ASDS Statistics, YSU, Fall 2020

Lecture 05

Michael Poghosyan

24 Sep 2020

Contents

Last Lecture Recap

- What we need to have to plot a Histogram?

Last Lecture Recap

- What we need to have to plot a Histogram?
- Give the Definition of the Frequency Histogram.

Last Lecture Recap

- What we need to have to plot a Histogram?
- Give the Definition of the Frequency Histogram.
- Give the Definition of the Density Histogram.

Histogram Example

Example: Plot the Frequency, Relative Frequency and Density Histograms for

$$0,\ 4,\ 2,\ 2,\ 0,\ 0.5,\ 1,\ 3$$

Example

To draw the Density Histogram in **R**, we will use the *freq=FALSE* parameter in the *hist* command.

Example

To draw the Density Histogram in **R**, we will use the *freq=FALSE* parameter in the *hist* command.

We use here the *discoveries* Standard Dataset from **R**, which gives us the numbers of "great" inventions and scientific discoveries in each year from 1860 to 1959:

Example

To draw the Density Histogram in **R**, we will use the *freq=FALSE* parameter in the *hist* command.

We use here the *discoveries* Standard Dataset from **R**, which gives us the numbers of “great” inventions and scientific discoveries in each year from 1860 to 1959:

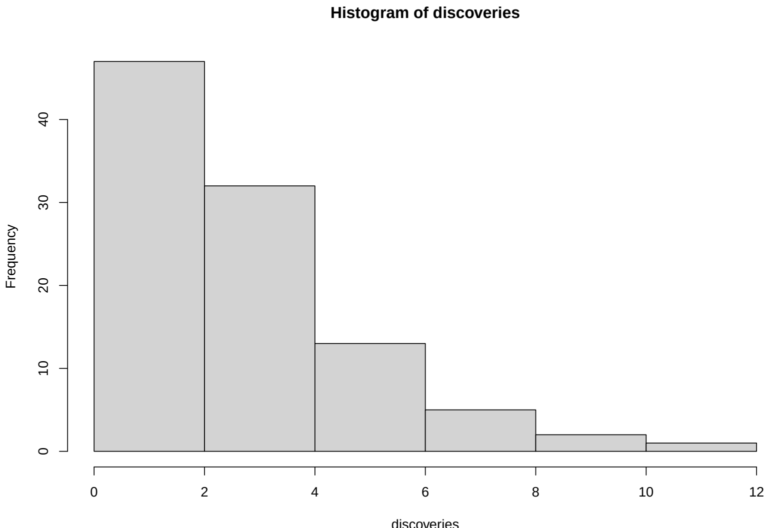
```
discoveries
```

```
## Time Series:
## Start = 1860
## End = 1959
## Frequency = 1
##   [1]  5  3  0  2  0  3  2  3  6  1  2  1  2  1  3  3
##  [26] 12  3 10  9  2  3  7  7  2  3  3  6  2  4  3  5
##  [51]  3  6  5  8  3  6  6  0  5  2  2  2  6  3  4  4
##  [76]  2  2  1  3  4  2  2  1  1  1  2  1  4  4  3  2
```

Example

First, the Frequency Histogram:

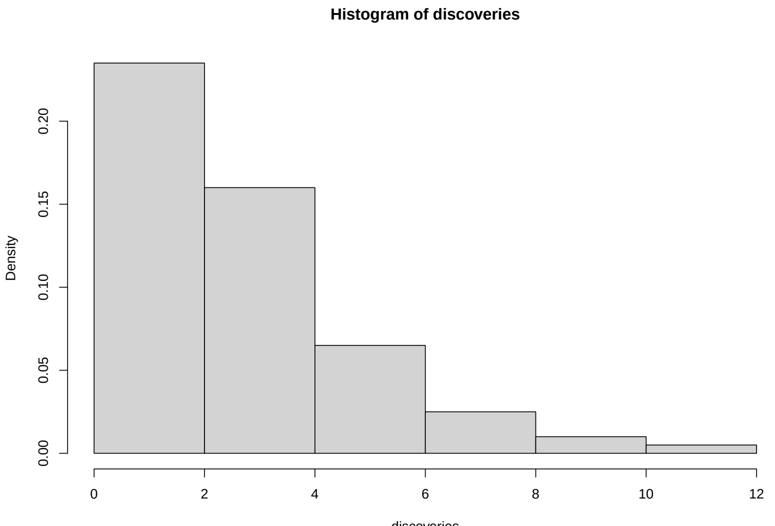
```
hist(discoveries)
```



Example

Now, the Density Histogram:

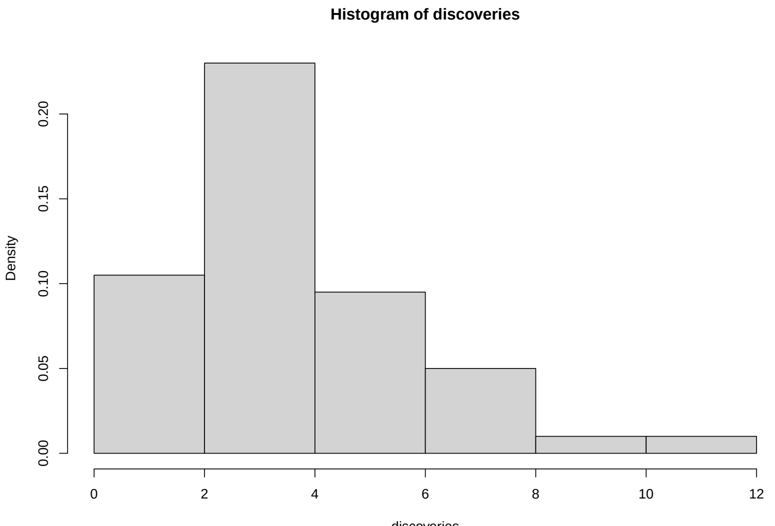
```r
hist(discoveries, freq = FALSE)
```



Example

Finally, the Density Histogram with the Bins left-endpoints included:

```
hist(discoveries, freq = FALSE, right = FALSE)
```



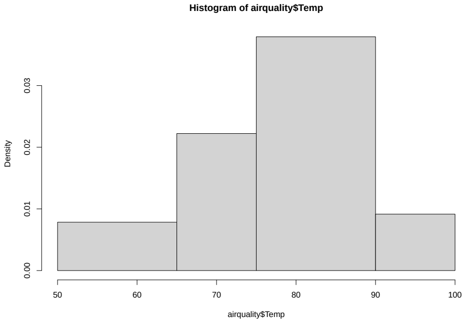
Example

Now let us change the default bins for a Histogram.

Example

Now let us change the default bins for a Histogram. We can use the following - first define the vector of our class interval (Bins) endpoints: (note that you need to cover all Datapoints!)

```
bins.endpoitns <- c(50, 65, 75, 90, 100)
hist(airquality$Temp, breaks = bins.endpoitns)
```



Notes

- By default, if we give custom bins with non-equal lengths, **R** is plotting the Density Histogram!

Notes

- By default, if we give custom bins with non-equal lengths, **R** is plotting the Density Histogram!
- You can give the *breaks* parameter either the vector of Bins' endpoints or the number of (equal-length) intervals

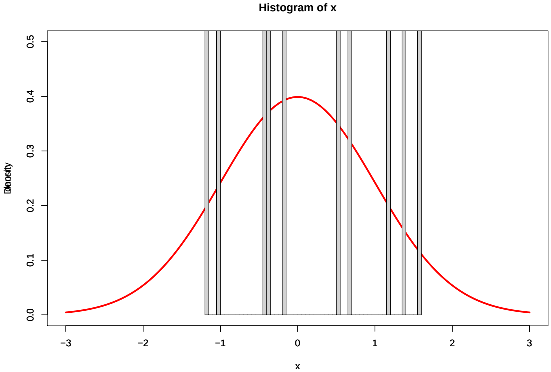
Estimation of the PDF through the Density Histogram

As it was stated above, the Density Histogram is an approximation (estimate) of the PDF of the Data unknown Distribution. To check this, let us take a synthetic Dataset from the Distribution we know:

Estimation of the PDF through the Density Histogram

As it was stated above, the Density Histogram is an approximation (estimate) of the PDF of the Data unknown Distribution. To check this, let us take a synthetic Dataset from the Distribution we know:

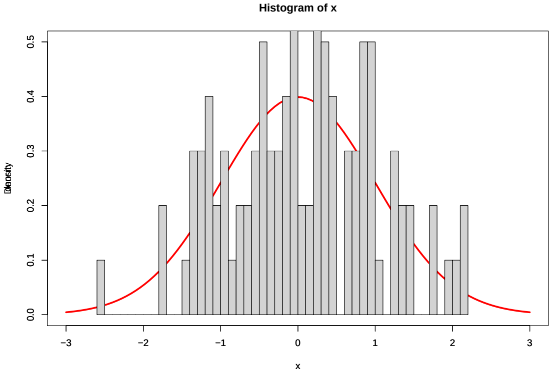
```
plot(dnorm, lwd = 3, col= "red", xlim=c(-3,3), ylim=c(0,0.5))
x <- rnorm(10)
par(new = TRUE)
hist(x, breaks = 40, freq = FALSE, xlim=c(-3,3), ylim=c(0,0.5))
```



Estimation of the PDF through the Density Histogram

As it was stated above, the Density Histogram is an approximation (estimate) of the PDF of the Data unknown Distribution. To check this, let us take a synthetic Dataset from the Distribution we know:

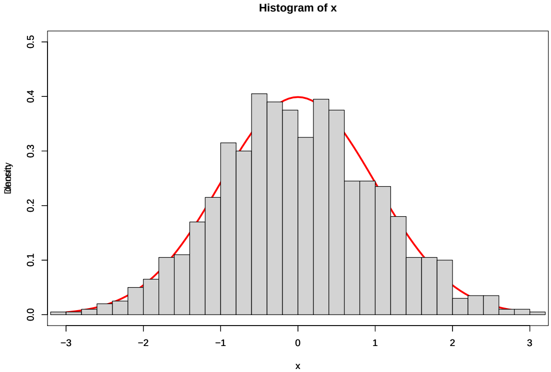
```
plot(dnorm, lwd = 3, col= "red", xlim=c(-3,3), ylim=c(0,0.5))
x <- rnorm(100)
par(new = TRUE)
hist(x, breaks = 40, freq = FALSE, xlim=c(-3,3), ylim=c(0,0.5))
```



Estimation of the PDF through the Density Histogram

As it was stated above, the Density Histogram is an approximation (estimate) of the PDF of the Data unknown Distribution. To check this, let us take a synthetic Dataset from the Distribution we know:

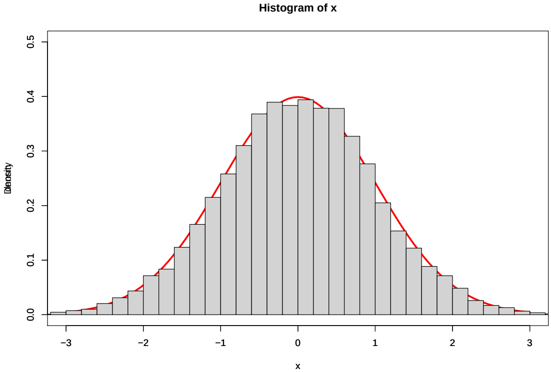
```
plot(dnorm, lwd = 3, col= "red", xlim=c(-3,3), ylim=c(0,0.5))
x <- rnorm(1000)
par(new = TRUE)
hist(x, breaks = 40, freq = FALSE, xlim=c(-3,3), ylim=c(0,0.5))
```



Estimation of the PDF through the Density Histogram

As it was stated above, the Density Histogram is an approximation (estimate) of the PDF of the Data unknown Distribution. To check this, let us take a synthetic Dataset from the Distribution we know:

```
plot(dnorm, lwd = 3, col= "red", xlim=c(-3,3), ylim=c(0,0.5))
x <- rnorm(10000)
par(new = TRUE)
hist(x, breaks = 40, freq = FALSE, xlim=c(-3,3), ylim=c(0,0.5))
```



Choosing Bin sizes correctly

It is important to choose the Bin sizes (lengths of the Bin, class, intervals) wisely. Otherwise you will skip some info or you will not get any valuable info.

Choosing Bin sizes correctly

It is important to choose the Bin sizes (lengths of the Bin, class, intervals) wisely. Otherwise you will skip some info or you will not get any valuable info.

Let us use another **R** standard dataset to show the effect of the choice of the bin size: *precip*. This Dataset shows the average amount of precipitation (rainfall) in inches for each of 70 United States (and Puerto Rico) cities.

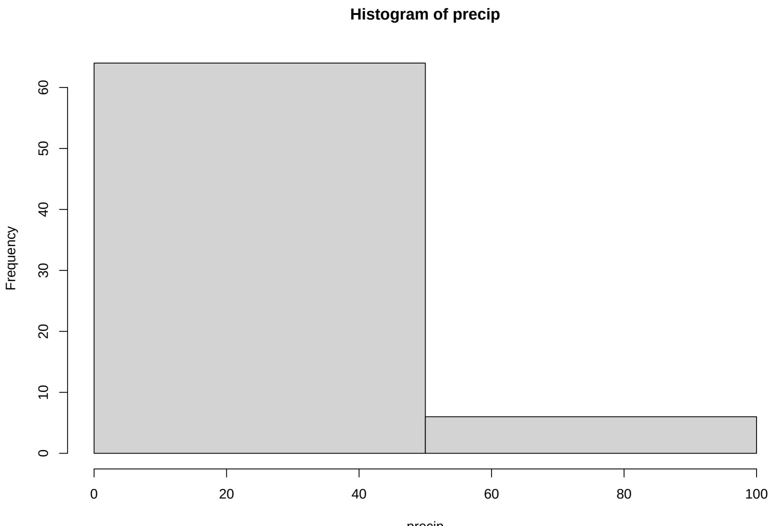
```r
head(precip)
```

```
##       Mobile       Juneau      Phoenix Little Rock Los Ange
##         67.0         54.7          7.0        48.5        1
```

Version 1, Small bins

Here, we just use 2 bins:

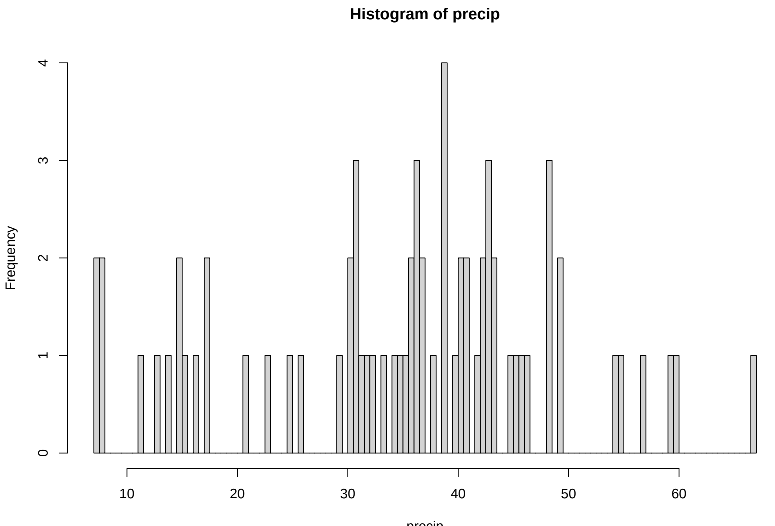
```
hist(precip, breaks = 2)
```



Version 2, large bins

Here, we use 200 bins:

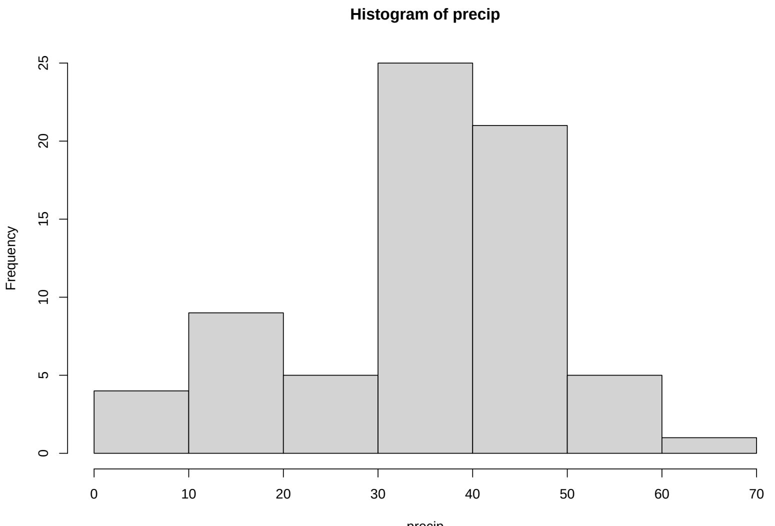
```r
hist(precip, breaks = 200)
```



Version 2, large bins

Now, the default:

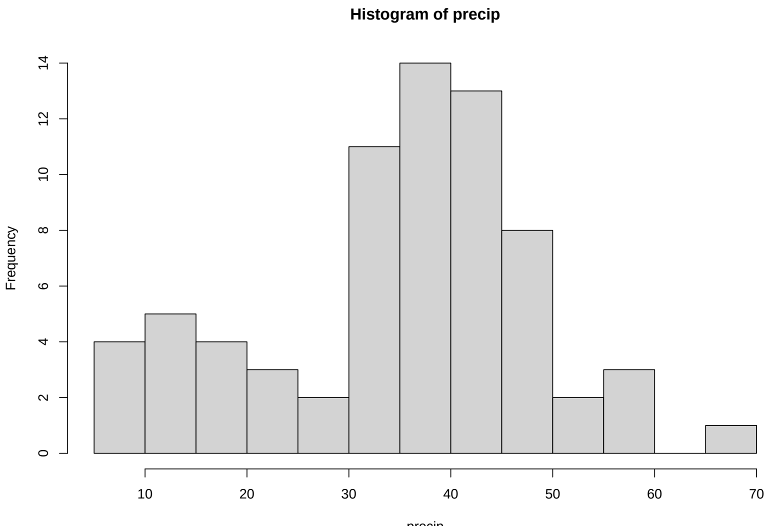
```r
hist(precip)
```



Version 3

Now, let us change to 20 bin intervals:

```r
hist(precip, breaks = 20)
```



Choosing the Bin Length

In fact, choosing the correct Bin width is not an easy job. See, for example, the Histogram Wiki page.

Differences between the Barplot and Histogram

- Can you give some differences?

Differences between the Barplot and Histogram

- Can you give some differences?

Here are some:

- *Barplot*'s rectangles widths are arbitrary, do not mean anything, rectangles are not adjacent; *Histogram*'s rectangles are adjacent, and the choice of the Bin widths is changing the graph

Differences between the Barplot and Histogram

- Can you give some differences?

Here are some:

- *Barplot*'s rectangles widths are arbitrary, do not mean anything, rectangles are not adjacent; *Histogram*'s rectangles are adjacent, and the choice of the Bin widths is changing the graph
- *Barplot* is for a categorical or Discrete Data, *Histogram* is for both Discrete and Continuous

Differences between the Barplot and Histogram

- Can you give some differences?

Here are some:

- *Barplot*'s rectangles widths are arbitrary, do not mean anything, rectangles are not adjacent; *Histogram*'s rectangles are adjacent, and the choice of the Bin widths is changing the graph
- *Barplot* is for a categorical or Discrete Data, *Histogram* is for both Discrete and Continuous
- We can exactly reconstruct the Dataset from the *Barplot*, but not the *Histogram*

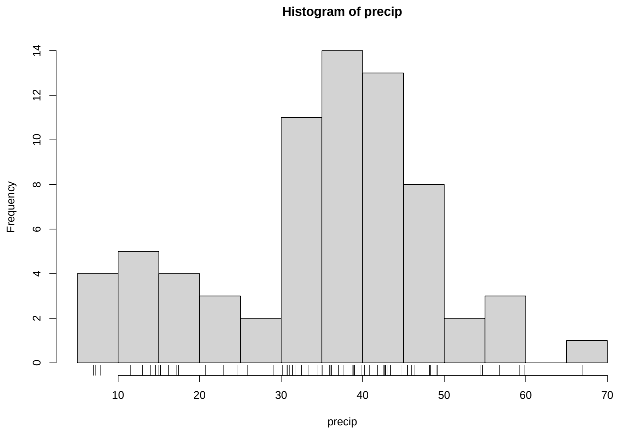
Addition to the Histogram

Nice addition to your Histogram Plot is to add, in some way, the Datapoints:

Addition to the Histogram

Nice addition to your Histogram Plot is to add, in some way, the Datapoints:

```
hist(precip, breaks = 20)
rug(precip)
```



What we can see from the Histogram

If we will not look at the Histogram as being an estimate for the unknown Distribution behind the Data, and if we will just try to get some info about our Dataset, Histogram is helping us to say if the Data:

What we can see from the Histogram

If we will not look at the Histogram as being an estimate for the unknown Distribution behind the Data, and if we will just try to get some info about our Dataset, Histogram is helping us to say if the Data:

- is symmetric about some point or is skewed to the left or right

What we can see from the Histogram

If we will not look at the Histogram as being an estimate for the unknown Distribution behind the Data, and if we will just try to get some info about our Dataset, Histogram is helping us to say if the Data:

- is symmetric about some point or is skewed to the left or right
- is spread out or concentrated at some point

What we can see from the Histogram

If we will not look at the Histogram as being an estimate for the unknown Distribution behind the Data, and if we will just try to get some info about our Dataset, Histogram is helping us to say if the Data:

- is symmetric about some point or is skewed to the left or right
- is spread out or concentrated at some point
- has some gaps

What we can see from the Histogram

If we will not look at the Histogram as being an estimate for the unknown Distribution behind the Data, and if we will just try to get some info about our Dataset, Histogram is helping us to say if the Data:

- is symmetric about some point or is skewed to the left or right
- is spread out or concentrated at some point
- has some gaps
- has values far apart from others, has outliers (anomalies)

What we can see from the Histogram

If we will not look at the Histogram as being an estimate for the unknown Distribution behind the Data, and if we will just try to get some info about our Dataset, Histogram is helping us to say if the Data:

- is symmetric about some point or is skewed to the left or right
- is spread out or concentrated at some point
- has some gaps
- has values far apart from others, has outliers (anomalies)
- is unimodal, bimodal or multimodal

KDE

Another estimate for the unknown Distribution PDF is the **Kernel Density Estimator**, KDE.

KDE

Another estimate for the unknown Distribution PDF is the **Kernel Density Estimator**, KDE. It is, in some sense, the smoothed version of the Histogram: Histogram is a piecewise-constant function, with jumps, so it is not a smooth function.

KDE

To define the KDE, we first choose a smooth Kernel function $K(t)$, here, a function with

$$K(t) \geq 0, t \in \mathbb{R}, \qquad \text{and} \qquad \int_{-\infty}^{+\infty} K(t)dt = 1.$$

KDE

To define the KDE, we first choose a smooth Kernel function $K(t)$, here, a function with

$$K(t) \geq 0, t \in \mathbb{R}, \qquad \text{and} \qquad \int_{-\infty}^{+\infty} K(t)dt = 1.$$

For example, we can take the Gaussian Kernel

$$K(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-t^2/2}, \qquad t \in \mathbb{R},$$

or any other PDF.

KDE

To define the KDE, we first choose a smooth Kernel function $K(t)$, here, a function with

$$K(t) \geq 0, t \in \mathbb{R}, \qquad \text{and} \qquad \int_{-\infty}^{+\infty} K(t)dt = 1.$$

For example, we can take the Gaussian Kernel

$$K(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-t^2/2}, \qquad t \in \mathbb{R},$$

or any other PDF.

Next, one defines the Kernel Density Estimator with Kernel K as

$$KDE_K(x) = KDE(x) = \frac{1}{nh} \cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$