Data Analysis 2: Fundaments of Statistics

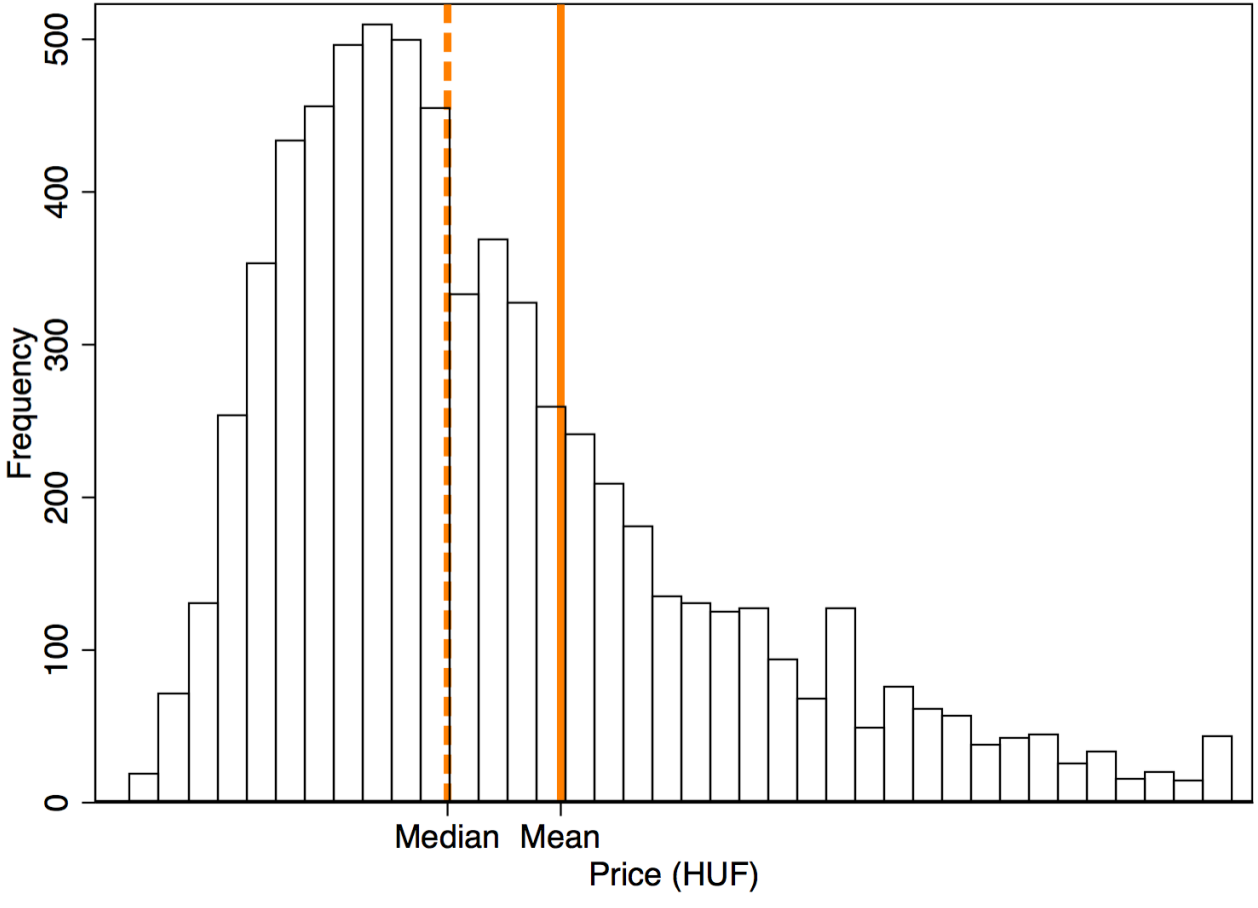
Instructor: Arieda Muço, Fall 2017

Types of Random Variables

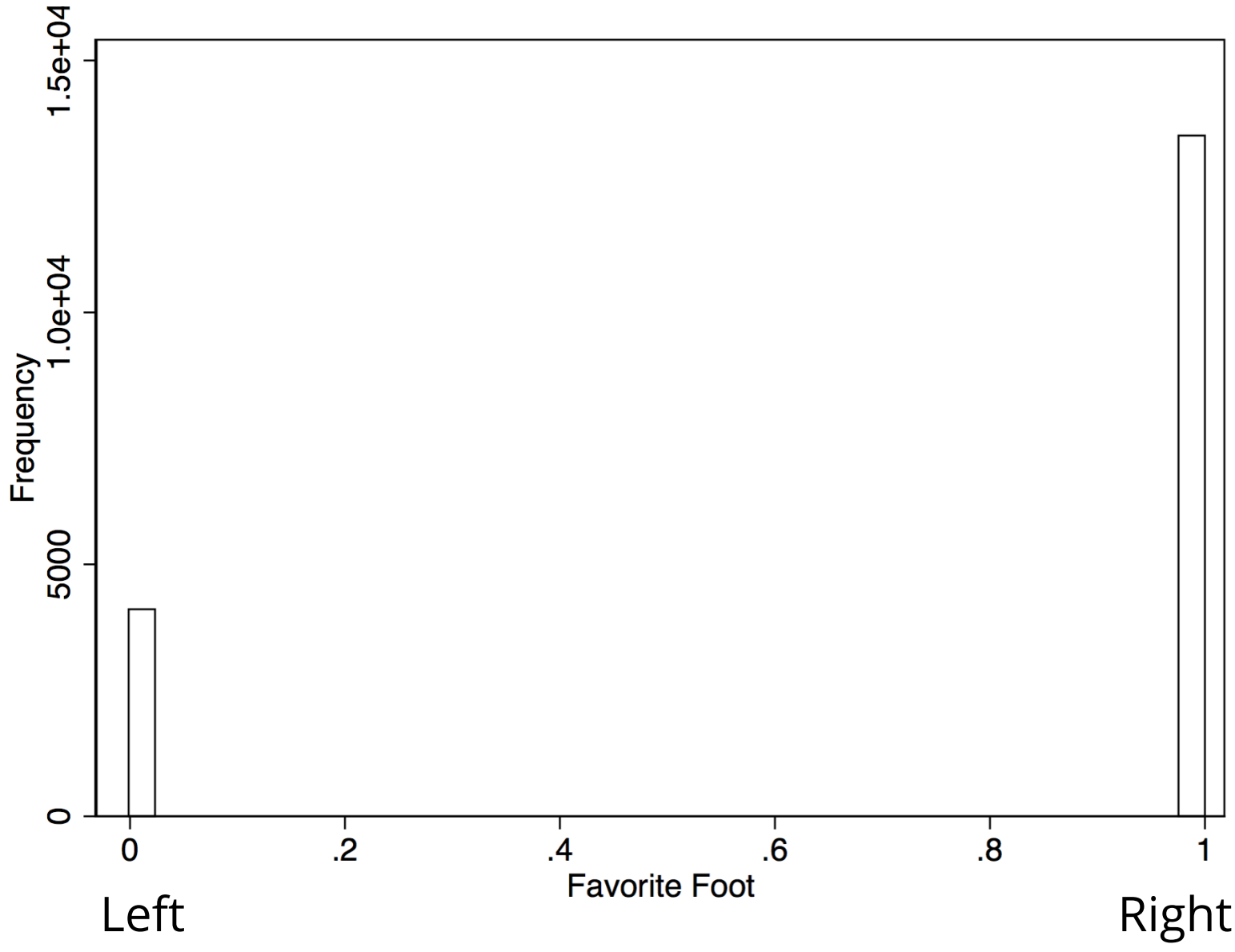
- Continuous, takes values in any interval
  - i.e prices, temperature, grades...
- Discrete, takes no more than a countable number of values
  - i.e hotel stars, gender, number of rooms

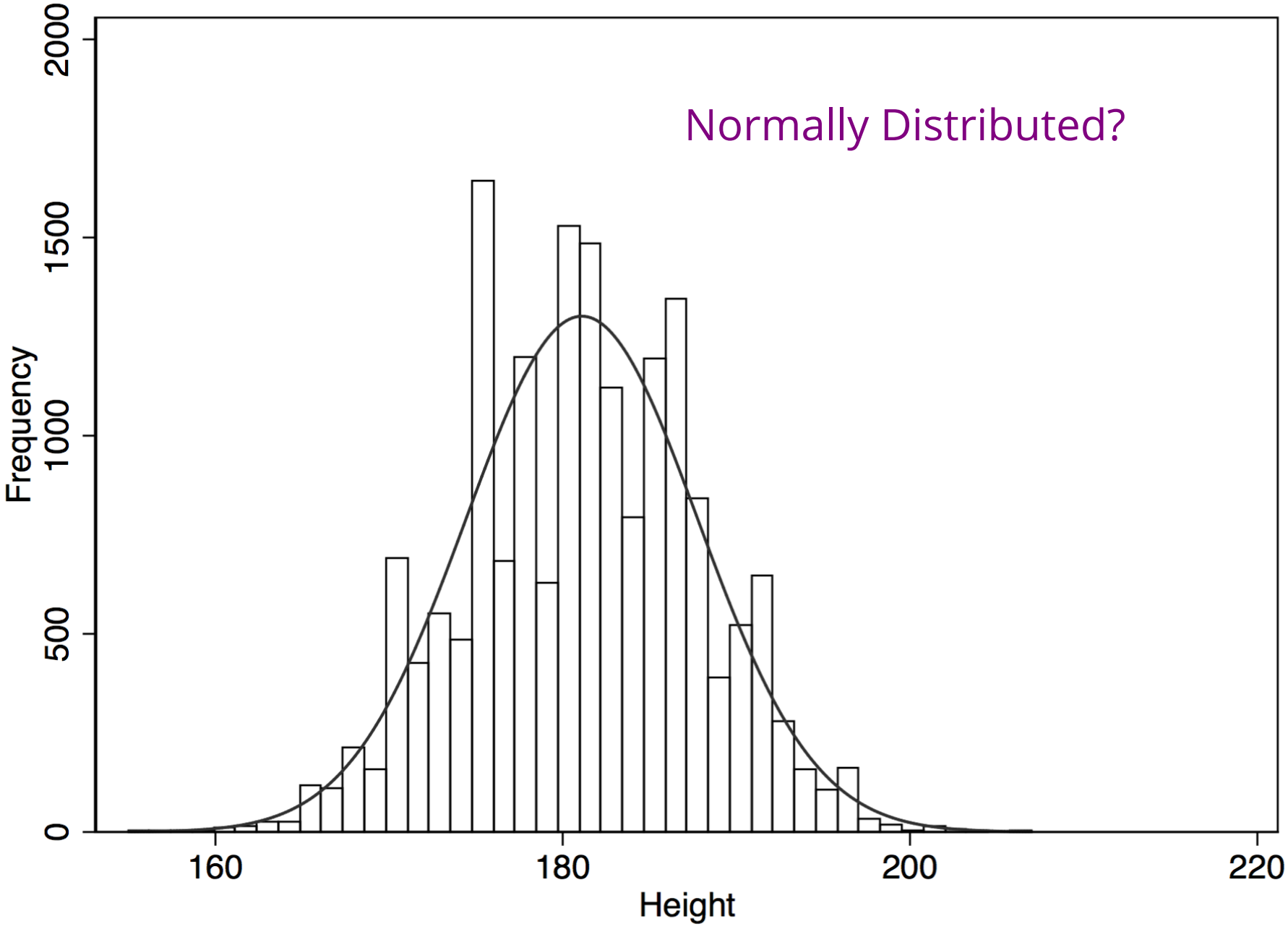
Each have their own distributions

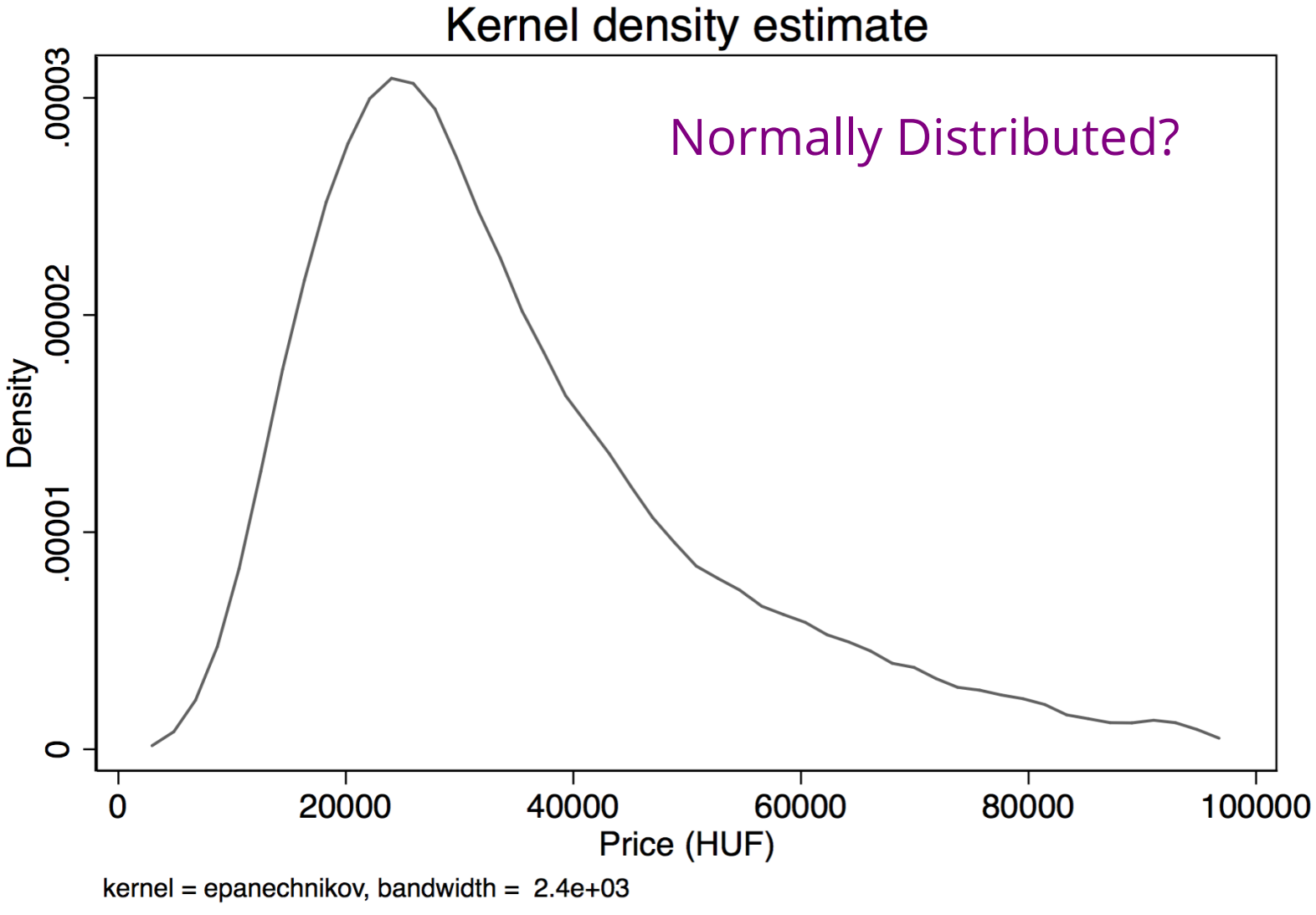
Histograms in practice

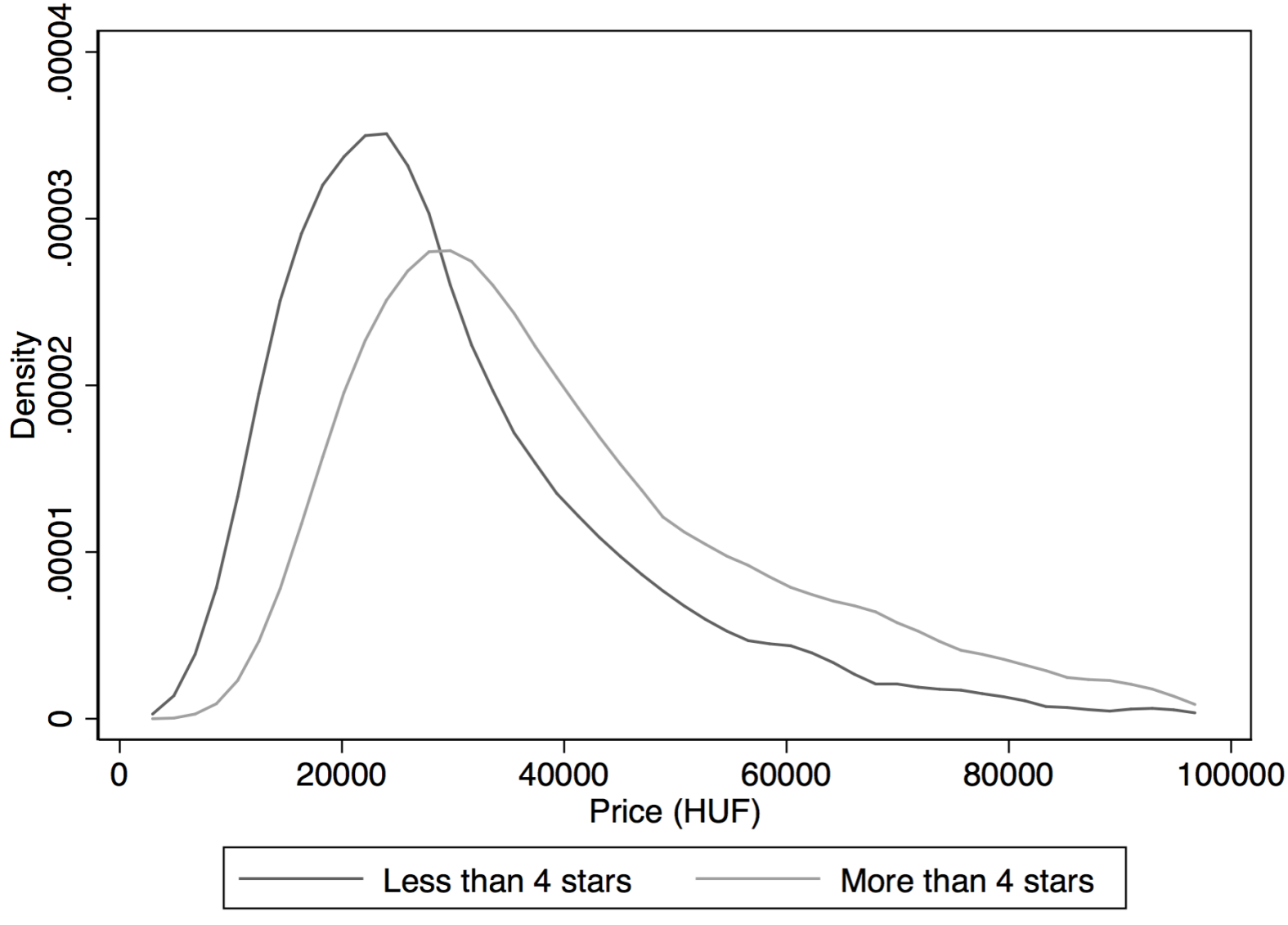


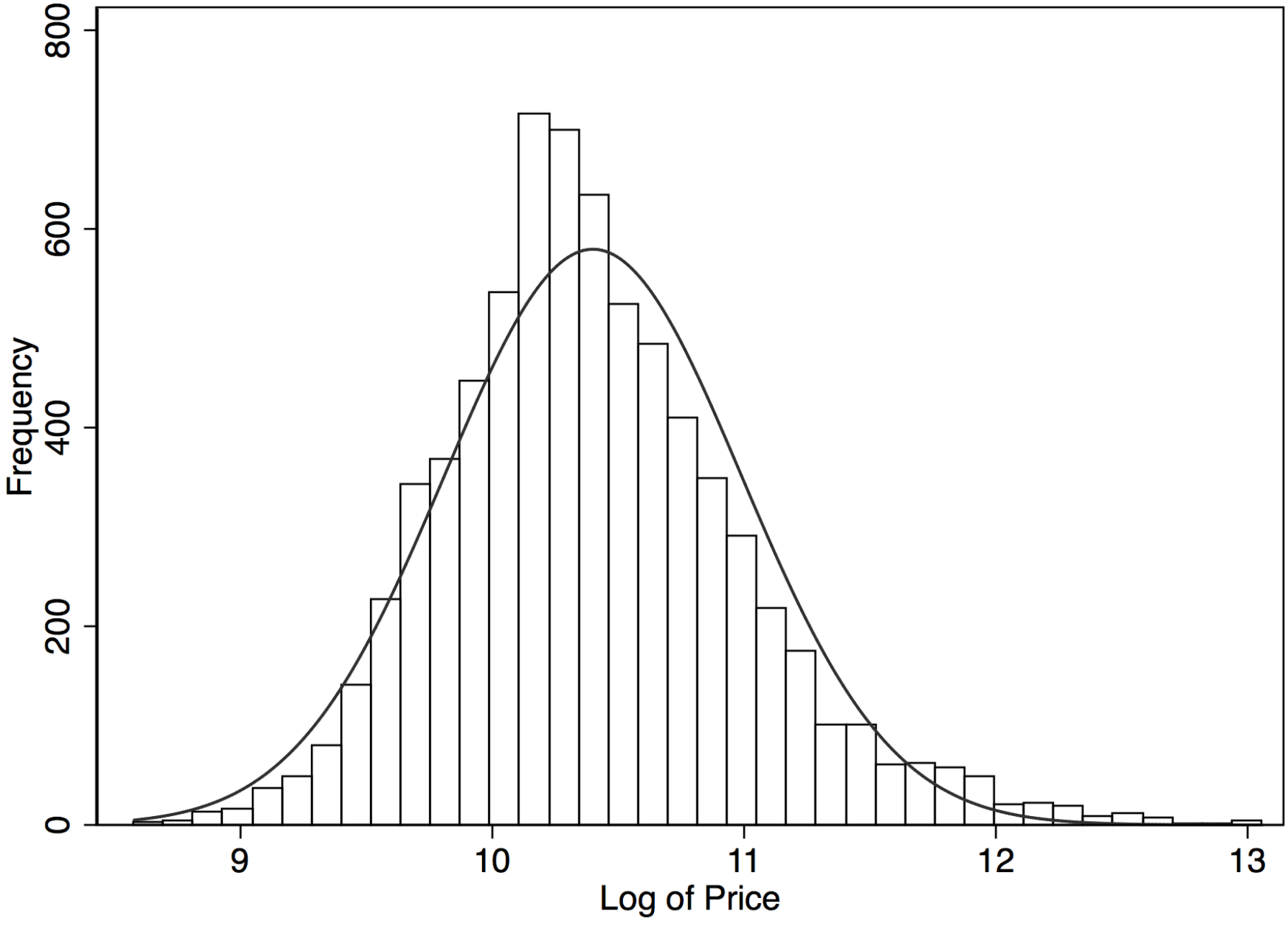
If I tell you that football players have a favorite foot?

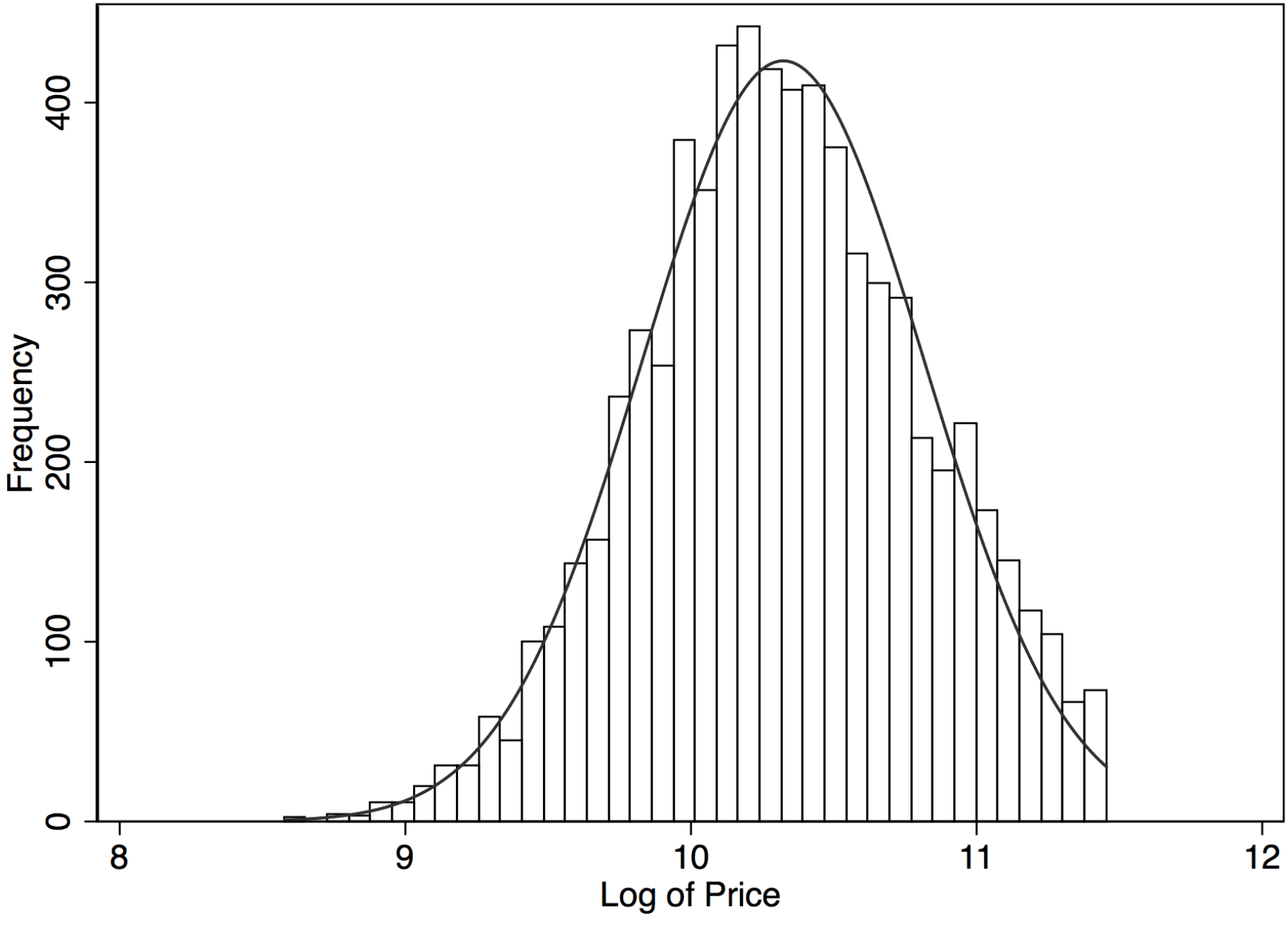

Frequency
0
5000
1.0e+04
1.5e+04
0
.2
.4
.6
.8
1
Favorite Foot
Left
Right


Normally Distributed?
Frequency
0
500
1000
1500
2000
160
180
200
220
Height


Kernel density estimate
Normally Distributed?
Density
0
.00001
.00002
.00003
0
20000
40000
60000
80000
100000
Price (HUF)
kernel = epanechnikov, bandwidth = 2.4e+03


Density
0
.00001
.00002
.00003
.00004
0
20000
40000
60000
80000
100000
Price (HUF)
Less than 4 stars
More than 4 stars


Frequency
0
200
400
600
800
9
10
11
12
13
Log of Price


Frequency
0
100
200
300
400
8
9
10
11
12
Log of Price

Covariance

- Is a measure of the linearity of relationships between two paired variables.
- It provides an indication of the linear relationship between the two variables

$$Cov(x, y) = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n}$$

Covariance

- In case: $y_i = a + bx_i$

$$Cov(x, y) = \frac{b(\sum_i (x_i - \overline{x})(y_i - \overline{y}))}{n}$$

Correlation

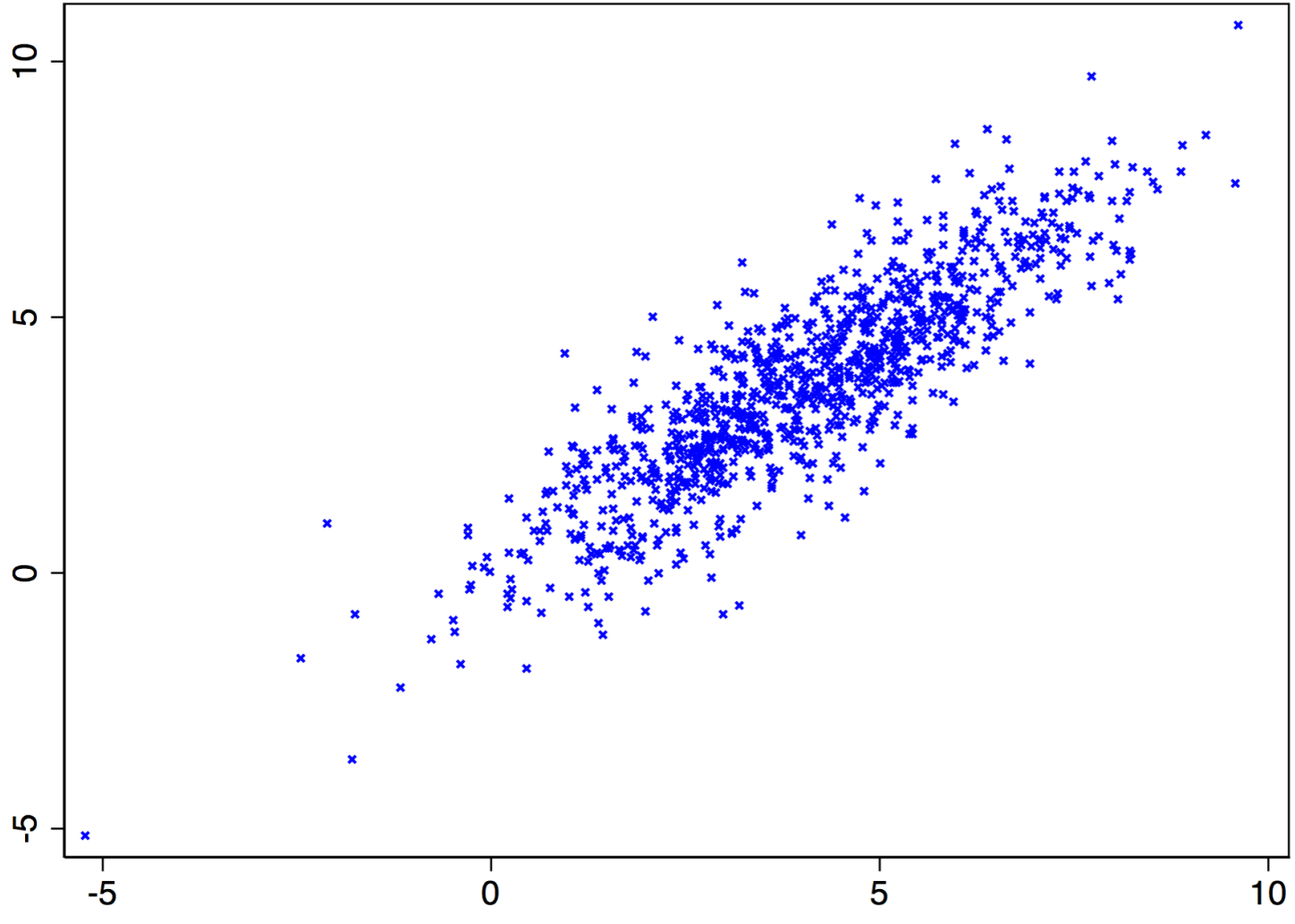
- Is computed by dividing the covariance by the standard deviation of each variable

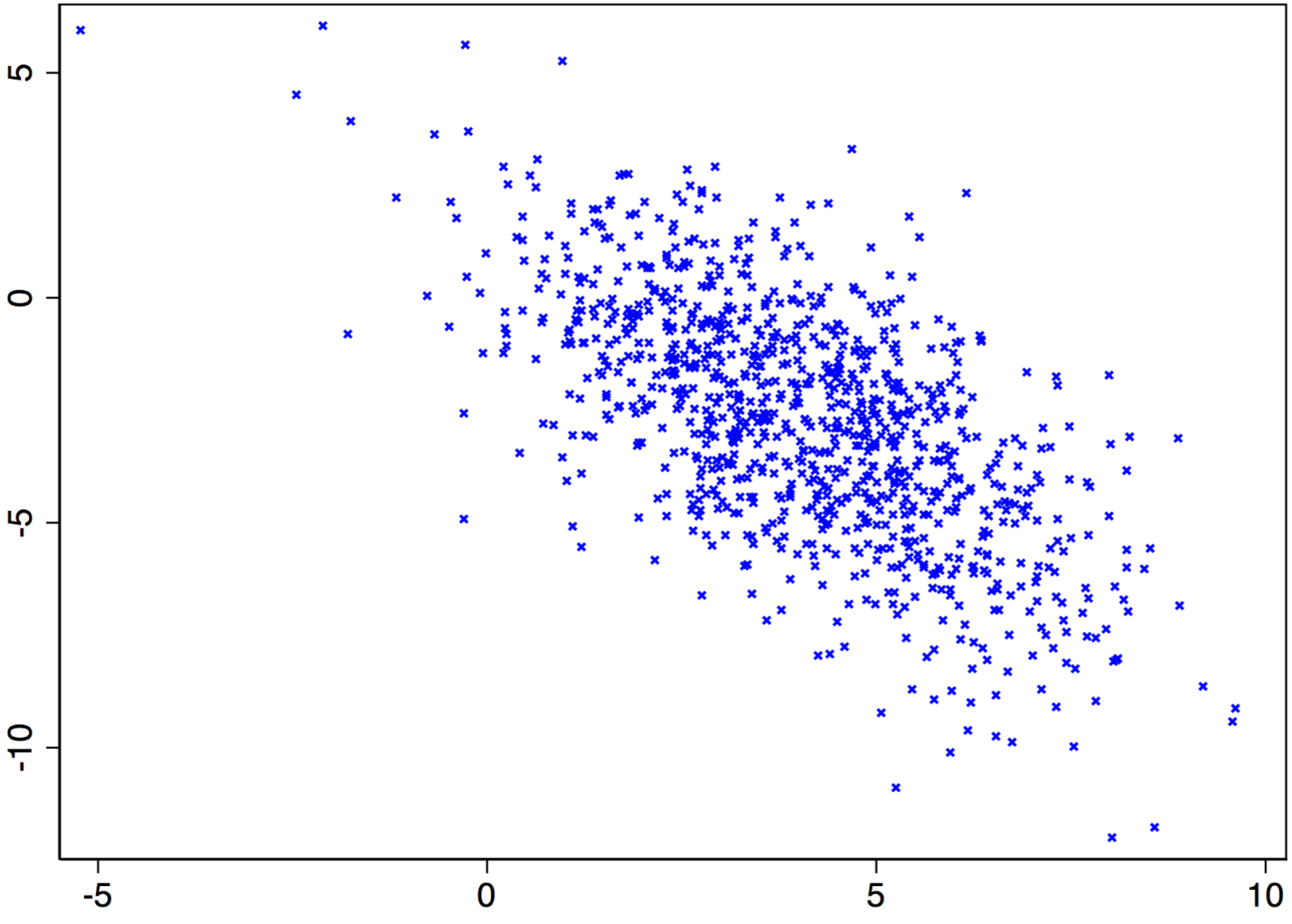
$$\rho = Corr(x, y) = \frac{Cov(x,y)}{Std(x) \cdot Std(y)}$$

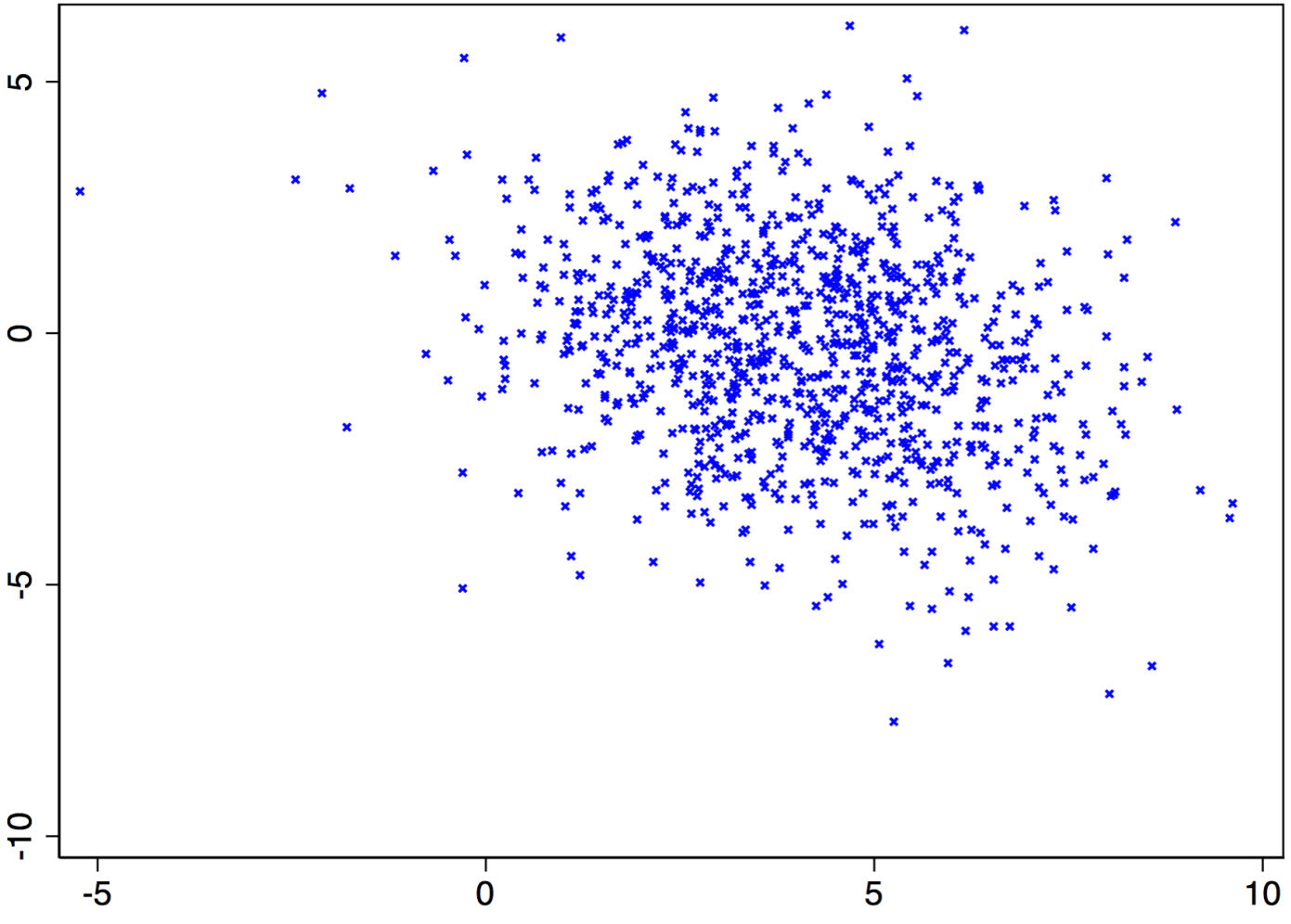
$$\rho \in [-1, 1]$$

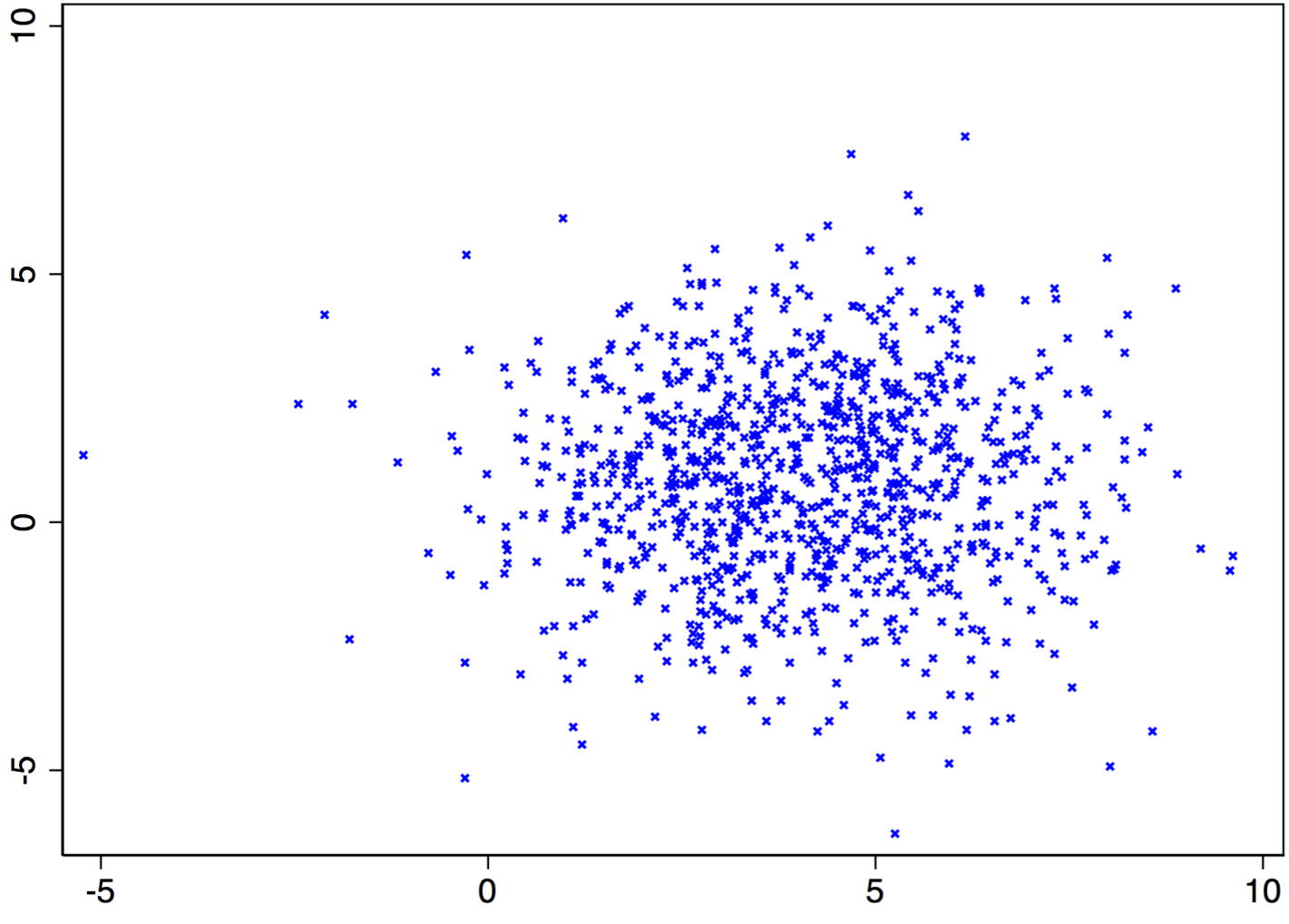
Scatterplot

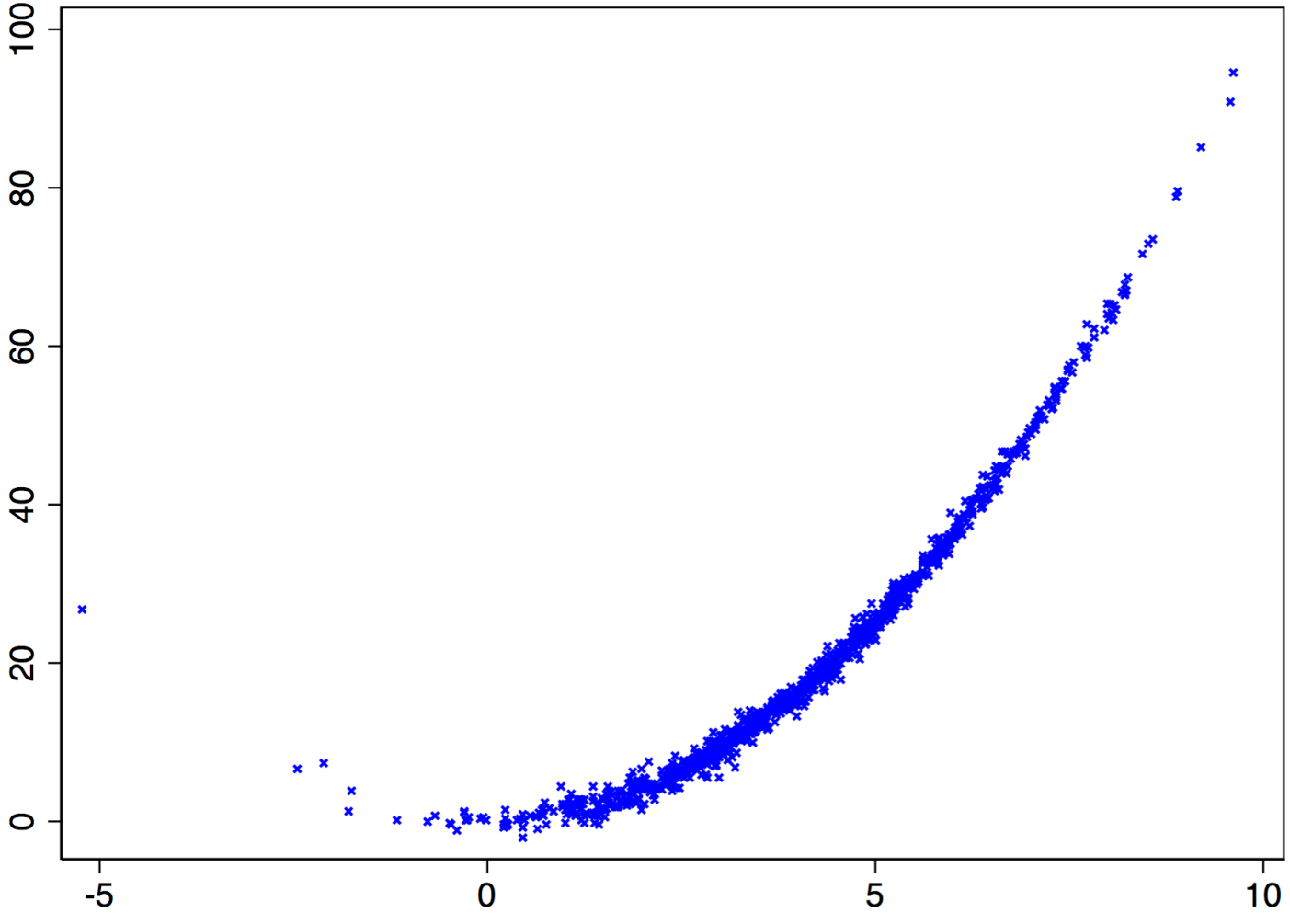
- Tells us if there is a relationship among the variables
- We investigate if there is a linear relationship, nonlinear, or no relationship


10
5
0
-5
-5
0
5
10


5
0
-5
-10
-5
0
5
10


5
0
-5
-10
-5
0
5
10

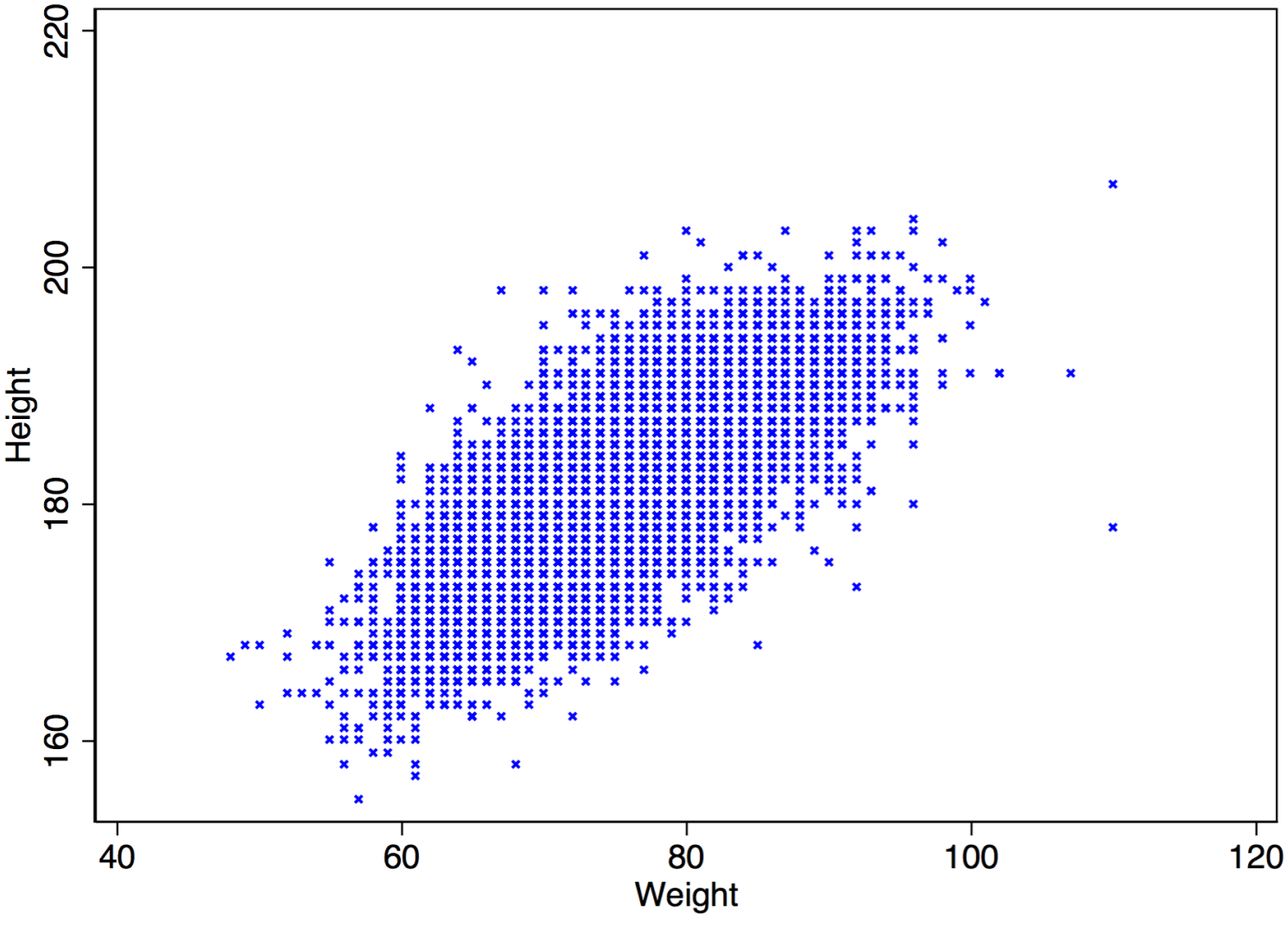

10
5
0
-5
-5
0
5
10

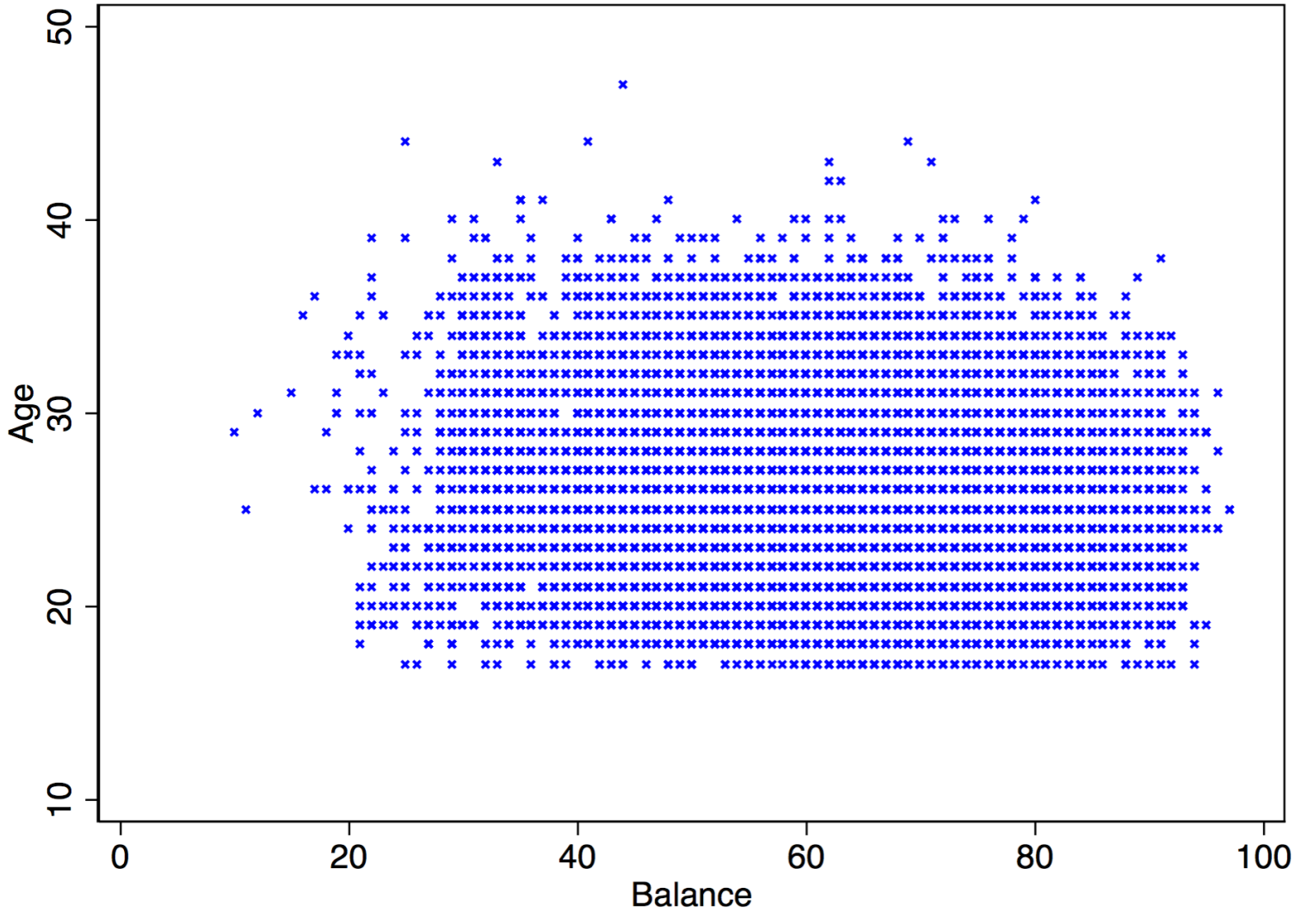

100
80
60
40
20
0
-5
0
5
10

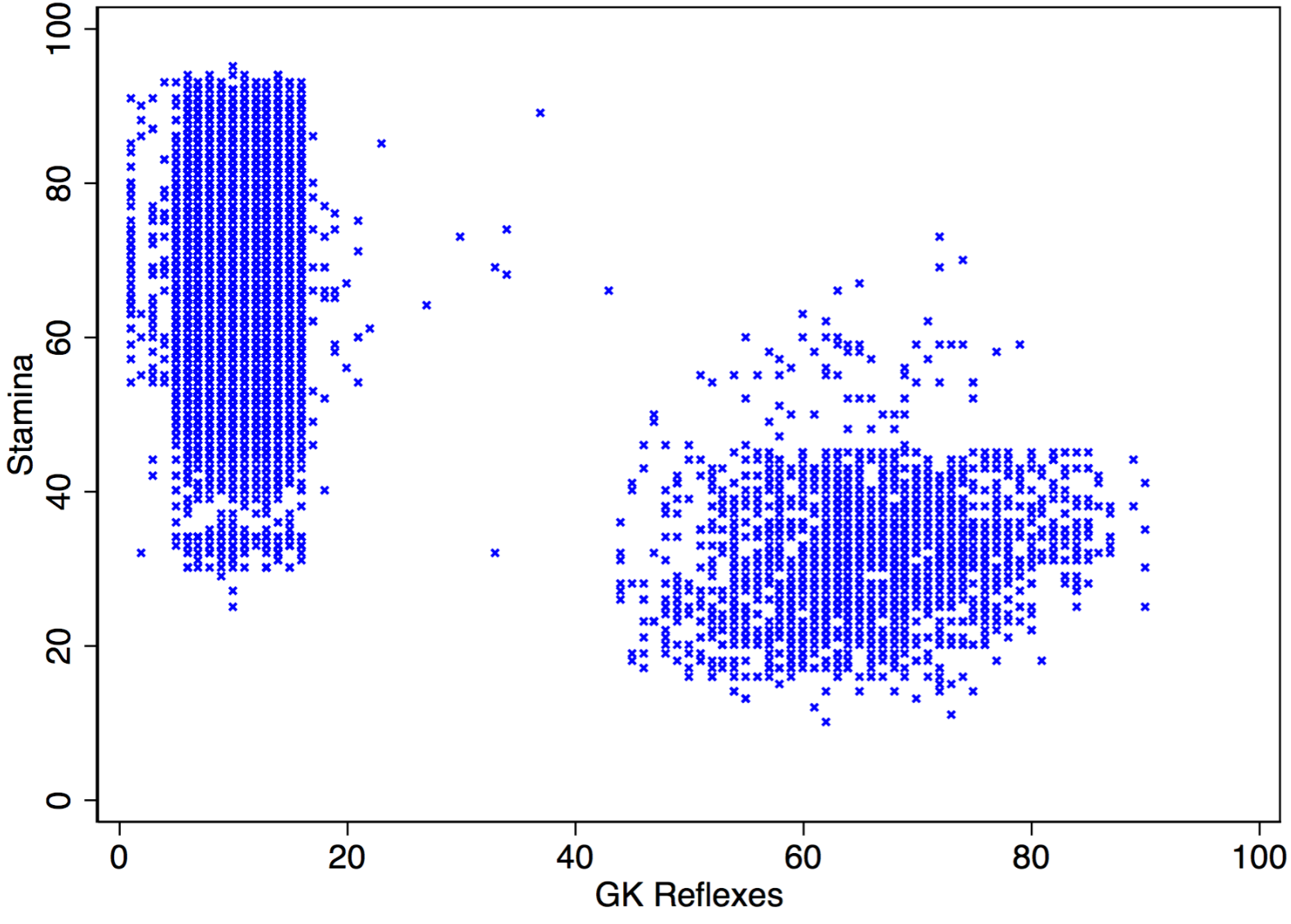
- Height and weight of football players on Fifa database

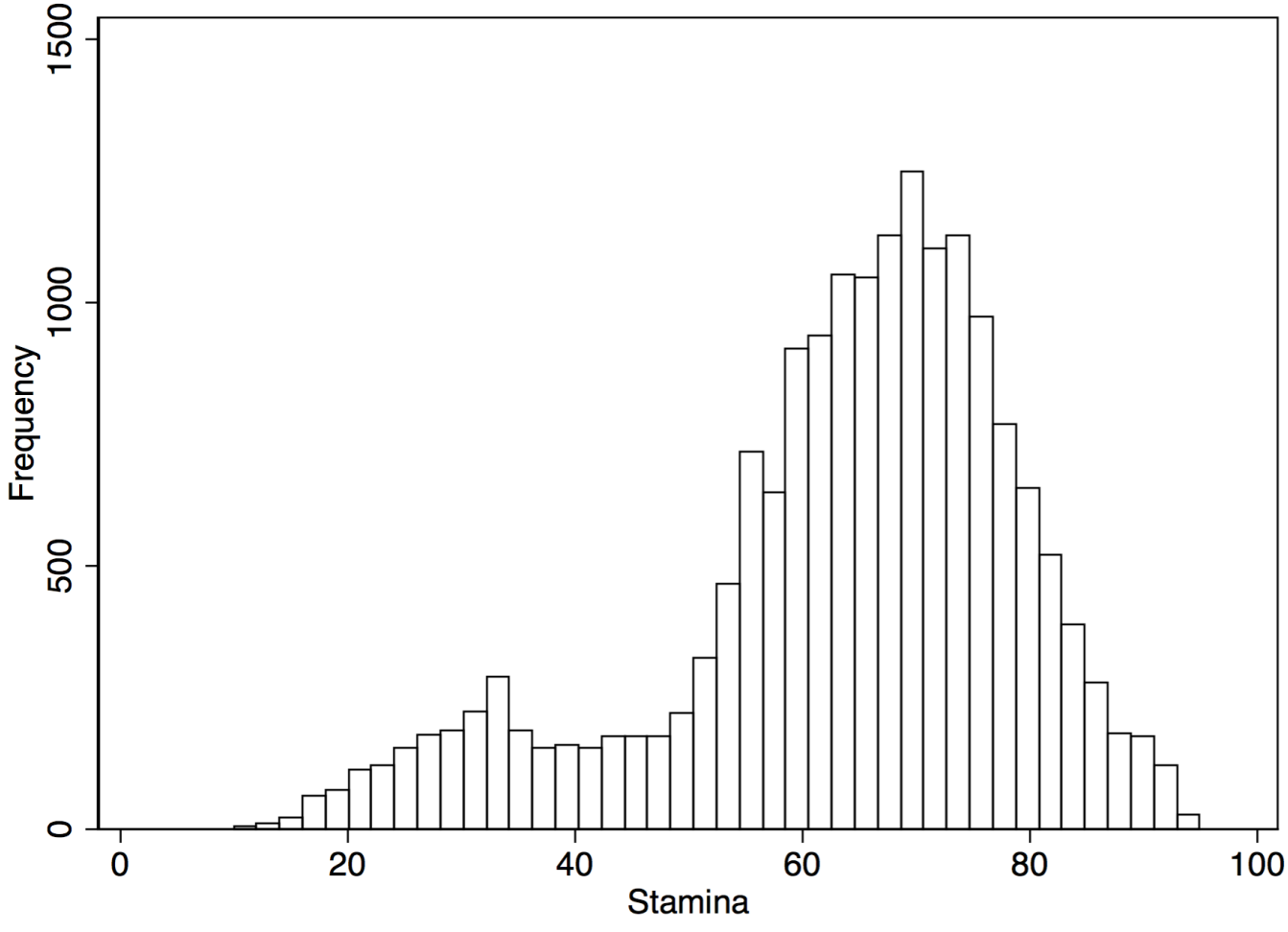
Downloaded from Kaggle

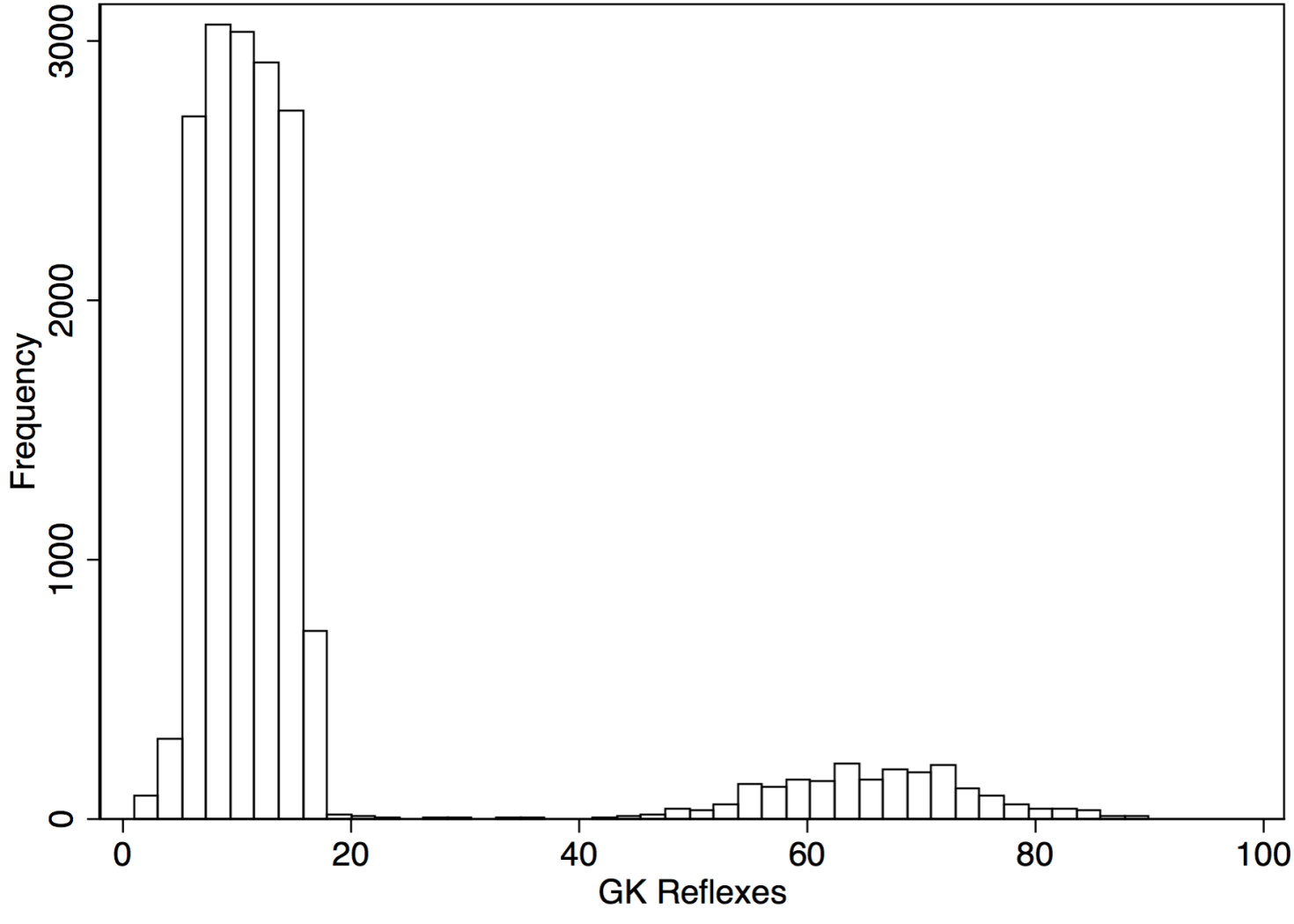
https://www.kaggle.com/hiteshp/exploring-fifa-2017-dataset/notebook

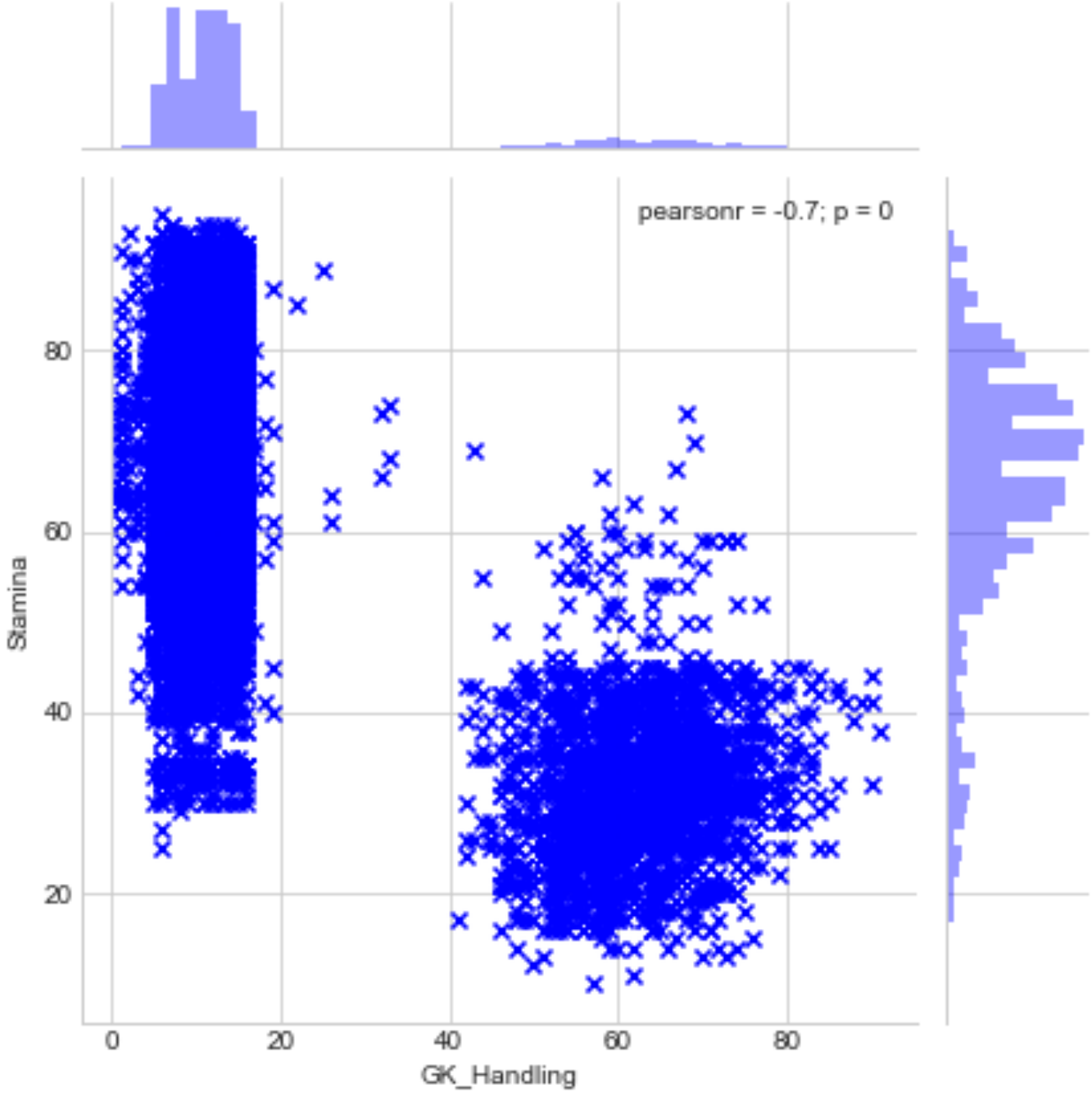

Height
220
200
180
160
40
60
80
100
120
Weight


Age
50
40
30
20
10
0
20
40
60
80
100
Balance


Stamina
GK Reflexes
0
20
40
60
80
100


Frequency
0
500
1000
1500
0
20
40
60
80
100
Stamina


Frequency
0
1000
2000
3000
0
20
40
60
80
100
GK Reflexes


pearsonr = -0.7; p = 0
Stamina
GK_Handling
0
20
40
60
80
20
40
60
80

Sampling

- Sampling theory is a study of relationships between a population and the draws from that population.
- There plenty of questions we want to ask when it comes to sampling such as:
  - Is the sample representative?
  - Are differences in statistics across the samples due to chance or due to measurement error (or any other type of error?)

Dangers of non representative sampling

- We cannot draw reliable statistics from the data
- Our findings are not generalizable

Order of random sampling

- We want the sample we are dealing with to be representative of the population.
- In order to do this, we should sample randomly from the populations.
  - Examples of sampling gone wrong?

Extreme values (Outliers)

- In statistics, an **outlier** is an observation point that is distant from other observations
- There are several reasons for observing extreme values in the data as well as several types
- You deal with them if you know your context and data very well

Examples?