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.

$$\textit{population} \rightarrow \textit{sample} \quad \text{(Probability)}$$

$$\textit{sample} \leftarrow \textit{population} \quad \text{(Inferential statistics)}$$

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

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

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):

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

Page lengths of books published 5 years from now:

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

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

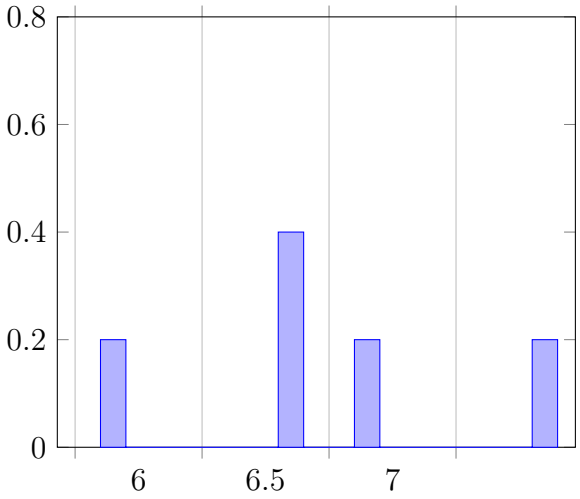
The frequency of 6.1 is 1.

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 $\frac{RelativeFrequency}{Width}$
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

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

If we represent the function x as:

$$\Big\{x \mid A \to B\Big\}$$

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

$$\Big\{x \mid A \to C\Big\}$$

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

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

Let $W = \big\{0, 1, 2, 3, \ldots\big\}$. 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:

$$\Big\{James, Tina, Christine\Big\}$$

$$\Big\{1, 8, 13, 17\Big\}$$

$$\Big\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\Big\}$$

$$\Big\{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\Big\}$$

Uncountable sets:

$$\Big[0, 1\Big]$$

$$\left(0, 1\right]$$

$$\left(0, 1\right)$$

$$\mathbb{R}$$

Power set of a set

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

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

$$2^A = \left\{\varnothing, 1, 2, 3, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\right\}$$

Sample Mean

As sets:

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

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:

$$\left\{\ldots, (2, 6.8), (3, 6.8), \ldots\right\}$$

$$\left(6.1, 6.8, 6.8, 7.1, 7.4\right)$$

Sample Median

With the set:

$$\left(6.1, 6.8, 6.8, 7.1, 7.4\right)$$

The median of the set is:

$$\tilde{x} = 6.8$$

In the case that the sample size is even:

$$\Big(5.3, 6.8, 6.8, 7.1, 7.4, 7.9\Big)$$

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.

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

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

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

Proposition

Let $x_1, x_2, x_3, \ldots, x_n$ 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 $i \in \{1, 2, 3, \ldots, n\}$, then the variance and standard deviation of the new sample $y_1, y_2, y_3, \ldots, y_n$ 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:

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