 Lagunita is retiring and will shut down at 12 noon Pacific Time on March 31, 2020. A few courses may be open for self-enrollment for a limited time. We will continue to offer courses on other online learning platforms; visit <http://online.stanford.edu>.


Course > EDA: Examining Distributions > One Quantitative Variable: Graphs >
Statistics Package Exercise: Creating and Describing Histograms

 Bookmark this page

Statistics Package Exercise: Creating and Describing Histograms

Learning Objective: Relate measures of center and spread to the shape of the distribution, and choose the appropriate measures in different contexts.

We will use the Best Actor Oscar winners (1970-2013) to learn how to create a histogram using a statistics package, and practice what we've learned about describing the histogram.

Click here  to see the entire dataset.

-     

R Instructions

To open R with the dataset preloaded, right-click here and choose "Save Target As" to download the file to your computer. Then find the downloaded file and double-click it to open it in R.

The data have been loaded into the data frame

```
actor_age
```

. Enter the command

```
actor_age
```

to see the data. The only variable (column title) in the data frame

```
actor_age
```

is

```
Age
```

.

To create a histogram of the actors' age data we can use the following code:

```
hist(actor_age$Age)
```

Notice the default settings in R are to use the variable name, in this case

```
actor_age$Age
```

in the title and x-axis label. In addition, the default y-axis is "Frequency." A good graphic is a well-labeled graphic. We can modify all of these settings with a few additional parameters added into the

```
hist()
```

command.

For example, if you want to add an x-axis label and remove the title of the histogram, use the following code:

```
hist(actor_age$Age, xlab="Age of Best Actor Oscar Winners  
(1970-2013)", main="")
```

If you want to modify the x-axis and y-axis label and the title use the following code:

```
hist(actor_age$Age, xlab="Age of Best Actor Oscar Winners  
(1970-2013)", ylab="Number of Actors", main="Best Actor  
Oscar Winners Ages")
```

Try replacing the x-axis label, y-axis label, and title with your own modified labels.

Another possible modification to the histogram is the number of bins. R uses an algorithm to determine the optimal number of bins based on the data, but in some cases you may want to

modify the number of bins yourself. You can add the parameter

```
breaks=
```

into the

```
hist()
```

command which will tell R to make that many breaks in the data. R will not always do the exact number of breaks if it is not possible, but it will provide a close approximation. For example, let's try 8 breaks:

```
hist(actor_age$Age, breaks=8, xlab="Age of Best Actor Oscar  
Winners (1970-2013)", main=" ")
```

Try replacing the number of “breaks” with 5 or 20. Which histogram gives the right amount of detail-neither too little nor too much?

Note: Using R-If you are looking at a graph in R, you may find that the command window (the one labeled "R Console") is not responsive. That is because the graph window is the “active” window. Click on the command window to make it the active window. In addition, notice that R will always overwrite your current graphic with a new graphic. Enter the code

```
x11()
```

and press prior to each new graphic command and R will create a new window for that graphic.

Learn By Doing (1/1 point)

In the textbox below, describe the distribution of the ages of the Best Actor Oscar winners. Be sure to address shape, center, spread and outliers. When you are done, compare your answer to ours.

Your Answer:

symmetric, unimodal, right skewed, centered around 40, range is 45 years, 1 outlier

Our Answer:

Shape: the distribution is skewed right. This means that most actors receive the best acting Oscar at a relatively younger age (before age 48), and fewer at an older age. **Center:** The distribution seems to be centered at around 42-43. This means that about half the actors are 42 or younger when they receive the Oscar, and about half are older. **Spread:** The age distribution ranges from about 30 to about 75. The entire dataset is covered, then, by a range of 45 years. It should be noted, though, that there is one high outlier at around age 75, and the rest of the data ranges only from 30 to 60. **Outliers:** As mentioned above, there is one high outlier at around age 75.

[Resubmit](#)[Reset](#)[Open Learning Initiative](#)

Unless otherwise noted this work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

© All Rights Reserved