MITx: 14.310x Data Analysis for Social Scientists

Hel

Bookmarks

- Module 1: The Basics of R and Introduction to the Course
- Entrance Survey
- Module 2: Fundamentals of Probability, Random Variables, Distributions, and Joint Distributions
- **Module 3: Gathering and Collecting Data, Ethics, and Kernel Density Estimates**

**Gathering and Collecting Data**
Finger Exercises due Oct 17, 2016 at 05:00 IST

**Summarizing and Describing Data**

Module 3: Gathering and Collecting Data, Ethics, and Kernel Density Estimates > Module 3: Homework > Question 11 - 16

Bookmark

Now, we are going to focus on how the distribution of the Adolescent Fertility Rate has changed from 1960 to 2000. The following code in R plots the histogram of these two variables in the same graph. Please take a look at the code and try to understand what it is doing.

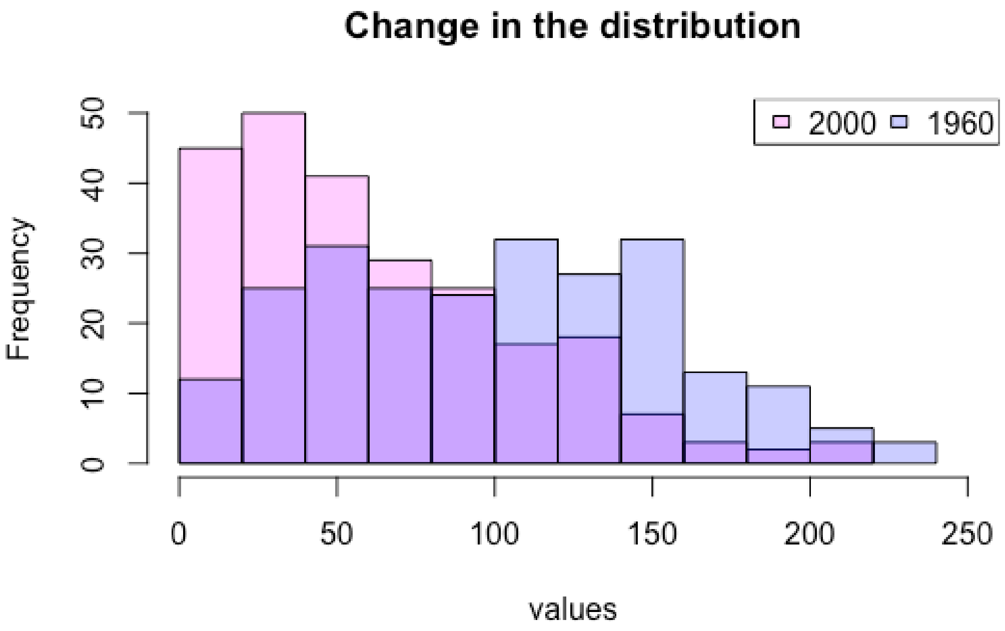
```
p1 <- hist(teenager_fr$X1960)
p2 <- hist(teenager_fr$X2000)
plot( p2, col=rgb(1,0,1,1/4), xlim = c(0, 250), main = "Change
in the distribution", xlab = "values")
plot( p1, col=rgb(0,0,1,1/4), xlim = c(0, 250), add = TRUE)
legend("topright", ncol = 2, legend = c("2000", "1960"),
fill=c(rgb(1,0,1,1/4), rgb(0,0,1,1/4)), text.width = 20)
png("histogram")
```

Here is the figure that this code has produced:

Finger Exercises due Oct 17, 2016 at 05:00 IST

**Module 3: Homework**
Homework due Oct 10, 2016 at 05:00 IST

▸ Exit Survey



## Question 11

(1/1 point)

The color of the bins was chosen using the option rgb(0,0,1,1/4)? What does the fourth argument in this vector represent in the plot?

○ a. The red level in the color of the bin.

○ b. The green level in the color of the bin.

○ c. The blue level in the color of the bin.

● d. The level of transparency in the color of bin. ✓

EXPLANATION

If you go to the R documentation for the rgb() function, you will see that the first arguments represent the level of red, green and blue, respectively. The fourth argument corresponds to alpha that sets the level of transparency in the color. By changing this parameter we are able to see the two histograms in the region where there is overlap.

*You have used 1 of 2 submissions*

## Question 12

(1/1 point)

As you can see, we have a certain number of bins in the figure. Go to the R documentation and look for the option in the command **hist** that will allow you to change the number of bins in the figure. What is the option that will allow you to do this?

breaks ✓ **Answer:** breaks

EXPLANATION

If you go to the R documentation, in particular just by typing help(hist) you will be able to see that the option breaks will allow you to change the number of bins. In particular, the documentation states that: breaks corresponds to one of the following: (i) a vector giving the breakpoints between histogram cells, (ii) a function to compute the vector of breakpoints, (iii) a single number giving the number of cells for the histogram, (iv) a character string naming an algorithm to compute the number of cells (see ‘Details’), (v) a function to compute the number of cells.In the last three cases the number is a suggestion only; the breakpoints will be set to pretty values. If breaks is a function, the x vector is supplied to it as the only argument.

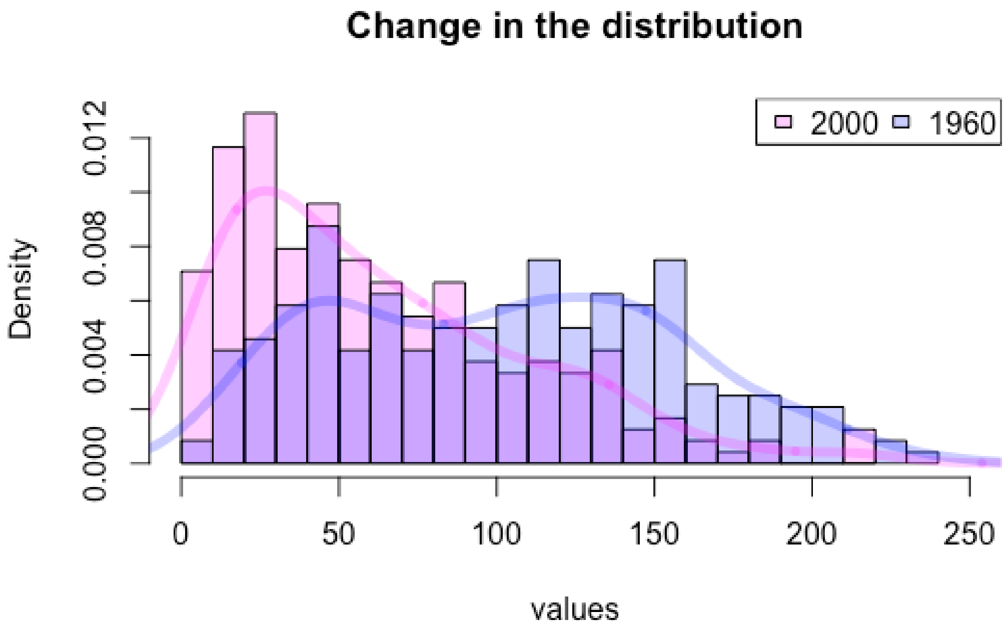
*You have used 1 of 2 submissions*

Now, we are going to add some kernels to the histogram. The kernels were done using the command density and all the default options in R. Again, take a look at the code, run it on your computer and try to understand what it is doing.

```
p1 <- hist(teenager_fr$X1960, freq = FALSE, breaks = 20)
p2 <- hist(teenager_fr$X2000, freq = FALSE, breaks = 20)
p1$counts = p1$density
p2$counts = p2$density
p3 <- density(teenager_fr$X1960, na.rm = TRUE)
p4 <- density(teenager_fr$X2000, na.rm = TRUE)

plot( p2, col=rgb(1,0,1,1/4), xlim = c(0, 250), main = "Change in the
distribution", xlab = "values", ylab = "Density")
plot( p1, col=rgb(0,0,1,1/4), xlim = c(0, 250), add = TRUE)
lines( p4, col=rgb(1,0,1,1/4), xlim = c(0, 250), lwd = 5)
lines(p3, col=rgb(0,0,1,1/4), xlim = c(0, 250), lwd = 5)
legend("topright", ncol = 2, legend = c("2000", "1960"),
fill=c(rgb(1,0,1,1/4), rgb(0,0,1,1/4)), text.width = 20)
legend("topright", ncol = 2, legend = c("2000", "1960"),
fill=c(rgb(1,0,1,1/4), rgb(0,0,1,1/4)), text.width = 20)
```

The below figure is produced by running this code:



## Question 13

(1/1 point)

As it was stated before, the plot was done using the default options in R. For the kernel, the default option is to use gaussian. There are other options that the user can choose when running the density command in R. Of the following list, which one doesn't have a bell-shaped? In other words, which one doesn't put less weight on the observations in the extremes of the bandwidth?

○ a. gaussian

○ b. epanechnikov

● c. rectangular ✓

○ d. triangular

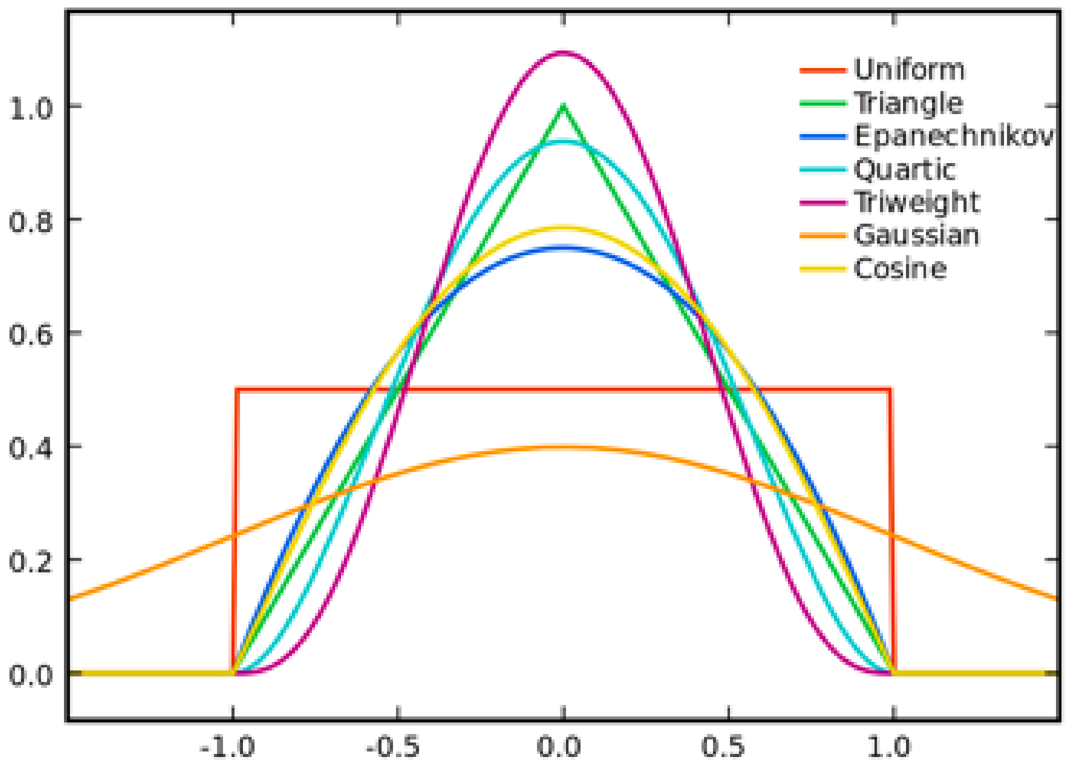
○ e. biweight

○ f. cosine

○ g. optcosine

**EXPLANATION**

The following plot shows the different shapes of the kernel functions. As you can see, the only one without a bell-shaped function is the rectangular one. This kernel is also called the uniform kernel.



*You have used 1 of 2 submissions*

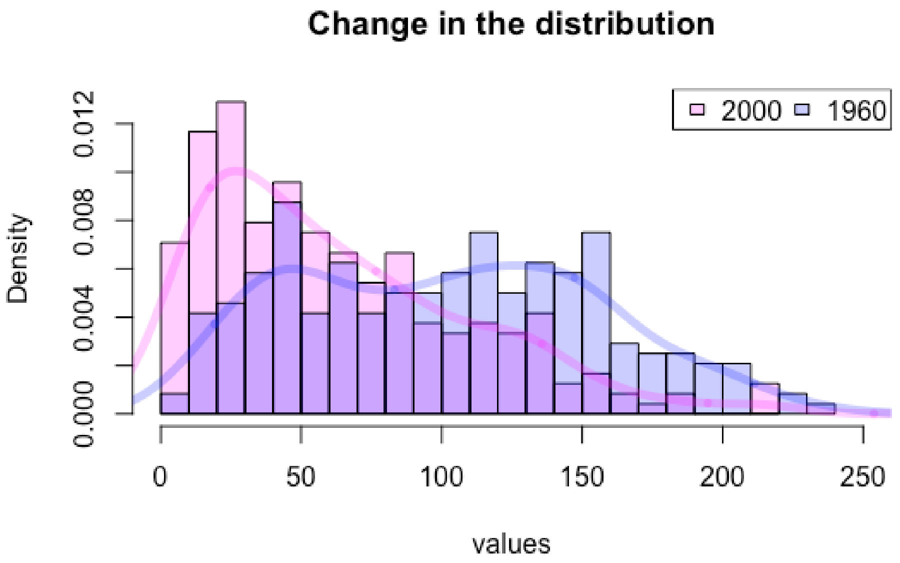
## Question 14

(1/1 point)

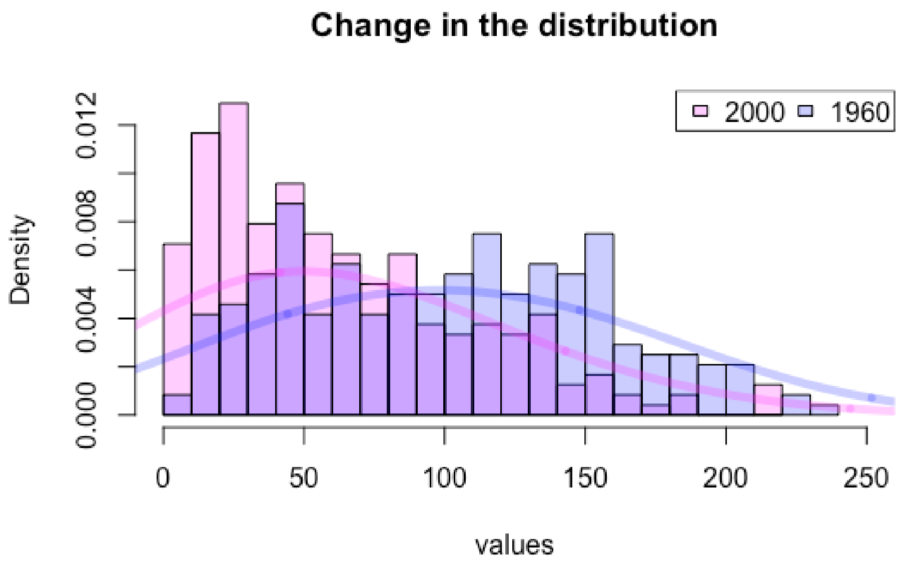
The following plots were made by changing the bandwidth of the kernel function in R. Which one of them was made with the largest bandwidth?

○ a. It is not possible to tell just by looking at the figure.

○ b.



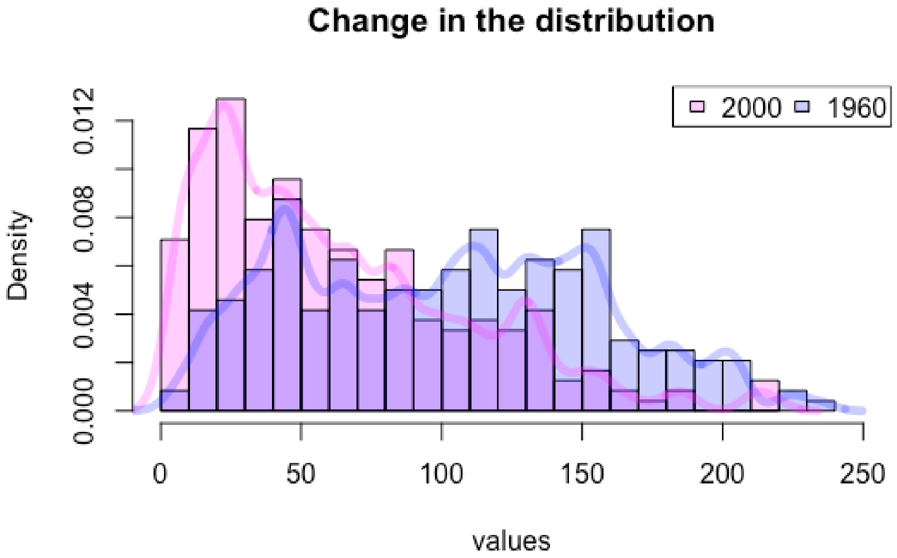
● c.


Change in the distribution
2000
1960
Density
0.000
0.004
0.008
0.012
values
0
50
100
150
200
250

✓

○

d.

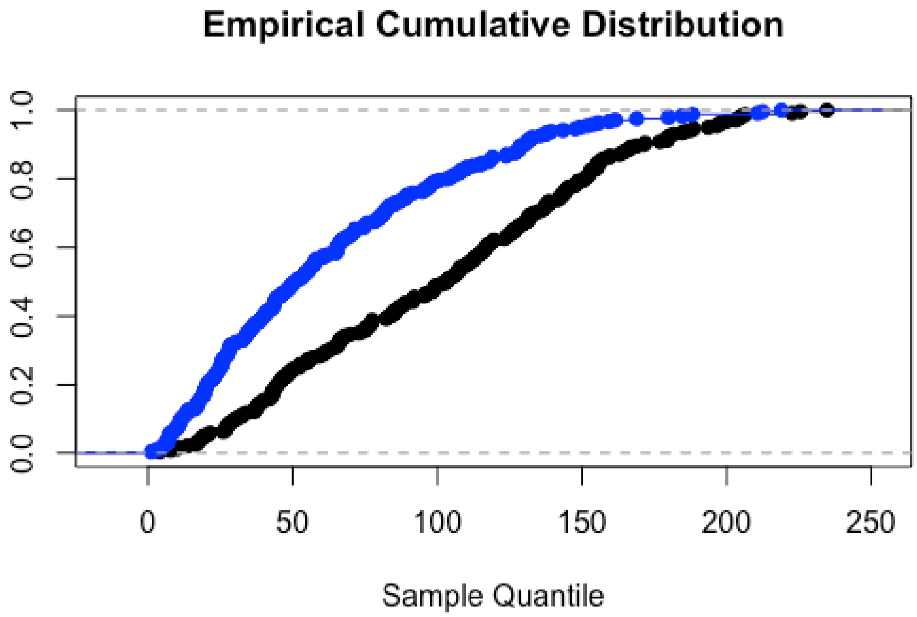


EXPLANATION

As Professor Duflo discussed in the class, the optimal bandwidth balances a trade-off between bias and variance. The larger the bandwidth, the larger the bias of the density calculated at each point is, and the smoother the function looks. As you can see, the kernel in answer (c) is the furthest from the histogram, which suggests it is the one that was constructed with the largest bandwidth.

*You have used 1 of 2 submissions*

One of the things that Professor Duflo also discussed in the lecture, was the construction of the Empirical Cumulative Distribution (ECD). The following figures shows the ECD for the Adolescent Fertility Rate in the World in 1960 and in 2000. However, as you can see the person who made the graph forgot to properly label it.



## Question 15

(1/1 point)

Can you infer from the histograms that were plotted before, which one corresponds to the Adolescent Fertility Rate in 2000 and which one to the same indicator in 1960? (Select all that apply)

☑ a. Blue corresponds to 2000 ✓

□ b. Black corresponds to 2000

□ c. Blue corresponds to 1960

☑ d. Black corresponds to 1960 ✓

□ e. It is not possible to tell from the plot

✓

EXPLANATION

Looking at the histogram you know that since the empirical pdf of the 2000 is at the left of the one in 1960, then it is accumulating mass faster. Thus, the empirical cdf of 2000 is always above the empirical cdf of 1960. For this reason the blue series corresponds to the distribution in 2000, and the black to the distribution in 1960.

*You have used 1 of 2 submissions*

## Question 16

(1/1 point)

Can you infer from the figure whether the distribution used to construct the black series satisfy the First Order Stochastic Dominance property over the distribution used to construct the blue series?

● Yes ✓

○ No

EXPLANATION

From the figure you can see that the black line is always below the blue line. This is precisely the definition of first order stochastic dominance since for all value of k, it is satisfied that $Pr(x \leq k|black) \leq Pr(x \leq k|blue)$.

*You have used 1 of 1 submissions*

POWERED BY OPENedX