Data Analysis 2: Foundations of Statistics

Instructor: Arieda Muço, Fall 2017

Types of Random Variables

- Continuous, takes values in any interval
    - i.e prices, temperature, grades...
- Discrete, countable number of values
    - i.e hotel stars, gender, number of rooms...

Each have their own distributions

Distributions in Practice

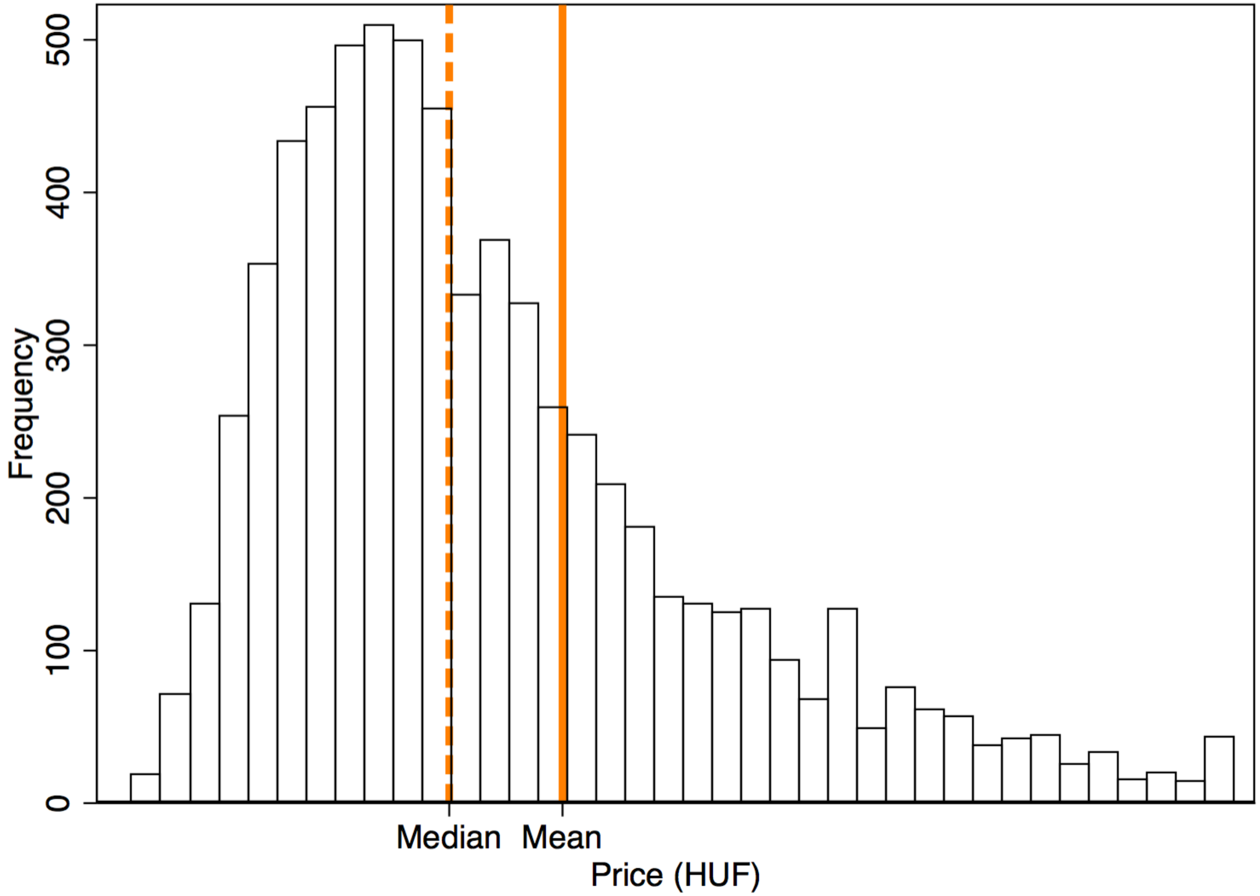
- The distribution of a variable tells the number of times each possible value of the variable occurs in the data
  - Can be expressed as frequency, percentage

- It does so in isolation from other variables.
  - It does not tell if certain values are more likely to occur when some other variable, or variables, take certain values

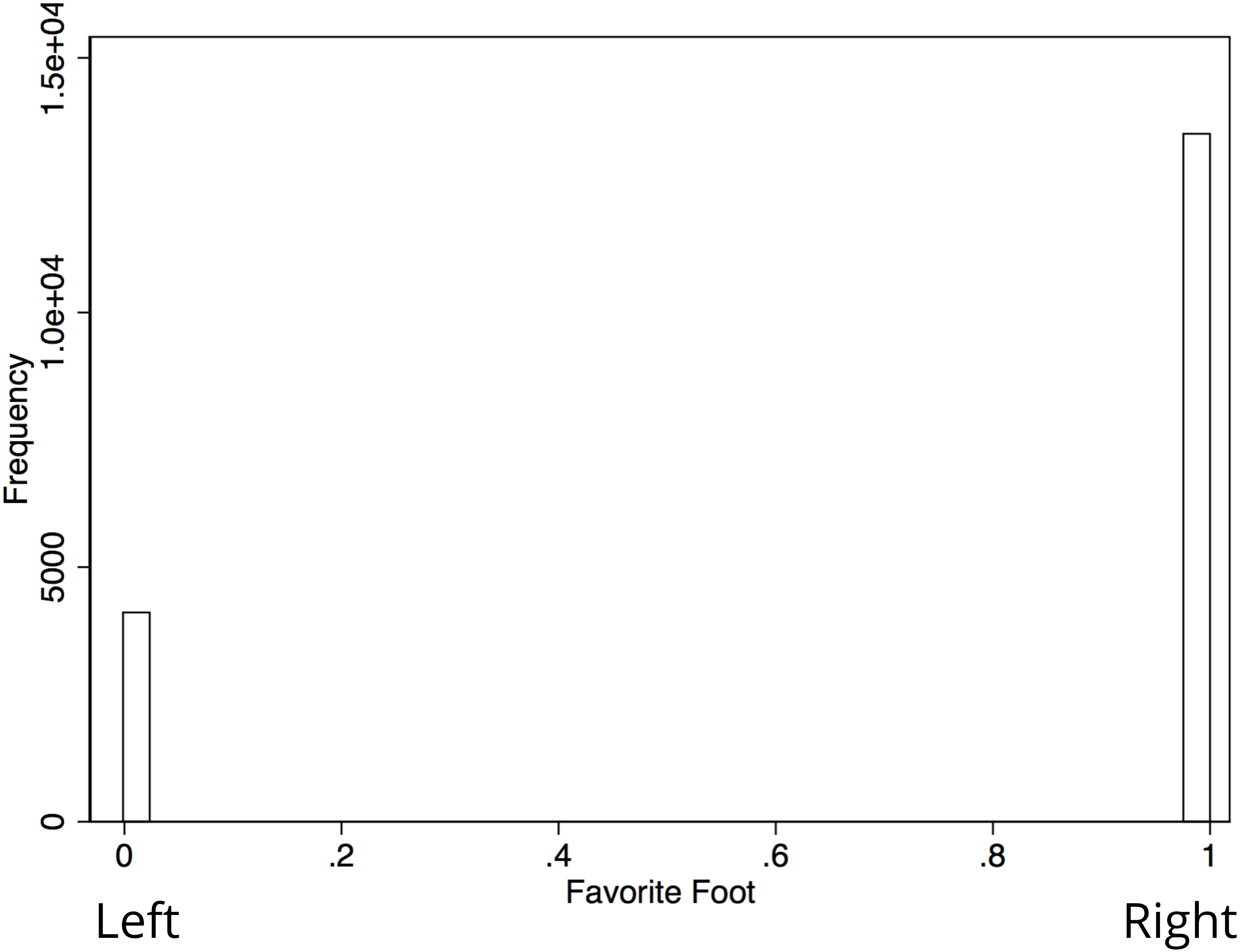
- The simplest way to visualize a distribution is through an histogram
- The histogram may takes on as many bars as the number of possible values
- From visual inspection of histograms we find out many interesting properties
  - the peaks, the immediate neighbourhood, if they have tails etc

Let's go back to our hotel price example.
What do we see?

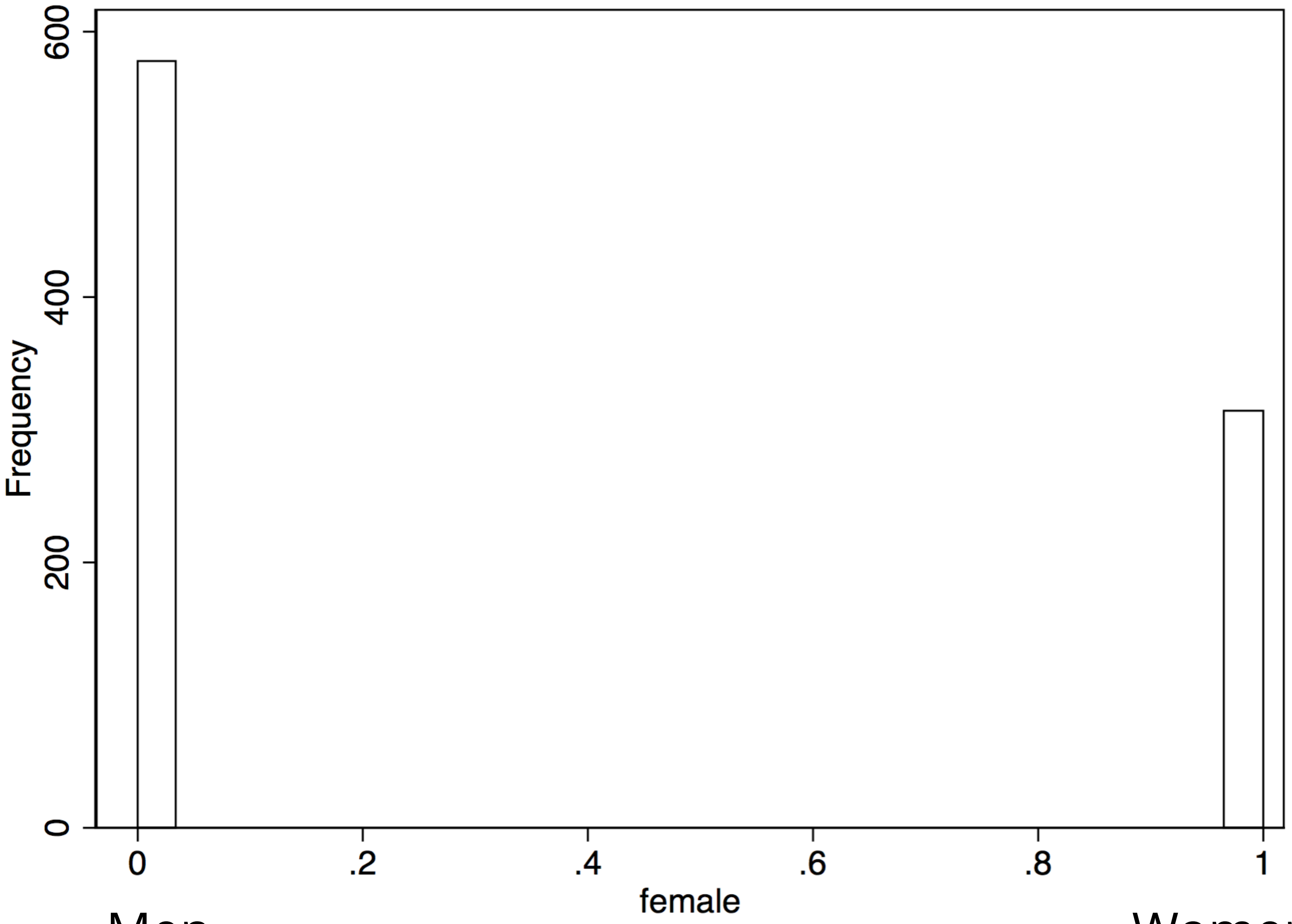
Histograms in practice



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

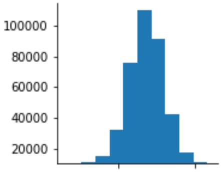

Frequency
1.5e+04
1.0e+04
5000
0
0
.2
.4
.6
.8
1
Favorite Foot
Left
Right

Titanic passengers data

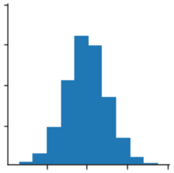


Men Women

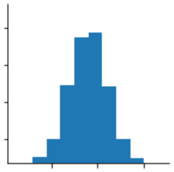
- Some of the properties of the distributions change if we change their bin size
    - Very wide bins may lump together multiple modes.
    - The statistical softwares you use, will compute the histogram with the default bin size
        - Try to figure out on your own the bin size that Stata or R use for their histograms.
        - Play with the simulations of distributions I provided you with by changing mean, variance, number of observarions, and binsize
        - Is this also the case for discrete variables?



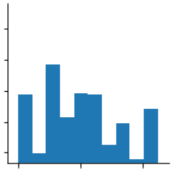
Average Area Income



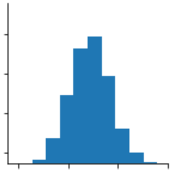
Average Area House age



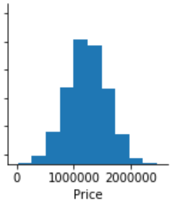
Number of rooms



Number of bedrooms



Area Population

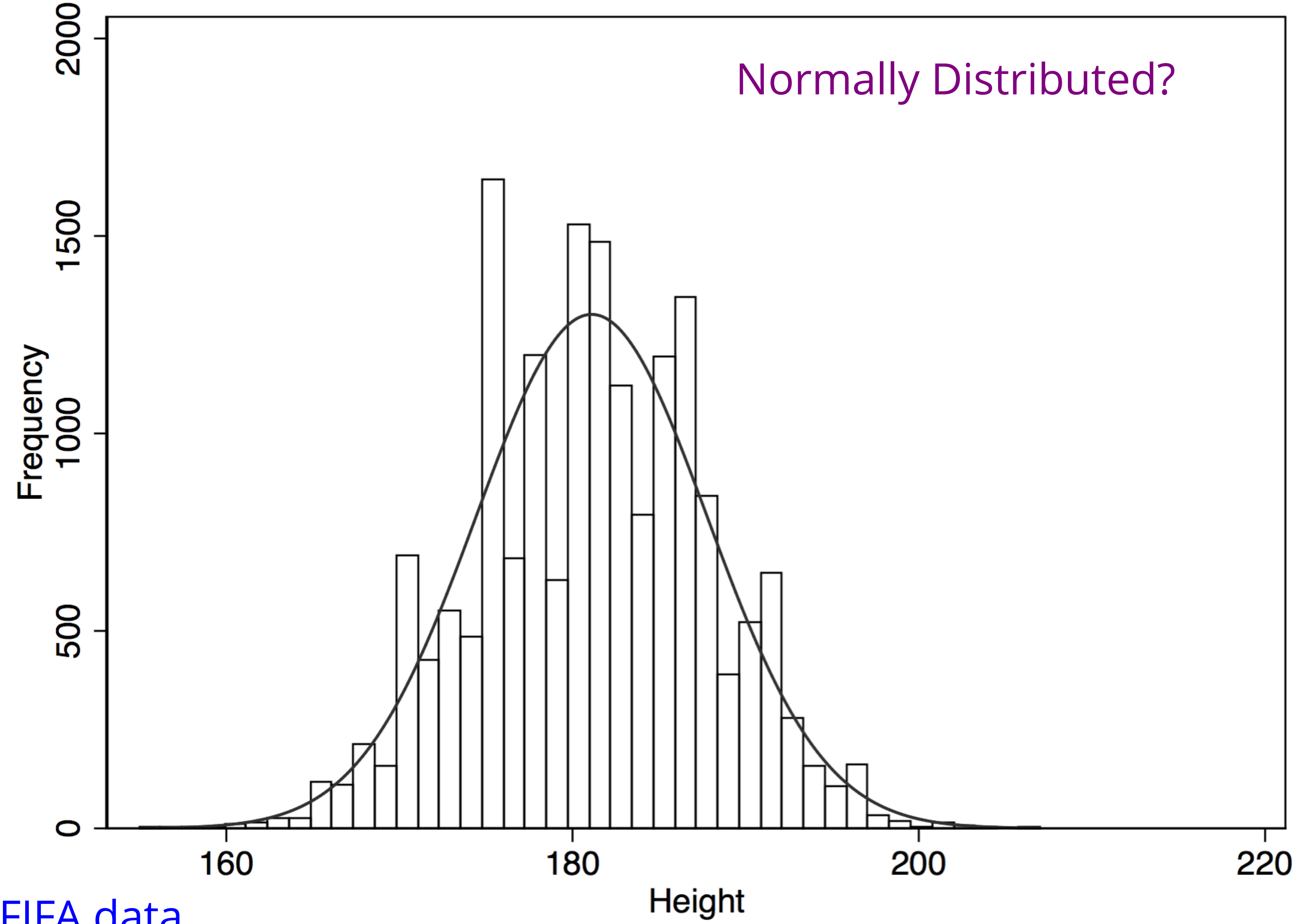


Price

- Kernel densities are an alternative way to histograms
    - for variables with many potential values

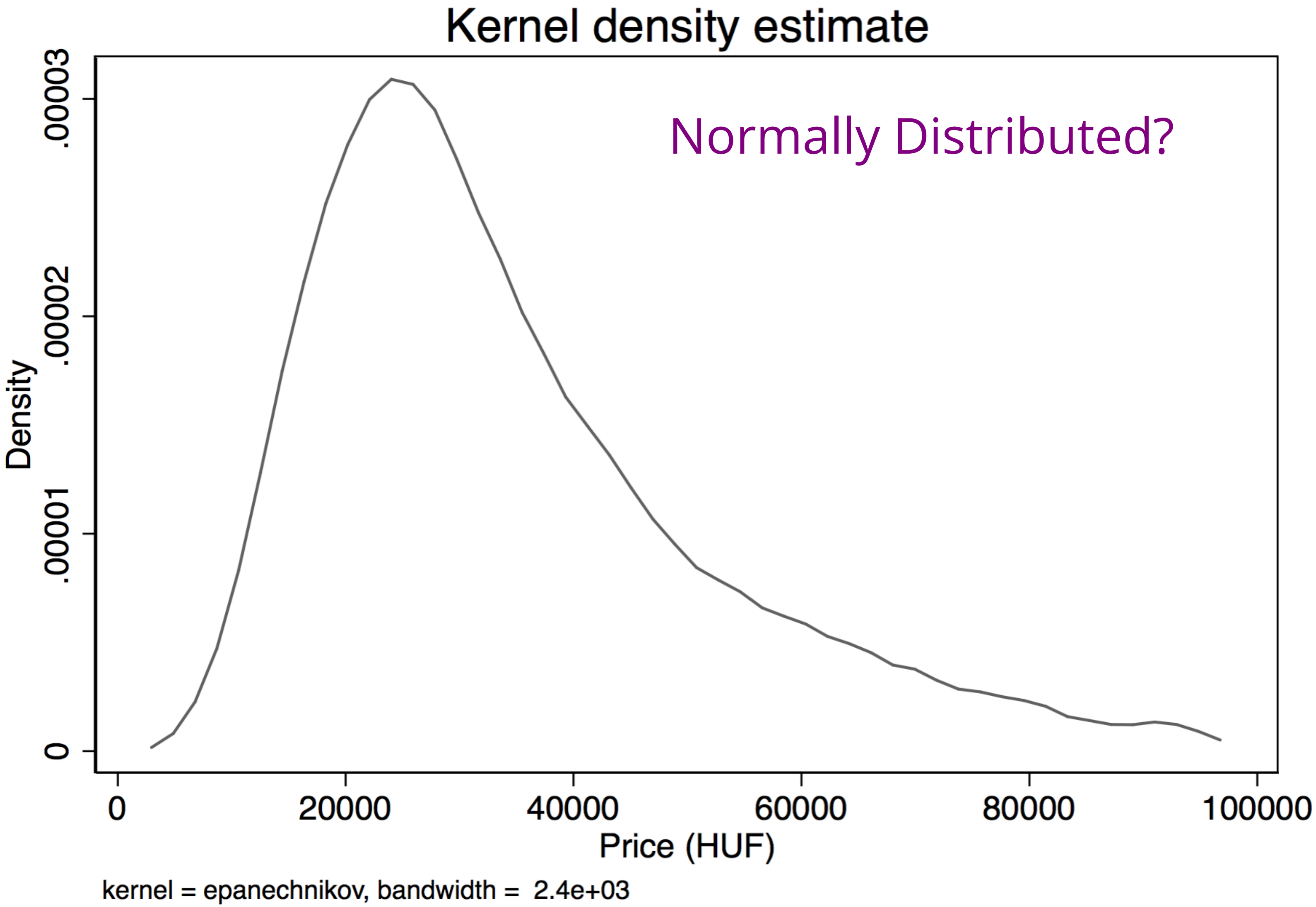
- A way to think about them is like curves that wrap around the corresponding hisogram

- The most important parameter to set is the bandwidth, which is similar to setting the binsize in the histogram



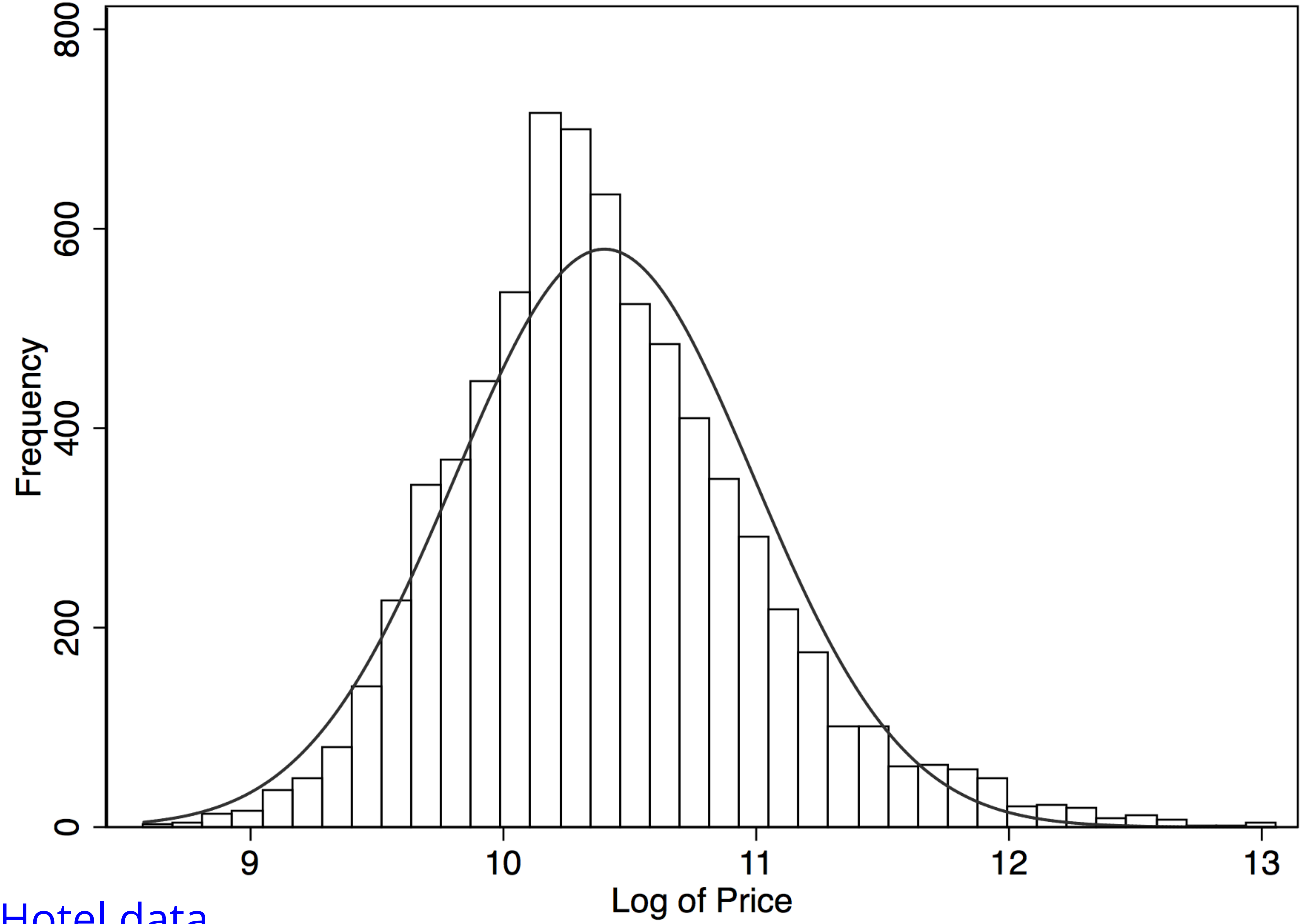
Normally Distributed?

FIFA data

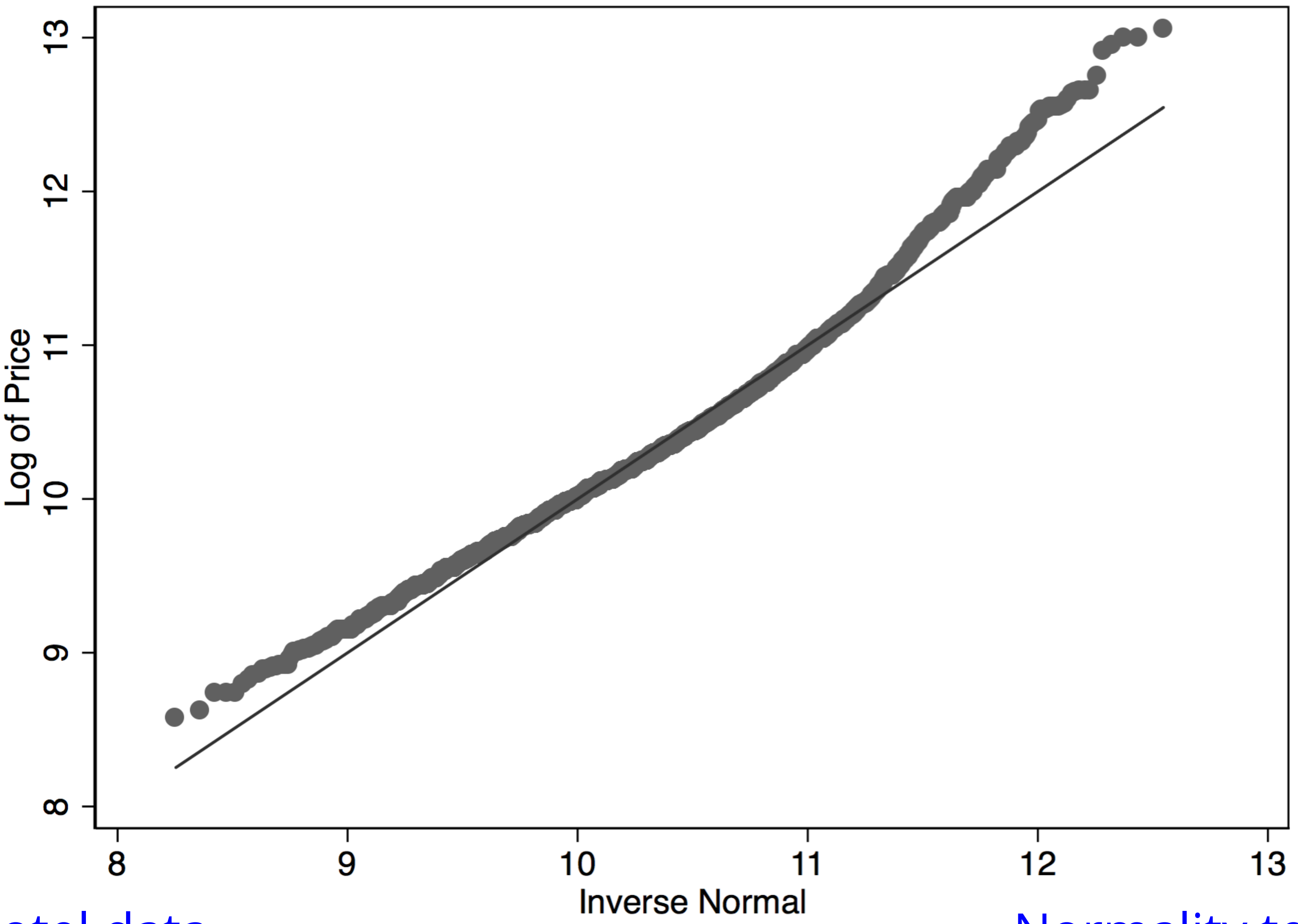


Normally Distributed?

Hotel data



Hotel data



Hotel data

Normality test

Joint Distributions

- In real life, we are often interested how variables that are related to each other. For example, number of rooms or bedrooms and average price of the house
    - The joint distribution shows the probabilities of each value combination of these variables we are interested in

Conditional Distributions

- Conditional distributions are distributions of one variable for one (or more) values of the other variable

Covariance

- It provides an indication of the dependence between 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

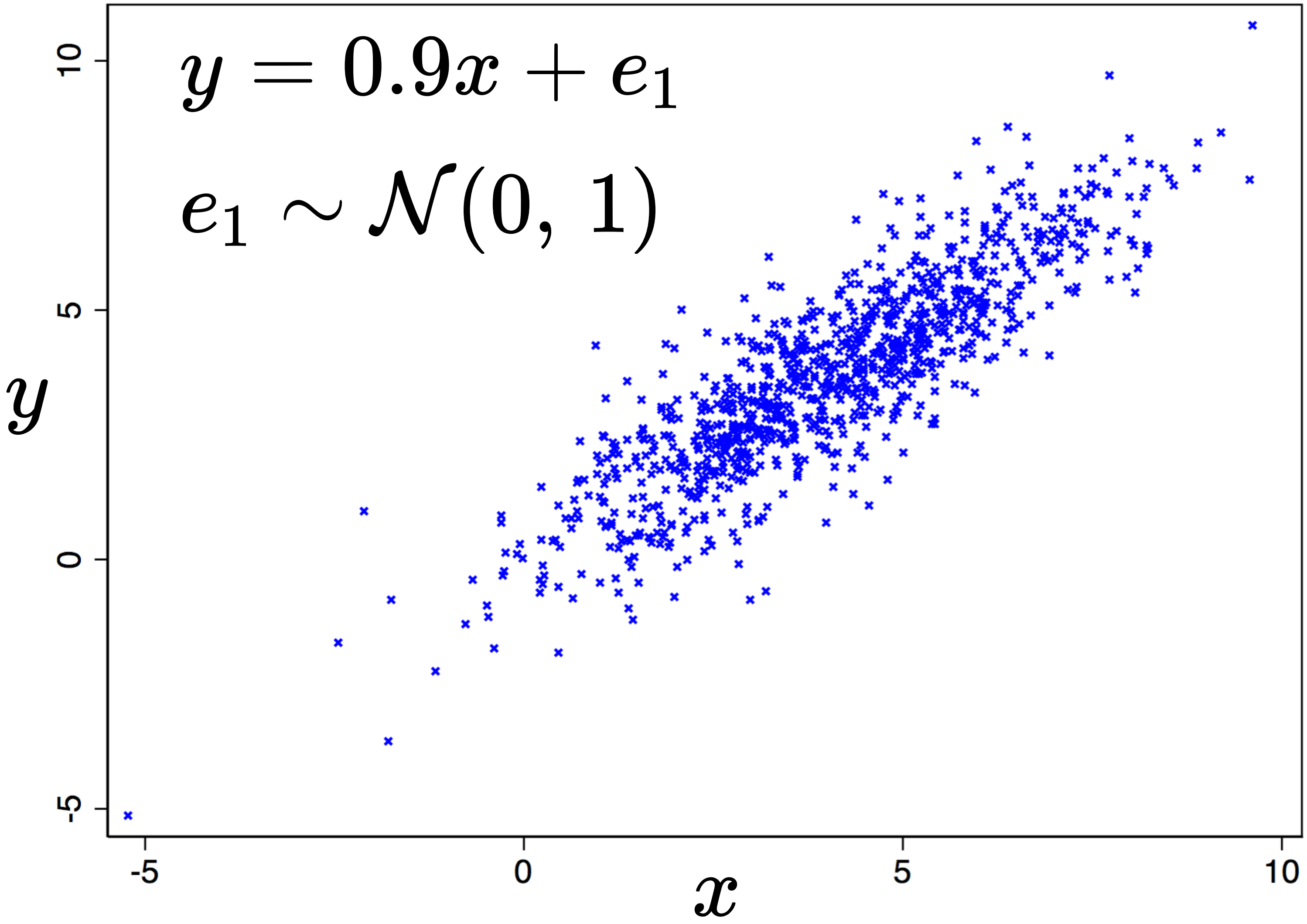
$$\rho = Corr(x, y) = \frac{Cov(x,y)}{Std(x) \cdot Std(y)}$$

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

Scatterplot

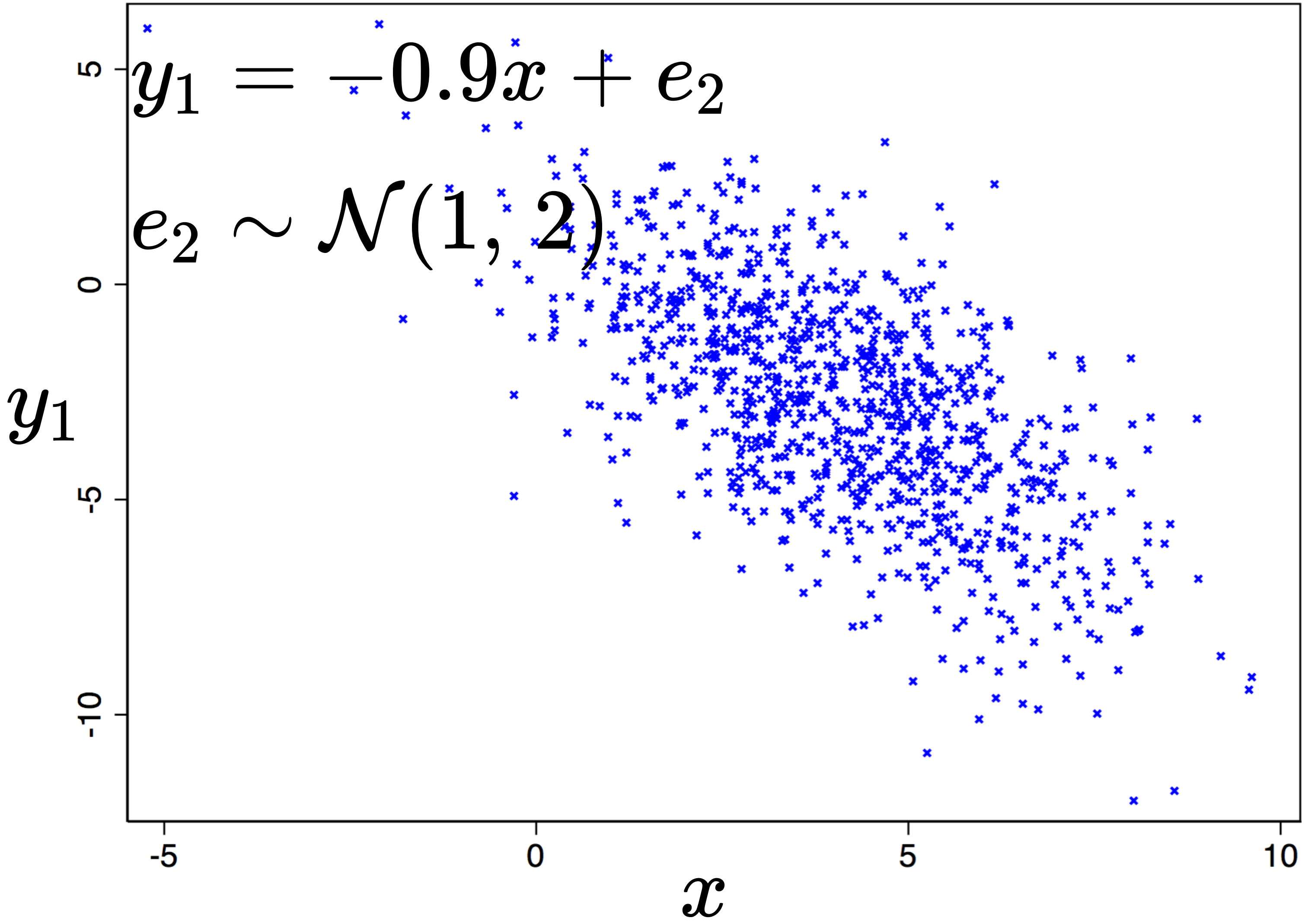
- Allows us to tell if there is a relationship among pairs of variables under consideration
  - We investigate if there is a linear relationship, nonlinear, or no relationship

Simulated data



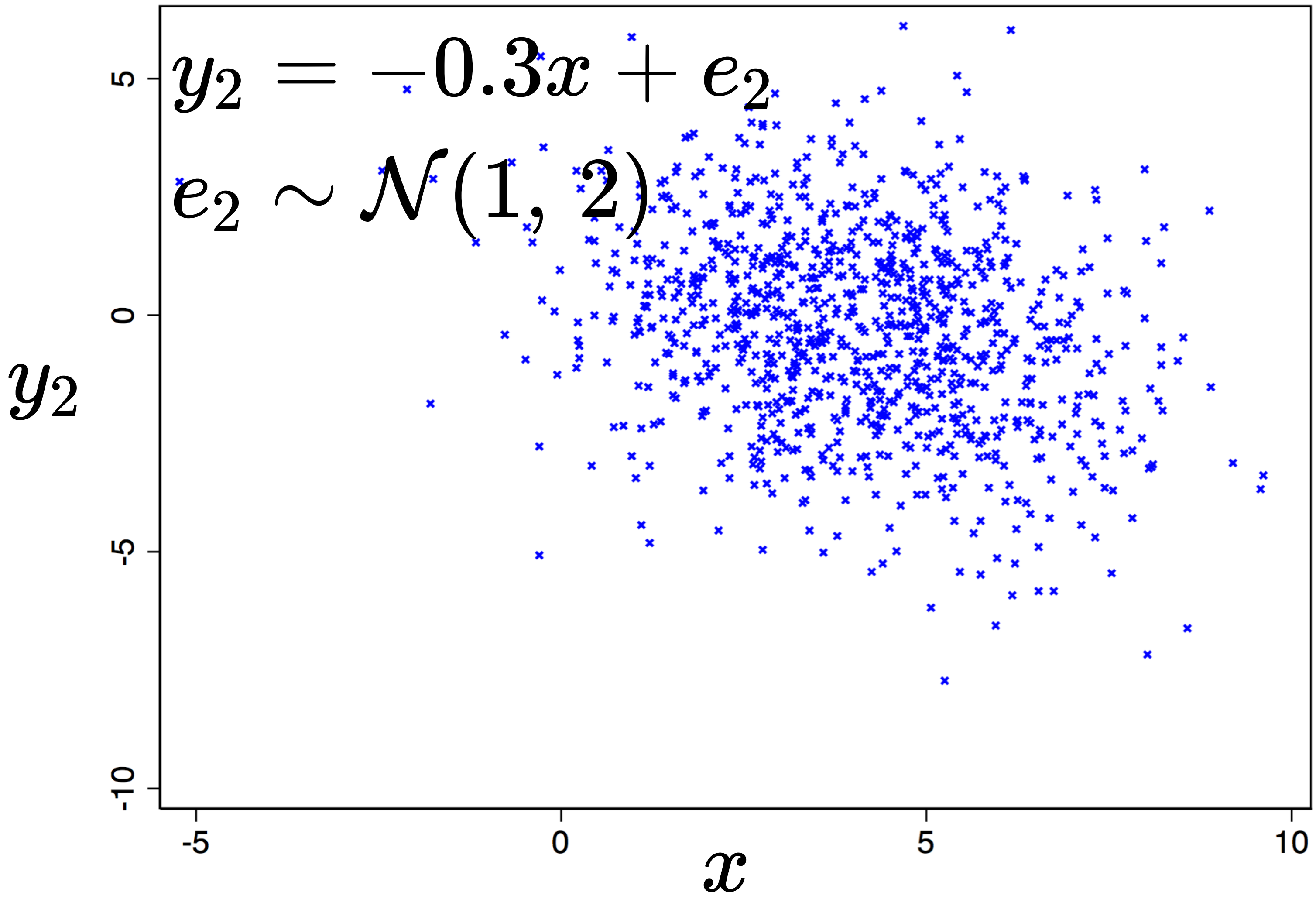
$$y = 0.9x + e_1$$

$$e_1 \sim \mathcal{N}(0, 1)$$



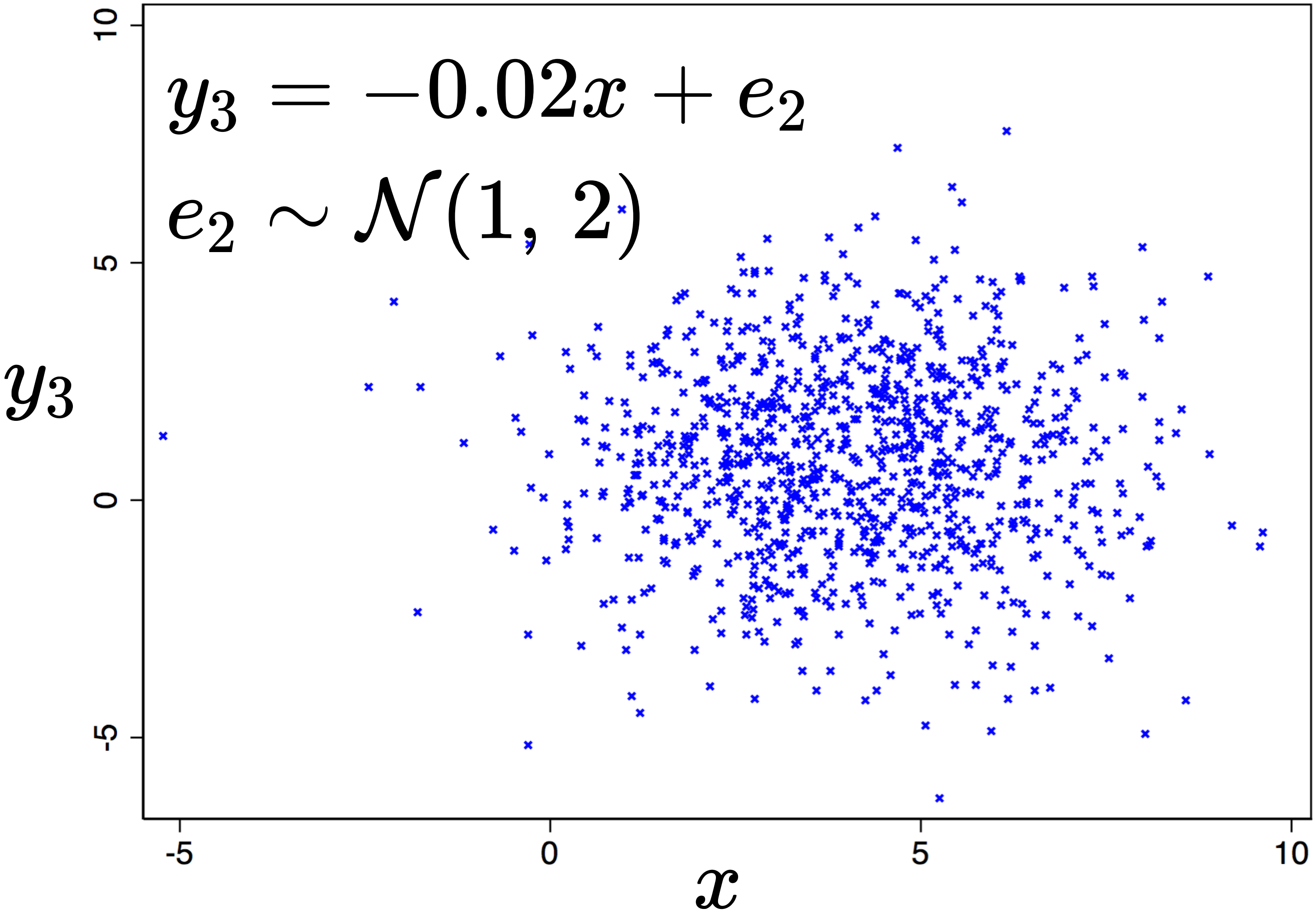
$$y_1 = -0.9x + e_2$$

$$e_2 \sim \mathcal{N}(1, 2)$$

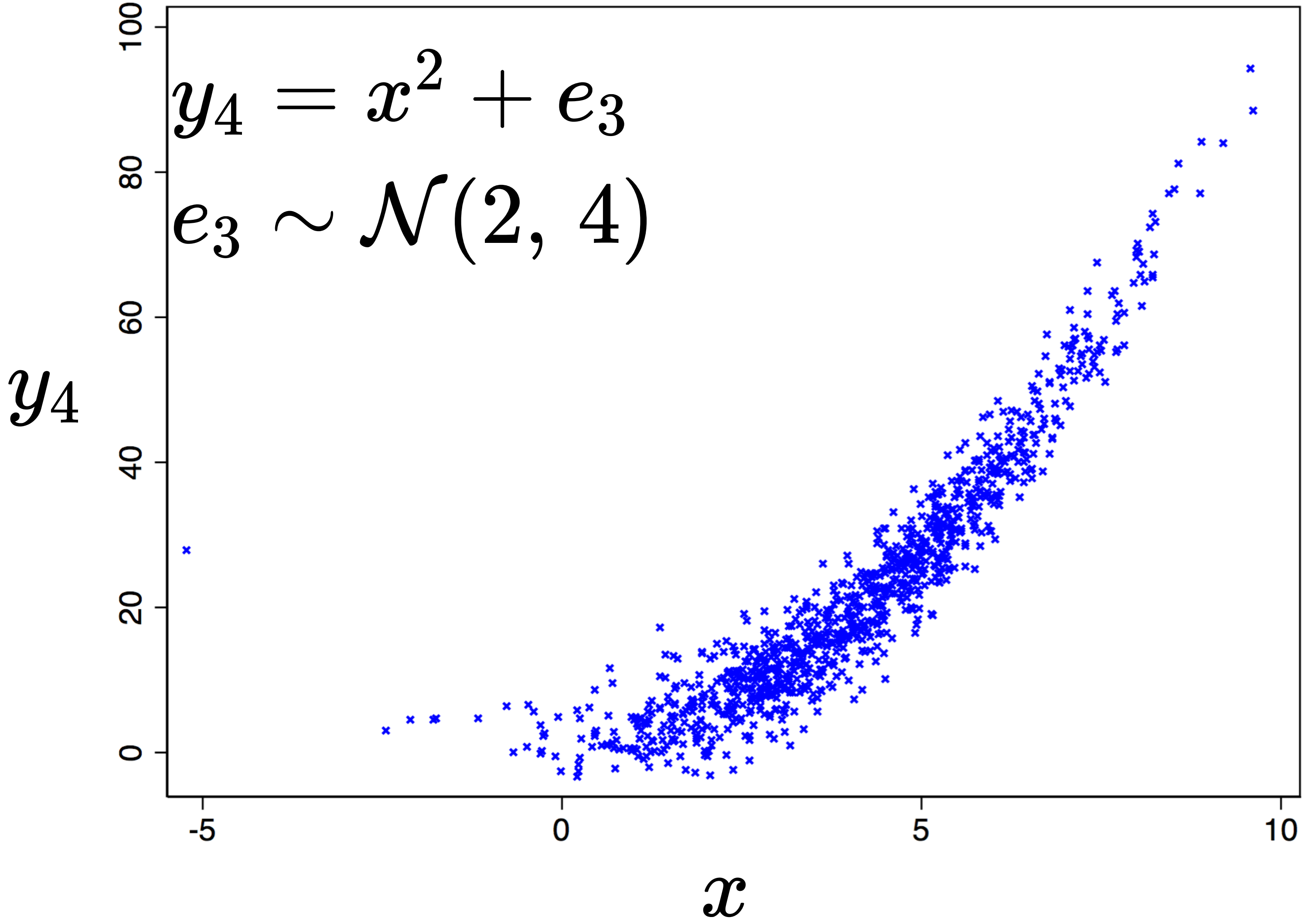


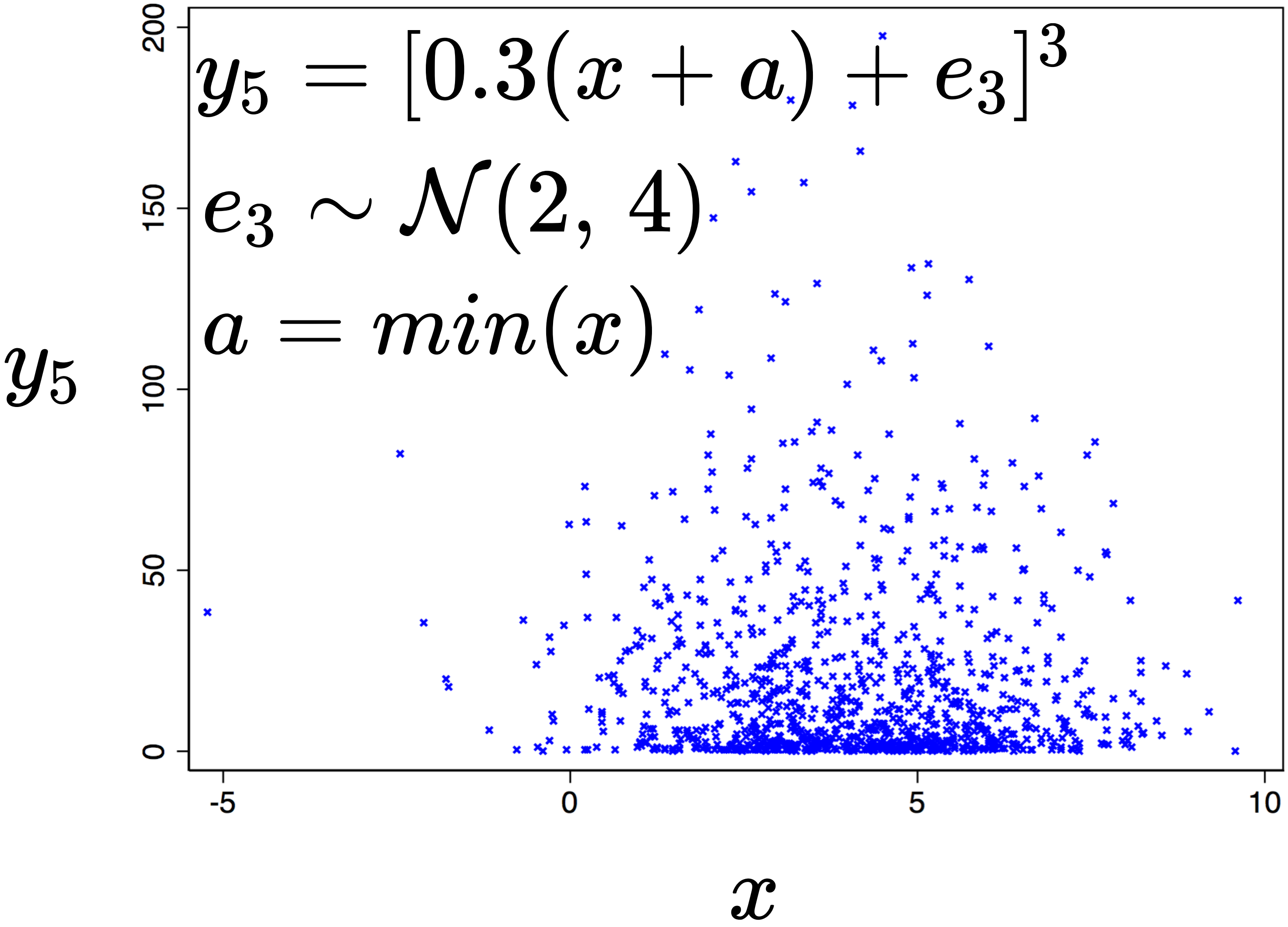
$$y_2 = -0.3x + e_2$$

$$e_2 \sim \mathcal{N}(1, 2)$$



$y_3 = -0.02x + e_2$
$e_2 \sim \mathcal{N}(1,\, 2)$
y_3
x
10
5
0
-5
-5
0
5
10


$y_4 = x^2 + e_3$
$e_3 \sim \mathcal{N}(2,\, 4)$
y_4
100
80
60
40
20
0
-5
0
5
10
x

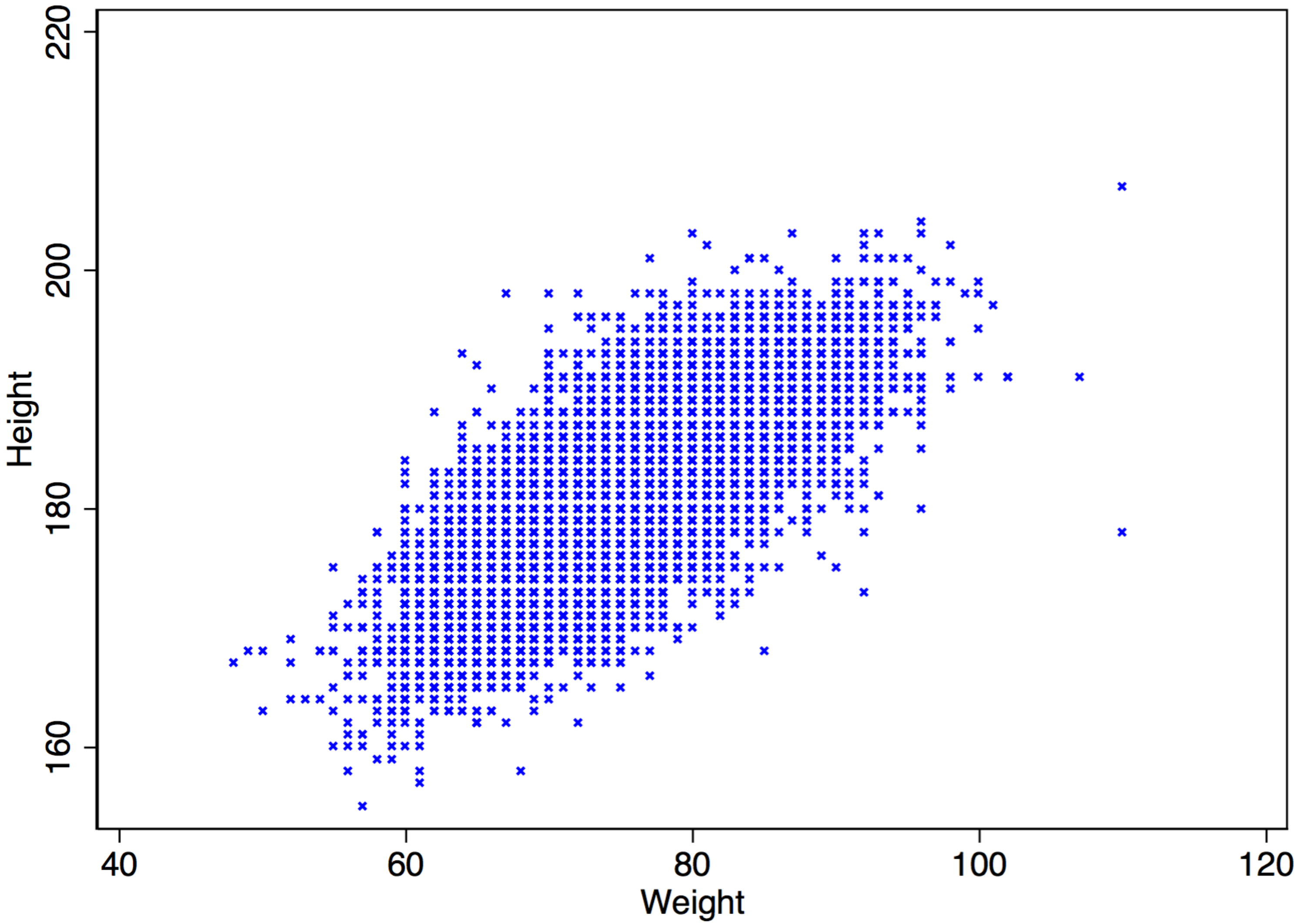


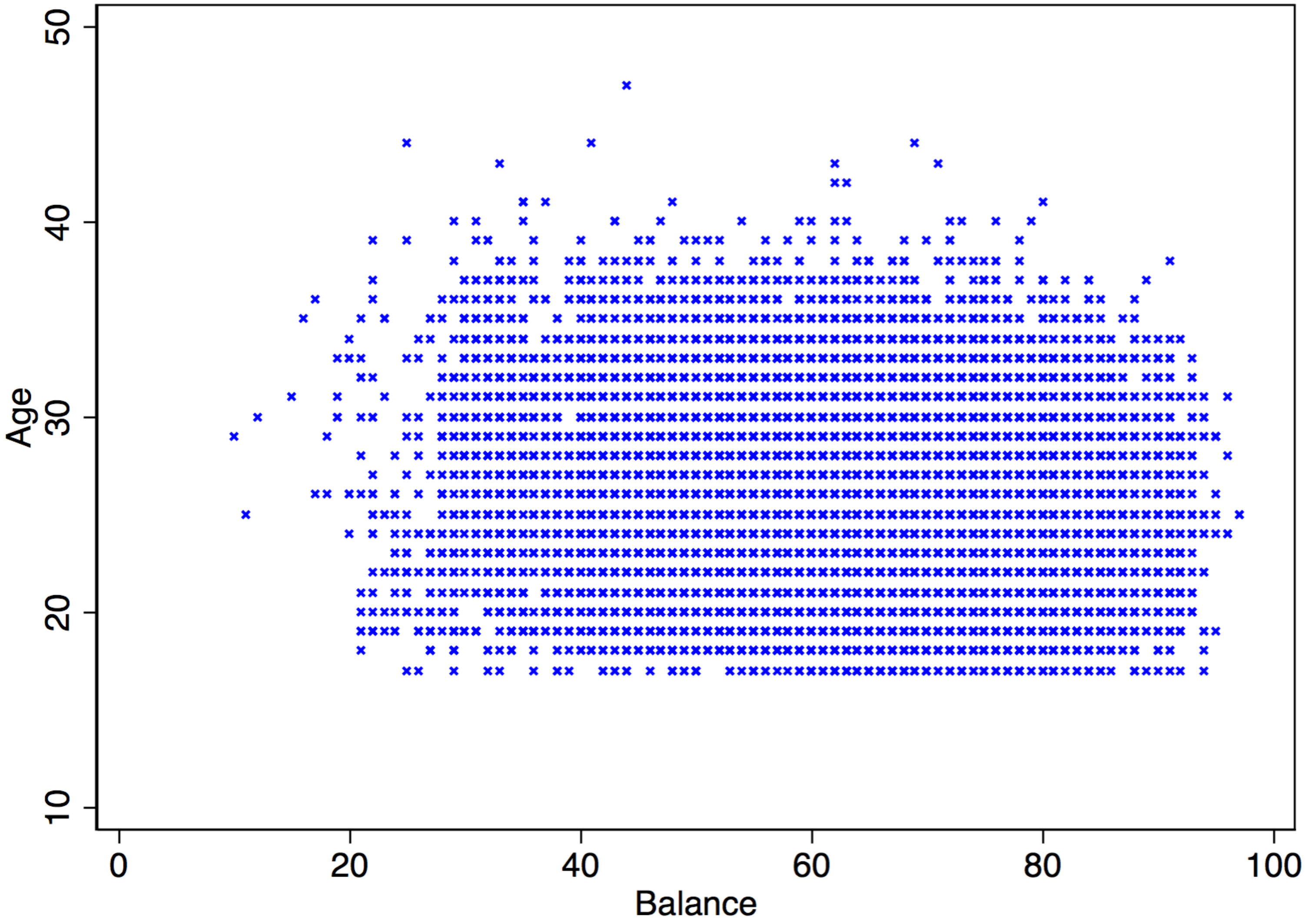
$$y_5 = [0.3(x+a) + e_3]^3$$

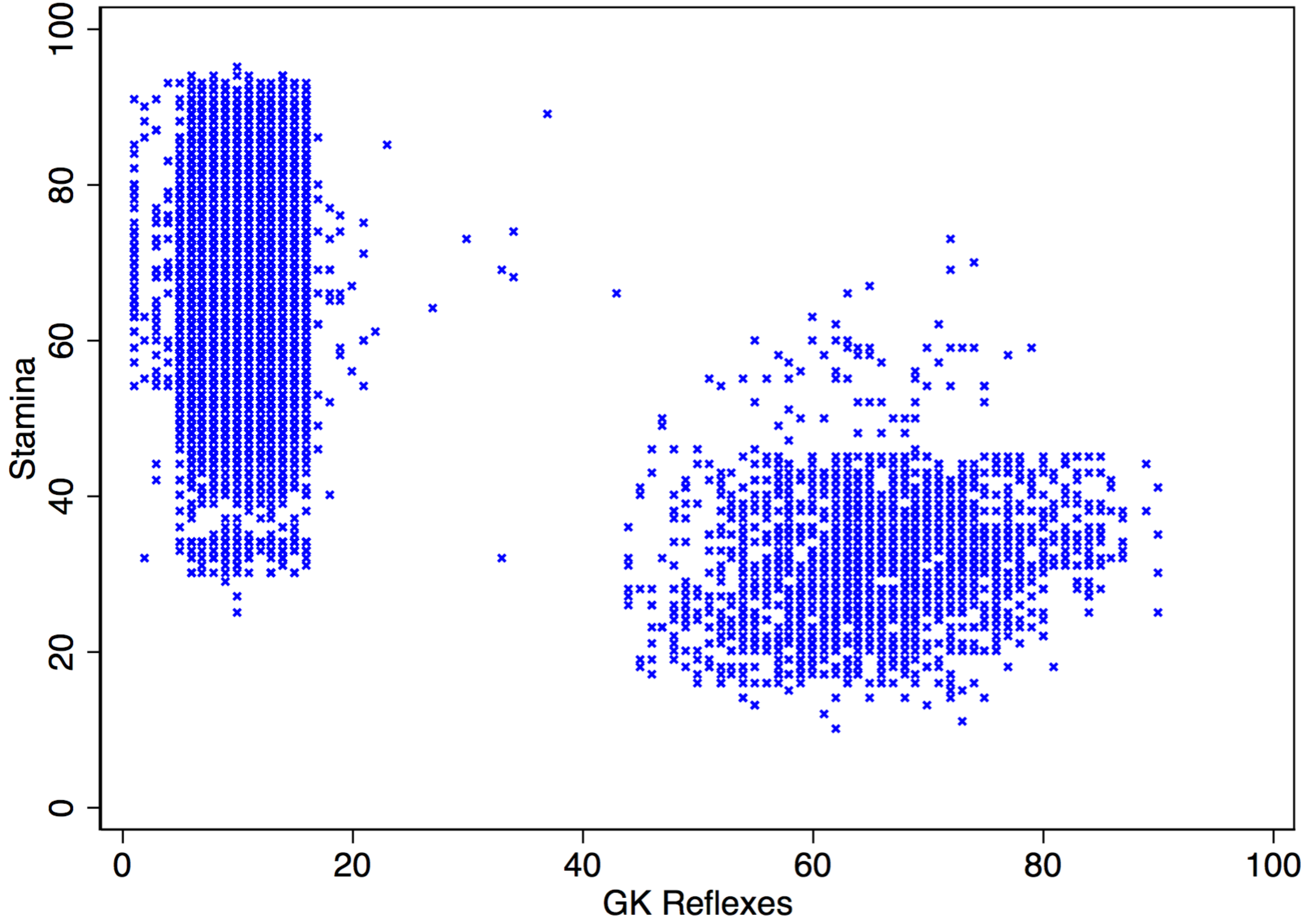
$$e_3 \sim \mathcal{N}(2,\ 4)$$

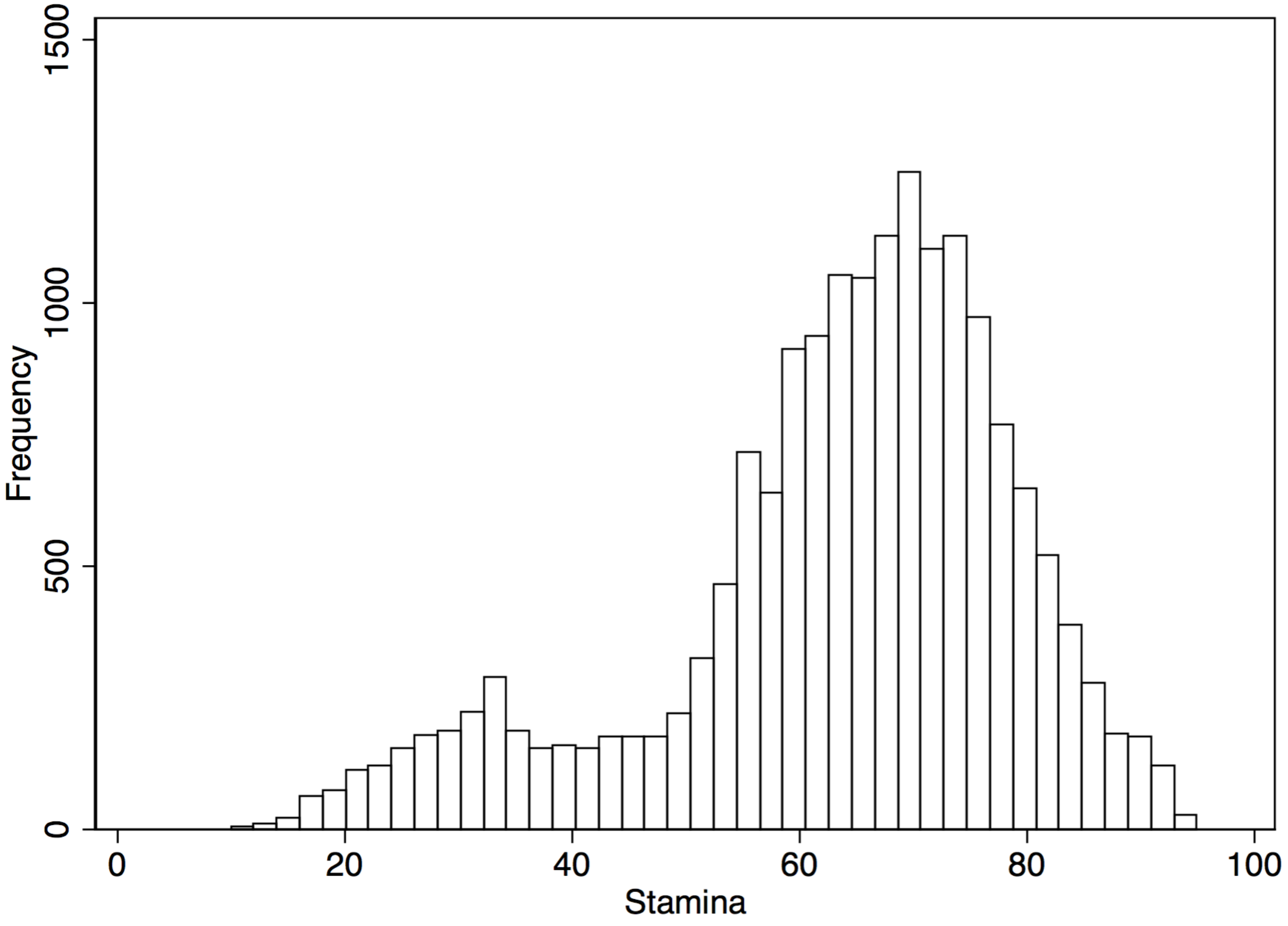
$$a = min(x)$$

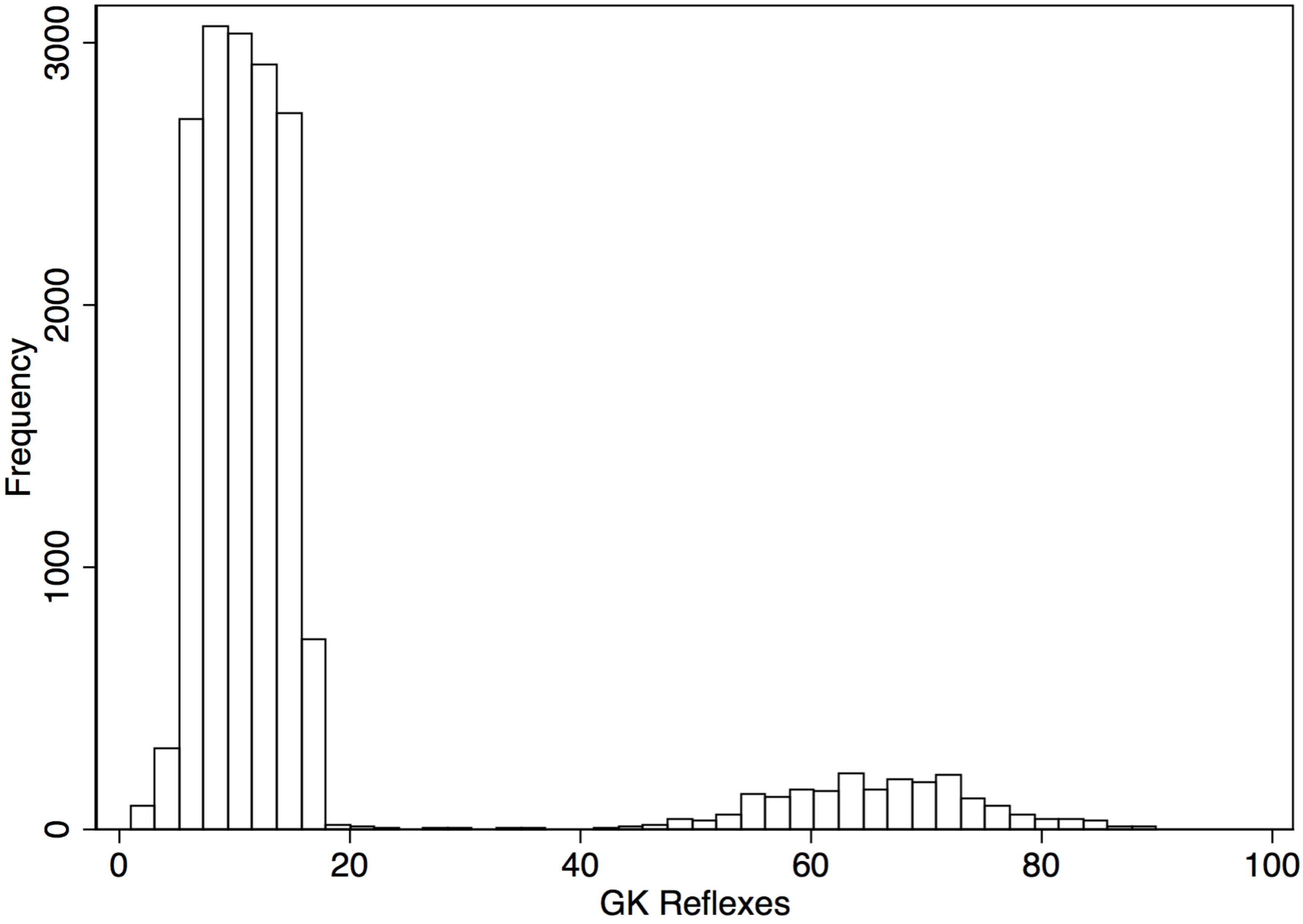
FIFA data


220
200
180
160
Height
40
60
80
100
120
Weight

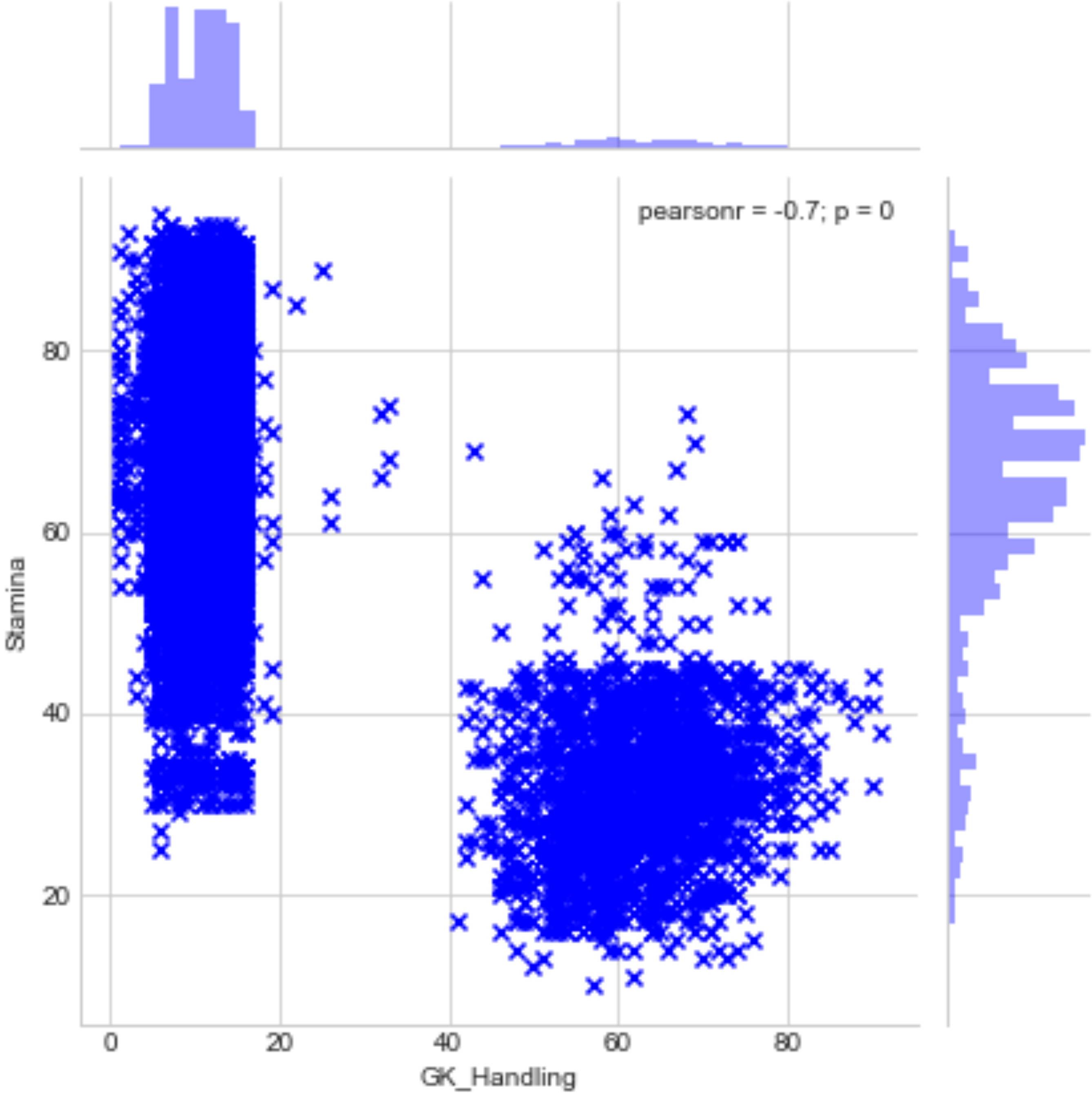

Age
10
20
30
40
50
0
20
40
60
80
100
Balance

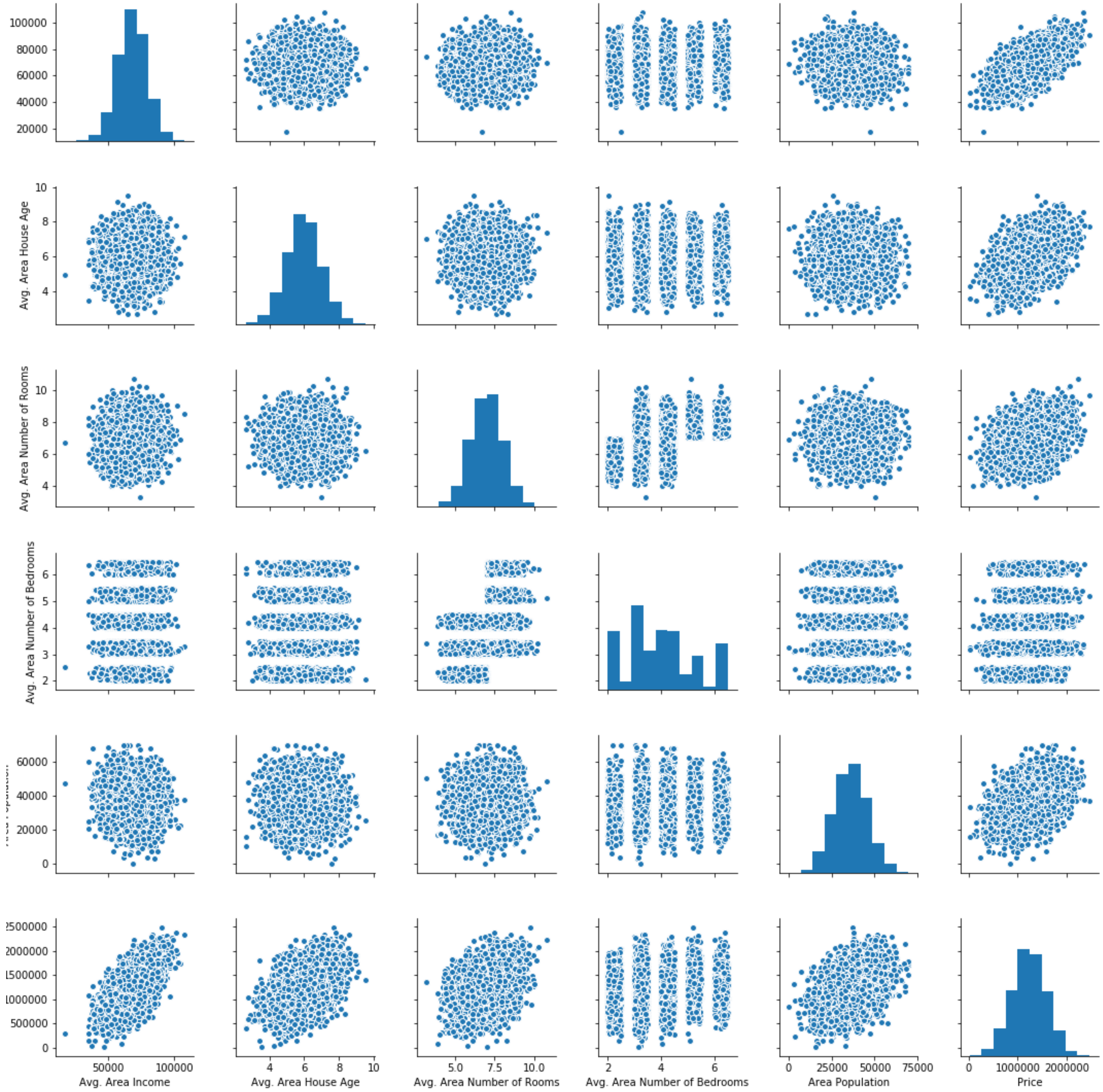

Stamina
0
20
40
60
80
100
GK Reflexes
0
20
40
60
80
100


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

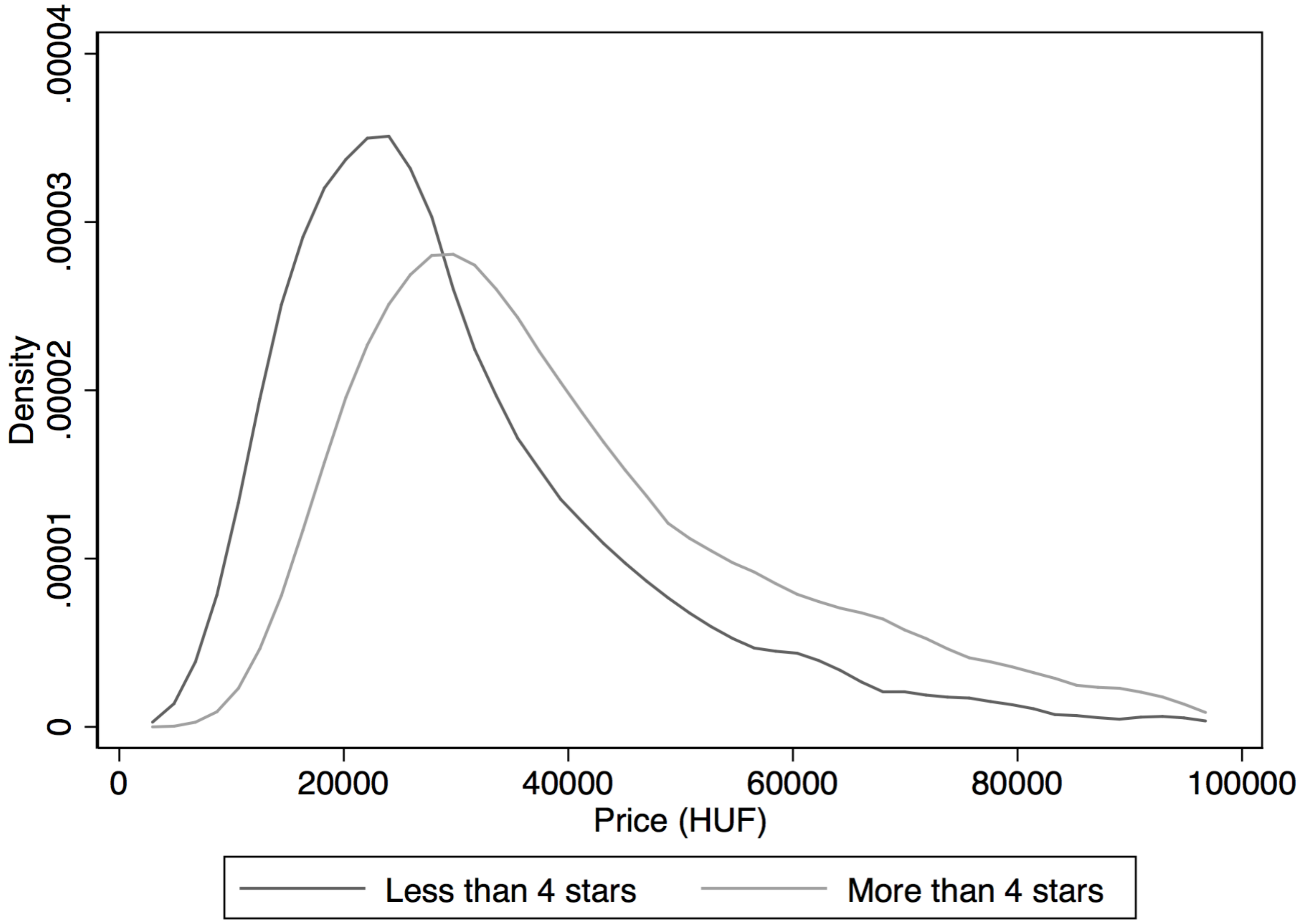


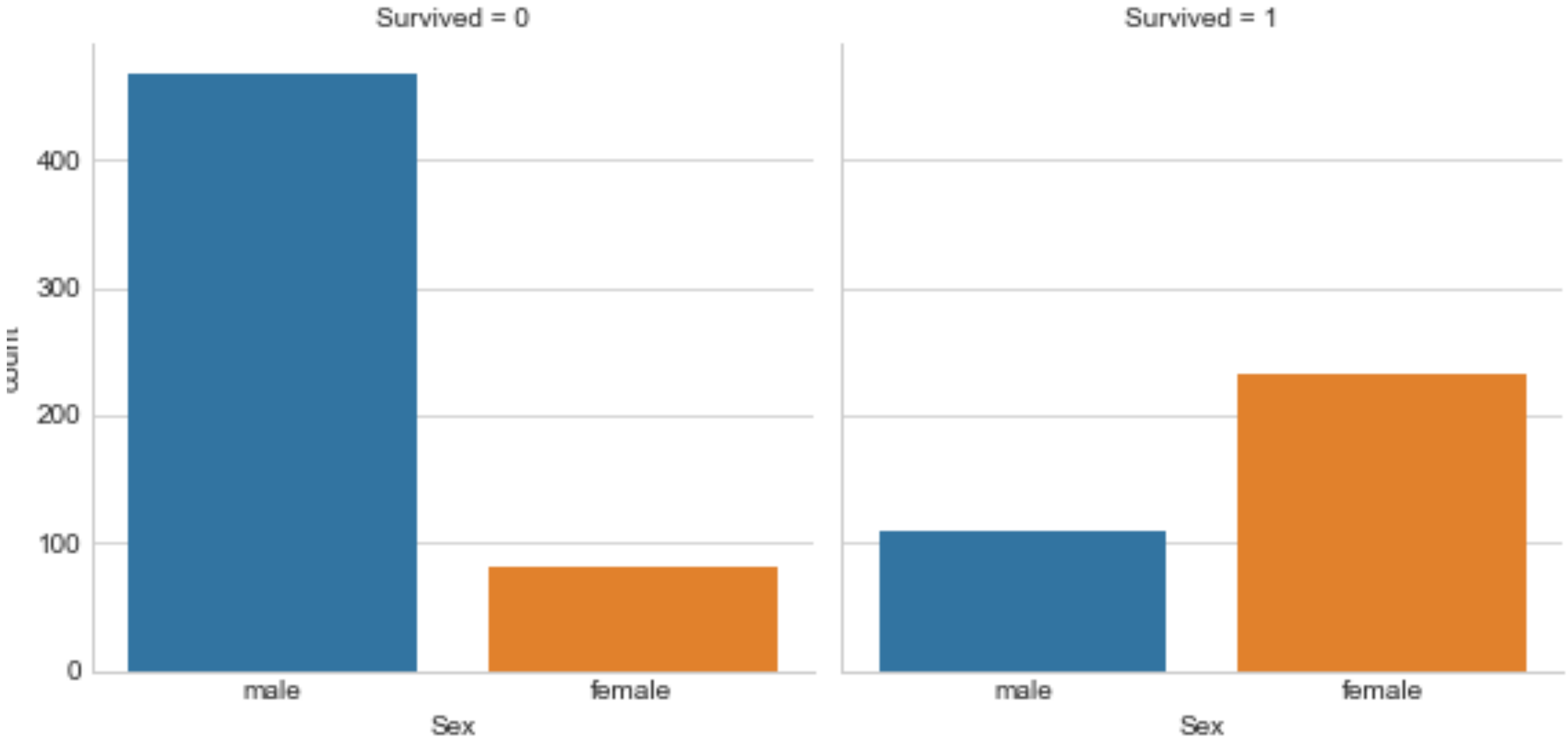
Frequency
0
1000
2000
3000
0
20
40
60
80
100
GK Reflexes

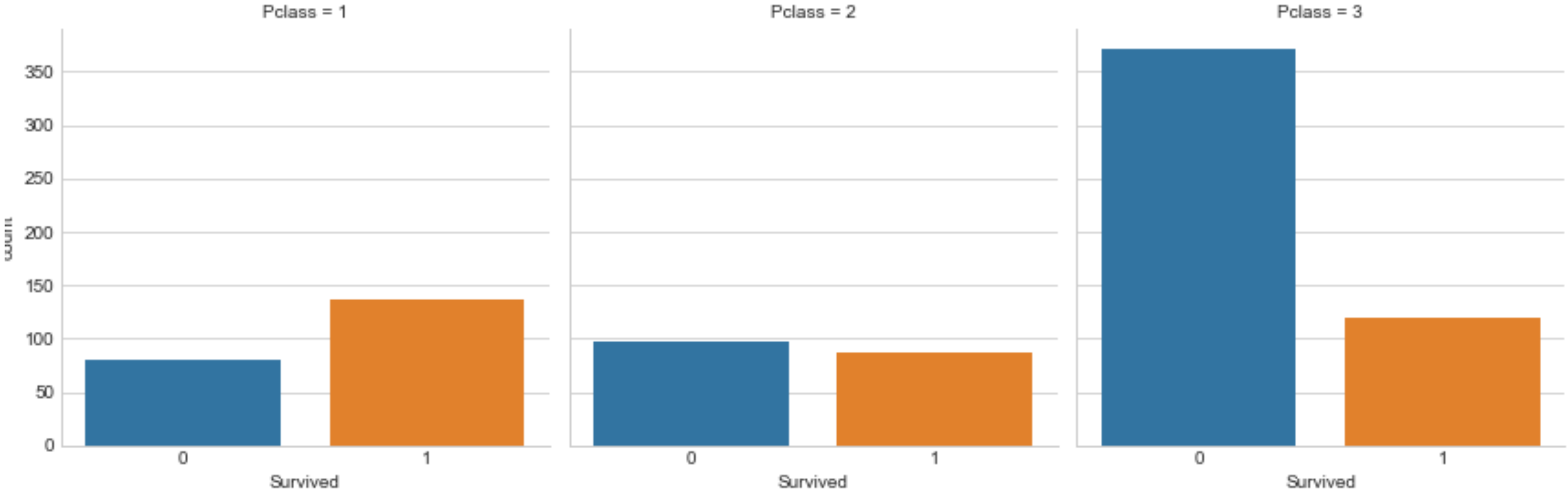

pearsonr = -0.7; p = 0
Stamina
20
40
60
80
0
20
40
60
80
GK_Handling


100000
80000
60000
40000
20000
Avg. Area House Age
10
8
6
4
Avg. Area Number of Rooms
10
8
6
4
Avg. Area Number of Bedrooms
6
5
4
3
2
60000
40000
20000
0
2500000
2000000
1500000
1000000
500000
0
50000
100000
Avg. Area Income
4
6
8
10
Avg. Area House Age
5.0
7.5
10.0
Avg. Area Number of Rooms
2
4
6
Avg. Area Number of Bedrooms
0
25000
50000
75000
Area Population
0
1000000
2000000
Price

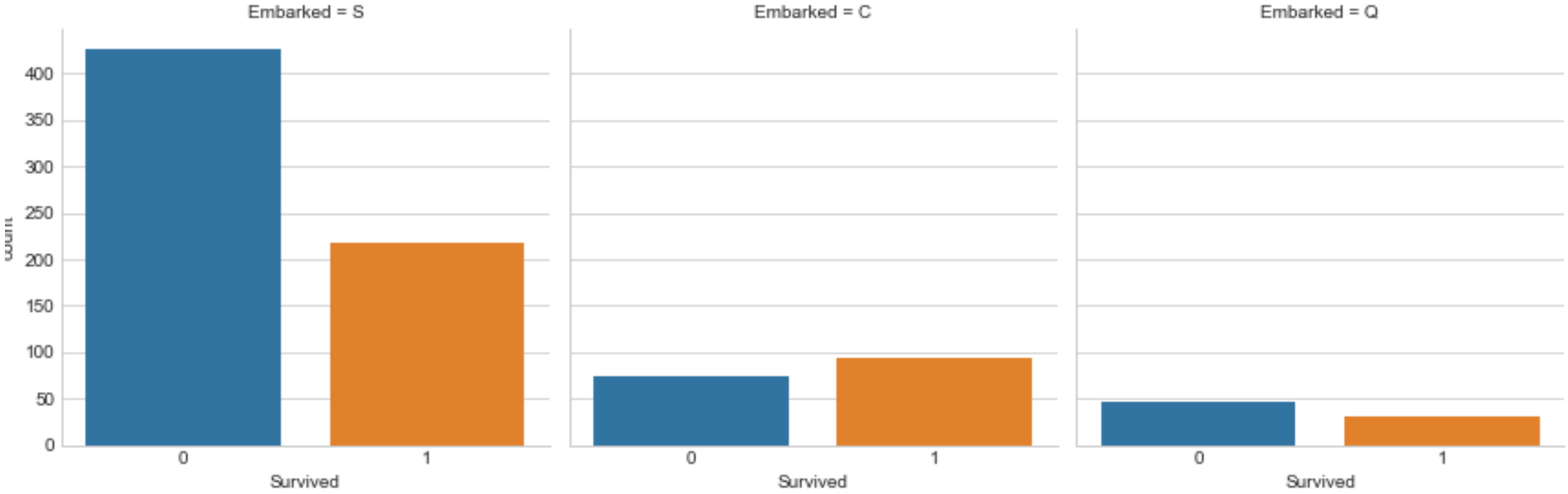
Conditional Distributions in Practice


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


Survived = 0
Survived = 1
count
400
300
200
100
0
male
female
Sex
male
female
Sex

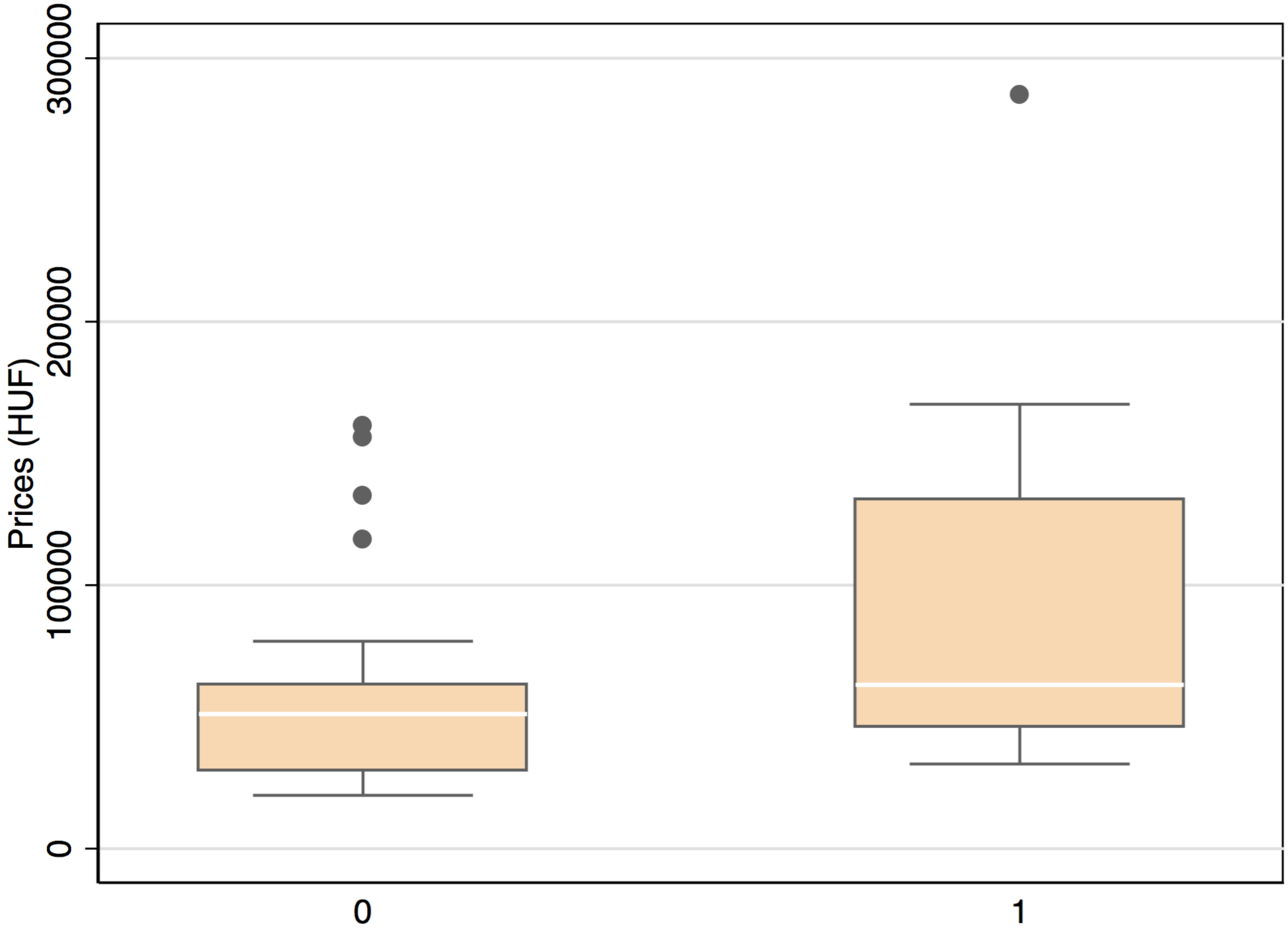


pclas=passanger class



Embarked=port of embarkation

Vienna Hotel Prices (again)



Sources of data used in these slides

- Hotel data, already available (only to be shared in Moodle)
- Simulated variables will be posted (Moodle and GitHub), try it on yourself, now you have the codes
- Fifa database and Titanic passengers database are going to also be posted (Moodle and GitHub).
  - Downloaded from Kaggle