Econ 5023: Statistics for Decision Making

Univariate Statistics (II): Intro to Random Variables and Probability Distributions (Discrete Variables)

Le Wang

2018-10-08

A Prediction Problem

An experiment from **Prediction Machines: The Simple Economics of Artificial Intelligence** by Argawal, Gans, and Goldfarb

OXXOXOXOXOXXOOXXOXOXXXOXX

Question: Suppose that the data are telling you about the true underlying pattern

1. What would be your prediction for the next one?
2. What would be your prediction for the next 10?

A Prediction Problem

Answer: 60 percent of X, and always X!

The goal

Imagine that the distribution is exactly the jar containing some Xs and Os, but how many Xs and Os or their ratio is unknown to us!

Note that this is nothing but a **classification** problem in **machine learning**

Classification refers to the problem of predicting a categorical (or discrete) outcome.

Examples

1. Good credit risk/bad credit risk
2. Heart attack/no heart attack,
3. Spam/not spam

Road Map

We have so far considered various events and their probability values. We could introduce the concept of *random variables* and their *probability functions*, which further widens the scope of mathematical analyses of these events.

1. Random Variables
2. Probability Distribution Functions

Some definitions are based on **Chapter 2**.

Outcome of Interest: Random Variables

Random Variables are variables that are not perfectly preditable but whose repeated realizations are described by a *probabilitiy distribution function*. A random variable assigns a number to each event of the experiment.

1. The values of random variables must represent *mutually exclusive* and *exhaustive* events.
2. All these values together form the entire sample space.

That is, different values *cannot* represent the same event, and all events should be represented by some values.

Examples

1. Two outcomes of X and O.
2. Two outcomes of a coin flip can be represented by a binary random variable where 1 indicates landing on heads and 0 landing on tails.
3. Gender: male denoted by 0 and female 1.
4. etc.

Examples (more complicated yet realistic case)

Racial group using five unique categories

1. black $= 1$
2. white $= 2$
3. hispanic $= 3$
4. asian $= 4$
5. others $= 5$

Examples (more complicated yet realistic case)

Let's look at what people have reported

Label	ASEC 17	ASEC 16	ASEC 15	ASEC 14	ASEC 13	ASEC 12	ASEC 11	ASEC 10	ASEC 09	ASEC 08	ASEC 07
White	X	X	X	X	X	X	X	X	X	X	X
Black/Negro	X	X	X	X	X	X	X	X	X	X	X
American Indian/Aleut/Eskimo	X	X	X	X	X	X	X	X	X	X	X
Asian or Pacific Islander	.	.	.	.	.	.	.	.	.	.	.
Asian only	X	X	X	X	X	X	X	X	X	X	X
Hawaiian/Pacific Islander only	X	X	X	X	X	X	X	X	X	X	X
Other (single) race, n.e.c.	.	.	.	.	.	.	.	.	.	.	.
Two or more races											
White-Black	X	X	X	X	X	X	X	X	X	X	X
White-American Indian	X	X	X	X	X	X	X	X	X	X	X
White-Asian	X	X	X	X	X	X	X	X	X	X	X
White-Hawaiian/Pacific Islander	X	X	X	X	X	X	X	X	X	X	X
Black-American Indian	X	X	X	X	X	X	X	X	X	X	X
Black-Asian	X	X	X	X	X	X	X	X	X	X	X
Black-Hawaiian/Pacific Islander	X	X	X	X	X	X	X	X	X	X	X
American Indian-Asian	X	X	X	X	X	X	X	X	X	X	X
Asian-Hawaiian/Pacific Islander	X	X	X	X	X	X	X	X	X	X	X
White-Black-American Indian	X	X	X	X	X	X	X	X	X	X	X
White-Black-Asian	X	X	X	X	X	X	X	X	X	X	X
White-American Indian-Asian	X	X	X	X	X	X	X	X	X	X	X
White-Asian-Hawaiian/Pacific Islander	X	X	X	X	X	X	X	X	X	X	X
White-Black-American Indian-Asian	X	X	X	X	X	X	X	X	X	X	X
American Indian-Hawaiian/Pacific Islander	X	X	X	X	X	.	.	.	.	.	.
White-Black--Hawaiian/Pacific Islander	X	X	X	X	X	.	.	.	.	.	.
White-American Indian-Hawaiian/Pacific Islander	X	X	X	X	X	.	.	.	.	.	.

Label	ASEC 17	ASEC 16	ASEC 15	ASEC 14	ASEC 13	ASEC 12	ASEC 11	ASEC 10	ASEC 09	ASEC 08	ASEC 07
Black-American Indian-Asian	X	X	X	X	X	.	.	.	.	.	.
White-American Indian-Asian-Hawaiian/Pacific Islander	X	X	X	.	X	.	.	.	.	.	.
Two or three races, unspecified	X	X	X	X	X	X	X	X	X	X	X
Four or five races, unspecified	X	X	X	X	X	X	X	X	X	X	X

Examples (more complicated yet realistic case)

What should you do in practice?

One solution is just code all these values that do not fall into the first four categories as 5.

Question What is the problem?

Advantages of Random Variables

We can easily manipulate the **numerical** outcome of an experiment in a format amenable to further analysis.

Example: Consider an experiment where we flip a coin 10 times. Let 0 represent tails and 1 heads. Our sample space is

$$\Omega = \{(b_1, b_2, \ldots, b_{10}) : b_n \in \{0, 1\} \text{ for each } n\}$$

Whether any heads occur in the outcome can be defined as follows

$$y(\omega) = y(b_1, b_2, \ldots, b_{10}) = \min\{\sum_{n=1}^{10} b_n, 1\}$$

An Important Random Variable (and Indicator Function)

A **Binary (or Bernoulli)** random variable is a random variable taking values in the set $\{0, 1\}$. It can be constructed through **indicator function**.

If Q is a statement, such as "Last name is Wang", then $\mathbb{I}[Q]$ is set to 1 when the statement Q is true, and zero otherwise. For an event $A \in \mathcal{F}$,

$$\mathbb{I}[\omega \in A] = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

Properties of Indicator Function or Binary Random Variables

Fact If $A_1, A_2, \ldots A_N$, then 1. $\mathbb{I}[\cap A_N] = \Pi_{n=1}^{N} \mathbb{I}[A_n]$ (This is useful for an important result for our analysis next semester). 2. $\mathbb{I}[\cup A_N] = \sum_{n=1}^{N} \mathbb{I}[A_n]$ whenever the sets are disjoint

Random Variables and Probability Distrbution Functions

Because (unlike events) a random variable takes numeric values, we can develop function-like mathematial rules mapping the events to their corresponding probability.

$\Longrightarrow$ a fomalized *probability model* using the distribution *function* of the random variable.

Outline

We will use two Ways to Discover Patterns/truth/distribution in our sample

1. A complete distribution approach
2. Only finite parameters of the distribution (Parameter, from the Greek for “almost measurement”)

What is our goal?

Before we examine how each method could help us discover the patterns or the distribution for forecasting, let's first understand

What should our forecasts be? Among all these values??

Suppose that you already obtain this function

What is our goal?

What should our forecasts be? Among all these values? Intuitively speaking

1. “Most likley” values
2. “Typical”" values

Some of these intuitions will fall apart.

The Distributional Approach

Now, let's see how this approach can help us find the most likely values.

1. We will first look at **discrete variables** to motivate our strategy
2. We will then see the issues with **continuous variables**, which will motivate our second approach.
3. Through this process, we will also see how R treats discrete variables.

The Distributional Approach

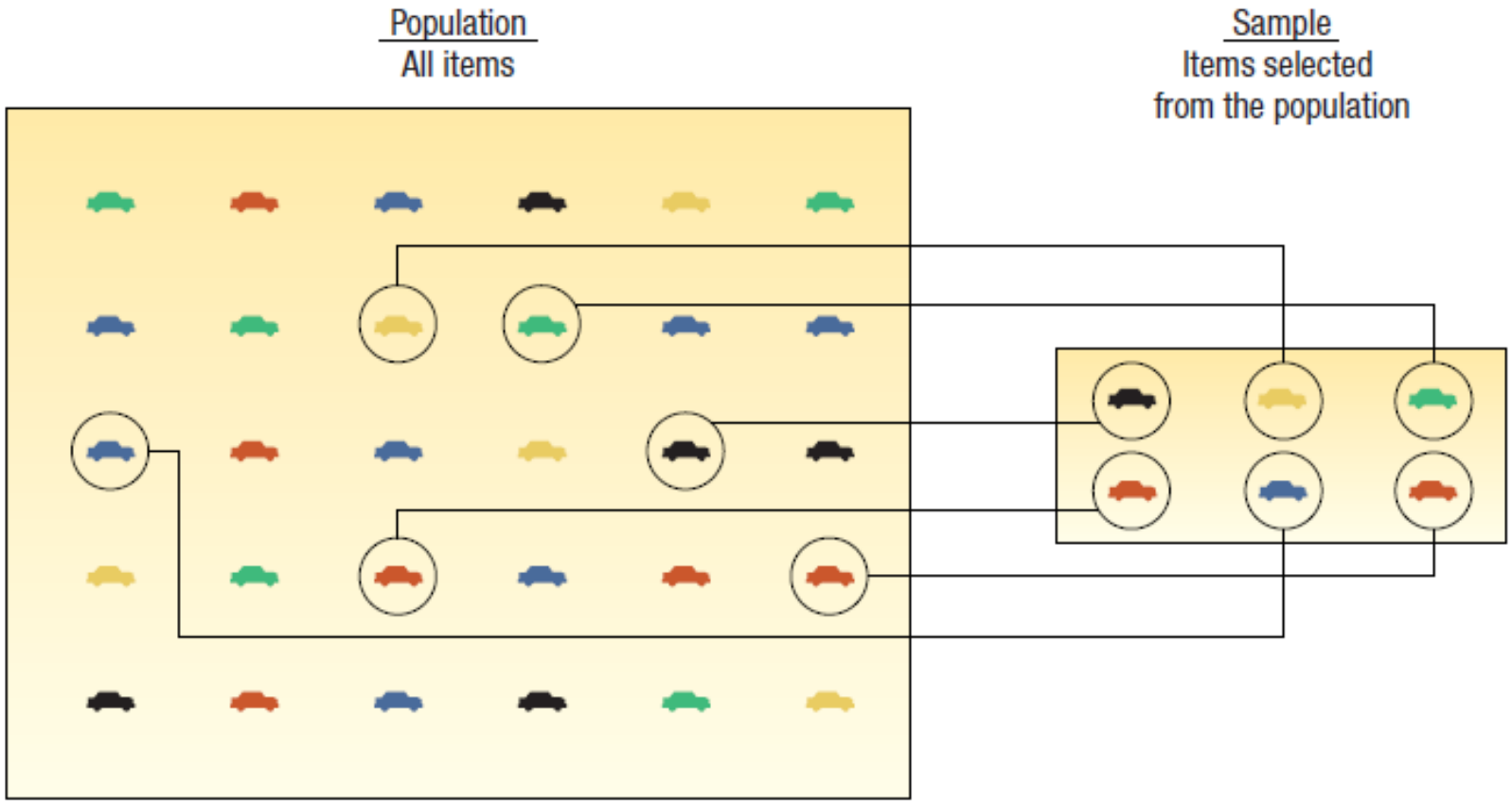
To uncover the distribution function (or the pattern), we will examine our data. It is important to distinguish between two concepts

Population vs. **Sample**

Population vs Sample

Population: The entire set of individuals or objects of interest

Sample: A portion, or part, of the population of interest.



Let's visualize the relationship between population and sample

Chapter 3: Probability Distributions – Random Variable

Population vs Sample

Note that: population does not change but sample may

- The size of a sample may vary
- The draws of the sample may vary

Since all we can know about the population depends on the sample available to us, every time when we have a different sample, we have a different answer. What to do??!!

Problem: Sampling Errors (we will discuss these more later)

Frequency Tables (Book's definition):

A grouping of qualitative data into **mutually exclusive** and **collectively exhaustive classes** showing the number of observations in each class.

Example:

```
x <- c(6, 7, 6, 3, 10, 3, 2, 1, 10, 7)
x
```

```
##  [1]  6  7  6  3 10  3  2  1 10  7
```

1. Class one: any values between 1 and 7
2. Class two: any values greater than 7.

By hand

Step 1: For discrete variables, we can classify each value as one class and look at all values/classes.

```
x
##  [1]  6  7  6  3 10  3  2  1 10  7
sort(x)
##  [1]  1  2  3  3  6  6  7  7 10 10
```

Step 2: And then count the unique values. From this step, we can alread know which value is unlikely.

```
unique(x)
```

```
## [1]  6  7  3 10  2  1
```

Step 3: We can then count

Use R

```
table(x)
```

```
## x
##  1  2  3  6  7 10
##  1  1  2  2  2  2
```

Visualization of the Data: Histogram

When there are many data points, it becomes infeasible to just count by hand. We need to use computers to accomplish this task for us.

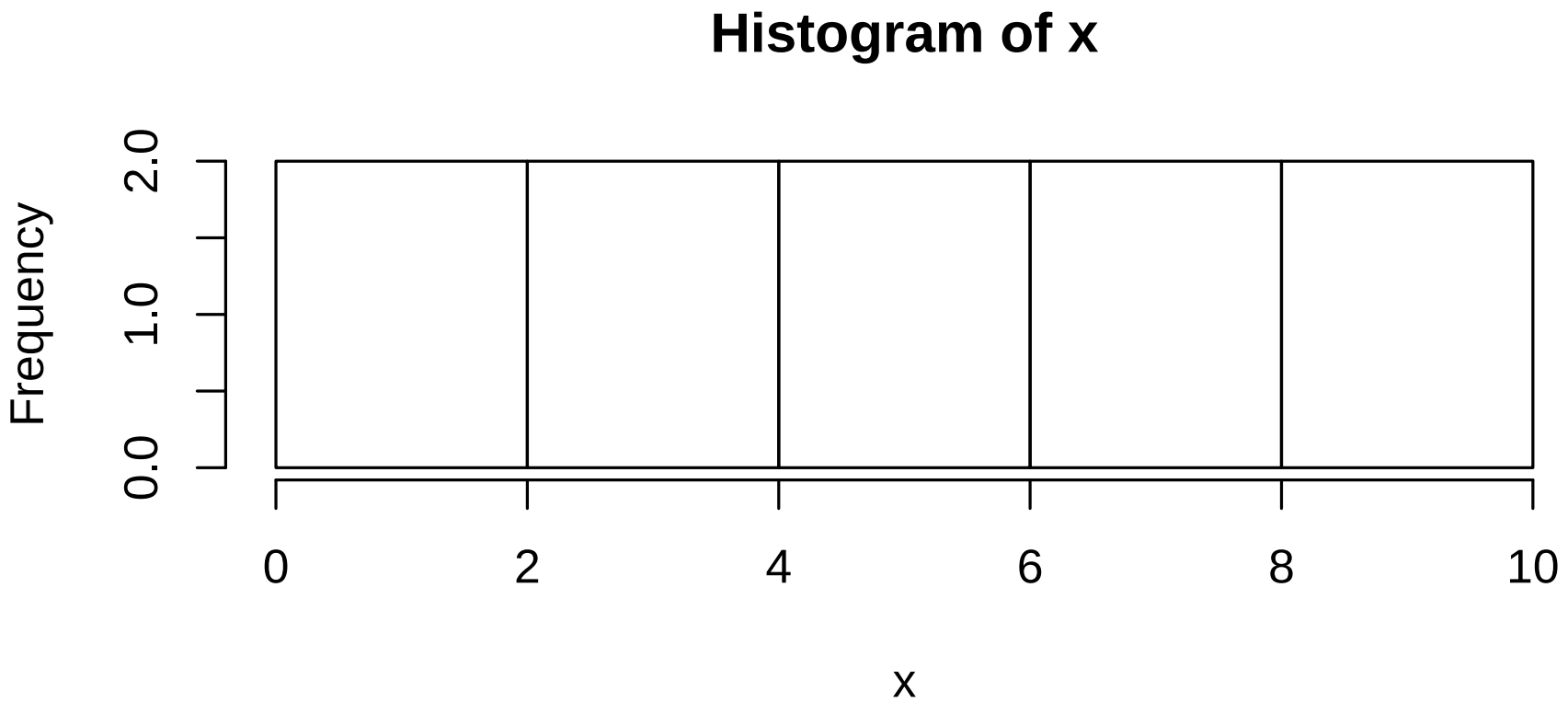
HISTOGRAM A graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars, and the bars are drawn adjacent to each other.

Density: The height of each class (or bin).

Visualization of the Data: Histogram

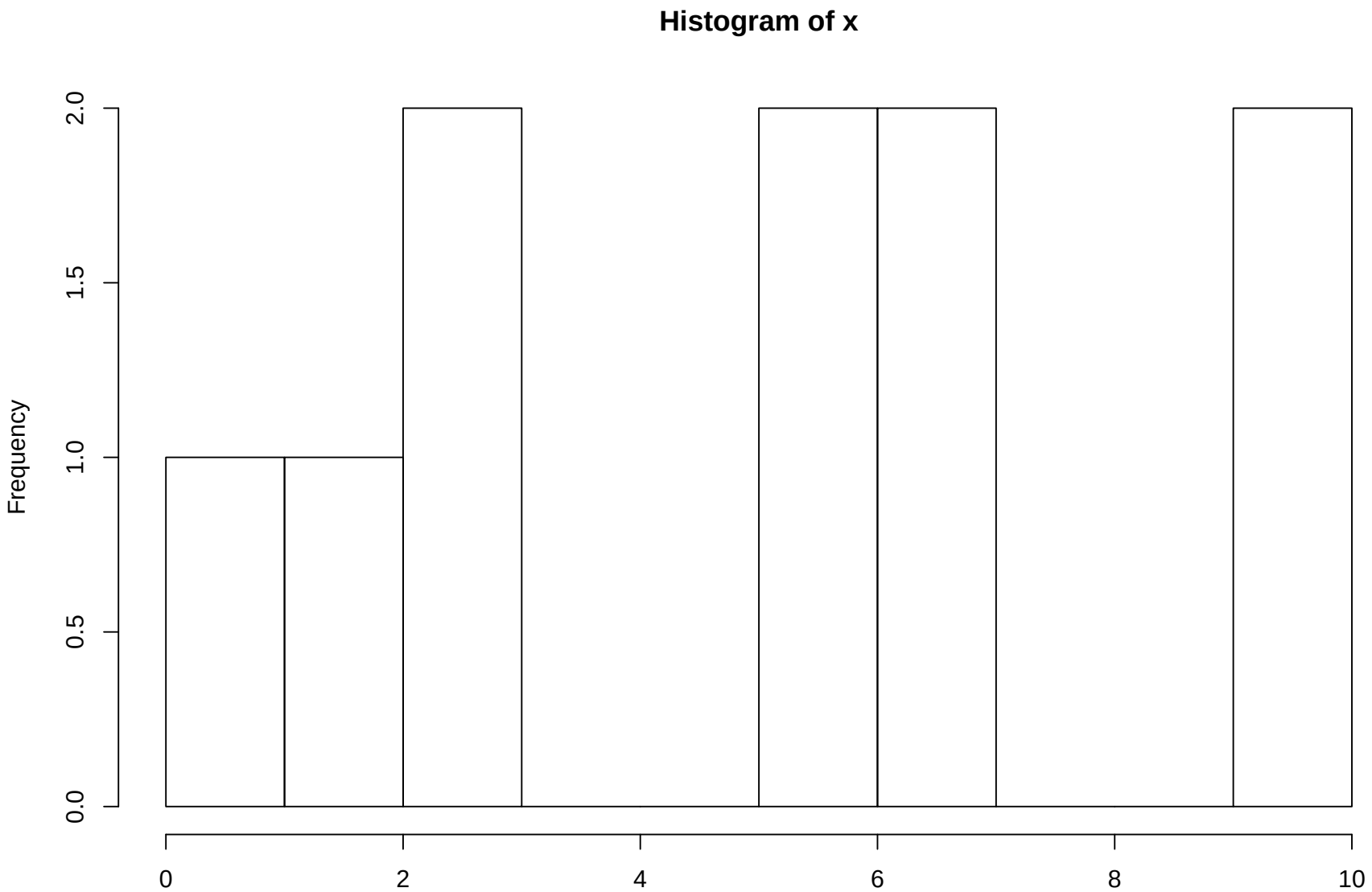
The command is called: `hist()`

```
hist(x)
```



Not very helpful if we want to find the most likely value(s) for forecasting!

```
hist(x, breaks = c(0:10), freq = TRUE)
```



We can also calculate the **percentage** of frequency or **relative frequency** of each value. This is an estimate of

Probability Mass Function:

$$f(y) = Pr[Y = y] = p_i \quad i = 1, 2, \ldots, k$$

Estimator of Probability Mass Function:

$$\widehat{p_i} = \frac{\sum_{i=1}^{N} I(Y_i = y)}{N}$$

```
x <- c(1, 2, 3, 4, 5)
x
## [1] 1 2 3 4 5
```

$$\Pr[Y = 1] = \frac{\sum_{i}^{N} I(Y_i=1)}{5} = \frac{I(1=1)+I(2=1)+I(3=1)+I(4=1)+I(5=1)}{5}$$
$$= \frac{1+0+0+0+0}{5} = .20$$

$$\Pr[Y = 1] = \frac{\sum_i^N I(Y_i{=}1)}{5} = \frac{I(1{=}1){+}I(2{=}1){+}I(3{=}1){+}I(4{=}1){+}I(5{=}1)}{5}$$
$$= \frac{1{+}0{+}0{+}0{+}0}{5} = .20$$

We can achieve this in R with following code

```
sum(x == 1)/length(x)
```

```
## [1] 0.2
```

Logical Operator in R

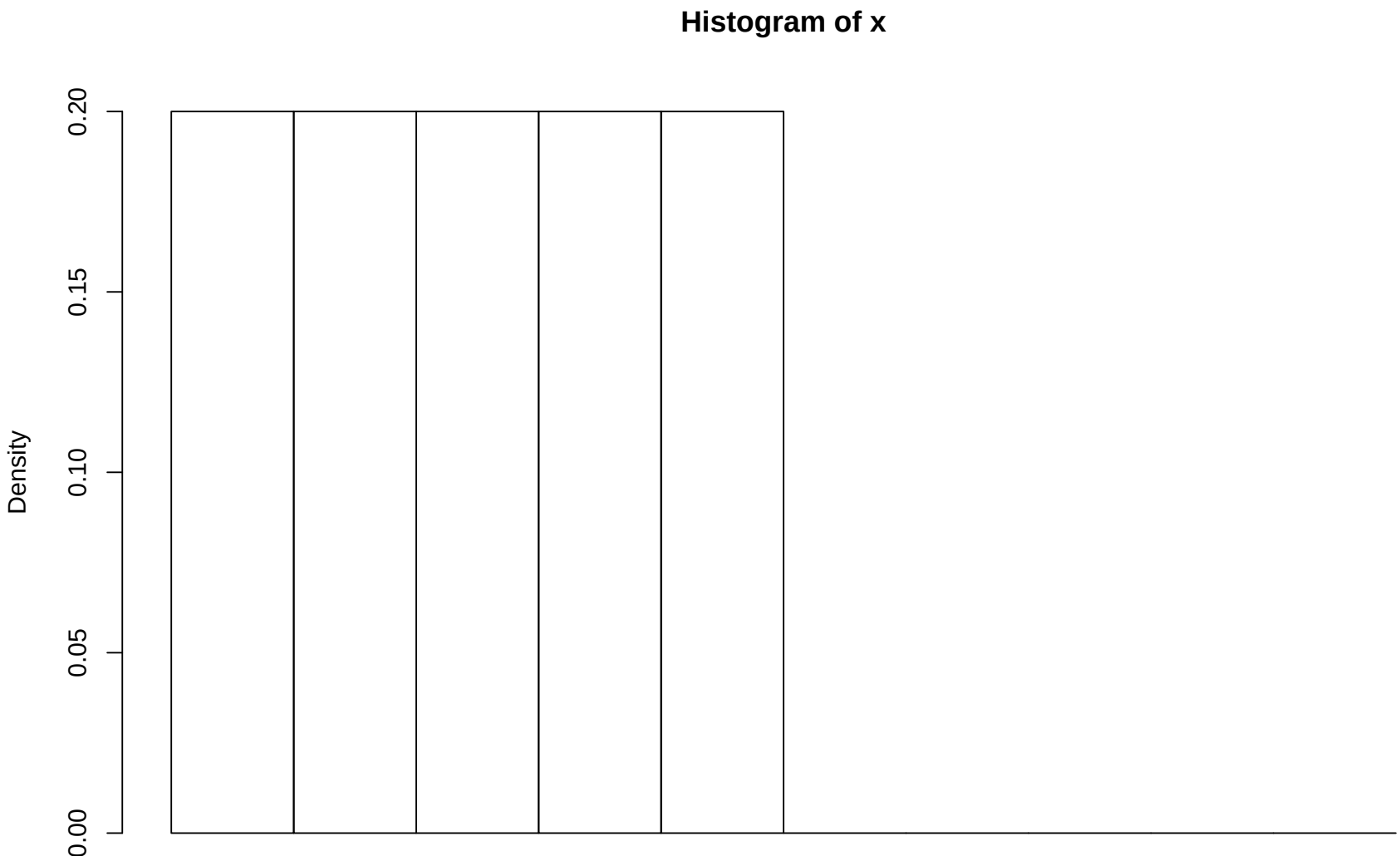
Let's break it down

```
x
## [1] 1 2 3 4 5
x == 1
## [1]  TRUE FALSE FALSE FALSE FALSE
sum(x == 1)
## [1] 1
length(x)
## [1] 5
sum(x == 1)/length(x)
## [1] 0.2
```

Histogram in R (Which value(s) are likely?)

We can repeat the process above for each value, or we can use the built-in program

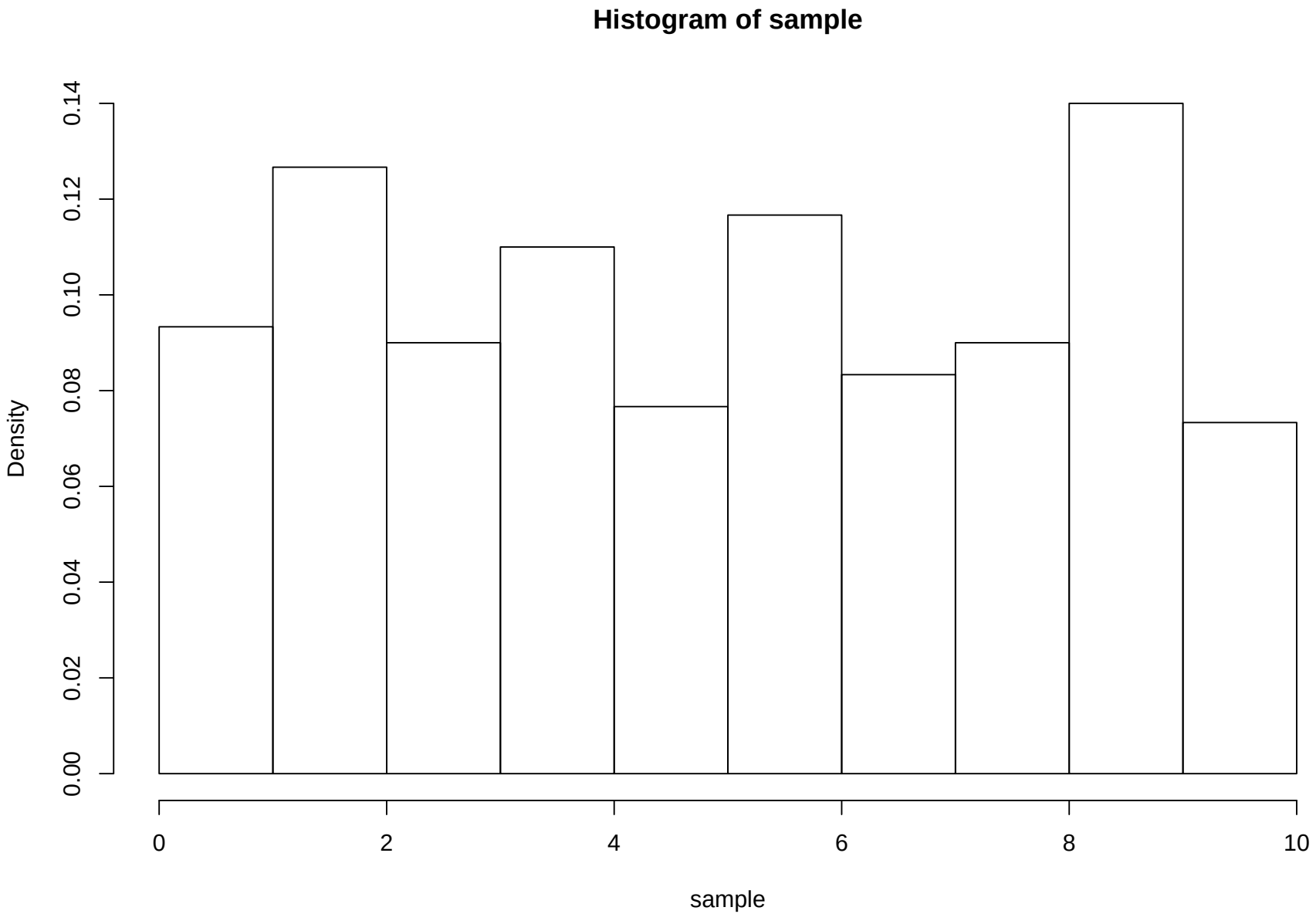
```
hist(x, breaks = c(0:10), probability = TRUE)
```



Example: What is your forecast?

```
##   [1]  8  8  4  4  4  2  6  1 10  2  8  6 10  9 10  9  9
##  [24]  2  5  8  9  9  9  2  6  9  3  2  9  1 10  9  8  7
##  [47]  6  8  7  4  4  8  6  9  8 10  9  9  3  2  2  1  6
##  [70]  1  2  7 10  4  9  2  9  3  2  7  2  1  9 10  6  5
##  [93]  2  2  5  1  9  8  2  4  1  7  4  2  7  6 10  1 10
## [116]  4  5  2  5 10  3  3  7 10  6  5  3  1  8  3  5  6
## [139]  3  1 10  7  8  9  1  5  7  8  1  6  9  7  3  8  6
## [162]  3  3  8  1  5  7  9  2  2  6  2  9  9  7 10  7  1
## [185]  8  8  4  4  8  4  9  6  7  7 10  9  6 10  6  4  2
## [208]  1  9  3  9  3  2  9  5  7  2  4  9  6  2 10  6  5
## [231]  4  6  2  8  9  3  4  5  5  6  9  3  1  4  1  5  1
## [254]  3  7  3  2  9  7  6  6  8  2  8  8  9  7  6  1  2
## [277]  6  4  3  7 10 10  6  4  4  3  9  1  9 10  3  6  8
## [300]  4
```

```
hist(sample, breaks = c(0:10), probability = TRUE)
```



Applications: Football Analytics

Football Analytics Video (Click here)

Starts at 4:02

Histogram and Classification

Classification refers to the problem of predicting a categorical (or discrete) outcome.

Examples

1. Good credit risk/bad credit risk
2. Heart attack/no heart attack,
3. Spam/not spam

Histogram and Classification

Extension:

Common ways to model (to exploit additional information)

1. linear regressions
2. logistic regression,
3. decision trees,
4. random forests,
5. support vector machines,
6. and neural networks

Predictions and Errors

Classification is either correct or incorrect.

In a binary classification problem, there are two types of **misclassification**:

1. false positive, incorrectly predicted positive outcomes
2. false negative, incorrectly predicted negative outcomes

Predictions and Errors

	Actual	Actual
Predicted Outcome	Positive	Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Classification vs Clustering

It is different from clustering problems, which is also trying to identify group memberships

For example, in a marketing setting, we might have demographic information for a number of current or potential customers. We may wish to understand which types of customers are similar to each other by grouping individuals according to their observed characteristics. This is known as a **clustering problem** [an example of unsupervised learning].

Cumulative Distribution Function (CDF)

$$\begin{aligned} F(y) = & \quad Pr[Y \leq y] \\ = & \quad Pr[Y = y_1] + \cdots + Pr[Y = y] \quad \forall Y \leq y \end{aligned}$$

Estimator of CDF

$$\widehat{F}(y) = \frac{\sum_{i=1}^{N} I(Y_i \leq y)}{N}$$

```
x
```

```
## [1] 1 2 3 4 5
```

$$\Pr[Y \leq 4] = \frac{\sum_{i}^{N} I(Y_i \leq 4)}{5} = \frac{I(1 \leq 4) + I(2 \leq 4) + I(3 \leq 4) + I(4 \leq 4) + I(5 \leq 4)}{5}$$
$$= \frac{1+1+1+1+0}{5} = .80$$

How will you program it yourself?

First approach

```
sum(x <= 4)/length(x)
```

```
## [1] 0.8
```

```
x <= 4
```

```
## [1]  TRUE  TRUE  TRUE  TRUE FALSE
```

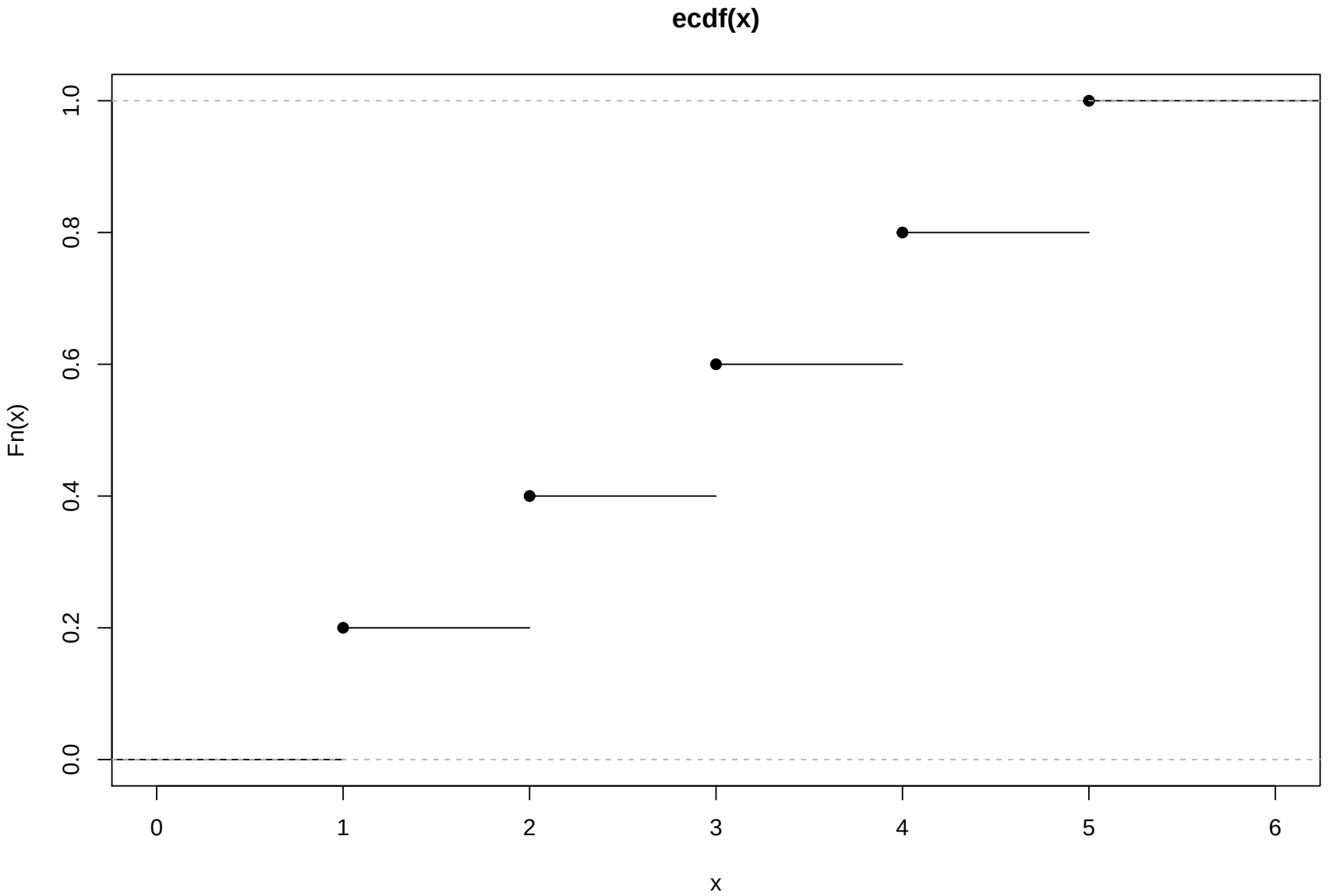
Another approach: use of `ecdf()` in R

```
ecdf(x)(4)
```

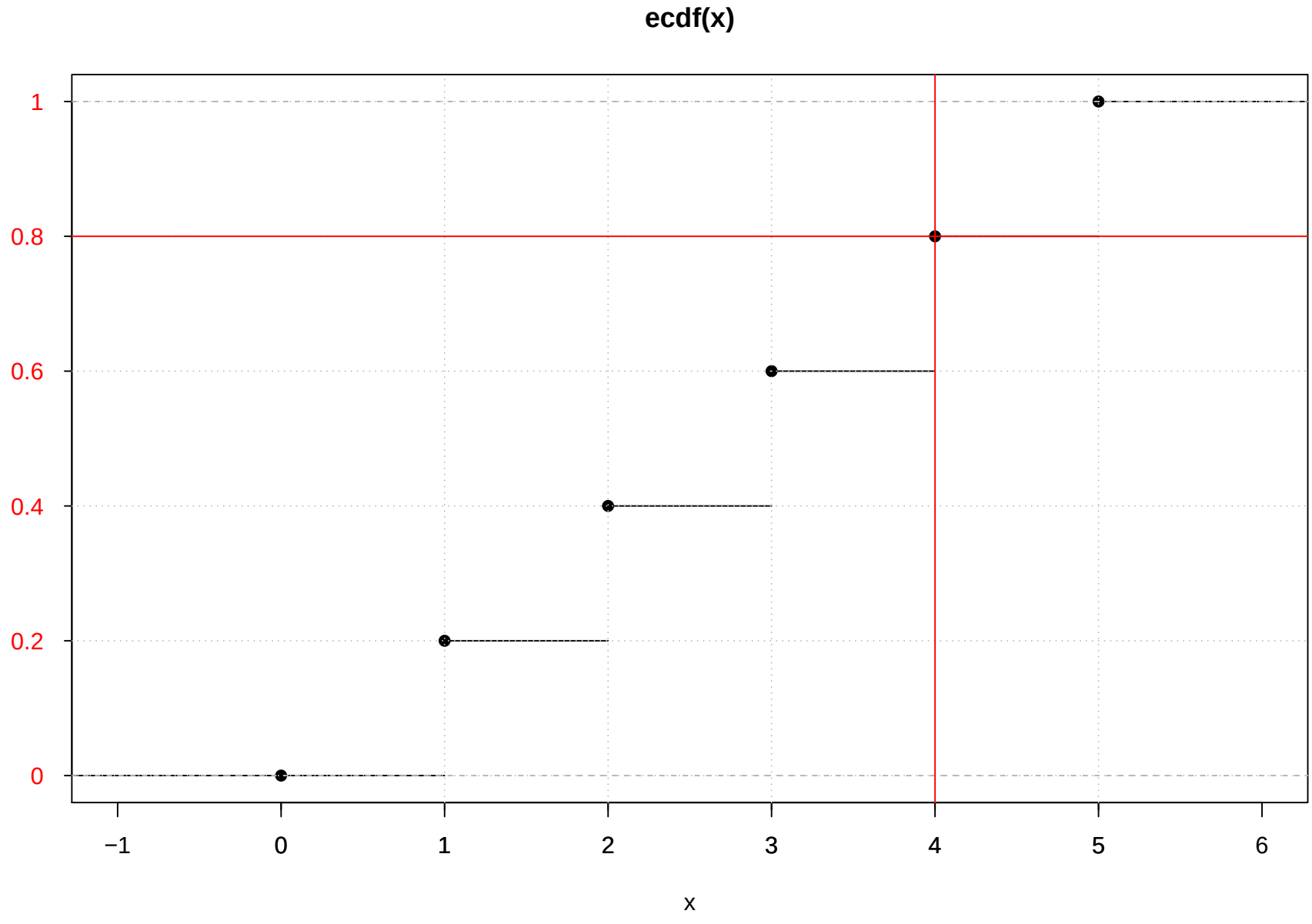
```
## [1] 0.8
```

You can also calculate it for every value, and plot it

```
plot(ecdf(x))
```



How to interpret this graph?



CDF and PMF

For discrete variables, CDF represents all the information about the distribution. For example, you can easily obtain (using the addition rule)

$$\Pr[Y = 4] = \Pr[Y \leq 4] - \Pr[Y \leq 3]$$

Note: You can think of PMF as the marginal change in CDF! (this would become useful way to intuitively define the probability function for continuous variables).

Alternative Ways to chacterize distributions

1. Moment Generating Functions (MGF): Laplace transformation of the density of a random variable (do not exist for many variables)
2. Characteristics Functions (CF): Fourier Transformation of the density
3. Entropy and Information Theory