

Data Analysis 2: Fundamentals of Statistics

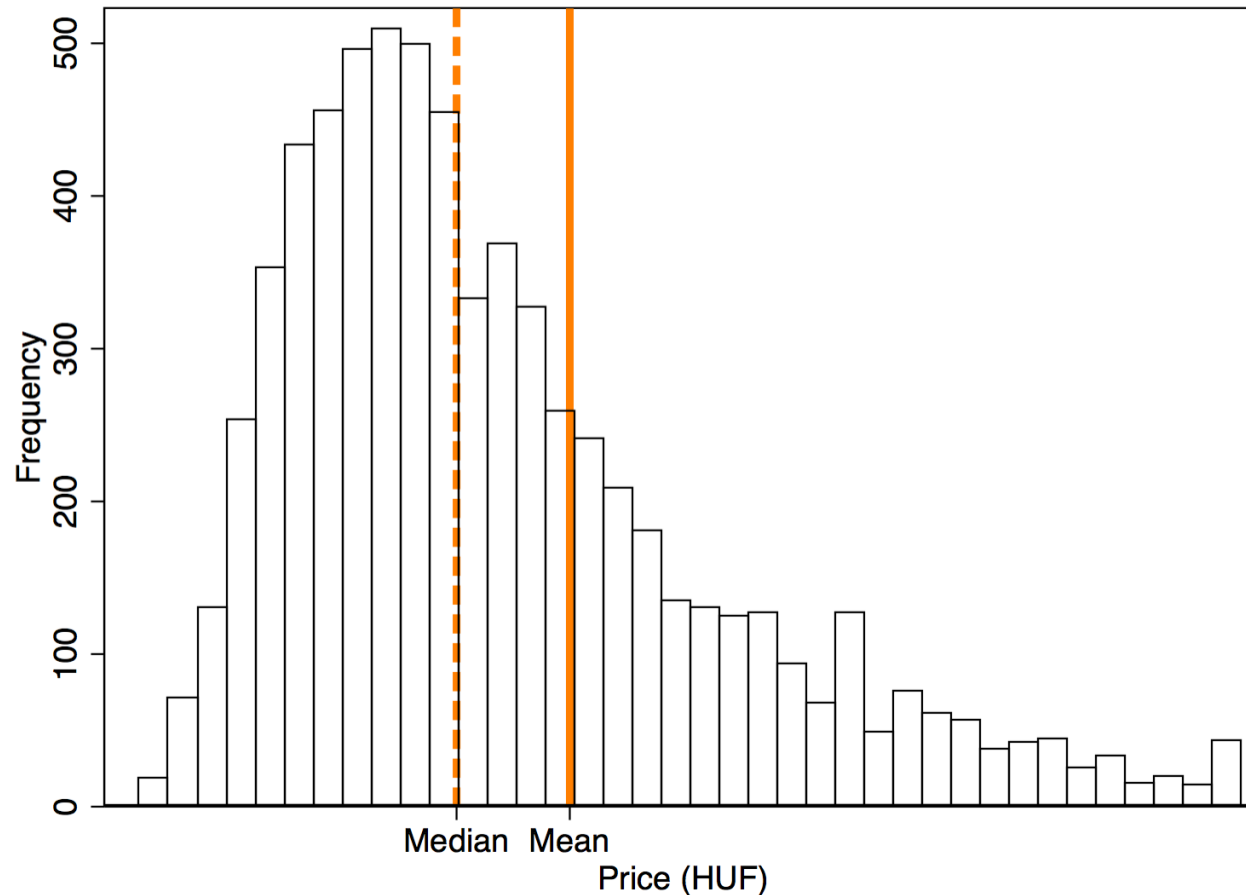
Instructor: Arieda Muço, Fall 2017

Types of Random Variables

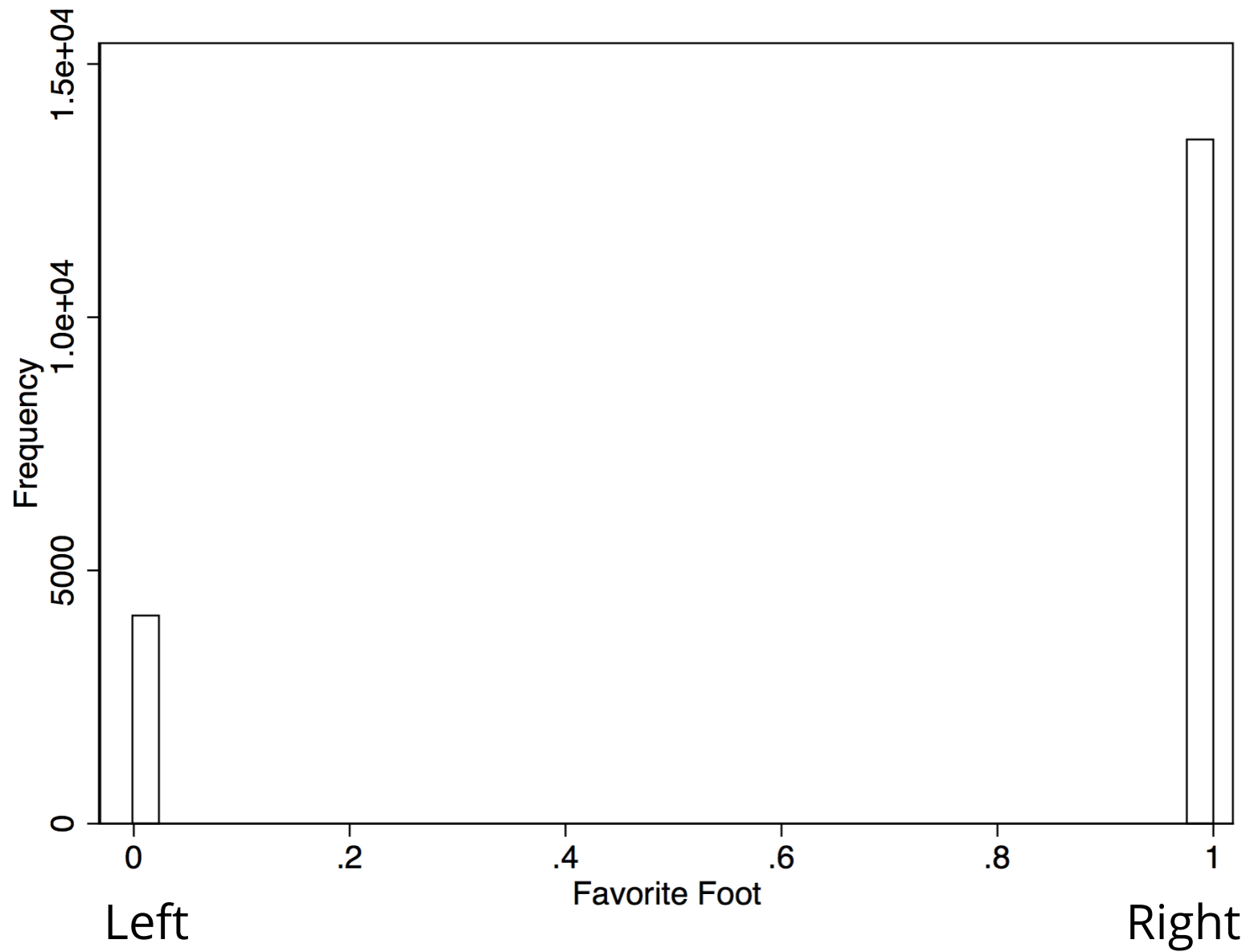
- Continuous, takes values in any interval
 - i.e prices, temperature, grades...
- Discrete, takes no more than a countable number of values
 - i.e hotel stars, gender, number of rooms

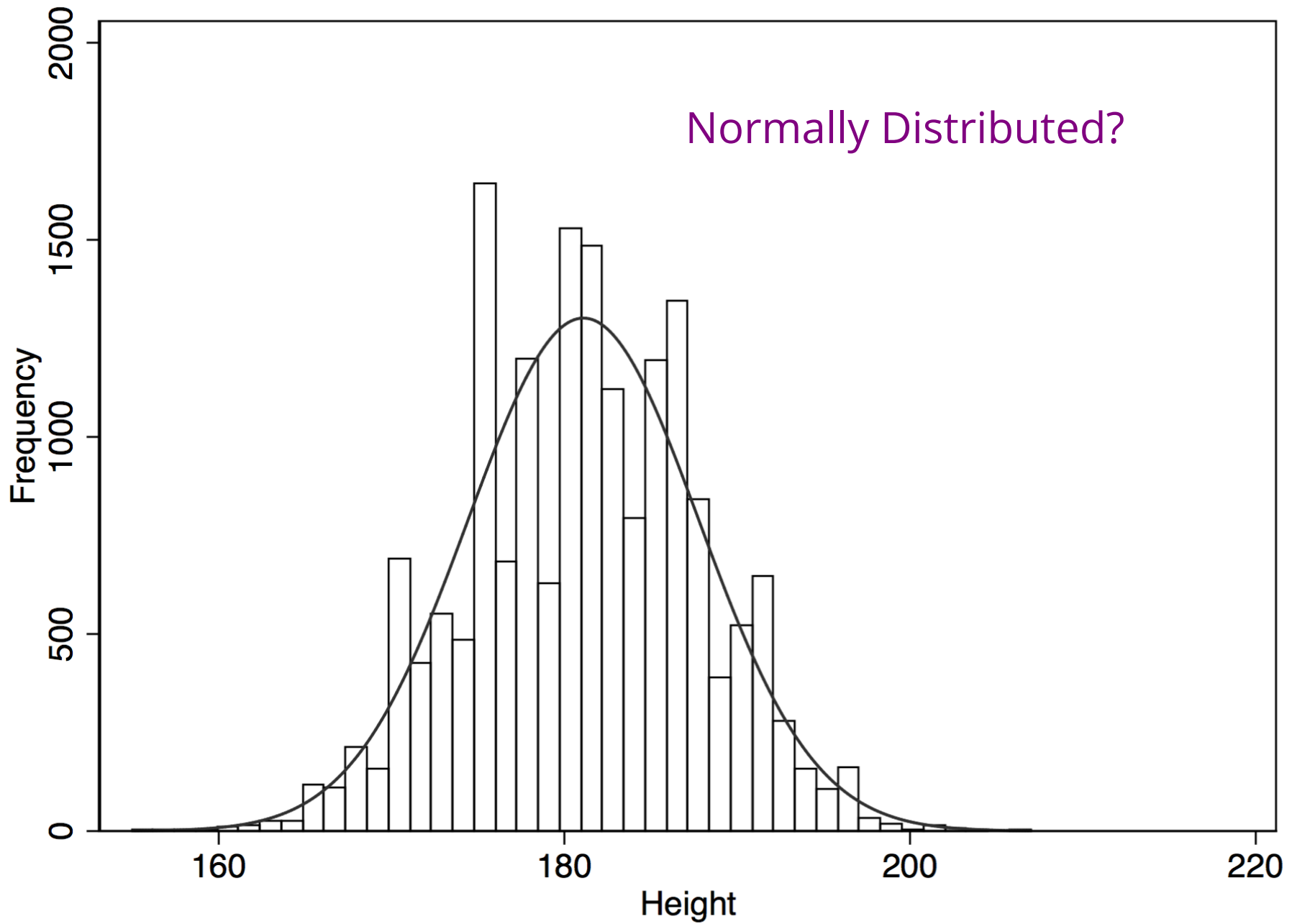
Each have their own distributions

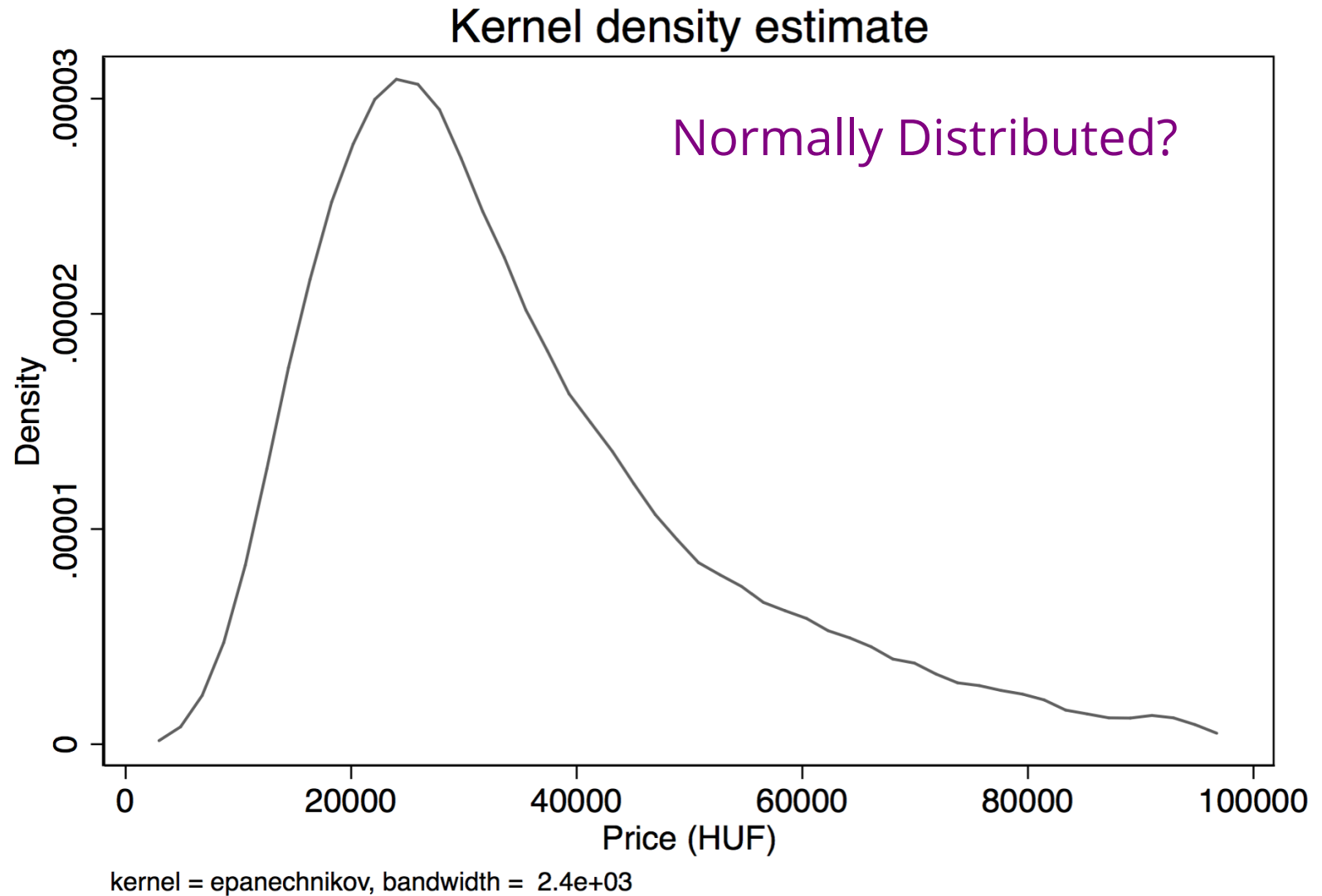
Histograms in practice

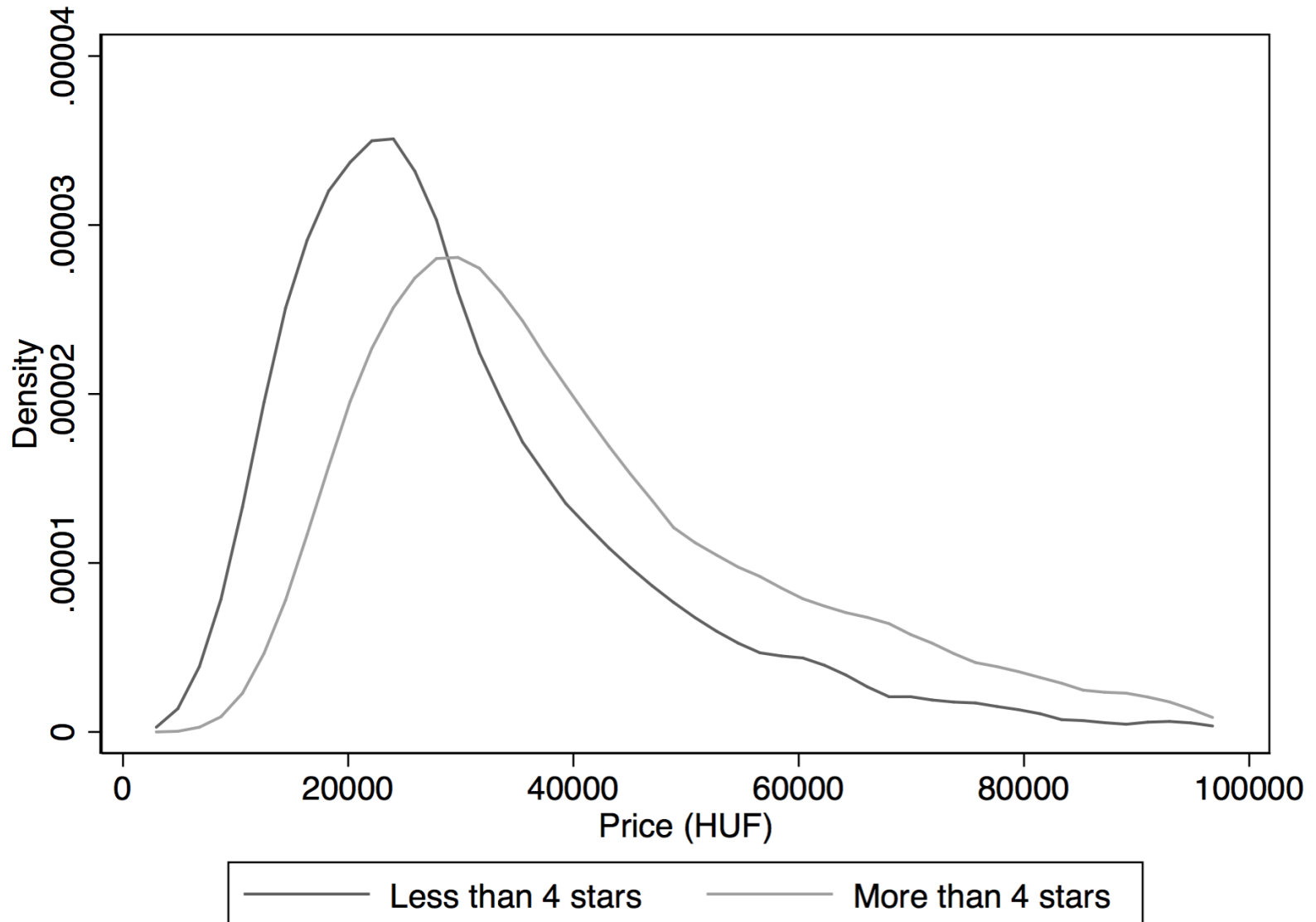


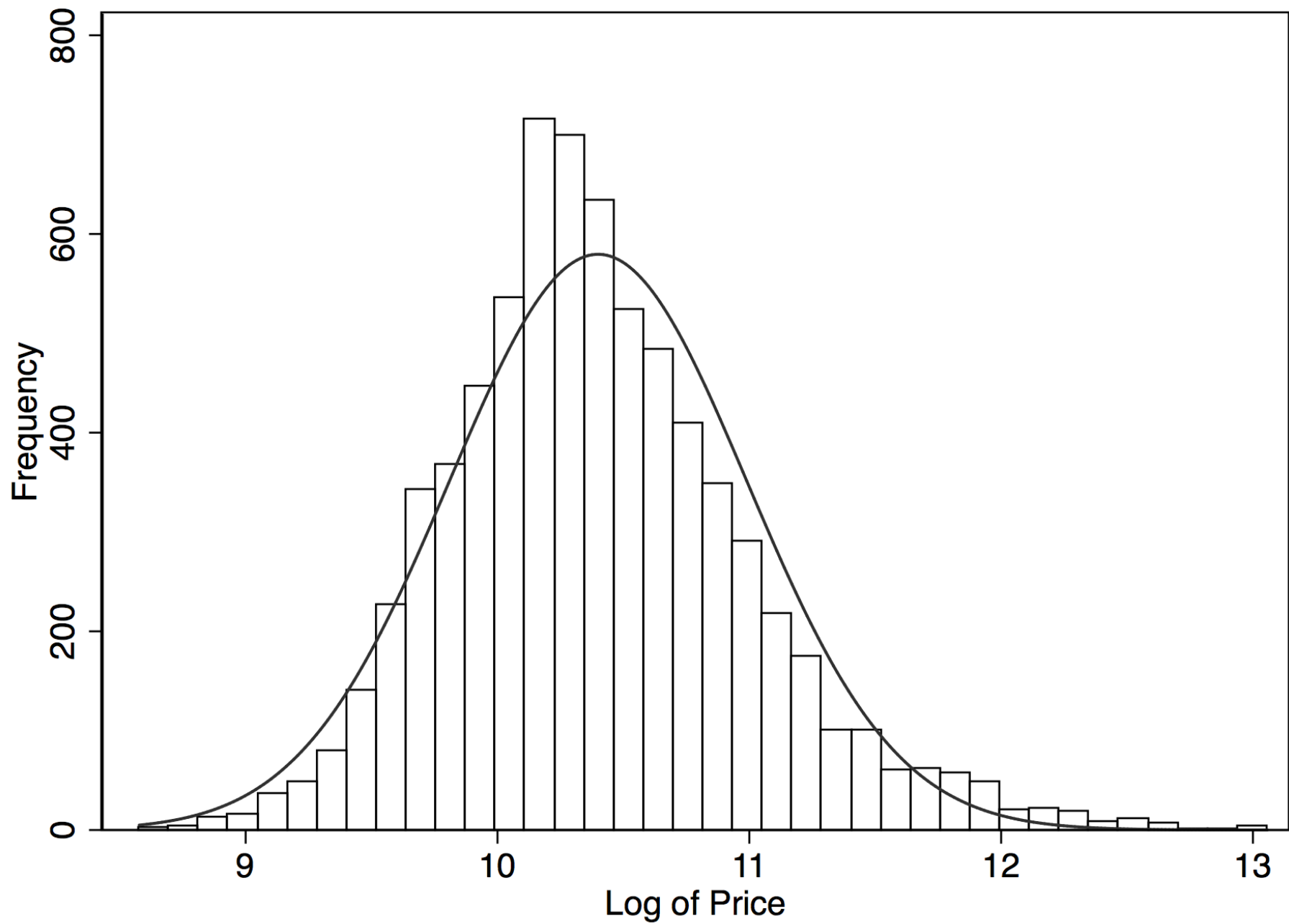
**If I tell you that football players have a
favorite foot?**

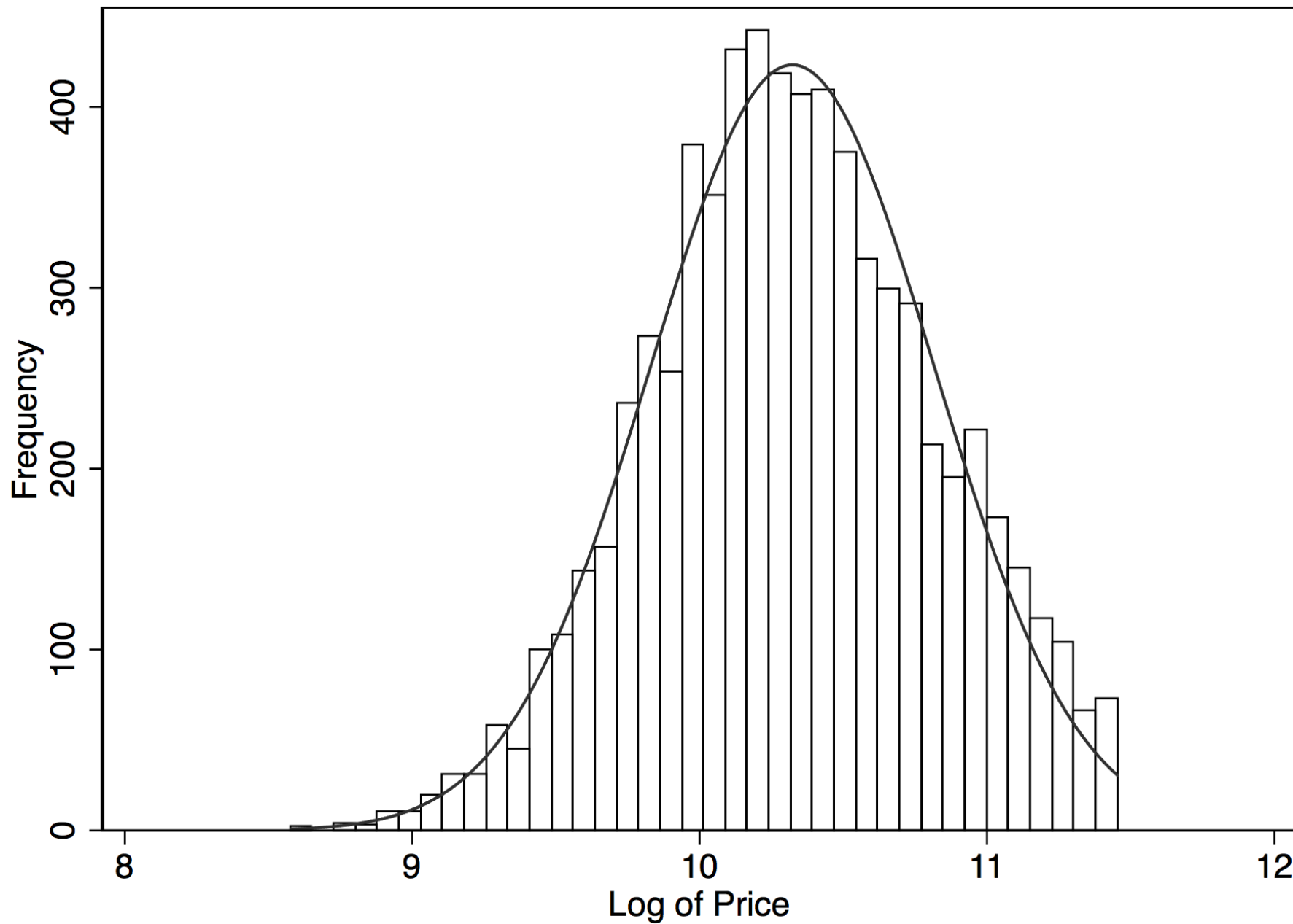












Covariance

- It provides an indication of the dependence between two variables

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

Covariance

- In case: $y_i = a + bx_i$

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

Correlation

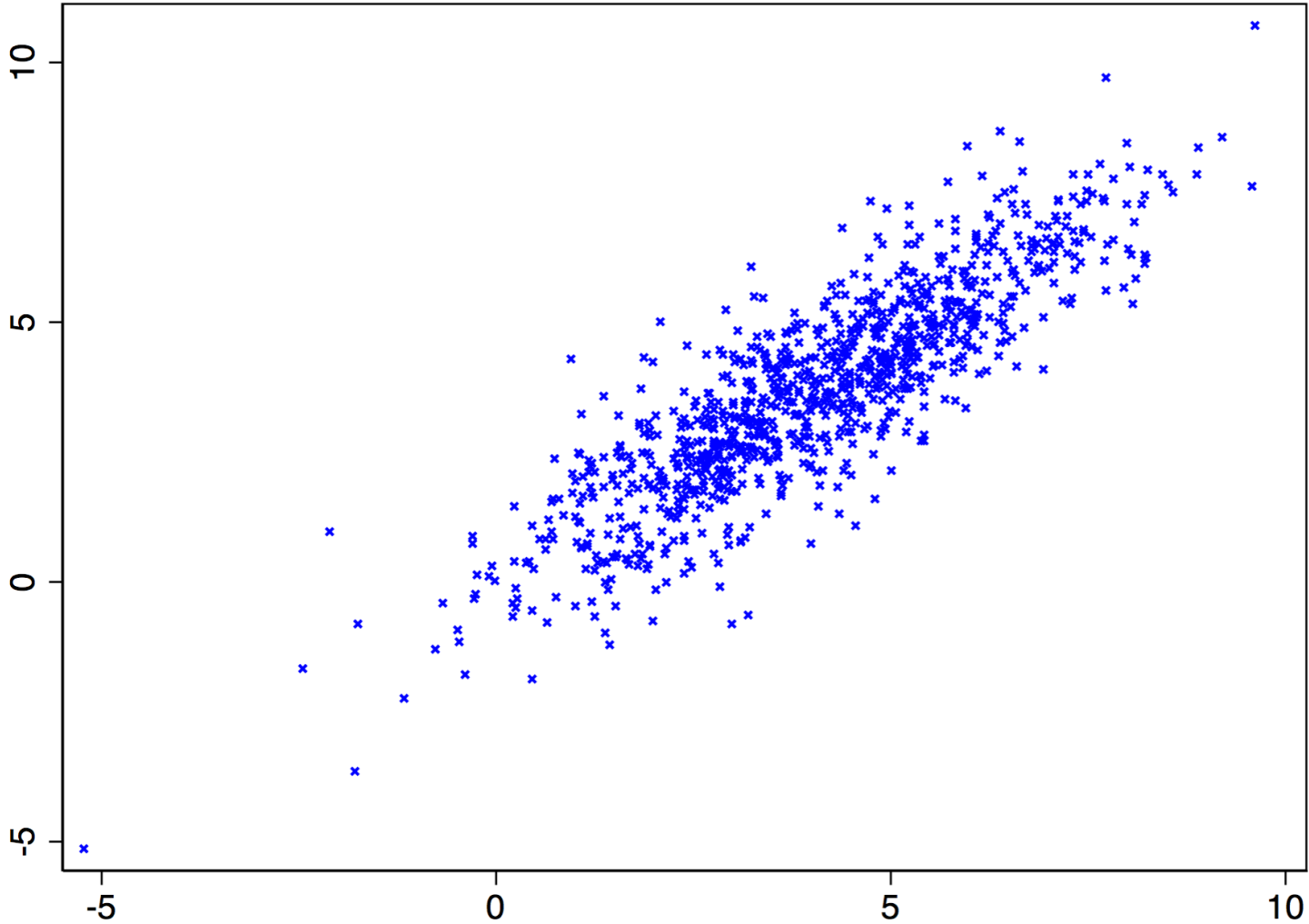
- Is computed by dividing the covariance by the standard deviation of each variable

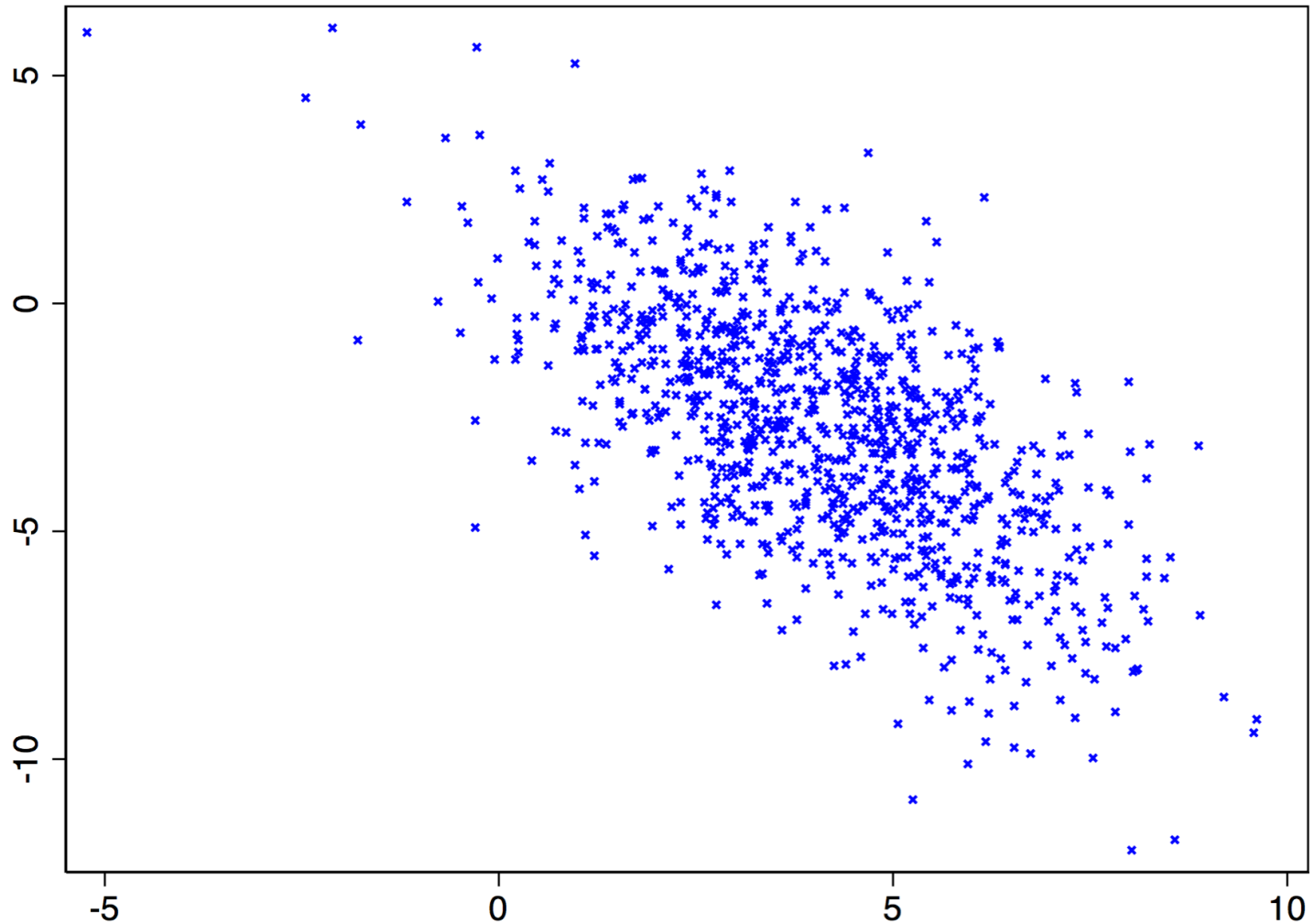
$$\rho = \text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\text{Std}(x) \cdot \text{Std}(y)}$$

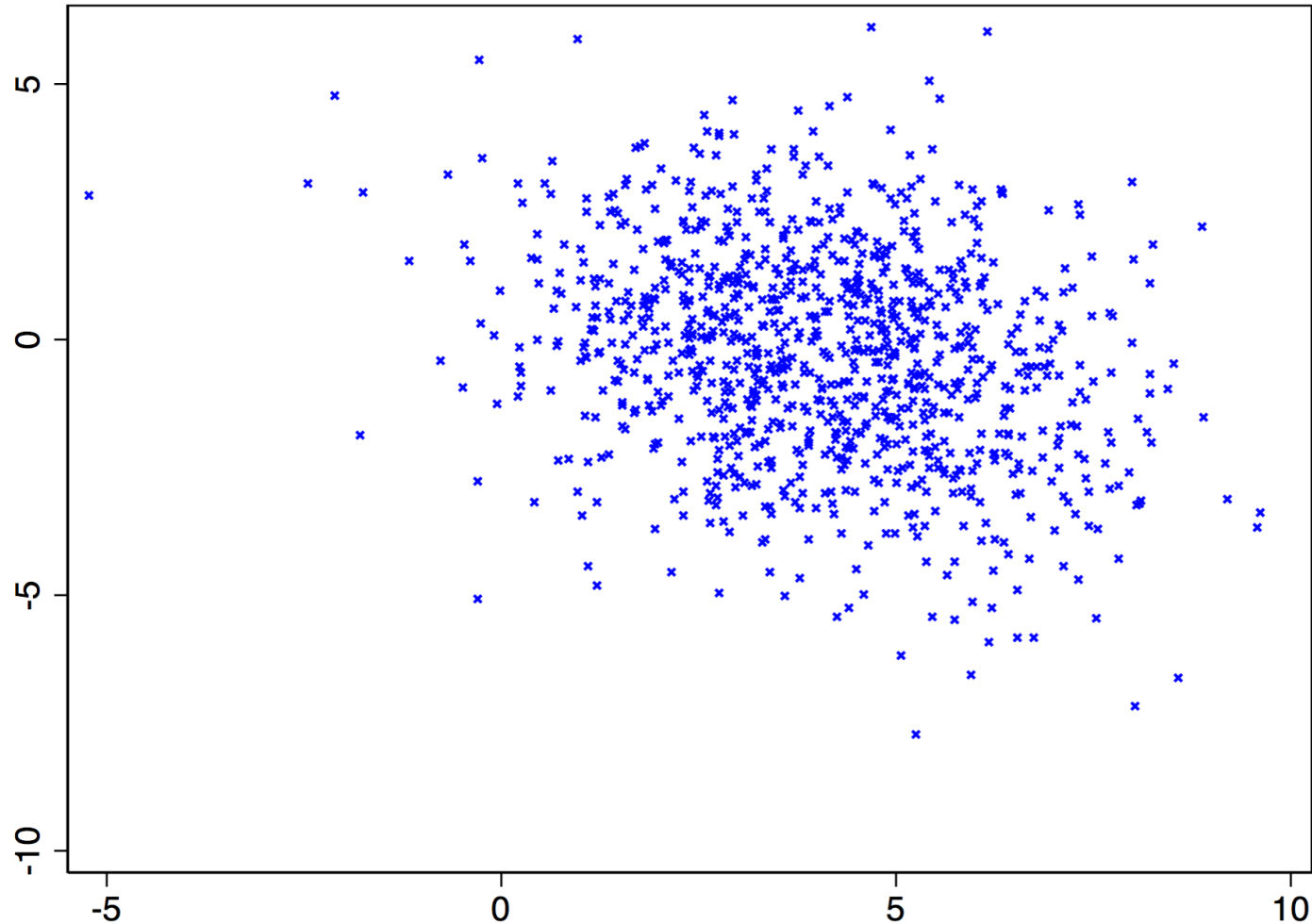
$$\rho \in [-1, 1]$$

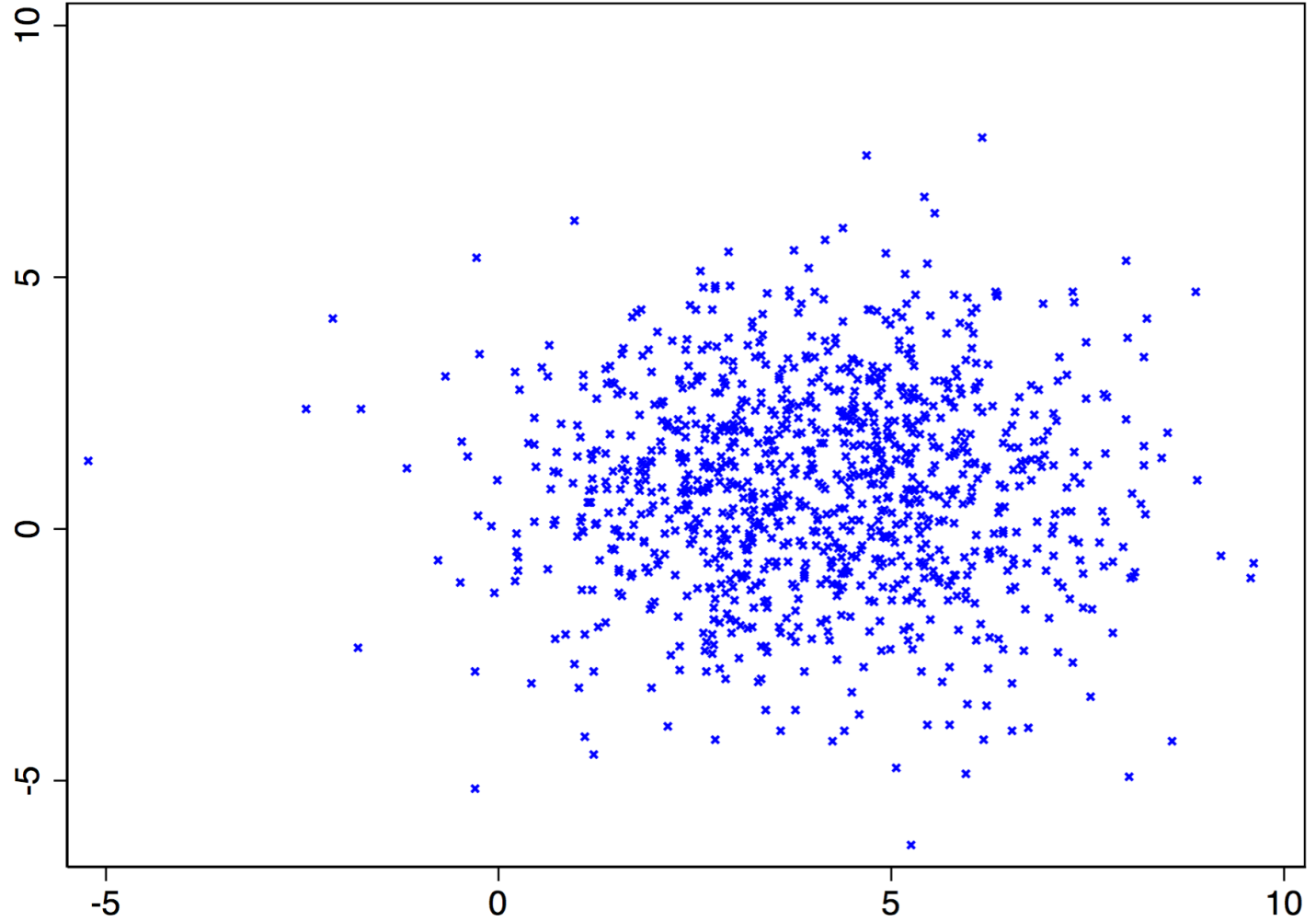
Scatterplot

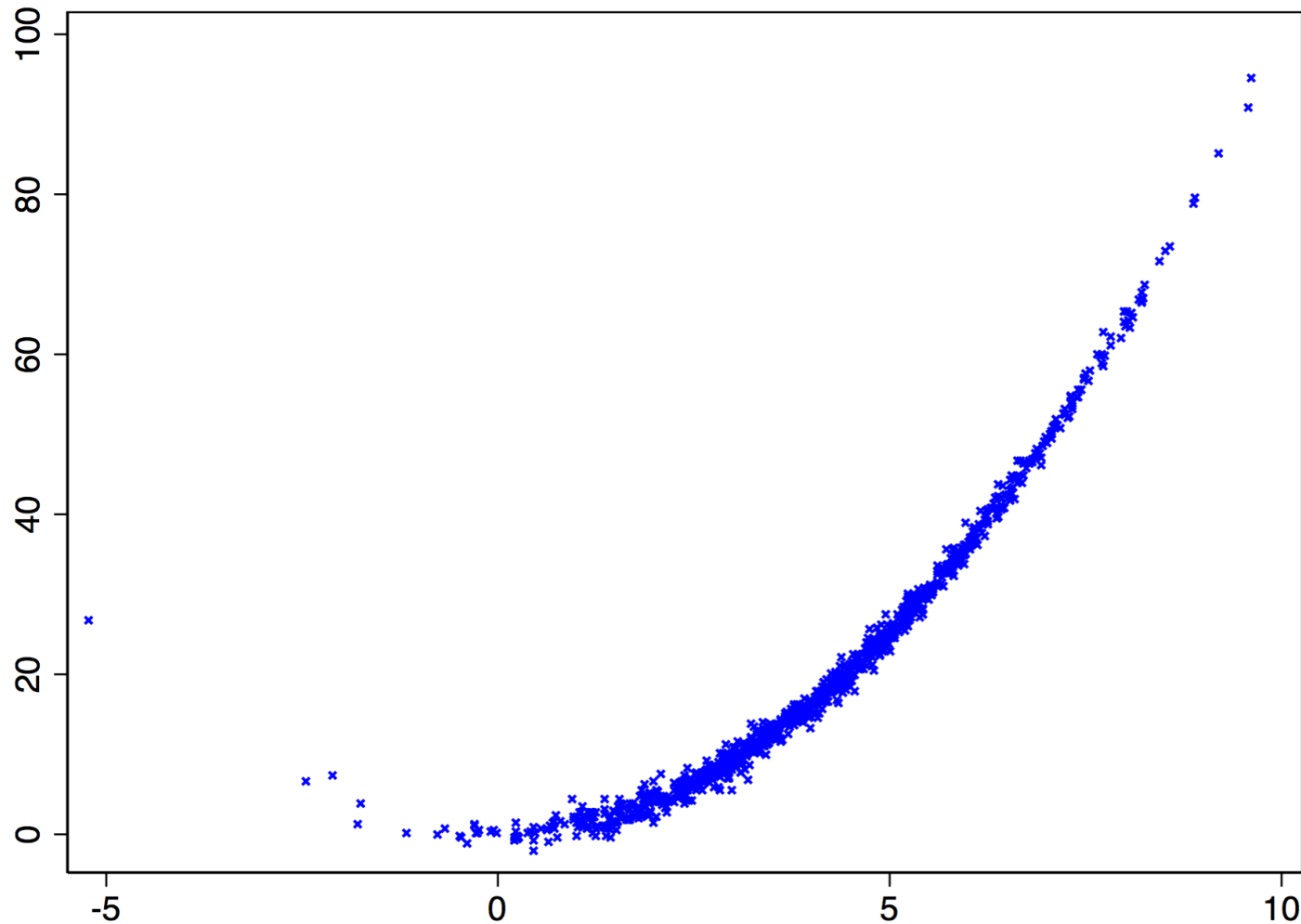
- Tells us if there is a relationship among the variables
- We investigate if there is a linear relationship, nonlinear, or no relationship







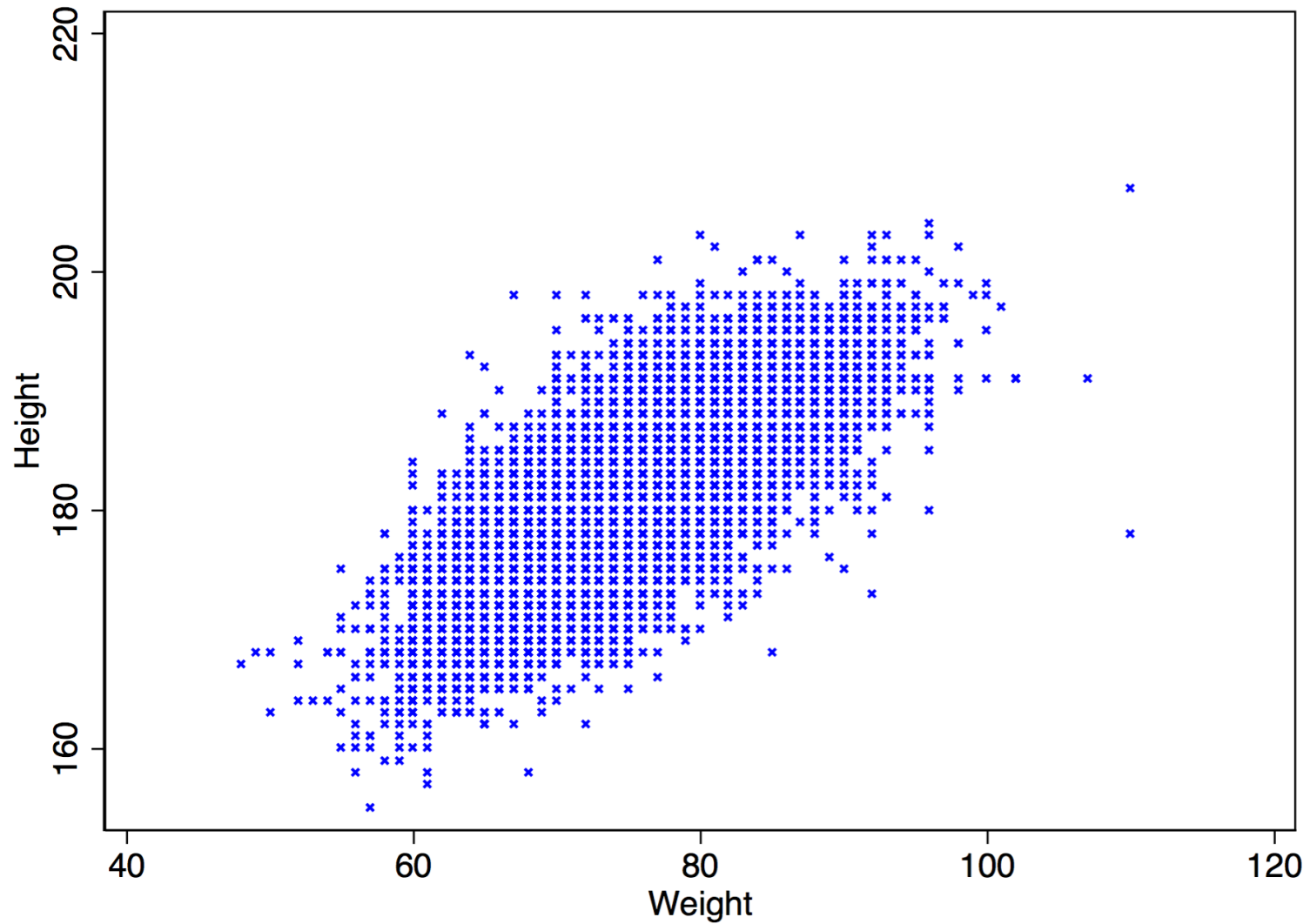


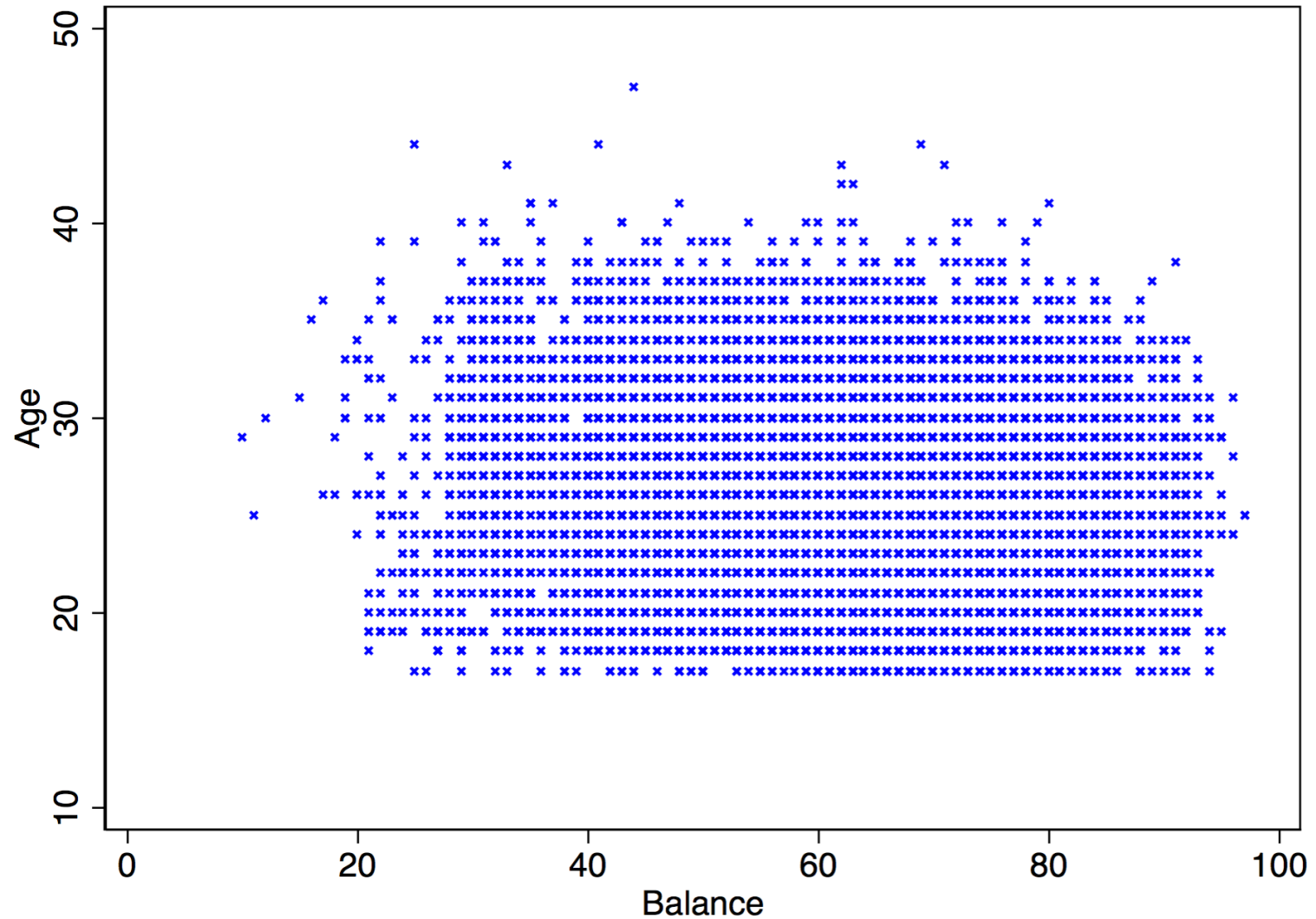


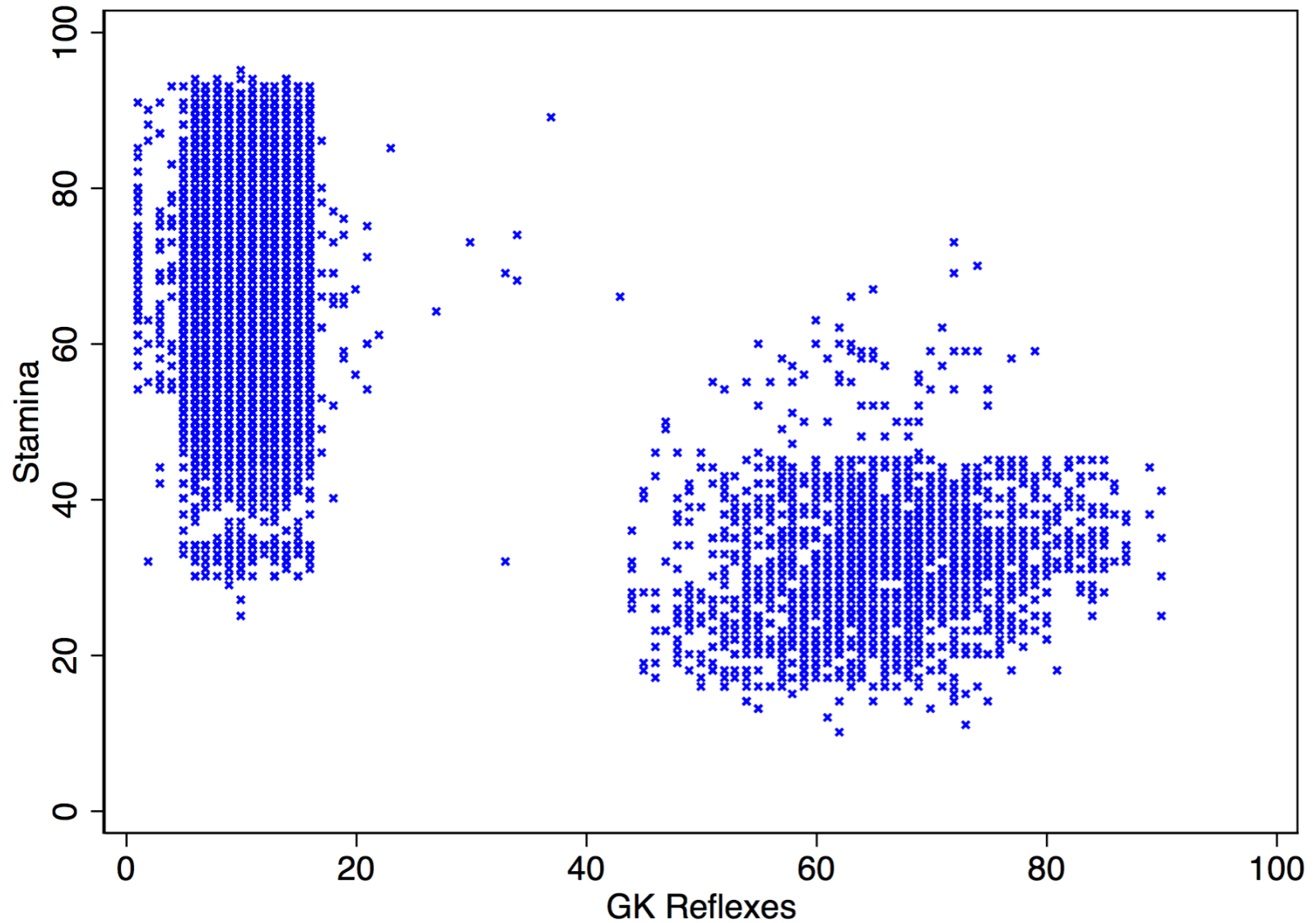
- Height and weight of football players on Fifa database

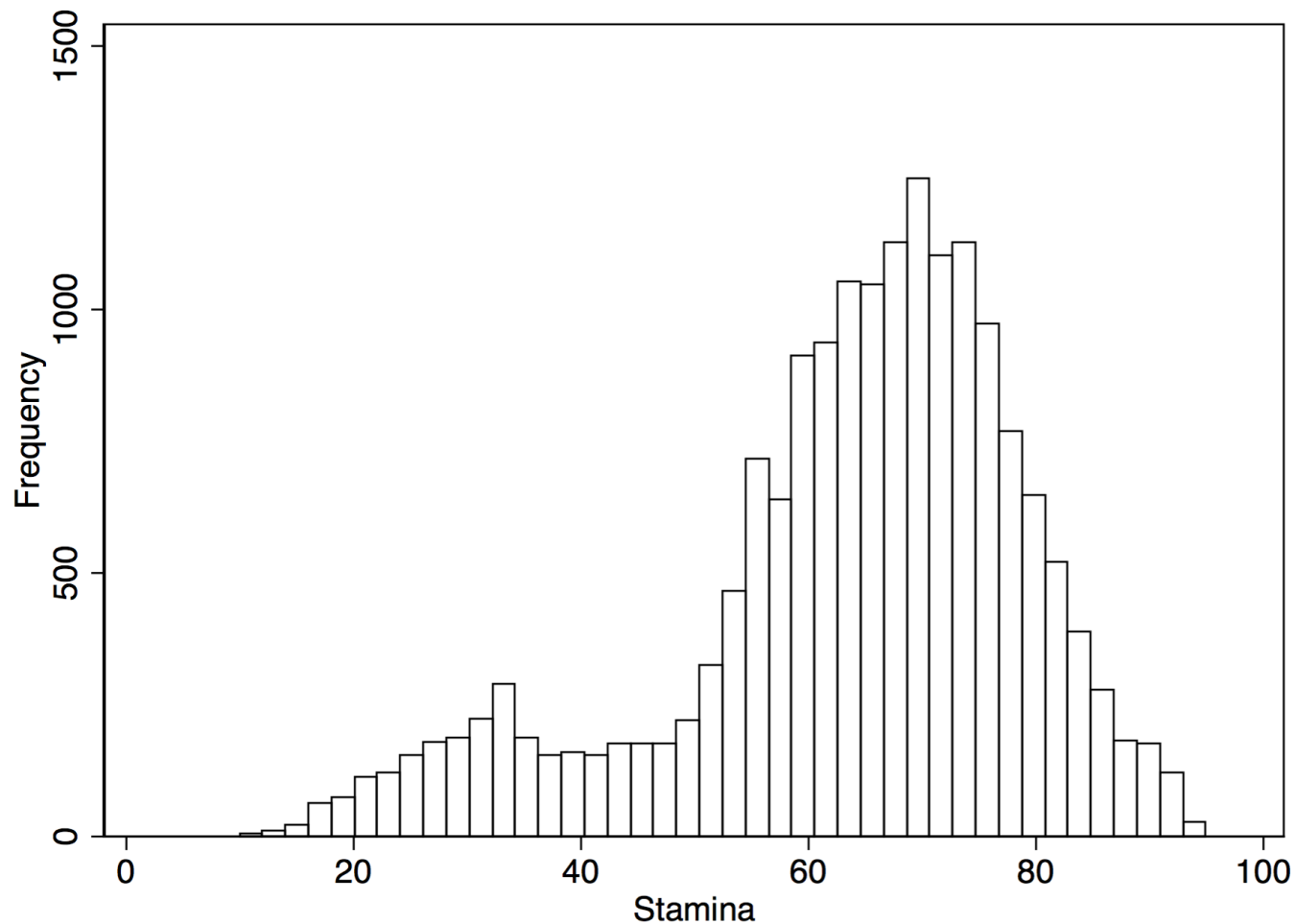
Downloaded from Kaggle

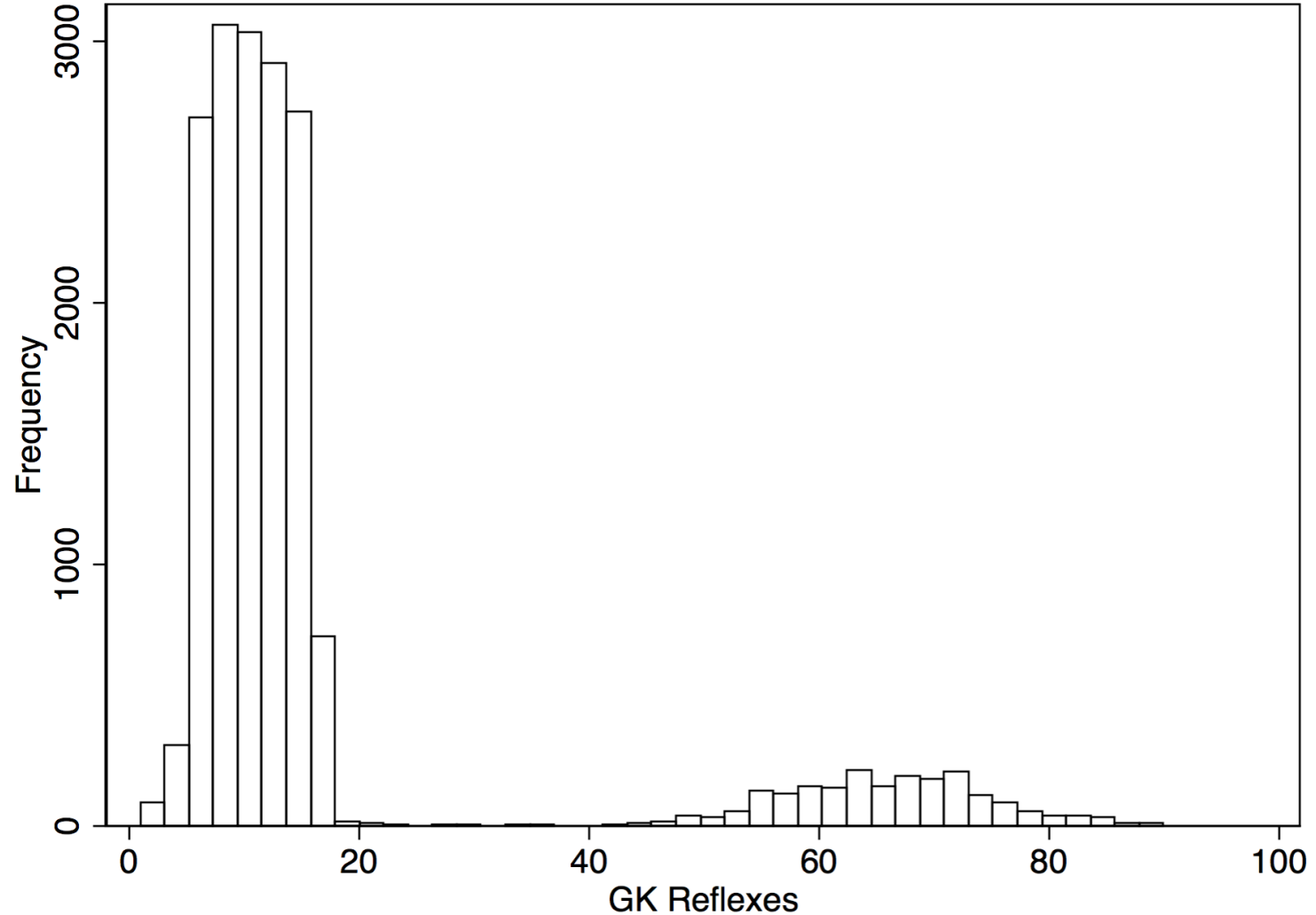
<https://www.kaggle.com/hiteshp/exploring-fifa-2017-dataset/notebook>

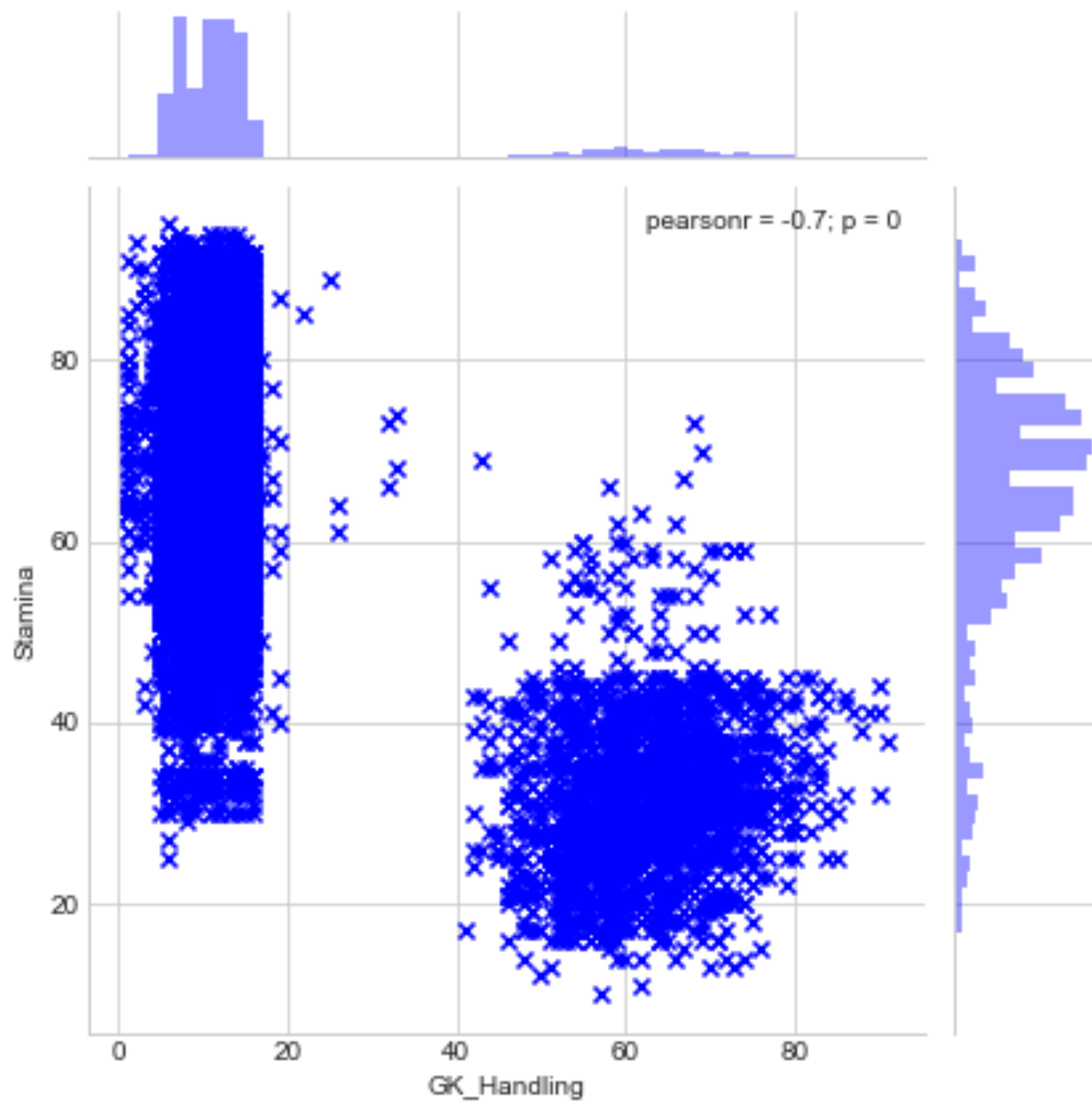












Sampling

- 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?