Lecture 5: Let's look at some data: Exploratory Data analysis

Prof. Esther Duflo

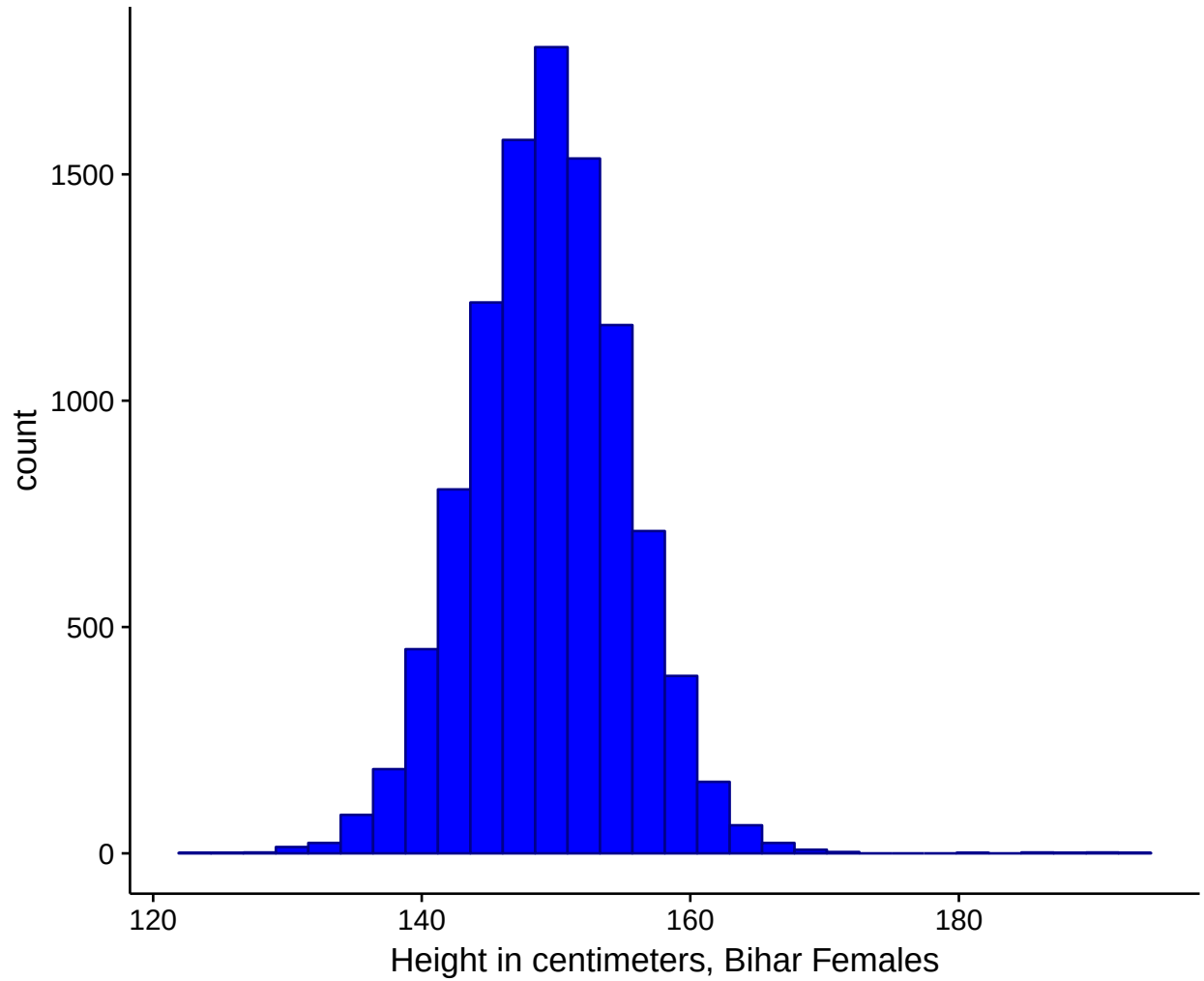
14.310x

OK, we have data, shall we look at it?

- Plotting Histograms
- Plotting Kernels density
- Comparing two (or more) distribution
- Plotting estimates of CDFs
- Bivariate distributions

- A histogram is a rough estimate of the probability distribution function of a continuous variable.
- We obtain it by binning the data (typically in bins of equal size) and simply counting the number of observations within each bin.
- Formally, a histogram is a function that counts the number of observations that fit into each bin. Let n be the total number of observations and k the number of bins, the histogram meets the definition: $n = \sum_{i=1}^{k} m_i$
- Graph of a histogram: We draw, for each bin, a rectangle proportional to the number of such cases.
- You can also divide by the total number of observations to obtain the density: the proportion of cases that within each bin.

Example: Women's height in Bihar



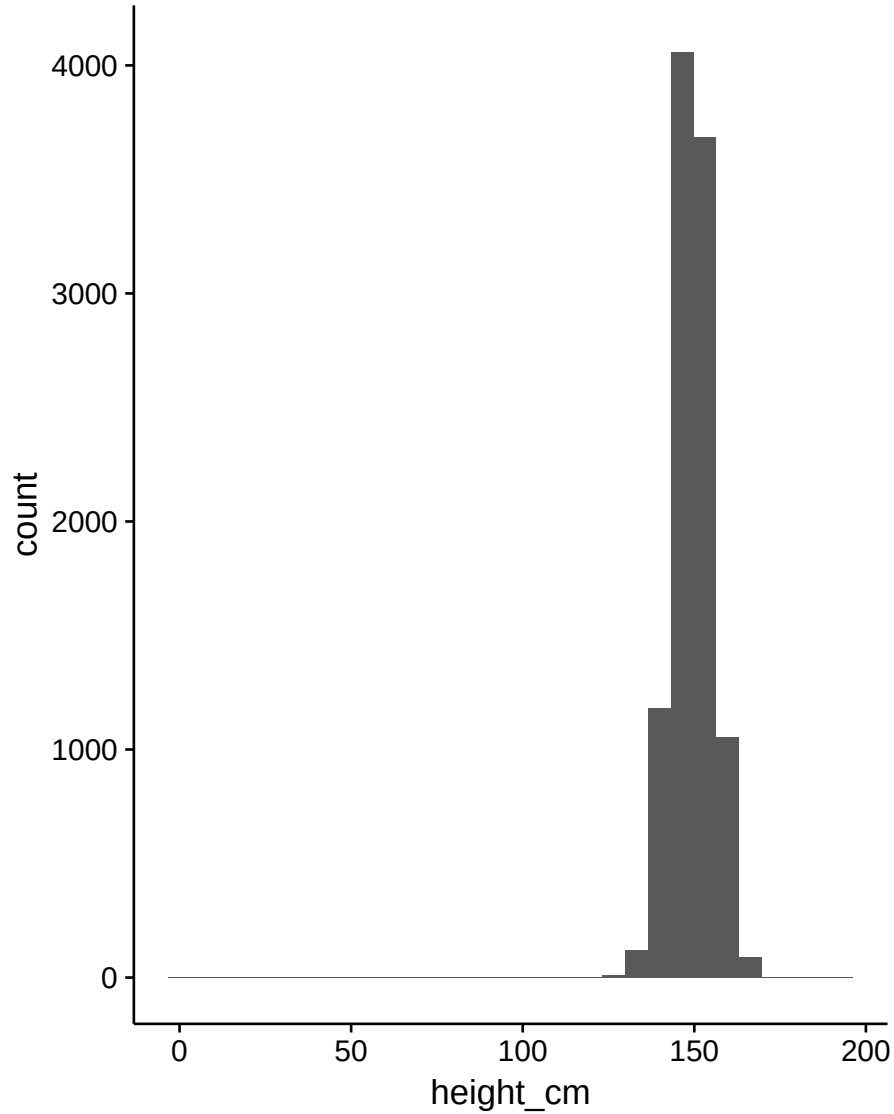
Our universe in R

- In this class we will use R studio
- And make heavy use of packages developed by Hadley Wickam (and described in *R for data science*)
- Specifically
  - ggplot2 to make graphs
  - tidyverse and dplyr to manipulate data sets
- ggplot2
  - ggplot2 implements the "grammar of graphics"
  - you specify the data, a "geom function" and an aesthetic
  - ggplot(data, aes(mapping))+ geomfunction+options
  - A very handy page to print is the ggplot cheat sheet `https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf`

The code for the basic histogram

```
require(cowplot)

#load in the data
bihar_data<-read_csv("data/Bihar_sample_data.csv")

#have a look
print(bihar_data)

#keep only females
bihar_adult_females <-filter(bihar_data, adult==1,female==1)

#have a look

print(bihar_adult_females)

# default histogram in ggplot

ggplot(bihar_adult_females, aes(height_cm))+
  geom_histogram()
ggsave("output/bihar_raw.pdf")

```

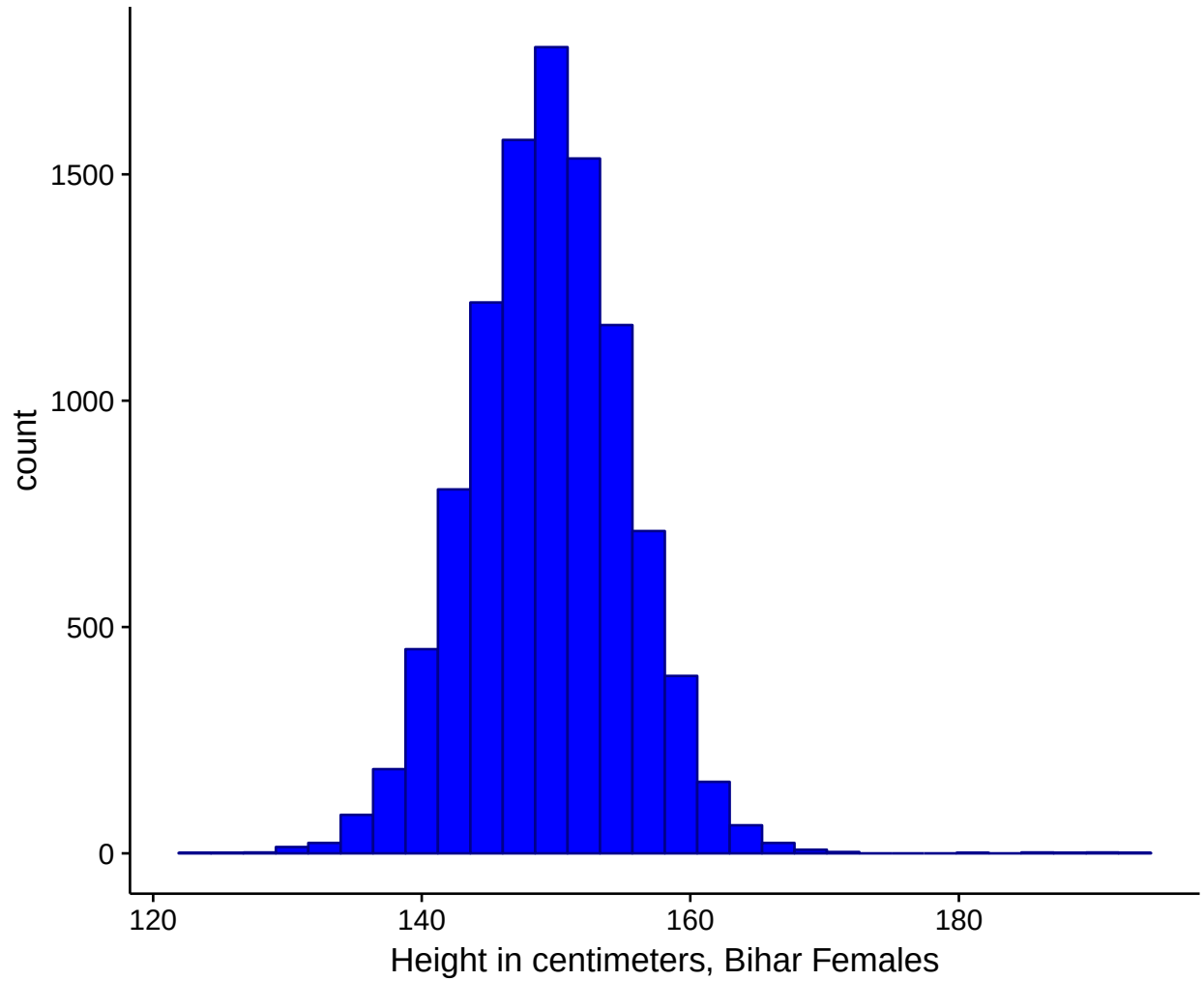
Mmmmm..Not too pretty



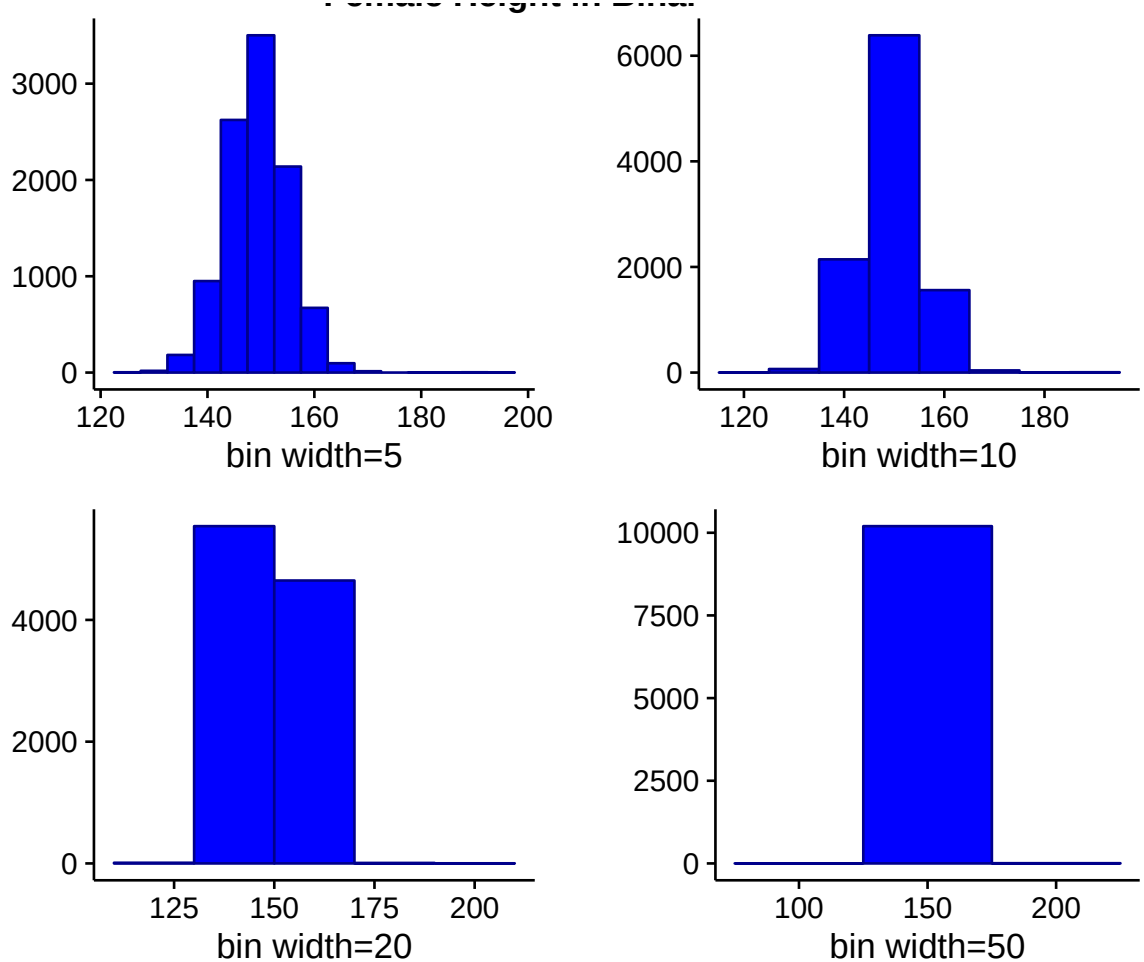
Let's try again

```
25  #some people look like they are very small: filtering
26  bihar_adult_females_trunc <-filter(bihar_adult_females, height_cm>120, height_cm<200)
27
28  #Plotting again, with a nicer label, and some color
29  ggplot(bihar_adult_females_trunc, aes(height_cm))+
30    geom_histogram(fill="blue", color="darkblue")+
31    xlab("Height in centimeters, Bihar Females")
32  ggsave("output/bihar_better.pdf")
33
```

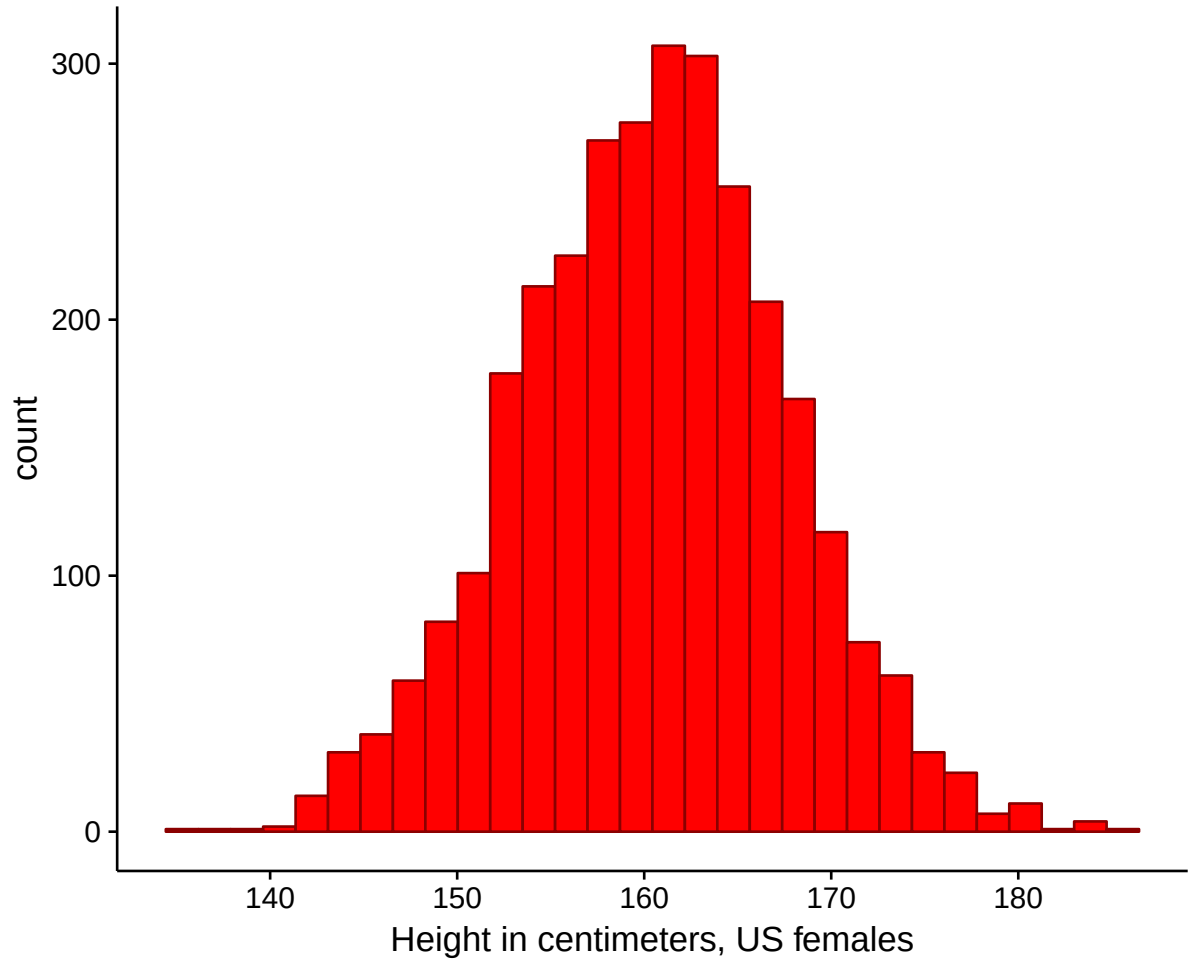
That is better



Playing with the bins



Same thing for US women



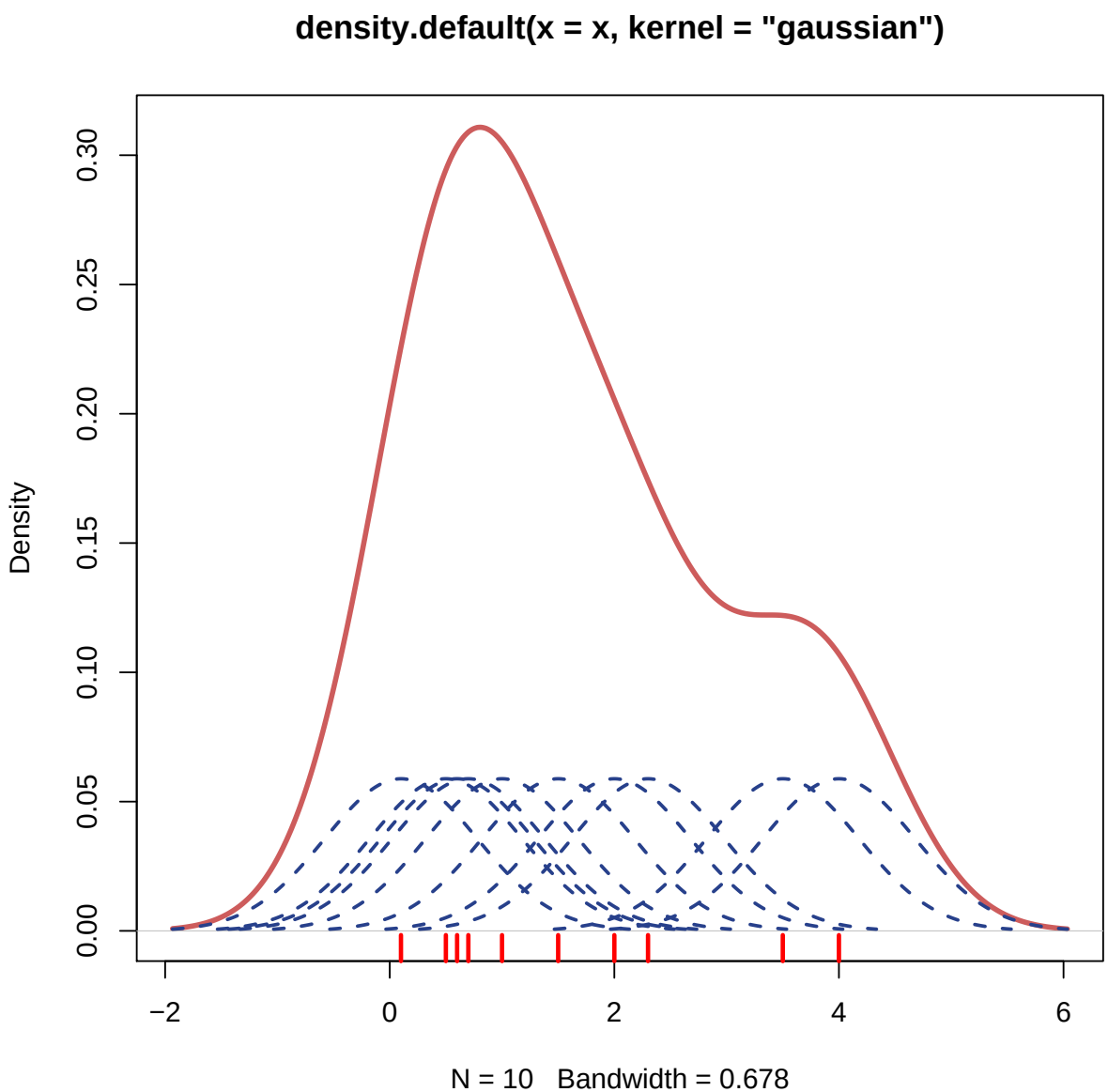
The Kernel Density Estimation

- The histogram is a little a bit bumpy ...
- Kernel density estimation is a non-parametric way to estimate the probability density function of a random variable.
- Straightforward extension of the histogram: at each point we take a weighted average of the frequency of the observations.
- Formally: Let $(x1, x2, ..., xn)$ be an independent and identically distributed sample drawn from some distribution with an unknown PDF f. We are interested in estimating the shape of this function f. Its kernel density estimator is given by:

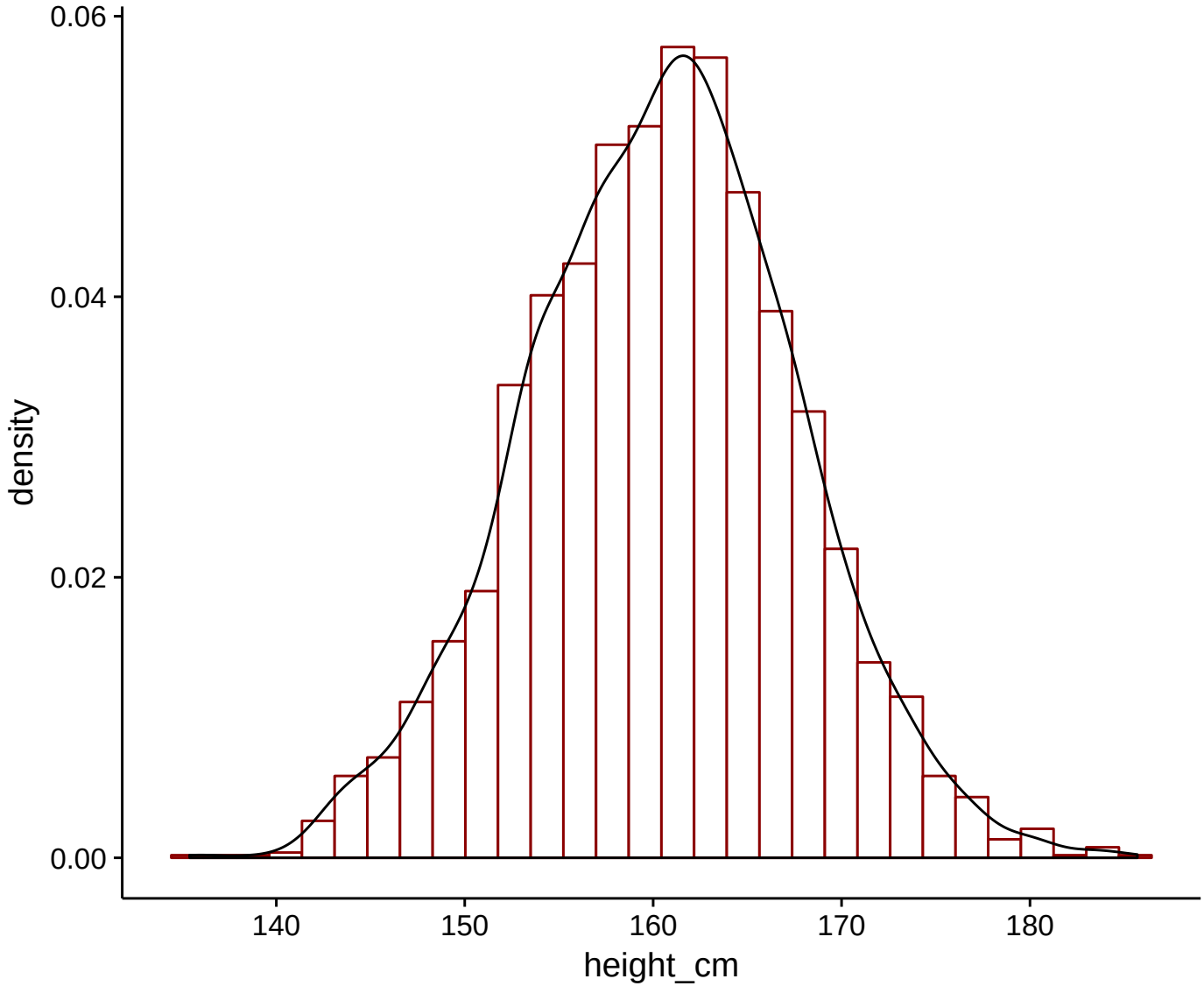
$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

- where $K()$ is the *kernel*, a non-negative function that integrates to one and has mean zero , and $h > 0$ the *bandwidth*
- Many choices for $K()$, but it is typically bell shaped

The Kernel Density Estimation



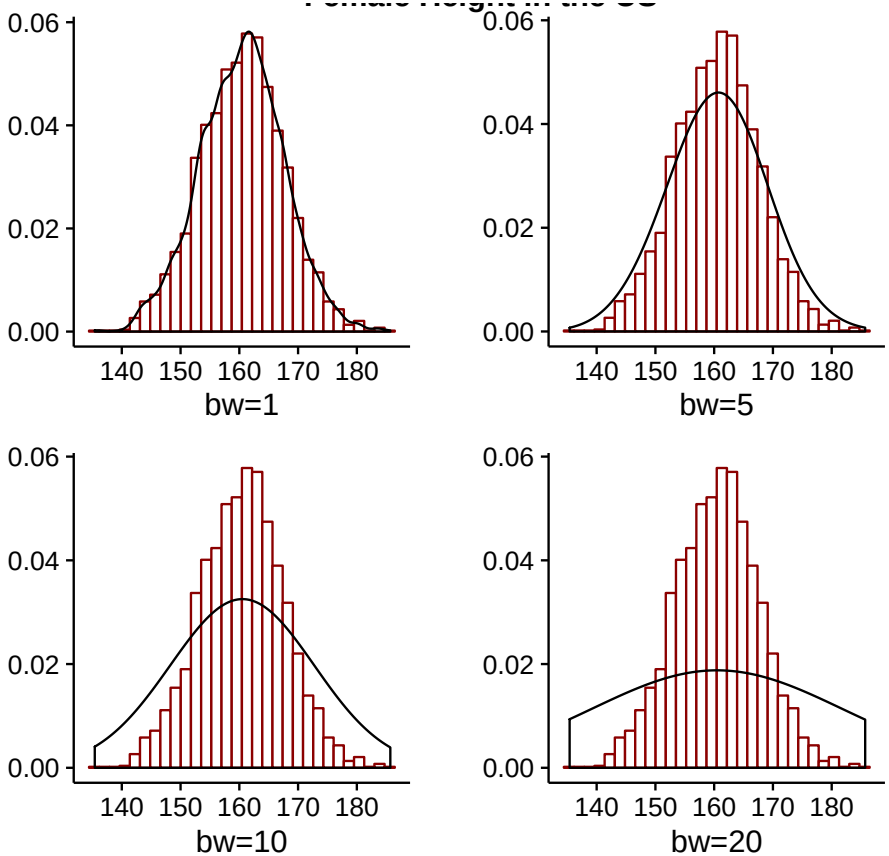
```r
# kernel density estimation

ggplot(us_adult_females_trunc, aes(height_cm))+
  geom_histogram(data=us_adult_females_trunc, aes(height_cm , ..density..), fill="white" , color="darkred")+
  geom_density(kernel="gaussian", aes(height_cm))

ggsave("output/US_kernel.pdf")

```


0.06
0.04
0.02
0.00
density
140
150
160
170
180
height_cm

Things to choose

- The Kernel function (Epanechnikov or Normal are frequent)
- The bandwidth: too small and the function will be squiggly... too large and you will miss important features of the distribution
- The optimal bandwidth minimizes the sum of these two errors (Mean Square Error).
- The default in R density (nrd0) is a rule of thumb optimal bandwith that should suit you well for most applications.

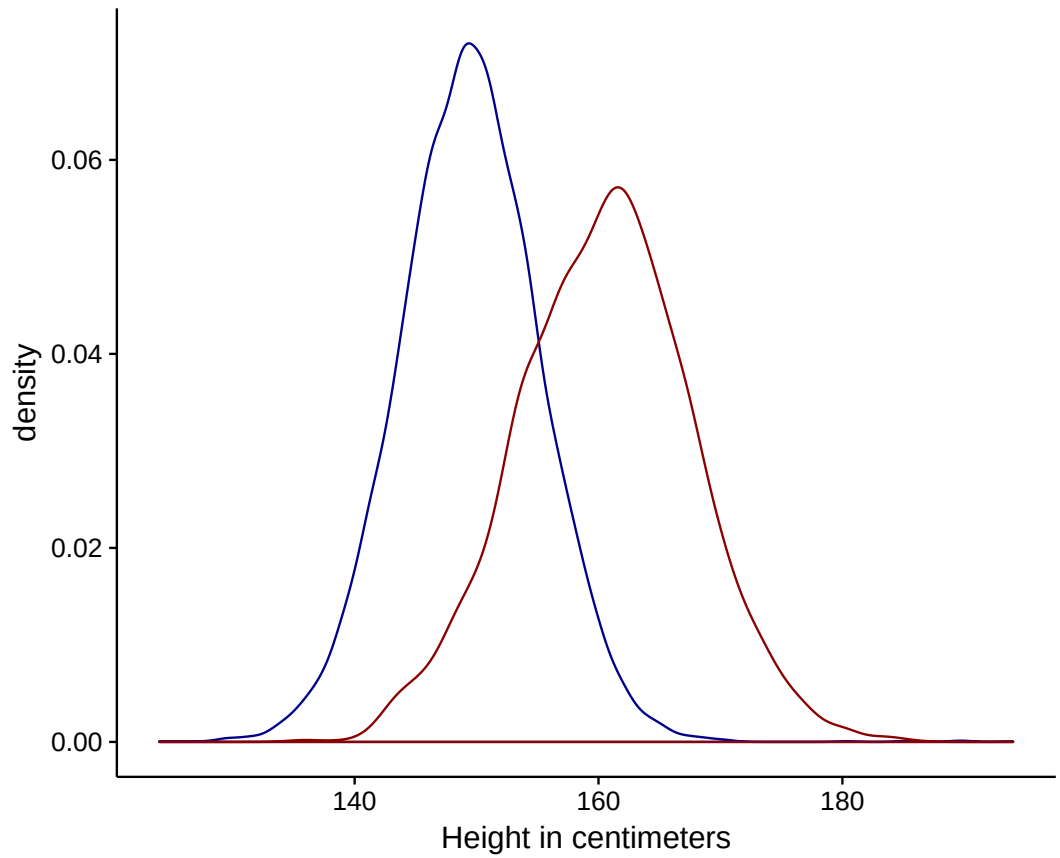
Varying the bandwidth



More about this distribution

- What do you think about the shape of this curve?
- It is unimodal, thin-tailed, symmetric
- Does it remind you of something we saw on the board of a previous lecture?
- With large enough n, the binomal distribution started to have this shape, no?
- The binomial distribution $B(n, p)$ is approximately normal with mean np and variance $np(1-p)$ for large n and for p not too close to zero or one.
- We will come back to the Normal distribution in great detail later, because it will turn out that many random variables can be conveniently assumed to have a normal distribution.
- Height (in a particular population) is a canonical example in textbooks, but why should it be normally distributed? or not? when we define the normal distribution more formally, and discuss where it comes from, we will discuss why heights should or should not be normally distributed.

Beginning to play with a bit more information: Female Heights in US and Bihar

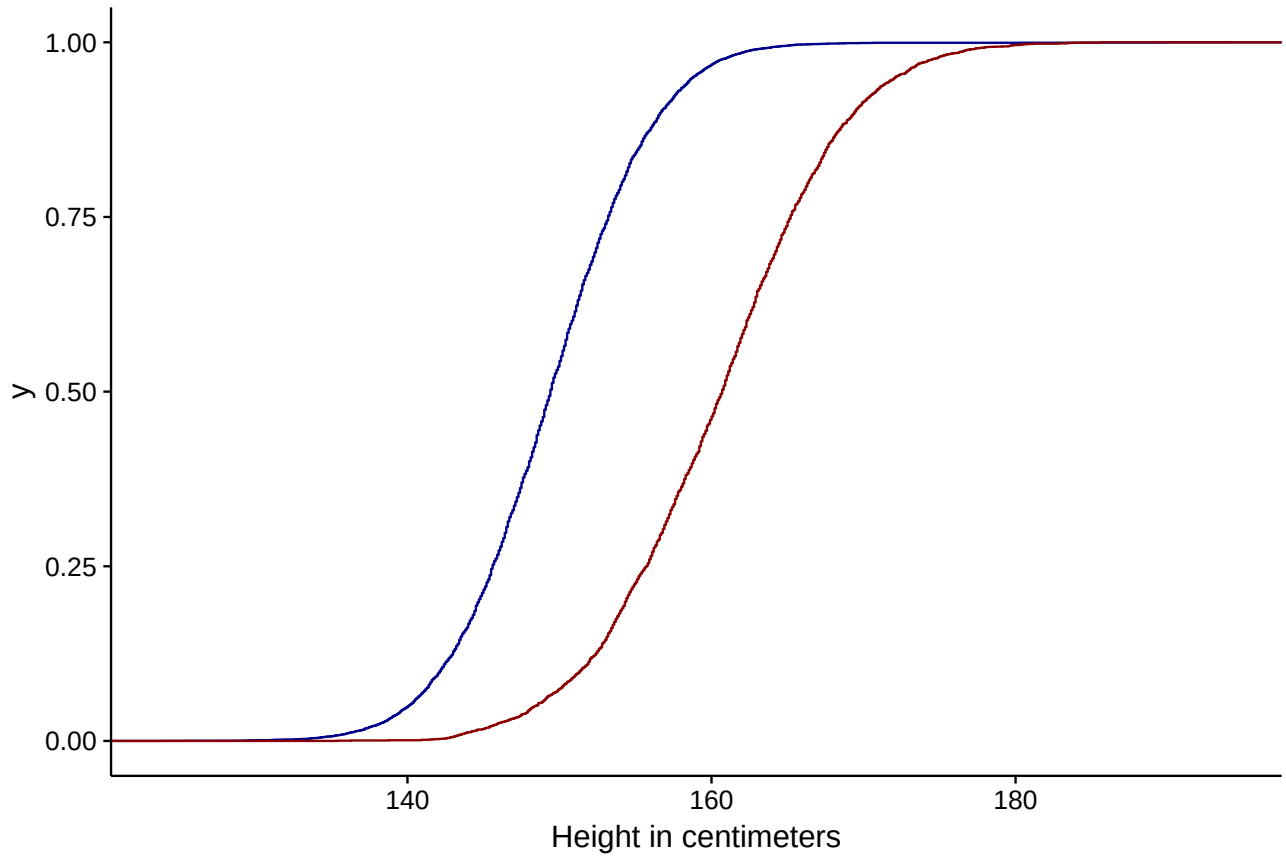


Cumulative histogram, cumulative CDF

- You may want to plot the CDF instead
- Then you can plot a cumulative histogram: the number / frequency of cases that are smaller or equal to the value for a particular bin
- Or you can get a smoothed version of a CDF (e.g. using ecdf in R)

```
152 # Representing the CDF
153
154 ggplot(bihar_adult_females_trunc, aes(height_cm))+
155   stat_ecdf(data=bihar_adult_females_trunc, aes(height_cm), color="darkblue" )+
156   stat_ecdf(data=us_adult_females_trunc, aes(height_cm),  color="darkred" )+
157   xlab("Height in centimeters")
158
159
160
161 ggsave("output/heightcdf.pdf")
```

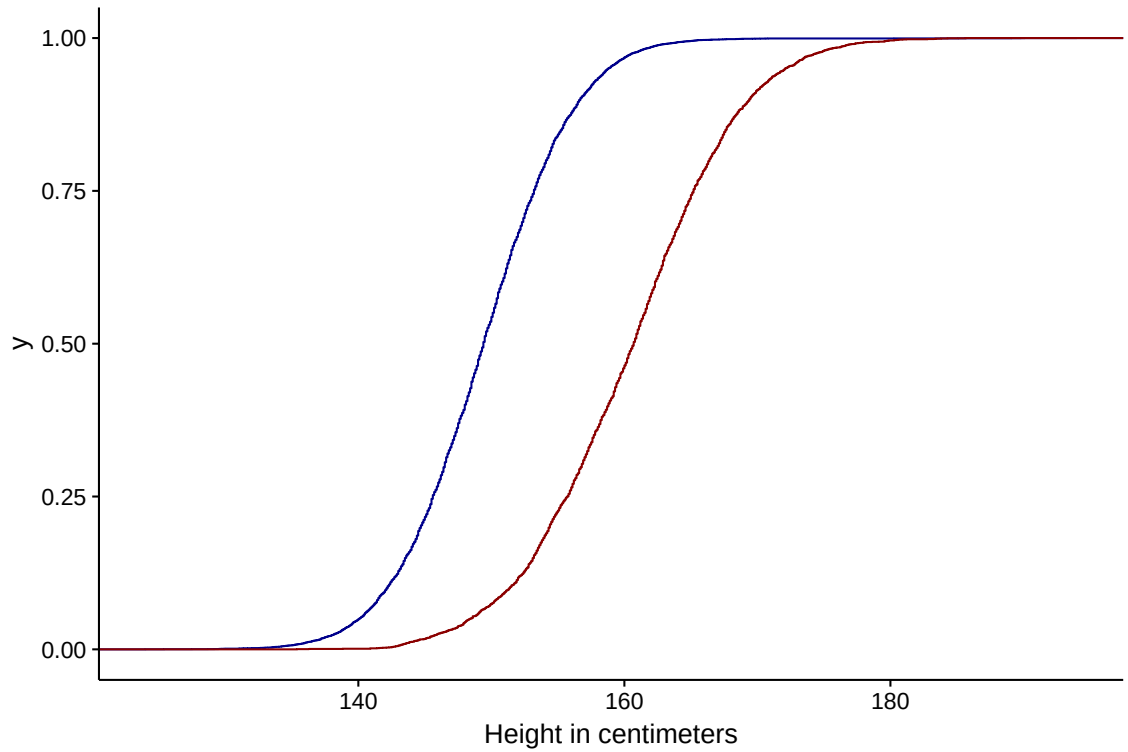
Beginning to play with a bit more information: Female Heights in US and Bihar



When do we want to plot pdf vs cdf?

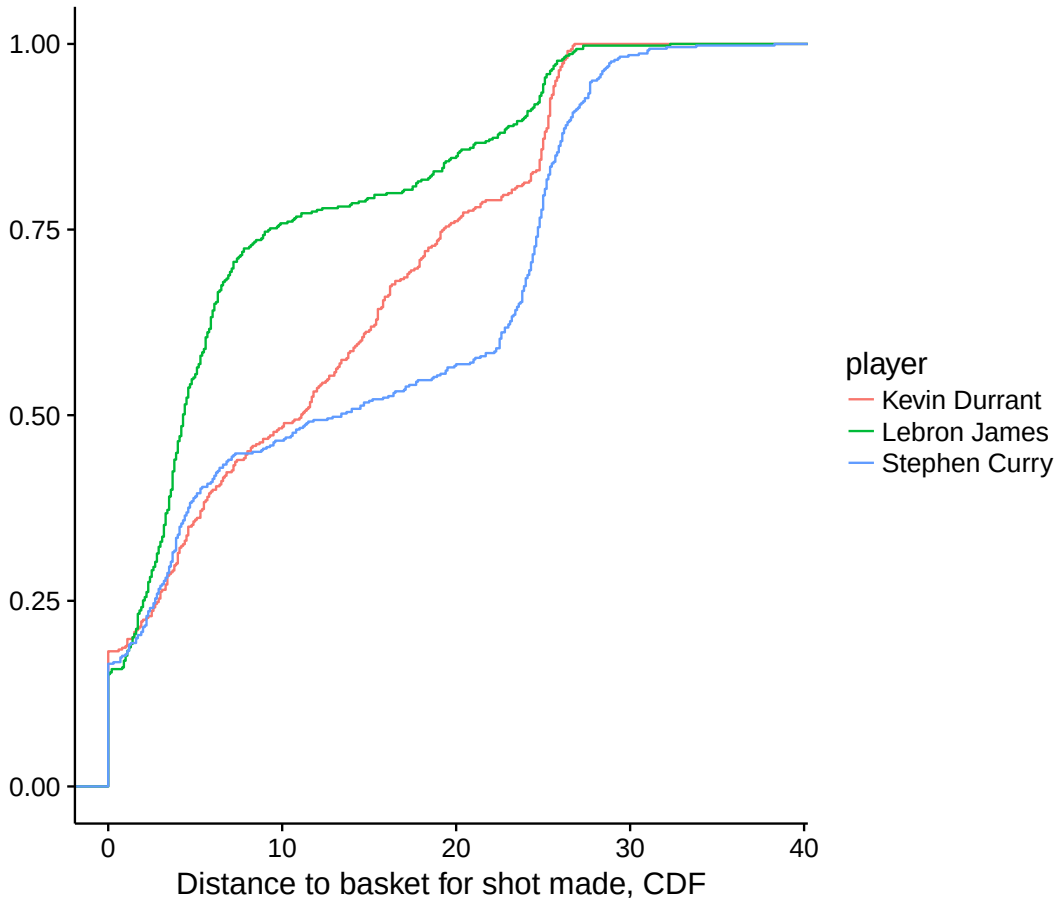
- They display the same underlying information...
- When you are interested in probabilities, representing them with CDF is more conventional, why?
  - A pdf represents probability with areas while a cdf represents probability with (vertical) distances.
  - It is much easier for the eye to compare distance than areas: the CDF is good to compare two distributions
  - In particular you can very easily visually assess *first order stochastic dominance* : for any size, the probability that a woman in Bihar is smaller than that size is larger than the probability that a US woman is smaller.
- When you are interested in the density, the pdf shows it as distance, the cdf as a slope
  - If you are interested in mode, shape, etc, the pdf is easier.

Haven't you learnt something exciting??



US women are taller than women in Bihar! The height distribution in the US stochastically dominates that in Bihar!!

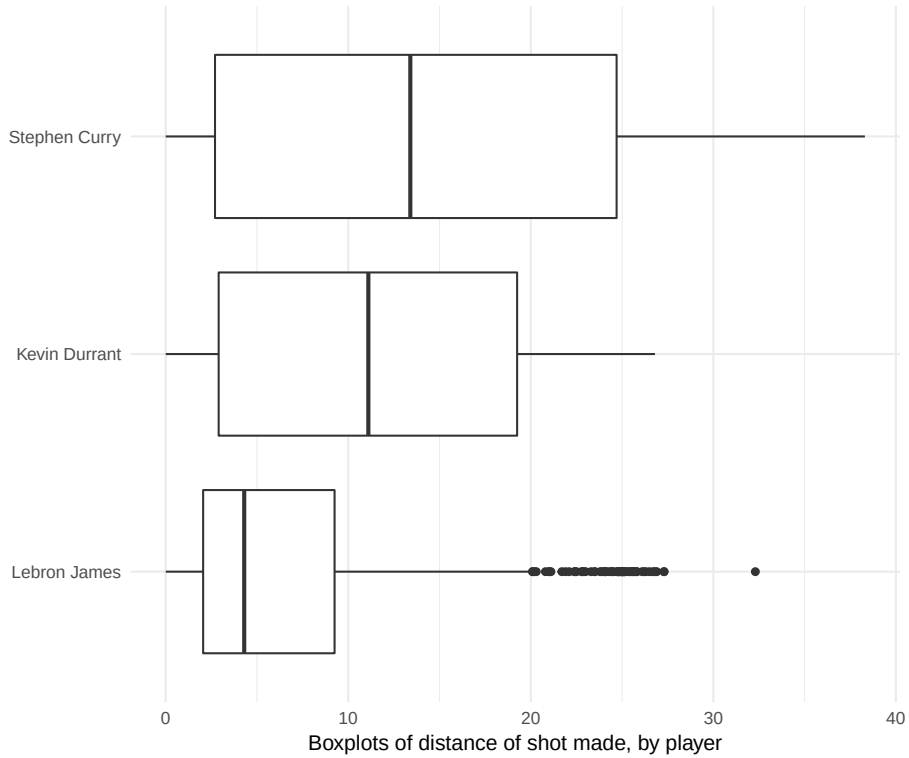
Not too surprised? OK , let's go back to Steph Curry



Code to produce this graph

```
# alternatively we can first append the three data sets, and then plot

threeplayers<-bind_rows(list( "Stephen Curry"=sc, "Lebron James"=lj, "Kevin Durrant"=kd) , .id="player")
threeplayers_shots_made <- filter(threeplayers, shot_made==1)

ggplot(threeplayers_shots_made, aes(shot_distance, colour=player))+
  stat_ecdf()+
  xlab("Distance to basket for shot made, CDF")+
  ylab("")
ggsave("Comparisonbetweenplayers_onedataset.pdf")

# The combined data set also allows to do a comparative boxplot
ggplot(threeplayers_shots_made)+
  geom_boxplot(
    mapping=aes(
      x = reorder(player, shot_distance, FUN = median),
      y = shot_distance
    )
  )+
  coord_flip()+
  xlab("")+
  ylab("Boxplots of distance of shot made, by player")+
  theme_minimal()
```

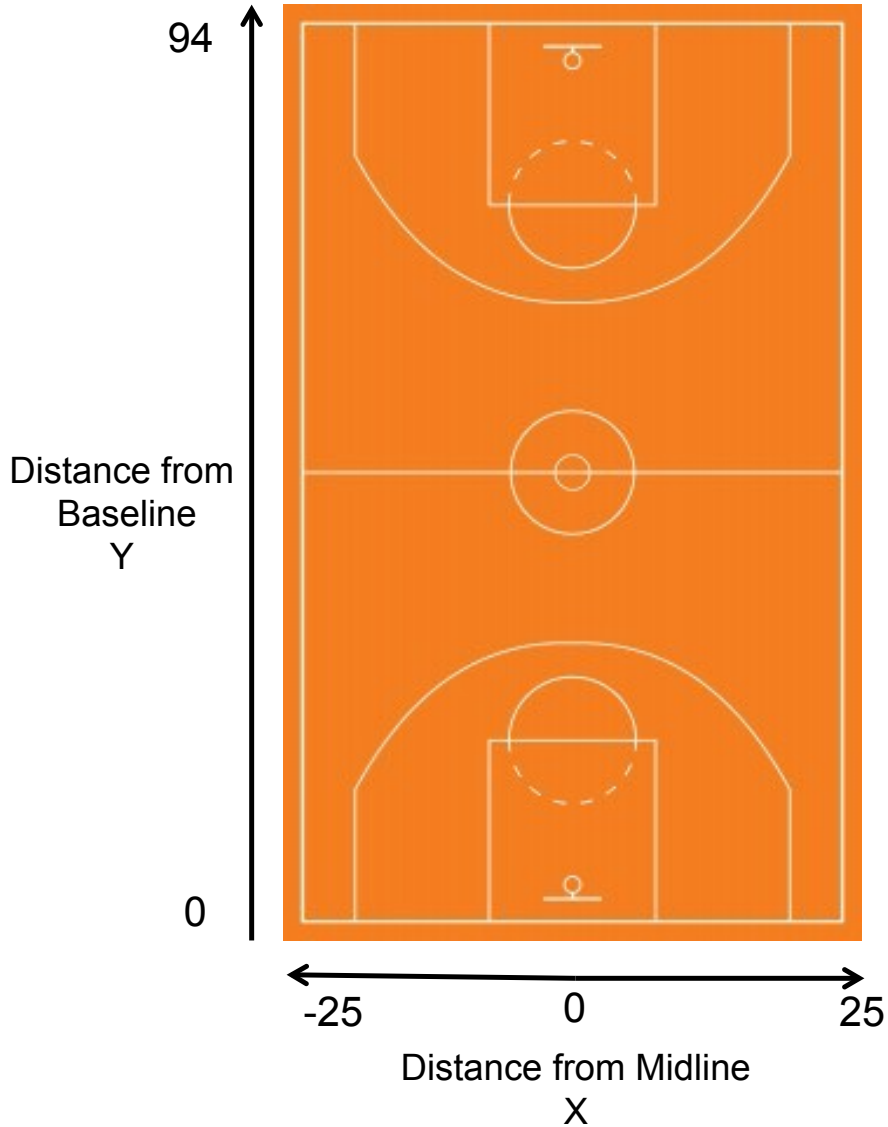
The boxplot


Stephen Curry
Kevin Durrant
Lebron James
0
10
20
30
40
Boxplots of distance of shot made, by player

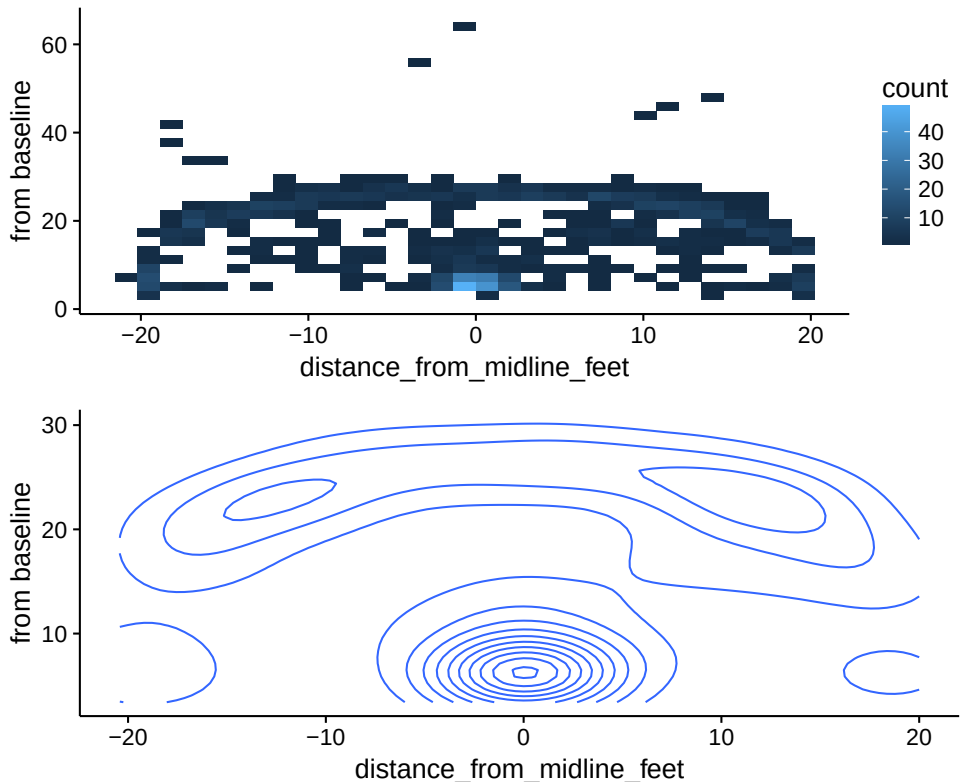
Representing joint distributions

- Suppose we want to represent the distribution of successful attempts by location
- There are actually two distances to consider: distance from baseline, and distance from the sideline
- If we plot each of them separately, what do we get?

A basketball court



A histogram of the joint density–or the map of a basketball court?



Now we see pretty clearly that there is bunching at the 3pt line!