

# Probability and Statistics

Alvin Lin

Probability and Statistics: January 2017 - May 2017

## Intro to Statistics

Suppose there is a set of people who can vote and a subset of them whom we want to survey. We can make a prediction about the people who can vote by surveying the subset. The set of people who can vote is called a **population** and the set of people surveyed is called a **sample**. This is known as inferential statistics.

$population \rightarrow sample$  (Probability)

$sample \leftarrow population$  (Inferential statistics)

Consider the experiment of tossing a coin twice. The **sample space** of this experiment is the set

$$\{(H, H), (H, T), (T, H), (T, T)\}$$

Each element (e.g. (T,H)) is a sample point.

## Example

For each of the following hypothetical populations, give a plausible sample of size 4.

All distances that might result when you throw a football (feet):

$$\{32, 40, 50, 46\}$$

Page lengths of books published 5 years from now:

$$\{520, 600, 670, 700\}$$

# Descriptive Statistics

In descriptive statistics, we do not make predictions. We use tools such as mean, median, and mode to describe data, and visualizations such as histograms and box plots.

## Example

Suppose we have 20 one-quart water bottles and of those 20, we select 5 water bottles. The pH values of water in the 20 water bottles is the population and the pH value of water in the 5 selected bottles is the sample.

The sample size is 5 ( $n = 5$ ). We can use the variable  $x$  to denote pH values. Individual observations can be denoted  $x_1, x_2, x_3, x_4, x_5$ .

Suppose the range of the function  $x$  is  $\{6.1, 6.8, 6.8, 7.1, 7.4\}$ .

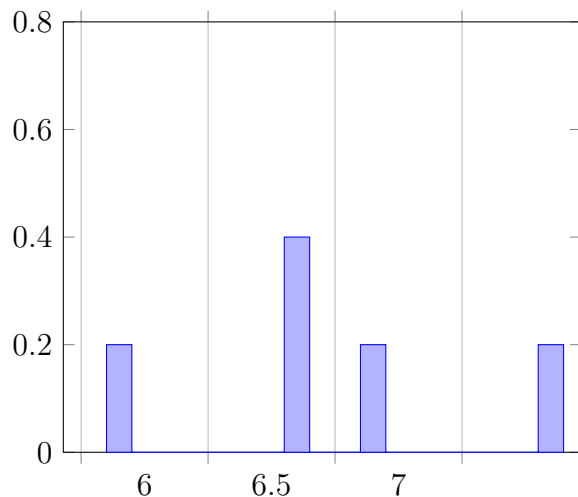
The frequency of 6.1 is 1.

The frequency of 6.8 is 2.

The frequency of 9.0 is 0.

The relative frequency of 6.1 is  $\frac{1}{5} = 0.2 = 20\%$ .

The relative frequency of 6.8 is  $\frac{2}{5} = 0.4 = 40\%$ .



There are two types of histograms in our textbook.

	Type 1	Type 2
Vertical Axis	Relative Frequency	Density $\frac{Relative\ Frequency}{Width}$
Width of the rectangles	Same	Do not have to be the same
Sum up to 1	Heights of the rectangles	Area of the rectangles

## Countable and Uncountable Sets

$$\begin{aligned} \text{Let } A &= \{1, 2, 3, 4, 5\} \\ B &= \{0.0, 0.1, 0.2, 0.3, \dots, 13.9, 14.0\} \\ C &= [0, 14] \end{aligned}$$

If we represent the function  $x$  as:

$$\{x \mid A \rightarrow B\}$$

$B$  is a countable set and  $x$  is a discrete variable. If we represent the function  $x$  as:

$$\{x \mid A \rightarrow C\}$$

$C$  is a countable set and  $x$  is a continuous variable. Set theory can be applied to functions of the following type:

$$f : A \rightarrow C \subseteq \mathbb{R}$$

Let  $W = \{0, 1, 2, 3, \dots\}$ . A set  $A$  is called countable if there is a bijective (one-to-one) function from  $A$  to a subset of  $W$ . Otherwise,  $A$  is uncountable. Countable sets:

$$\begin{aligned} &\{James, Tina, Christine\} \\ &\{1, 8, 13, 17\} \\ &\left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\right\} \\ &\{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\} \end{aligned}$$

Uncountable sets:

$$[0, 1]$$

$$\begin{array}{c} (0, 1] \\ (0, 1) \\ \mathbb{R} \end{array}$$

## Power set of a set

A power set of  $A$ ,  $P(A)$  or  $2^A$ , is the set containing all subsets of  $A$ .

$$\text{Let : } A = \{1, 2, 3\}$$

$$2^A = \{\emptyset, 1, 2, 3, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$$

## Sample Mean

As sets:

$$\{6.1, 6.8, 6.8, 7.1, 7.4\} = \{6.1, 6.8, 7.1, 7.4\}$$

But if we take the arithmetic mean:

$$\bar{x} \neq \frac{6.1 + 6.8 + 7.1 + 7.4}{4}$$

Thus, it is better for us to represent the set as an ordered 5-tuple:

$$\begin{array}{c} \{\dots, (2, 6.8), (3, 6.8), \dots\} \\ (6.1, 6.8, 6.8, 7.1, 7.4) \end{array}$$

## Sample Median

With the set:

$$(6.1, 6.8, 6.8, 7.1, 7.4)$$

The median of the set is:

$$\tilde{x} = 6.8$$

In the case that the sample size is even:

$$\left( 5.3, 6.8, 6.8, 7.1, 7.4, 7.9 \right)$$

The median is:

$$\tilde{x} = \frac{6.8 + 7.1}{2}$$

## Measure of Variability

How are the sample data spread out? We can express the spread in numbers or visually as scatter, dispersion, or variability.

$$\text{Let : } S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i)^2 - n(\bar{x}^2)$$

$$\text{Sample variance : } s^2 = \frac{S_{xx}}{n - 1}$$

$$\text{Sample standard deviation : } s = \sqrt{s^2}$$

### Proposition

Let  $x_1, x_2, x_3, \dots, x_n$  be a sample and  $(S_x)^2$  and  $S_x$  be the variance and standard deviation of the sample, respectively.

Let  $a$  and  $b$  be constants.

If  $y_i = ax_i + b$  for any  $i \in \{1, 2, 3, \dots, n\}$ , then the variance and standard deviation of the new sample  $y_1, y_2, y_3, \dots, y_n$  are  $(S_y)^2 = a^2(S_x)^2$  and  $S_y = |a|S_x$ , respectively.

### Usage Example

Suppose a sample consists of 1000 temperature measurements in Centigrade. We know its standard deviation. If we want to convert the temperature to Fahrenheit, we can apply  $F = \frac{9}{5}C + 32$  to each individual measurement.

We can also use the proposition above to determine the new standard deviation:

$$S_F = \left| \frac{9}{5} \right| S_C$$

You can find all my notes at <http://omgimanerd.tech/notes>. If you have any questions, comments, or concerns, please contact me at [alvin@omgimanerd.tech](mailto:alvin@omgimanerd.tech)