Statistics: Distribution – Normal

Nazgul Rakhimzhanova

IITU 2021

COURSE SCHEDULE

week	Mid Term (weeks 01-07)	End Term (weeks 08-14)	week
01	Intro: Data Science Area and open source tools for Data Science	Statistics: Distribution – Lognormal, Exponential	08
02	NumPy package for data science	Sampling and Estimation	09
03	Pandas package for data science	Correlation and Covariance	10
04	Visualization with matplotlib	Hypothesis testing	11
05	Statistics: Distribution – Normal	Decision Tree	12
06	Exploratory Data Analysis (EDA)	Linear Regression	13
07	**Summary for 6 weeks QA session**	**Summary for 6 weeks QA session**	**14**
15	**Course summary**		

OUTLINE

- Previously
- Distributions
- Statistics
- Readings

PREVIUOSLY

- Discussed about **data visualization**
- Did overview of the **matplotlib package**
- Practiced **matplotlib** functions

PREVIUOSLY

We have practiced to choose the visualization method and visualize the data using matplotlib library.

-Line plot

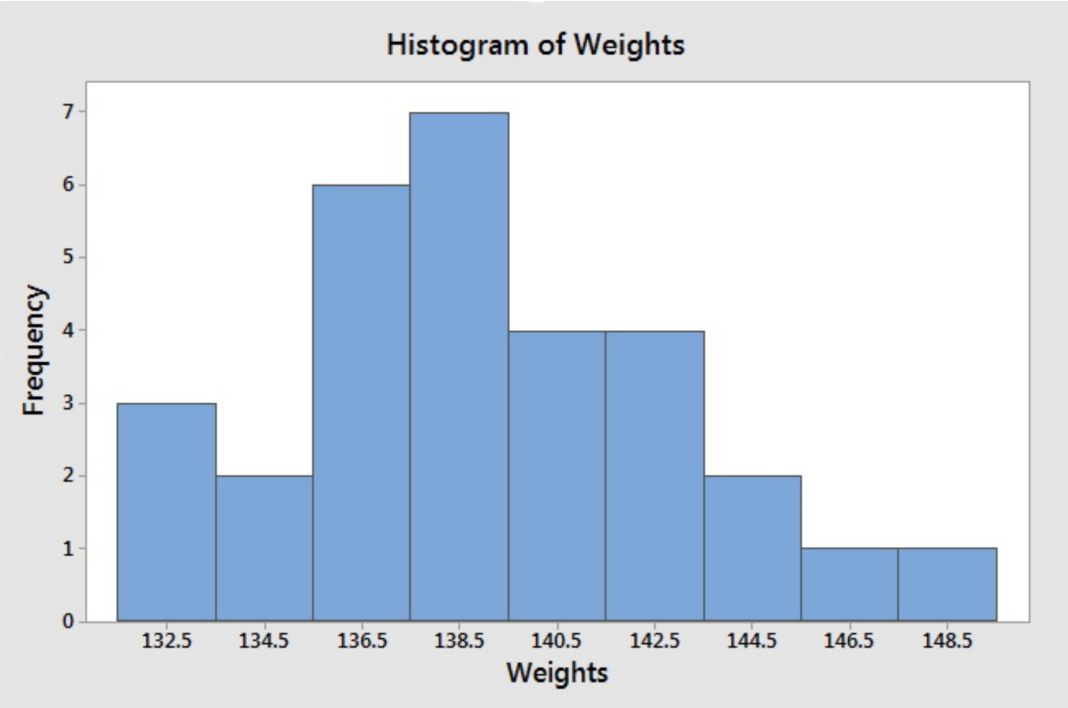
-Bar plot

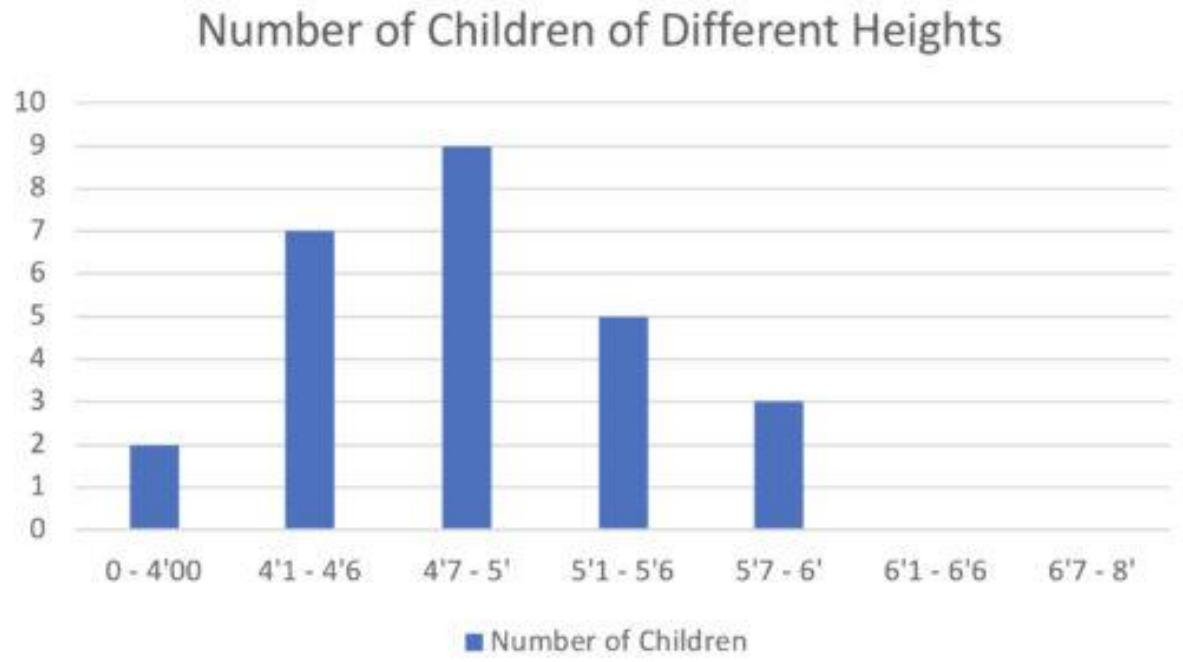
-Scatter plot

-Histogram plot

-Pie plot

HISTOGRAM VS. BAR PLOT





DISTRIBUTION

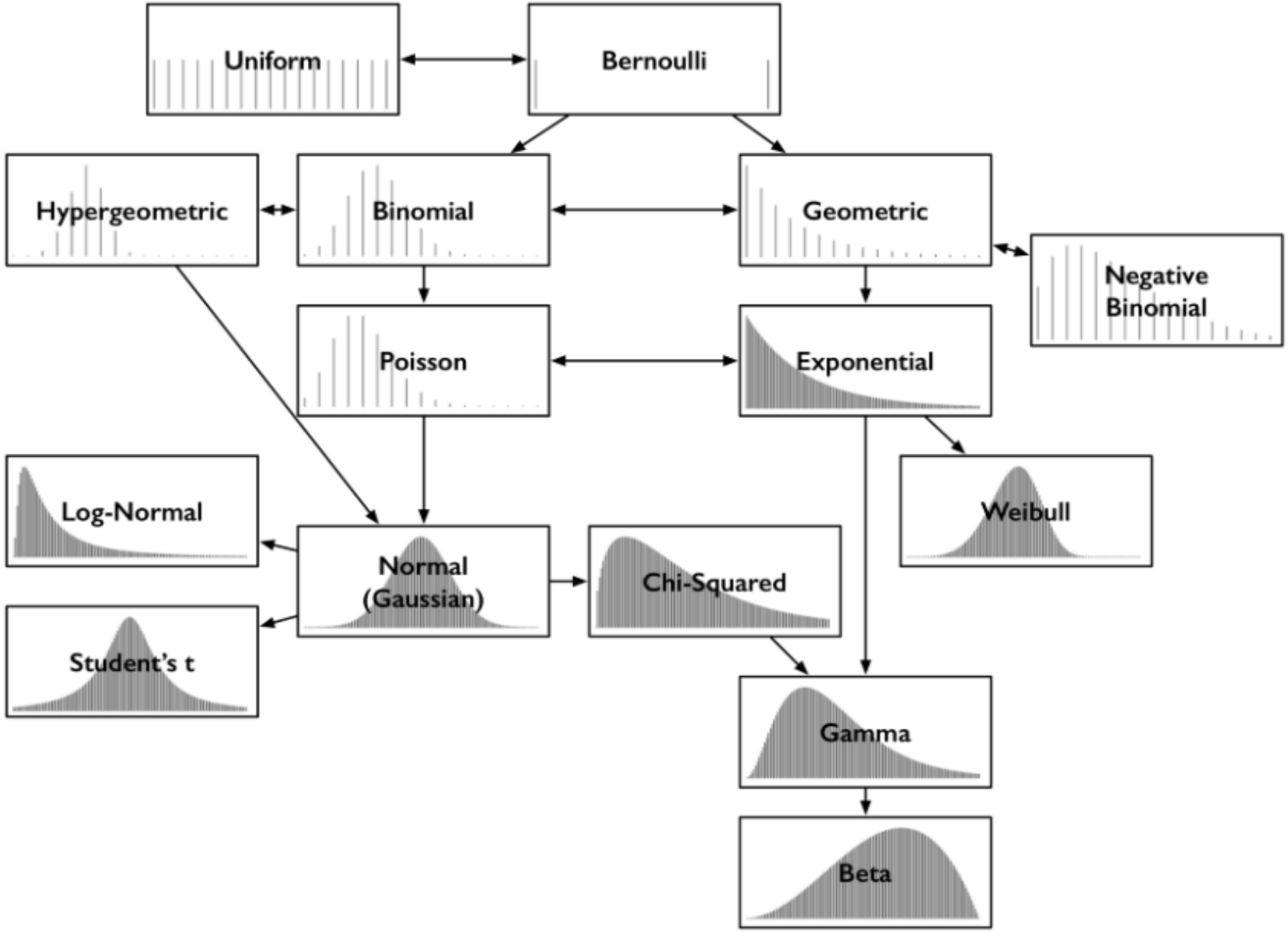
- One of the best ways to **describe a variable is to report the values** that appear in the dataset and how many times each value appears.
- This description is called the distribution of the variable.

DISTRIBUTION

- In statistics, a **frequency distribution is a list**, table or graph that displays the frequency of various outcomes in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval.

- In probability theory and statistics, **a probability distribution** is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).

DISTRIBUTION

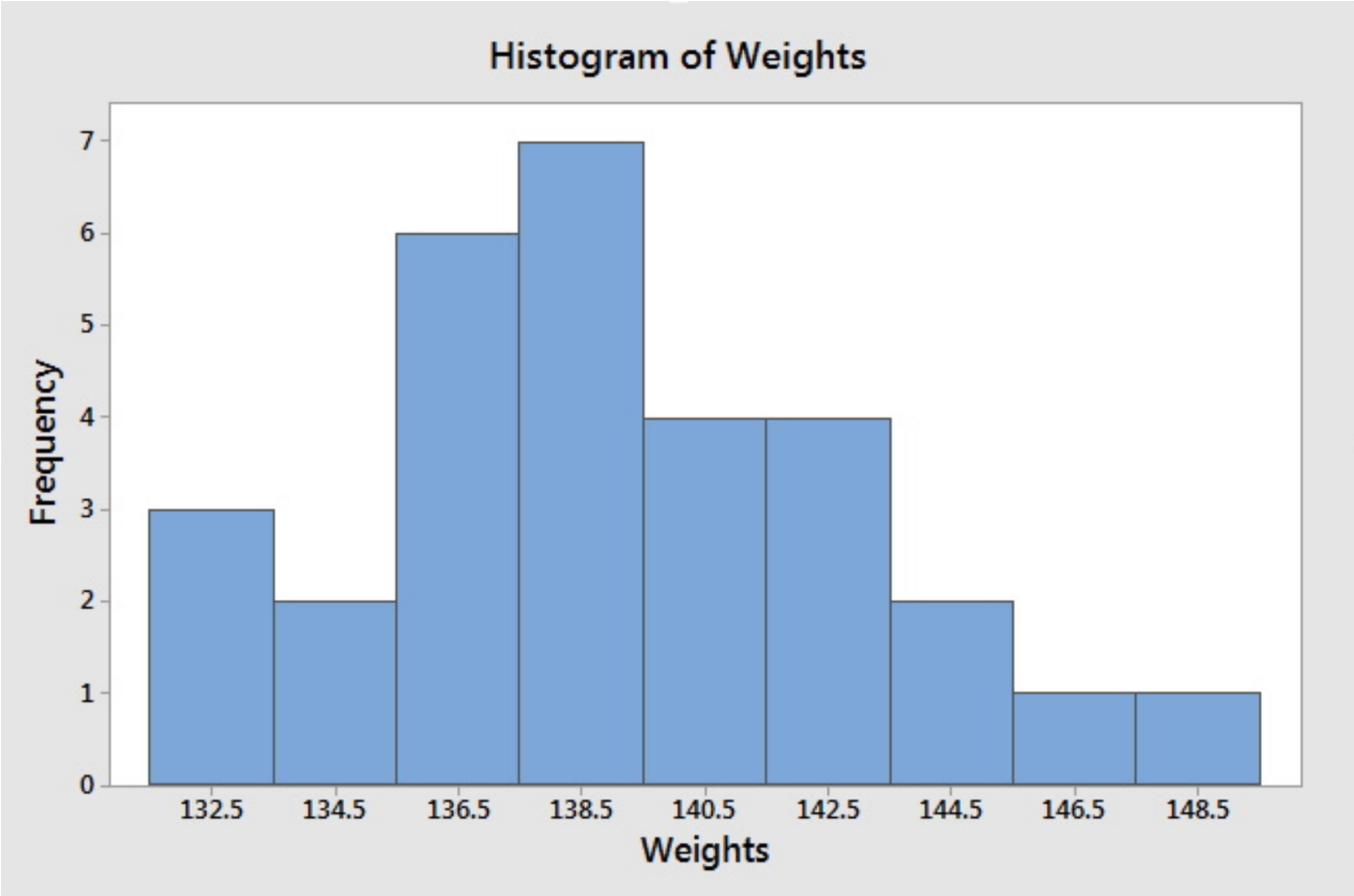


Common probability distributions and some key relationships

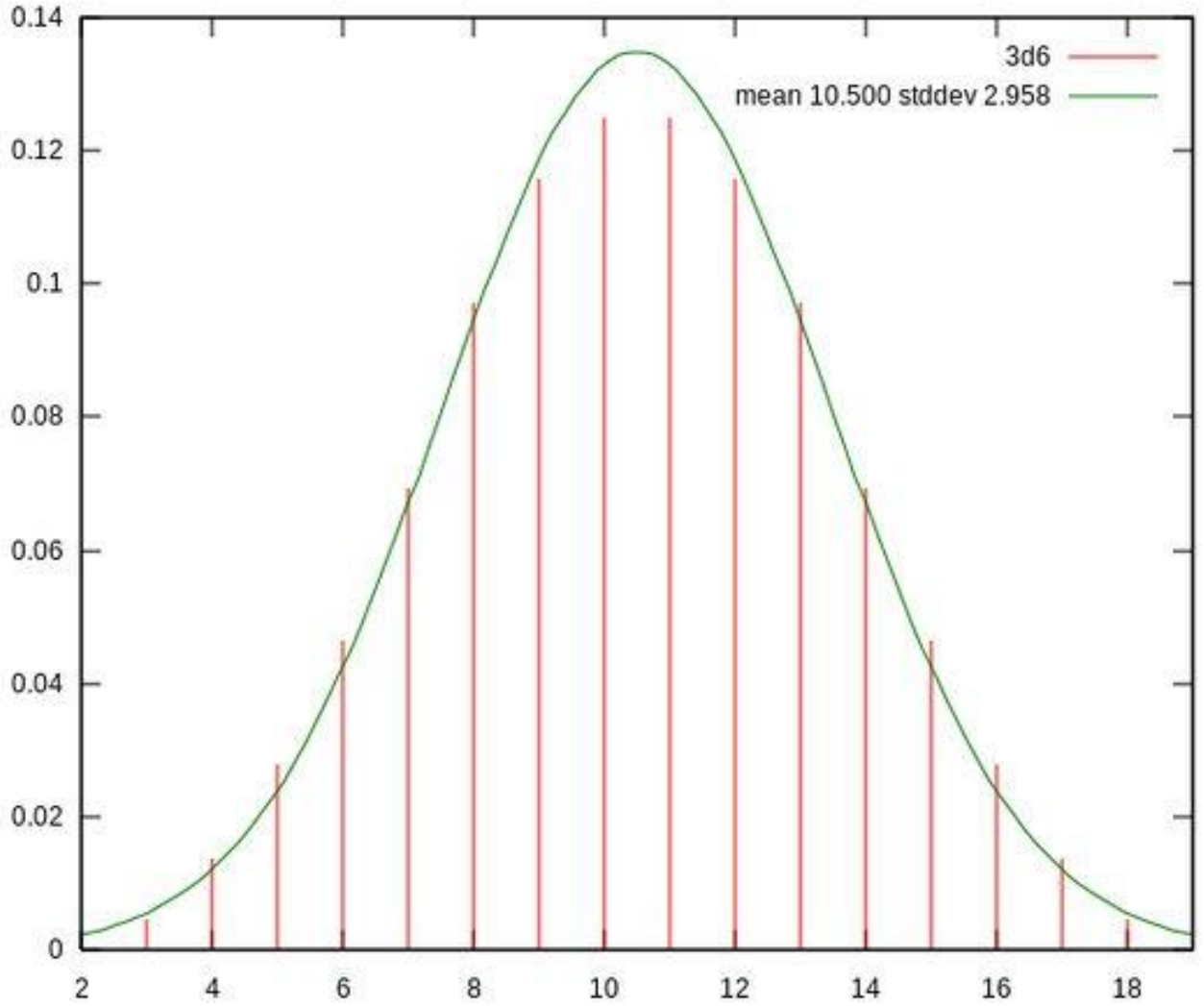
DISTRIBUTION

- Frequency
- Probability mass function
- Probability Density Function
- Cumulative Distribution Function
- Kernel Density Estimation

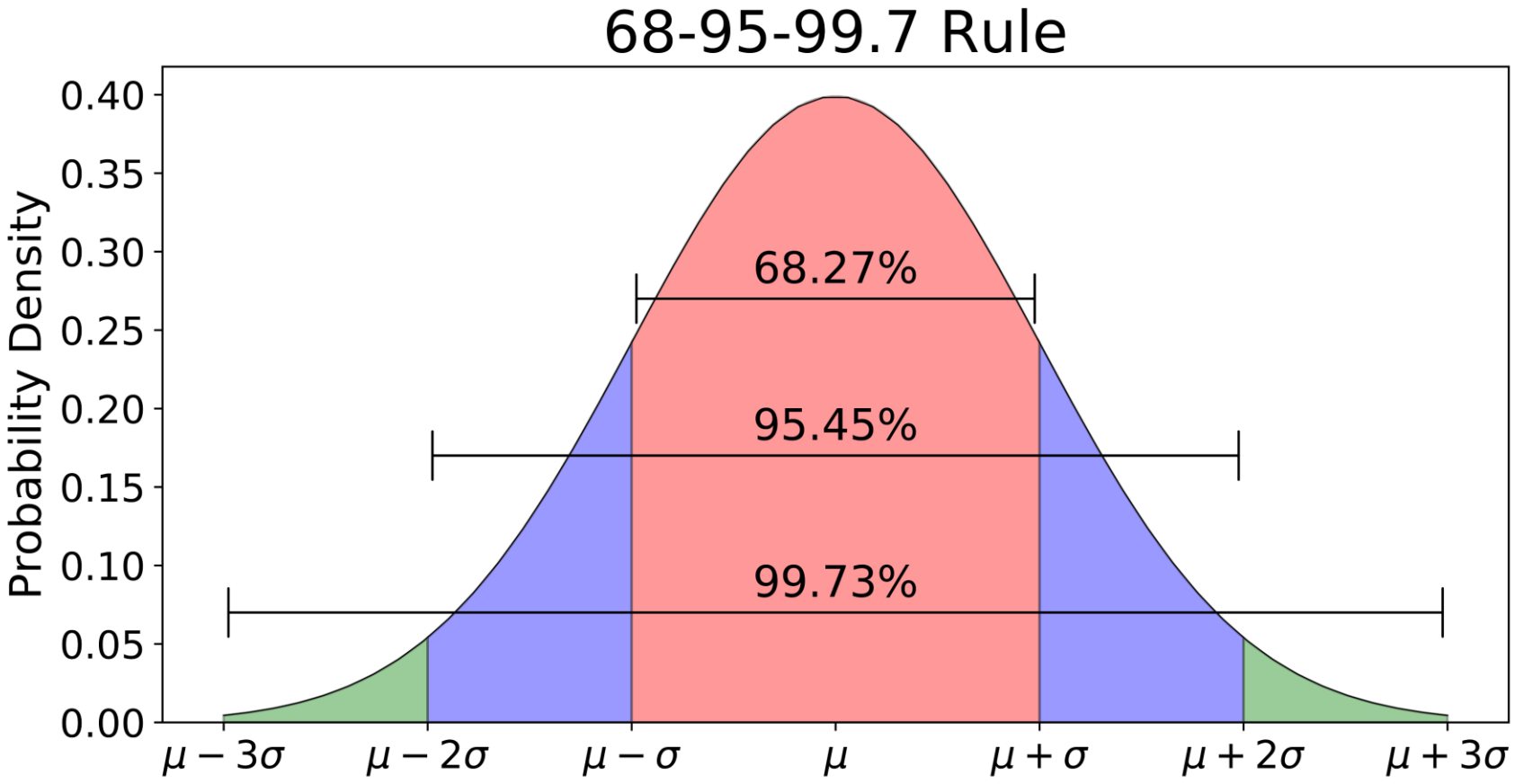
FREQUENCY


Histogram of Weights
Frequency
0
1
2
3
4
5
6
7
132.5
134.5
136.5
138.5
140.5
142.5
144.5
146.5
148.5
Weights

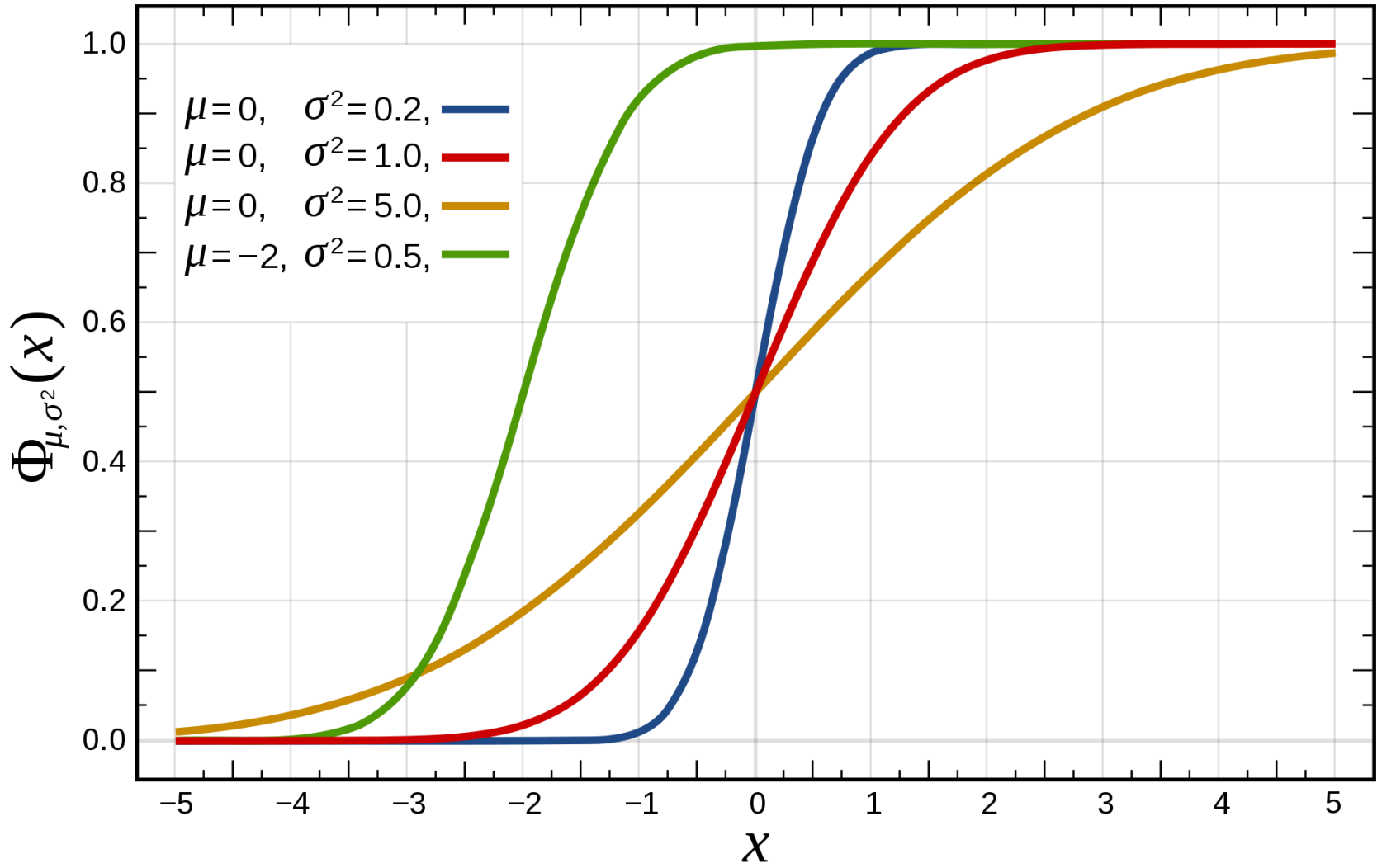
PROBABILITY MASS FUNCTION



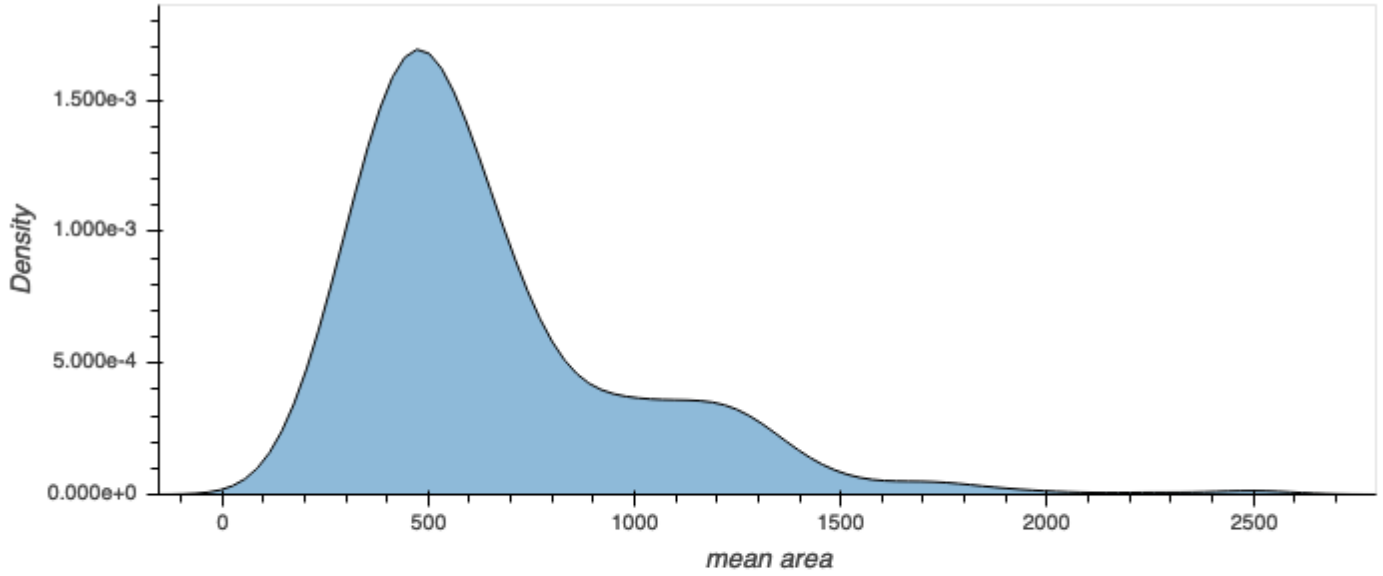
PROBABILITY DENSITY FUNCTION

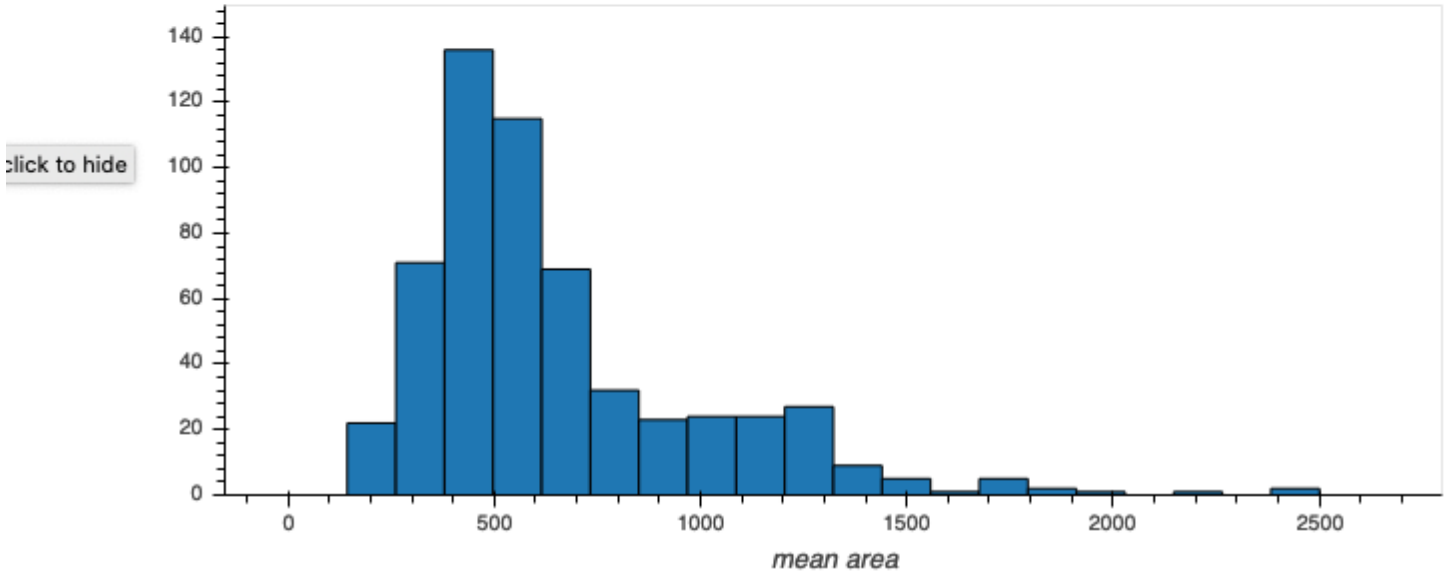


CUMULATIVE DISTRIBUTION FUNCTION



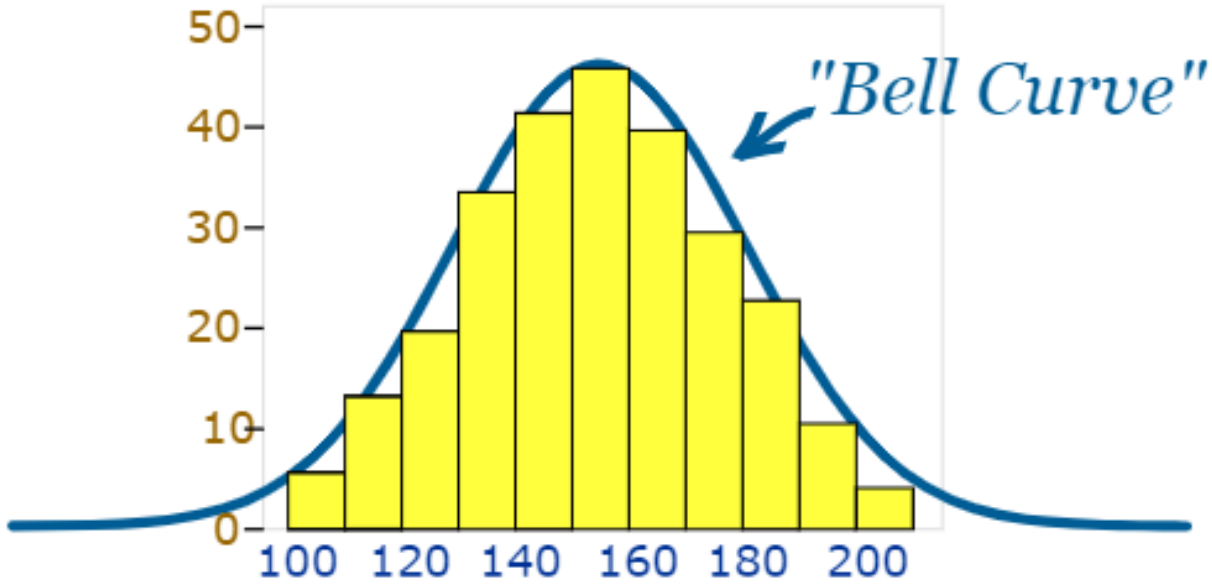
KERNEL DENSITY ESTIMATION





Normal or Gaussian Distribution

Normal or Gaussian Distribution



A Normal Distribution

STATISTICS

- Problem with unreliable statements/data/evidences:
  - Small number of observations
  - Selection bias
  - Confirmation bias
  - Inaccuracy

STATISTICS: terminology

1.population: A group we are interested in studying. "Population" often refers to a group of people, but the term is used for other subjects, too

2.sample: The subset of a population used to collect data

3.representative: A sample is representative if every member of the population has the same chance of being in the sample

STATISTICS

1. **Data** location
2. **Describe** variability and dispersion
3. **Range**
4. **Percentile,** IQR
5. **Data** distribution visualization

STATISTICS

- Basic step -> finding out **where** does **data located**, i.e., its **central tendency**
  - Mean
  - Weighted mean
  - Median
  - Weighted median
  - Trimmed mean

MEAN

If you have a sample of n values, x_i, the mean, μ, is the sum of the values divided by the number of values; in other words

$$\mu = \frac{1}{n}\sum_i x_i$$

MEAN vs. AVERAGE

- The "mean" of a sample is the summary statistic computed with the previous formula.
- An "average" is one of many summary statistics you might choose to describe the typical value or the central tendency of a sample.

MEAN vs. AVERAGE

- Sometimes the mean is a good description of a set of values. For example, apples are all pretty much the same size (at least the ones sold in supermarkets). So if I buy 6 apples and the total weight is 3 pounds, it would be a reasonable summary to say they are about a half pound each.
- But pumpkins are more diverse. Suppose I grow several varieties in my garden, and one day I harvest three decorative pumpkins that are 1 pound each, two pie pumpkins that are 3 pounds each, and one Atlantic Giant® pumpkin that weighs 591 pounds. The mean of this sample is 100 pounds, but if I told you "The average pumpkin in my garden is 100 pounds," that would be wrong, or at least misleading.
- In this example, there is no meaningful average because there is no typical pumpkin

MEAN vs. AVERAGE

WEIGHTED MEAN

- A weighted mean is a **kind of average**. Instead of each data point contributing equally to the final mean, some data points contribute more "**weight**" than others. If all the weights are equal, then the weighted mean equals the arithmetic mean.

$$\bar{x} = \frac{\sum_{i=1}^{n}(x_i \times w_i)}{\sum_{i=1}^{n} w_i}$$

w = the weights

x = the value

WEIGHTED MEAN

Example: Sam wants to buy a new camera, and decides on the following rating system:

- Image Quality **50%**
- Battery Life **30%**
- Zoom Range **20%**

The Sonu camera gets 8 (out of 10) for Image Quality, 6 for Battery Life and 7 for Zoom Range

The Conan camera gets 9 for Image Quality, 4 for Battery Life and 6 for Zoom Range

Which camera is best?

$$\text{Sonu: } 0{,}5 \times 8 + 0{,}3 \times 6 + 0{,}2 \times 7 = 4 + 1{,}8 + 1{,}4 = \mathbf{7{,}2}$$

$$\text{Conan: } 0{,}5 \times 9 + 0{,}3 \times 4 + 0{,}2 \times 6 = 4{,}5 + 1{,}2 + 1{,}2 = \mathbf{6{,}9}$$

Sam decides to buy the Sonu.

TRIMMED MEAN

- A variation of the mean is a trimmed mean, which you calculate by dropping a fixed number of sorted values at each end and then taking an average of the remaining values.

$$\bar{x} = \frac{\sum_{i=p+1}^{n-p}(x_i)}{n - 2p}$$

x = the value

p = trim

TRIMMED MEAN

VARIANCE

- If there is no single number that summarizes pumpkin weights, we can do a little better with two numbers: mean and variance.
- In the same way that the mean is intended to describe the central tendency, variance is intended to describe the **spread**. The variance of a set of values is

$$\sigma^2 = \frac{1}{n}\sum_i (x_i - \mu)^2$$

VARIANCE

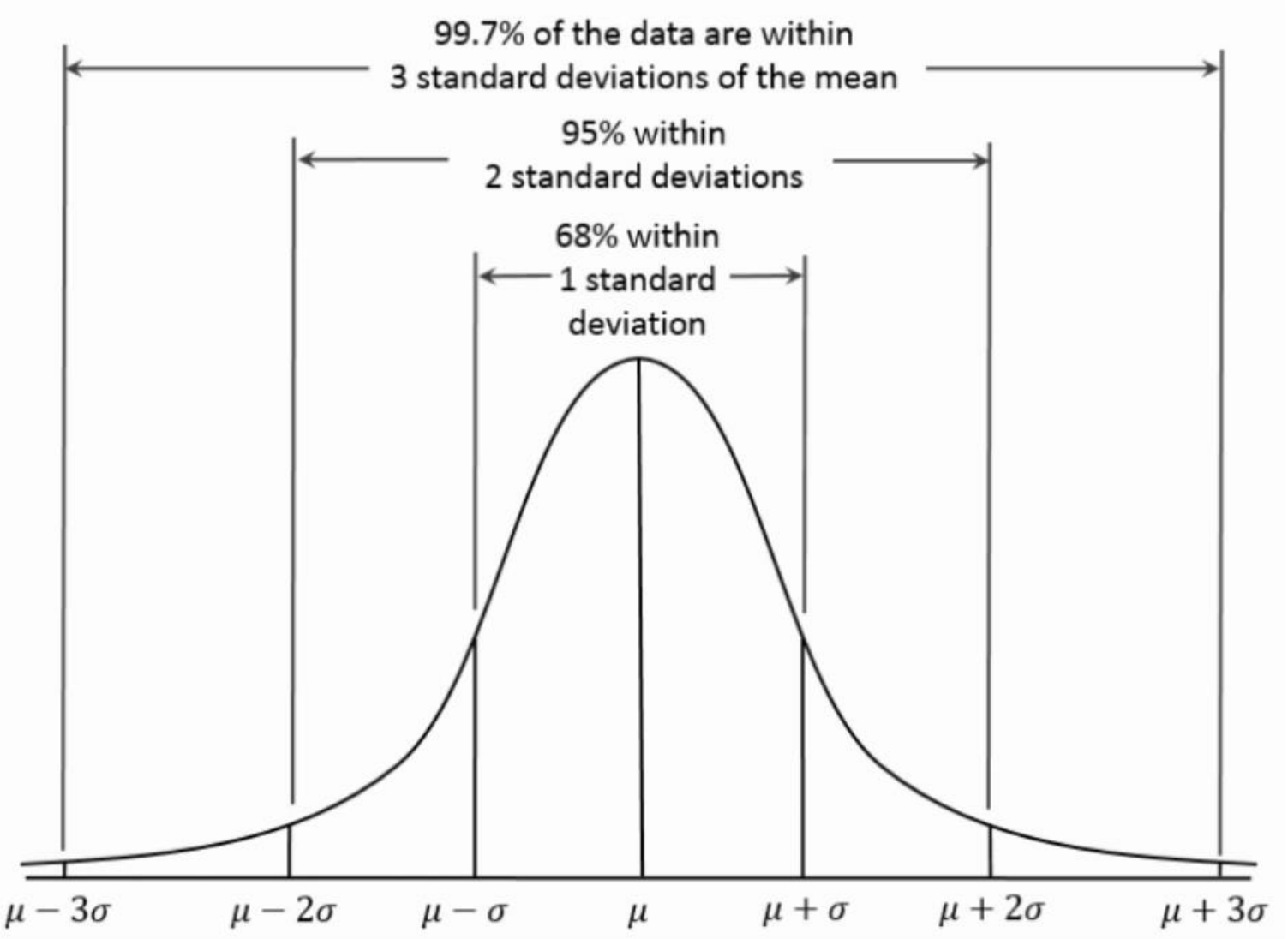
- The term $x_i - \mu$ is called the "deviation from the mean," so variance is the mean squared deviation, which is why it is denoted as σ^2. The square root of variance, σ, is called the **standard deviation**.
- σ^2 – variance
  σ – standard deviation

VARIANCE

- σ^2 – variance
  σ – standard deviation
- By itself, variance is hard to interpret. One problem is that the units are strange; in this case the measurements are in pounds, so the variance is in pounds squared. Standard deviation is more meaningful; in this case the units are pounds.

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

STD

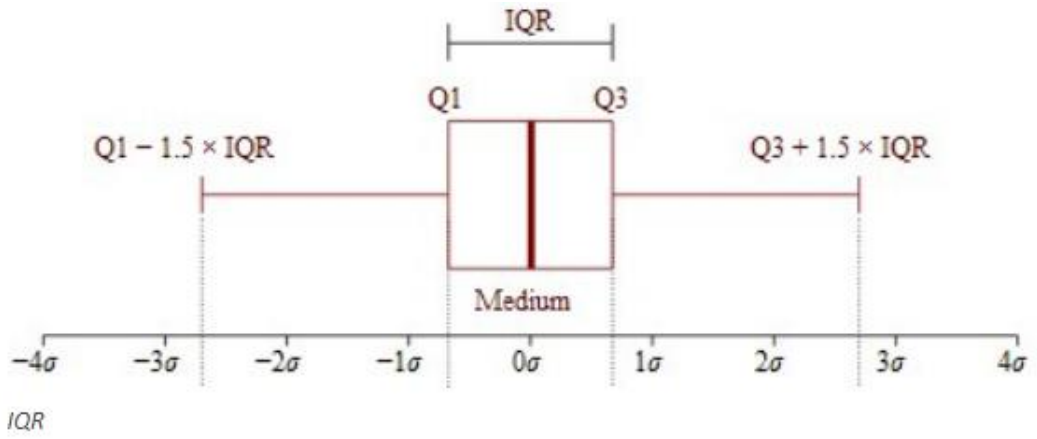

99.7% of the data are within
3 standard deviations of the mean
95% within
2 standard deviations
68% within
1 standard
deviation
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

PERCENTILE, IQR

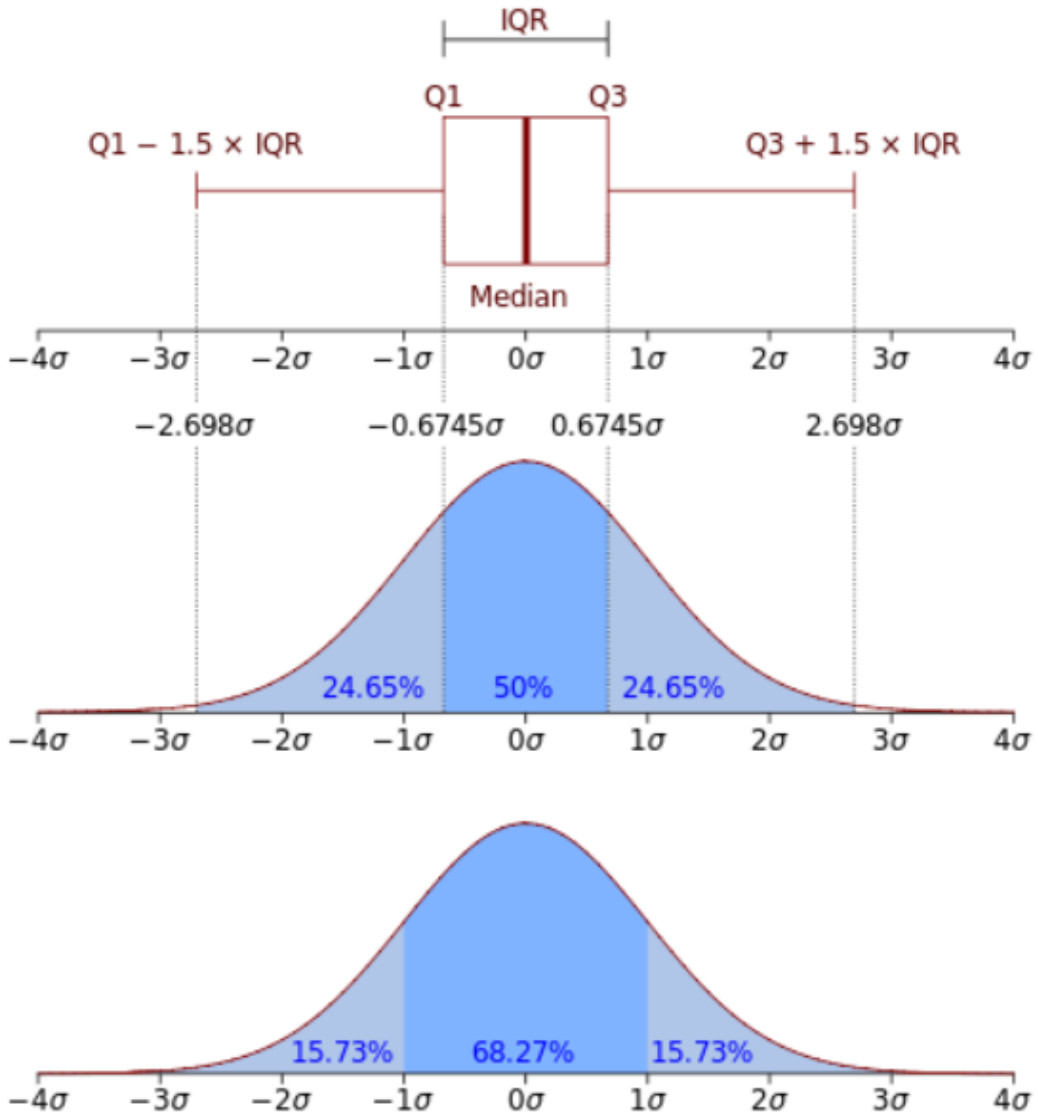
Interquartile range

A percentile is a value below which a certain percentage of observations lie. Quartile – 25%, 50%, 75% percent's

- **1st** quartile, **Q1**, or 25th percentile—the number that separates the lowest 25% of the group from the highest 75% of the group.
- **Median**, or 50th percentile—the number in the middle of the group, when arranged from smallest to largest.
- **3rd** quartile, **Q3**, or 75th percentile—the number that separates the lowest 75% of the group from the highest 25% of the group.



PERCENTILE, IQR

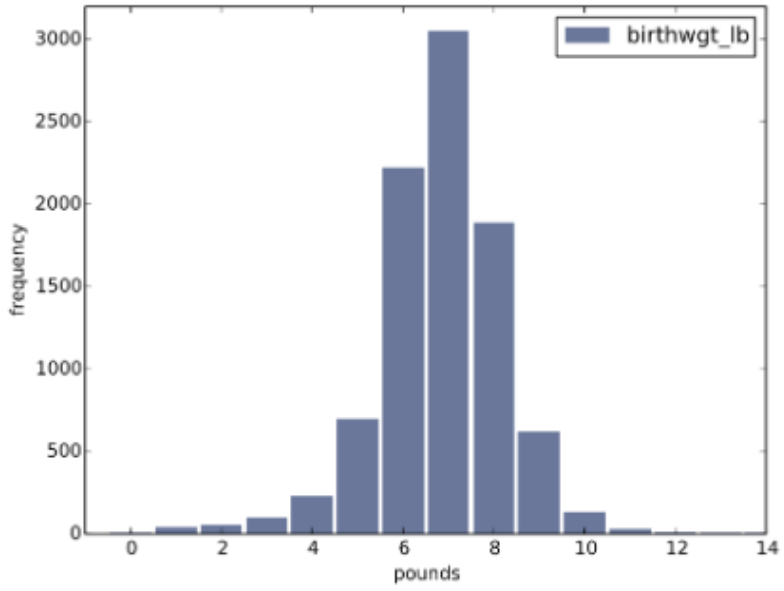


QUARTILES

- **Quartiles** are **used to summarize a group** of numbers. Instead of looking a big list of numbers (way too unwieldy!), you are looking at just a few numbers that give you a picture of what's going on in the big list. Quartiles are great for reporting on a set of data and for making box and whisker plots.
- **Quartiles** are especially useful when you're working with data that **isn't symmetrically distributed**, or a data set that has outliers.

DISTRIBUTION

The most common representation of a distribution is a histogram, which is a graph that shows the frequency of each value. In this context, "frequency" means the number of times the value appears in a dataset.

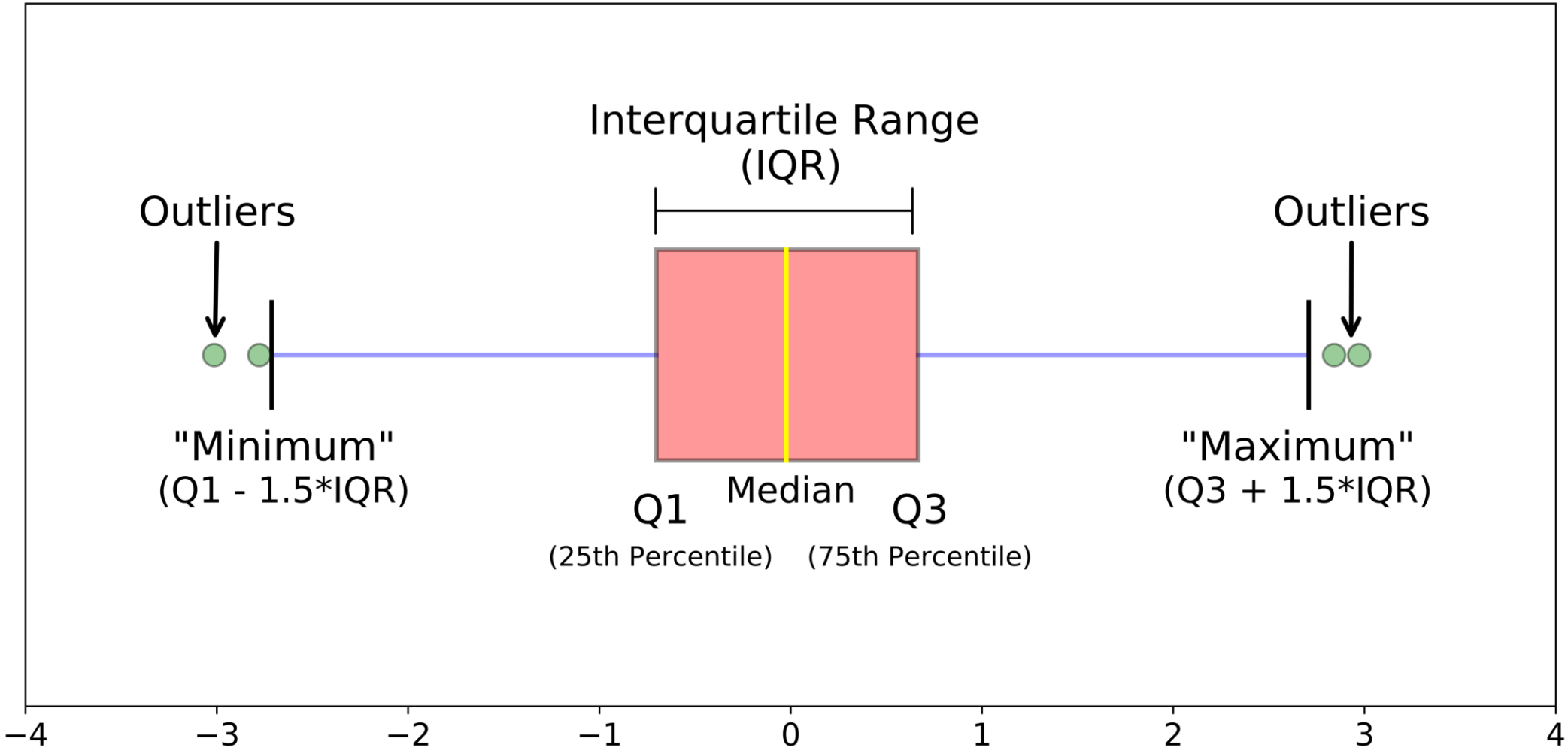


OUTLIERS

Outliers, are extreme values that might be errors in measurement and recording, or might be accurate reports of rare events.

In the list of pregnancy lengths for live births, the 10 lowest values are [0, 4, 9, 13, 17, 18, 19, 20, 21, 22]. Values below 10 weeks are certainly errors; the most likely explanation is that the outcome was not coded correctly. Values higher than 30 weeks are probably legitimate. Between 10 and 30 weeks, it is hard to be sure; some values are probably errors, but some represent premature babies.

OUTLIERS


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
"Maximum"
(Q3 + 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
−4
−3
−2
−1
0
1
2
3
4

DISTRIBUTION

When you start working with a new dataset, I suggest you explore the variables you are planning to use one at a time, and a good way to start is by looking at histograms.

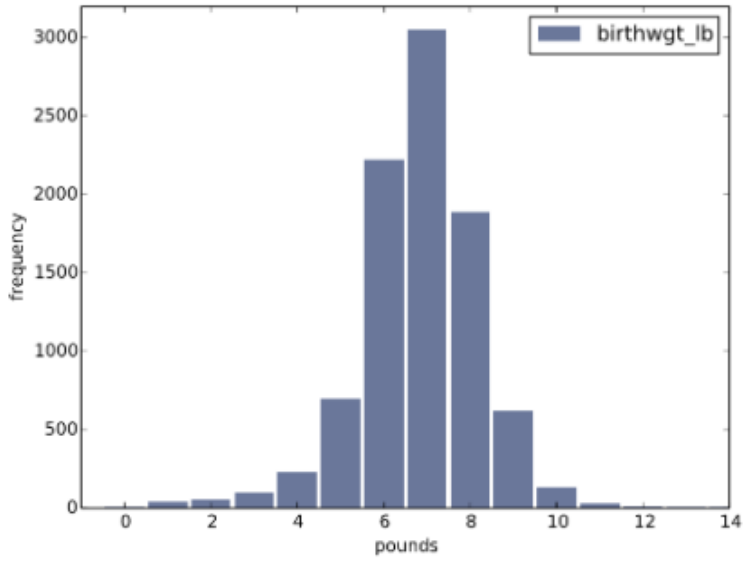


Figure 2.1: Histogram of the pound part of birth weight

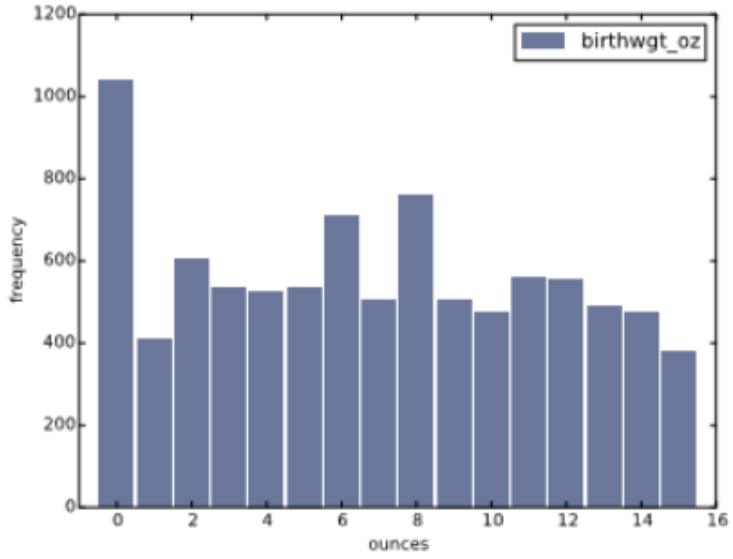
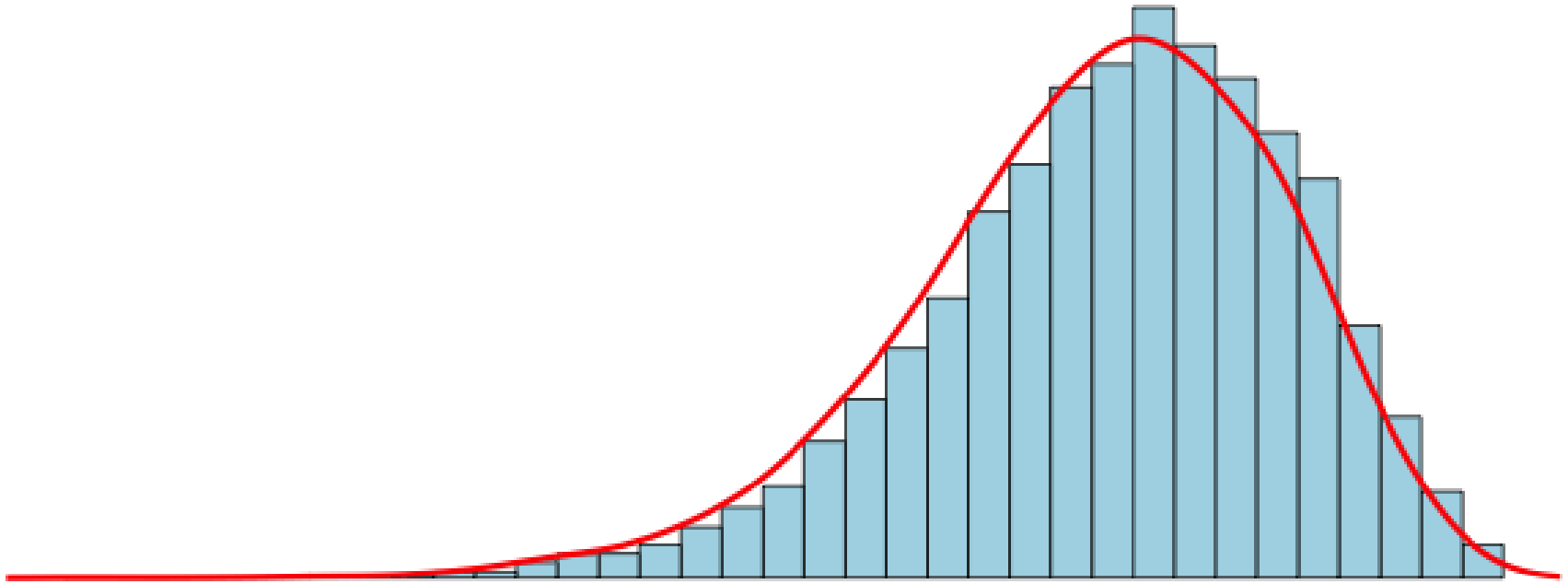
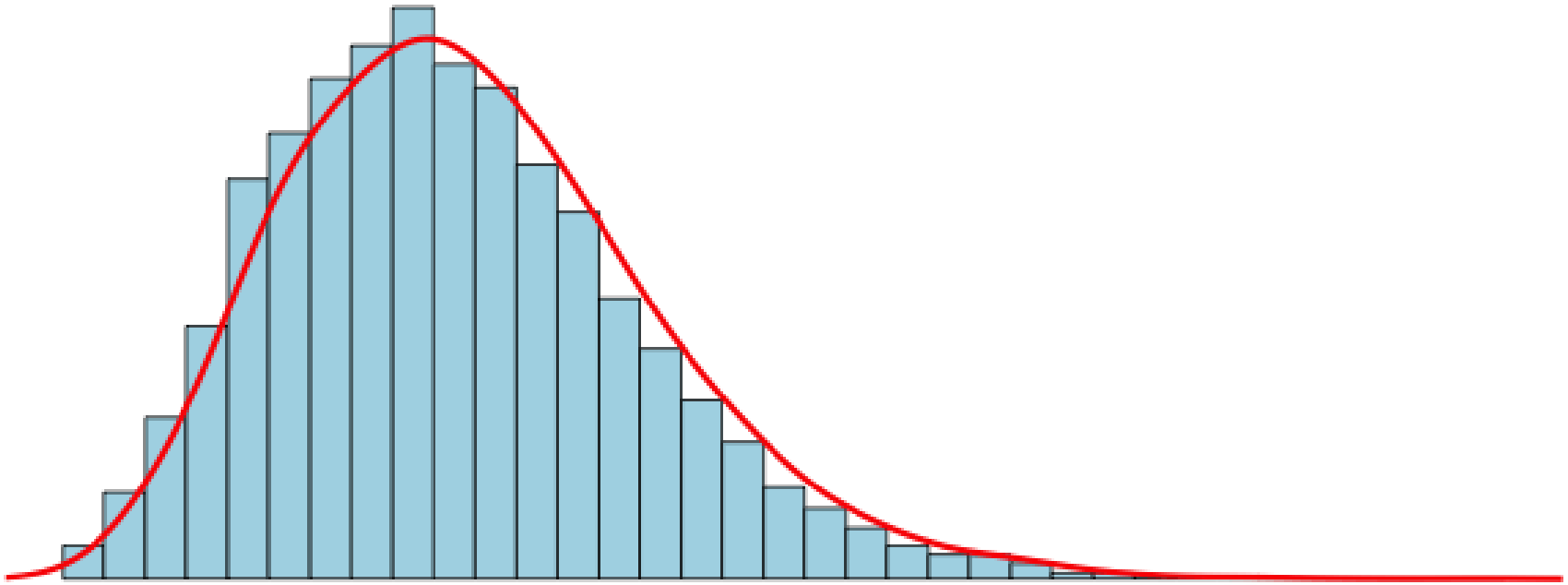


Figure 2.2: Histogram of the ounce part of birth weight.

NEGATIVE SKEWED DISTRIBUTION



POSITIVE SKEWED DISTRIBUTION



LOOK BACK to STATISTICS

- **Probability theory** — the mathematical foundation for statistics — was developed in the 17th to 19th centuries based on work by **Thomas Bayes, Pierre-Simon Laplace,** and **Carl Gauss.**
- **Modern statistics** as a rigorous scientific discipline traces its roots back to the late 1800s and **Francis Galton** and **Karl Pearson.**
- The field of **exploratory data analysis** was established with **Tukey's 1977** now-classic book Exploratory **Data Analysis.**

Readings

- **Practical Statistics for DS**
- https://medium.com/@srowen/common-probability-distributions-347e6b945ce4
- https://mathisonian.github.io/kde/
- https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51