Stats Fundamentals

Chad Worley

January 23, 2020

Data (sample) generation

- In the real world, data are generated by randomly sampling from a population.
- In this class, most data are generated with random number generators.
  - Coins
  - Dice
  - Spinners
  - Computer-based random number generators.
- Unfair coins are useful for generating binary data.
- Dice are useful for generating discrete uniform data.
- Spinners and computers can produce any type of data.

Example data and possible notations

I rolled a 4-sided die 7 times.

The results: 2, 4, 2, 4, 4, 1, 3

We let x_i represent the result of the ith roll. We call i the index.

$$x_1 = 2 \quad x_2 = 4 \quad x_3 = 2 \quad x_4 = 4 \quad x_5 = 4 \quad x_6 = 1 \quad x_7 = 3$$

We usually use a table.

i	x_i
1	2
2	4
3	2
4	4
5	4
6	1
7	3

The summation operator

The summation operator is defined with the following equation.

$$\sum_{i=a}^{b} x_i = x_a + x_{a+1} + \cdots + x_{b-1} + x_b$$

Simple examples:

$$\sum_{i=1}^{7} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7$$

$$\sum_{i=1}^{3} \sqrt{x_i + 6} = \sqrt{x_1 + 6} + \sqrt{x_2 + 6} + \sqrt{x_3 + 6}$$

The mean

We will always use n for the sample size (number of measurements).

$$n = \text{sample size}$$

The mean is the sum of the values divided by the sample size.

Formal definition:

$$\text{mean} = \frac{\sum_{i=1}^{n} x_i}{n}$$

In this class we can be a bit informal. We always sum over all values implicitly, so we can use a less decorated formula.

$$\text{mean} = \frac{\sum x}{n}$$

We will usually use $\bar{x}$ for a sample mean. We call this "x bar".

$$\bar{x} = \frac{\sum x}{n}$$

The mean: example

i	x_i
1	2
2	4
3	2
4	4
5	4
6	1
7	3

We calculate the mean by adding all the x values and dividing by the number of values.

$$\bar{x} = \frac{\sum x}{n} = \frac{2+4+2+4+4+1+3}{7} = 2.8571429$$

Data = Results = Sample = Measurements = Values

We are going to use a variety of words to mean basically the same thing in this class.

In the past, I have tried using only "measurements", but it feels unnatural. Like when rolling a die, it feels weird to call the results measurements. But, when discussing lizard lengths, it feels weird to call the measurements results.

In the context of rolling dice I will probably say "rolls".

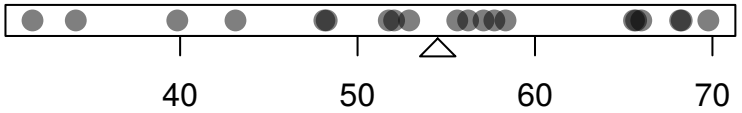
In the context of flipping a coin, I will say "flips".

In the context of spinning a spinner, I will say "spins".

Etc. . .

1-dimensional scatterplots (stripcharts)

Our textbook calls the following a dot plot:

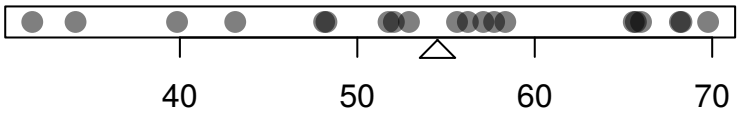


(made from the data below)

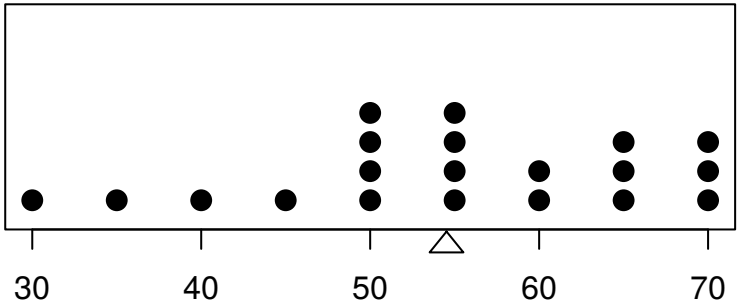
65.7, 52.06, 48.26, 68.27, 48.13, 57.1, 52.91, 34.12, 65.99, 39.84, 31.68, 43.12, 68.18, 65.58, 57.71, 55.62, 69.77, 56.23, 58.34, 51.76

The triangle marks the mean. Notice it marks the balance point (center of mass).

dot plots



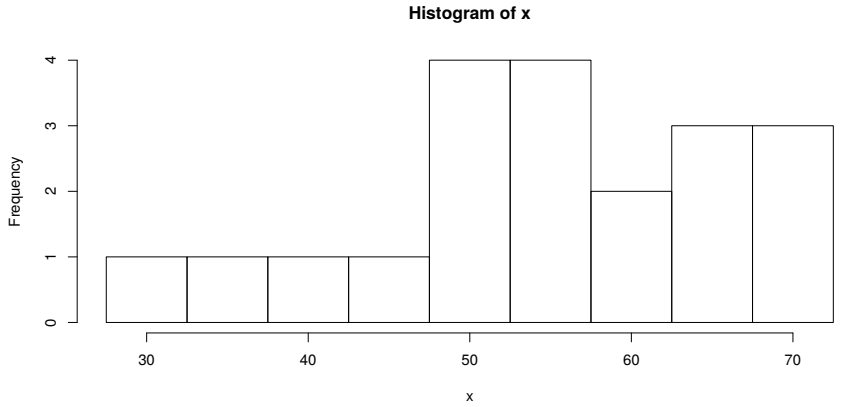
When we say **dot plot**, we will refer to what the textbook calls a stacked dotplot:



Notice, this was made by first rounding to the nearest multiple of 5. Stacking requires discrete (granular) data.

Histograms

A histogram is like a stacked dot plot, but uses bars instead of dots. **We will use histograms a lot.**



Center and Spread

We often want to summarize data with center and spread.

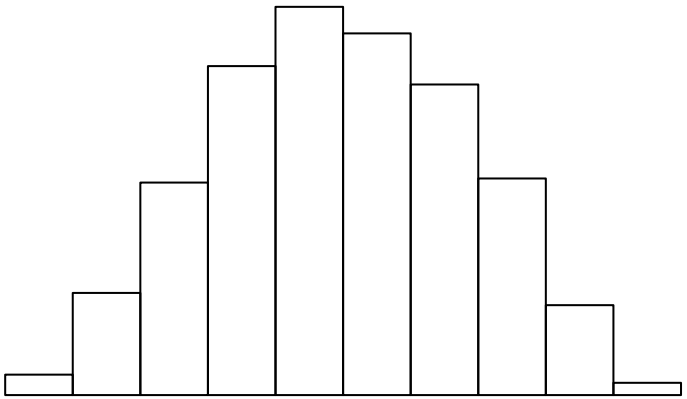
- Measures of center:
  - **Mean**
  - Median
- Measures of spread:
  - Range
  - Inter-quartile range
  - Mean absolute deviation
  - **Standard deviation**

The center indicates where the pile is located. The spread indicates how "wide" the pile is, or how much variation the variable has.

Shapes of histograms

The main shapes I want you to know are: bell-shaped, uniform, bimodal, left-skewed, and right-skewed.

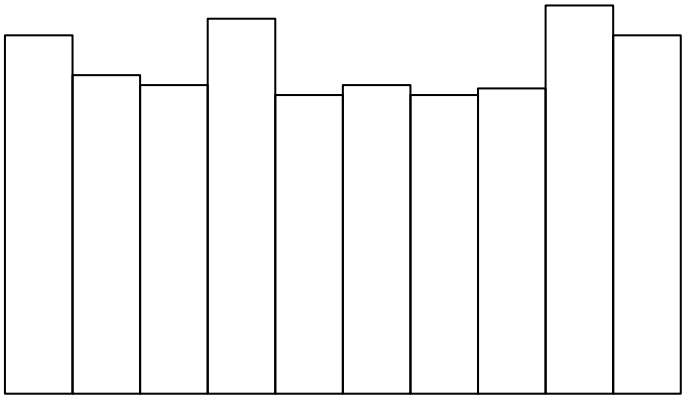
- Bell-shaped describes symmetric mounds with two tails.



Shapes of histograms

The main shapes I want you to know are: bell-shaped, uniform, bimodal, left-skewed, and right-skewed.

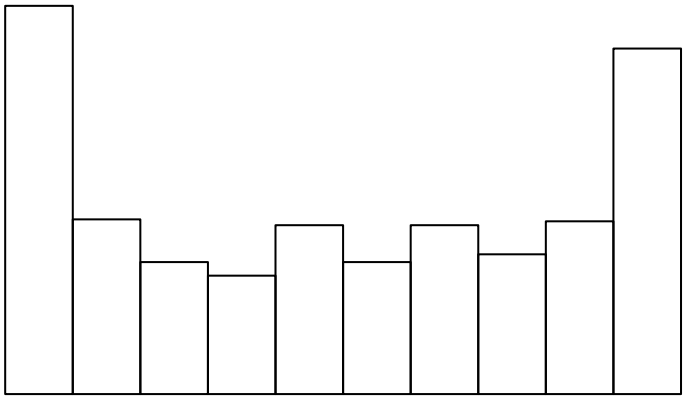
- Uniform describes symmetric mesas (flat-topped hill with steep sides).



Shapes of histograms

The main shapes I want you to know are: bell-shaped, uniform, bimodal, left-skewed, and right-skewed.

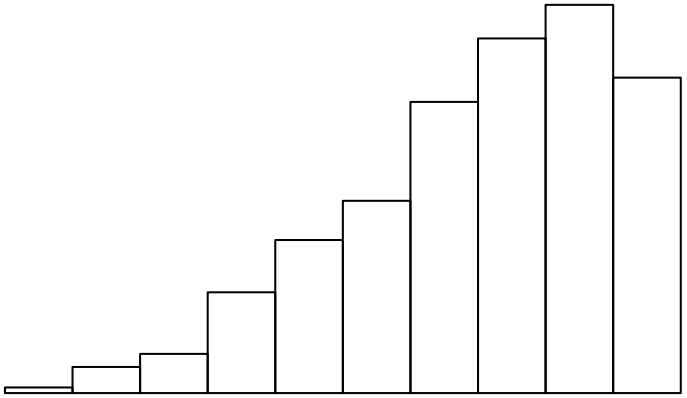
- Bimodal describes two hills around a valley.



Shapes of histograms

The main shapes I want you to know are: bell-shaped, uniform, bimodal, left-skewed, and right-skewed.

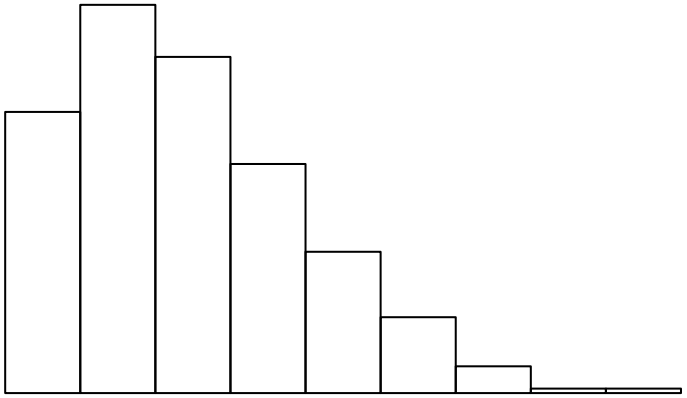
- Left-skewed describes a mound with a long left tail.



Shapes of histograms

The main shapes I want you to know are: bell-shaped, uniform, bimodal, left-skewed, and right-skewed.

- Right-skewed describes a mound with a long right tail.



The sample standard deviation

The sample standard deviation will always be denoted with s.

The standard deviation of a sample (data) is determined with the following formula.

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

Or, informally:

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}$$

The sample standard deviation: example calculation

Let's say we have the following data:

6, 9, 11, 8, 12

First, we determine the sample size by counting how many numbers are in the sample.

$$n = 5$$

We can calculate the mean.

$$\bar{x} = \frac{6+9+11+8+12}{5} = 9.2$$

We can use a table to determine the standard deviation.

x	$x - \bar{x}$	$(x - \bar{x})^2$
6	-3.2	10.24
9	-0.2	0.04
11	1.8	3.24
8	-1.2	1.44
12	2.8	7.84

The sample standard deviation: example calculation...

$$n = 5$$

$$\bar{x} = 9.2$$

x	$x - \bar{x}$	$(x - \bar{x})^2$
6	-3.2	10.24
9	-0.2	0.04
11	1.8	3.24
8	-1.2	1.44
12	2.8	7.84

$$\sum(x - \bar{x})^2 = 22.8$$

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}} = \sqrt{\frac{22.8}{5-1}} = \sqrt{5.7} \approx 2.387$$