# Probability and Statistics

Nels Grevstad

Metropolitan State University of Denver

*ngrevsta@msudenver.edu*

January 22, 2019

# Topics

# Objectives

Objectives:

- Identify sources of variation in data
- Recognize the four types of variables
- Construct a frequency distribution (table) and histogram

# Introduction (1.1)

- ***Statistics*** is the science of collecting, organizing, analyzing, and drawing conclusions from data.
  - Understanding variability in data
  - Distinguishing signal from noise
  - Assessing the evidence data provides
  - Using data to answer scientific questions

# Introduction (1.1)

- ***Statistics*** is the science of collecting, organizing, analyzing, and drawing conclusions from data.
  - Understanding variability in data
  - Distinguishing signal from noise
  - Assessing the evidence data provides
  - Using data to answer scientific questions
- Variation in data arises from many sources.

## Example

In a study to assess the effect of exercise on cholesterol levels, one group is assigned to an exercise regimen and other isn't. Is cholesterol reduced in exercise group?

Some **sources of variability**:

- People have naturally different cholesterol levels
- Respond differently to same amount of exercise (e.g. genetics)
- May vary in adherence to exercise regimen
- Diet may have an effect

## Example

Surface water specimens taken upstream and downstream of landfill are analyzed for certain toxins. Can these be used to show toxins are leaching from landfill?

Some **sources of variability**:

- Natural variation; where specimens were taken
- Specimens gathered under different conditions
- Lab error or variability in analyzing specimens

- Statistics fits into a broader **scientific process**:
  - Scientific question.
  - Design experiment / study.
  - Perform experiment / collect data.
  - Summarize and analyze data.
  - Draw statistical conclusions
  - Draw scientific conclusions.

- Drawing statistical conclusions may involve making ***statistical inferences*** (conclusions about a **population** based on a **sample**).

- Drawing statistical conclusions may involve making ***statistical inferences*** (conclusions about a **population** based on a **sample**).
- A ***population*** is a large group of individuals about which we're interested.

- Drawing statistical conclusions may involve making ***statistical inferences*** (conclusions about a **population** based on a **sample**).
- A ***population*** is a large group of individuals about which we're interested.
- A ***sample*** is a subset of the population, usually selected randomly.

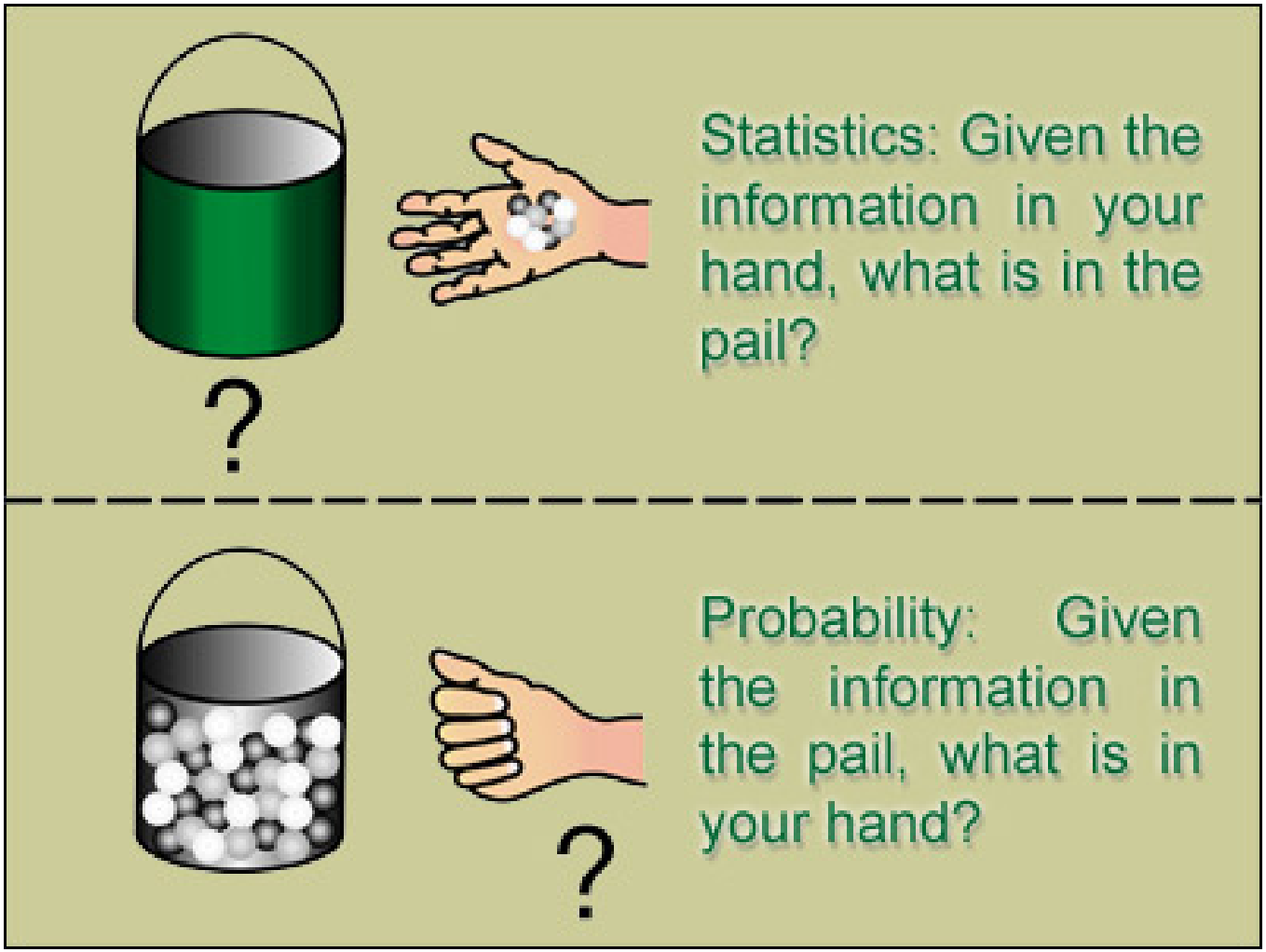
- **Statistical inference** and **probability** are **inverse** inference methods.

- **Statistical inference** and **probability** are **inverse** inference methods.
  - Statistical inference is **inductive** (from the specific to the general).

- **Statistical inference** and **probability** are **inverse** inference methods.
  - Statistical inference is **inductive** (from the specific to the general).
  - Probability is **deductive** (from the general to the specific).

- **Statistical inference** and **probability** are **inverse** inference methods.
  - Statistical inference is **inductive** (from the specific to the general).
  - Probability is **deductive** (from the general to the specific).

  Need to learn the language of probability in order to communicate statistical conclusions.


Statistics: Given the information in your hand, what is in the pail?
?
Probability: Given the information in the pail, what is in your hand?
?

# Variables and Data (1.2)

- ***Data*** are observed values of ***variables*** (characteristics that vary from one individual to the next).

# Variables and Data (1.2)

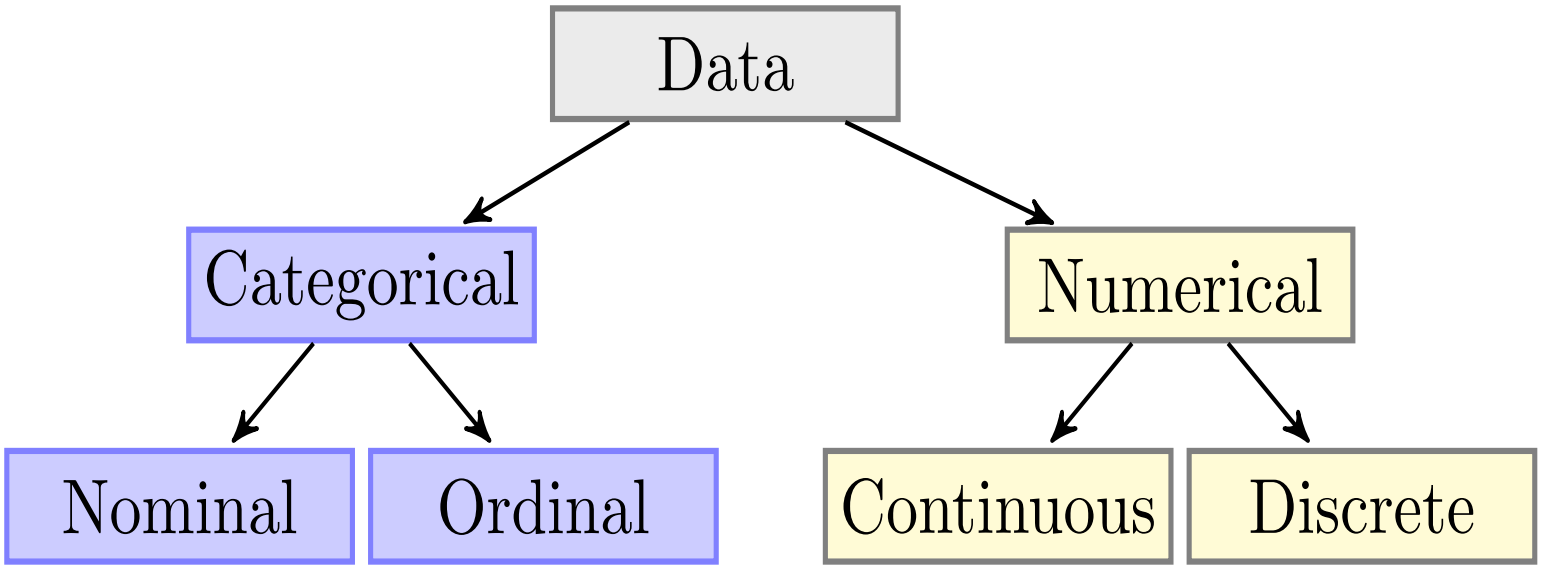
- ***Data*** are observed values of ***variables*** (characteristics that vary from one individual to the next).
- Variables can be ***categorical*** (or ***qualitative***) (taking values in a set of categories) or ***numerical*** (or ***quantitative***) (taking numerical values).

- Categorical variables can be:

- Categorical variables can be:
  - ***Ordinal*** (the categories have an inherent ordering, e.g. low, medium, high)
  - ***Nominal*** (the categories have no inherent ordering, e.g. red, green, and blue).

- Categorical variables can be:
  - ***Ordinal*** (the categories have an inherent ordering, e.g. low, medium, high)
  - ***Nominal*** (the categories have no inherent ordering, e.g. red, green, and blue).
- Numerical variables can be:

- Categorical variables can be:
  - ***Ordinal*** (the categories have an inherent ordering, e.g. low, medium, high)
  - ***Nominal*** (the categories have no inherent ordering, e.g. red, green, and blue).
- Numerical variables can be:
  - ***Discrete*** (the set of possible values is finite or countably infinite, i.e. they're isolated numbers with gaps between them, e.g. integers)
  - ***Continuous*** (the set of possible values is a continuum, or interval).


Data
Categorical
Numerical
Nominal
Ordinal
Continuous
Discrete

# Frequency Distributions and Histograms (1.2)

- ***Frequency Distribution for Discrete Data***: One way to summarize data.

# Frequency Distributions and Histograms (1.2)

- ***Frequency Distribution for Discrete Data***: One way to summarize data.

  Displays the number (***frequency***) and/or proportion (***relative frequency***) of times each value of a variable $x$ occurs.

## Example

Size of litters for $n = 36$ sows. Litter size (number of piglets) is an integer (discrete). The data:

| | | | | | |
|---|---|---|---|---|---|
| 10 | 12 | 10 | 7 | 14 | 11 |
| 14 | 11 | 10 | 13 | 10 | 10 |
| 8 | 11 | 7 | 13 | 12 | 13 |
| 10 | 8 | 5 | 11 | 11 | 12 |
| 11 | 11 | 9 | 8 | 12 | 10 |
| 9 | 11 | 10 | 12 | 10 | 9 |

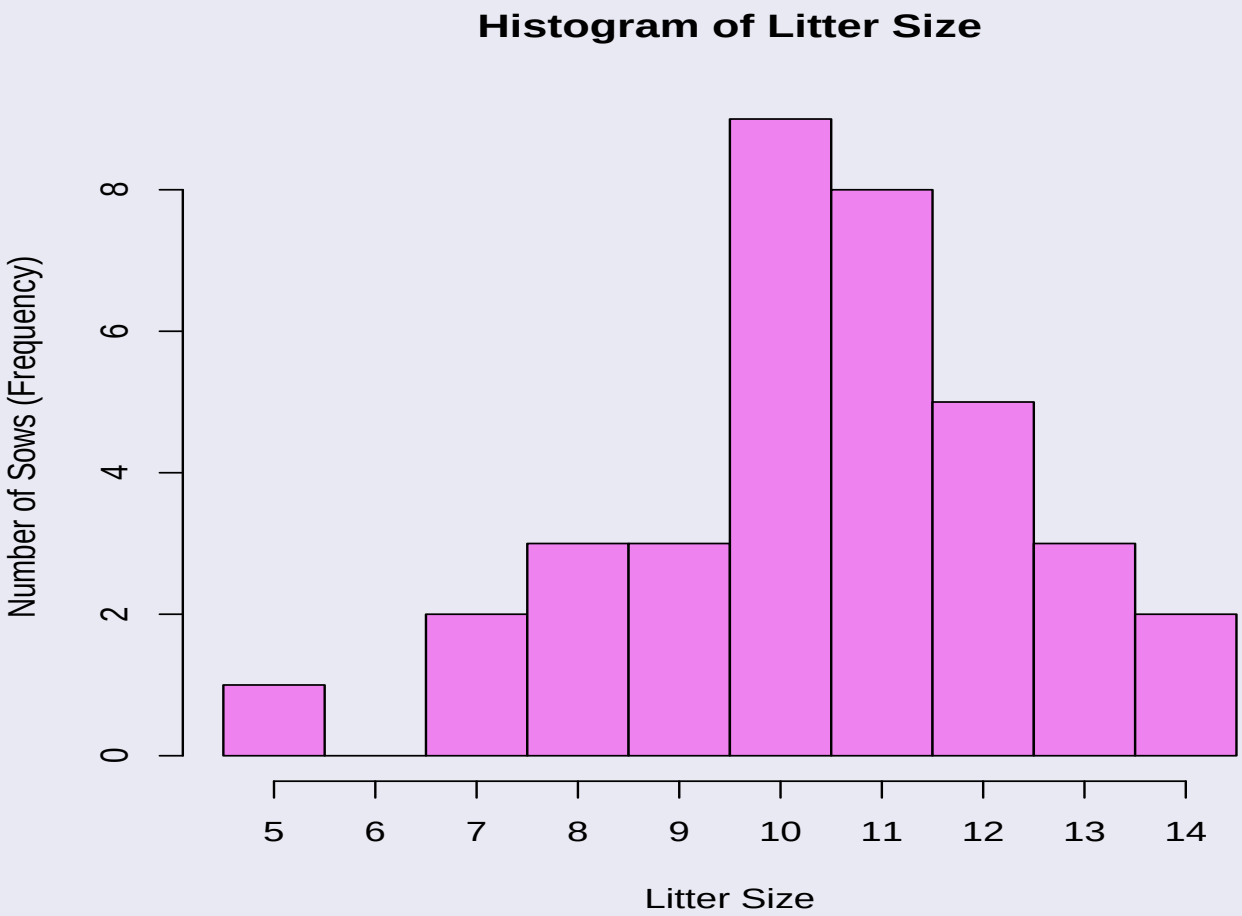
Here's the **frequency distribution** (table):

| Litter Size ($x$) | Number of Sows (Frequency) | Relative Frequency |
|---|---|---|
| 5 | 1 | 0.028 |
| 6 | 0 | 0.000 |
| 7 | 2 | 0.056 |
| 8 | 3 | 0.083 |
| 9 | 3 | 0.083 |
| 10 | 9 | 0.250 |
| 11 | 8 | 0.222 |
| 12 | 5 | 0.139 |
| 13 | 3 | 0.083 |
| 14 | 2 | 0.056 |

- ***Histogram for Discrete Data***: A graph of the frequency distribution.

- ***Histogram for Discrete Data***: A graph of the frequency distribution.
  - Mark the possible values $x$ on the horizontal axis.
  - Above each value, draw a rectangle whose height is the frequency (or relative frequency) of that value.

# Example (Cont'd)



- ***Frequency Distribution for Continuous Data***: A way to summarize the data.

- ***Frequency Distribution for Continuous Data***: A way to summarize the data.
  - Divide the measurement axis of $x$ into a suitable number of ***class intervals***.
  - Displays the number (***frequency***) and/or proportion (***relative frequency***) of data observations that fall into each interval.
  - Use right-open, left-closed class intervals (i.e. $a \leq x < b$) so that a data observation falling on a boundary goes in interval to the *right* of the boundary.

## Example

Here are surgery times (in hours) for emergency surgeries of $n = 50$ animals at a local animal hospital (ordered from shortest to longest).

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.33 | 0.33 | 0.33 | 0.33 | 0.50 | 0.50 | 0.50 | 0.67 | 0.75 |
| 0.75 | 0.75 | 0.83 | 0.92 | 0.92 | 1.00 | 1.00 | 1.00 | 1.00 |
| 1.08 | 1.08 | 1.17 | 1.25 | 1.27 | 1.33 | 1.42 | 1.42 | 1.50 |
| 1.50 | 1.50 | 1.50 | 1.58 | 1.67 | 1.67 | 1.67 | 1.67 | 1.67 |
| 1.70 | 1.75 | 1.75 | 1.83 | 1.83 | 2.00 | 2.00 | 2.00 | 2.33 |
| 2.42 | 2.50 | 2.67 | 3.08 | 4.50 | | | | |

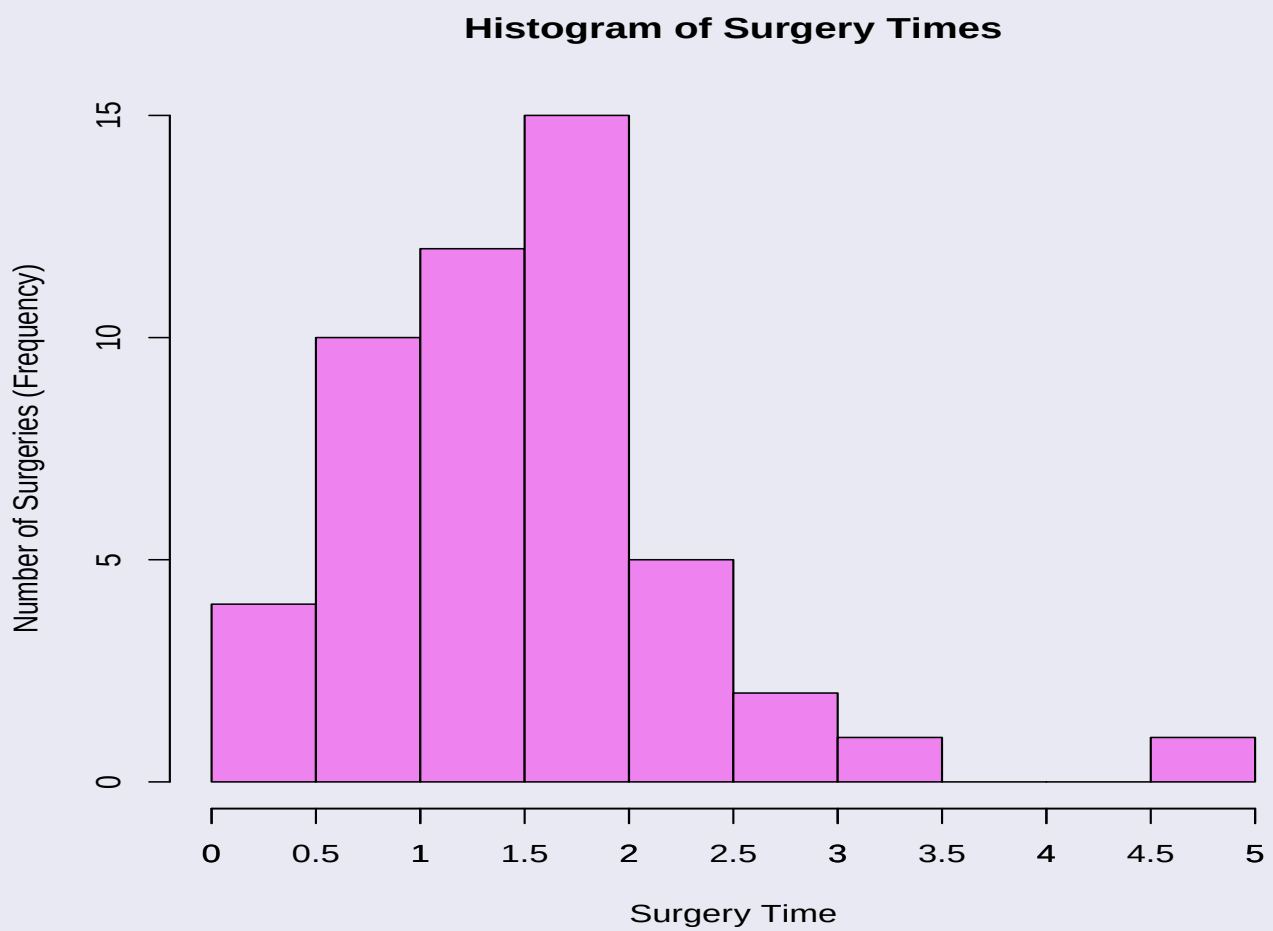
Here's the **frequency distribution** (table):

| Class Interval | Number of Surgeries (Frequency) | Relative Frequency |
|---|---|---|
| 0.0 - $<$ 0.5 | 4 | 0.08 |
| 0.5 - $<$ 1.0 | 10 | 0.20 |
| 1.0 - $<$ 1.5 | 12 | 0.24 |
| 1.5 - $<$ 2.0 | 15 | 0.30 |
| 2.0 - $<$ 2.5 | 5 | 0.10 |
| 2.5 - $<$ 3.0 | 2 | 0.04 |
| 3.0 - $<$ 3.5 | 1 | 0.02 |
| 3.5 - $<$ 4.0 | 0 | 0.00 |
| 4.0 - $<$ 4.5 | 0 | 0.00 |
| 4.5 - $<$ 5.0 | 1 | 0.02 |

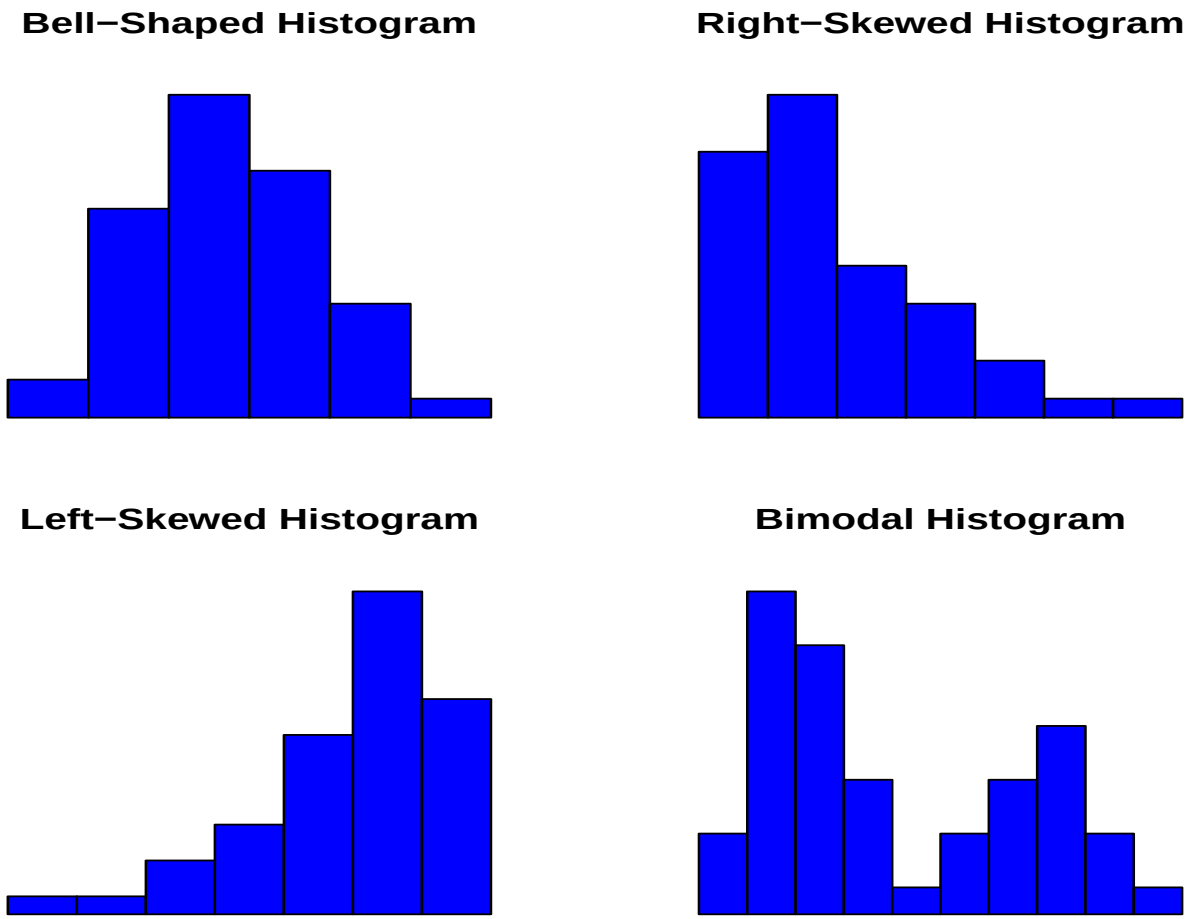
- ***Histogram for Continuous Data***: A graph of the frequency distribution.

- ***Histogram for Continuous Data***: A graph of the frequency distribution.
  - Mark the class interval endpoints on the horizontal $x$ axis.
  - Above each class interval, draw a rectangle whose height is the frequency (or relative frequency) of that interval.

# Example (Cont'd)



The figure below illustrates some common **histogram shapes**.



- **Terminology** used to describe **histogram shapes**:

- **Terminology** used to describe **histogram shapes**:
  - ***Symmetric*** – left and right halves are mirror images, most commonly ***bell-shaped***

- **Terminology** used to describe **histogram shapes**:
  - ***Symmetric*** – left and right halves are mirror images, most commonly ***bell-shaped***
  - ***Right skewed*** (or ***positively skewed***) – long ”tail” on the right

- **Terminology** used to describe **histogram shapes**:
  - ***Symmetric*** – left and right halves are mirror images, most commonly ***bell-shaped***
  - ***Right skewed*** (or ***positively skewed***) – long ”tail” on the right
  - ***Left skewed*** (or ***negatively skewed***) – long ”tail” on the left

- **Terminology** used to describe **histogram shapes**:
  - ***Symmetric*** – left and right halves are mirror images, most commonly ***bell-shaped***
  - ***Right skewed*** (or ***positively skewed***) – long ”tail” on the right
  - ***Left skewed*** (or ***negatively skewed***) – long ”tail” on the left
  - ***Bimodal*** – two distinct ”mounds”

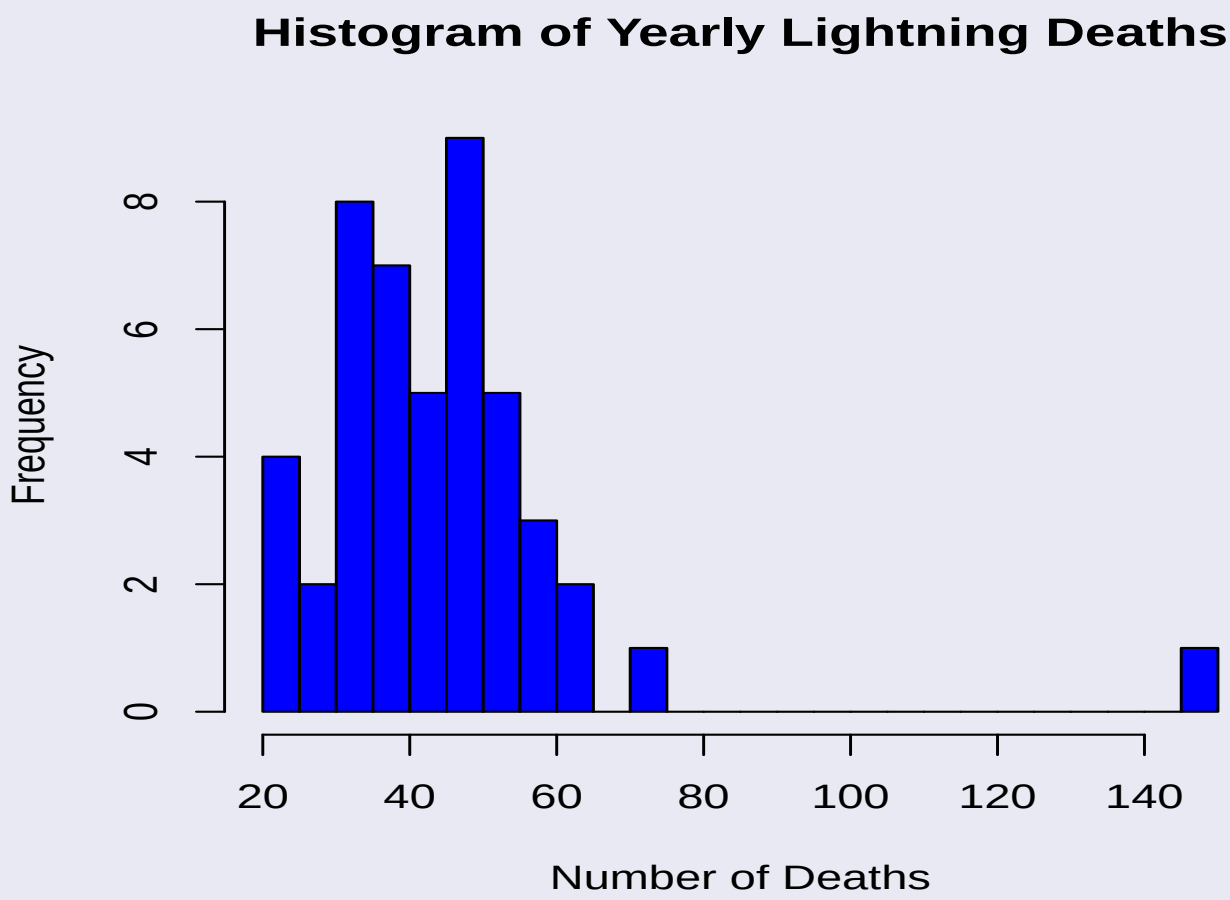
- **Terminology** used to describe **histogram shapes**:
  - ***Symmetric*** – left and right halves are mirror images, most commonly ***bell-shaped***
  - ***Right skewed*** (or ***positively skewed***) – long ”tail” on the right
  - ***Left skewed*** (or ***negatively skewed***) – long ”tail” on the left
  - ***Bimodal*** – two distinct ”mounds”
  - ***Multimodal*** – multiple ”mounds”

- **Terminology** used to describe **histogram shapes**:
  - ***Symmetric*** – left and right halves are mirror images, most commonly ***bell-shaped***
  - ***Right skewed*** (or ***positively skewed***) – long ”tail” on the right
  - ***Left skewed*** (or ***negatively skewed***) – long ”tail” on the left
  - ***Bimodal*** – two distinct ”mounds”
  - ***Multimodal*** – multiple ”mounds”
- Histograms can also reveal ***outliers*** (values far away from the rest of the data).

## Example

Here's (part of) a data set on the numbers deaths by lightning strikes in the U.S. for each of the years 1959 - 2005, as compiled from reports by the National Weather Service.

| Year | Deaths |
|---|---|
| 1959 | 75 |
| 1960 | 48 |
| 1961 | 61 |
| 1962 | 48 |
| 1963 | 150 |
| 1964 | 49 |
| ⋮ | ⋮ |
| 2005 | 38 |


Histogram of Yearly Lightning Deaths
Frequency
0
2
4
6
8
20
40
60
80
100
120
140
Number of Deaths

Regarding the **outlier**, the National Weather Service report states:

> *On December 8, 1963 the crash of a jetliner killing 81 people near Elkin, Maryland, was attributed to lightning by the Civil Aeronautics Board investigators.*

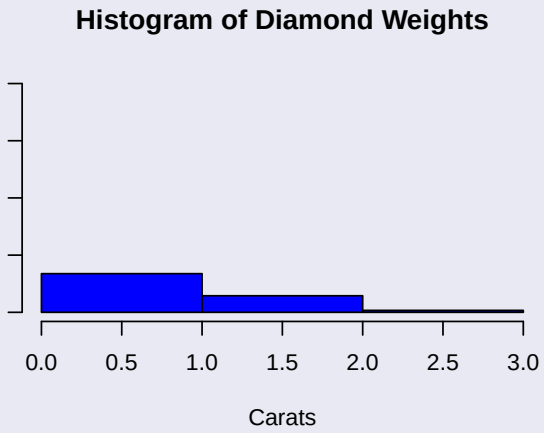
- Choosing the number of class intervals involves some judgment, but there are some rules and guidelines:

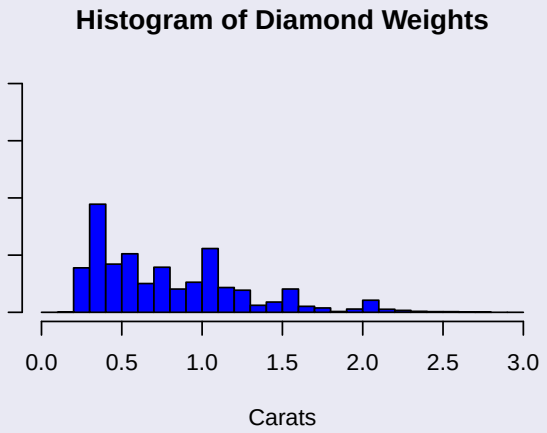
- Choosing the number of class intervals involves some judgment, but there are some rules and guidelines:
  - Each data value must go in exactly one class (i.e. no overlapping classes and no gaps between neighboring classes)
  - Classes should (usually) be all the same width
  - Use sensible and convenient boundaries, e.g. 20-30 not 19.83-28.87.
  - Usually 5-15 intervals works well. For larger data sets, more than 15 might be better.
    - Using too few can hide important details in the data.
    - Using too many intervals can show too much detail.

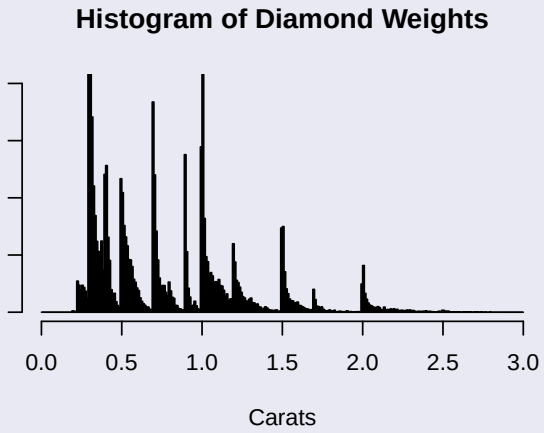
## Example

Here's (part of) a data set containing the weights in carats of $n = 50,000$ cut diamonds from reputable dealers around the world. A carat is 0.2 of a gram.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.36 | 1.20 | 0.72 | 0.54 | 0.33 | 0.71 | 1.01 | 1.20 | 2.22 |
| 0.90 | 0.55 | 0.77 | 1.21 | 1.22 | 2.01 | 0.31 | 0.32 | 0.73 |
| 0.32 | 1.47 | 2.00 | 0.33 | 1.01 | 1.50 | 0.72 | 0.33 | 0.27 |
| | | | | ⋮ | | | | |
| 0.44 | 0.50 | 0.31 | 0.34 | 0.90 | 0.70 | 1.52 | 1.10 | 0.76 |


Histogram of Diamond Weights
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Carats


Histogram of Diamond Weights
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Carats


Histogram of Diamond Weights
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Carats