STA130H1 - Class # 1: Introduction to R, Histograms and Density Functions

Prof. Nathan Taback

2018-01-08

Welcome to STA130H1

- Login to Portal to get the location of your tutorial room (look under My Groups).
- Let's explore the course website

A Brief Introduction to R

What is R?

R Coding basics

Go to console . . .

Histograms and Density Functions

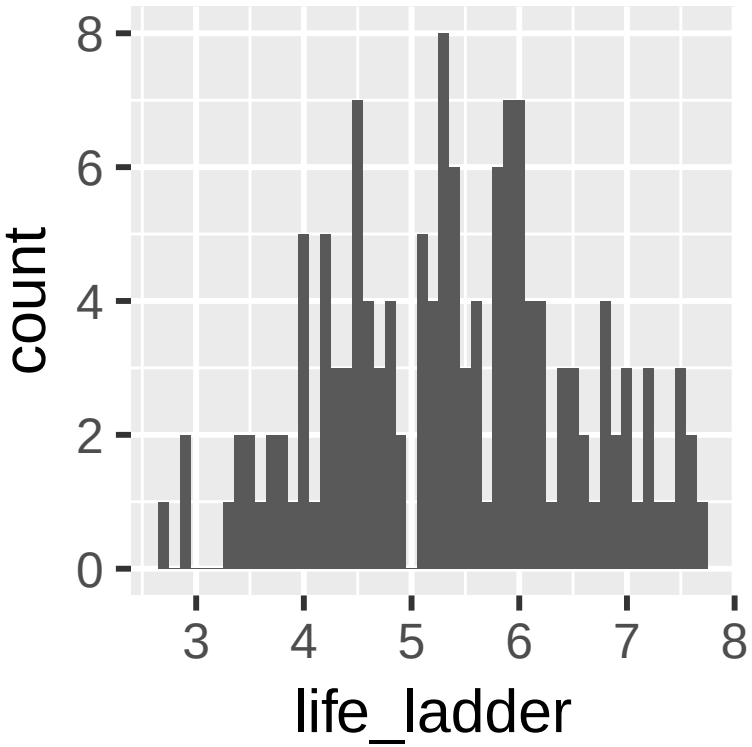
Histograms and Density Functions

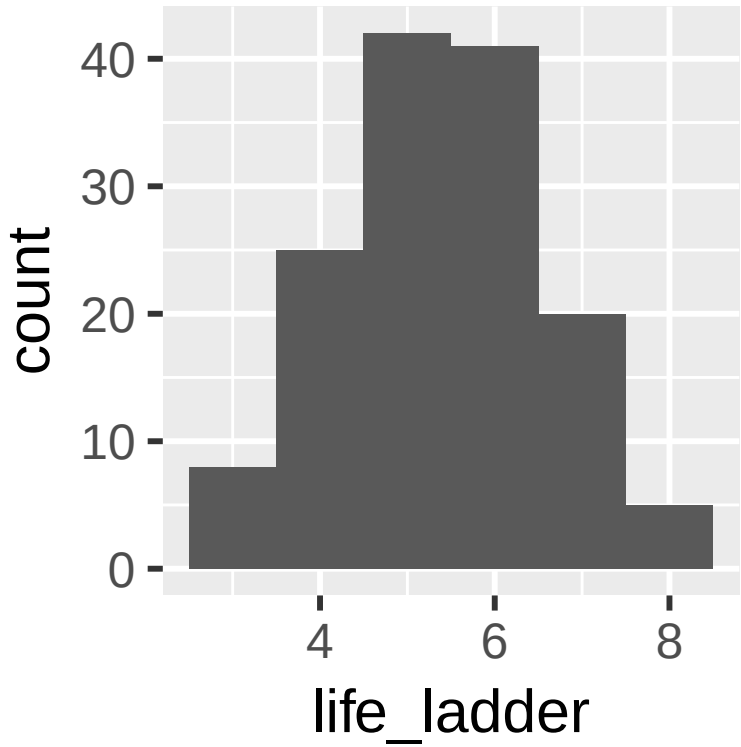
- The histogram of a variable is a graphical method to vizualize the distribution of a single variable.

Histograms and Density Functions

- Different bin width will yield different histograms

```
p1 <- ggplot(data = happinessdata2016, aes(x = life_ladder)
  geom_histogram(binwidth = 0.1)
p2 <- ggplot(data = happinessdata2016, aes(x = life_ladder)
  geom_histogram(binwidth = 1.0)
grid.arrange(p1,p2,nrow = 1)
```





Mathematical Definition of Histogram

- The bins of the histogram are the intervals:

$$[x_0 + mh, x_0 + (m+1)h).$$

  x_0 is the origin, $m = \ldots, -1, 0, 1, \ldots$ indexes the bins, and $h = (x_0 + (m+1)h) - (x_0 + mh)$ is the bin width.
- The bins can be used to construct rectangles with width h and height $\hat{f}(x)$.
- The area of these rectangles is $h\hat{f}(x)$.
- The area of the rectangles is the same as the proportion of data in the same bin as x.

Example - Mathematical Definition of Histogram

```
dat <- data_frame(x = c(1,2,2.5,3,7))
dat$x
```

```
[1] 1.0 2.0 2.5 3.0 7.0
```

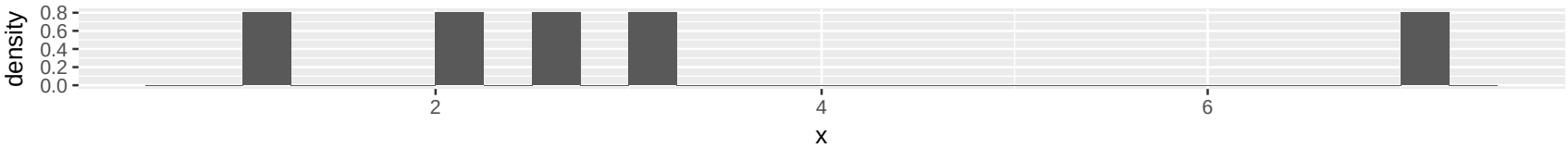
Let $x_0 = 0.5, h = 0.25, m = 1, \ldots, 29$

```
seq(0.5,7.5,by = 0.25)
```

```
 [1] 0.50 0.75 1.00 1.25 1.50 1.75 2.00 2.25 2.50 2.75 3.00
[15] 4.00 4.25 4.50 4.75 5.00 5.25 5.50 5.75 6.00 6.25 6.50
[29] 7.50
```

The bins are: $[0.50, 0.75), [0.75, 1.00), [1.00, 1.25), \ldots, [7.25, 7.50)$.

Example - Mathematical Definition of Histogram



y	count	x	xmin	xmax	density
0.0	0	0.625	0.50	0.75	0.0
0.0	0	0.875	0.75	1.00	0.0
0.8	1	1.125	1.00	1.25	0.8
0.0	0	1.375	1.25	1.50	0.0
0.0	0	1.625	1.50	1.75	0.0
0.0	0	1.875	1.75	2.00	0.0

Mathematical Definition of Histogram

- Suppose we have data: $X_1, X_2, \ldots, X_n$.
- Let $\#\{X_i \text{ in same bin as } x\}$ be the number of data points X_i in the same bin as x.
- Let n be the total number of data points. So, $\frac{\#\{X_i \text{ in same bin as } x\}}{n}$ is the proportion of data in the same bin as x.
- Area of rectangle containing $x \approx \frac{\#\{X_i \text{ in same bin as } x\}}{n}$.
- $$h\hat{f}(x) = \frac{\#\{X_i \text{ in same bin as } x\}}{n}.$$

Mathematical Definition of Histogram

$$\hat{f}(x) = \frac{1}{hn} \#\{X_i \text{ in same bin as } x\}$$

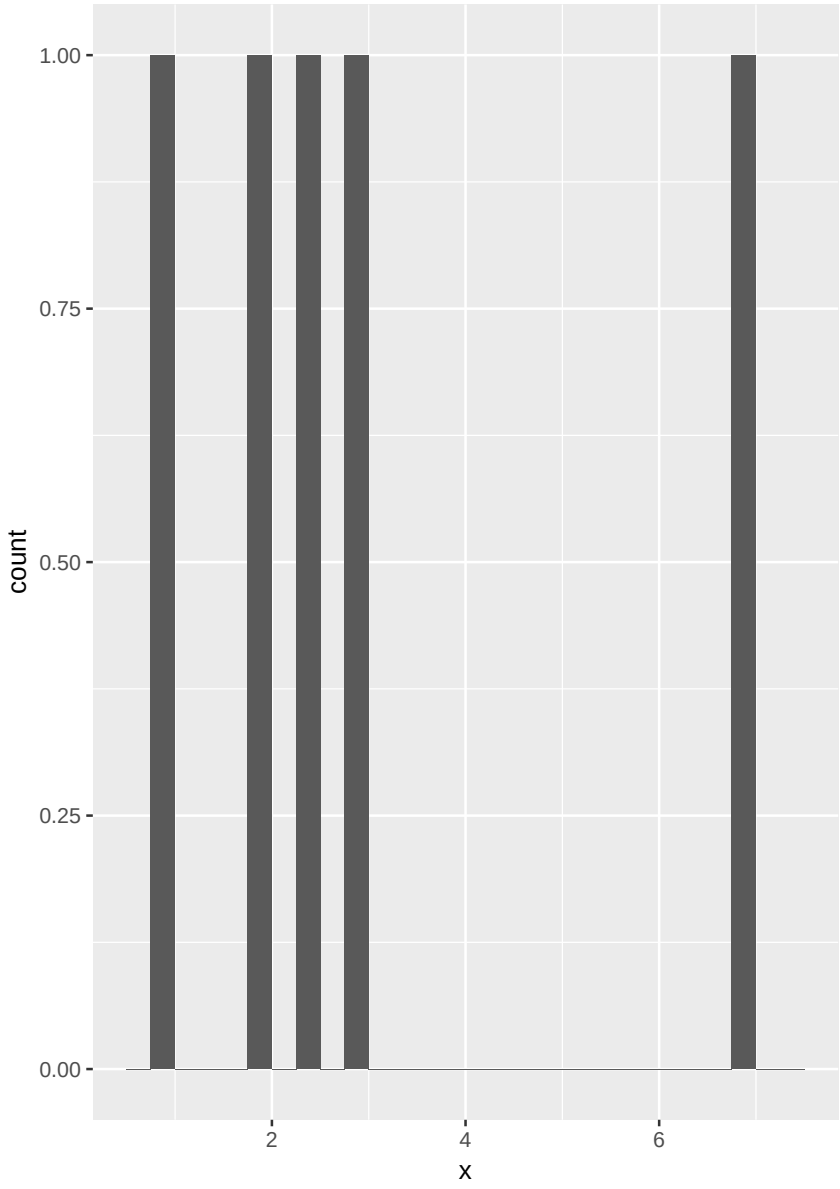
is called the **histogram estimator**.

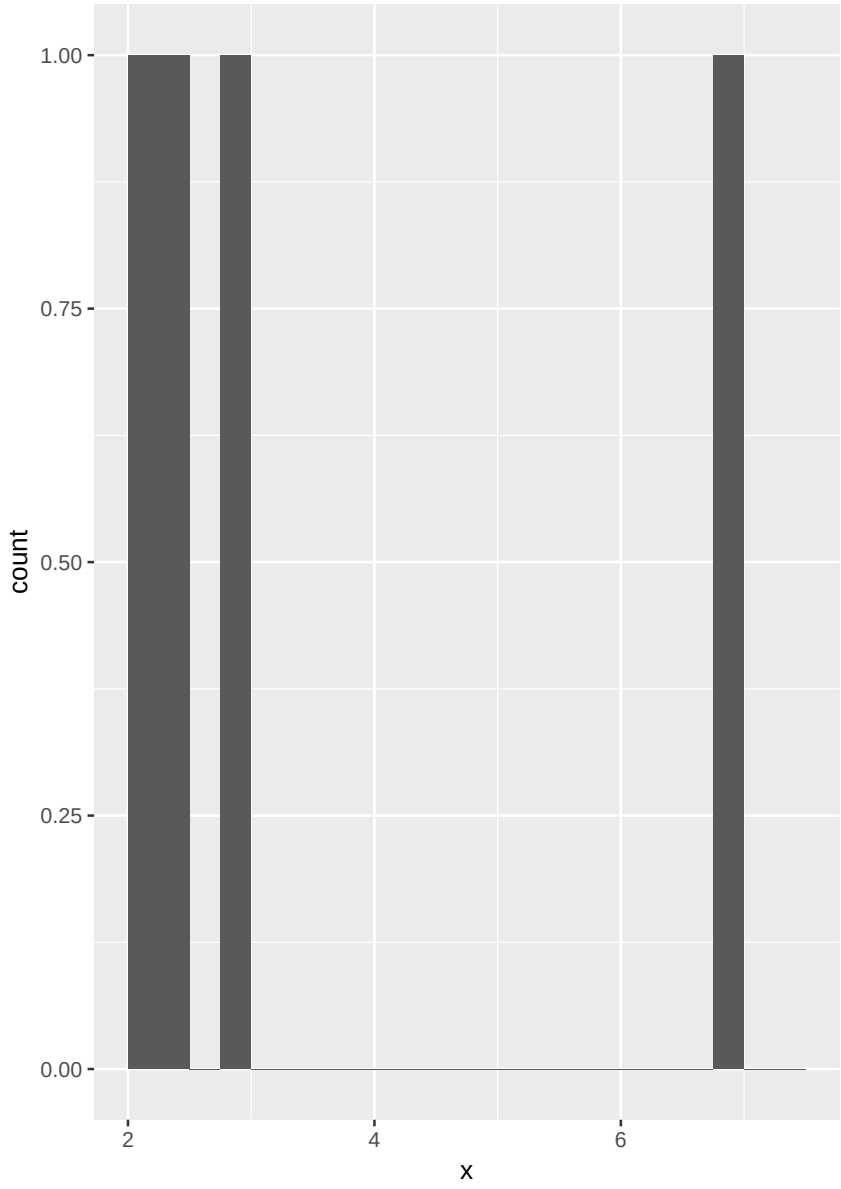
$\hat{f}(x)$ is an estimate of the density at a point x.

To construct the histogram we have to choose an origin x_0 and bin width h.

Choosing Origin and Bin Width in R

Same bin width but different origin





Naive Estimator of Density

The histogram can be centered on a point x.

$$\hat{f}(x) = \frac{1}{2hn} \#\{X_i \in (x-h, x+h)\}$$

This estimate is often called the **naive estimator** of the density.

This can be expressed by defining the weight function $w(x)$ by

$$w(x) = \begin{cases} 1/2 & \text{if } |x| < 1 \\ 0 & \text{if } |x| \geq 1. \end{cases}$$

Then the naive estimator can be written

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - X_i}{h}\right).$$

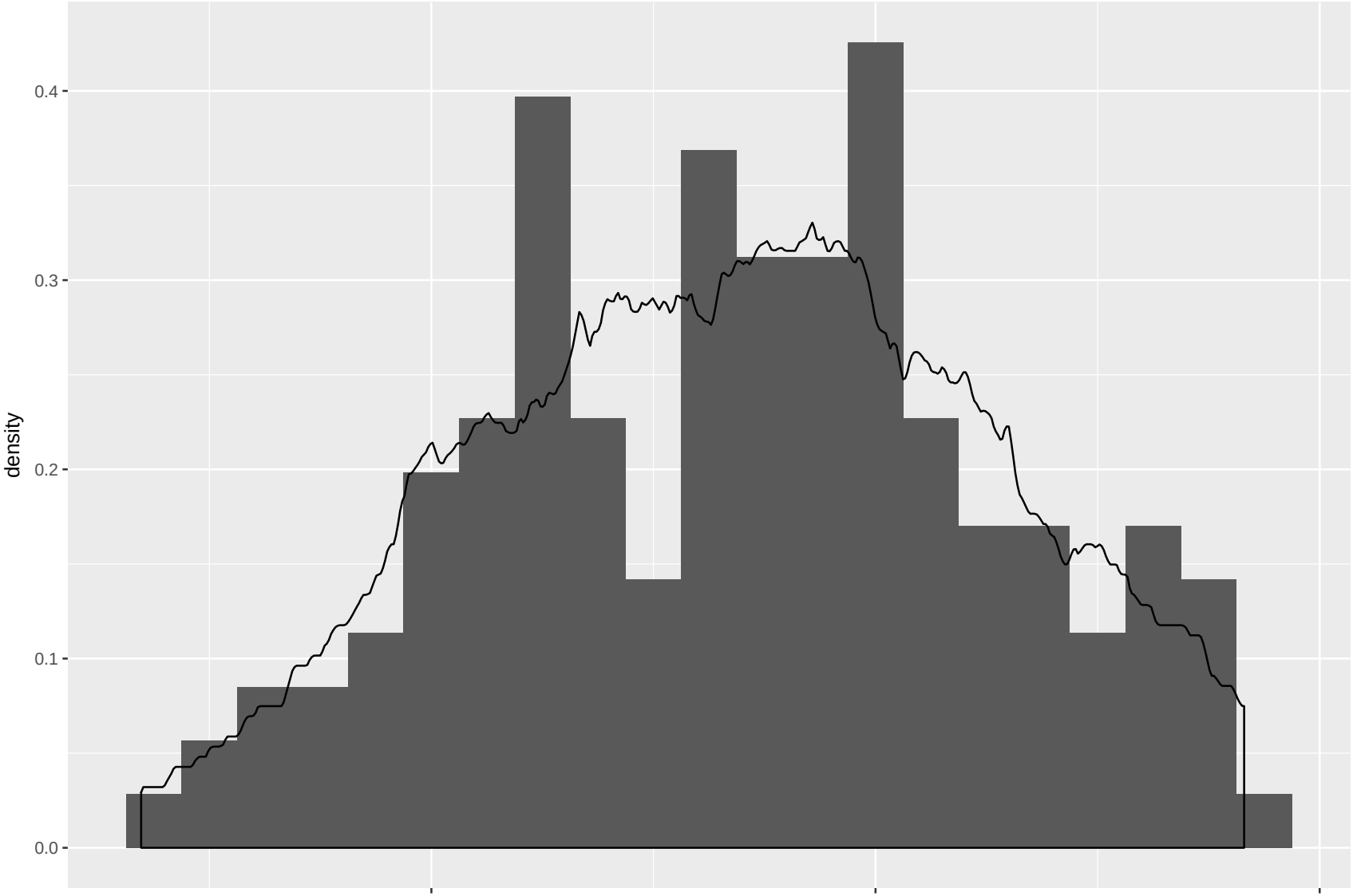
Naive Estimator of Density

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - X_i}{h}\right).$$

- Consider the histogram constructed from data using bins of width $2h$.
- Assume that x is at the centre of one of the histogram bins then the naive estimate will be the same as the y value of the histogram estimate.

Naive Estimator in R

```
ggplot(data = happinessdata2016, aes(x = life_ladder, ..der
  geom_histogram(binwidth = 0.25) + geom_density(kernel = '
```



Kernel Estimator

Replace the weight function is by a kernel function $K(x) \geq 0$ which satisfies $\int_{-\infty}^{\infty} K(x) = 1$. The **kernel estimator** of the density function is defined by

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

The Gaussian kernel is a popular choice

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right), -\infty < x < \infty.$$

Kernel Estimator in R

```
p_gauss <- ggplot(data = happinessdata2016, aes(x = life_la
  geom_histogram(binwidth = 0.25) + geom_density(kernel = '
p_rect <- ggplot(data = happinessdata2016, aes(x = life_lad
  geom_histogram(binwidth = 0.25) + geom_density(kernel = '
grid.arrange(p_gauss,p_rect)
```