Course 2 - Gaussian Versus Pareto Distributions

UCLA - Econ 19 - Fall 2018

François Geerolf

Contents

Introduction

During this course, we shall try to understand a technical passage in Rosen [1983]'s *The American Scholar* piece:

> Of particular interest here is an observation, first studied systematically by the great Italian economist Vilfredo Pareto in the late nineteenth century, that the distribution of income contains an unusually large proportion of top earners: that is, among the rich rather than the poor. A visual image will perhaps clarify what is meant by "unusual" in this connection. Imagine a graph plotting IQ scores on the horizontal and the frequency of scores on the vertical. The result is a familiar bell-shaped curve. The peak of the bell occurs at a score arbitrarily scaled at 100 and the curve falls symmetrically on either side of 100. Now picture a similar graph, except with earnings on the horizontal. The resulting curve is unbalanced and nonsymmetrical - a bell that is definitely out of whack. To the left of the modal (peak) value it appears much like the IQ frequency curve. However, to the right of the mode it does not fall as fast as it does to the left. It looks as if someone had stood at the right end of the curve, placed it over his back like a rope, and dragged and stretched it out a very long distance. The upper or right-hand tail of the distribution of income is much thicker than the lower, left-hand tail. The extra weight on the right lends a certain skewness to the distribution of income. What this comes down to is that the distribution of earnings is far from proportionate to the distribution of ability. Amazingly, Pareto's observations have been qualitatively duplicated in virtually every era of every society for which data on income distributions can be found.

In this passage, Sherwin Rosen draws a sharp distribution between Gaussian distributions on the one hand (characterized by the well known bell-shaped curve) and Pareto distributions on the other hand:

1. "Imagine a graph plotting IQ scores on the horizontal and the frequency of scores on the vertical. The result is a familiar Bell-shaped curve."
2. "The upper or right-hand tail of the distribution of income is much thicker than the lower, left-hand tail. The extra weight on the right lends a certain skewness to the distribution of income. What this comes down to is that the distribution of earnings is far from proportionate to the distribution of ability."

We first investigate the mathematics of these different distributions, before proceeding to describing some real-world statistical distributions, and connect them to Bell-shaped curves on the one hand and Pareto distributions on the other hand.

1 Mathematics of Statistical Distributions

In order to understand Sherwin Rosen's above comment, I need to take you through some mathematics. Do not panic ! I am going to take you through everything, and a prerequisite of mathematics from high school should be sufficient.

1.1 Bell-Shaped Distributions

The Bell shape curve is defined by a density function given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right).$$

One implication is that the density of a Bell Shaped curve goes very rapidly to zero as x goes to infinity. When premultiplied by any power function x^a, no matter how large a, the density of a Bell-shaped curve still converges to zero, which means that the density is negligible compared to any power function when x to infinity:

$$\text{for all } a > 0, \quad \lim_{x \to +\infty} x^a f(x) = 0.$$

Intuitively, this means that the Gaussian Distribution goes "very fast" to zero as x becomes large, faster in fact than usual functions which are thought to go very fast to zero (thing, for example of $x^1 0000$ when x becomes large).

Here is a link to the Google Sheets that we created in order to look at the Gaussian distribution. In particular, we were able to plot the density function of a Normal Distribution with $\mu = 0$ and $\sigma = 1$, using the formula above. Note: this Google Sheet is read only. However, you may copy and paste from this Google Sheet, and choose your own values for μ and σ.

1.2 Pareto Distributions

A key feature of the Pareto distribution is that the density distribution does not go as fast to 0 as with the Gaussian Distribution, as x becomes large.

In the context For concreteness, if x is population, then this would mean that there are relatively many cities with a large size, especially when assessed against the average city size, as well as its standard deviation. Similarly, there are relatively many incomes that are much larger than the mean. The Pareto Distribution is in fact defined by:

$$f(x) = a \frac{x_m^a}{x^{a+1}}.$$

For the cumulative distribution function, this implies:

$$1 - F(x) = \left(\frac{x_m}{x}\right)^a.$$

2 Some Real-Life Distributions

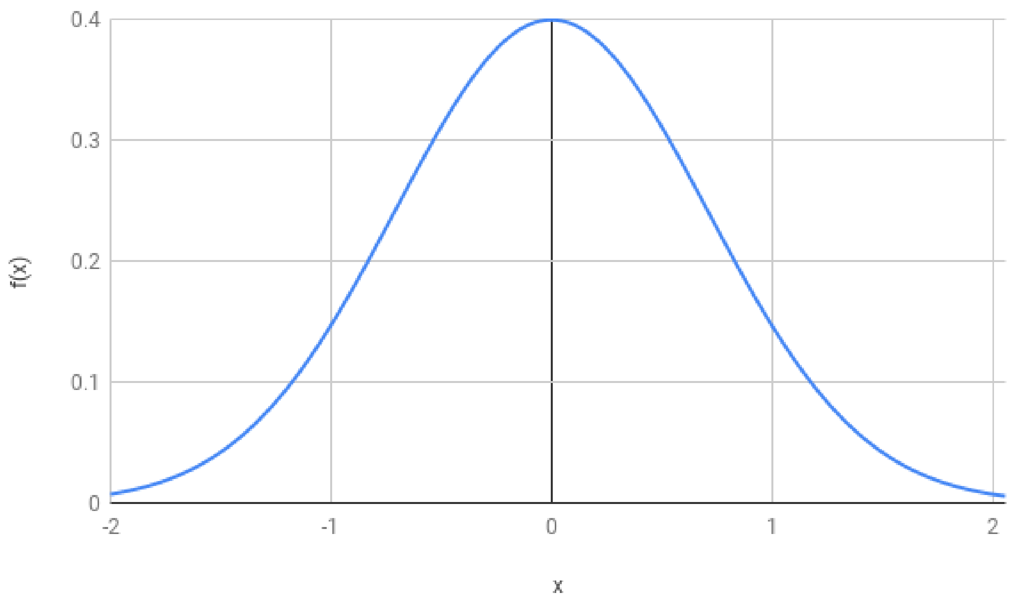


Figure 1: Bell Shaped Curve

```
pklist <- c("tidyverse", "rvest", "scales")
source("https://fgeerolf.github.io/code/load-packages.R")
options(tibble.print_max = 100)
```

2.1 Natural Sciences

Many distributions in the natural sciences are well described by a Bell-shaped curve. In order to illustrate this, let us use the National Longitudinal Surveys (NLS) from the Bureau of Labor Statistics which tracks the income, education, and life circumstances of a large cohort of Americans across several decades. We use `summary` in order to summarise our dataset.

```
height <- read.csv("https://raw.githubusercontent.com/hadley/r4ds/master/data/heights.csv")
height %>%
  summary
```

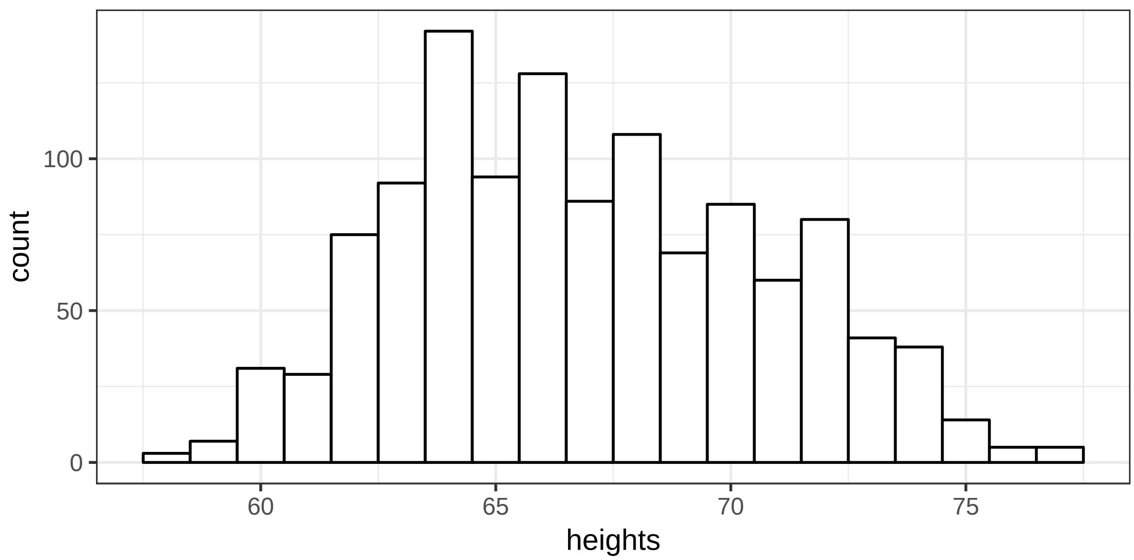
```
##       earn            height          sex              ed       
##  Min.   :   200   Min.   :57.50   female:687   Min.   : 3.0  
##  1st Qu.: 10000   1st Qu.:64.01   male  :505   1st Qu.:12.0  
##  Median : 20000   Median :66.45                Median :13.0  
##  Mean   : 23155   Mean   :66.92                Mean   :13.5  
##  3rd Qu.: 30000   3rd Qu.:69.85                3rd Qu.:16.0  
##  Max.   :200000   Max.   :77.05                Max.   :18.0  
##       age             race    
##  Min.   :18.00   black   :112  
##  1st Qu.:29.00   hispanic: 66  
##  Median :38.00   other   : 25  
##  Mean   :41.38   white   :989  
##  3rd Qu.:51.00
```

```
## Max.   :91.00
```

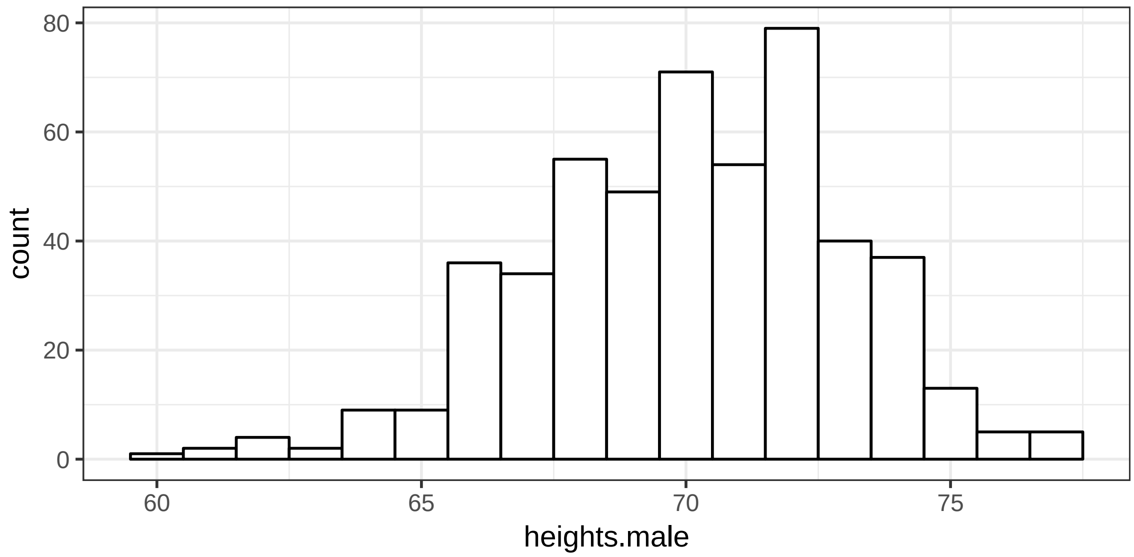
The variable names are pretty self-explanatory. We are in particular interested by the distribution of height, possibly by gender.

```r
heights <- height %>%
  select(height) %>%
  unlist %>%
  unname

heights.male <- height %>%
  filter(sex == "male") %>%
  select(height) %>%
  unlist %>%
  unname

heights.female <- height %>%
  filter(sex == "female") %>%
  select(height) %>%
  unlist %>%
  unname
```

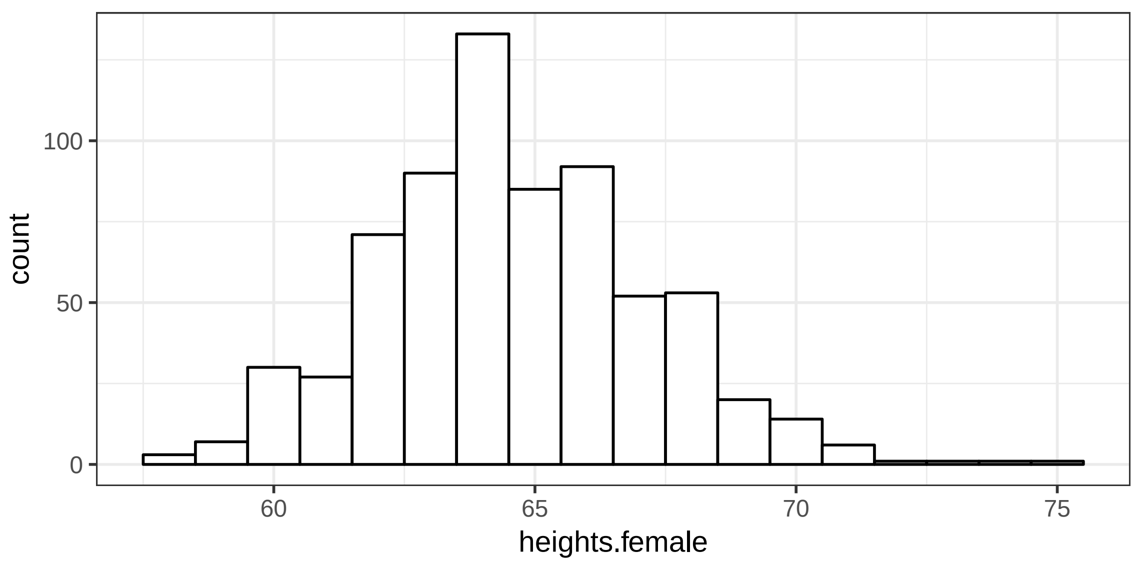
```r
ggplot() +
  aes(heights) +
  geom_histogram(binwidth = 1, colour = "black", fill = "white") +
  theme_bw()
```




```r
ggplot() +
  aes(heights.male) +
  geom_histogram(binwidth = 1, colour = "black", fill = "white") +
  theme_bw()
```




```
ggplot() +
  aes(heights.female) +
  geom_histogram(binwidth = 1, colour = "black", fill = "white") +
  theme_bw()
```



2.2 Data on cities using Google Sheets

We have used this Google Spreadsheet in order to plot the city size distribution of cities.

We note that the result is something that is close to a linear relationship, suggestive of Pareto like behavior.

2.3 Data on cities using R Statistical software

The data comes from the following Wikipedia entry: List of United States cities by population.

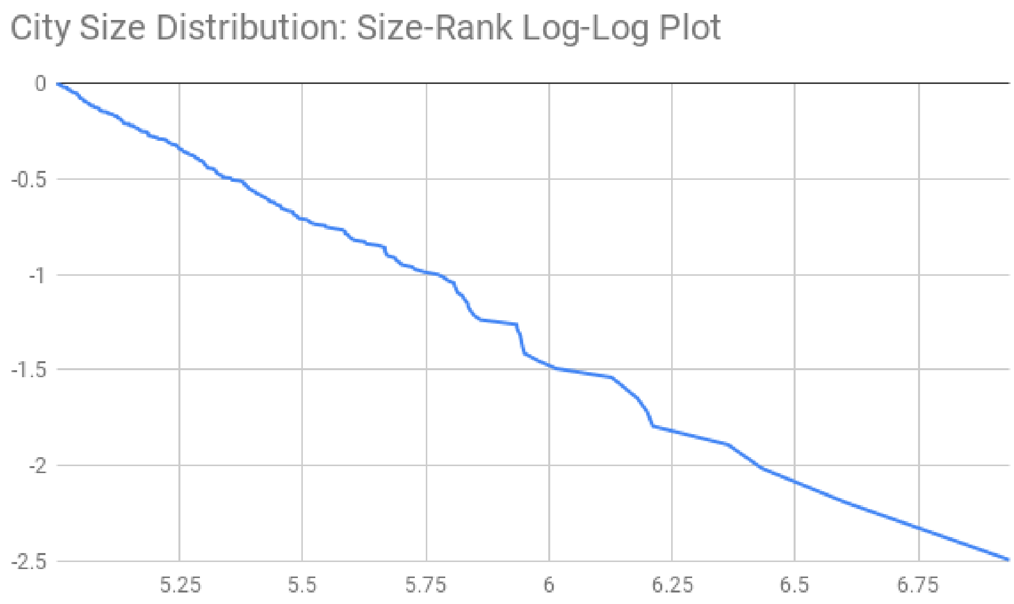


Figure 2: City Size Distribution

```
data <- "https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population" %>%
  read_html %>%
  html_table(header = TRUE, fill = TRUE)
```

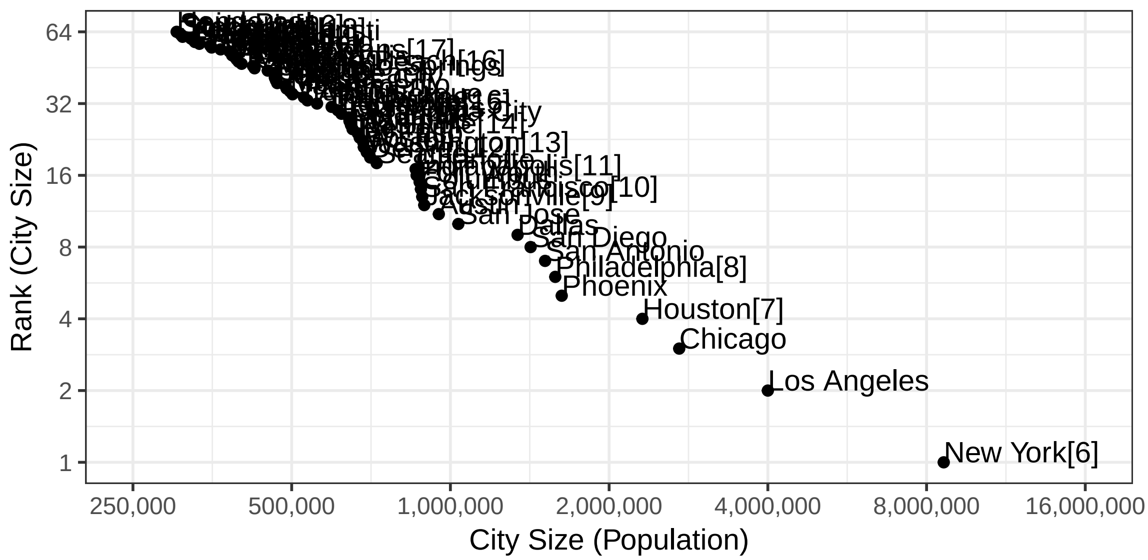
Biggest cities:

```
data[[5]][,c(1:4)] %>%
  as.tibble %>%
  select(rank = "2017rank", "City", state = "State[5]", pop = "2017estimate") %>%
  head(28)
```

```
## # A tibble: 28 x 4
##     rank City              state              pop
##    <int> <chr>             <chr>              <chr>
##  1     1 New York[6]       New York           8,622,698
##  2     2 Los Angeles       California         3,999,759
##  3     3 Chicago           Illinois           2,716,450
##  4     4 Houston[7]        Texas              2,312,717
##  5     5 Phoenix           Arizona            1,626,078
##  6     6 Philadelphia[8]   Pennsylvania       1,580,863
##  7     7 San Antonio       Texas              1,511,946
##  8     8 San Diego         California         1,419,516
##  9     9 Dallas            Texas              1,341,075
## 10    10 San Jose          California         1,035,317
## 11    11 Austin            Texas              950,715
## 12    12 Jacksonville[9]   Florida            892,062
## 13    13 San Francisco[10] California         884,363
## 14    14 Columbus          Ohio               879,170
## 15    15 Fort Worth        Texas              874,168
## 16    16 Indianapolis[11]  Indiana            863,002
```

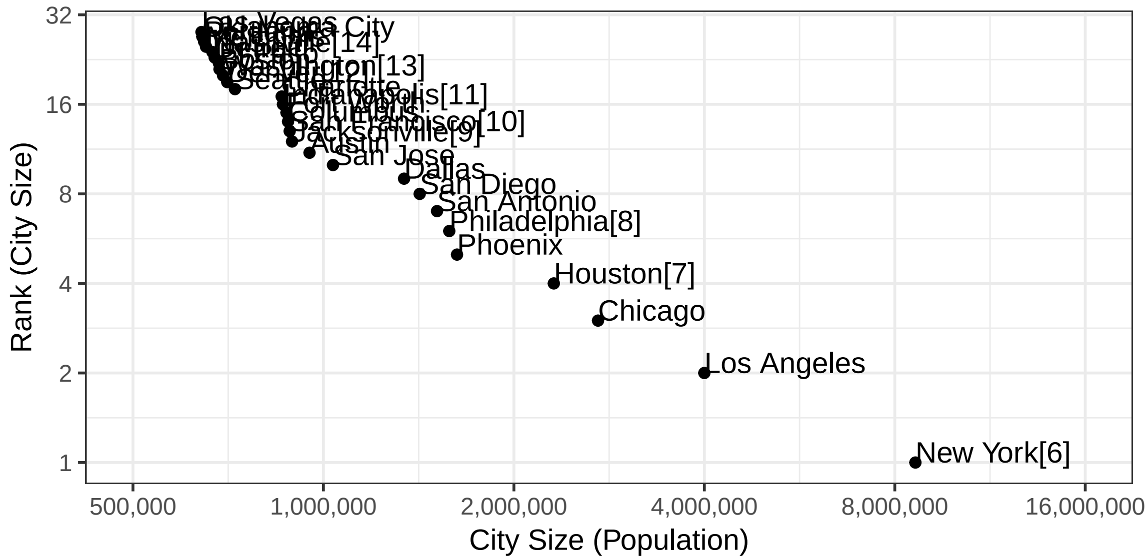
```
## 17     17 Charlotte         North Carolina       859,035
## 18     18 Seattle           Washington           724,745
## 19     19 Denver[12]        Colorado             704,621
## 20     20 Washington[13]    District of Columbia 693,972
## 21     21 Boston            Massachusetts        685,094
## 22     22 El Paso           Texas                683,577
## 23     23 Detroit           Michigan             673,104
## 24     24 Nashville[14]     Tennessee            667,560
## 25     25 Memphis           Tennessee            652,236
## 26     26 Portland          Oregon               647,805
## 27     27 Oklahoma City     Oklahoma             643,648
## 28     28 Las Vegas         Nevada               641,676
```

```
data[[5]][,c(1:4)] %>%
  as.tibble %>%
  select(rank = "2017rank", pop = "2017estimate", "City") %>%
  mutate(pop = pop %>% gsub(",", "", .) %>% as.numeric) %>%
  arrange(-pop) %>%
  mutate(rank = 1:n()) %>%
  ggplot(aes(x = pop, y = rank)) + geom_point() + theme_bw() +
  geom_text(aes(label = City), hjust = 0, vjust = 0) +
  scale_y_log10(breaks = 2^(seq(0, 10, 1)),
                limits = c(1, 64)) +
  scale_x_log10(breaks = 250000*2^seq(0, 10, 1),
                limits = c(250000, 16000000),
                labels = comma) +
  ylab("Rank (City Size)") + xlab("City Size (Population)")
```




```
data[[5]][,c(1:4)] %>%
  as.tibble %>%
  select(rank = "2017rank", pop = "2017estimate", "City") %>%
  mutate(pop = pop %>% gsub(",", "", .) %>% as.numeric) %>%
  arrange(-pop) %>%
  mutate(rank = 1:n()) %>%
  filter(pop >= 500000) %>%
  ggplot(aes(x = pop, y = rank)) + geom_point() + theme_bw() +
  geom_text(aes(label = City), hjust = 0, vjust = 0) +
```

```
  scale_y_log10(breaks = 2^(seq(0, 10, 1)),
                limits = c(1, 28)) +
  scale_x_log10(breaks = 500000*2^seq(0, 10, 1),
                limits = c(500000, 16000000),
                labels = comma) +
  ylab("Rank (City Size)") + xlab("City Size (Population)")
```



References

Sherwin Rosen. The Economics of Superstars. *The American Scholar*, 52(4):449–460, 1983. ISSN 0003-0937. URL http://www.jstor.org/stable/41210977.