Lecture 4: Let's look at some data!

Prof. Esther Duflo

14.31/310

Game plan

1. Where can we find data?
2. Visualizing distributions
3. Exploiting our knowledge of distribution functions to extract the information we want

1. *Where can we find data?*
2. Visualizing distributions
3. Exploiting our knowledge of distribution functions to extract the information we want

Where can we find data?

1. Existing data Libraries
2. Collecting your own data
3. Extracting data from the internet

Existing data libraries

- A Great resource for MIT students and others: `http://libguides.mit.edu/ssds`
- Popular sources of data
  - Data.gov: Datasets generated by the executive branch of the US government `http://www.data.gov/`
  - IPUMS : censuses from the US and many more! `https://www.ipums.org/`
  - International IPUMS `https://international.ipums.org/international/`
  - ICPSR `http://www.icpsr.umich.edu/icpsrweb/ICPSR/`: A data repository with many data sets on lots of subjects
  - Harvard-MIT Data center and Harvard Data verse `https://dataverse.harvard.edu/` where many researchers archive their data
  - Amazon dataverse `http://aws.amazon.com/public-data-sets/`

International household survey data

- Demographic and Health surveys
  `http://www.dhsprogram.com/`
- World bank `http://data.worldbank.org/`
- LSMS (search for LSMS on the world bank data page)
- Rand public-use databases
  `http://www.rand.org/labor/data.html`

Replication data from researchers

- Randomized control trials
  - `https://dataverse.harvard.edu/dataverse/jpal`
  - `https://dataverse.harvard.edu/dataverse/socialsciencercts`
- The American Economic Association journals require posting of any data used for research: there is lots of data on the AEA website

The internet!

- Many websites that are data intensive are making that data directly available to people
  - 538
  - Yahoo data dump : "a sample of anonymized user interactions on the news feeds of several Yahoo properties". `http://webscope.sandbox.yahoo.com/catalog.php?datatype=r&did=75`
- Some sites are specializing in aggregating data
  - Sports data sets `http://www.opensourcesports.com/`, `http://nbasavant.com/`
  - Web pages: Way back machine `https://archive.org/web/` [this may not be free]
- There is much more... search in library catalog, google, etc.

But what if this is not what I am looking for?

- Sometimes what you are looking for is available, but not free: your library may be able to purchase it... or you may need to get an agreement.
- Sometimes it is available but the access is restricted (for example for confidentiality reasons).
  - For administrative data, see `https://www.povertyactionlab.org/admindata`
  - The entity that owns the data may be interested in sharing it with you if this is part of a research project (prospective or retrospective)
  - You will then have to comply with the partner's requirements for data security, and also go through an Human Research Board Review (IRB) at your institution.
- And sometimes you will have to harvest it yourself!

Harvesting data

- Scraping data from the internet.
- Collecting your own data.

Scraping data from the internet

- General web scraping
- Collecting data from social networks

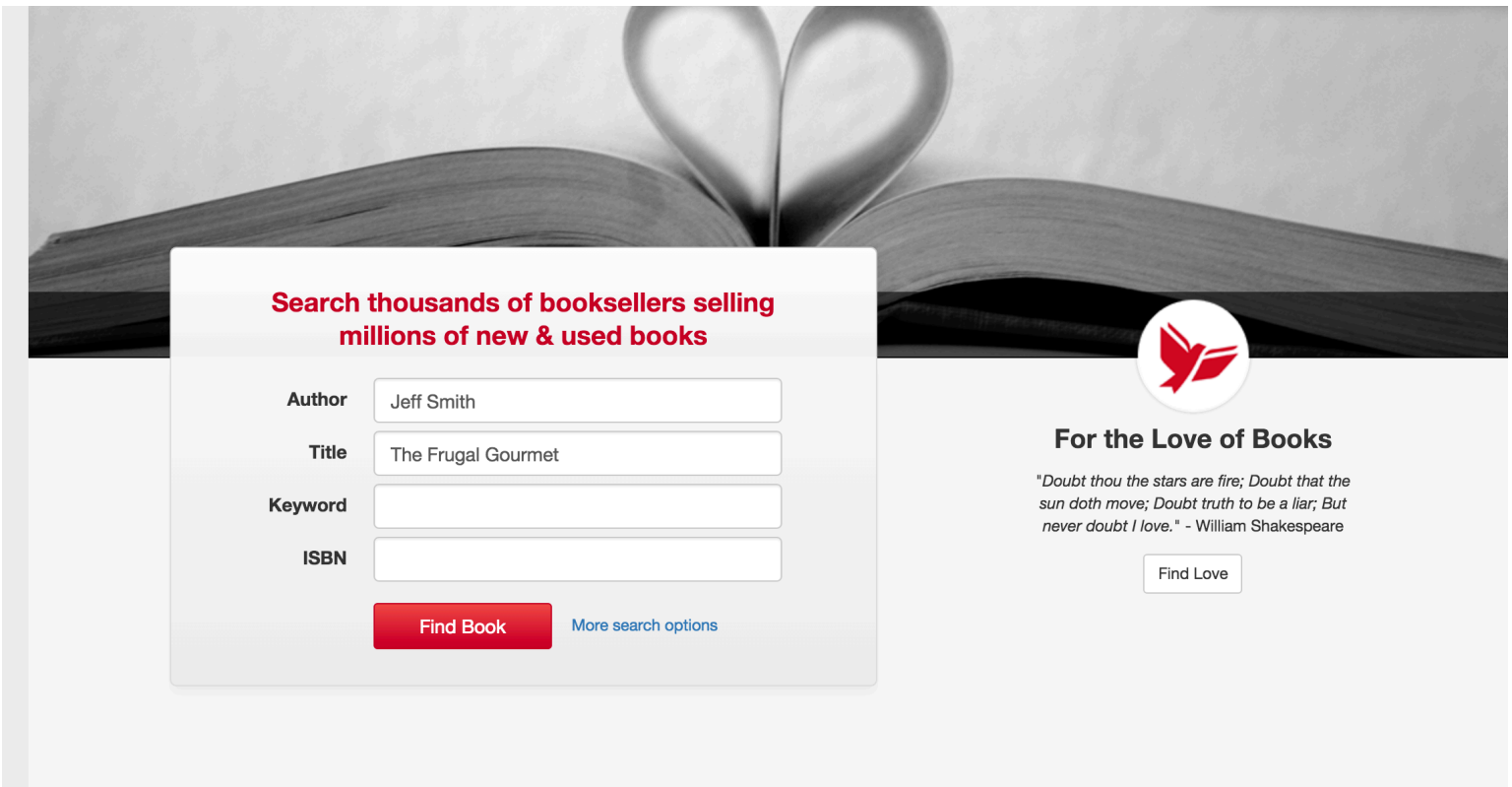
What is Web Scraping?

- Pull data form one page
- Crawl an entire web page
- A set of forms running in the background
- Any of the above in an ongoing fashion

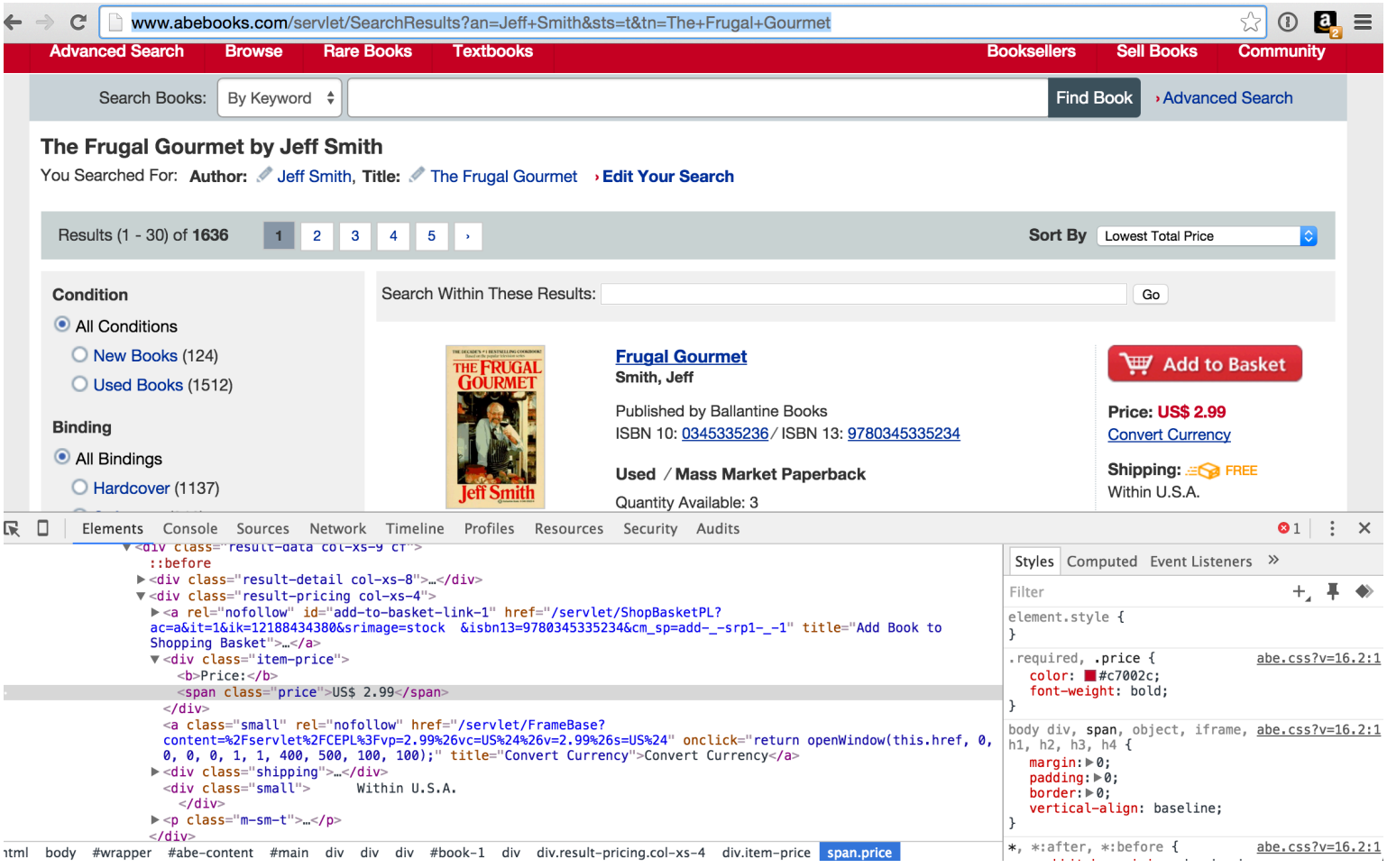
Example: Ellison and Ellison: Did the internet change the price of used books?

- Want to compare the price of the same used book in stores and on line
- Need to collect data at regular interval on the prices of a bunch of used books.
- From the web site `http://www.abebooks.com/`

What we start with



What we start with



What we want to get

A nice table we can import in R or another software with

- Name of title (for lots of titles)
- Date
- Price

Web sites as API

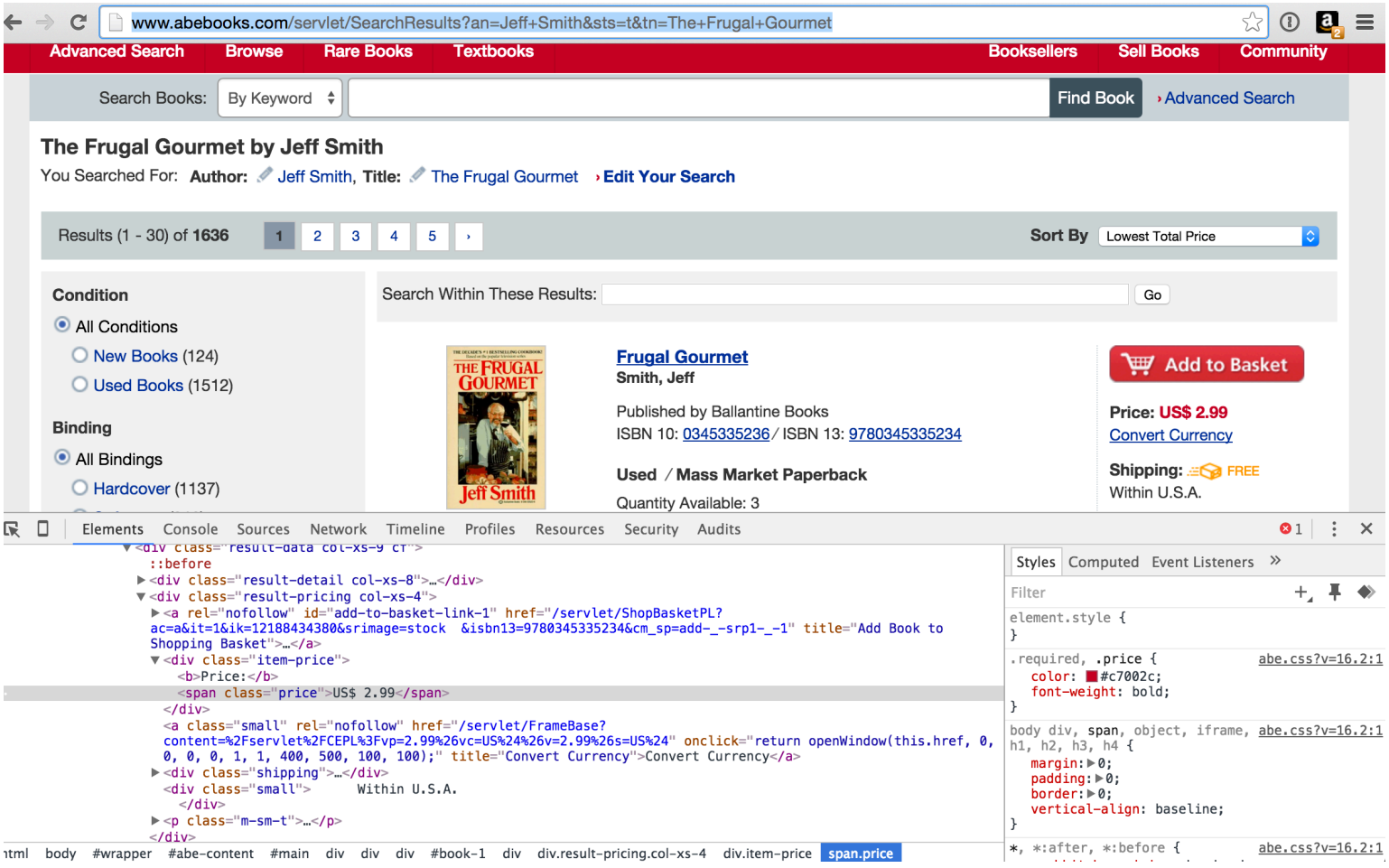
- API (Application Programme interface) are programs that help a particular program (e.g. Twitter) to communicate with other programs
- Some web sites provide and invest in API (Twitter, Facebook, etc.) and you will typically use those to harvest the data from those sites
- Some don't, and even when they do , the website is usually better maintained
- So then you need to invest in getting the data from the site directly: the site is the API
- There are many set of tools that will help you extract the information you are looking for in a page

Web scraping in R

- Really? *Really!*
- For simple tables it will work: use XML library, and the command readHTMLTable
- a video `https://www.youtube.com/watch?v=lAyE0qEXc6w`
- Another library is rvest.
- See: `http://blog.rstudio.org/2014/11/24/rvest-easy-web-scraping-with-r/`
- It could work for easy projects

Web scraping with Python

- Is certainly more conventional! Some entry cost but not too much (and the internet is full of tutorials)
- you will need: Python 2.7 and pip [both free]
- you will work using the request library and the BeautifulSoup library
- With those you will write simple routine that will extract what you are looking for.
- In the used book example, we need to pull up the page for each book at specified date, and instruct python to search for the the price (which is nicely identified to as a class).
- And export them into a table.


www.abebooks.com/servlet/SearchResults?an=Jeff+Smith&sts=t&tn=The+Frugal+Gourmet
Advanced Search
Browse
Rare Books
Textbooks
Booksellers
Sell Books
Community
Search Books:
By Keyword
Find Book
Advanced Search
The Frugal Gourmet by Jeff Smith
You Searched For: Author: Jeff Smith, Title: The Frugal Gourmet
Edit Your Search
Results (1 - 30) of 1636
1
2
3
4
5
Sort By
Lowest Total Price
Condition
All Conditions
New Books (124)
Used Books (1512)
Binding
All Bindings
Hardcover (1137)
Search Within These Results:
Go
Frugal Gourmet
Smith, Jeff
Published by Ballantine Books
ISBN 10: 0345335236 / ISBN 13: 9780345335234
Used / Mass Market Paperback
Quantity Available: 3
Add to Basket
Price: US$ 2.99
Convert Currency
Shipping: FREE
Within U.S.A.
Elements
Console
Sources
Network
Timeline
Profiles
Resources
Security
Audits
::before
…

<a rel="nofollow" id="add-to-basket-link-1" href="/servlet/ShopBasketPL?
ac=a&it=1&ik=12188434380&srimage=stock &isbn13=9780345335234&cm_sp=add-_-srp1-_-1" title="Add Book to
Shopping Basket">…

Price:
US$ 2.99

<a class="small" rel="nofollow" href="/servlet/FrameBase?
content=%2Fservlet%2FCEPL%3Fvp=2.99%26vc=US%24%26v=2.99%26s=US%24" onclick="return openWindow(this.href, 0,
0, 0, 0, 1, 1, 400, 500, 100, 100);" title="Convert Currency">Convert Currency
…

Within U.S.A.

…

html body #wrapper #abe-content #main div div div #book-1 div div.result-pricing.col-xs-4 div.item-price span.price
Styles
Computed
Event Listeners
Filter
element.style {
}
.required, .price {
color: #c7002c;
font-weight: bold;
}
abe.css?v=16.2:1
body div, span, object, iframe, h1, h2, h3, h4 {
margin:0;
padding:0;
border:0;
vertical-align: baseline;
}
abe.css?v=16.2:1
*, *:after, *:before {
abe.css?v=16.2:1

Collecting your own data

- It is not as infeasible as it sounds!
- Survey tool on the internet
- Install Apps on willing participants that will track their movements (or other things)
- Sit in the science center and administer some questionnaires
- and of course if you have more money, organize a data collection team to collect whatever you would like!

Steps for collecting your own data

- Obtain the funding you may need
- Prepare a data management plan : how will you keep the data safe? will you share it?
- Obtain Human subject Approval
- Design your data collection instrument
- Pilot your data collection instrument
- Implement!

Game Plan

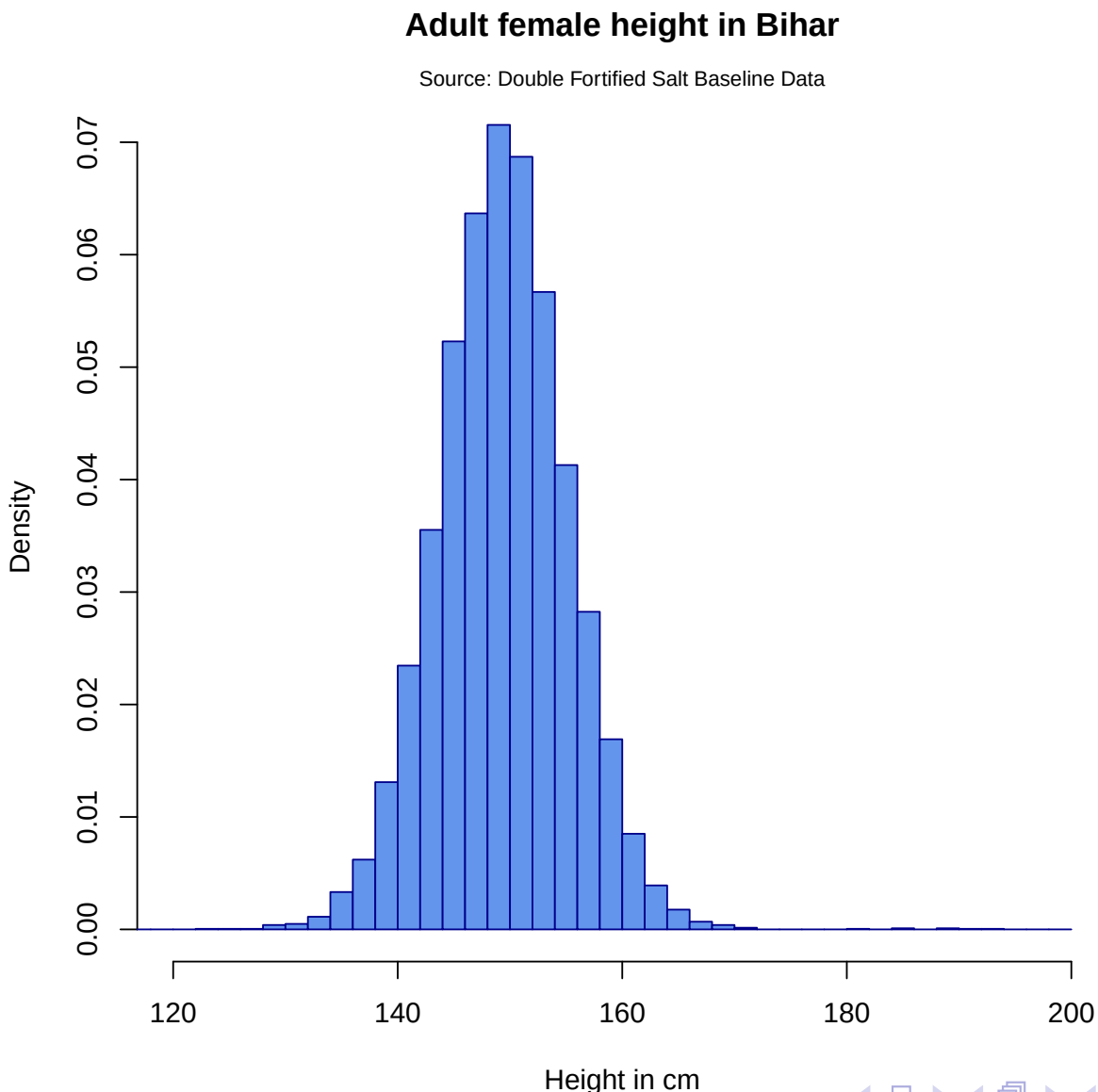
1. Where can we find data?
2. *Visualizing distributions*
3. Exploiting our knowledge of distribution functions to extract the information we want

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

- Plotting Histograms
- Plotting Kernels density
- 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



How did I do that?

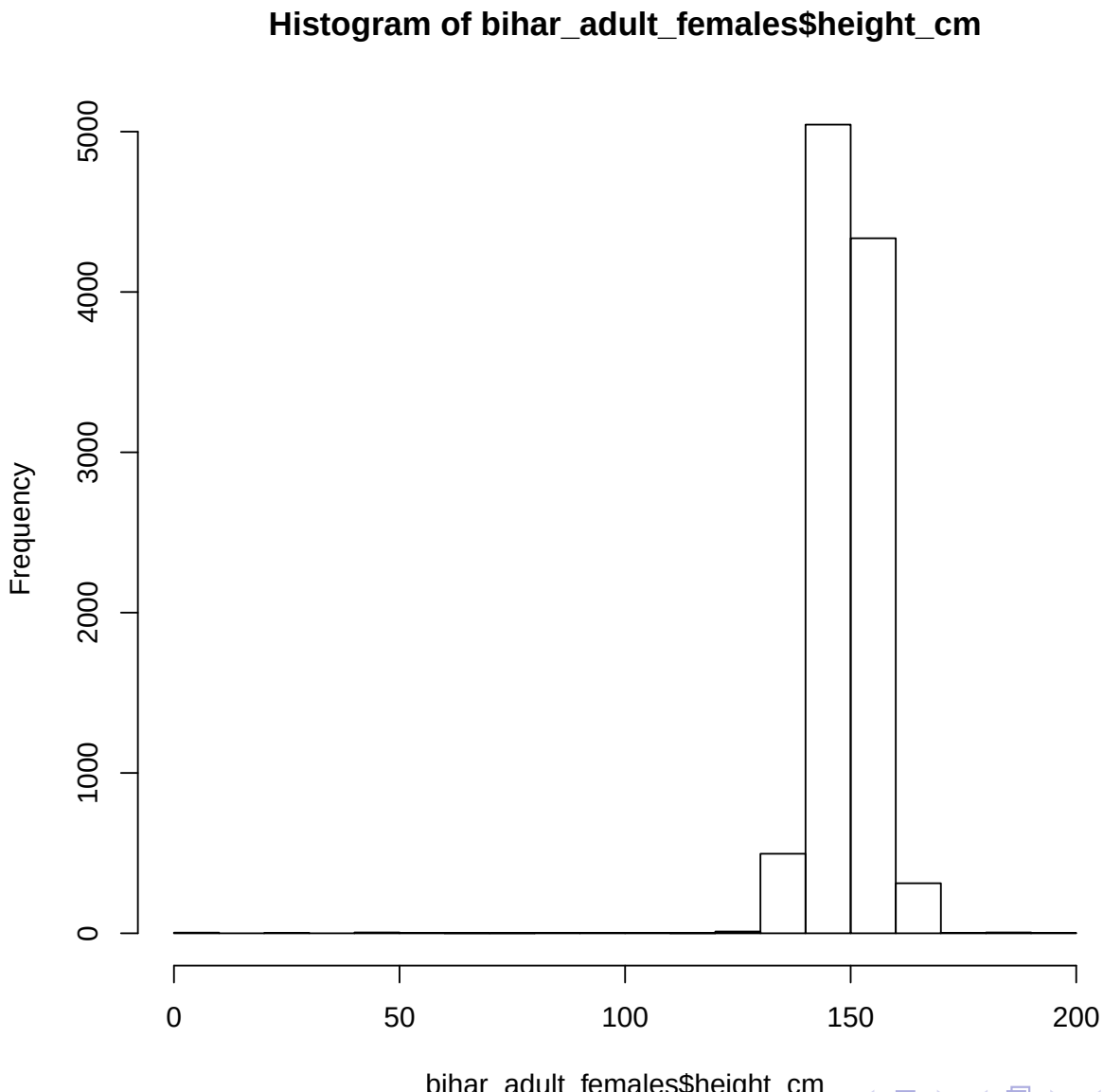
```
# Load in the Bihar height and weight data
bihar_data <- read.csv("data/Bihar_sample_data.csv")
colnames(bihar_data)

# Take a look at the data
summary(bihar_data)
head(bihar_data)

# Subset the data for adult females
bihar_adult_females <- subset(bihar_data,female==1 & adult==1)
head(bihar_adult_females)

# The simplest default histogram
pdf("output/2_bihar_female_height_default.pdf")
hist(bihar_adult_females$height_cm)
hide<-dev.off()
```

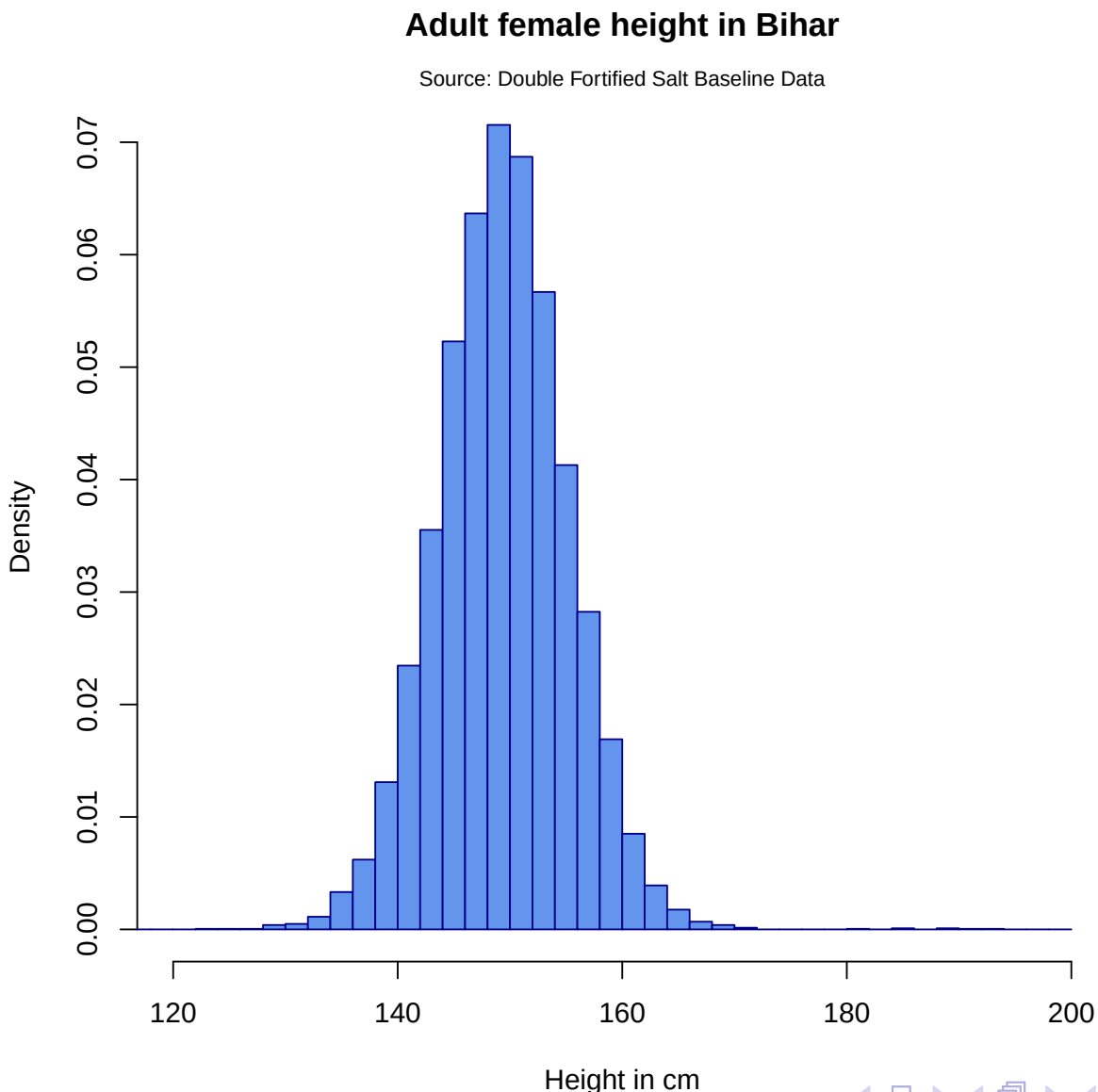
Mmmmm..Not too pretty



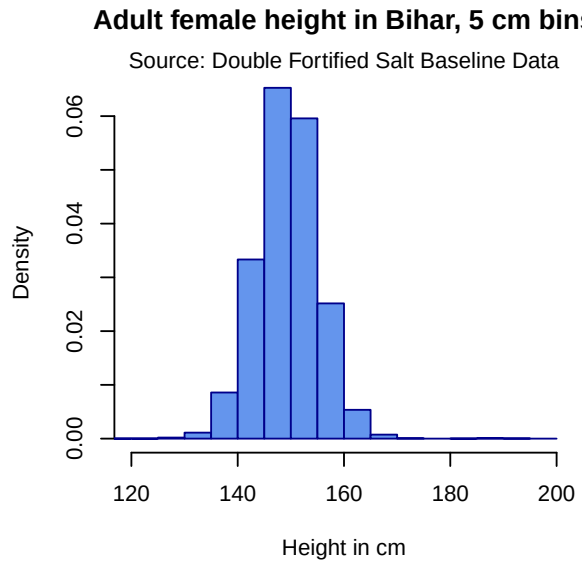
Let's try again

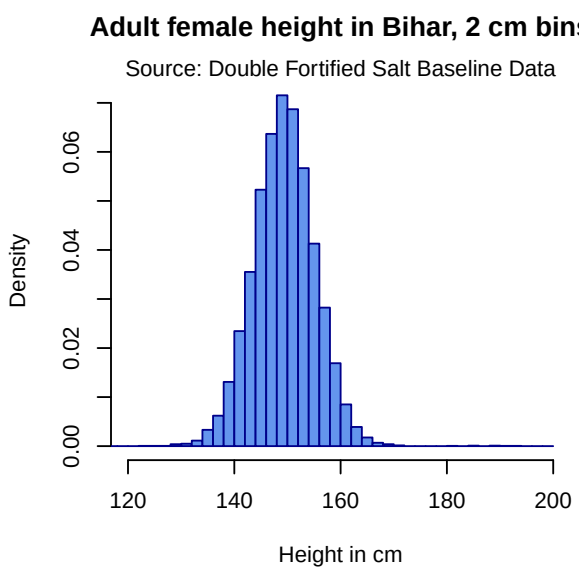
```
# Simple histogram with some customization
pdf("output/3_bihar_female_height.pdf")
hist(bihar_adult_females$height_cm,
     main="Adult female height in Bihar",
     freq=FALSE,
     xlab="Height in cm",
     xlim=c(120,200),
     breaks=seq(0,200,by=2),
     col="cornflowerblue",border="darkblue")
hide<-dev.off()
```

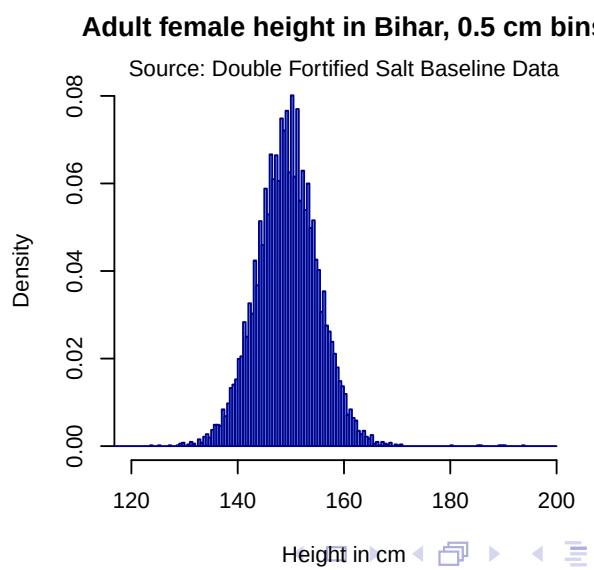
That is better



Playing with the bins







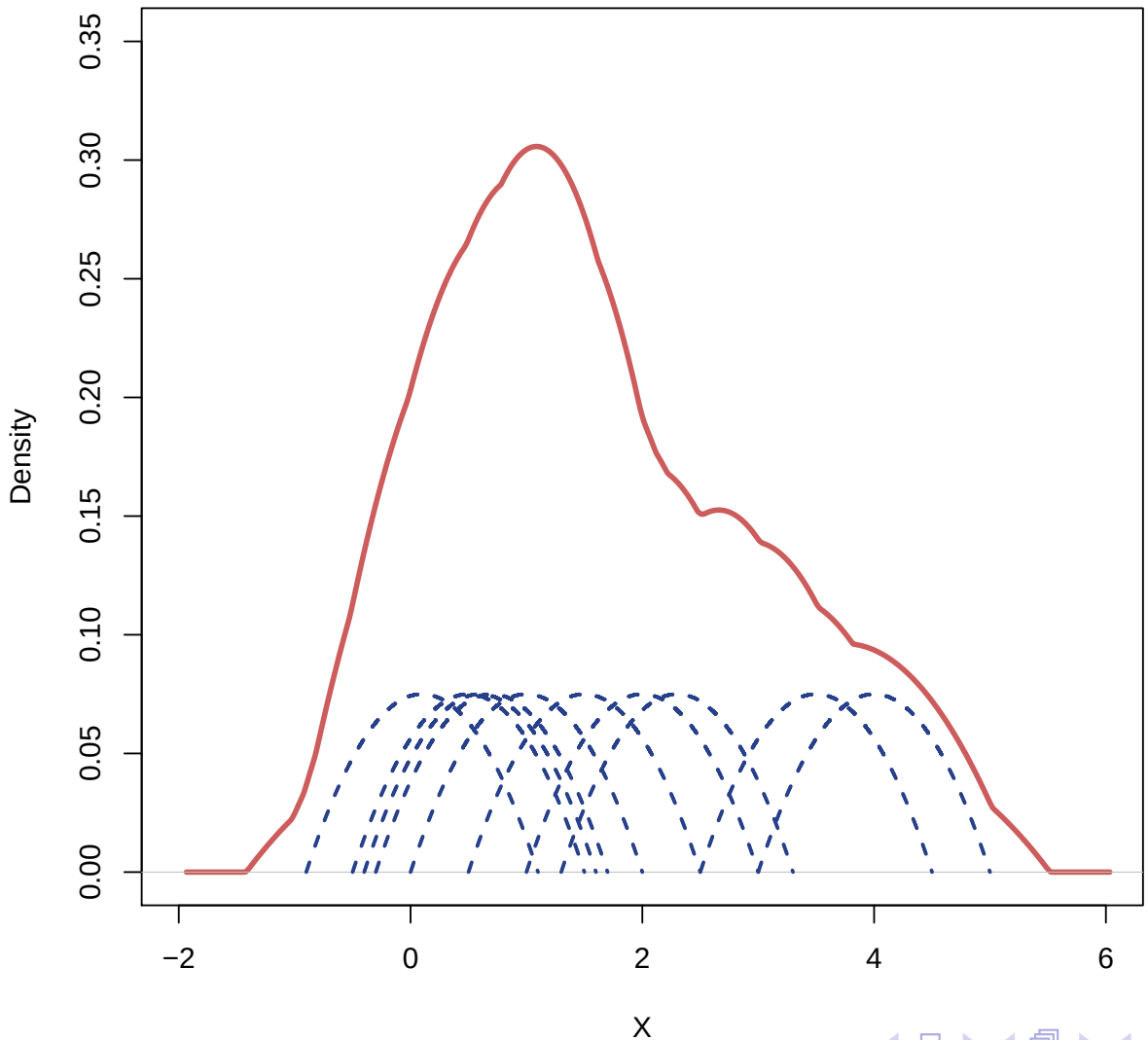
The Kernel Density Estimation

- 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:

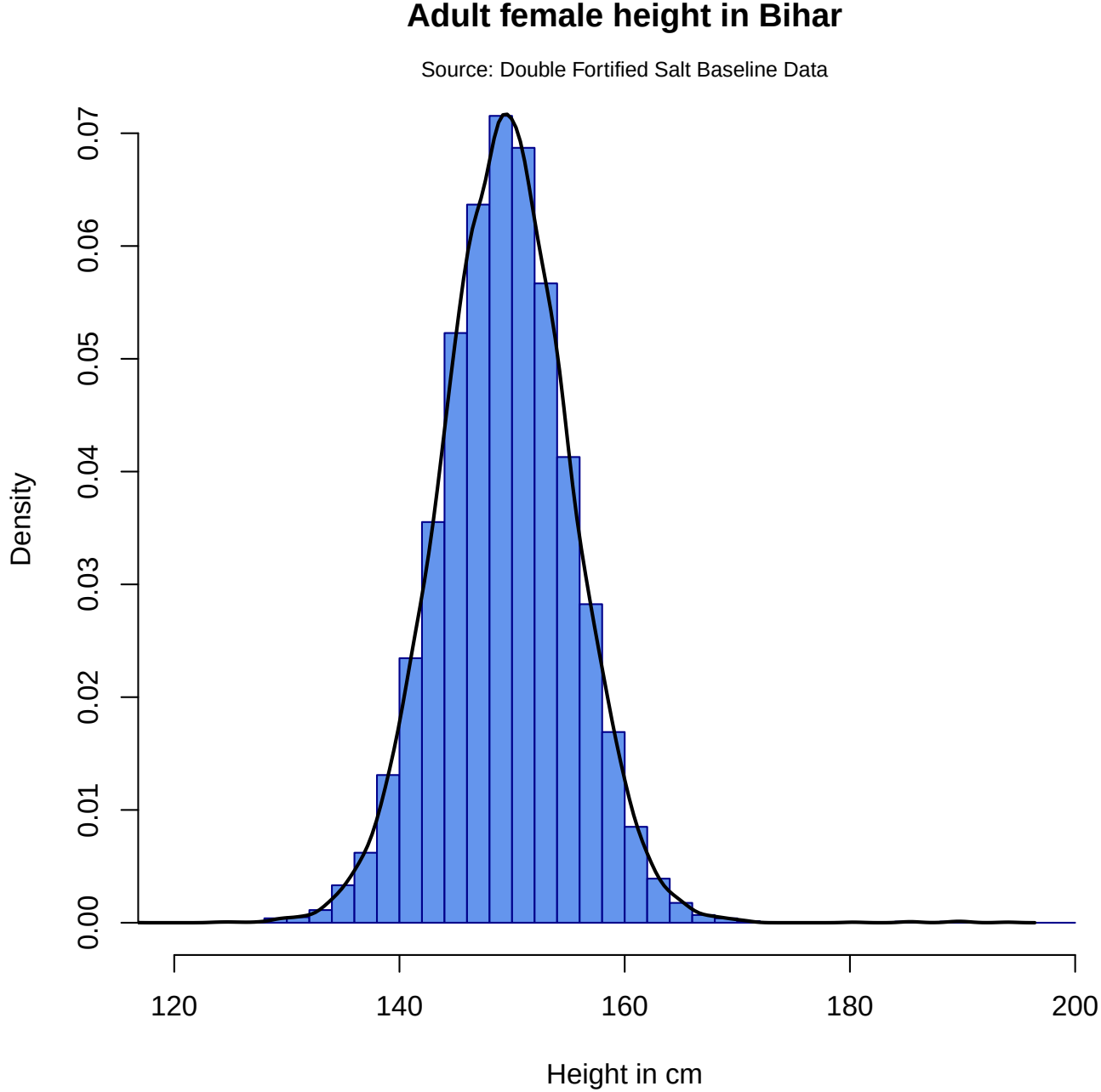
$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

- 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



```
# Add a kernel estimation
pdf("output/5_bihar_female_height_kernel.pdf")
hist(bihar_adult_females$height_cm,
    main="Adult female height in Bihar",
    freq=FALSE,
    xlab="Height in cm",
    xlim=c(120,200),
    breaks=seq(0,200,by=2),
    col="cornflowerblue",border="darkblue")
    lines(density(bihar_adult_females$height_cm,
        bw="nrd0", kernel="epanechnikov",
        na.rm=TRUE),col="black",lwd=3)
    mtext("Source: Double Fortified Salt Baseline Data",side=3,cex=0.75)
hide<-dev.off()
```


Adult female height in Bihar
Source: Double Fortified Salt Baseline Data
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
120
140
160
180
200
Height in cm

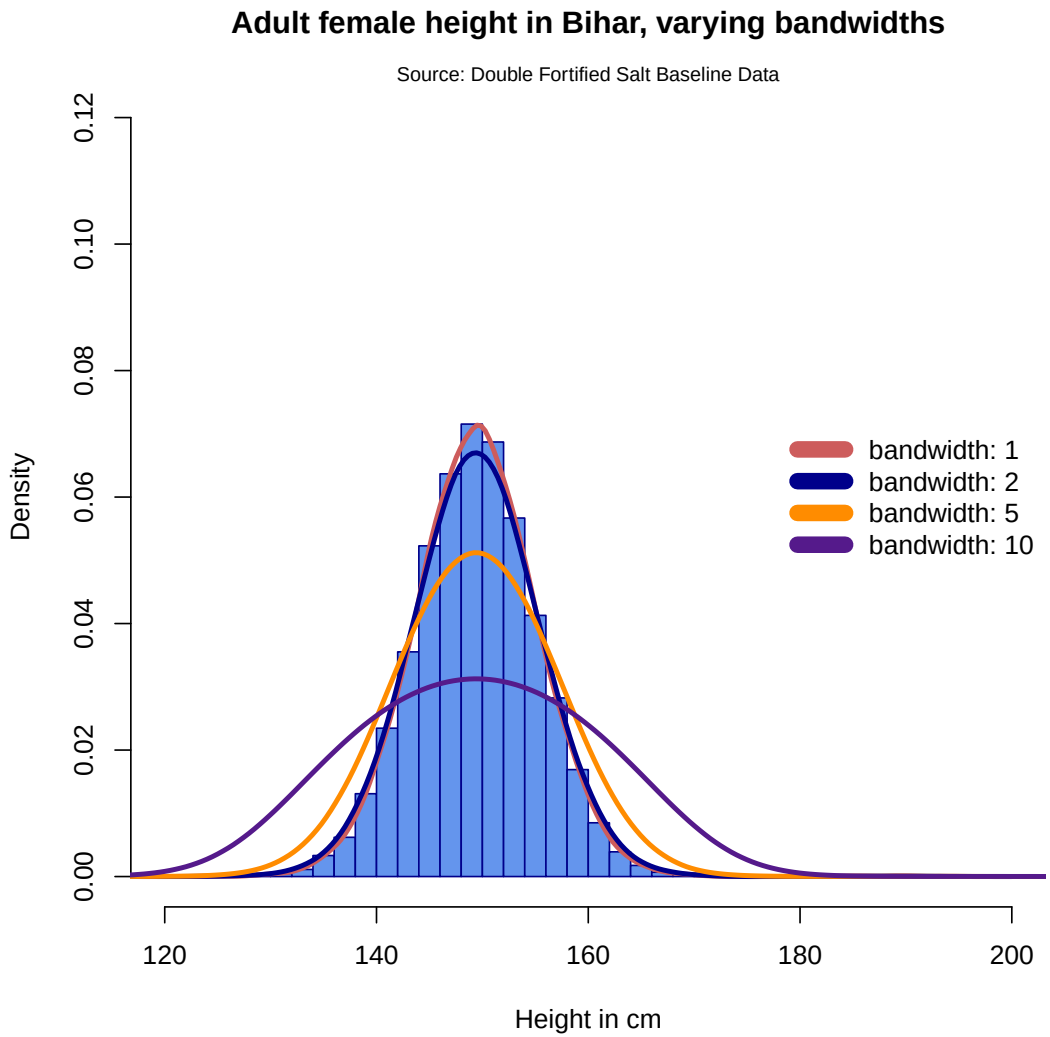
Things to choose

- 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

```
# Testing out different bandwidths (4 lines)
pdf("output/5.5_bihar_female_height_4lines_bandwidth.pdf")
colors <- c("indianred", "darkblue", "darkorange", "purple4")
bandwidths <- c(1,2,5,10)
hist(bihar_adult_females$height_cm,
     main=paste("Adult female height in Bihar, varying bandwidths"),
     xlab="Height in cm",
     xlim=c(120,200),
     ylim=c(0,0.12),
     prob=TRUE,
     breaks=seq(0,200,by=2),
     col="cornflowerblue",border="darkblue")
     for (i in 1:4)
     {
          lines(density(bihar_adult_females$height_cm,
          bw=bandwidths[i],kernel="epanechnikov",window="epanechnikov",
          na.rm=TRUE), col=colors[i], lwd=3)
     }
     legend("right",c("bandwidth: 1", "bandwidth: 2", "bandwidth: 5", "bandwidth: 10"),
          col=colors, lwd=10, bty="n")
     mtext("Source: Double Fortified Salt Baseline Data",side=3,cex=0.75)
hide<-dev.off()
```

Varying the bandwidth



More about this distribution

- What do you think about the shape of this curve?

More about this distribution

- What do you think about the shape of this curve?
- It is unimodal, thin-tailed, symmetric

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 in the previous lecture?

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 in the previous lecture?
- With large enough n, the binomal distribution started to have this shape, no?

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 in the 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.

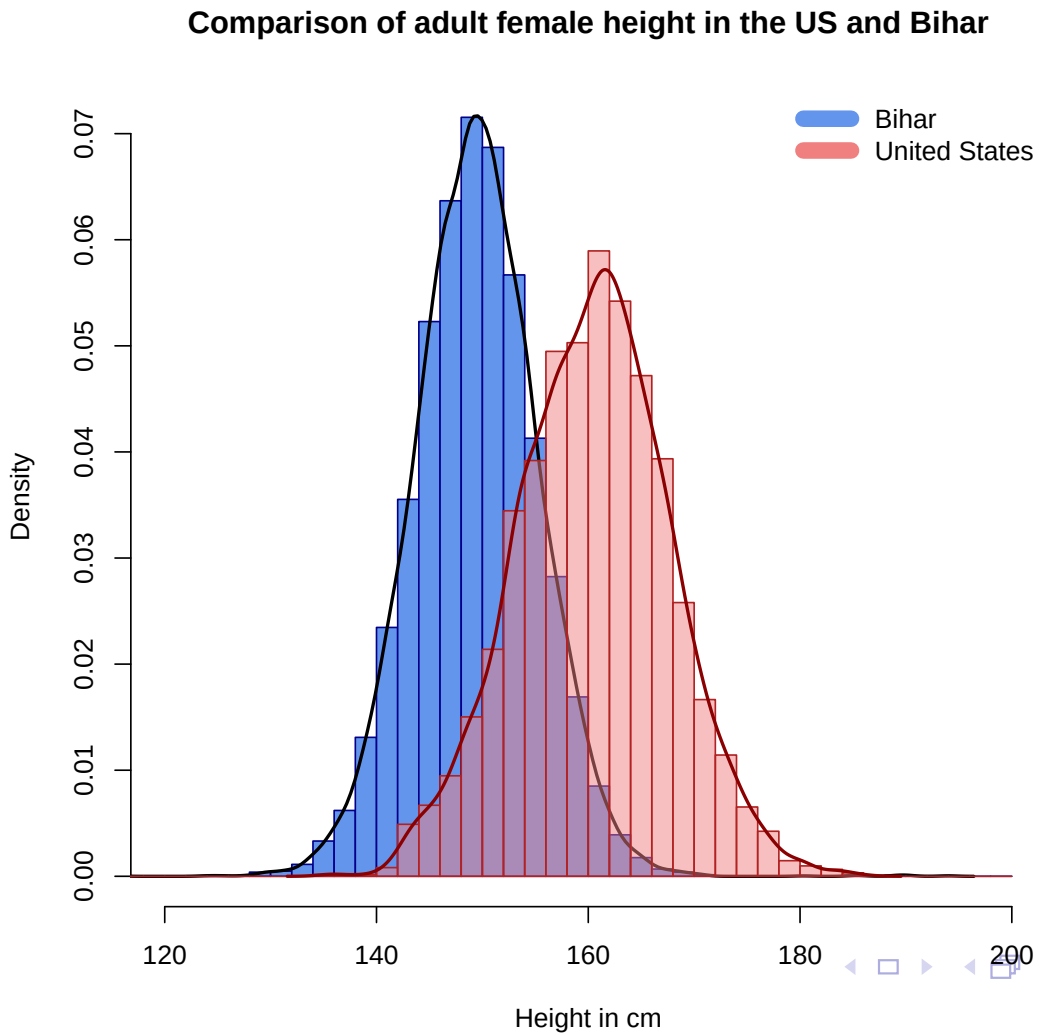
More about this distribution

- 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 in the 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.

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 in the 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

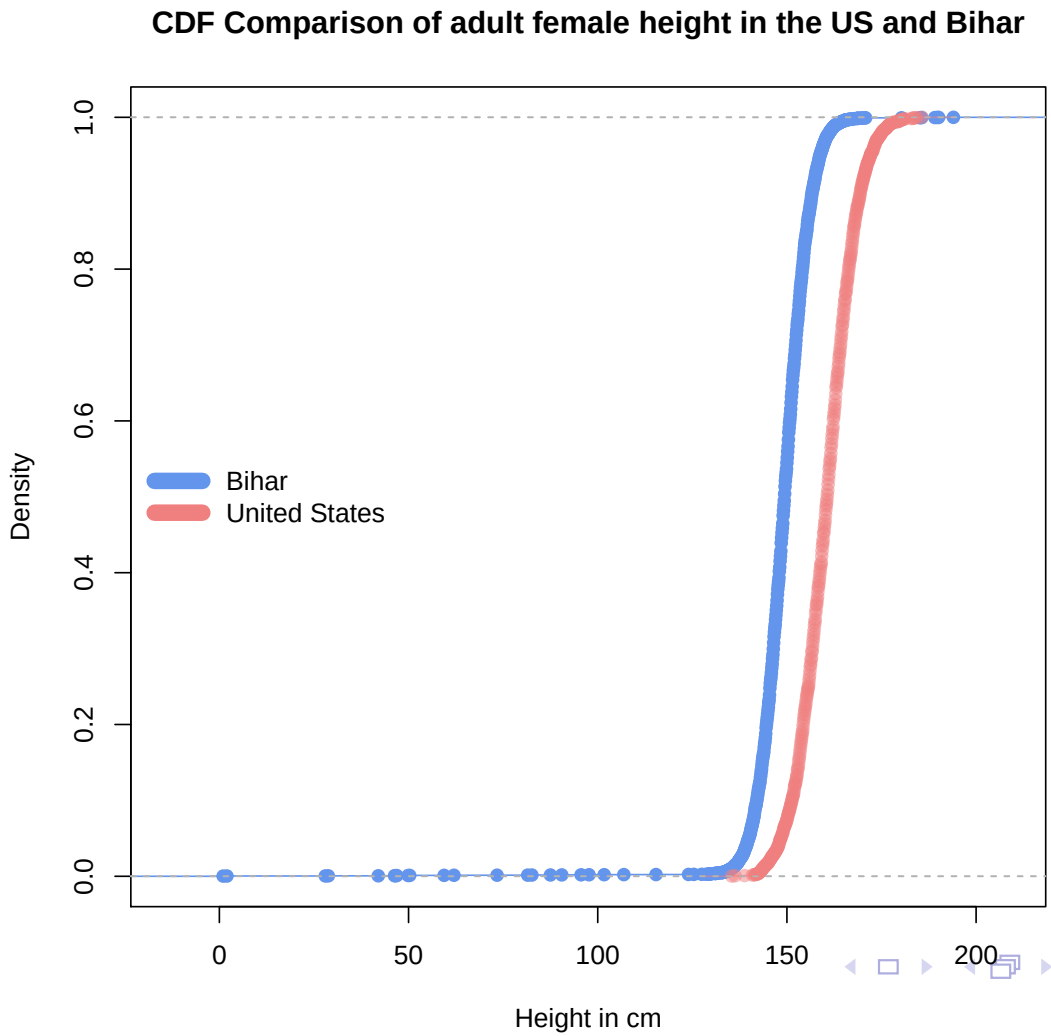


Cumulative histogram, cumulative CDF

- 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)

```
# Combined CDF
pdf("output/CDF Female Height.pdf")
plot(ecdf(bihar_adult_females$height_cm), col="cornflowerblue",
    main="CDF Comparison of adult female height in the US and Bihar",
    xlab="Height in cm", ylab="Density")
plot(ecdf(us_adult_females$height_cm),
    col=scales::alpha("lightcoral",.5), add=T)
legend("left", c("Bihar", "United States"),
    col=c("cornflowerblue", "lightcoral"),
    lwd=10, bty="n")
hide<-dev.off()
```

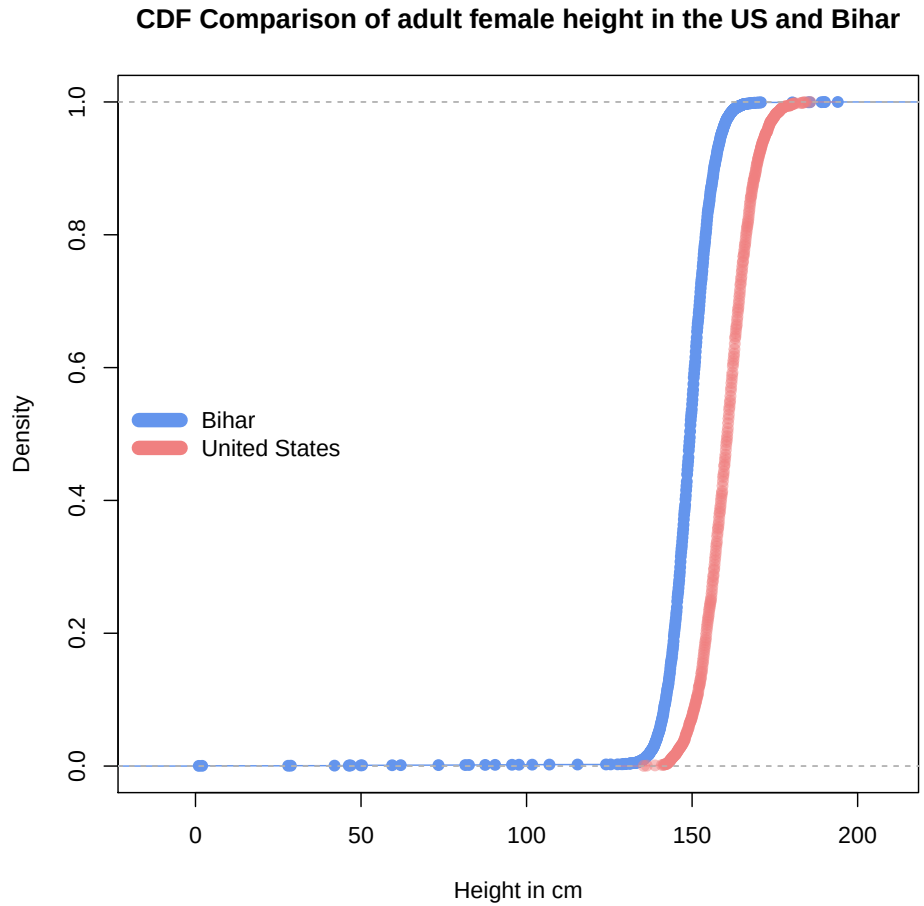
Beginning to play with a bit more information: Female Heights in US and Bihar



When do we want to plot pdf vs cdf?

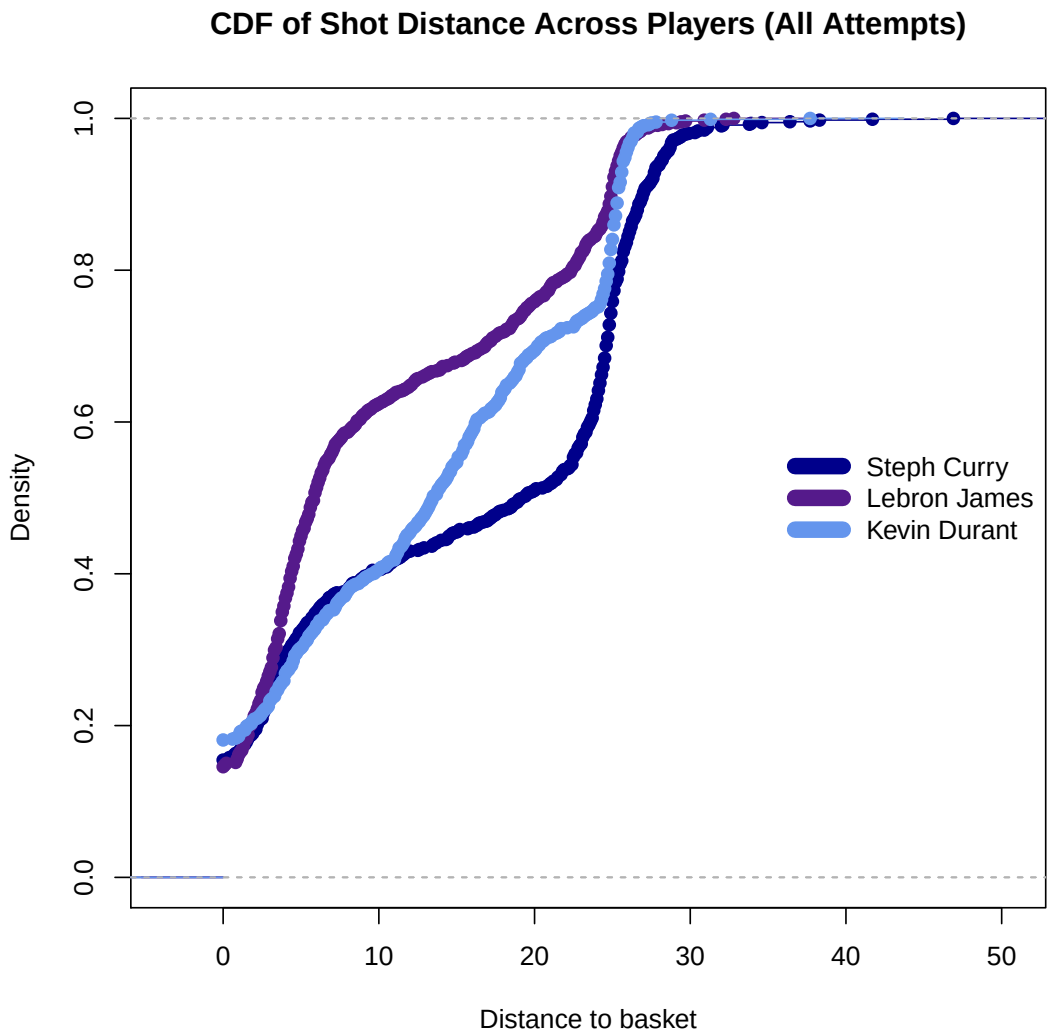
- They display the same underlying information...
- When you are interested in probabilities, representing them with CDF is more conventional, why?
  - A pdf represents probability with areas while a cdf represents probability with (vertical) distances.
  - It is much easier for the eye to compare distance 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??

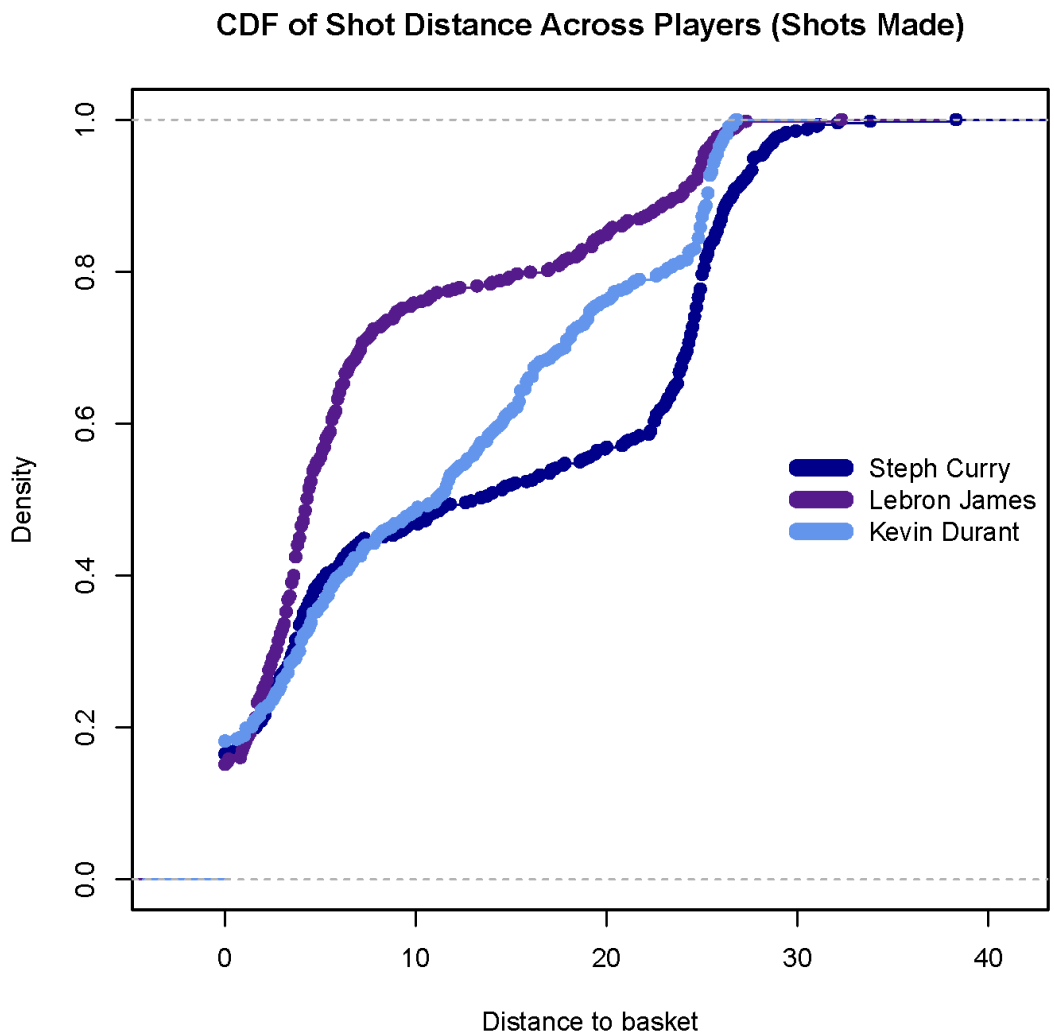


US women are taller than women in Bihar! The height distribution in the US stochastically dominates that in Bihar!!

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



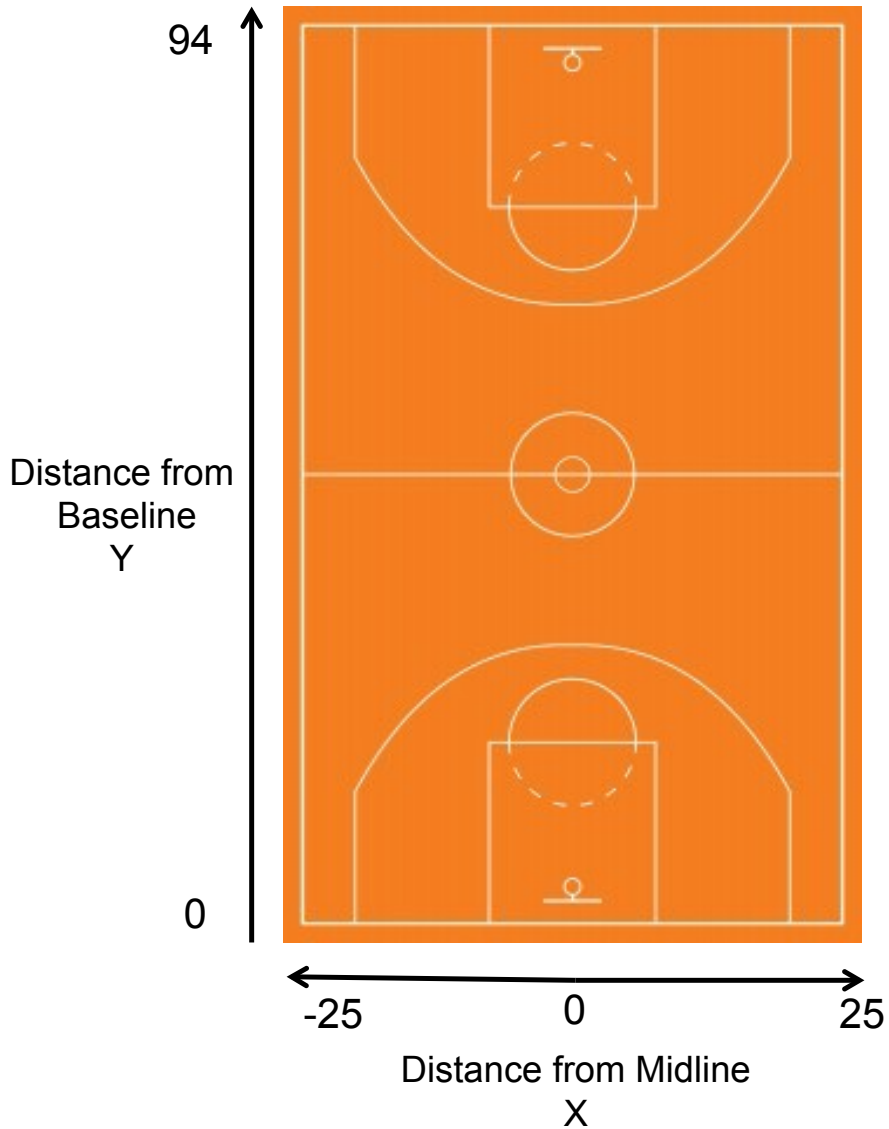
Not too surprised? OK , let's go back to Steph Curry



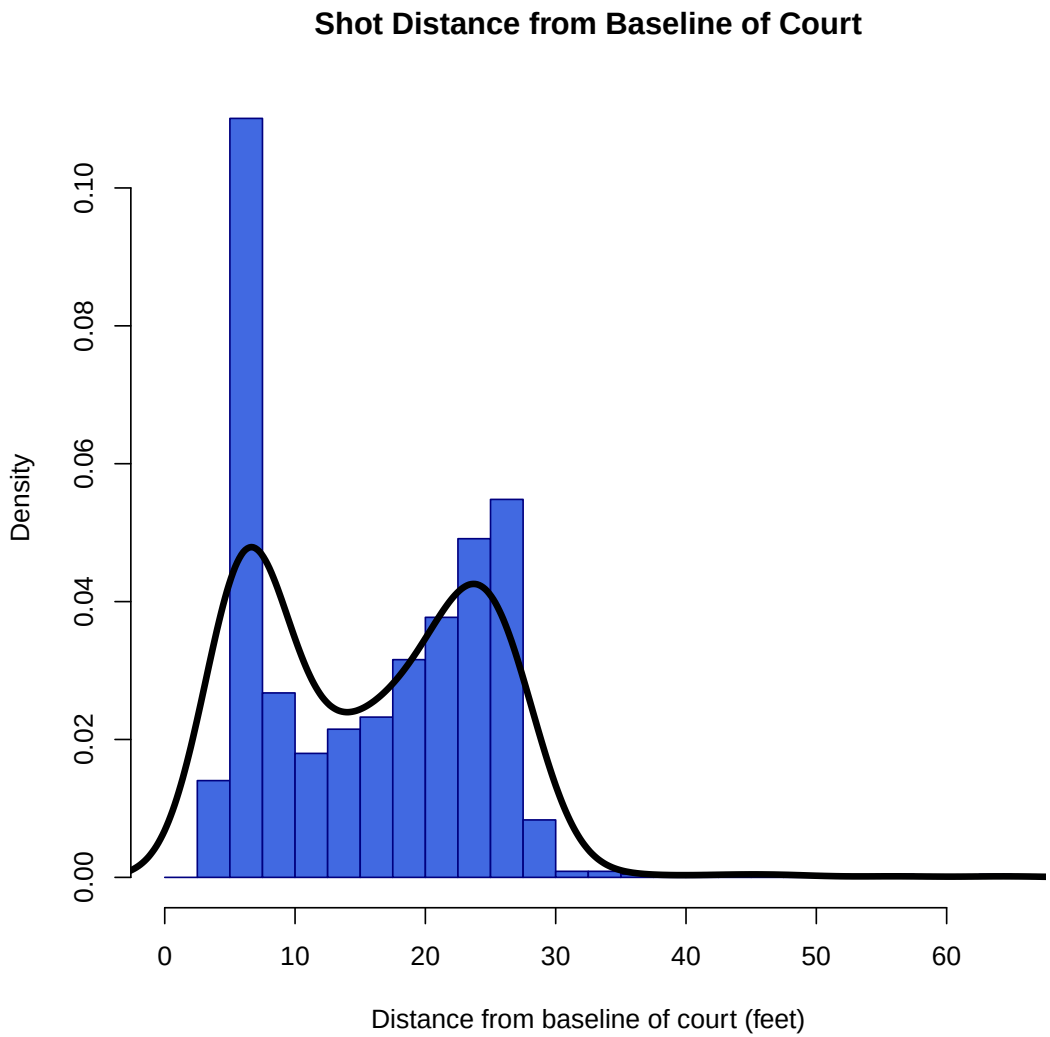
Representing joint distributions

- 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?

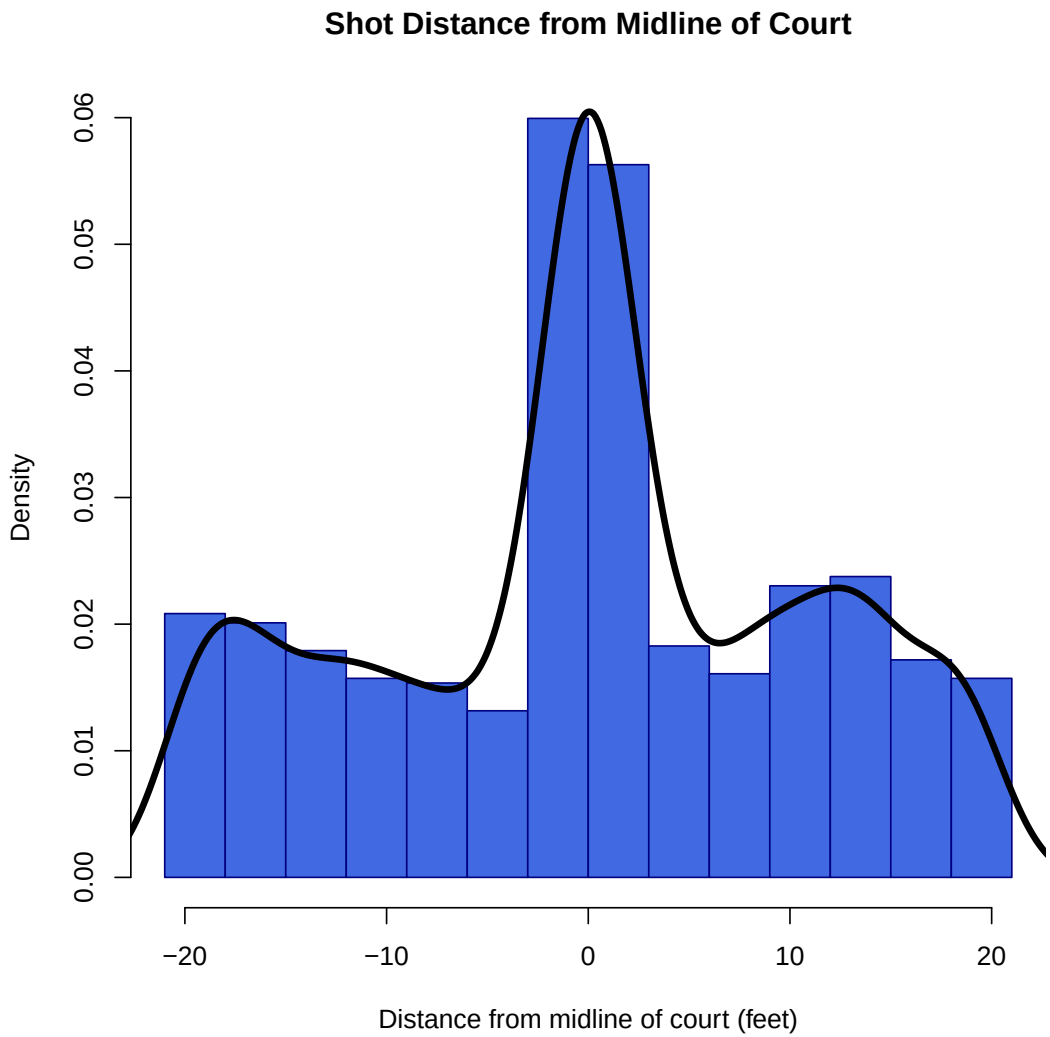
A basketball court



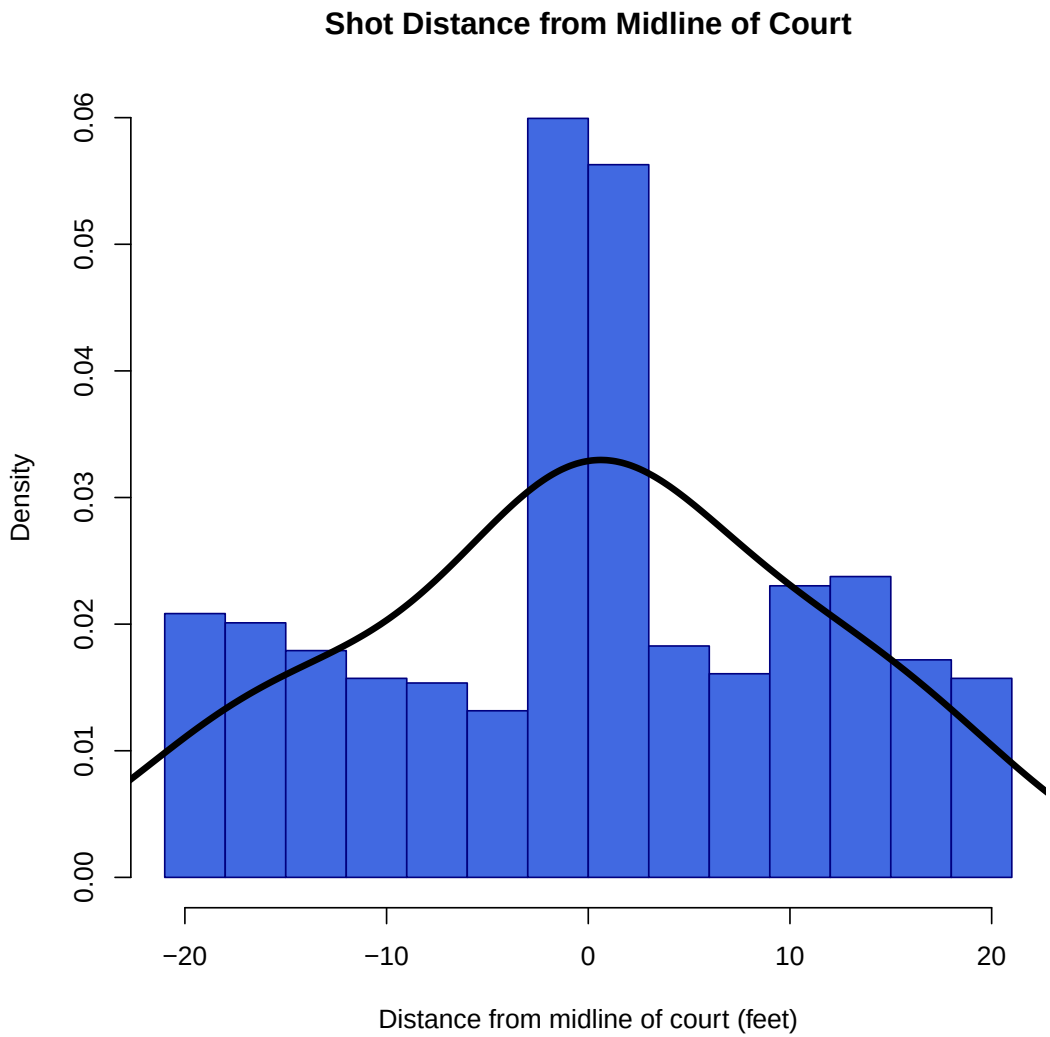
frequency, of shot, by distance from baseline



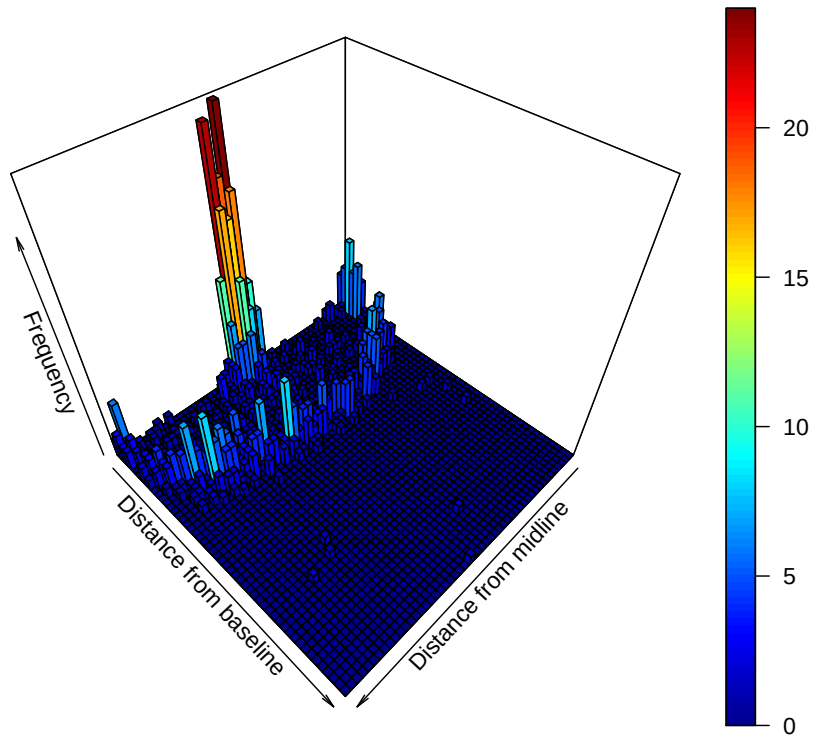
frequency of shots, by distance from midline



frequency of shots, by distance from midline



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



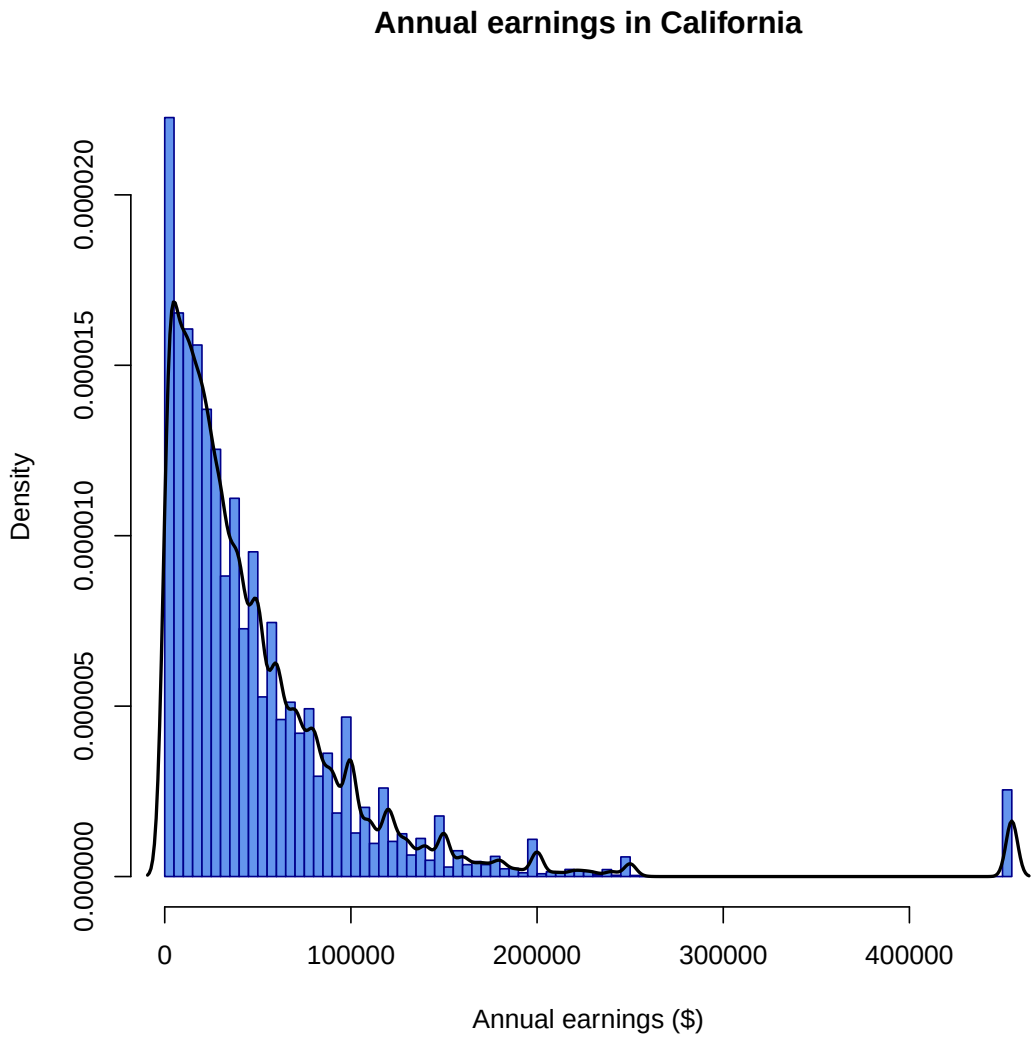
Now we see pretty clearly that there is bunching at the 3pt line!

1. Where can we find data?
2. Visualizing distributions
3. *Exploiting our knowledge of distribution function to extract the information we want*

Top income distribution

- How are wages distributed?
- let's look at California, from the CPS, year 2014

Distribution of earnings in Califonia



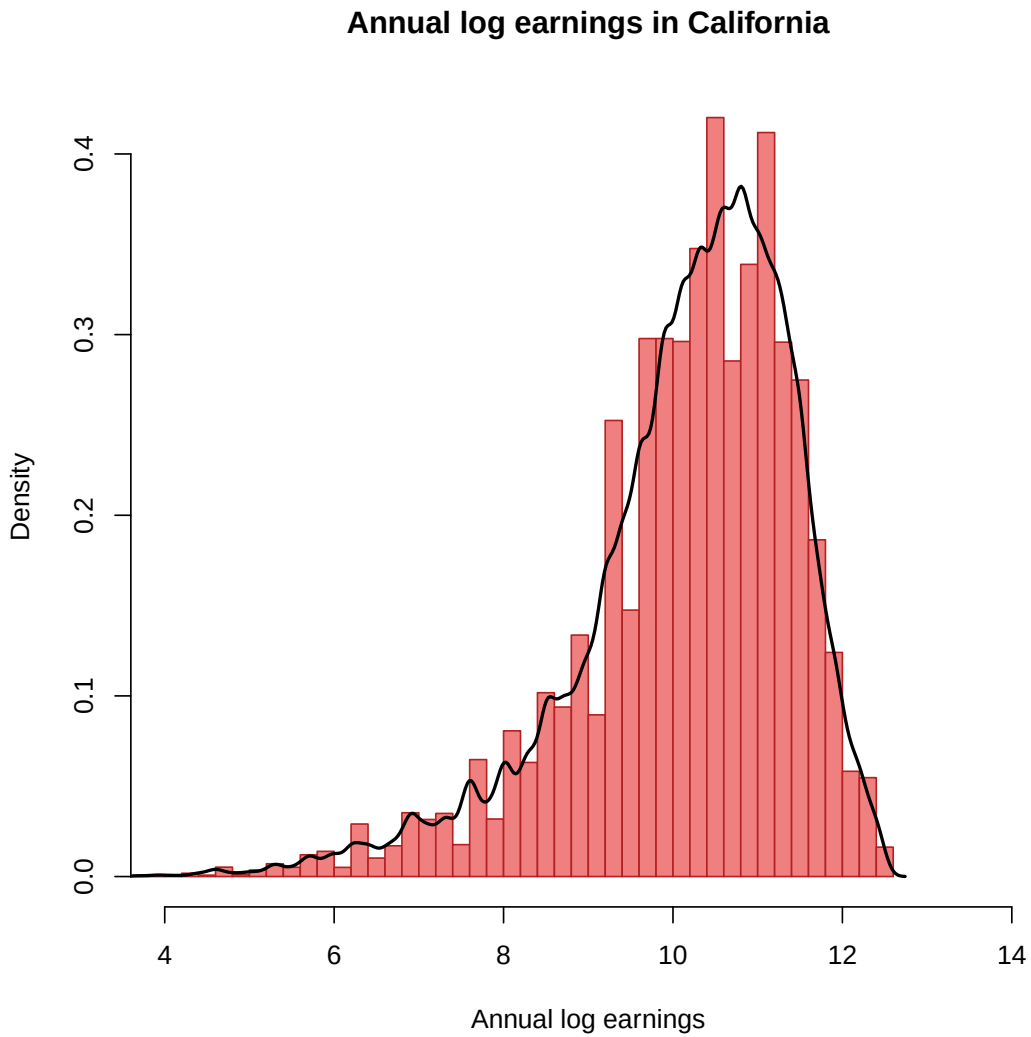
Top income distribution

- How are wages distributed?
- I said that lots of economic variables appear to be approximately normally distributed but that does not appear the case for earnings...
- This is very *skewed*...: lots of mass in the low values, but a long tail.
- Any idea what to do?

Top income distribution

- How are wages distributed?
- I said that lots of economic variables appear to be approximately normally distributed but that does not appear the case for earnings...
- This is very *skewed*...: lots of mass in the low values, but a long tail.
- Any idea what to do?
- Yes... let's try to take logs

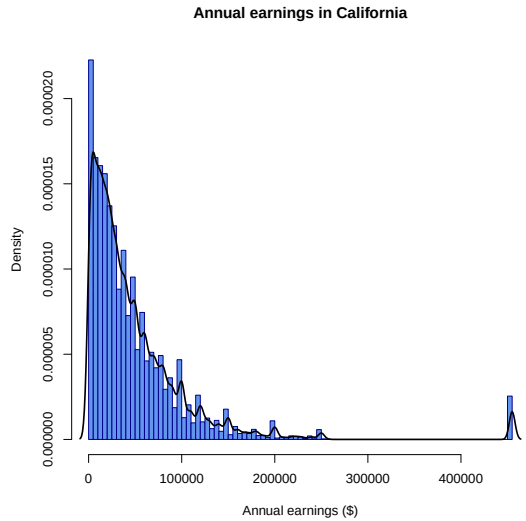
Distribution of log (earnings) in Califonia



Log earnings

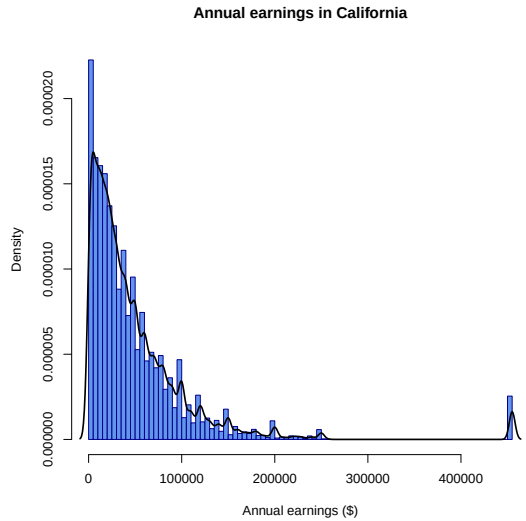
- Hmm... it is a bit squewed the other way, but closer to something normally distributed
- Often, before doing anything with data, it is worth plotting it to see if we should take logs.
- Now... something else was a bit strange in the first distibution

Distribution of earnings in California



- what is this bump at the end?

Distribution of earnings in California



- what is this bump at the end?
- wages are top coded in the census and the CPS to protect privacy
- but what if we worry precisely about the highest earnings? What if our paper is on CEO pay in Sillicon valley?!


WE ARE THE
99%
www.wearethe99percent.us

Where did this come from?

Bernie Sanders, December 10, 2010 ("Filibernie")

> *We cannot give tax breaks to the rich when we already have the most unequal distribution of income of any major country on Earth...The percentage of income going to the top 1 percent nearly tripled since the 1970s...The top 1 percent now owns more wealth than the bottom 90 percent. That is not the foundation of a democratic society...The fact is, 80 percent of all new income earned from 1980 to 2005 has gone to the top 1 percent. People should be mindful of this fact: The last time that type of income disparity took place was in 1928. I think we all know what happened in 1929.*

And where did HE take it from? Them!

Share of income going to the top 1 percent

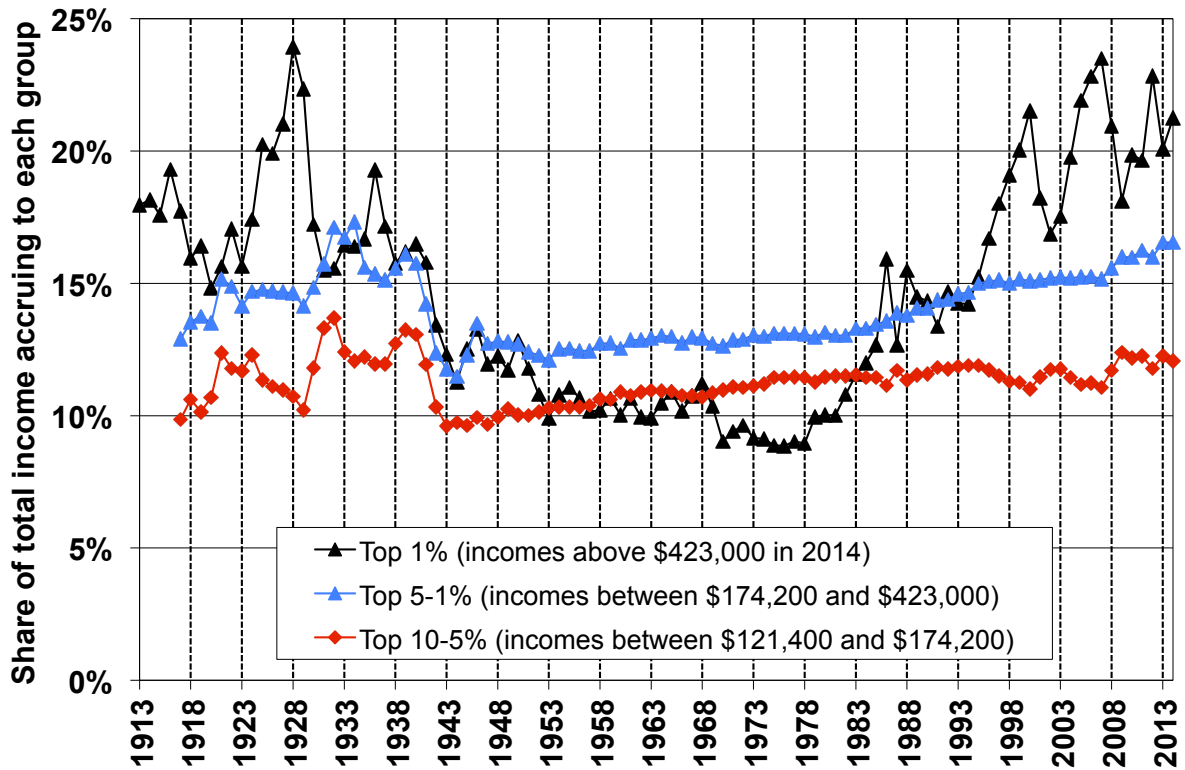


FIGURE 2
Decomposing the Top Decile US Income Share into 3 Groups, 1913-2014

How could they make these graphs?

- The only source of data on the distribution of income that is consistently available over long time periods is tax data.
- Tax data is publicly available in the form of tabulation: These statistics report the number of taxpayers and their total income and tax liability for a large number of income brackets.
- Combining these data with population census data and aggregate income sources, Piketty and Saez calculate the share of total personal income accruing to various upper-income groups, such as the top 10 percent or top 1 percent.
- How? By exploiting a known fact about the distribution of top income: it tends to be well approximated by a Pareto distribution

Pareto Distribution

- The Pareto distribution emerges in many settings, both in economics (size distribution, sizes of cities, income distribution) and in other social sciences (language, family names, popularity, certain social network patterns, crime per convict), and even in various physical environments (e.g., sizes of large earthquakes, power outages, etc.)
- 1-F(x)=$\overline{F}(x) = \Pr(X > x) = \begin{cases} \left(\frac{x_{\mathrm{m}}}{x}\right)^{\lambda} & x \geq x_{\mathrm{m}}, \\ 1 & x < x_{\mathrm{m}}. \end{cases}$
- The average income of those whose income is greater than y is y times the constant $\frac{\lambda}{\lambda-1}$

Zipf's Law

- The Zipf distribution is a special case with $\lambda = 1$ (or sometimes it is used for the case where it is approximately equal to 1).
- The empirical distribution of city size and firm size appear to be well approximated by this distribution, and for city size distributions, this is generally referred to Zipf's law
- This is also equivalent to a relationship of slope -1 between log rank of the city (according to city size) and log of the population

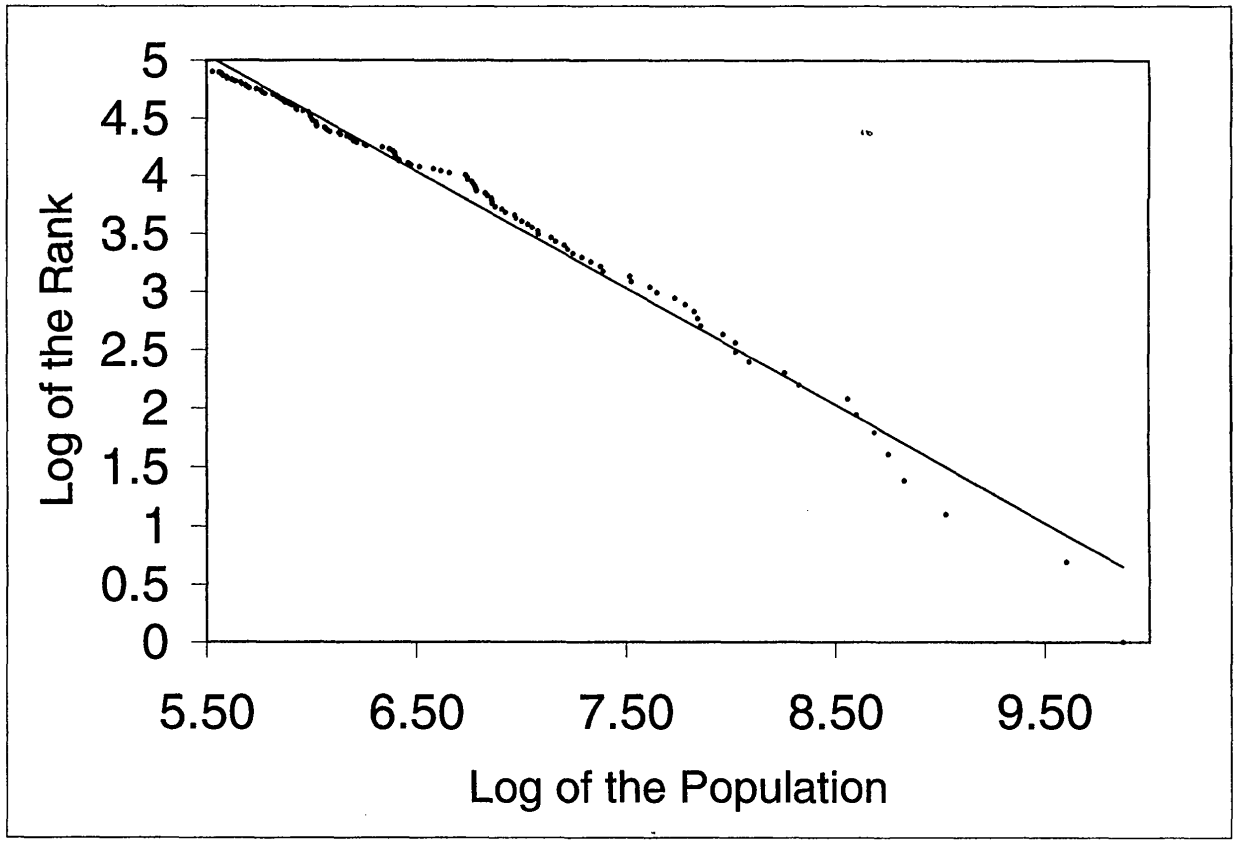


FIGURE I
Log Size versus Log Rank of the 135 largest U. S. Metropolitan Areas in 1991
Source: Statistical Abstract of the United States [1993].

$$\ln \text{Rank} = 10.53 - 1.005 \ln \text{Size}, \tag{1}$$
$$(.010)$$

Top inequality and the Pareto distribution

- The income distribution is Pareto at the top, then one can derive simple expressions for the share of top 1%, or top 10% etc.
- For example, suppose that the entire income distribution is given by Pareto with a shape parameter λ
- Then the top qth percentile's share of total income can be derived as:

$$(\frac{q}{100})^{\frac{\lambda-1}{\lambda}}$$

- This expression makes it clear that a lower λ corresponds to a thicker tail of the Pareto distribution and thus to a greater share of total income being captured by individuals/households at higher percentiles of the distribution. For example, with $\lambda = 2$, the top 1% share is 10%, and with $\lambda = 3$, it is 4%.
- All we have to do now is to use the available data to estimate λ and x_m

Estimating pareto coefficients for each year

- While the top income distribution is approximately Pareto, it is best to estimate the Pareto coefficient as closely as you can from the percentile you are interested
- The tables give in each tax bracket the number of people in this bracket, and the sum of the income of people in this bracket
- For any income s, Piketty-Saez calculate the minimum income for people above the line (s), the fraction of people above s (p), and their average income (y)
- In each braket: $b = \frac{y}{s}$, $\lambda = \frac{b}{b-1}$, and $x_m = sp^{\frac{1}{\lambda}}$
- You can then calculate the share going to the relevant percentile using the closest bracket.

References

- Saez, Emmanuel "Striking it Richer: The Evolution of Top Incomes in the United States (Updated with 2014 preliminary estimates)", Emmanuel Saez, UC Berkeley? June 25, 2015
- Xavier Gabaix Zipf's Law for Cities: An Explanation. The Quarterly Journal of Economics, Vol. 114, No. 3 (Aug., 1999), pp. 739-767