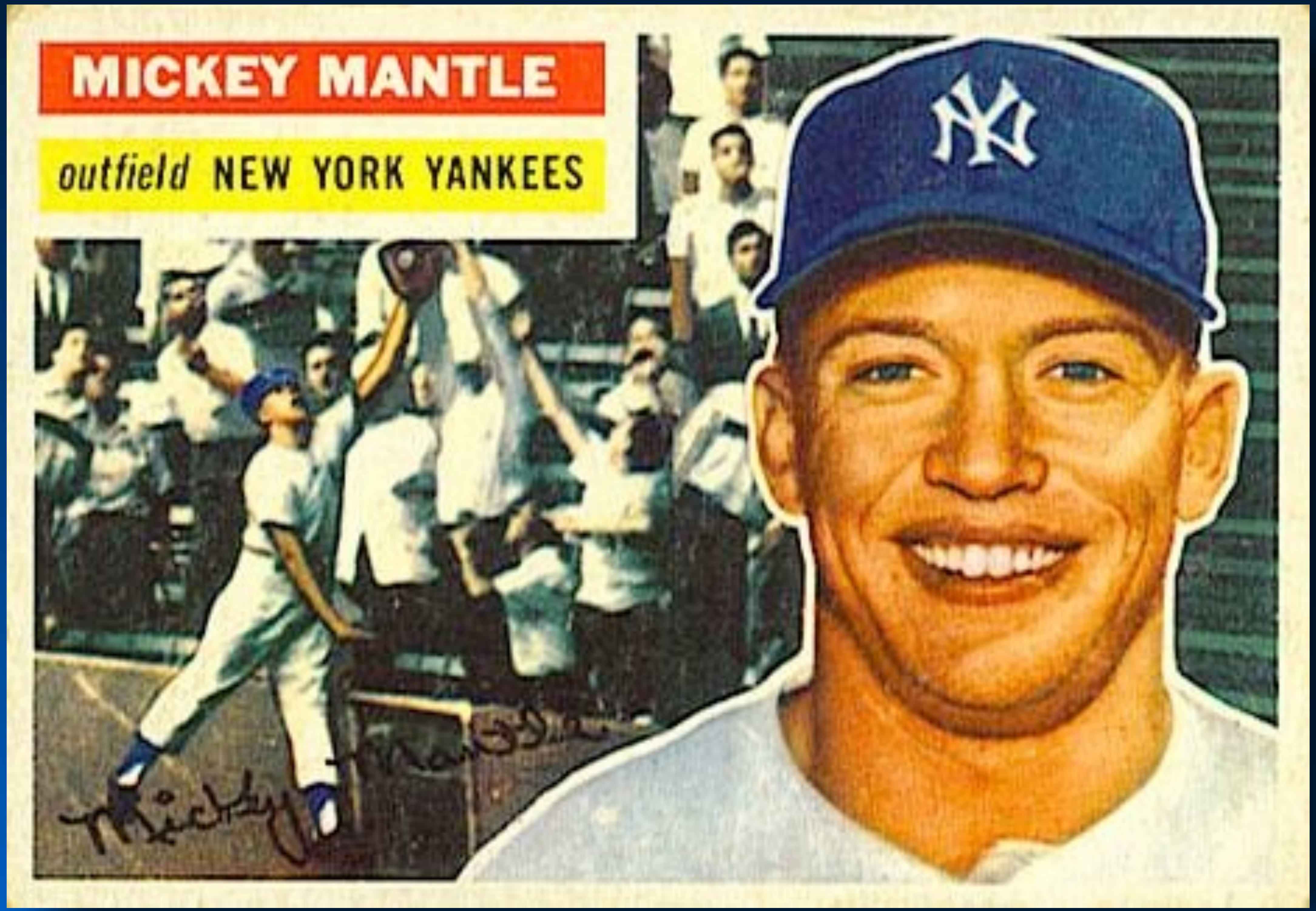


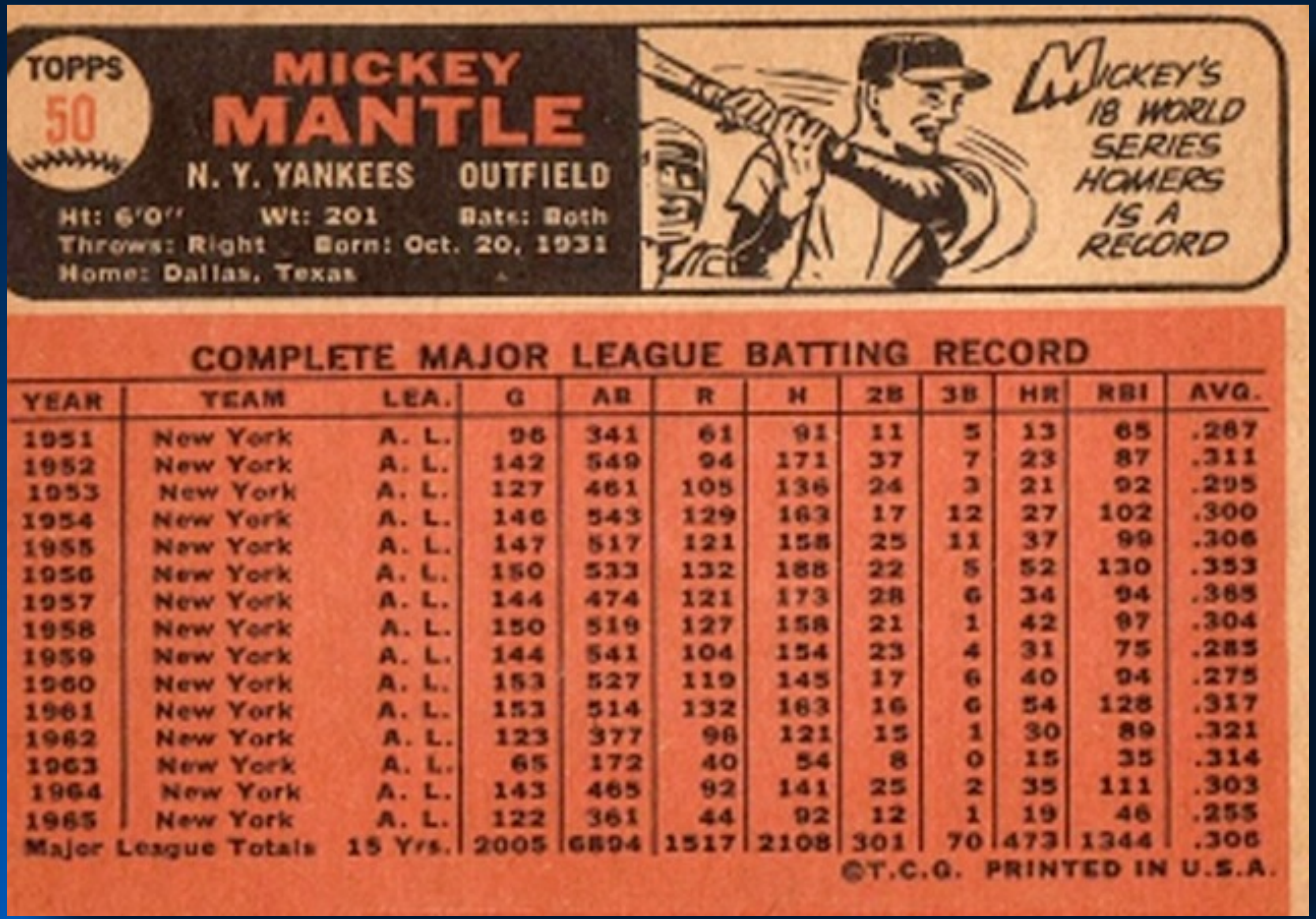
Definitions of “Statistics”

- Any list of numerical data.
- The science of making effective use of numerical data relating to individual samples, groups of samples or experiments.
- A set of concepts, rules, and procedures that help us to:
  - Organize numerical information in the form of tables, graphs, and charts;
  - Understand statistical techniques underlying decisions that affect our lives and well-being; and make informed decisions.
- The set of possible values for a random variable together with a probability measure quantifying the likelihood of those values.
- The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling.

Statistics as Data

Baseball Cards





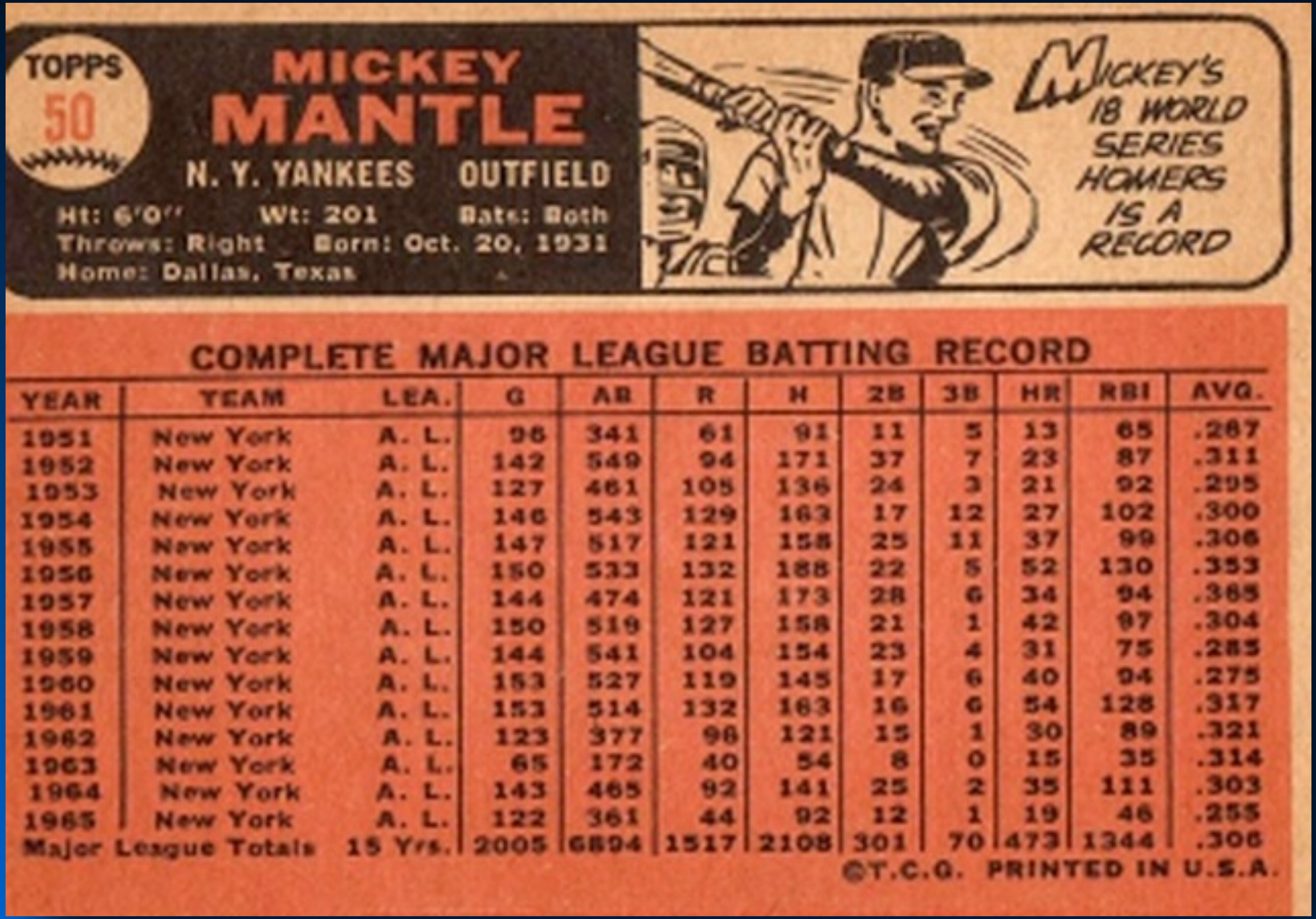
COMPLETE MAJOR LEAGUE BATTING RECORD

YEAR	TEAM	LEA.	G	AB	R	H	2B	3B	HR	RBI	AVG.
1951	New York	A. L.	96	341	61	91	11	5	13	65	.267
1952	New York	A. L.	142	549	94	171	37	7	23	87	.311
1953	New York	A. L.	127	461	105	136	24	3	21	92	.295
1954	New York	A. L.	146	543	129	163	17	12	27	102	.300
1955	New York	A. L.	147	517	121	158	25	11	37	99	.306
1956	New York	A. L.	150	533	132	188	22	5	52	130	.353
1957	New York	A. L.	144	474	121	173	28	6	34	94	.365
1958	New York	A. L.	150	519	127	158	21	1	42	97	.304
1959	New York	A. L.	144	541	104	154	23	4	31	75	.285
1960	New York	A. L.	153	527	119	145	17	6	40	94	.275
1961	New York	A. L.	153	514	132	163	16	6	54	128	.317
1962	New York	A. L.	123	377	96	121	15	1	30	89	.321
1963	New York	A. L.	65	172	40	54	8	0	15	35	.314
1964	New York	A. L.	143	465	92	141	25	2	35	111	.303
1965	New York	A. L.	122	361	44	92	12	1	19	46	.255
Major League Totals		15 Yrs.	2005	6894	1517	2108	301	70	473	1344	.306

©T.C.G. PRINTED IN U.S.A.

Collectable memorabilia of the players of various American baseball teams, usually printed on card stock, showing an image of the player on one side and their summary performance statistics on the other.

Statistics as Data


TOPPS 50

MICKEY MANTLE

N. Y. YANKEES OUTFIELD

Ht: 6'0" Wt: 201 Bats: Both

Throws: Right Born: Oct. 20, 1931

Home: Dallas, Texas

COMPLETE MAJOR LEAGUE BATTING RECORD

YEAR	TEAM	LEA.	G	AB	R	H	2B	3B	HR	RBI	AVG.
1951	New York	A. L.	96	341	61	91	11	5	13	65	.267
1952	New York	A. L.	142	549	94	171	37	7	23	87	.311
1953	New York	A. L.	127	461	105	136	24	3	21	92	.295
1954	New York	A. L.	146	543	129	163	17	12	27	102	.300
1955	New York	A. L.	147	517	121	158	25	11	37	99	.306
1956	New York	A. L.	150	533	132	188	22	5	52	130	.353
1957	New York	A. L.	144	474	121	173	28	6	34	94	.365
1958	New York	A. L.	150	519	127	158	21	1	42	97	.304
1959	New York	A. L.	144	541	104	154	23	4	31	75	.285
1960	New York	A. L.	153	527	119	145	17	6	40	94	.275
1961	New York	A. L.	153	514	132	163	16	6	54	128	.317
1962	New York	A. L.	123	377	96	121	15	1	30	89	.321
1963	New York	A. L.	65	172	40	54	8	0	15	35	.314
1964	New York	A. L.	143	465	92	141	25	2	35	111	.303
1965	New York	A. L.	122	361	44	92	12	1	19	46	.255
Major League Totals		15 Yrs.	2005	6894	1517	2108	301	70	473	1344	.306

©T.C.G. PRINTED IN U.S.A.

Variable - a characteristic that may differ from one individual to another.

Datum - the qualitative or quantitative value of a variable (*pl.* data).

Statistics as Data

Data Types: Discrete Data

- Attributes (Nominal Data) - names of mutually exclusive groups of objects.
  - Usually alpha-numeric names
    - Example: short, long, medium, medium, short, short
- Ranked Variables (Ordinal Data) - codes that can be ranked via reference to an external scale (e.g., 0 = small, 1 = medium, 2 = long).
  - Usually integers
    - Example: 0, 2, 1, 1, 0, 0

Statistics as Data

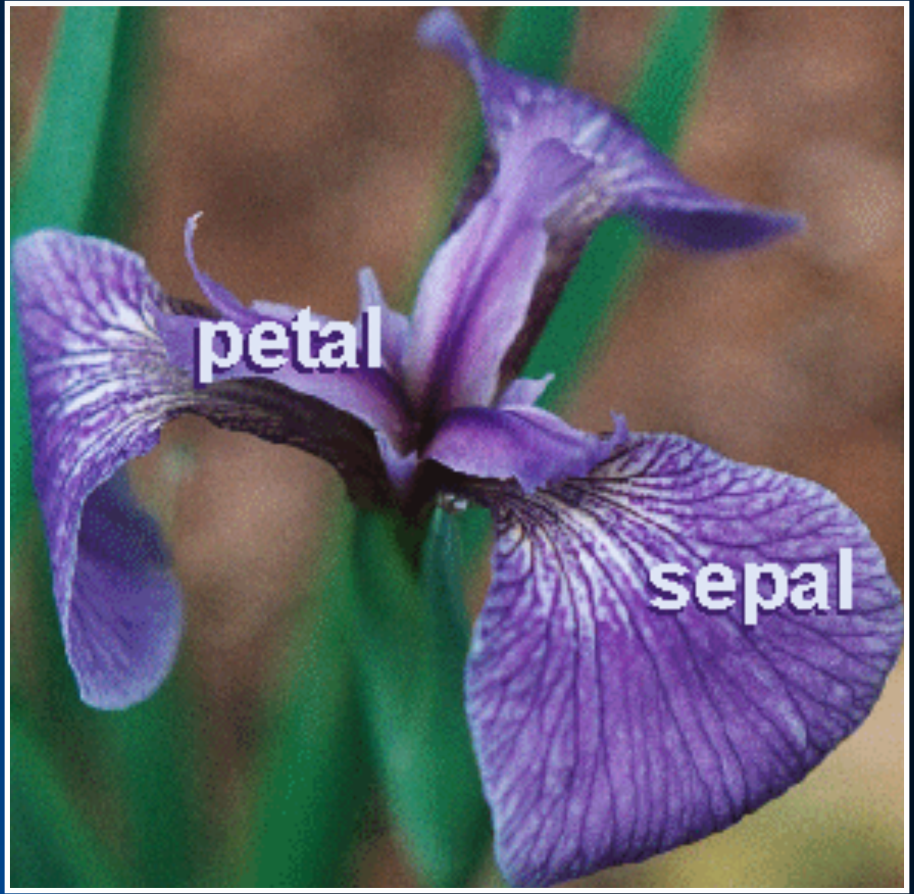
Data Types: Continuous Data

- Measurement Variables (Interval Data) - values that reference a consistent numerical scale (equivalent steps in magnitude), but one with an arbitrary origin (e.g., °F, °C °K).
  - Usually real numbers
    - Examples: -4.7°C, 49.39°F, 2.8°C, 22.3°F, 273.5°K
- Measurement Variables (Ratio Data) - values that reference a consistent numerical scale with a true zero point.
  - Real numbers
    - Examples: 4.14 mm, 2.24 mm, 1.75 mm, 3.03 mm
- Mixed-mode Data - Any combination of names, integers, and real numbers.

Statistics as Data

An Example: The Fisher *Iris* Data

Edgar S. Anderson
(1897-1969)



R. A. Fisher
(1890-1962)

Iris setosa

Iris versicolor

Iris virginica

Fisher *Iris* Data

Iris setosa

Iris versicolor

Iris virginica

Iris setosa

	Petal		Sepal	
n	Leng	Widt	Leng	Widt
1	1.4	0.2	5.1	3.5
2	1.4	0.2	4.9	3.0
3	1.3	0.2	4.7	3.2
4	1.5	0.2	4.6	3.1
5	1.4	0.2	5.0	3.6
6	1.7	0.4	5.4	3.9
7	1.4	0.3	4.6	3.4
8	1.5	0.2	5.0	3.4
9	1.4	0.2	4.4	2.9
10	1.5	0.1	4.9	3.1
11	1.5	0.2	5.4	3.7
12	1.6	0.2	4.8	3.4
13	1.4	0.1	4.8	3.0
14	1.1	0.1	4.3	3.0

Iris versicolor

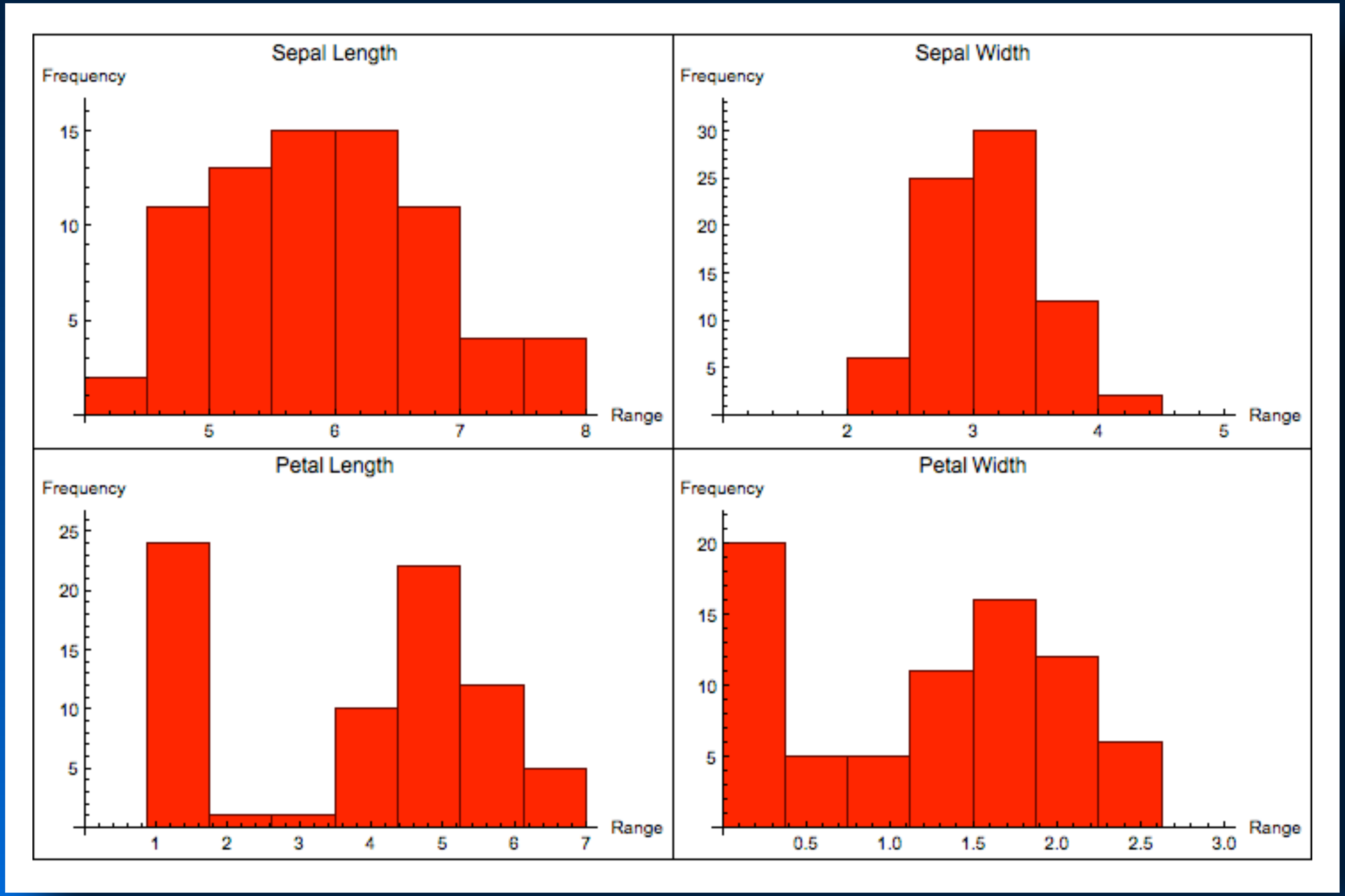
Petal		Sepal	
Leng	Widt	Leng	Widt
4.7	1.4	7.0	3.2
4.5	1.5	6.4	3.2
4.9	1.5	6.9	3.1
4.0	1.3	5.5	2.3
4.6	1.5	6.5	2.8
4.5	1.3	5.7	2.8
4.7	1.6	6.3	3.3
3.3	1.0	4.9	2.4
4.6	1.3	6.6	2.9
3.9	1.4	5.2	2.7
3.5	1.0	5.0	2.0
4.2	1.5	5.9	3.0
4.0	1.0	6.0	2.2

Iris virginica

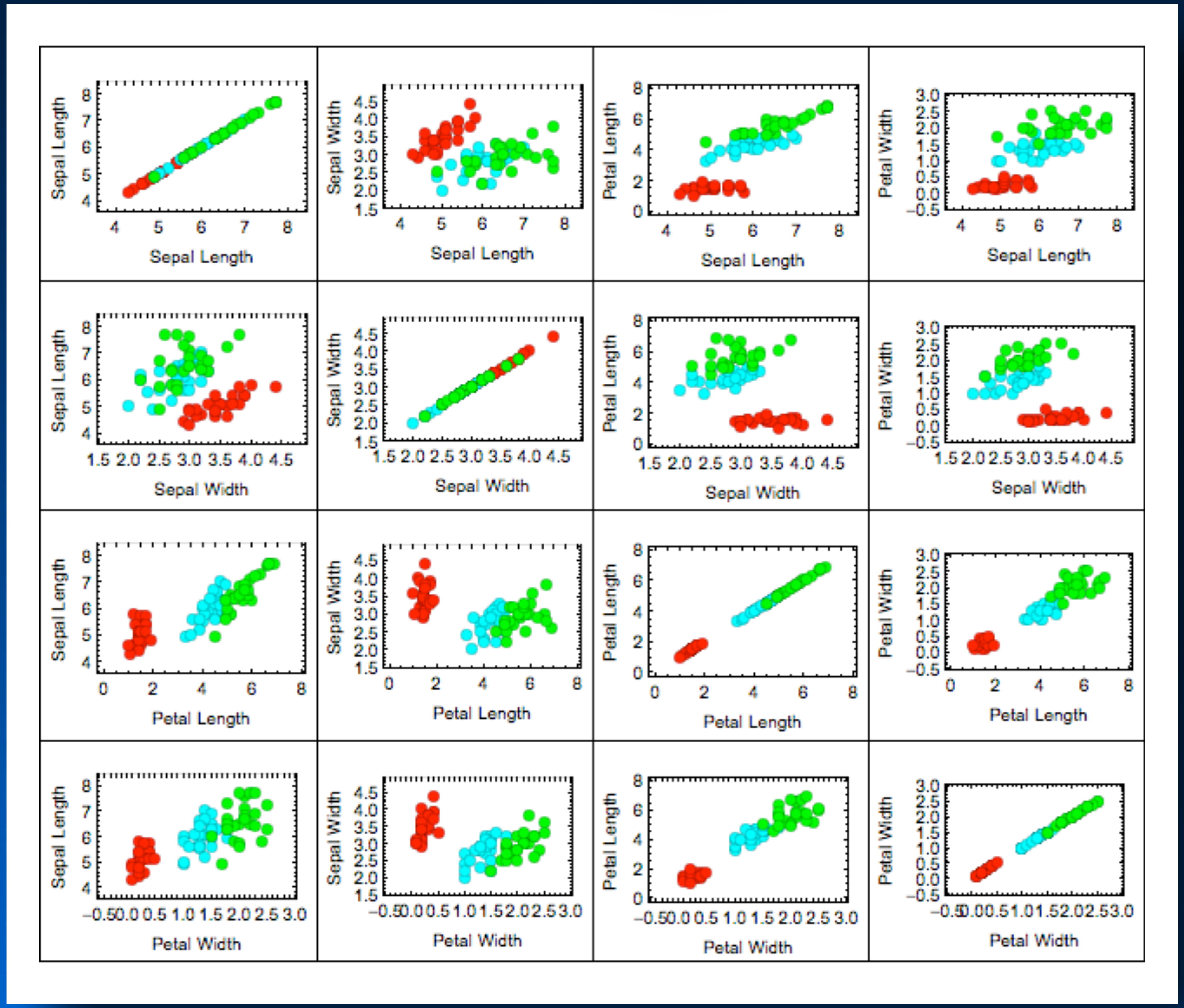
Petal		Sepal	
Leng	Widt	Leng	Widt
6.0	2.5	6.3	3.3
5.1	1.9	5.8	2.7
5.9	2.1	7.1	3.0
5.6	1.8	6.3	2.9
5.8	2.2	6.5	3.0
6.6	2.1	7.6	3.0
4.5	1.7	4.9	2.5
6.3	1.8	7.3	2.9
5.8	1.8	6.7	2.5
6.1	2.5	7.2	3.6
5.1	2.0	6.5	3.2
5.3	1.9	6.4	2.7
5.5	2.1	6.8	3.0

Fisher *Iris* Data

Frequency Histograms



Scatterplots



Statistics as Data Analysis

Describing a Collection

Statistics as Data Analysis

Describing a Collection

- Range - difference between highest and lowest observation.
- Median - middle number of an ordered set of data.
- Mode - the most frequent observation in a set of data.
- Mean - family of indices relating the sums or products of a dataset to the number of data values
  - Arithmetic Mean - the centroid of the distribution.
  - Geometric Mean - the n^{th} root of the product of n observations.
  - Harmonic Mean - reciprocal of the arithmetic mean of the set of reciprocals of a set of observations.
- Variance - mean of squared deviations from the mean.
- Standard Deviation - mean of deviations from the mean.
- Coefficient of Variation - ratio of std. deviation to the mean.
- Percentile - the value of a variable below which a certain percent of observations fall.

Statistics as Data Analysis

Locating Data

Range - difference between highest and lowest observation.

Raw Data

-1.145, 1.887, 2.270, 1.242, 0.825, 0.498, -1.600, 4.124, -0.083, -1.044

Ordered Data

-1.600, -1.145, -1.044, -0.083, 0.498, 0.825, 1.242, 1.887, 2.270, 4.124

Range

4.124 - -1.600 = 5.724

Statistics as Data Analysis

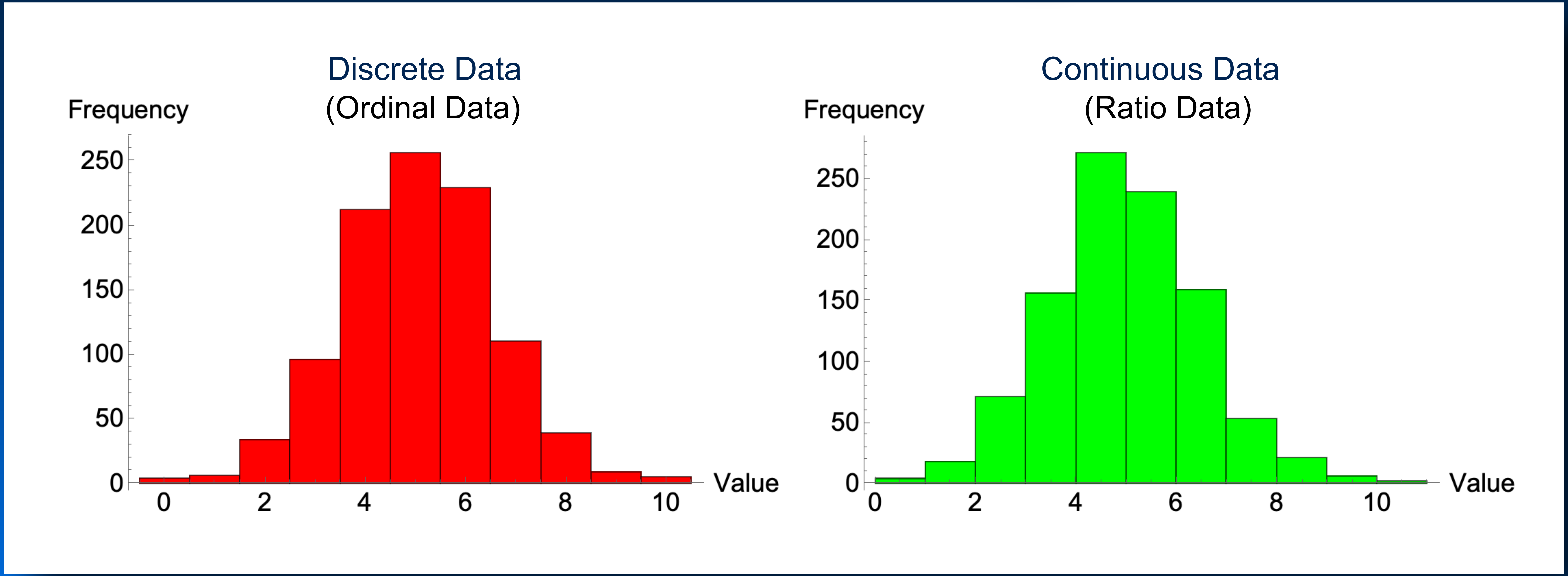
Describing a Collection

- Range - difference between highest and lowest observation.
- Median - middle number of an ordered set of data.
- Mode - the most frequent observation in a set of data.
- Mean - family of indices relating the sums or products of a dataset to the number of data values
  - Arithmetic Mean - the centroid of the distribution.
  - Geometric Mean - the n^{th} root of the product of n observations.
  - Harmonic Mean - reciprocal of the arithmetic mean of the set of reciprocals of a set of observations.
- Variance - mean of squared deviations from the mean.
- Standard Deviation - mean of deviations from the mean.
- Coefficient of Variation - ratio of std. deviation to the mean.
- Percentile - the value of a variable below which a certain percent of observations fall.

Statistics as Data Analysis

Measures of Central Tendency

Two Datasets of Different Types

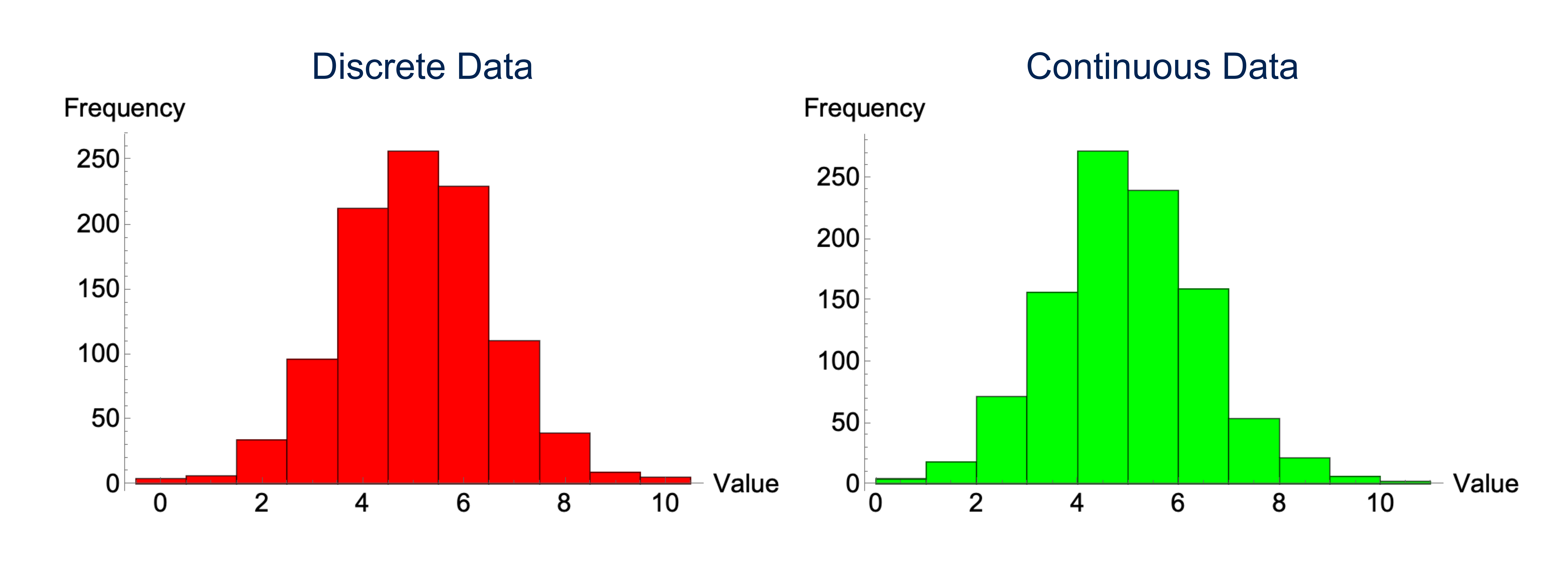


$N = 1{,}000$

Statistics as Data Analysis

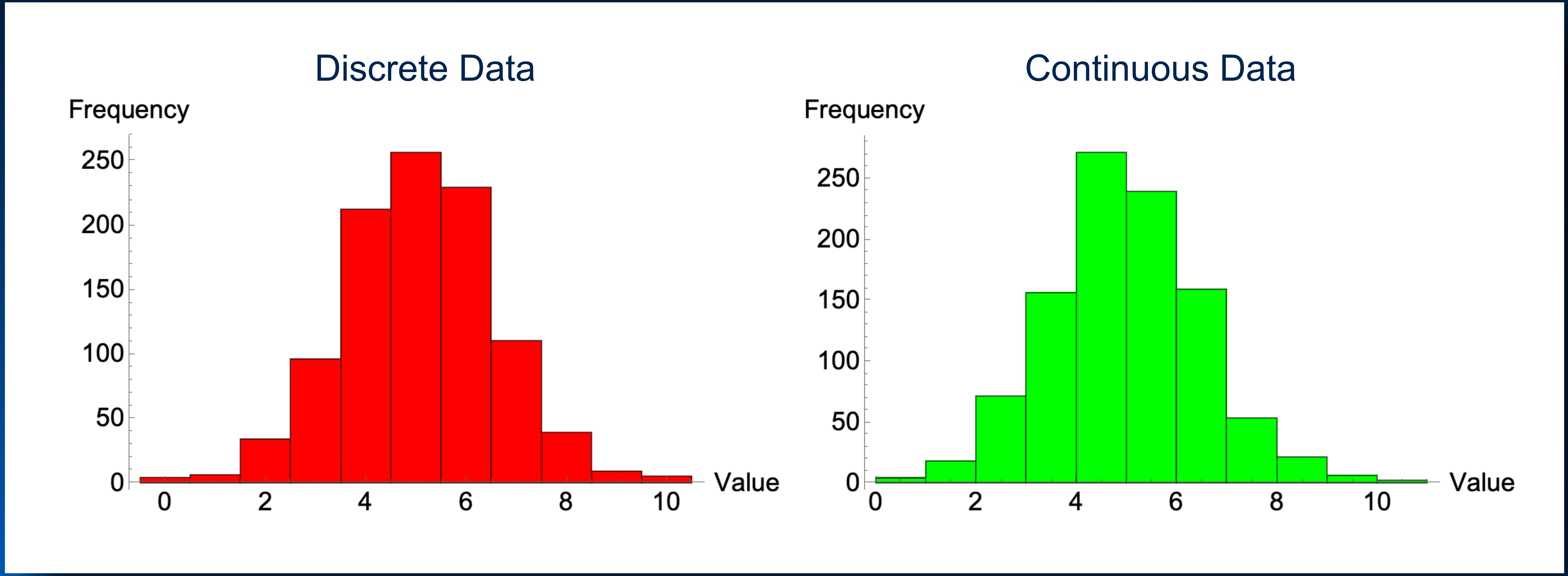
Measures of Central Tendency

- Mean = usually $\Sigma x_i/n$ (arithmetic mean).
- Median = lies at the (*n*+1)/2 position when *n* is odd and at the mean of the *n*/2 an (*n*/2)+1 positions when *n* is even.
- Mode = category or class with the highest *n*.



Statistics as Data Analysis

Measures of Central Tendency

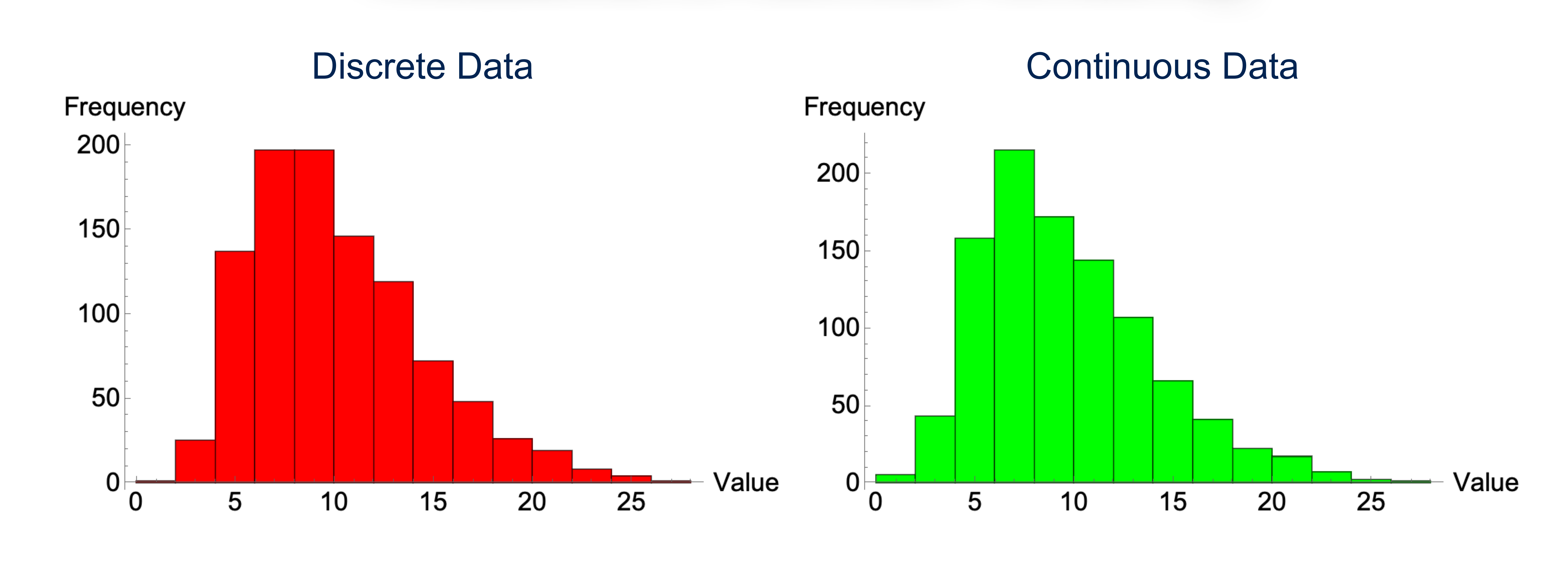


Mean	5
Median	5
Mode	5

Mean	4.7
Median	4.9
Mode	4.9

Statistics as Data Analysis

Measures of Central Tendency



Mode	6 - 7
Median	9
Mean	9

Mode	5.83 – 7.50
Median	8.9
Mean	9.7

Statistics as Data Analysis

Measures of Central Tendency

- Mode - used primarily to describe nominal and/or ordinal data. When used for continuous data its value is tied to the size of the bins used to aggregate the data.
- Median - Can only be used to describe data that can be arranged in rank order (e.g., ordinal, interval, ratio data).
- Mean - Assumes equal spacing between adjacent values. Can be used to describe ordinal, interval, ratio data, but when used on ordinal data it usually produces a non-discrete (= impossible) result.

None of these measures of central tendency can be applied usefully to datasets that include a variety of different data types (mixed-mode data).

Statistics as Data Analysis

Describing a Collection

- Range - difference between highest and lowest observation.
- Median - middle number of an ordered set of data.
- Mode - the most frequent observation in a set of data.
- Mean - family of indices relating the sums or products of a dataset to the number of data values
  - Arithmetic Mean - the centroid of the distribution.
  - Geometric Mean - the n^{th} root of the product of n observations.
  - Harmonic Mean - reciprocal of the arithmetic mean of the set of reciprocals of a set of observations.
- Variance - mean of squared deviations from the mean.
- Standard Deviation - mean of deviations from the mean.
- Coefficient of Variation - ratio of std. deviation to the mean.
- Percentile - the value of a variable below which a certain percent of observations fall.

Statistics as Data Analysis

The Arithmetic Mean

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Data

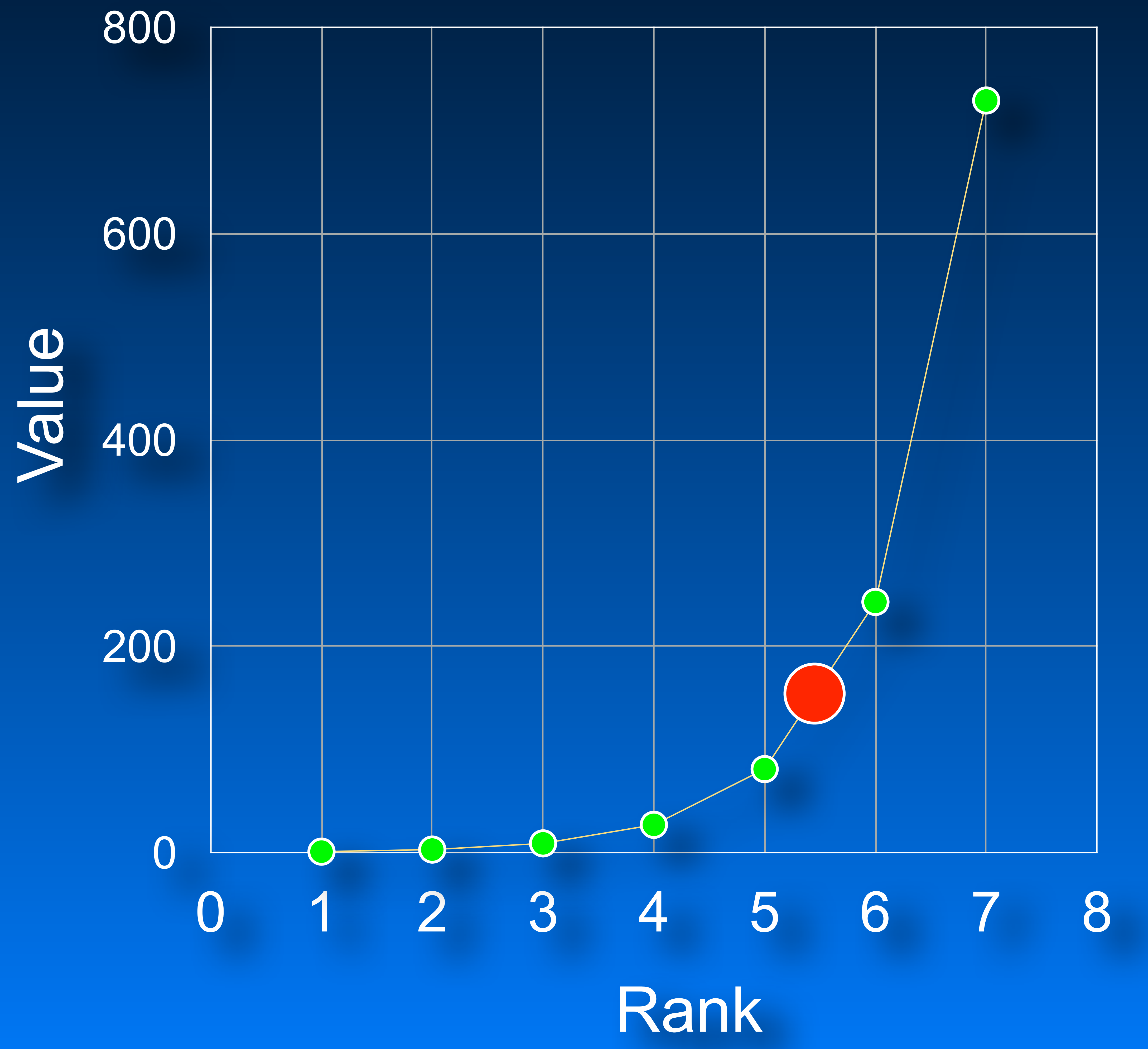
x	y
1	1
2	4
3	7
4	10
5	13
6	16
7	19

Evaluation

$$\bar{y} = \frac{70}{7} = 10$$

Statistics as Data Analysis

The Arithmetic Mean



$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Data

x	y
1	1
2	4
3	7
4	10
5	13
6	16
7	19

Evaluation

$$\bar{y} = \frac{1093}{7} = 156.14$$

Statistics as Data Analysis

The Geometric Mean

$$\bar{y} = \sqrt[n]{\Pi_{i=1}^{n} y_i}$$

Data

x	y
1	1
2	4
3	7
4	10
5	13
6	16
7	19

Evaluation

$$\bar{y} = \sqrt[7]{10,460,353,203}$$

$$\bar{y} = 27$$

Statistics as Data Analysis

What about rates?

Time (myr)	Interval	Length (mm)	Rate (mm/myr)
0	-	100	-
10	10	200	10
20	10	400	20

Which mean should I use to calculate average rates?

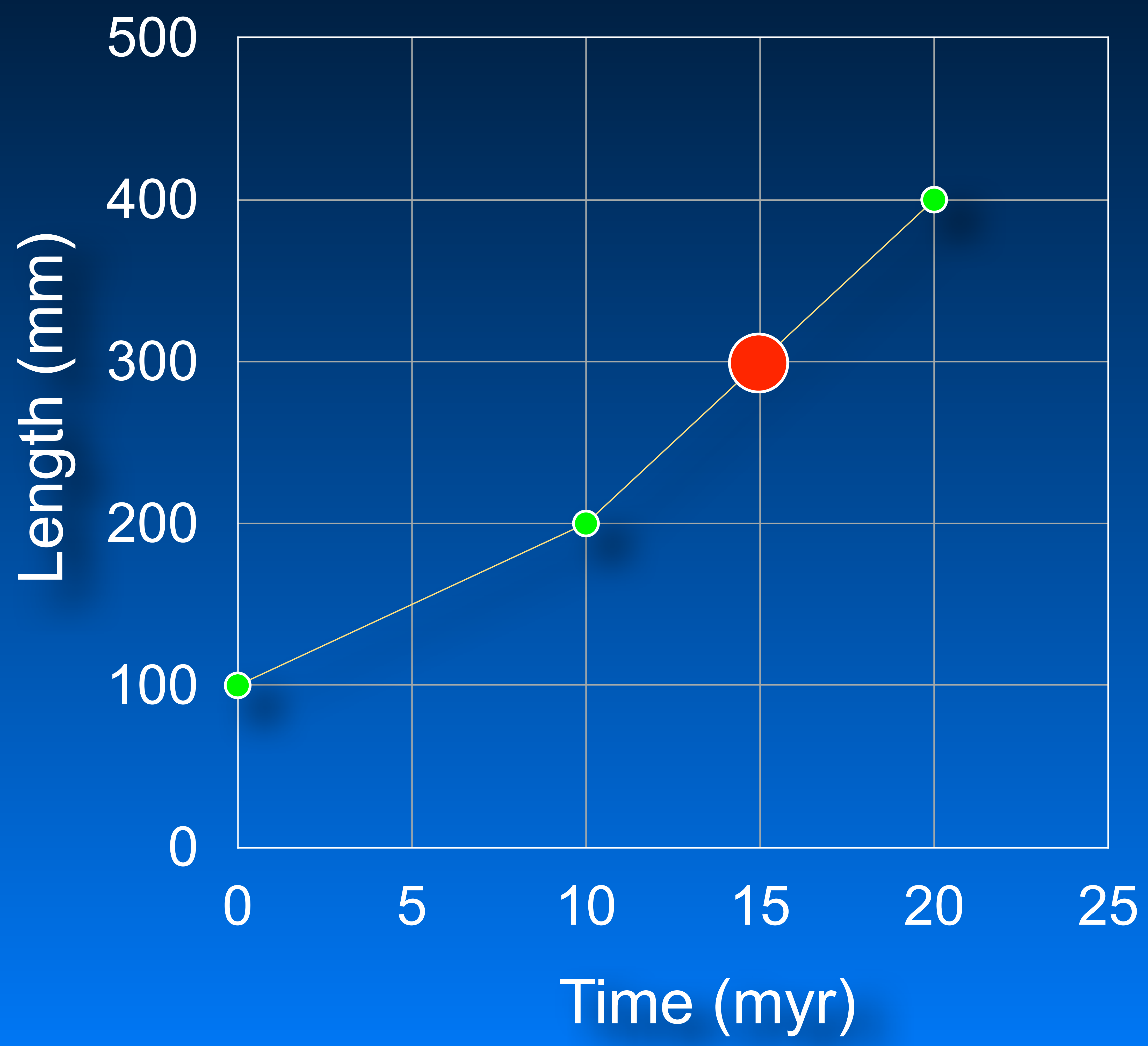
t = 10 myr

t = 0.0

t = 20 myr

Statistics as Data Analysis

The Geometric Mean



$$\bar{y} = \sqrt[n]{\Pi_{i=1}^{n} y_i}$$

Data

Time	Rate	Rate
0	100	-
10	200	10
20	400	20

Evaluation

$$\bar{y} = \frac{(10 + 20)}{2} = 15$$

Statistics as Data Analysis

What about rates?

Time (myr)	Interval	Length (mm)	Rate (mm/myr)
0	-	100	-
10	10	200	10
20	10	400	20

However, because the net change from *t* = 0 was unequal over the two time intervals, the average cannot be the midpoint.

It should be closer to 10 mm/myr.

t = 10 myr

t = 0.0

t = 20 myr

Statistics as Data Analysis

What about rates?

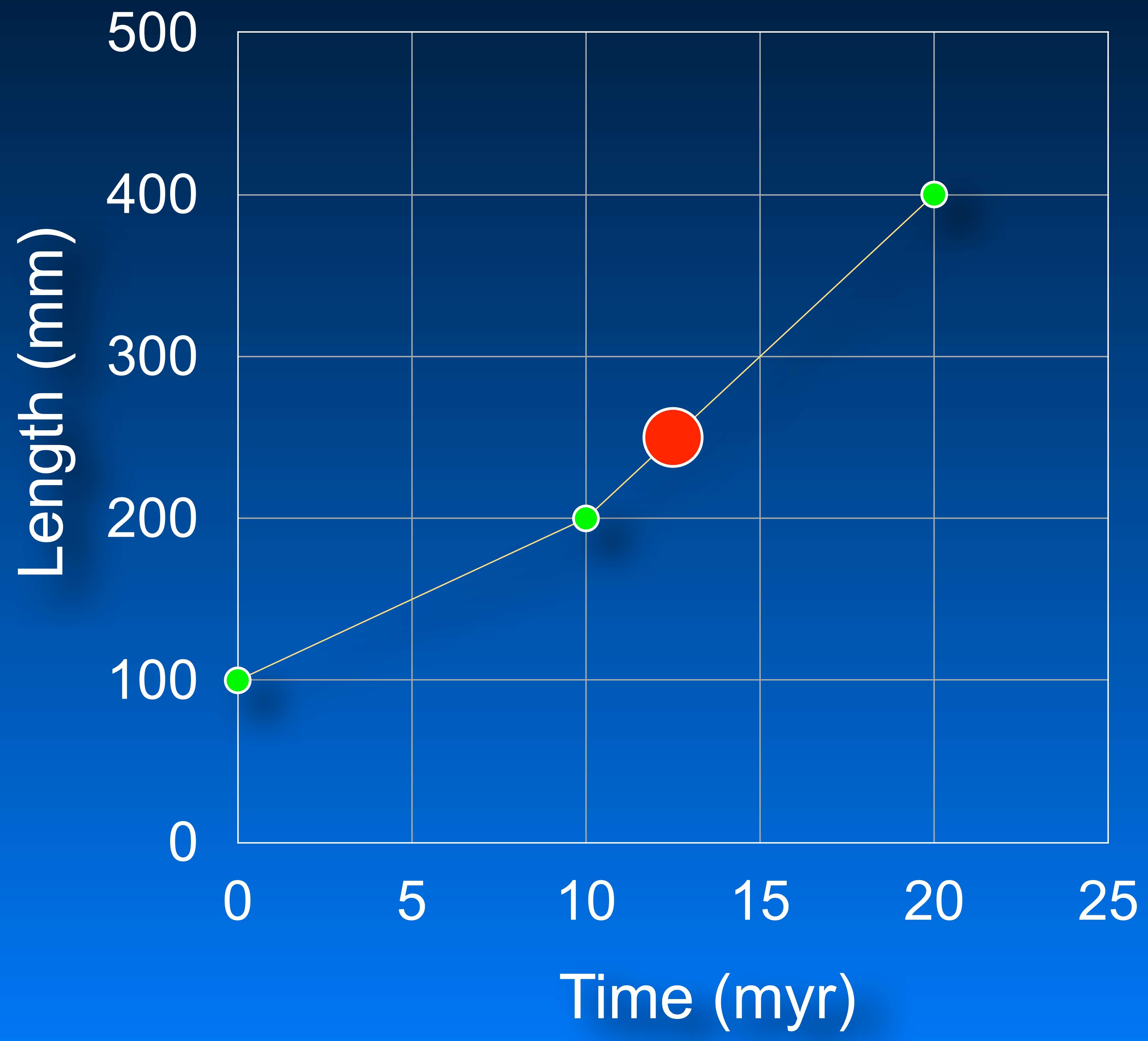
Time (myr)	Length (mm)	Rate (mm/myr)	Proportion of size change	Ratio
0	100	-	-	-
10	200	10	100/300	0.33
10	400	20	200/300	0.67

Harmonic (“Unweighted” Arithmetic) Mean

$$\bar{y} = (10 \times 0.67) + (20 \times 0.33)$$
$$\bar{y} = (6.7) + (6.6)$$
$$\bar{y} = 13.3$$

Statistics as Data Analysis

The Harmonic Mean



$$\bar{y} = \frac{n}{\sum_{i=1}^{n} \frac{1}{y_i}}$$

Data

Time	Rate	Rate
0	100	-
10	200	10
20	400	20

Evaluation

$$\bar{y} = \frac{2}{(\frac{1}{10} + \frac{1}{20})} = \frac{2}{(0.1 + 0.05)}$$

$$\bar{y} = \frac{2}{0.15} = 13.3$$

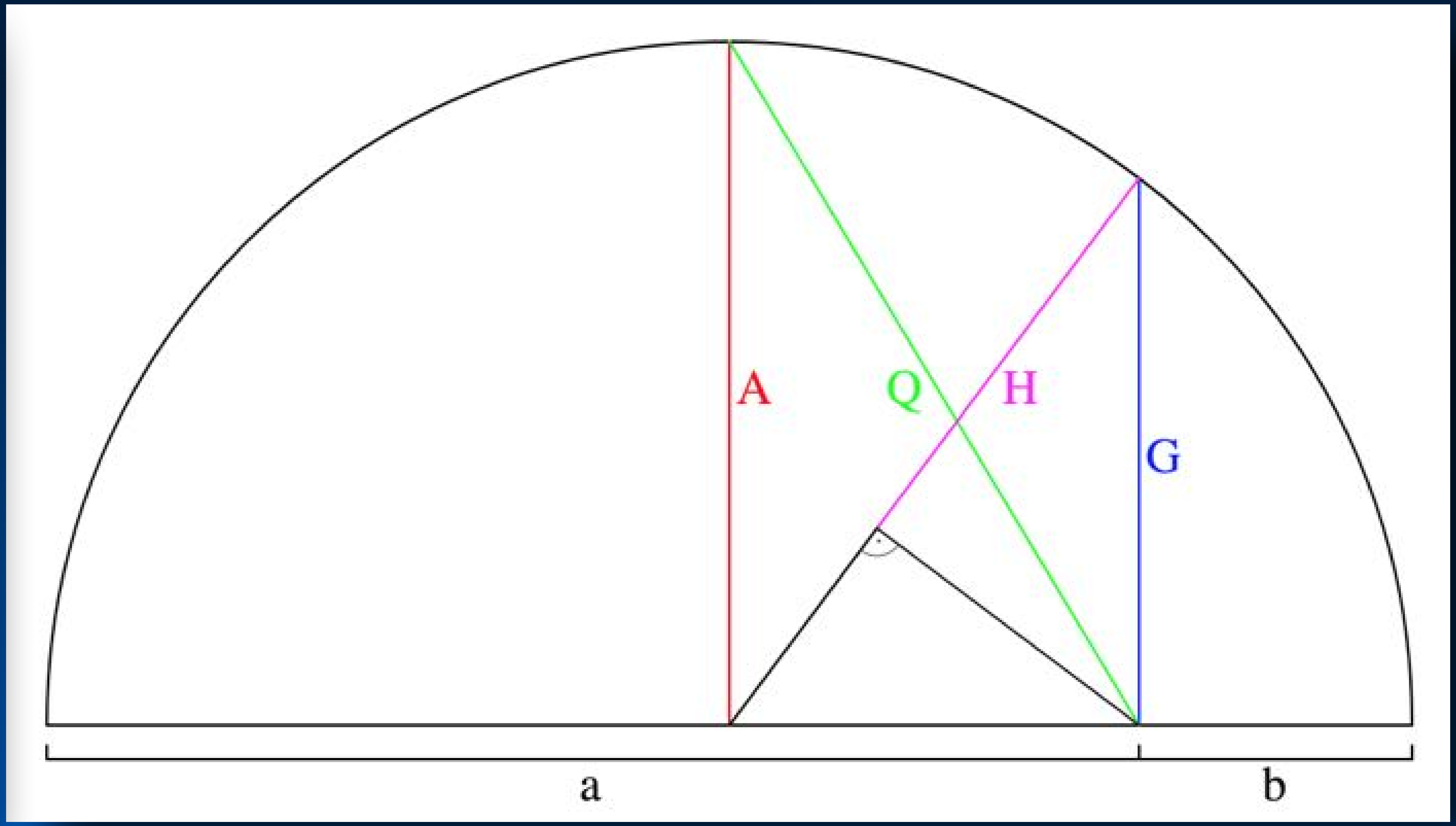
Statistics as Data Analysis

Describing Central Tendency: Summary

- Arithmetic Mean - the centroid of the distribution.
  - Use when data are linear, additive and referenced to a common scale.
- Geometric Mean - the n^{th} root of the product of n observations.
  - Use when data are non-linear, multiplicative and/or when variables have different scales.
- Harmonic Mean - reciprocal of the arithmetic mean of the set of reciprocals of a set of observations.
  - Use when data are ratios or speeds.

Statistics as Data Analysis

Wisdom of the Ancients



The Greek mathematician & philosopher Pythagoras' diagram of the geometric relation between the arithmetic mean (A), geometric mean (G), harmonic mean (H), and quadratic mean (Q) of two numbers (a & b) of unequal magnitude.

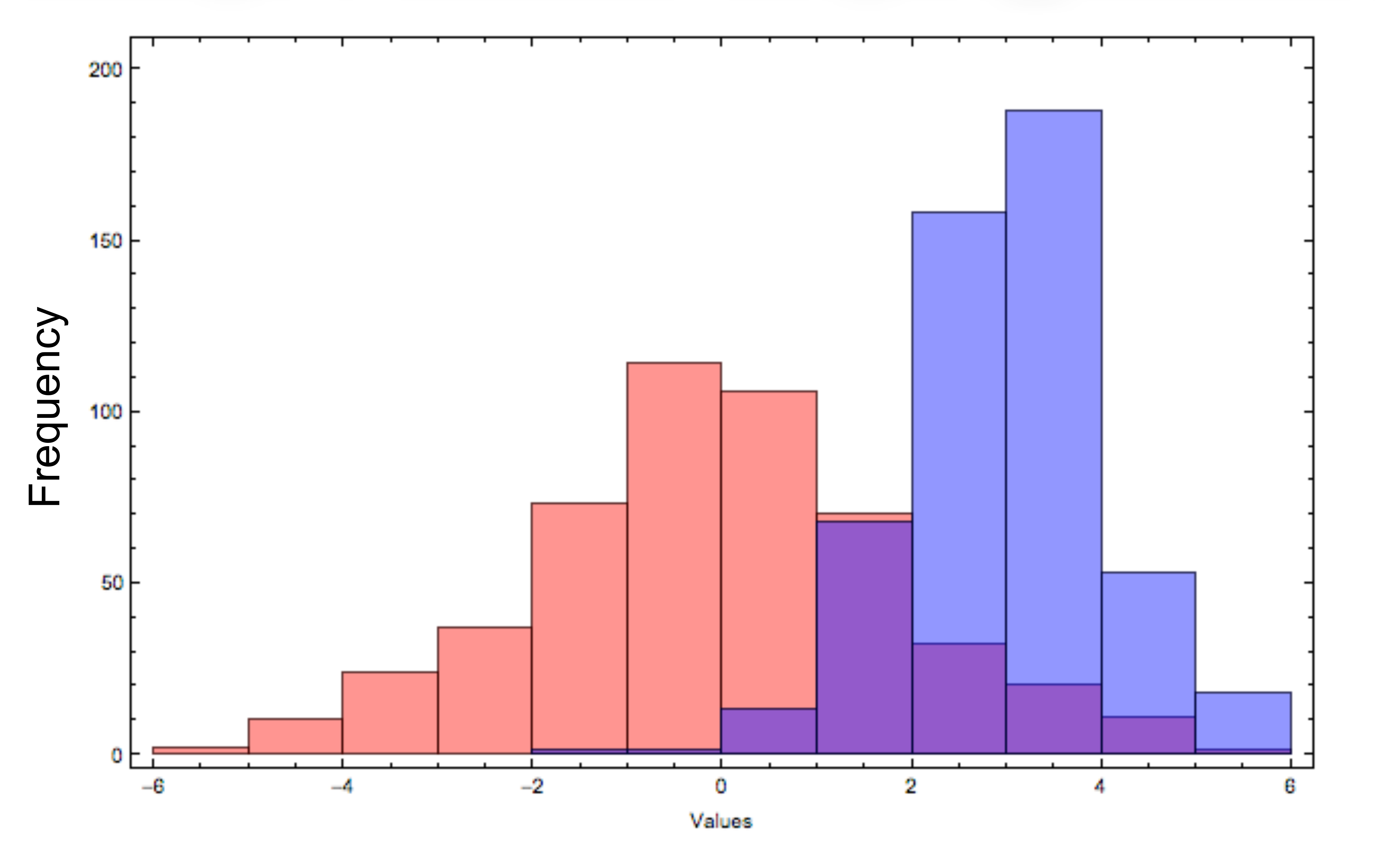
Statistics as Data Analysis

What's the point of a mean?

How can we use means to describe specimens with numbers?
And why would we want to?

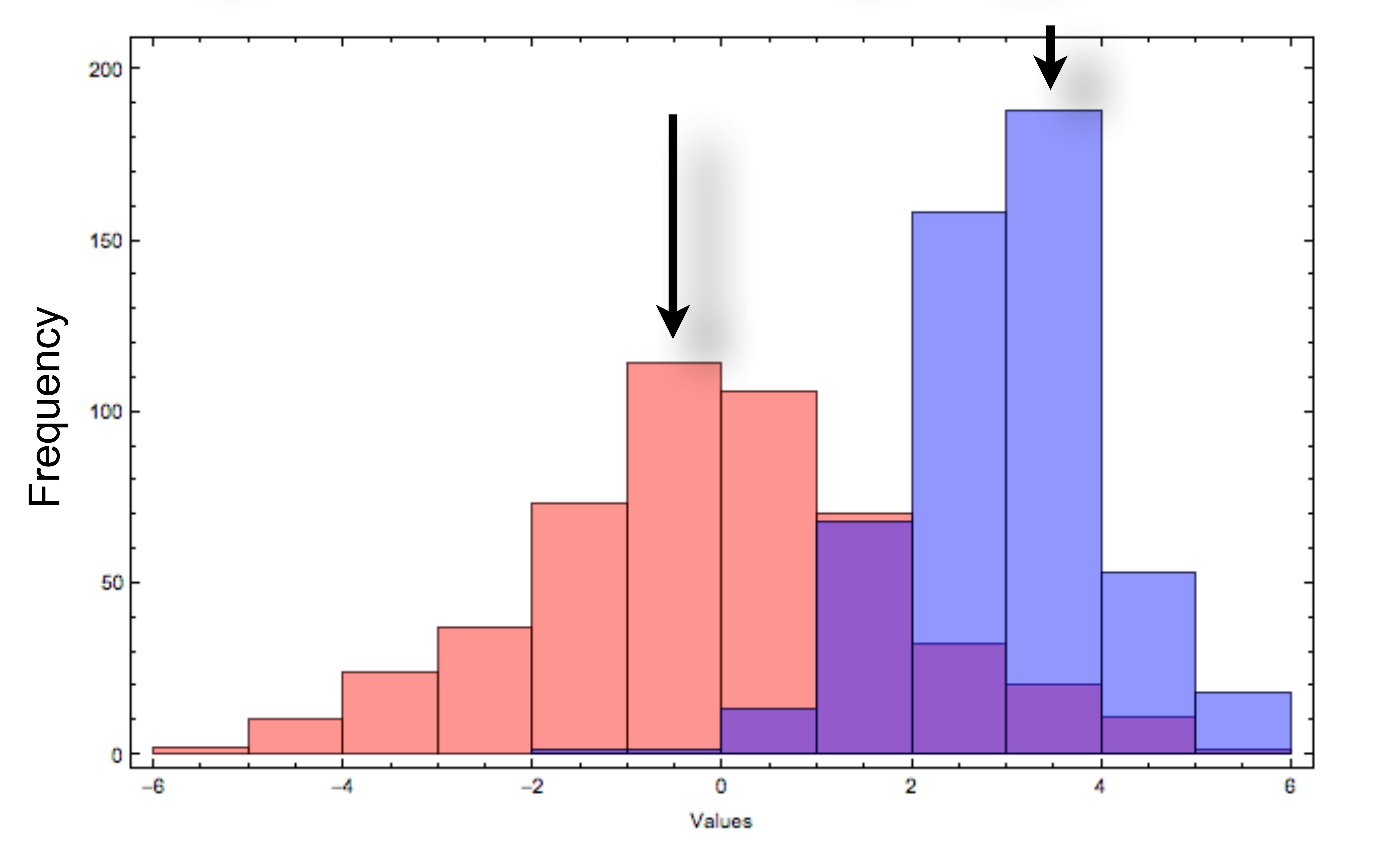
Statistics as Data Analysis

Descriptive Statistics: Comparing Collections



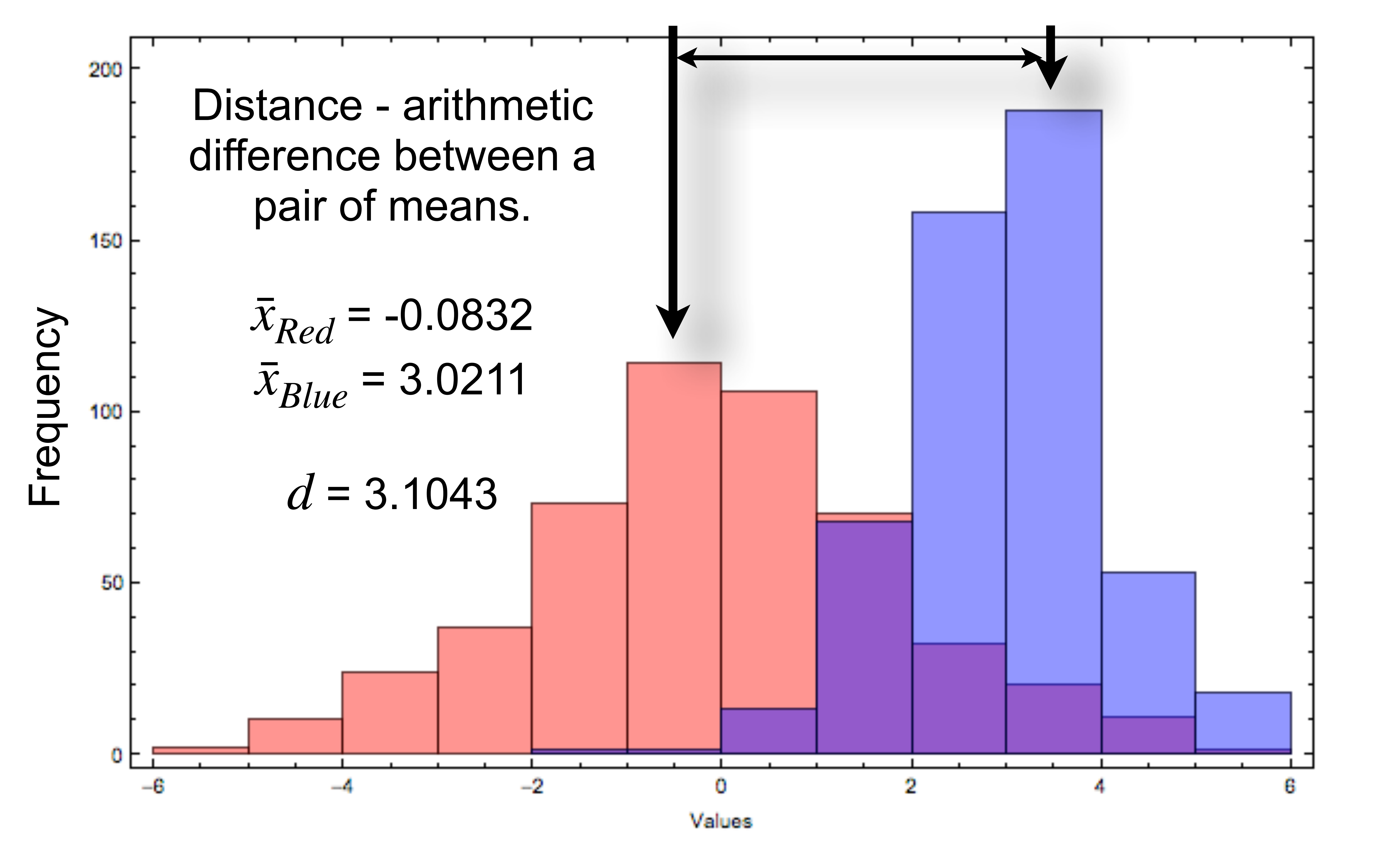
Statistics as Data Analysis

Descriptive Statistics: Comparing Collections



Statistics as Data Analysis

Descriptive Statistics: Comparing Collections



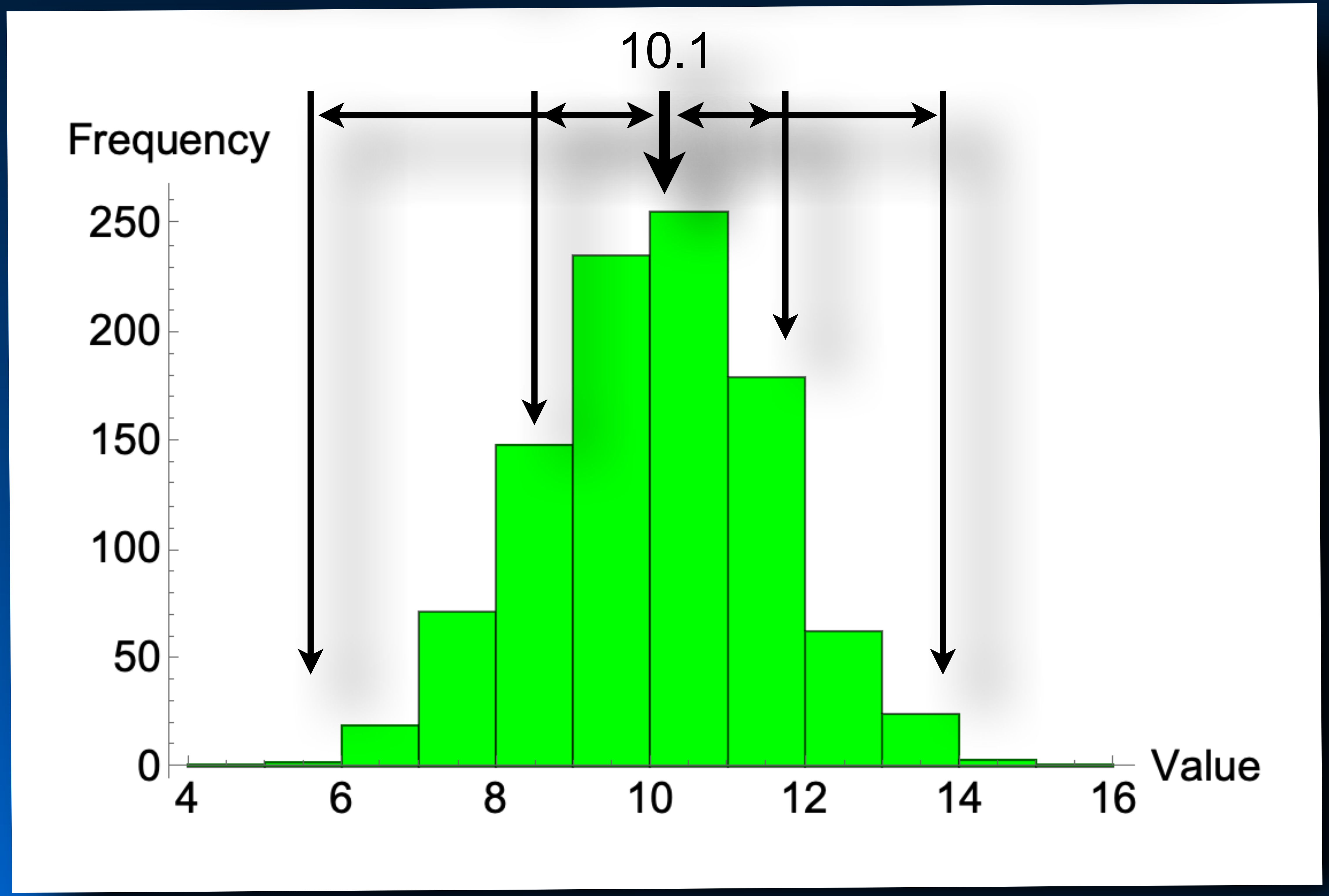
Statistics as Data Analysis

Describing a Collection

- Range - difference between highest and lowest observation.
- Median - middle number of an ordered set of data.
- Mode - the most frequent observation in a set of data.
- Mean - family of indices relating the sums or products of a dataset to the number of data values
  - Arithmetic Mean - the centroid of the distribution.
  - Geometric Mean - the n^{th} root of the product of n observations.
  - Harmonic Mean - reciprocal of the arithmetic mean of the set of reciprocals of a set of observations.
- Variance - mean of squared deviations from the mean.
- Standard Deviation - mean of deviations from the mean.
- Coefficient of Variation - ratio of std. deviation to the mean.
- Percentile - the value of a variable below which a certain percent of observations fall.

Statistics as Data Analysis

Variability About the Mean

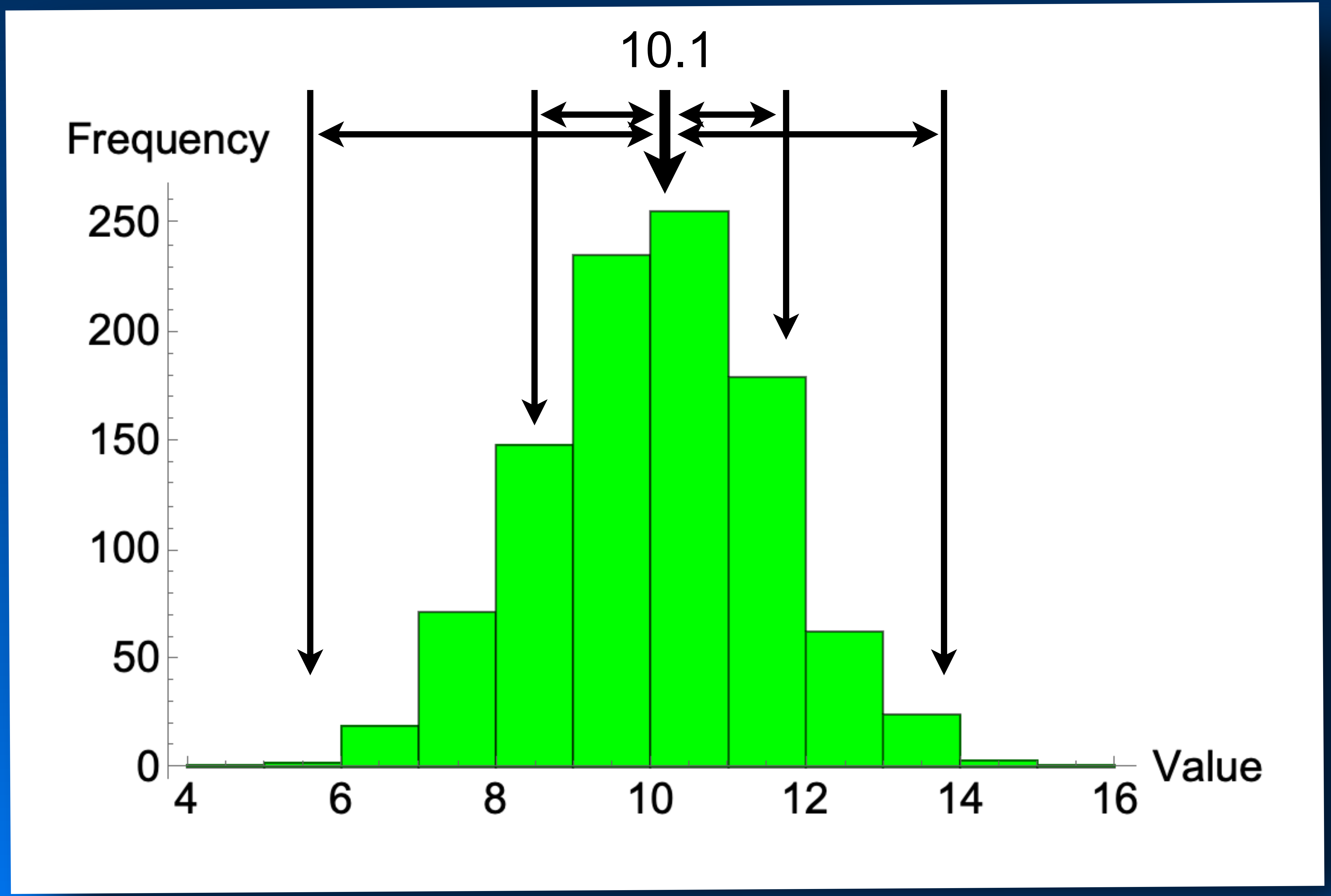


The variance is the average of the squared deviations from the arithmetic mean.

Statistics as Data Analysis

Variability About the Mean

The variance is the average of the squared deviations from the arithmetic mean.



Variance

$$s^2 = \frac{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n \cdot (n-1)}$$

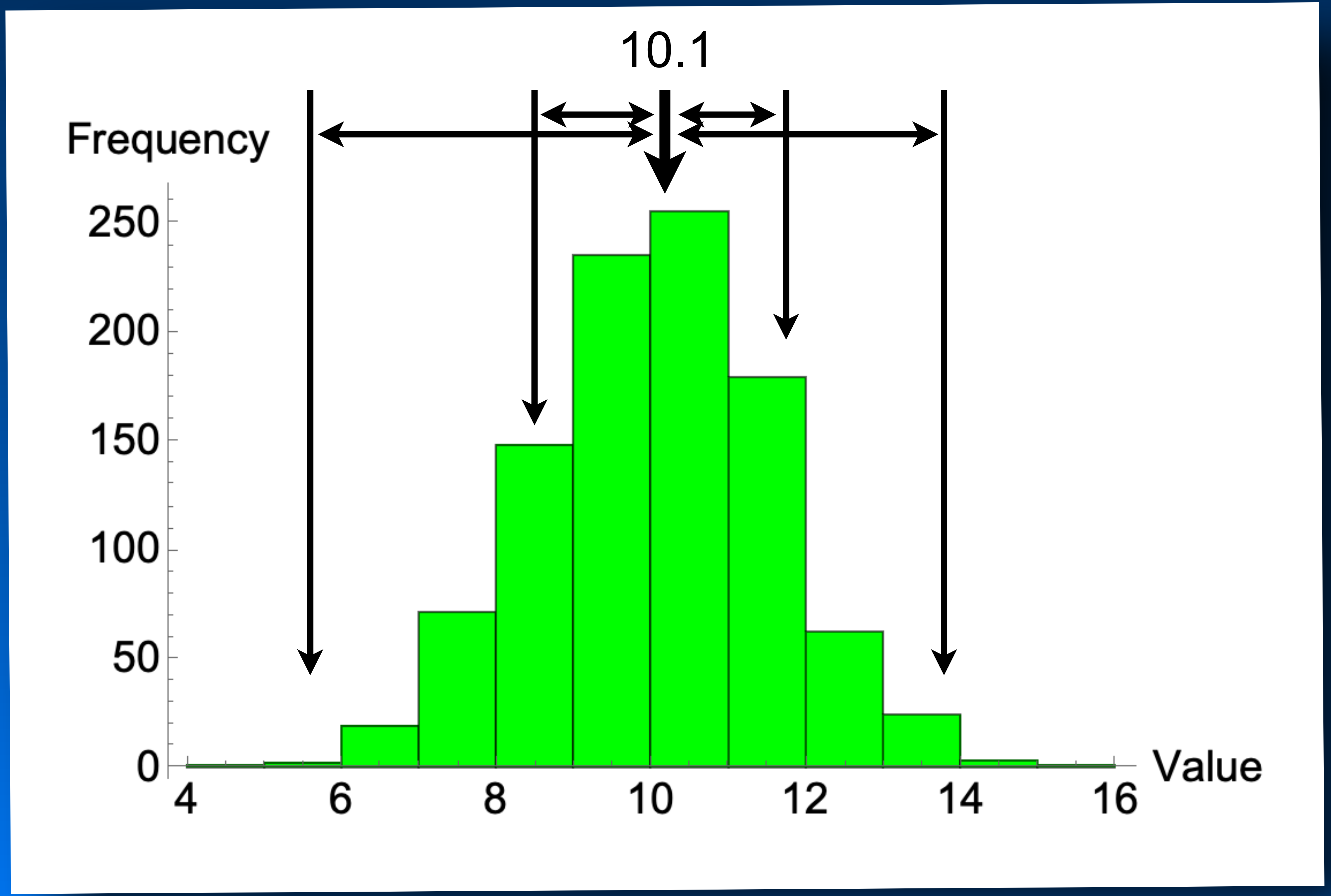
$$s^2 = 2.345$$

What are the units of the variance?

Statistics as Data Analysis

Variability About the Mean

The standard deviation is the square root of the variance.



Standard Deviation

$$s = \sqrt{\frac{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n \cdot (n-1)}}$$

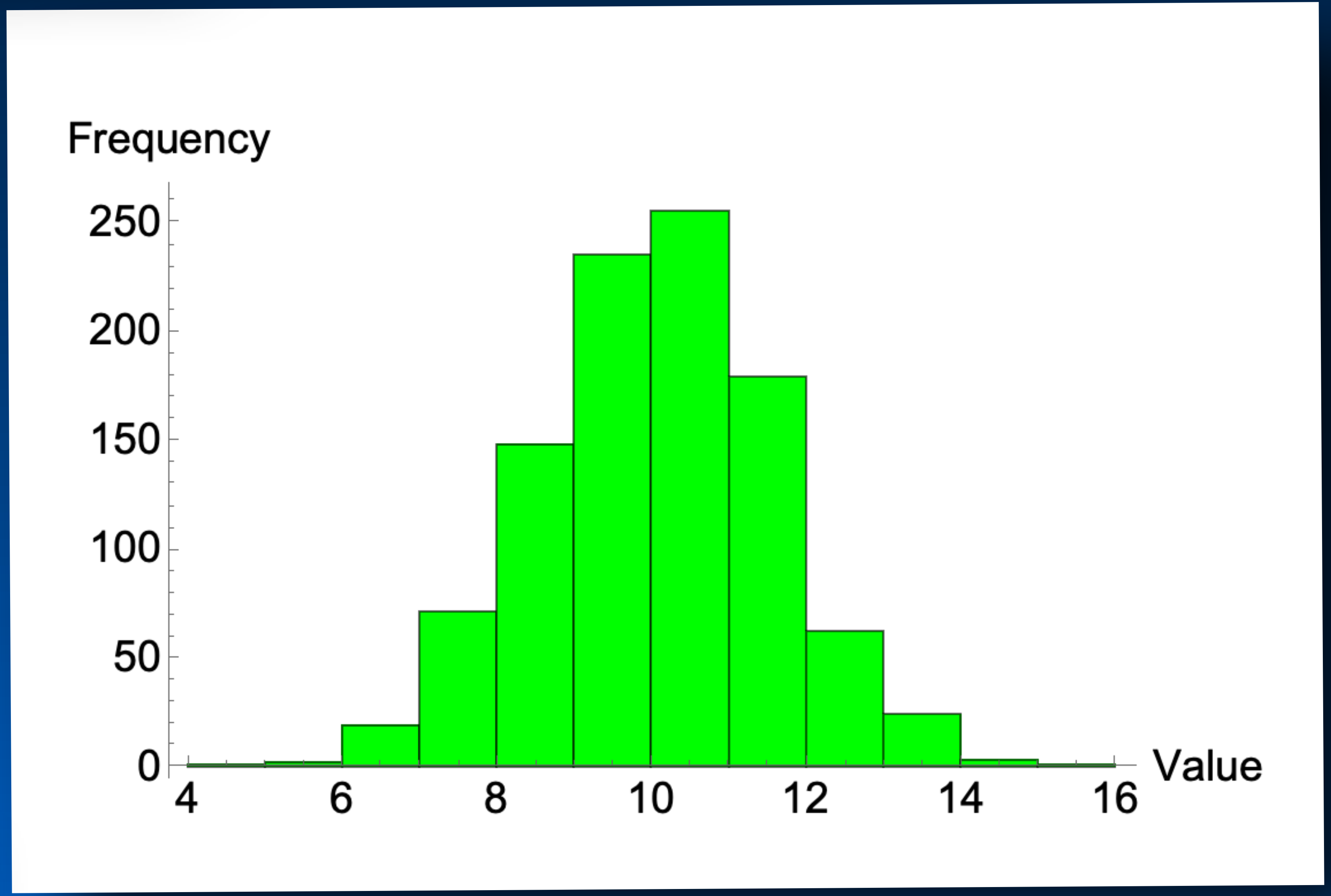
$$s = 1.531$$

What are the units of the standard deviation?

Statistics as Data Analysis

Variability About the Mean

What about assumptions?



- Variances and standard deviations are used to describe single sets of values.
- Their calculation assumes each data value is independent of all other data values in the sample or population of interest.
- If this is not true – if the data exhibits some type of structure (e.g., clustering) due to the presence of a constraint that limits variation in certain directions – the variance and standard deviation can still be estimated, but must be adjusted to compensate for the presence of the structure.

Statistics as Data Analysis

Describing a Collection

- Range - difference between highest and lowest observation.
- Median - middle number of an ordered set of data.
- Mode - the most frequent observation in a set of data.
- Mean - family of indices relating the sums or products of a dataset to the number of data values
  - Arithmetic Mean - the centroid of the distribution.
  - Geometric Mean - the n^{th} root of the product of n observations.
  - Harmonic Mean - reciprocal of the arithmetic mean of the set of reciprocals of a set of observations.
- Variance - mean of squared deviations from the mean.
- Standard Deviation - mean of deviations from the mean.
- Coefficient of Variation - ratio of std. deviation to the mean.
- Percentile - the value of a variable below which a certain percent of observations fall.

Statistics as Data Analysis

Measures of Variation

Coefficient of Variation

$$c_v = \frac{\sigma}{\mu}$$

Because the variance and standard variation take the mean as their points of reference, variables that have a high mean (μ or $\bar{x}$) will always have a tendency to return higher variance (σ^2 or s^2) and standard deviation (σ or s) values than those with low means. Therefore, тhe mean-standardized standard deviation – also called the coefficient of variation – is usually employed when comparisons between different variables, or the same variable whose values range over different intervals, are needed.

Care must be taken to use the coefficient of variation only to com-pare ratio variables with meaningful zero points.

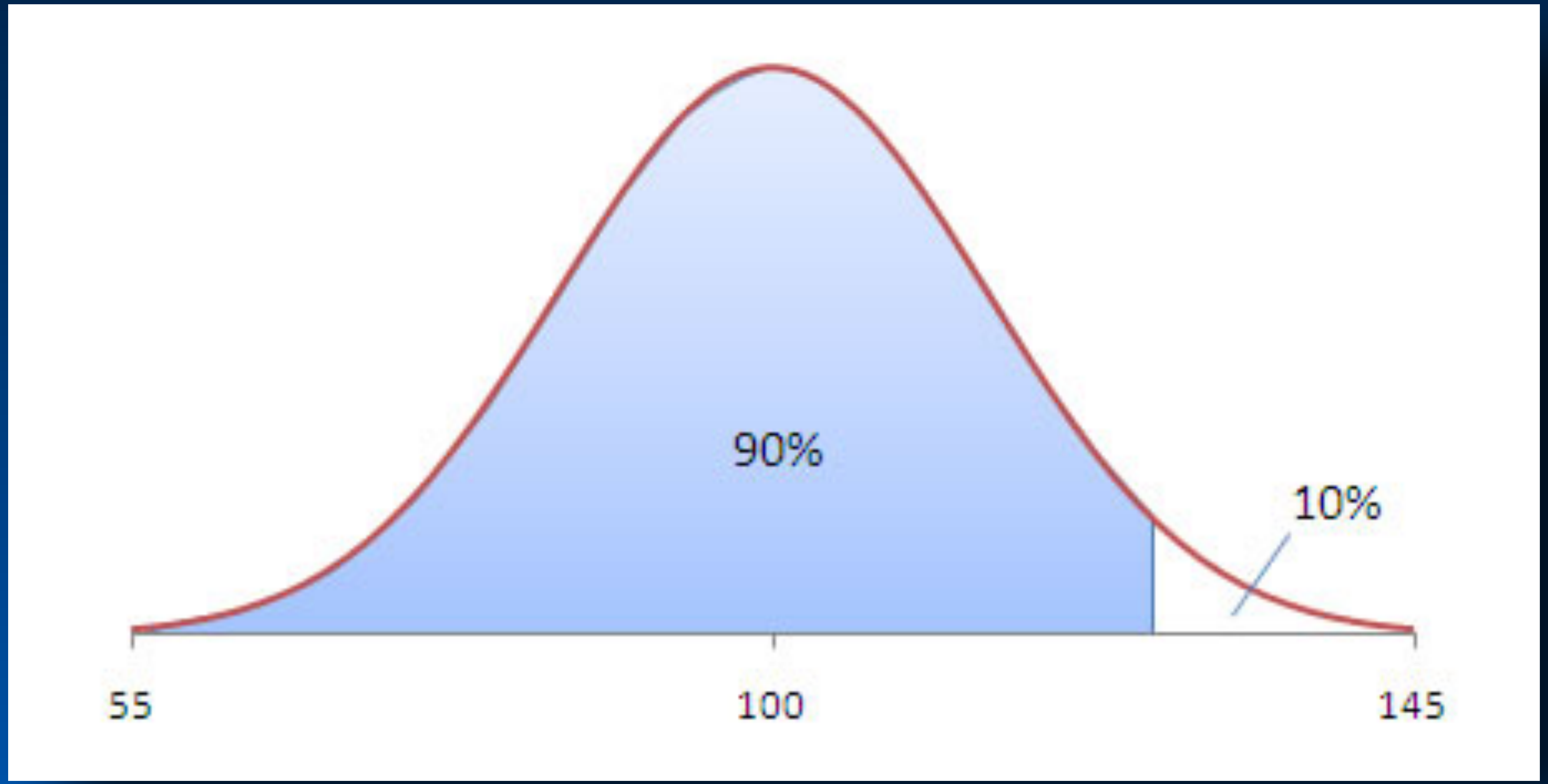
Statistics as Data Analysis

Measures of Variation

Percentile

The variable value below which a given percentage of the population or sample observations lie.

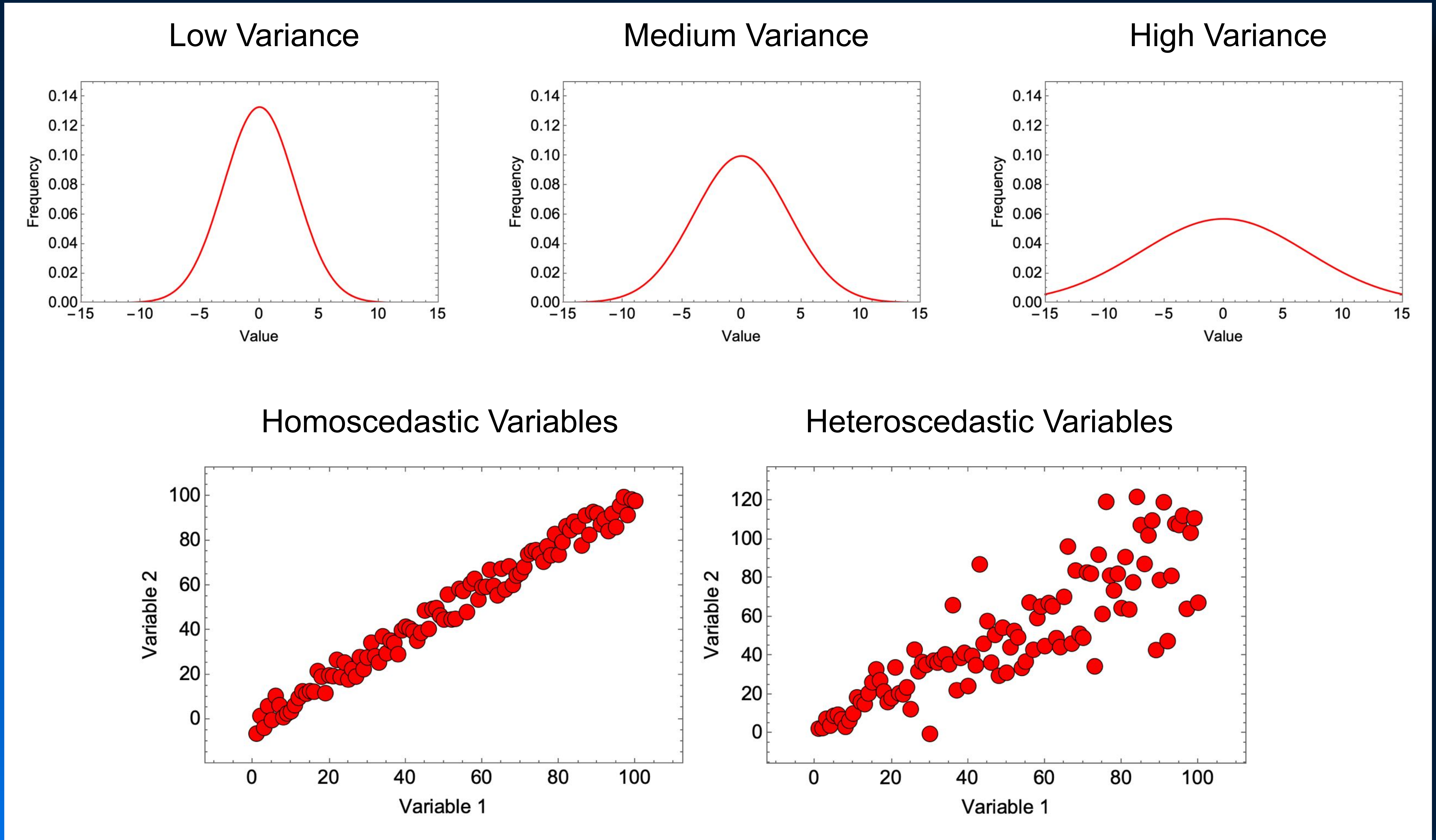
		Probability	
Percentile	Value	(p)	(α)
10%	65	0.10	0.90
25%	90	0.25	0.75
50%	100	0.50	0.50
90%	119	0.90	0.10
95%	125	0.95	0.05



The percentile can be calculated for one-tailed or two-tailed distributions and is the parameter that gives meaning to the concept of frequentist probability estimation.

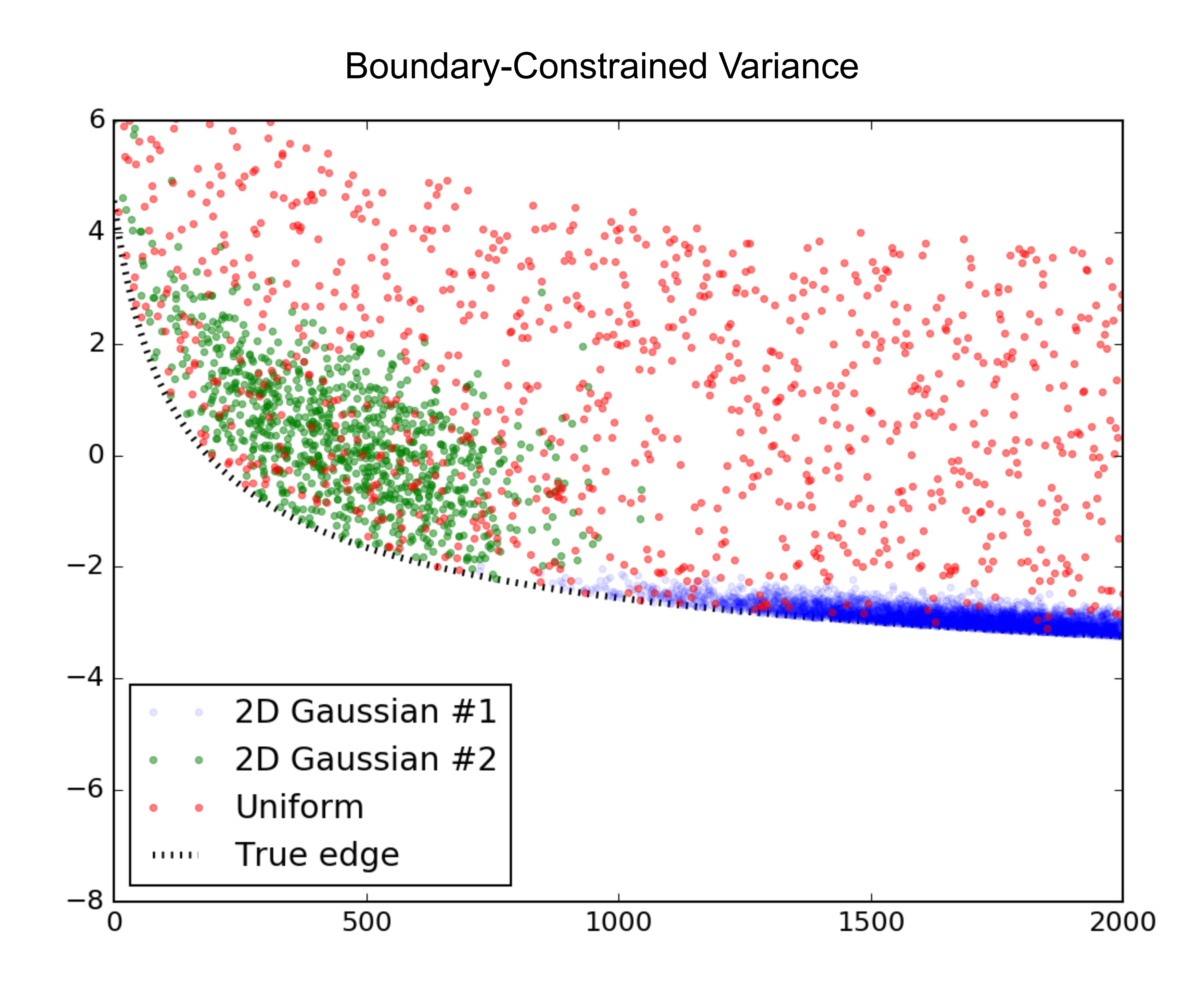
Statistics as Data Analysis

Describing Patterns of Variation



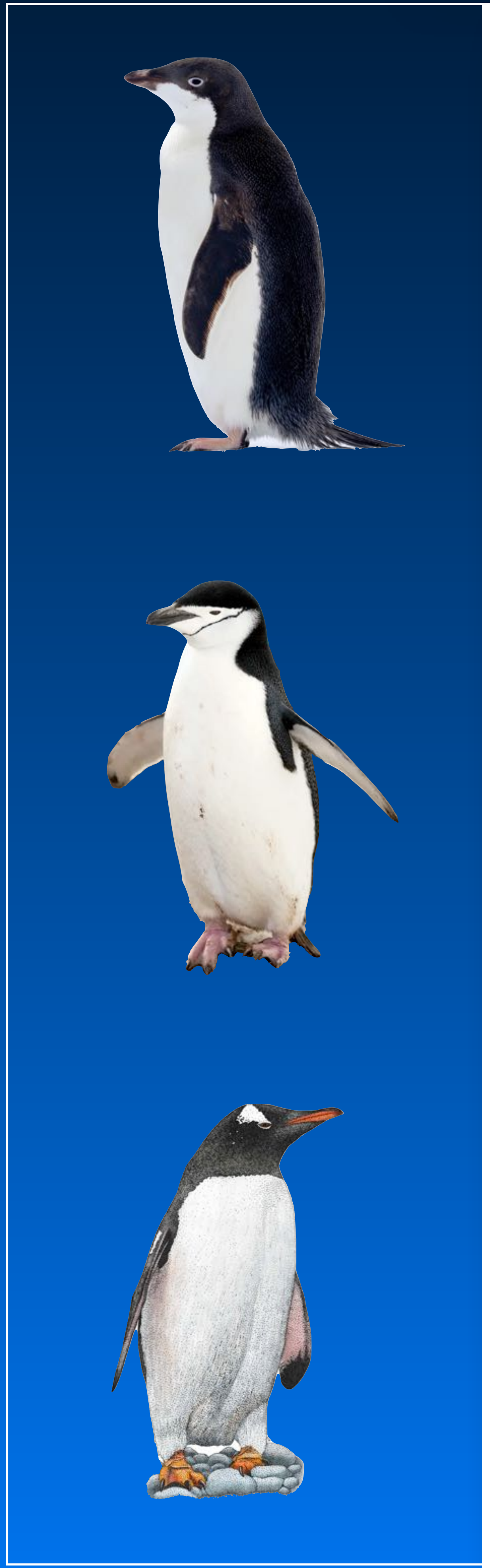
Statistics as Data Analysis

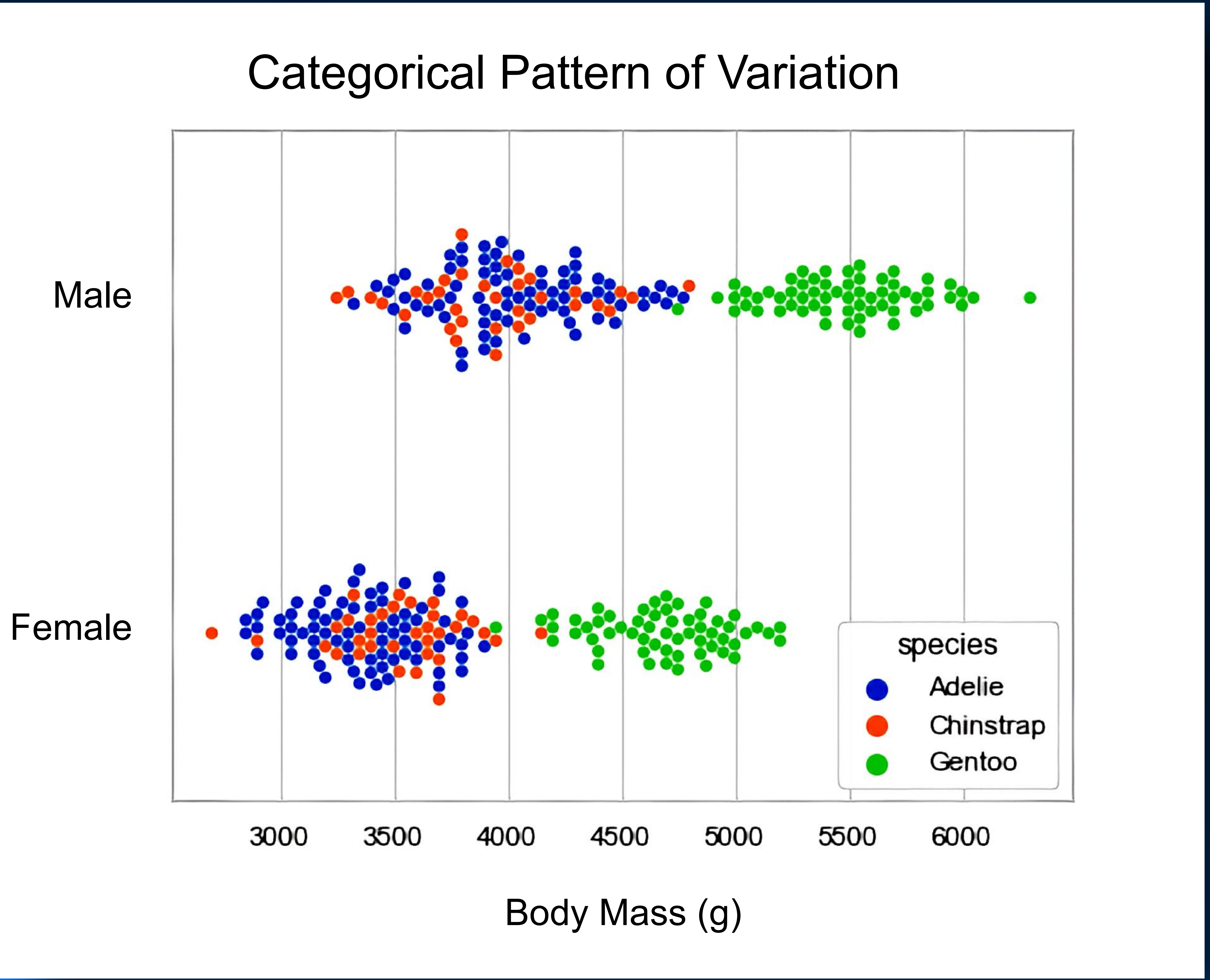
Describing Patterns of Variation



Statistics as Data Analysis

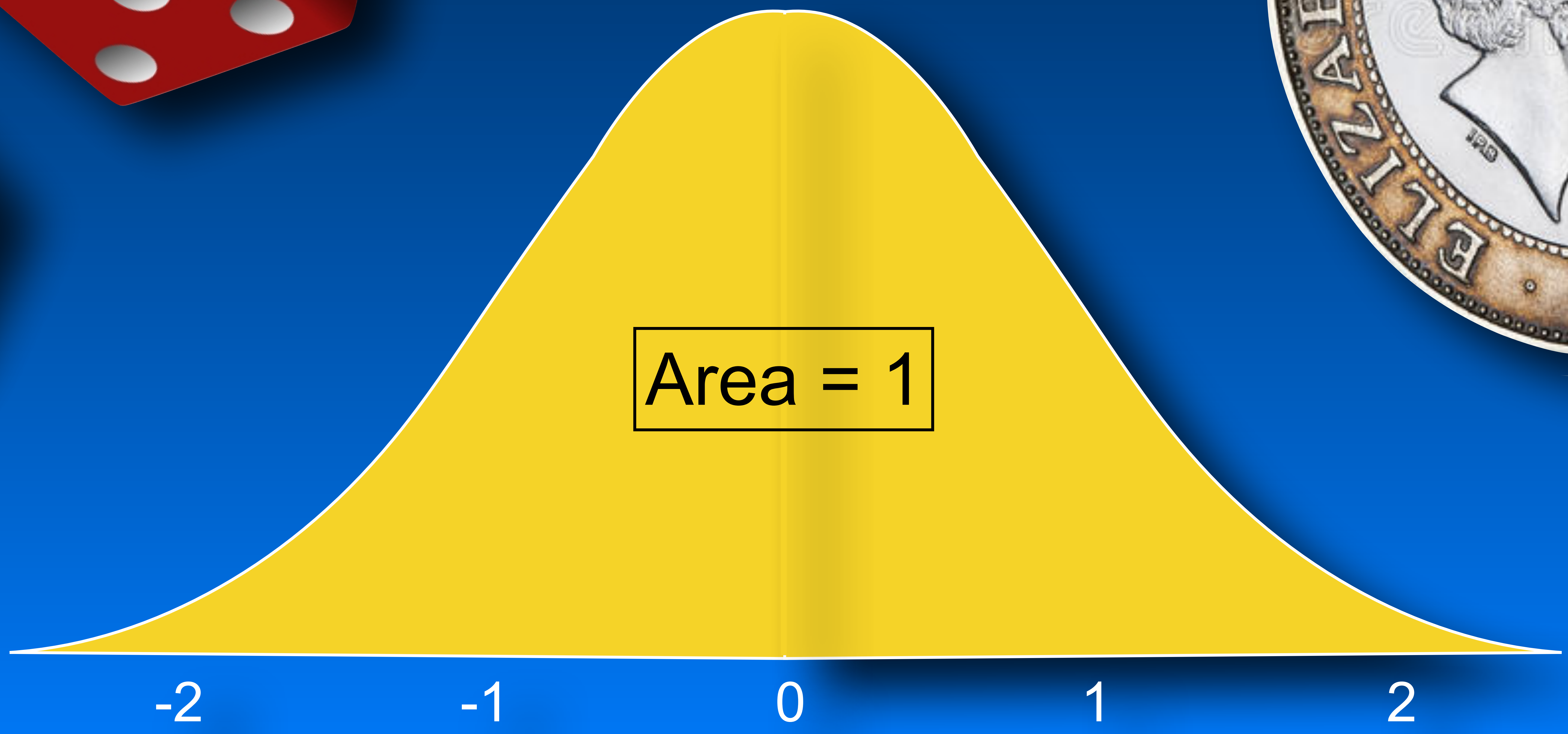
Describing Patterns of Variation





Statistics as Probabilistic Inference

Probability



Statistics as Probabilistic Inference

Probability

The branch of mathematics concerned with numerical descriptions of how likely an event will occur or a proposition is true.

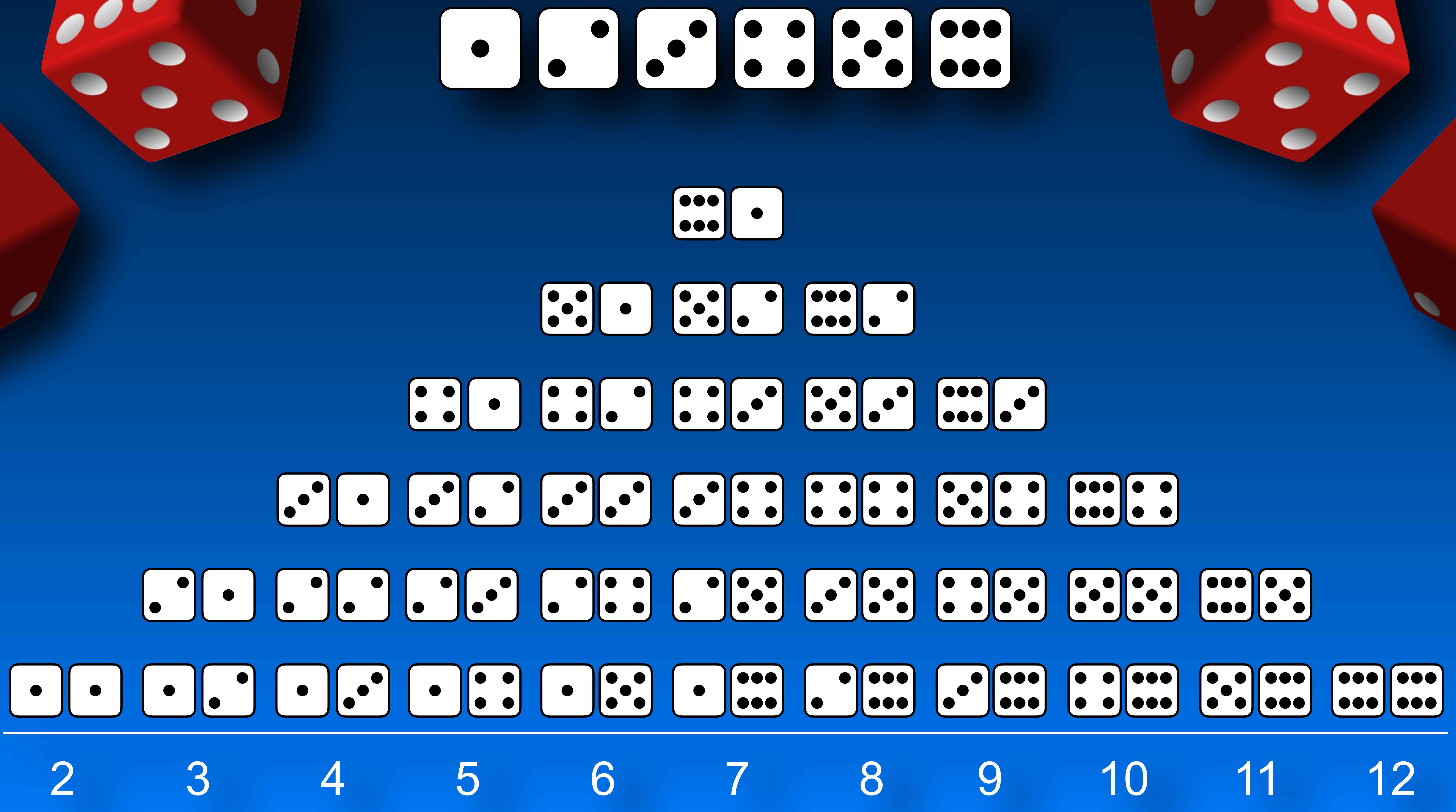
What is the most common result?

What is the chance we'll roll a 7?

What's the probability of rolling a 7?

Statistics as Probabilistic Inference

Probability



Statistics as Probabilistic Inference

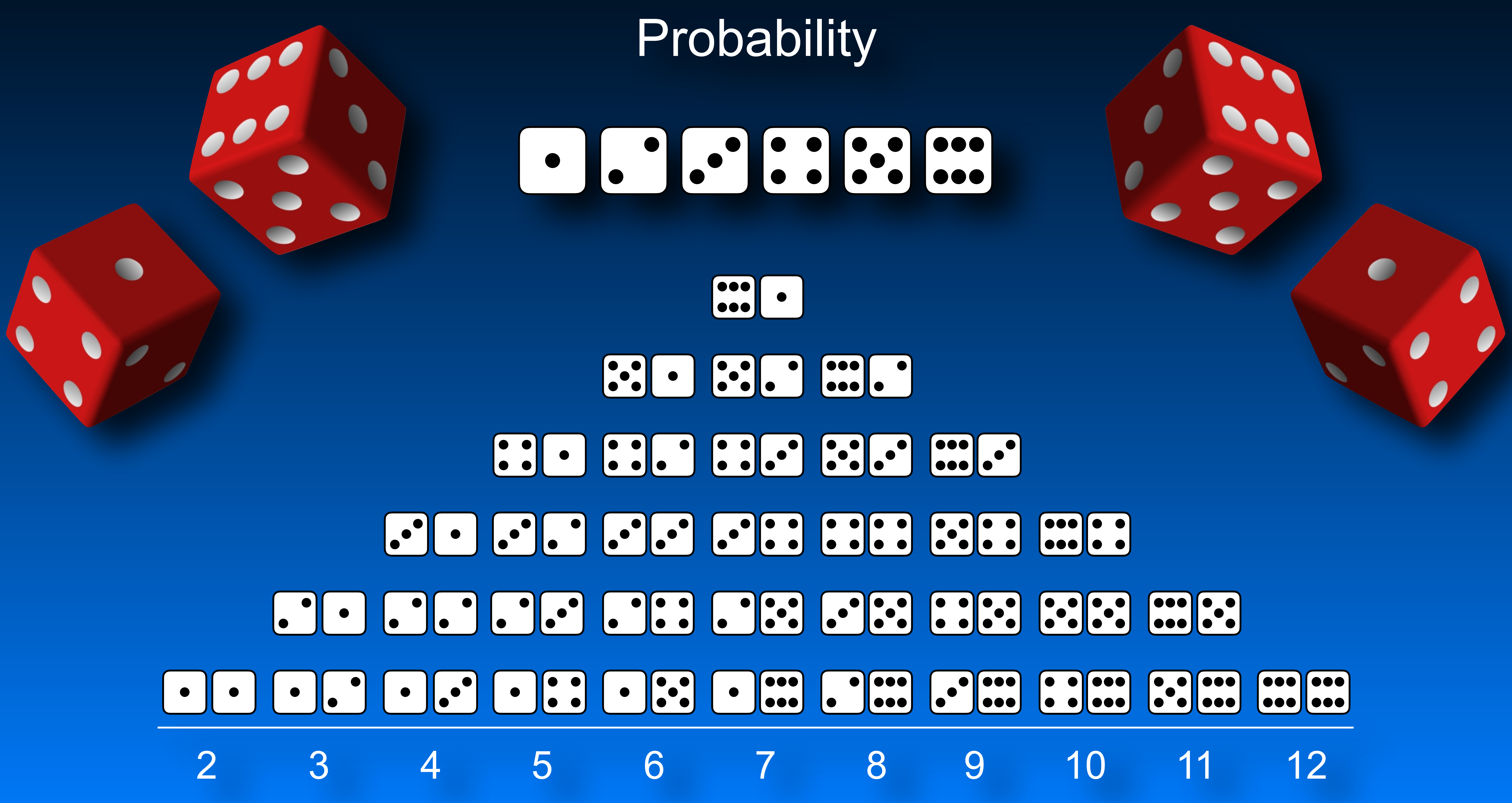
Probability

The branch of mathematics concerned with numerical descriptions of how likely an event will occur or a proposition is true.

What is the most common result?
7

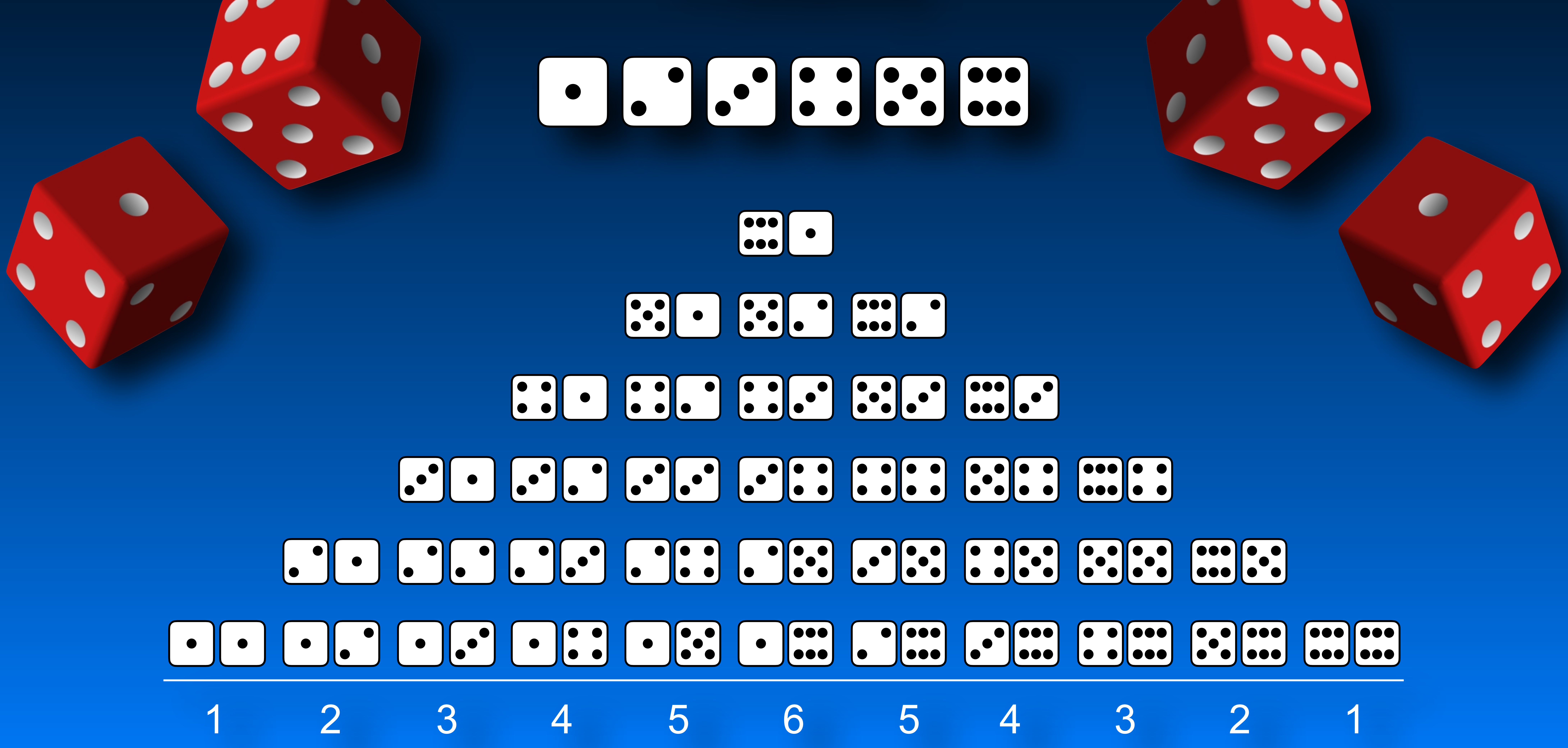
What is the chance we'll roll a 7?

Statistics as Probabilistic Inference



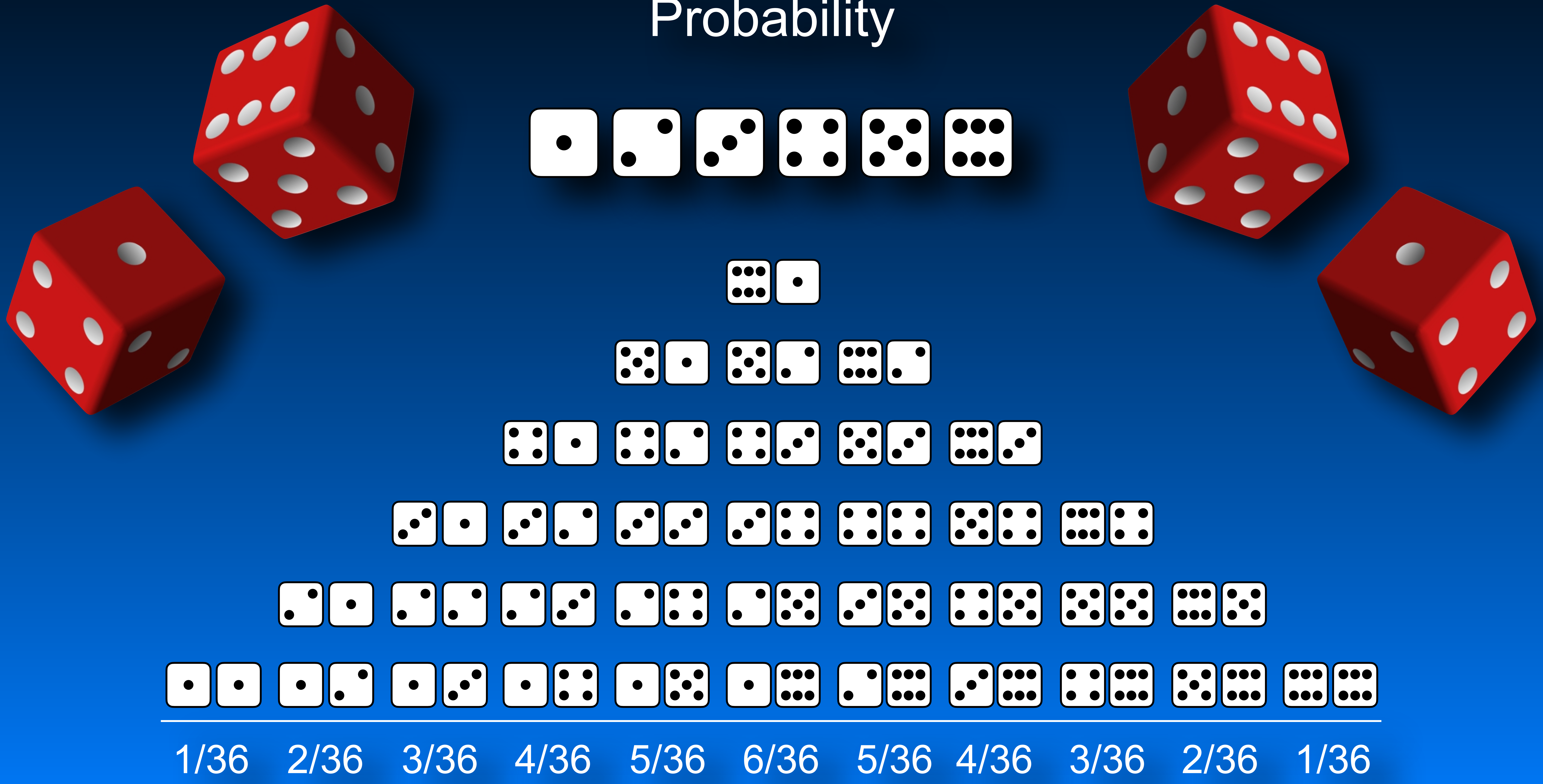
Statistics as Probabilistic Inference

Probability



Statistics as Probabilistic Inference

Probability



Statistics as Probabilistic Inference

Probability

The branch of mathematics concerned with numerical descriptions of how likely an event will occur or a proposition is true.

What is the most common result?

7

Statistics as Probabilistic Inference

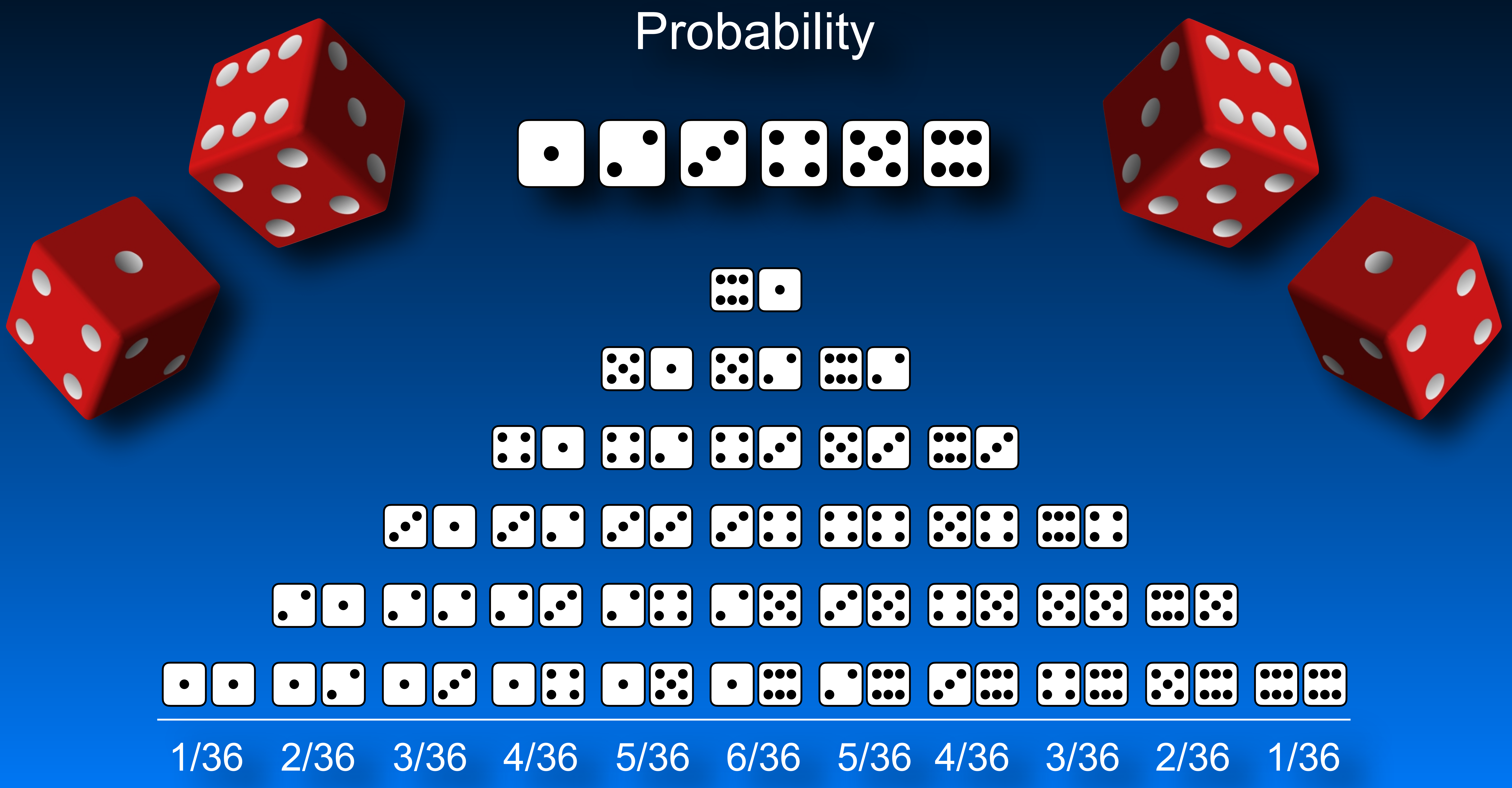
Probability

The branch of mathematics concerned with numerical descriptions of how likely an event will occur or a proposition is true.

What is the most common result? 7

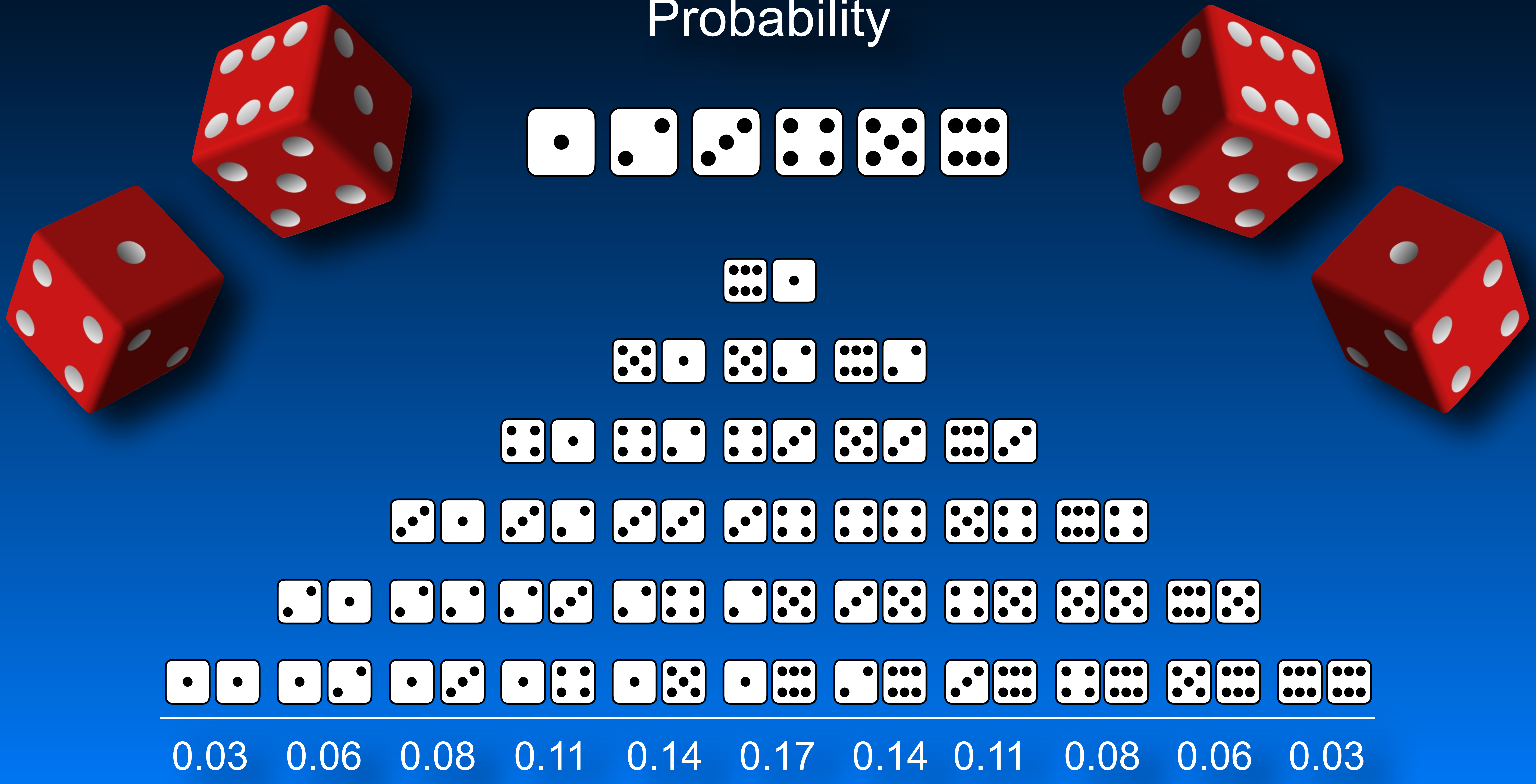
What is the chance we'll roll a 7? 6/36

Statistics as Probabilistic Inference



Statistics as Probabilistic Inference

Probability



Statistics as Probabilistic Inference

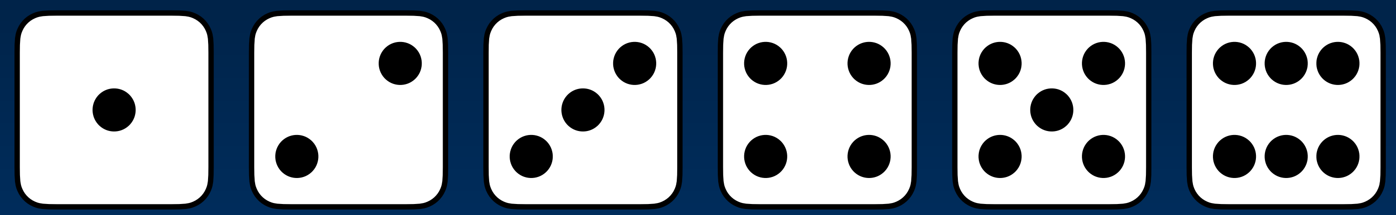
Probability

The branch of mathematics concerned with numerical descriptions of how likely an event will occur or a proposition is true.

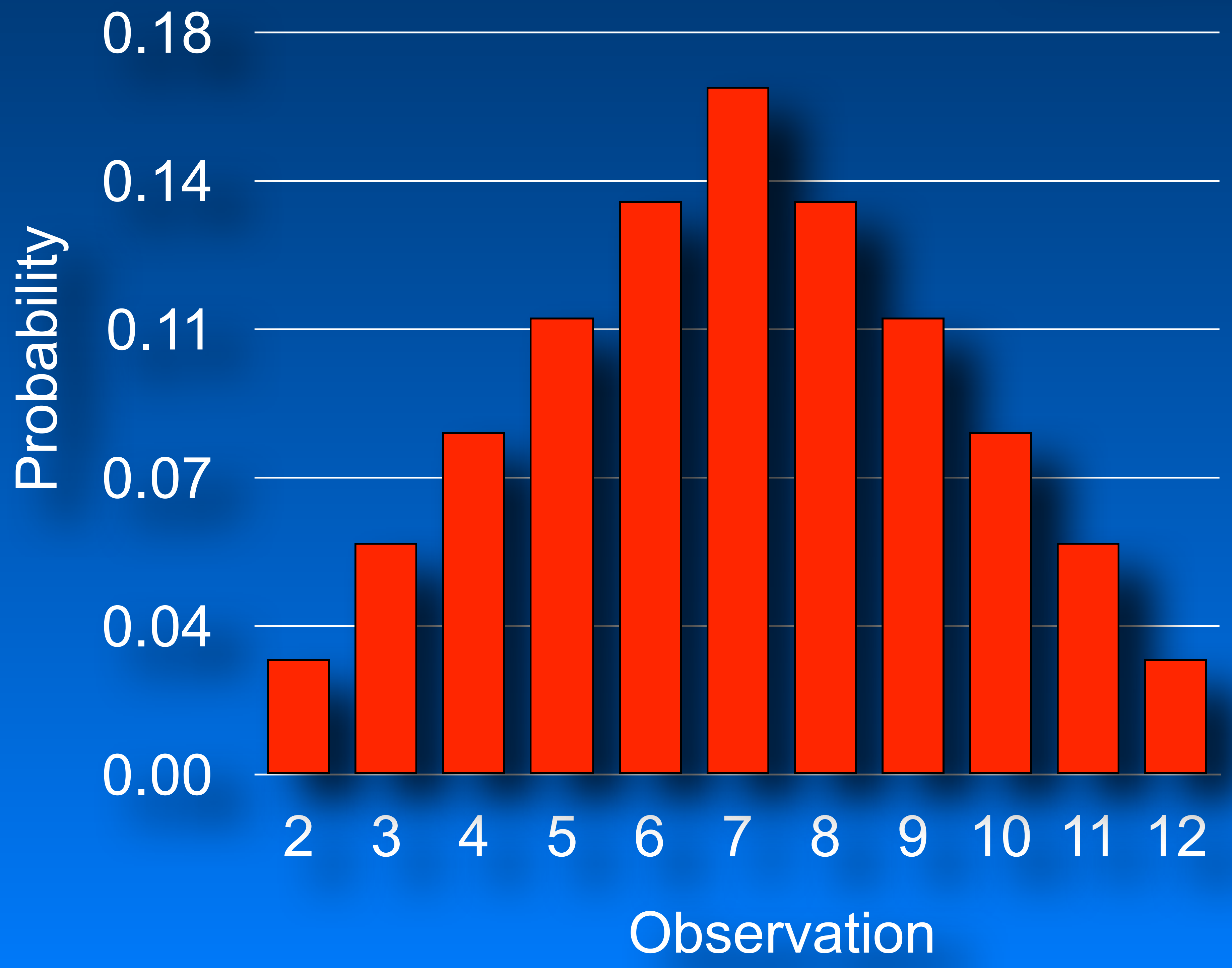
What is the most common result?	7
What is the chance we'll roll a 7?	6/36
What's the probability of rolling a 7?	16.67%

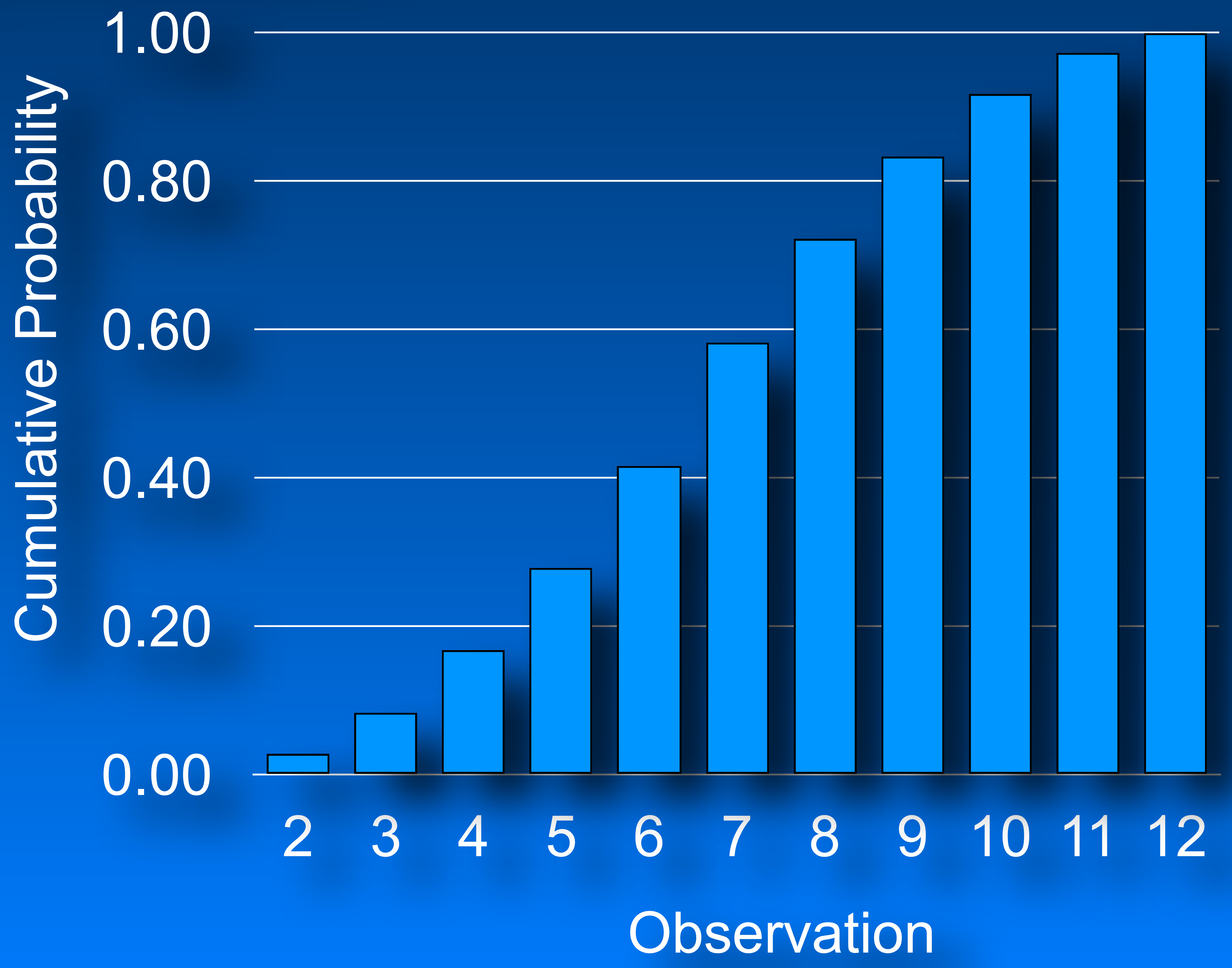
Statistics as Probabilistic Inference

Probability



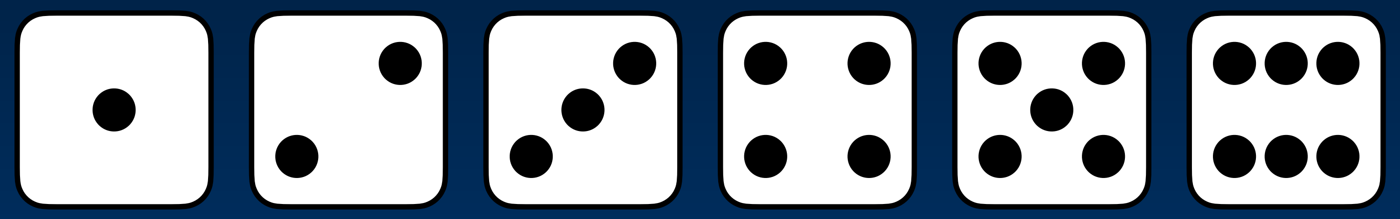
Discrete Probability Distribution



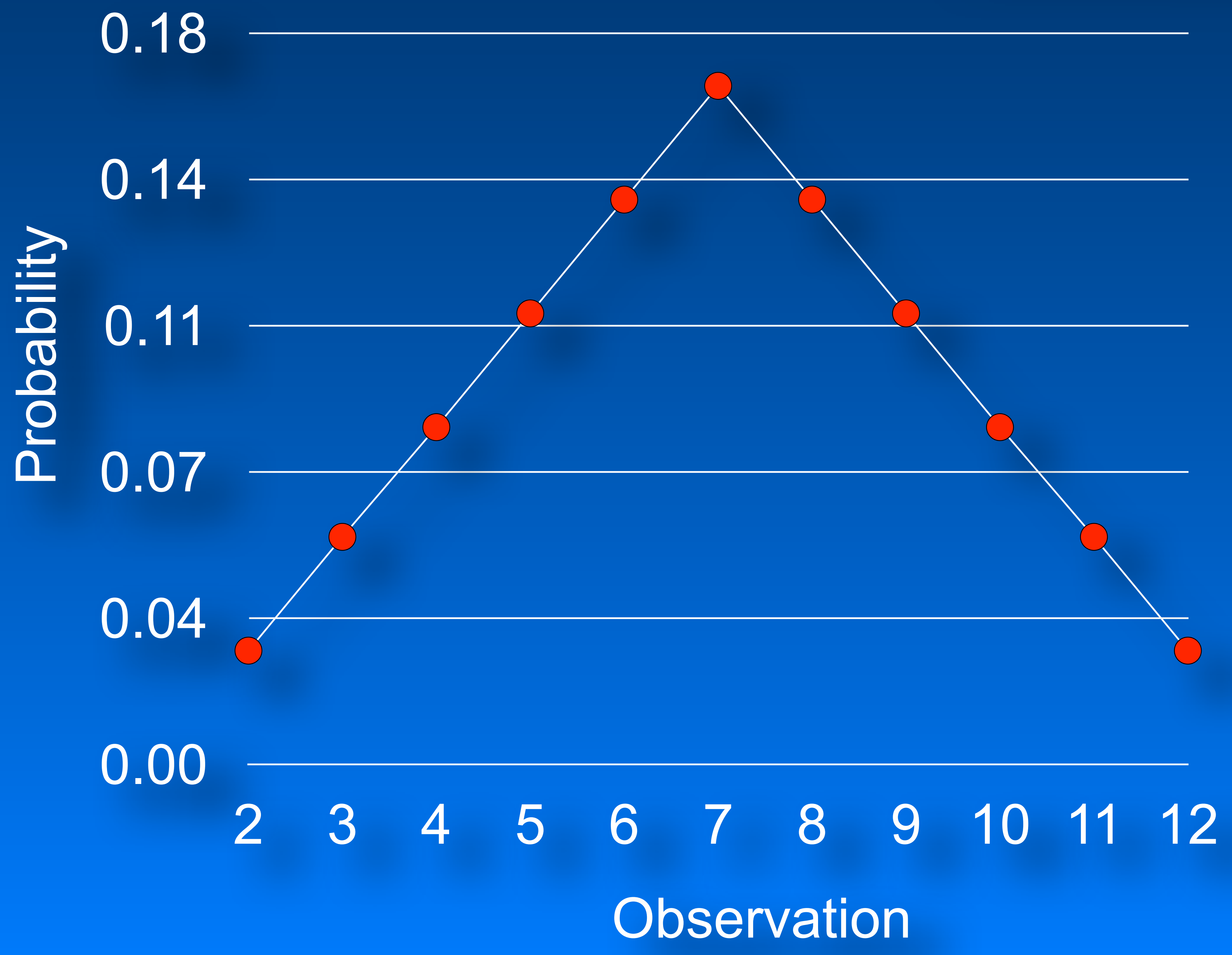


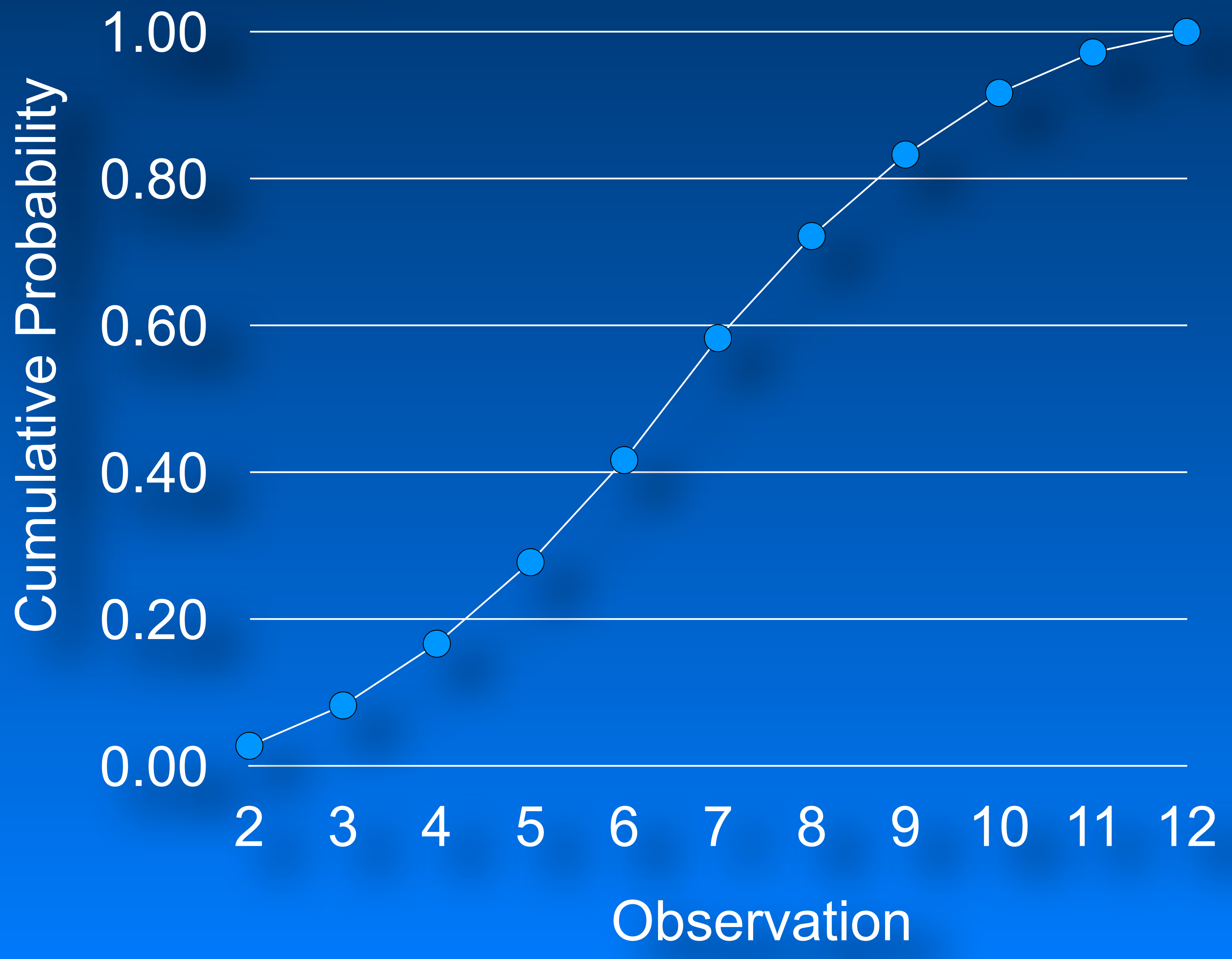
Statistics as Probabilistic Inference

Probability



Continuous Probability Distribution





Frequentist Inference

Assumes any given experiment can be considered as one of an infinite sequence of possible repetitions of the same experiment, each capable of producing a statistically independent result. Under this approach the conclusion is drawn by locating empirical results within this hypothetical set of repetitions.

Unknown parameters are usually treated as having fixed, but unknown, values that are not capable of being treated as random variates themselves. Accordingly, there is no way to associate probabilities with unknown parameters under a frequentist approach to probabilistic inference.

Siméon Poisson
(1781 – 1840)

Statistics for Hypothesis Testing

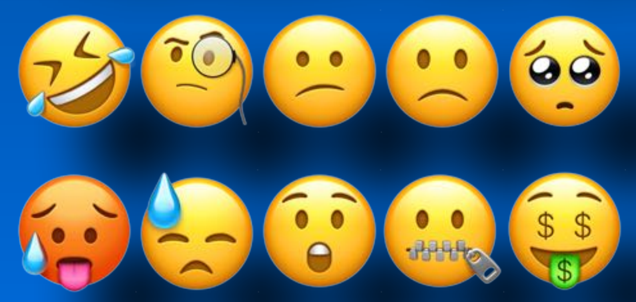
Basic Concepts

Population - the totality of individual observations existing within a specified area and/or time.

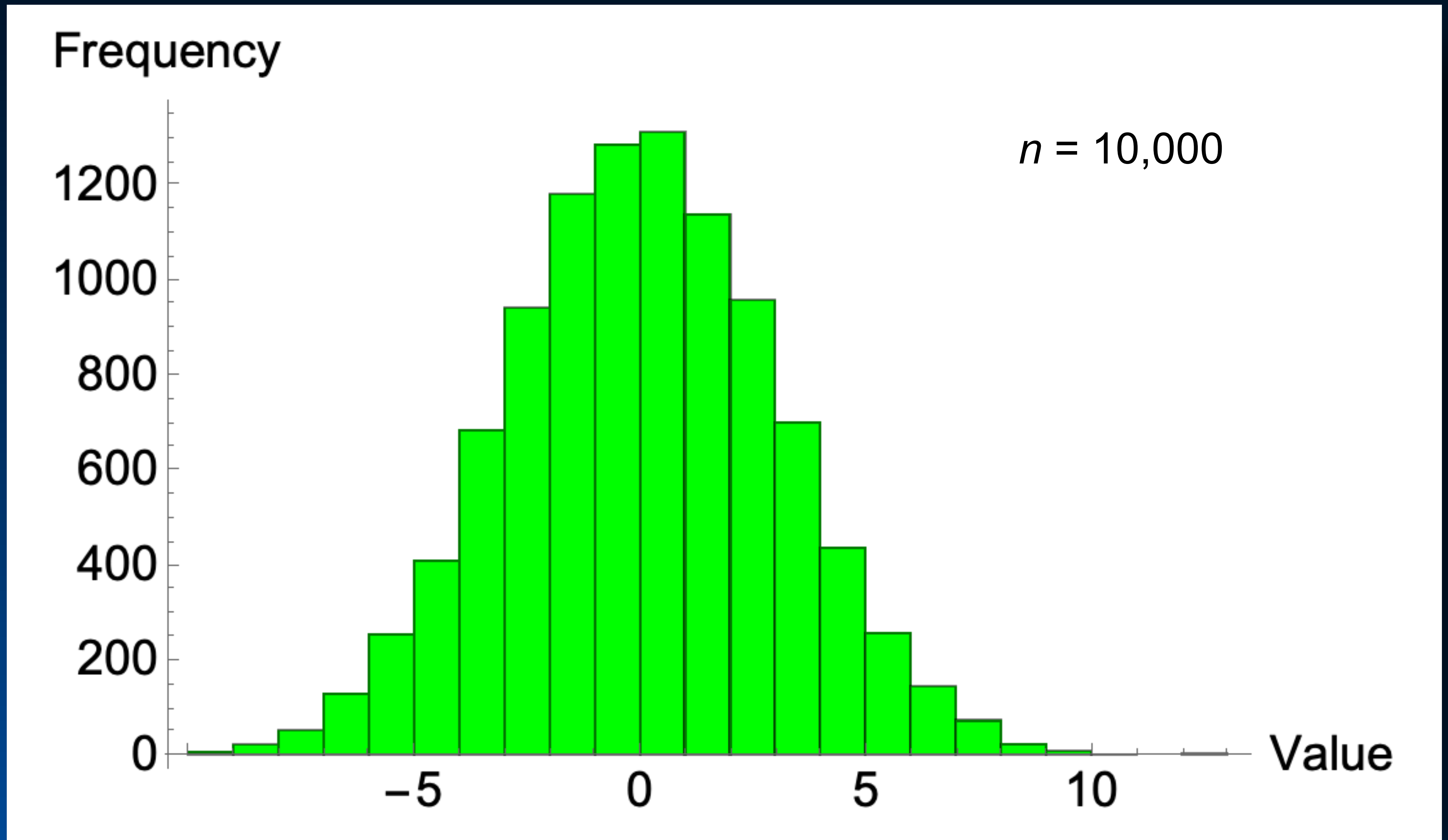
Sample - the subset of the population containing individual specimens from which data have been collected.

Biased Sample?

Random Sample



The Normal Distribution



- Approximates the expected distribution of observations being influenced by many random factors.
- Is the distribution of many derived descriptive statistical summaries (e.g., means)

Carl Friedrich Gauss
(1777 – 1855)

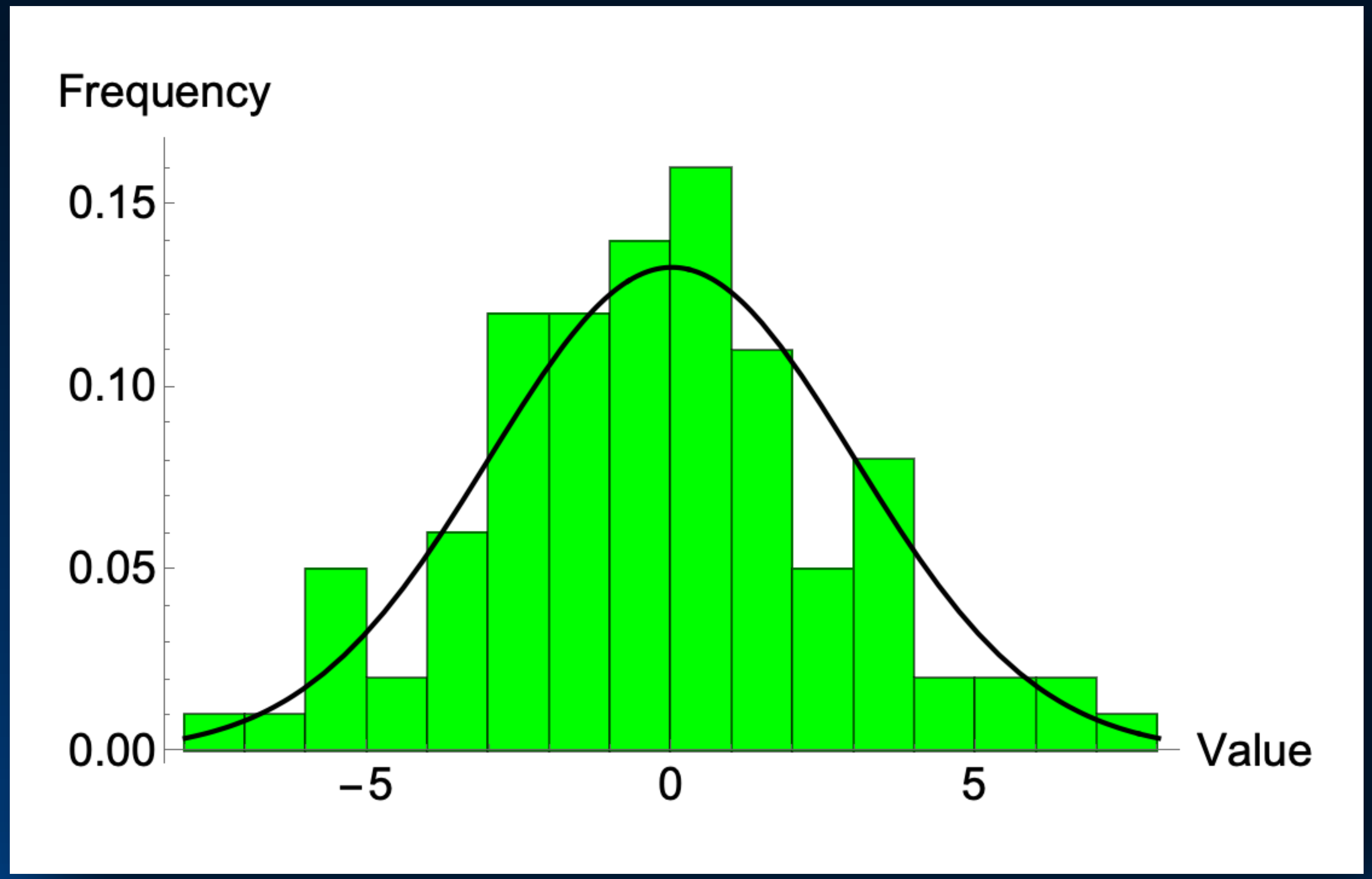
$$f_{(x)} = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

where: $i = i^{th}$ *observation*

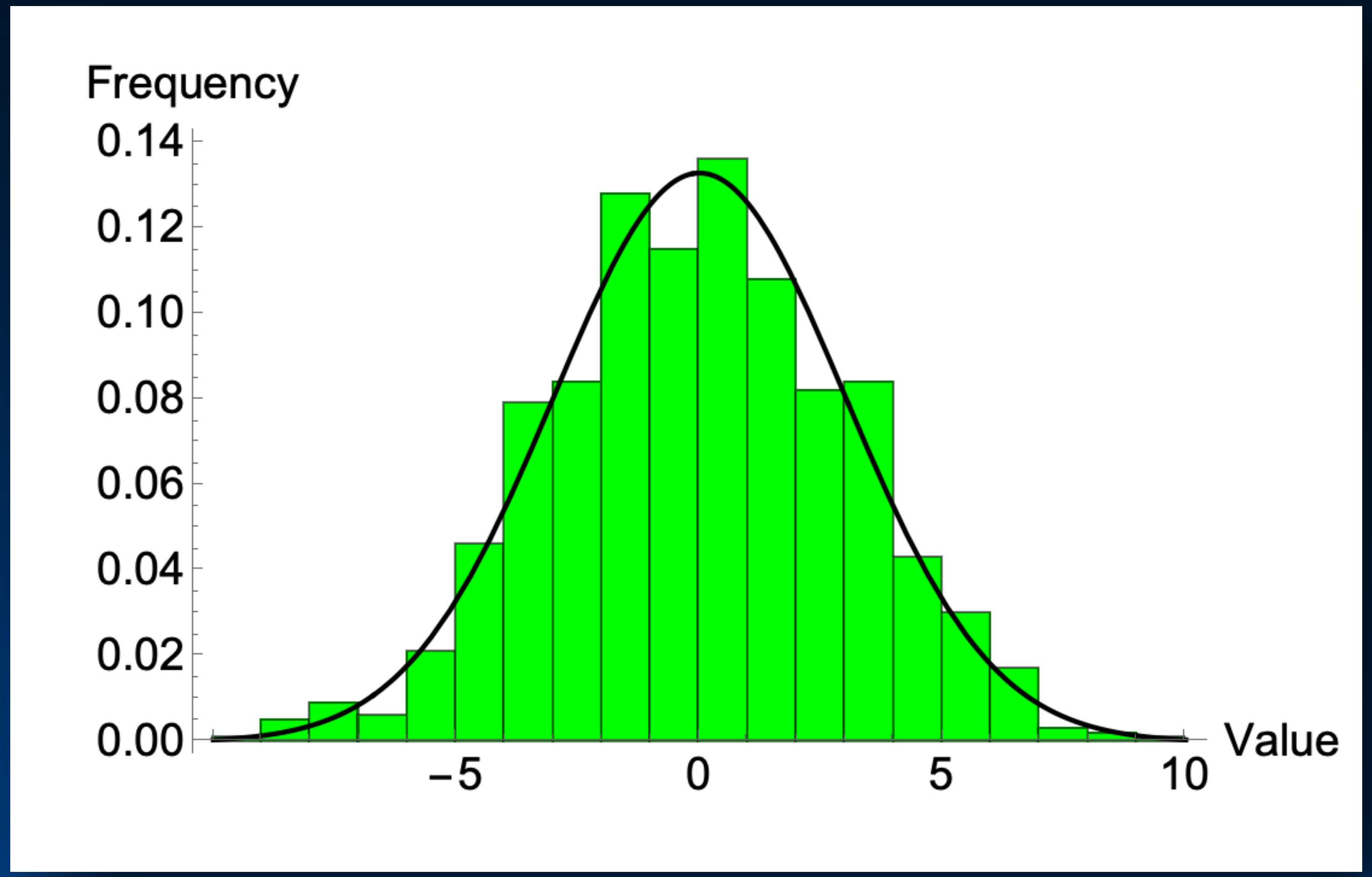
$\mu =$ *population mean*

$\sigma =$ *population standard deviation*

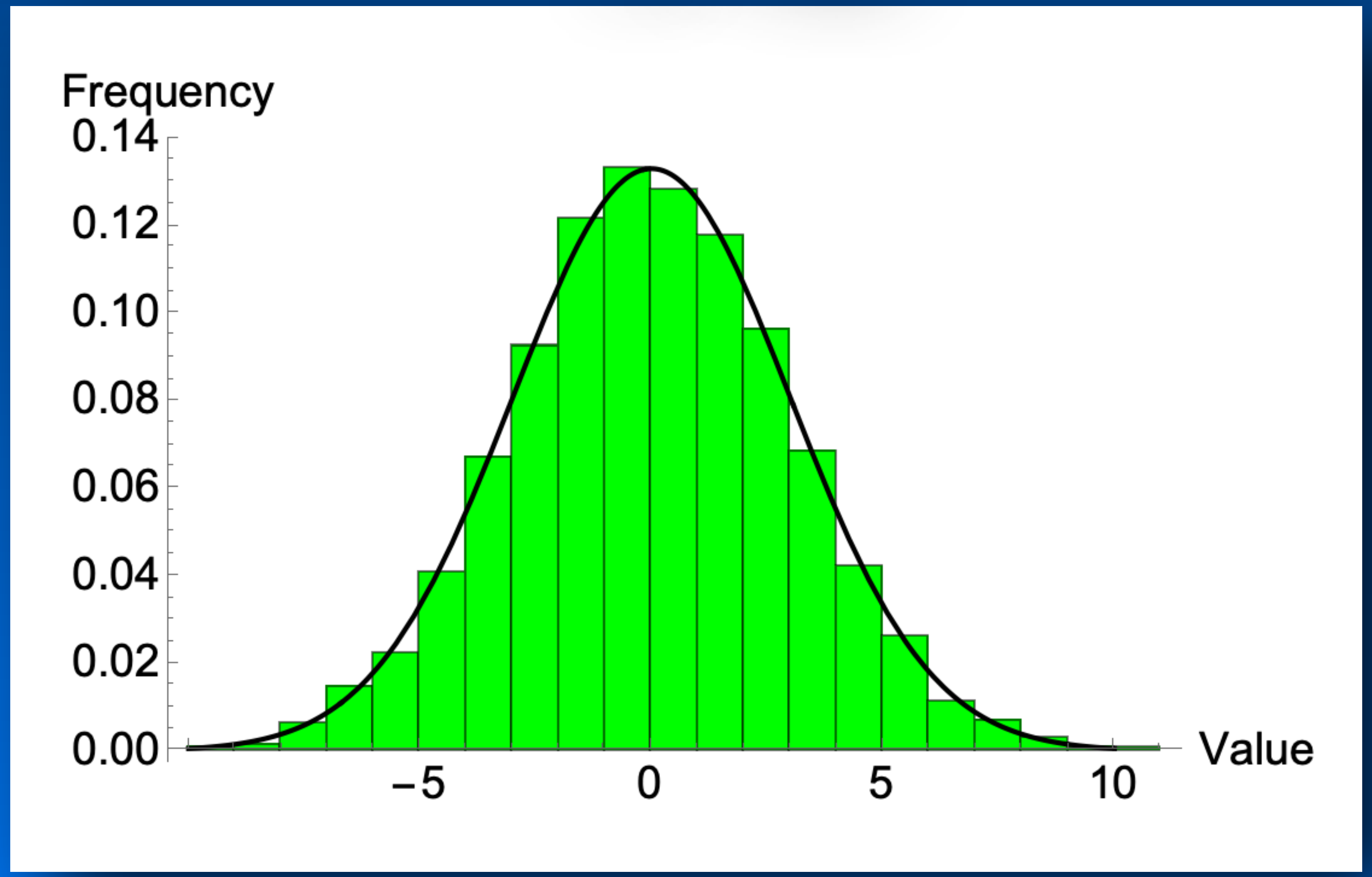
The Normal Distribution



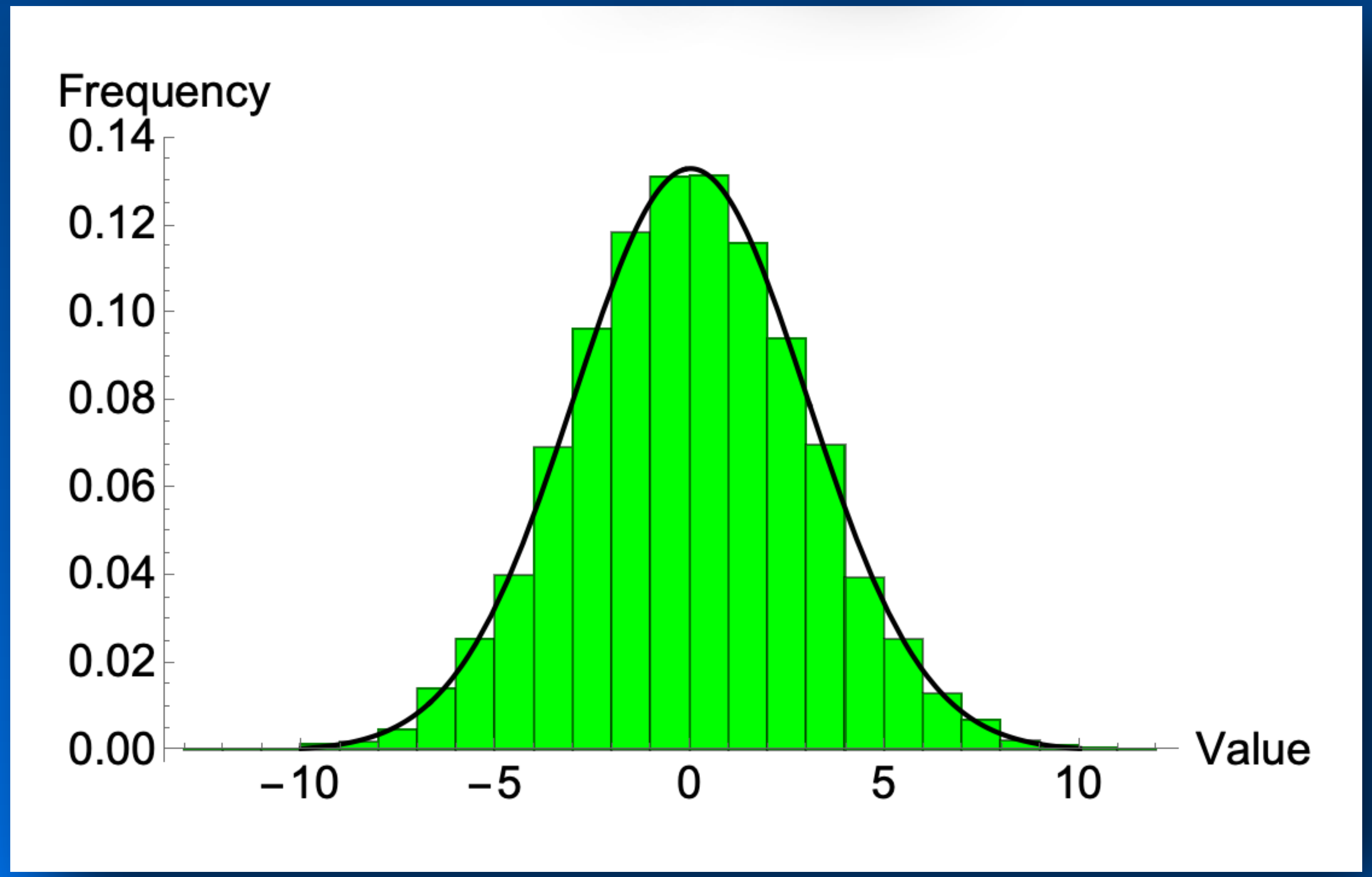
n = 100



n = 1,000

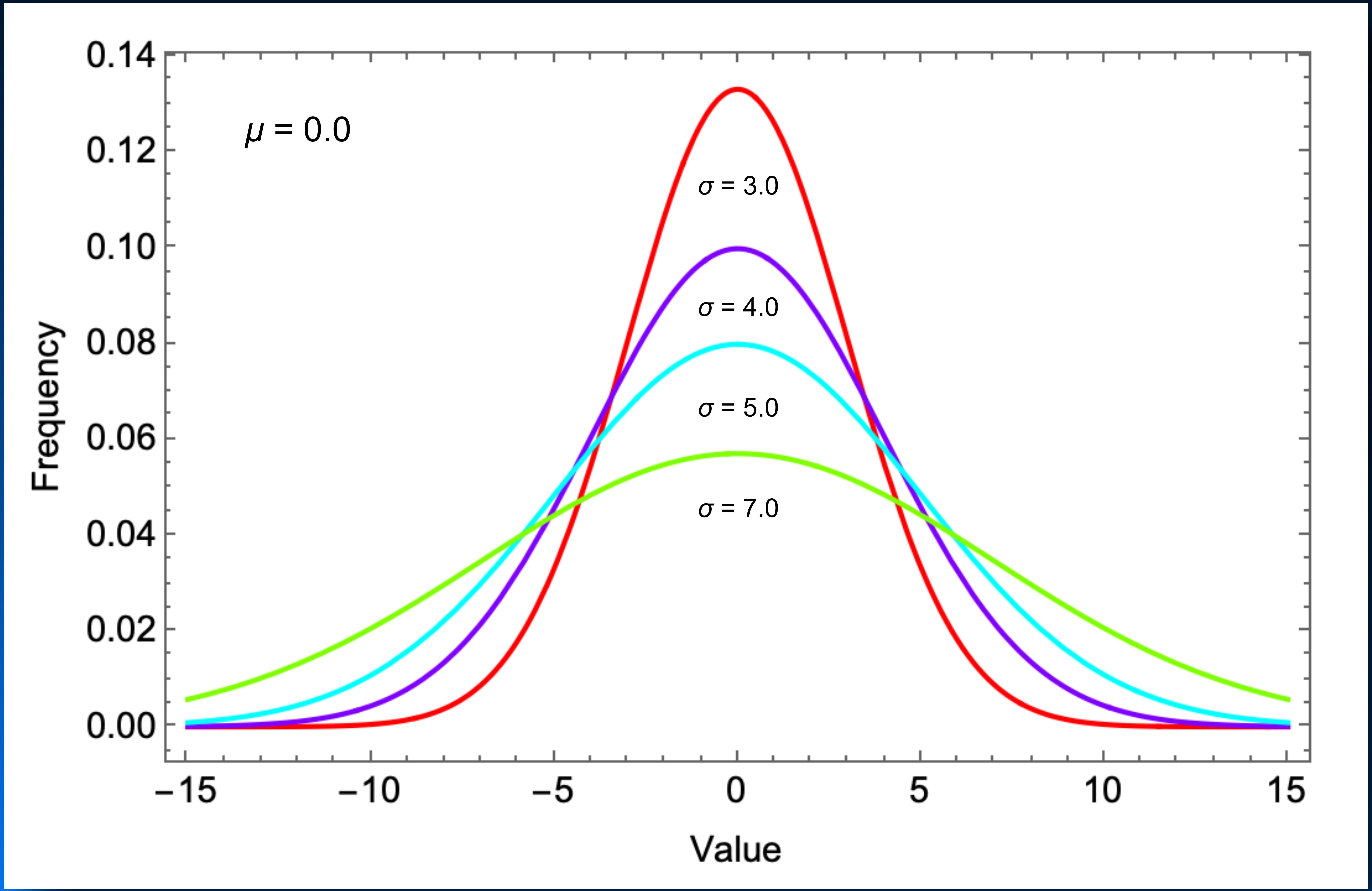


n = 5,000

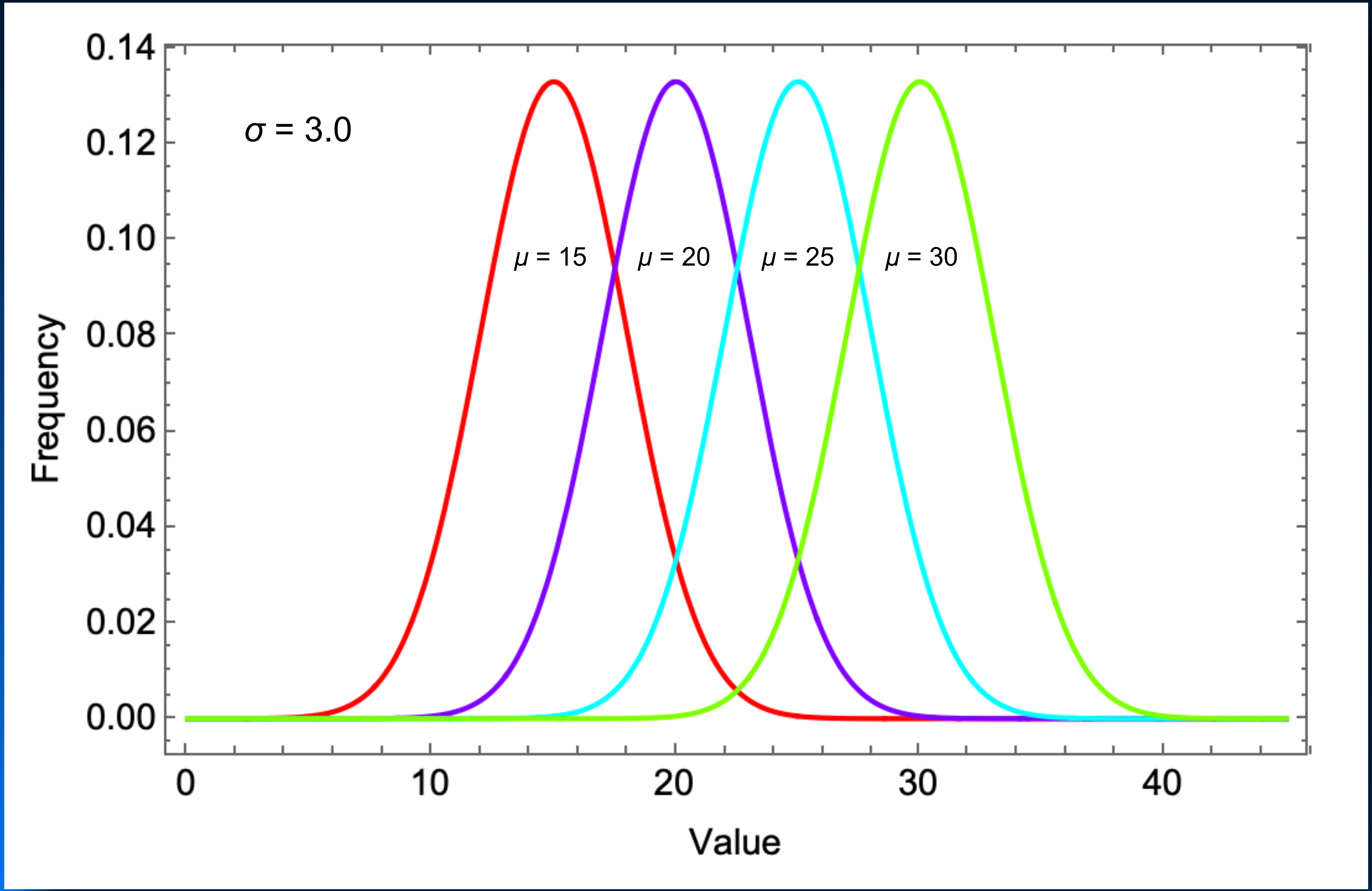


n = 10,000

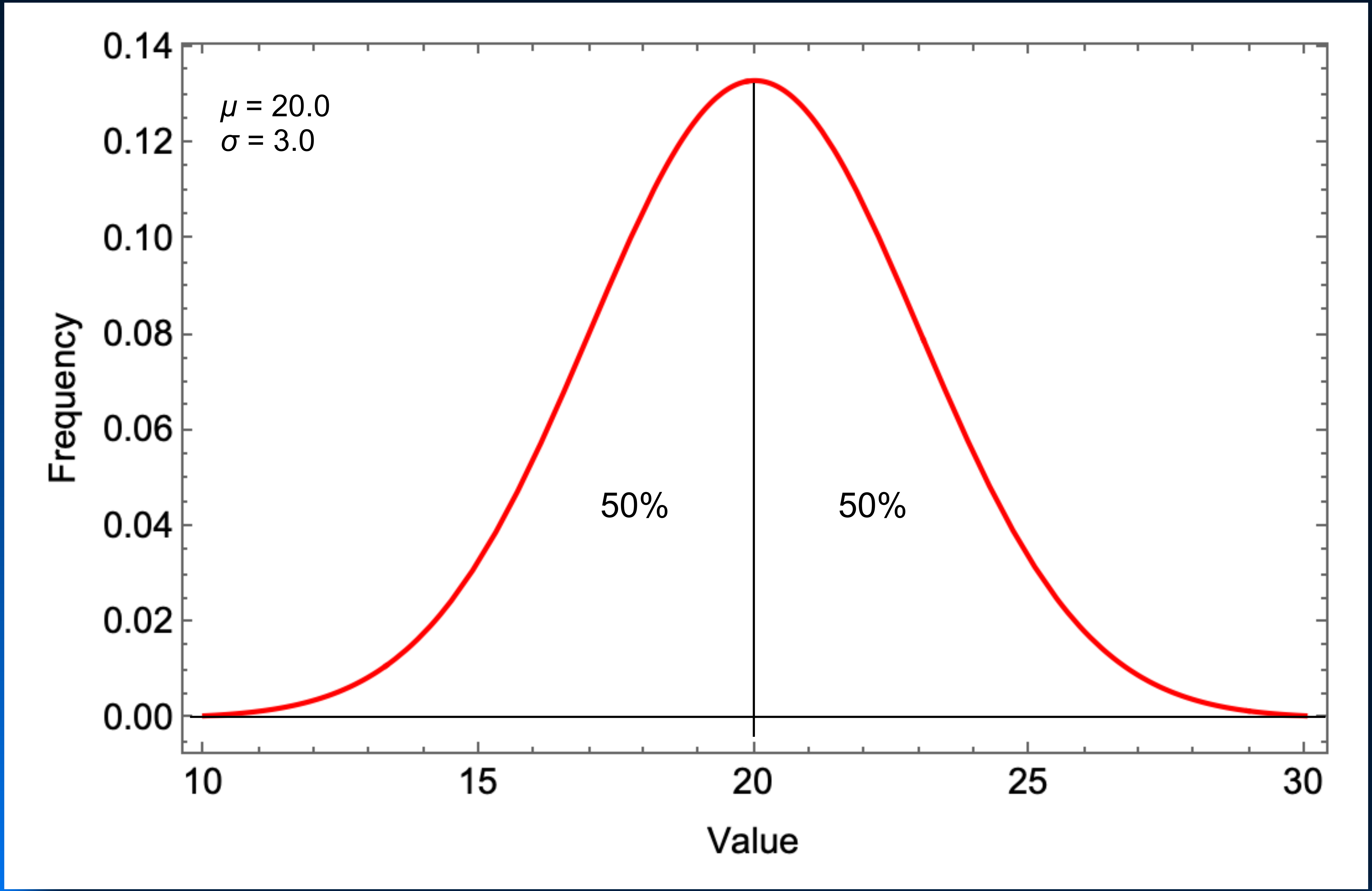
The Normal Distribution



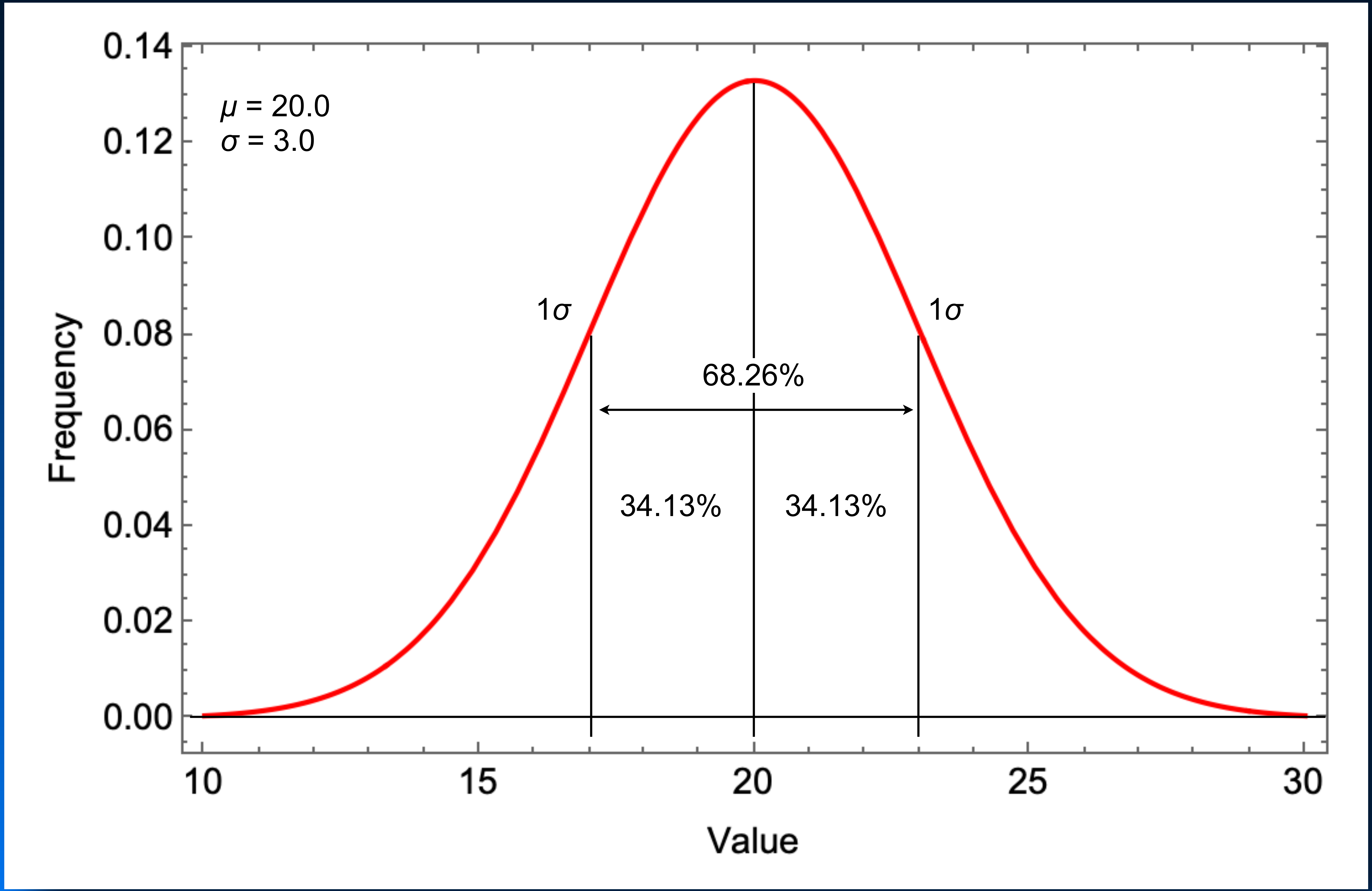
The Normal Distribution



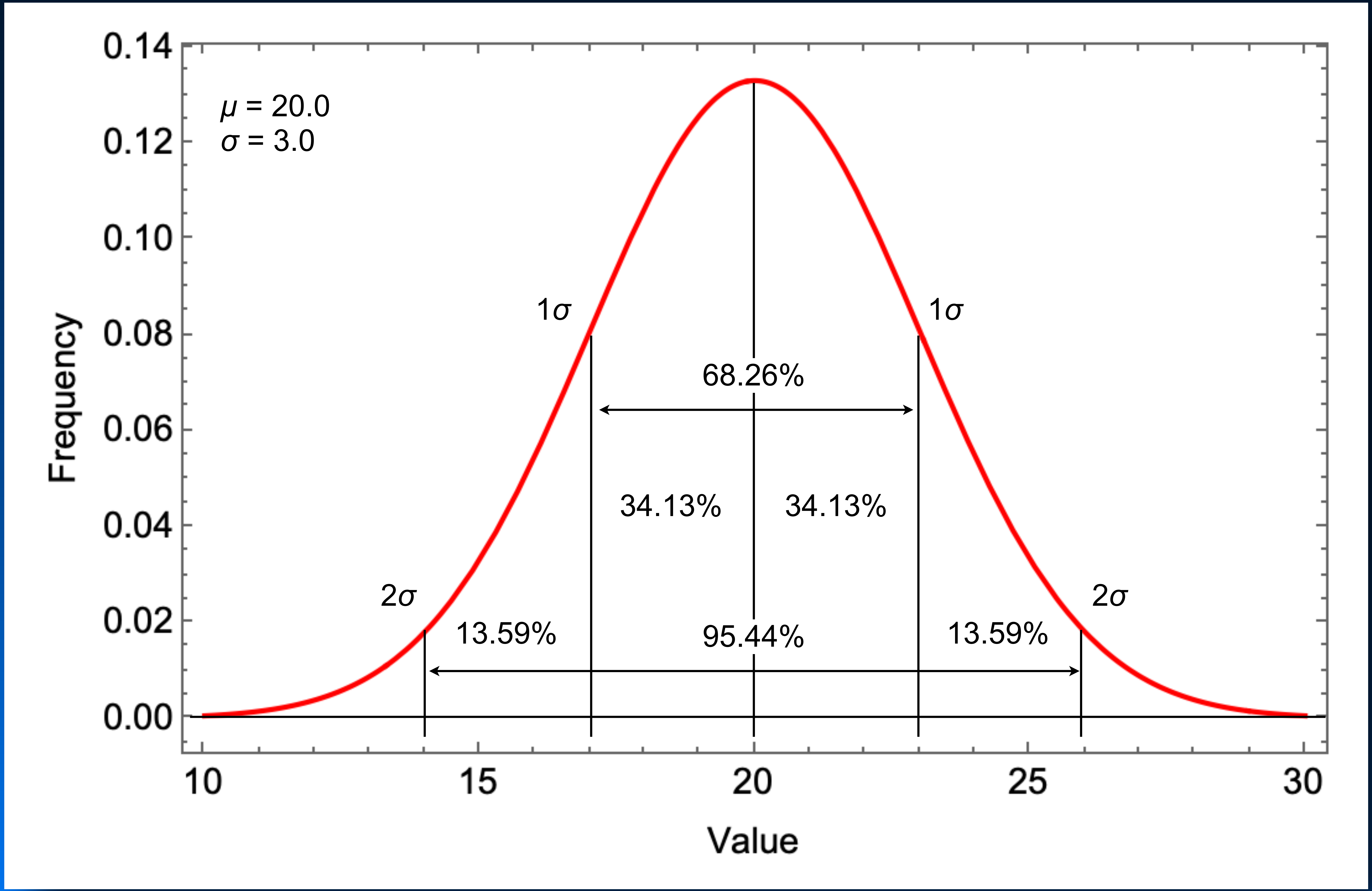
The Normal Distribution



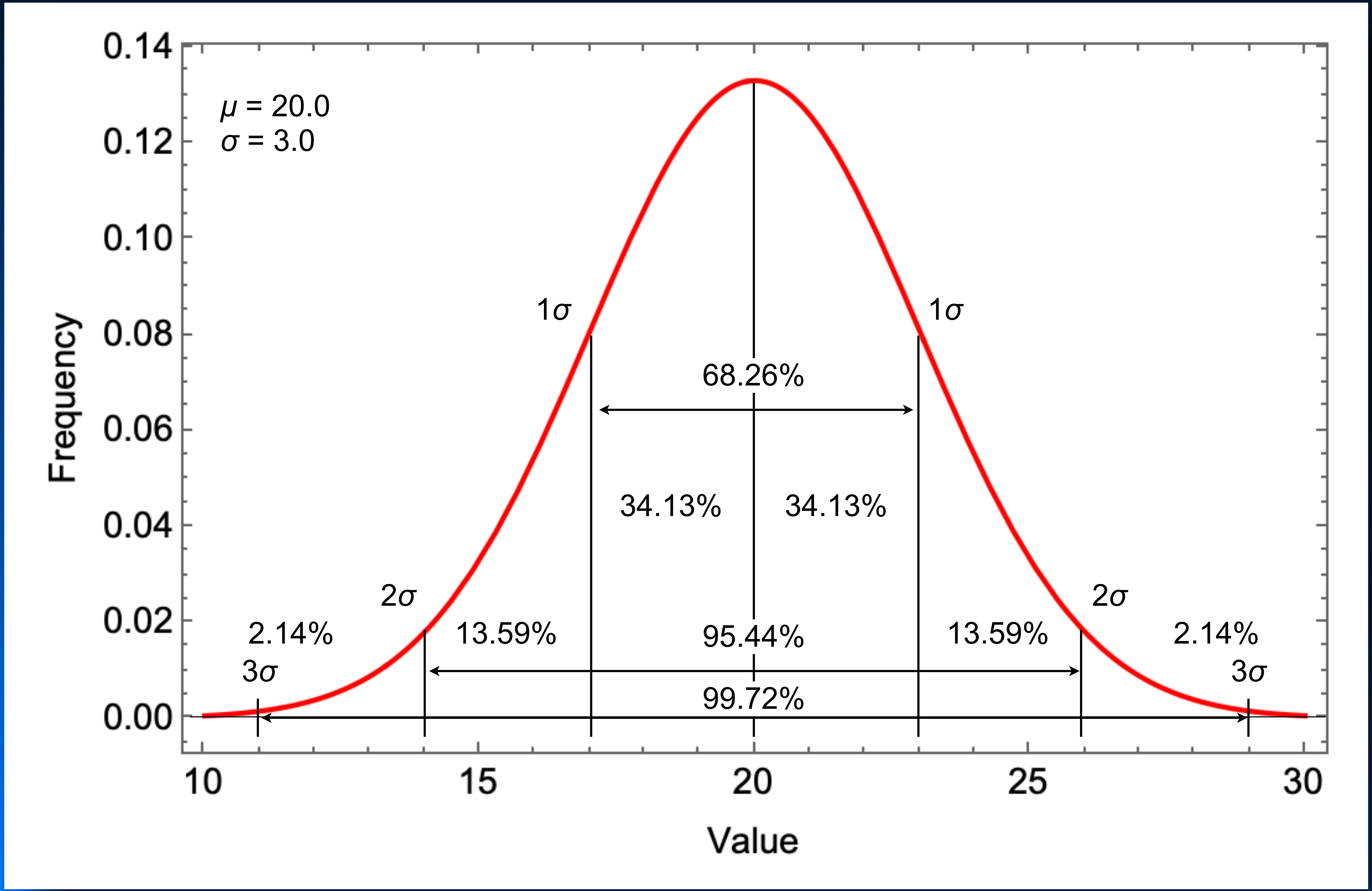
The Normal Distribution



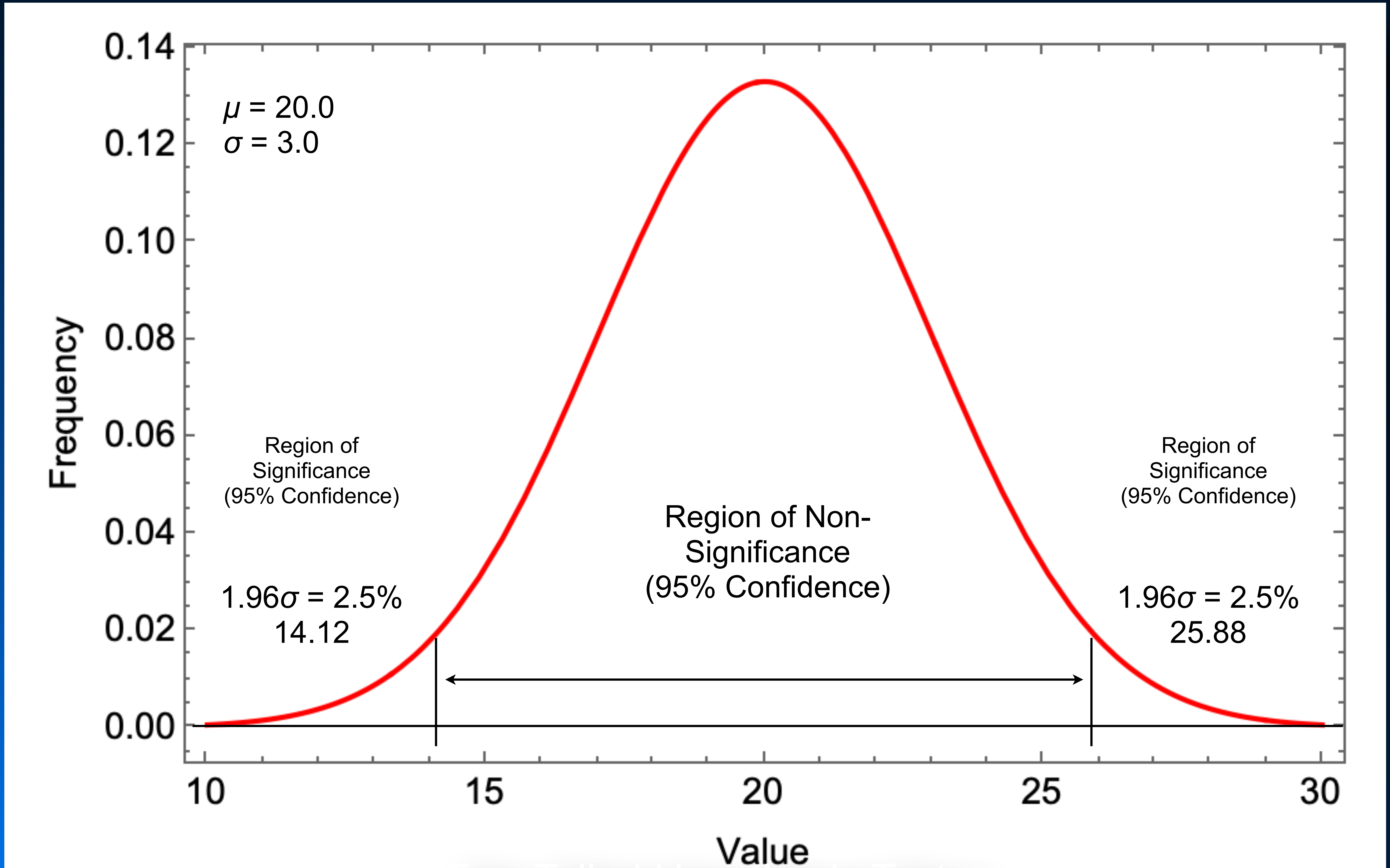
The Normal Distribution



The Normal Distribution

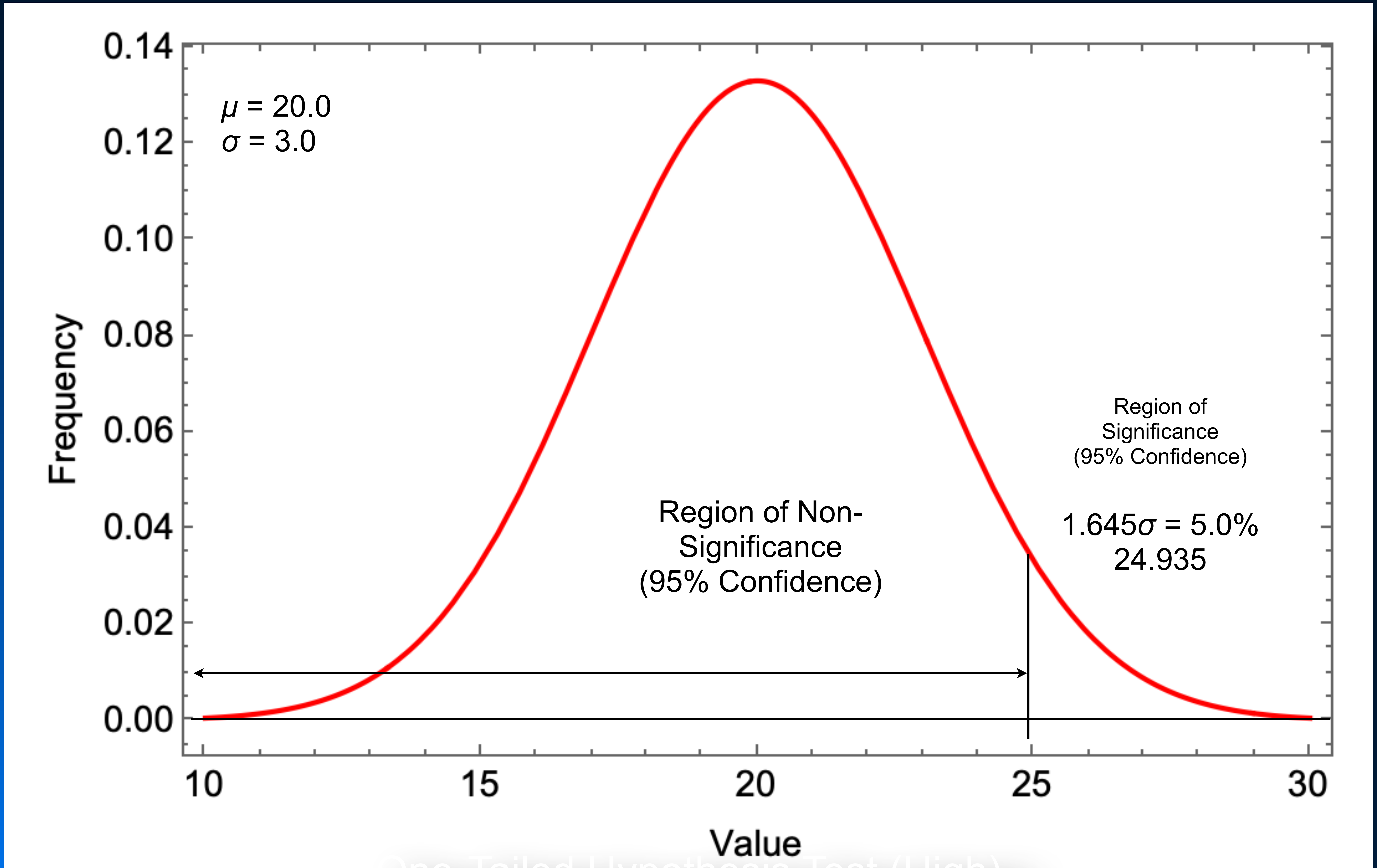


The Normal Distribution



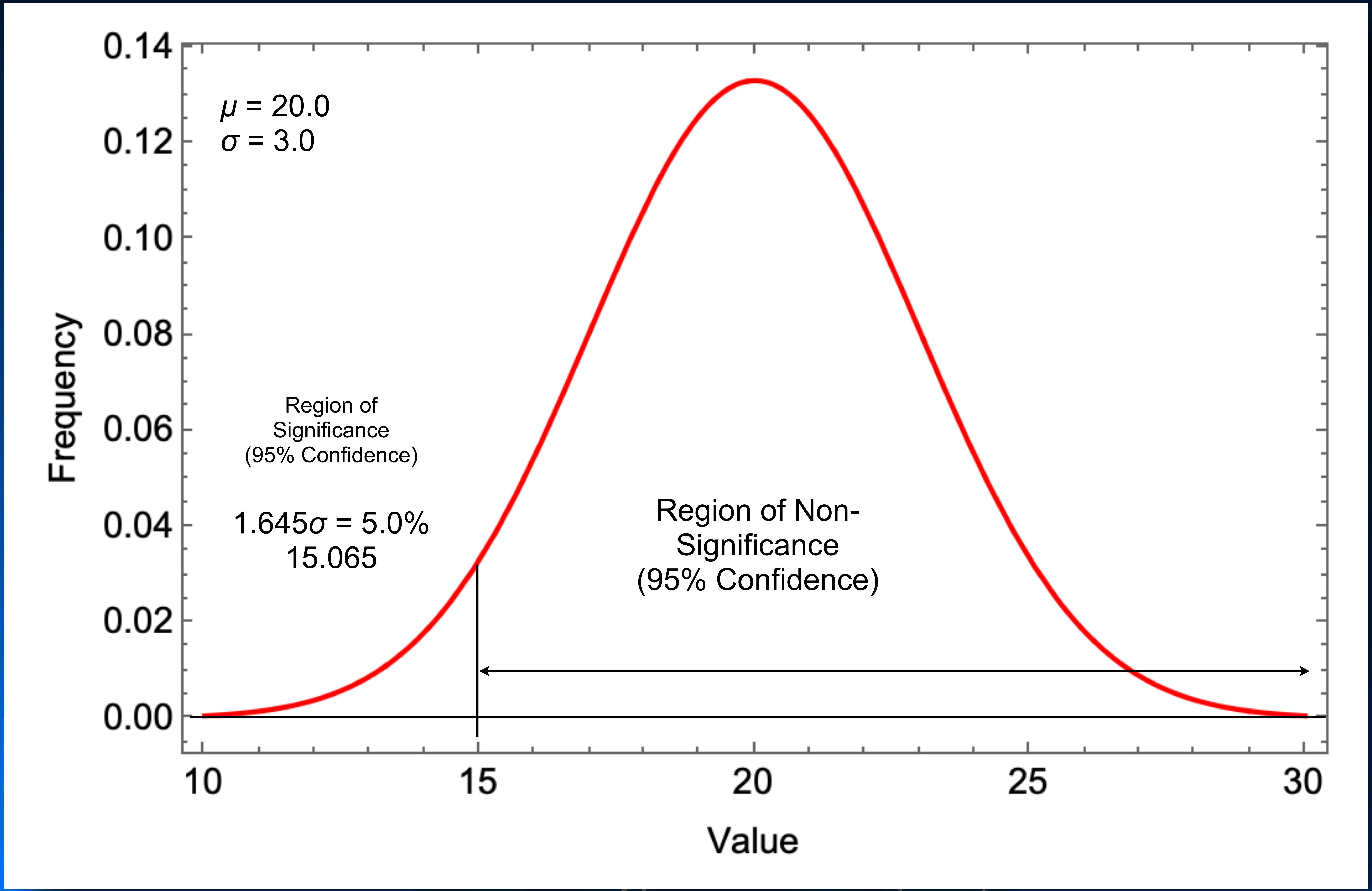
Two-Tailed Hypothesis Test

The Normal Distribution



One-Tailed Hypothesis Test (High)

The Normal Distribution



Statistics as Hypothesis Testing

An Example

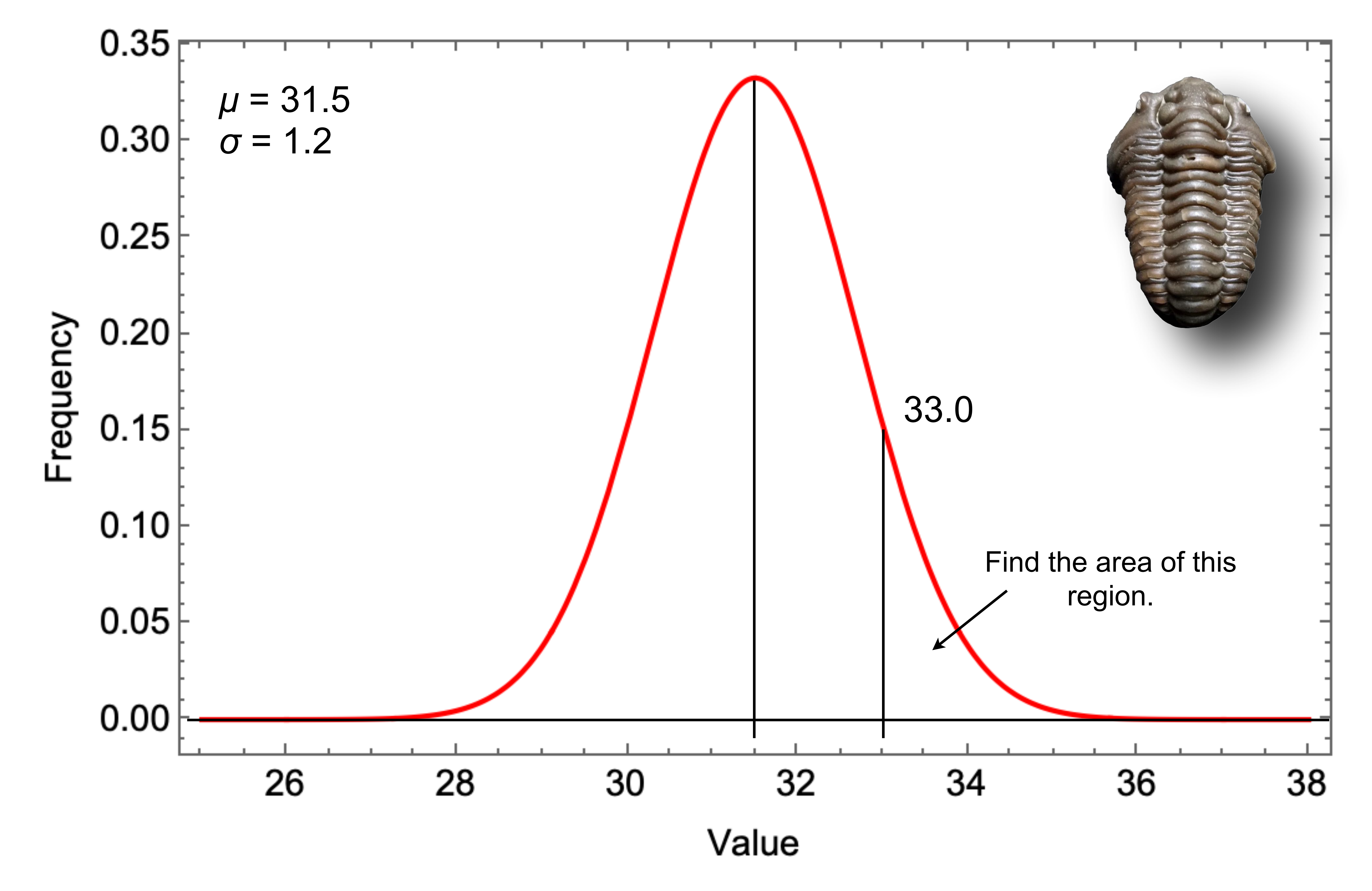
We have a sample of 48 trilobites that we have obtained length data from.

μ = 31.5 mm
σ = 1.2 mm

If we assume the distribution of trilobite lengths is normal, what is the probability we will find specimens whose lengths are ≥ 33.0 mm?

Statistics as Hypothesis Testing

An Example



One-Tailed Hypothesis Test (High)

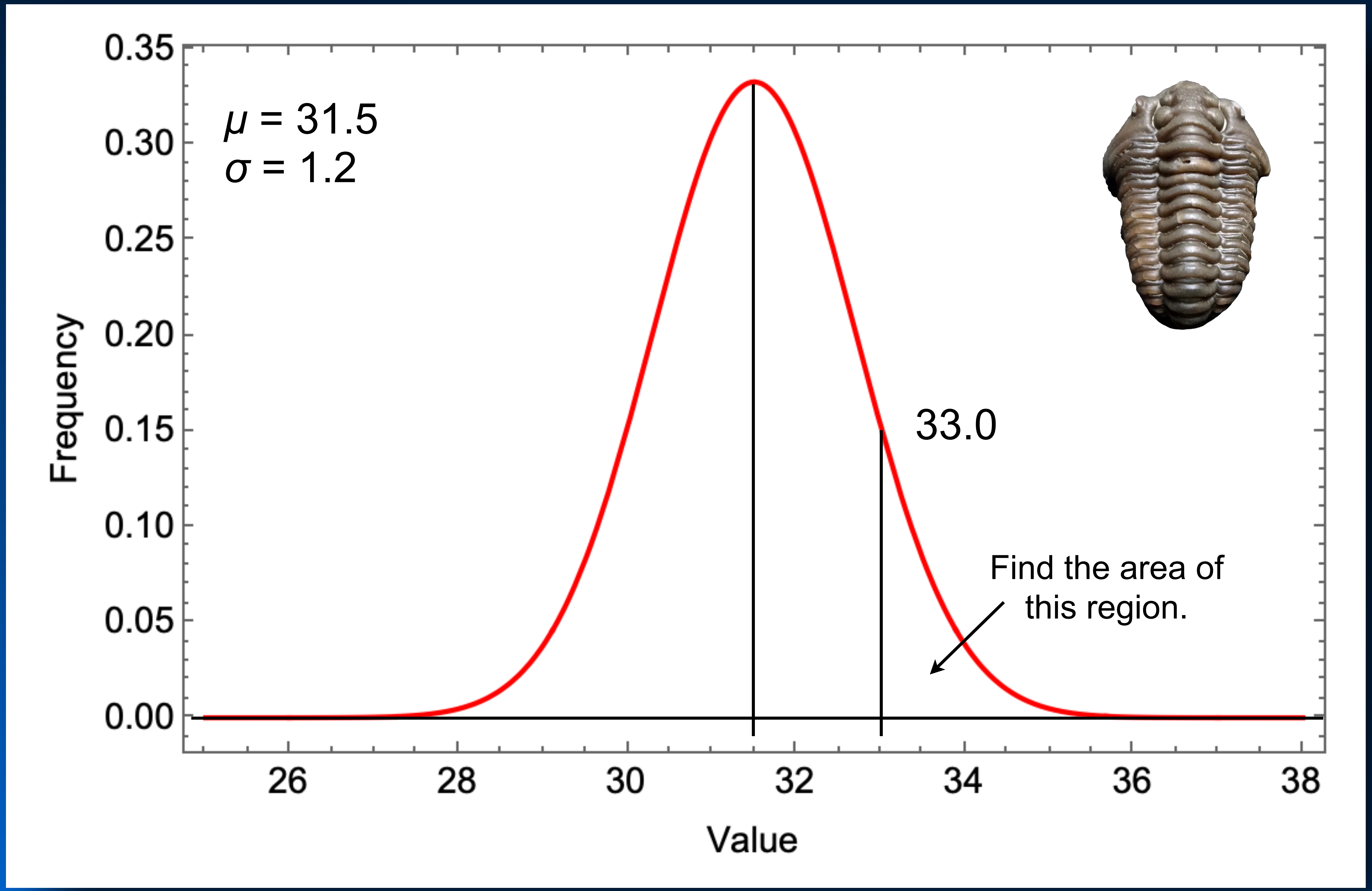
Statistics as Hypothesis Testing

An Example

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{33.0 - 31.5}{1.2}$$

$$z = 1.25$$



Statistics as Hypothesis Testing

An Example

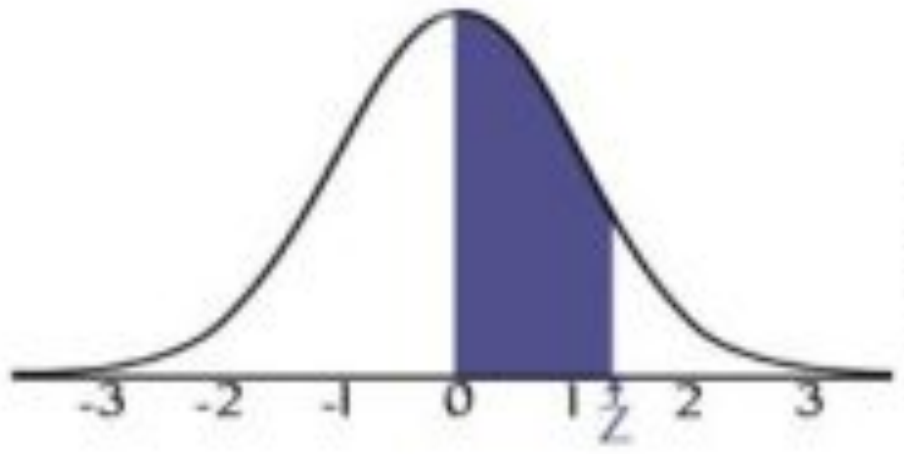
$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{33.0 - 31.5}{1.2}$$

$$z = 1.25$$

$$1.25 = 0.5 - 0.3944$$

$$1.25 = 0.1056 = 10.56\,\%$$



STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and *z* standard deviations above the mean. For example, for *z* = 1.25 the area under the curve between the mean (0) and *z* is 0.3944.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545

Statistics as Hypothesis Testing

An Example

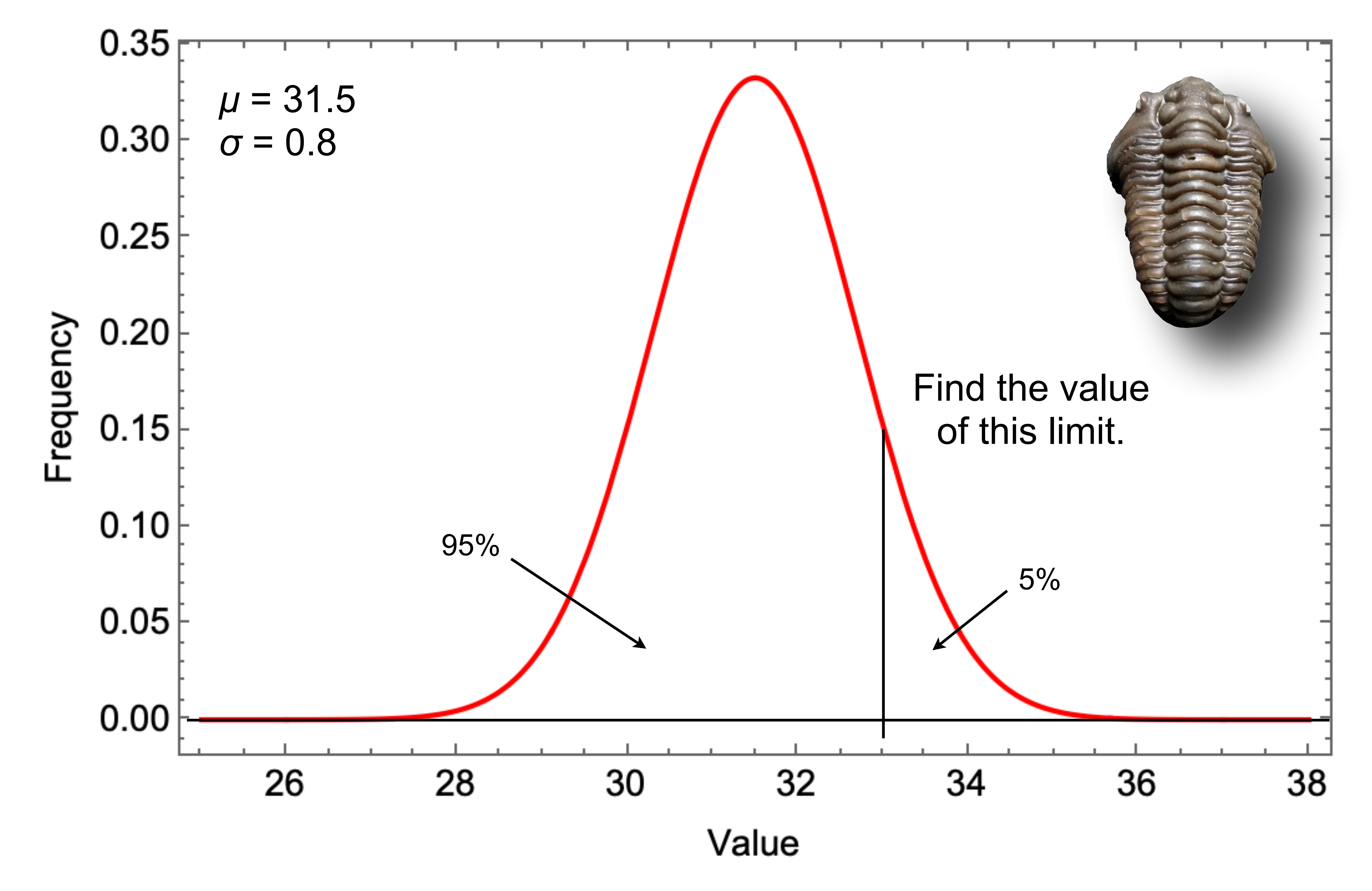
We have a sample of 48 trilobites that we have obtained length data from.

μ = 31.5 mm
σ = 0.8 mm

If we assume the distribution of trilobite lengths is normal, what is the limit beyond which we be 95% confident of not expecting to find trilobites of this species?

Statistics as Hypothesis Testing

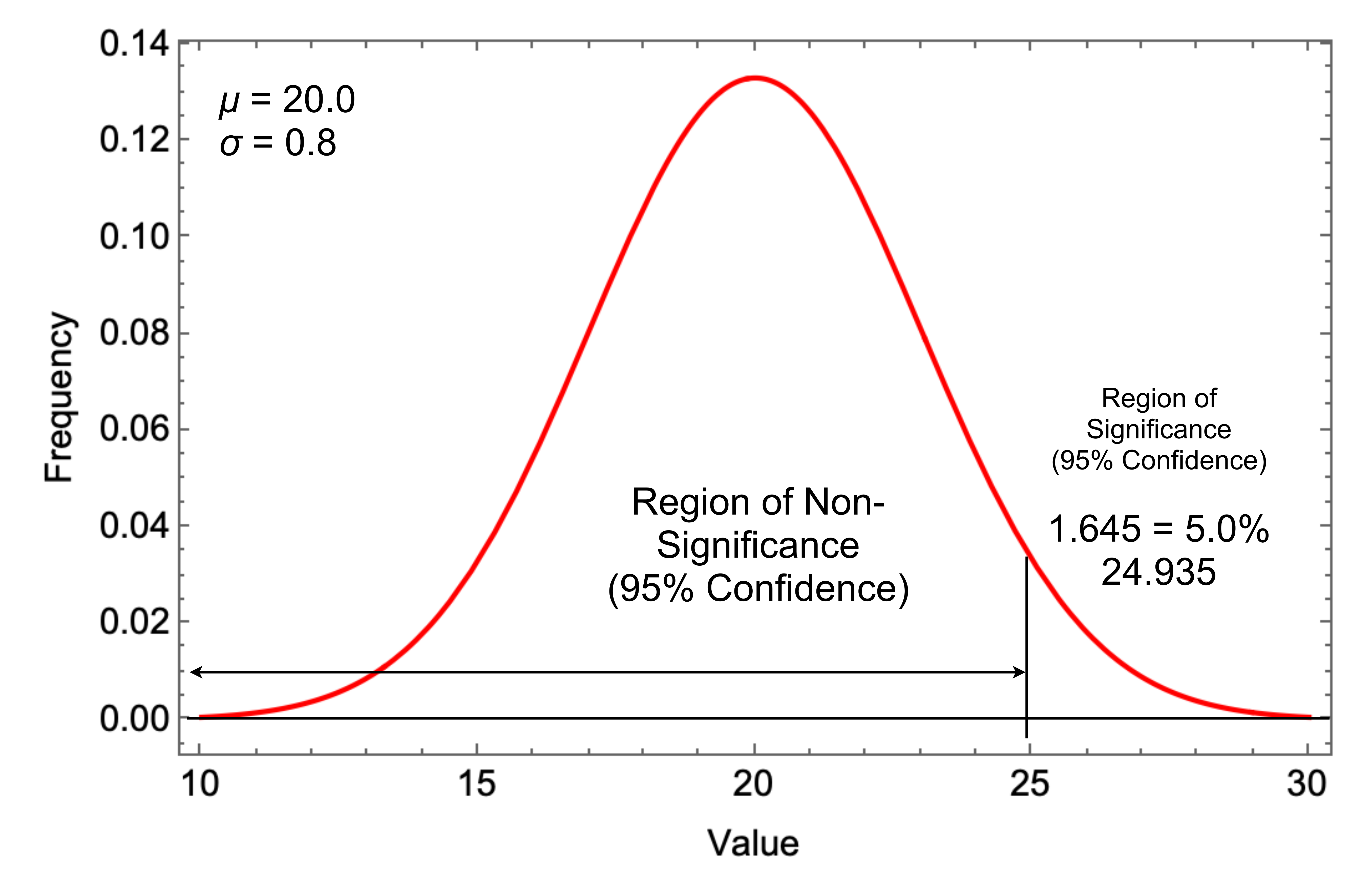
An Example



One-Tailed Hypothesis Test (High)

Statistics as Hypothesis Testing

The Normal Distribution



One-Tailed Hypothesis Test (High)

Statistics as Hypothesis Testing

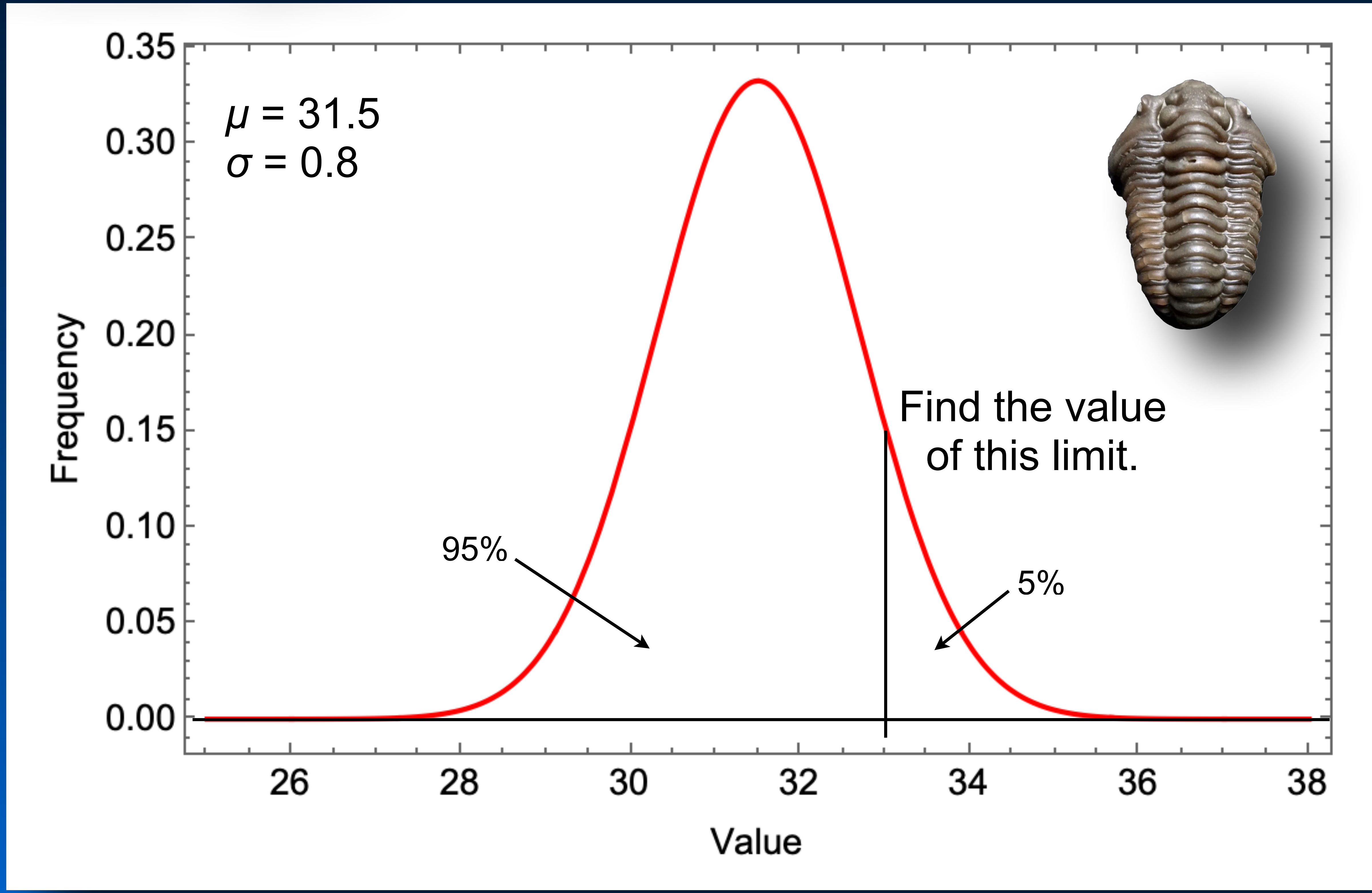
An Example

$$z = \frac{x - \mu}{\sigma}$$

$$1.645 = \frac{x - 31.5}{0.8}$$

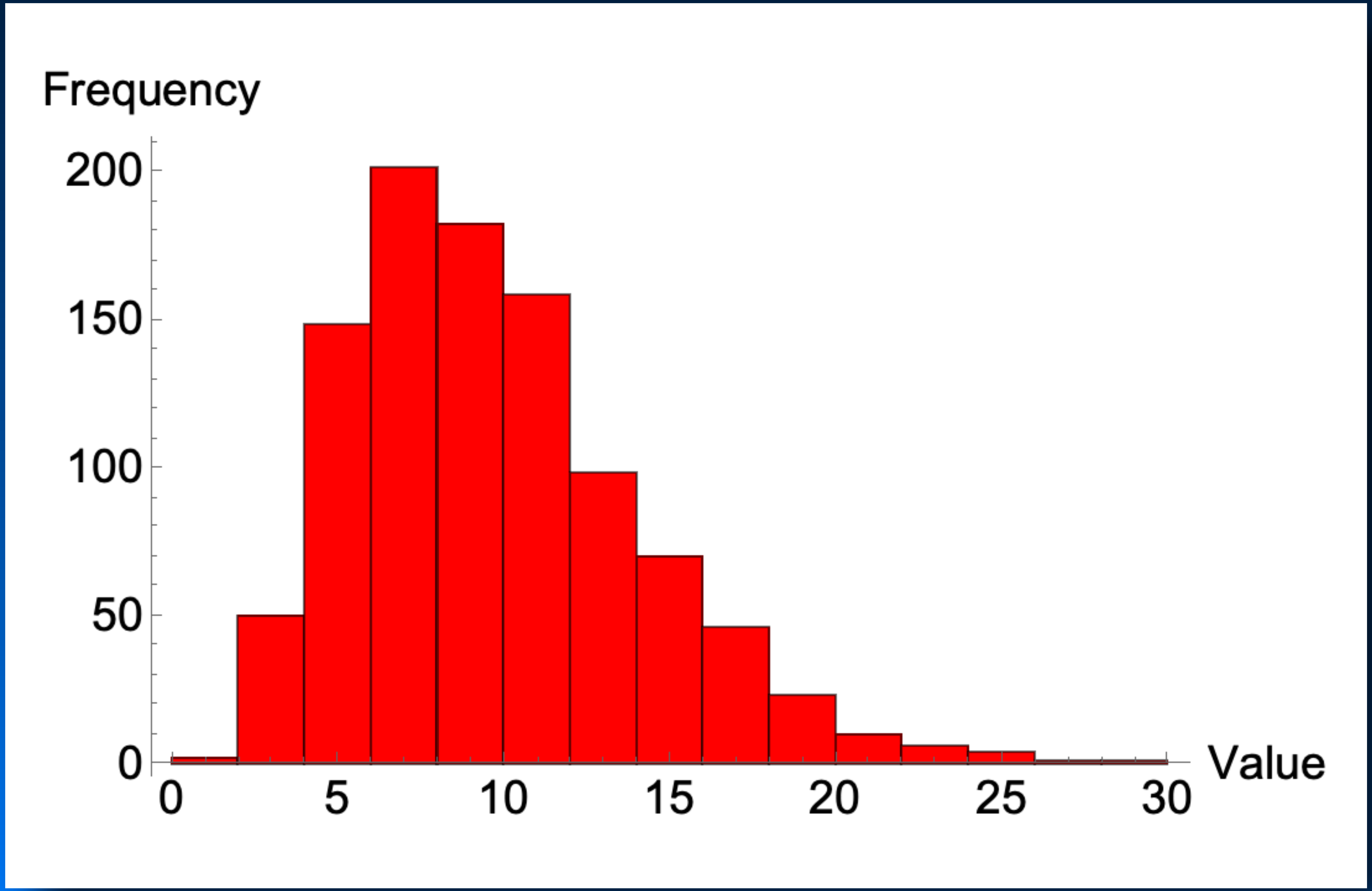
$$x = (1.645 \cdot 0.8) + 31.5$$

$$x = 32.816$$

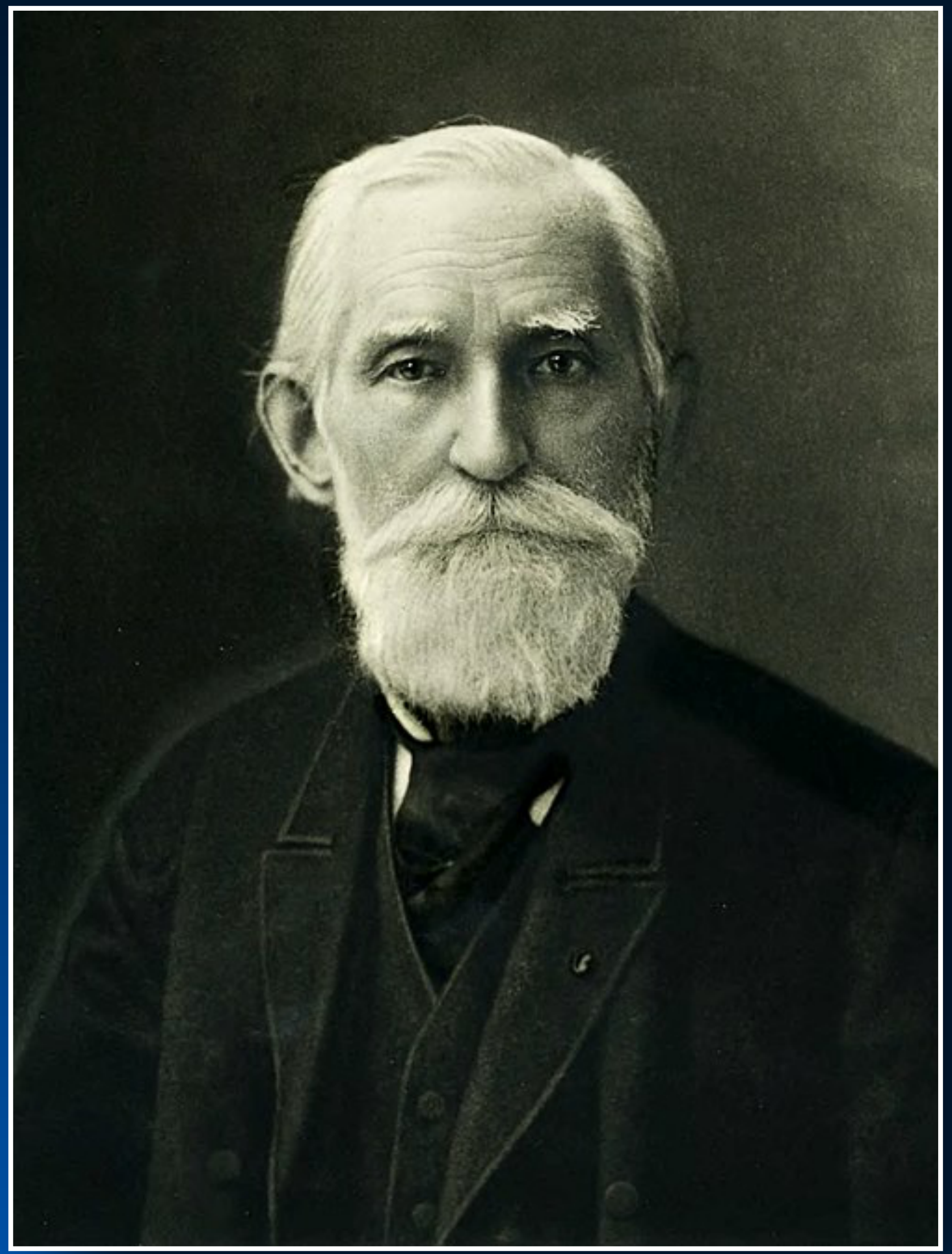


Statistics as Hypothesis Testing

What About Non-Normal Data?



Statistics as Hypothesis Testing



Pafnuty Chebyshev
(1821-1894)

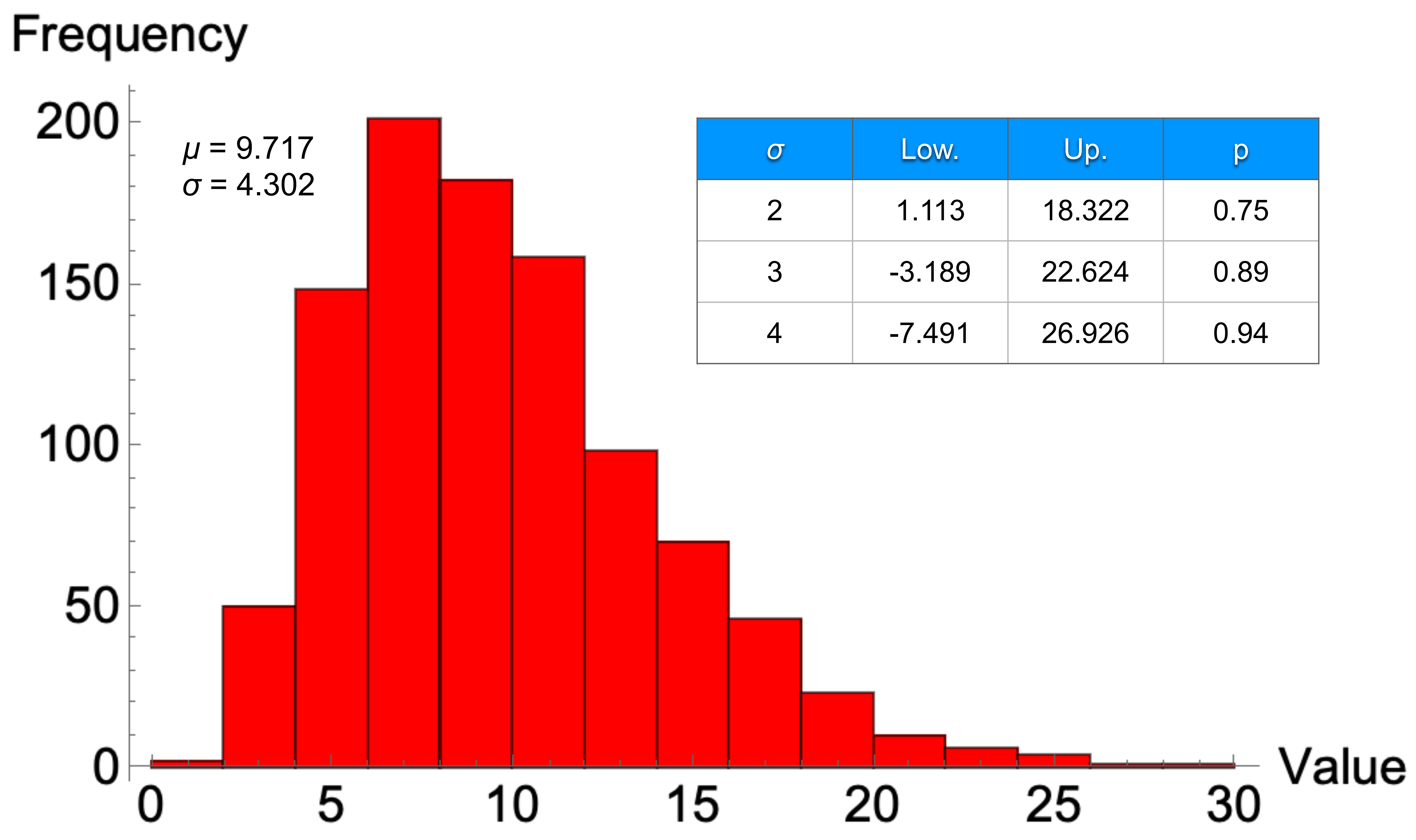
Chebychev's Theorem

For any distribution regardless of its shape, for intervals about the mean > 1 at least $1-1/\kappa^2$ of the values will lie within κ standard deviations of the mean.

$$P(|x-\mu| < \kappa\sigma > 1 - \frac{1}{\kappa^2}$$

Statistics as Hypothesis Testing

What About Non-Normal Data?



σ	Low.	Up.	p
2	1.113	18.322	0.75
3	-3.189	22.624	0.89
4	-7.491	26.926	0.94

Statistics & Probability for Earth Scientists - A Review?

Prof. Norman MacLeod
School of Earth Sciences & Engineering, Nanjing University