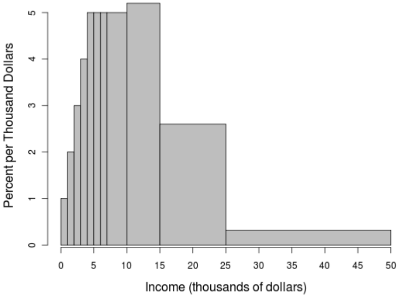
Math 207: Statistics

The Histogram



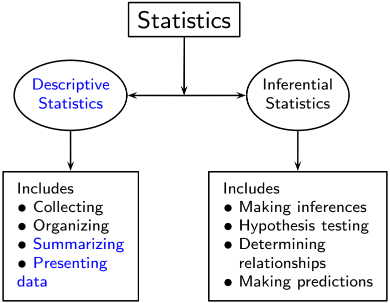
Dr. Ralph Wojtowicz

Mathematics Department

Shenandoah University

Introduction

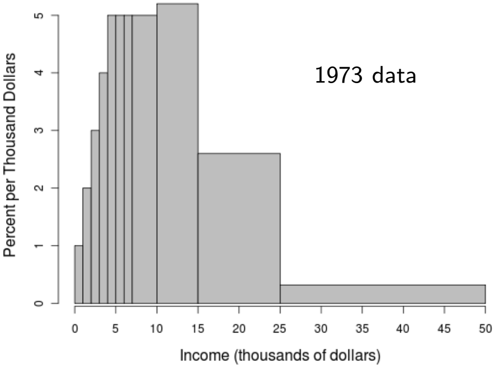
- **Graphical** tools (e.g., histograms and pie charts) can be used for summarizing and presenting data.
- These show the manner in which the data values are distributed.
- **Numerical** tools (mean, standard deviation, median and percentiles) are also used for describing data.



Reading a Histogram

- A **histogram** is a graph used to summarize data.
- The total area under the curve is 1 (that is, 100%).
- The horizontal axis is divided into **class intervals**.
- The area of a rectangle is proportional to the percentage of data values in the class interval.

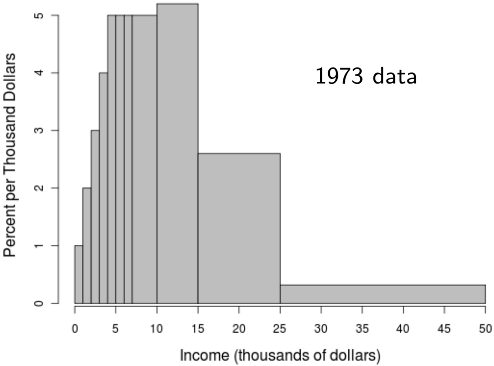
Income Level	Percent
\$0–\$1,000	1
\$1,000–\$2,000	2
\$2,000–\$3,000	3
\$3,000–\$4,000	4
\$4,000–\$5,000	5
\$5,000–\$6,000	5
\$6,000–\$7,000	5
\$7,000–\$10,000	15
\$10,000–\$15,000	26
\$15,000–\$25,000	26
\$25,000–\$50,000	8
\$50,000 and over	1



Drawing a Histogram from a Distribution Table

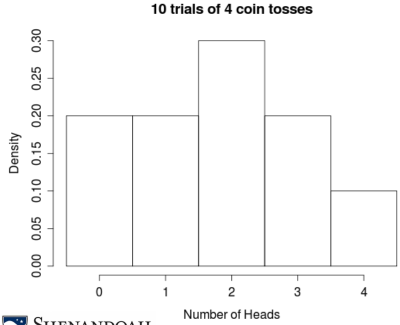
- A **distribution table** shows the percentage of data in each class interval.
- Choose an **endpoint convention** (e.g., put left endpoints in class intervals).
- Use the class intervals to draw horizontal axis.
- To figure out the height of a block over a class interval, divide the percentage by the length of the interval.

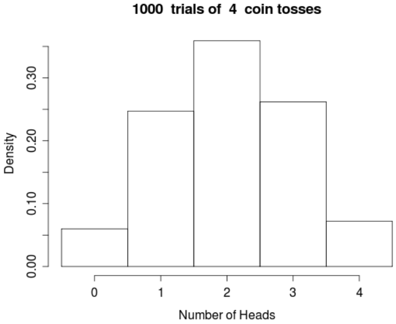
Income Level	Percent	Width	Height
\$0–\$1,000	1	1	1.0
\$1,000–\$2,000	2	1	2.0
\$2,000–\$3,000	3	1	3.0
\$3,000–\$4,000	4	1	4.0
\$4,000–\$5,000	5	1	5.0
\$5,000–\$6,000	5	1	5.0
\$6,000–\$7,000	5	1	5.0
\$7,000–\$10,000	15	3	5.0
\$10,000–\$15,000	26	5	5.2
\$15,000–\$25,000	26	10	2.6
\$25,000–\$50,000	8	25	0.3
\$50,000 and over	1		



Generating a Histogram from Data

- Toss a fair coin $n = 4$ times and count the number of heads.
- Repeat this experiment $N = 10$ times.
- Example: 3, 1, 3, 2, 0, 2, 1, 4, 2, 0 heads in the 10 trials gives the histogram below left.
- If we repeat the experiment $N = 1000$ times, we get a histogram such as the one shown below right.





Generating a Histogram from Data (Example)

- Simulate rolling 20 dice:

Generating a Histogram from Data (Example)

- Simulate rolling 20 dice:

```
> sample(1:6, 20, replace=T)
[1] 3 6 4 5 2 6 3 4 5 4 5 3 2 3 6 3 2 1 2 3
```

Generating a Histogram from Data (Example)

- Simulate rolling 20 dice:
```
> sample(1:6, 20, replace=T)
[1] 3 6 4 5 2 6 3 4 5 4 5 3 2 3 6 3 2 1 2 3
```
- Create a distribution table:

Generating a Histogram from Data (Example)

- Simulate rolling 20 dice:

```
> sample(1:6, 20, replace=T)
[1] 3 6 4 5 2 6 3 4 5 4 5 3 2 3 6 3 2 1 2 3
```

- Create a distribution table:

Number of Dots	Number of Rolls	Percent
1	1	$\frac{1}{20} = 0.05 = 5\%$
2	4	$\frac{4}{20} = 0.20 = 20\%$
3	6	$\frac{6}{20} = 0.05 = 30\%$
4	3	$\frac{3}{20} = 0.05 = 15\%$
5	3	$\frac{3}{20} = 0.05 = 15\%$
6	3	$\frac{3}{20} = 0.05 = 15\%$

Generating a Histogram from Data (Example)

- Simulate rolling 20 dice:

```
> sample(1:6, 20, replace=T)
[1] 3 6 4 5 2 6 3 4 5 4 5 3 2 3 6 3 2 1 2 3
```

- Create a distribution table:

Number of Dots	Number of Rolls	Percent
1	1	$\frac{1}{20} = 0.05 = 5\%$
2	4	$\frac{4}{20} = 0.20 = 20\%$
3	6	$\frac{6}{20} = 0.05 = 30\%$
4	3	$\frac{3}{20} = 0.05 = 15\%$
5	3	$\frac{3}{20} = 0.05 = 15\%$
6	3	$\frac{3}{20} = 0.05 = 15\%$

- Draw the histogram (called a bar plot since the variable isn't continuous):

Generating a Histogram from Data (Example)

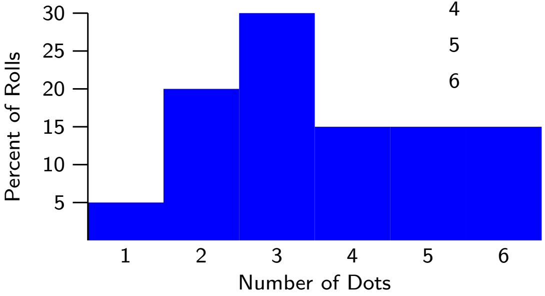
- Simulate rolling 20 dice:

```
> sample(1:6, 20, replace=T)
[1] 3 6 4 5 2 6 3 4 5 4 5 3 2 3 6 3 2 1 2 3
```

- Create a distribution table:

Number of Dots	Number of Rolls	Percent
1	1	$\frac{1}{20} = 0.05 = 5\%$
2	4	$\frac{4}{20} = 0.20 = 20\%$
3	6	$\frac{6}{20} = 0.05 = 30\%$
4	3	$\frac{3}{20} = 0.05 = 15\%$
5	3	$\frac{3}{20} = 0.05 = 15\%$
6	3	$\frac{3}{20} = 0.05 = 15\%$

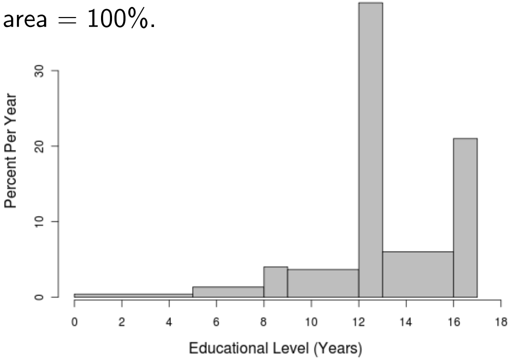
- Draw the histogram (called a bar plot since the variable isn't continuous):



The Density Scale

- The histogram below shows years of school completed by persons age 25 and older in the U.S. in 1991.
- Endpoint convention: years of school *completed* (e.g., people who dropped out part way through ninth grade are in the 8–9 block)
- Units on the vertical axis are percent (of people) per year (of schooling).
- Area represents percent. Total area = 100%.
- Box heights show *crowding*.
- Peaks: 8–9, 12–13 and 16–17

Years	Percent
0–5	2
5–8	4
8–9	4
9–12	11
12–13	39
13–16	18
16 or more	21



Variables

- A (random) **variable** is a measurement that depends on the outcome of a (random) event.
- **Quantitative** variables have numeric values.
    - **Continuous** variables can assume a continuum of values: Examples include income, temperature, pressure, mass, and speed.
    - A **discrete** variable can assume only finitely (or countably) many values. Examples include: family size, and number of engine cylinders.
- **Qualitative** variables are non-numeric.
    - They can be **Ordered**: good, better, best; or sometimes, always, never
    - or they can be **Unordered** such as eye color, marital status or automobile transmission type