STATISTICAL THINKING IN PYTHON II

Welcome to the course!

You will be able to...

- Estimate parameters
- Compute confidence intervals
- Perform linear regressions
- Test hypotheses

with *real* data!


Caltech

We use hacker statistics

- Literally simulate probability
- Broadly applicable with a few principles

Statistical analysis of the beak of the finch

Geospiza fortis

Geospiza scandens

Source: John Gould, public domain

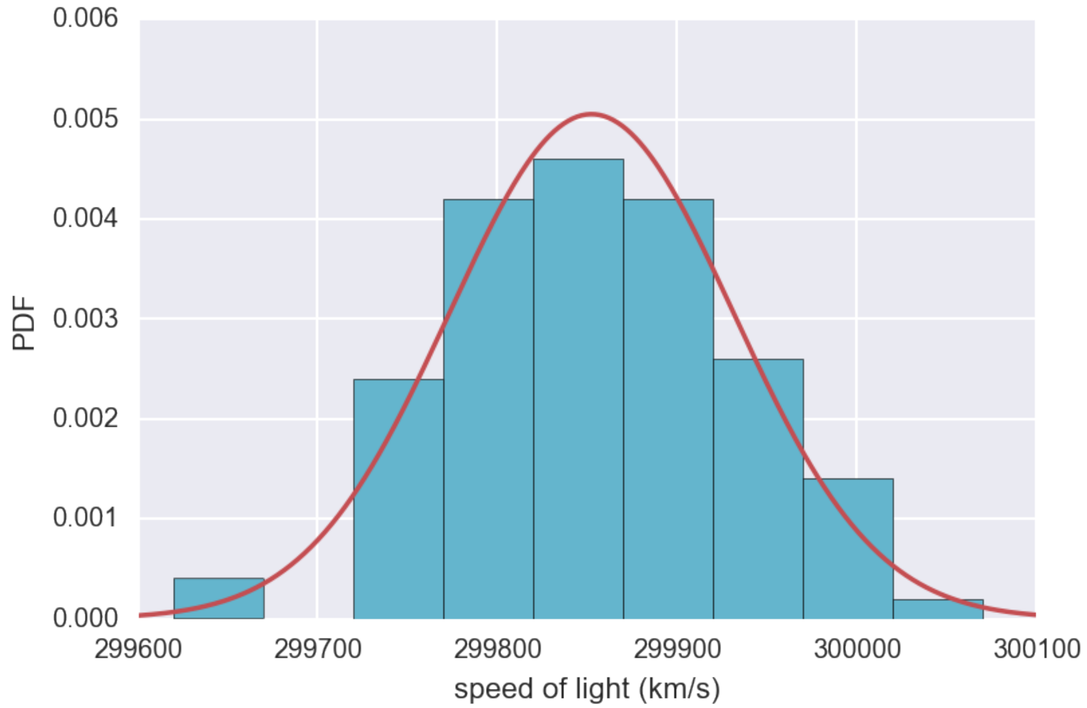
STATISTICAL THINKING IN PYTHON II

Let's start thinking statistically!

STATISTICAL THINKING IN PYTHON II

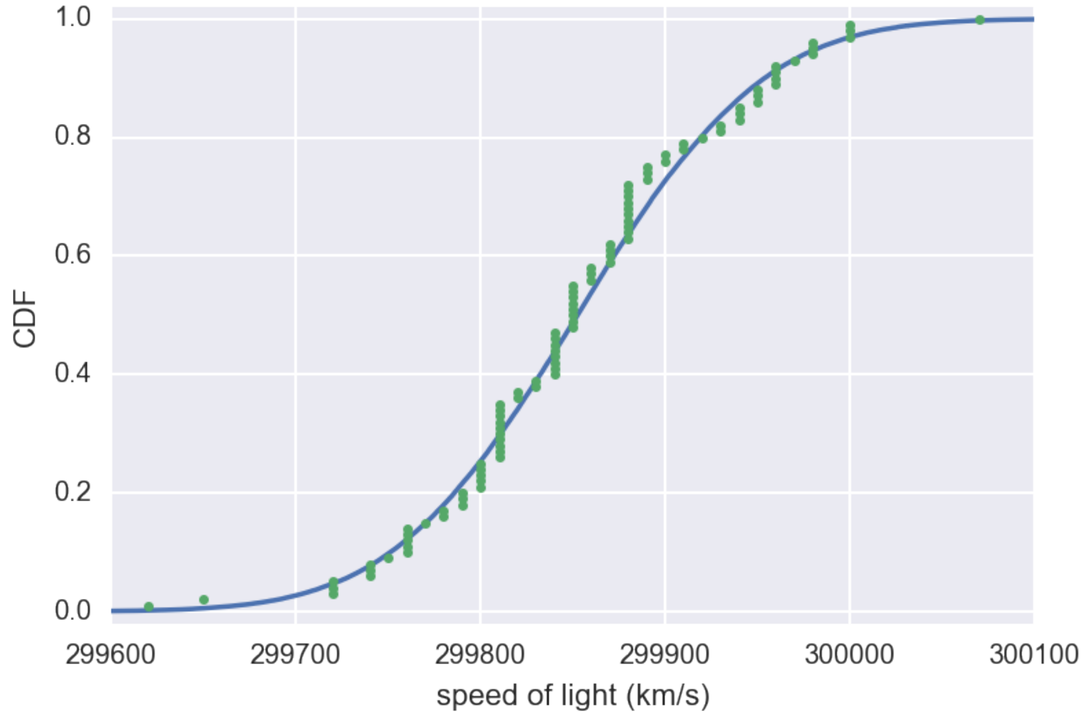
Optimal parameters

Histogram of Michelson's measurements



Data: Michelson, 1880

CDF of Michelson's measurements

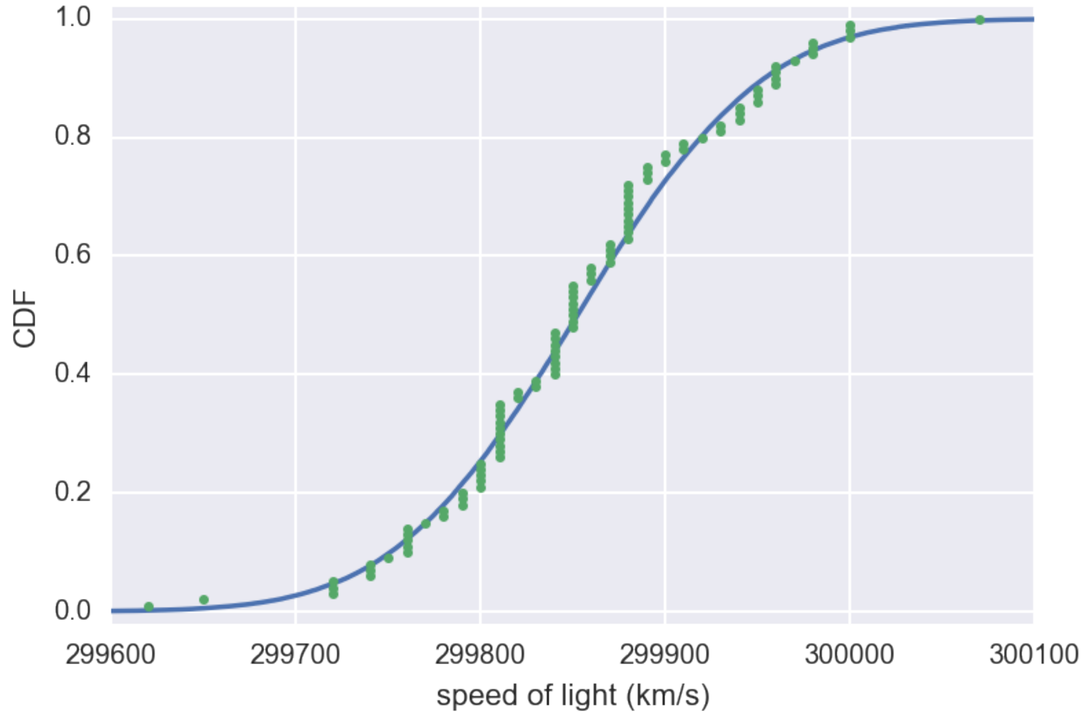


Data: Michelson, 1880

Checking Normality of Michelson data

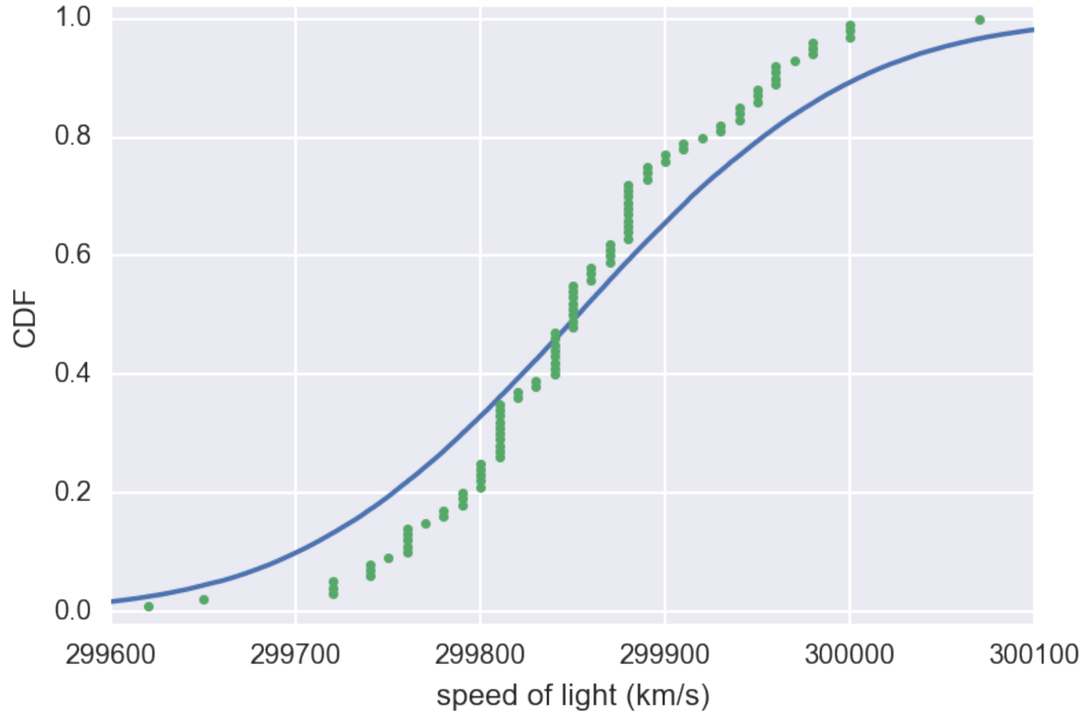
```
In [1]: import numpy as np

In [2]: import matplotlib.pyplot as plt

In [3]: mean = np.mean(michelson_speed_of_light)

In [4]: std = np.std(michelson_speed_of_light)

In [5]: samples = np.random.normal(mean, std, size=10000)
```

CDF of Michelson's measurements



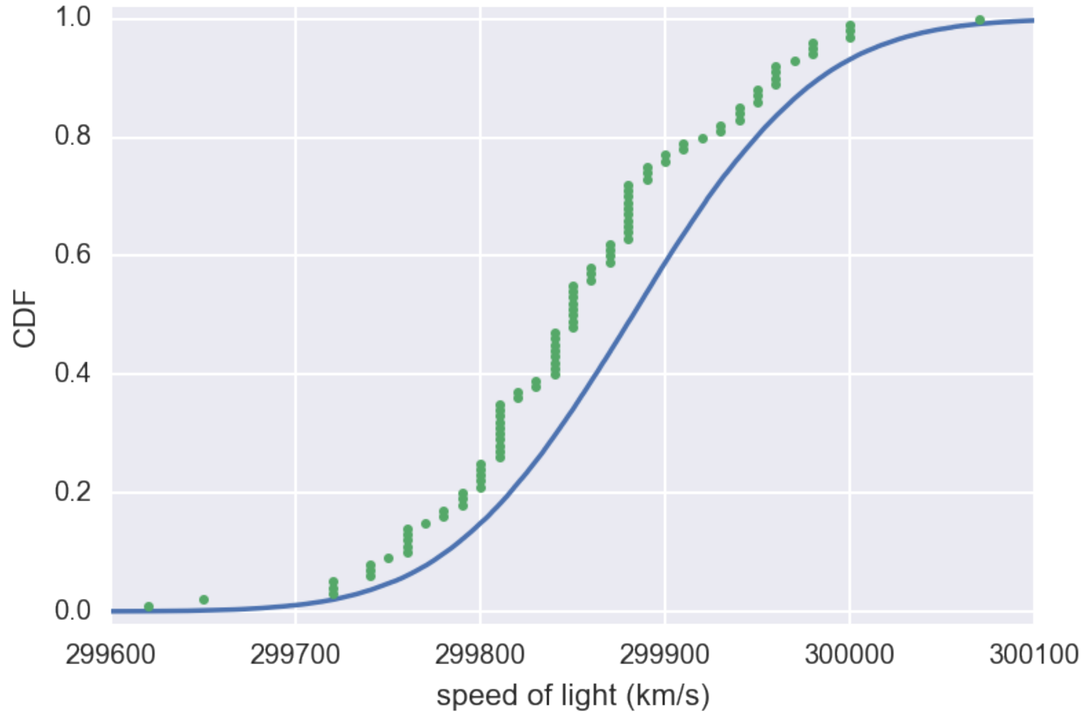
Data: Michelson, 1880

CDF with bad estimate of st. dev.



Data: Michelson, 1880

CDF with bad estimate of mean

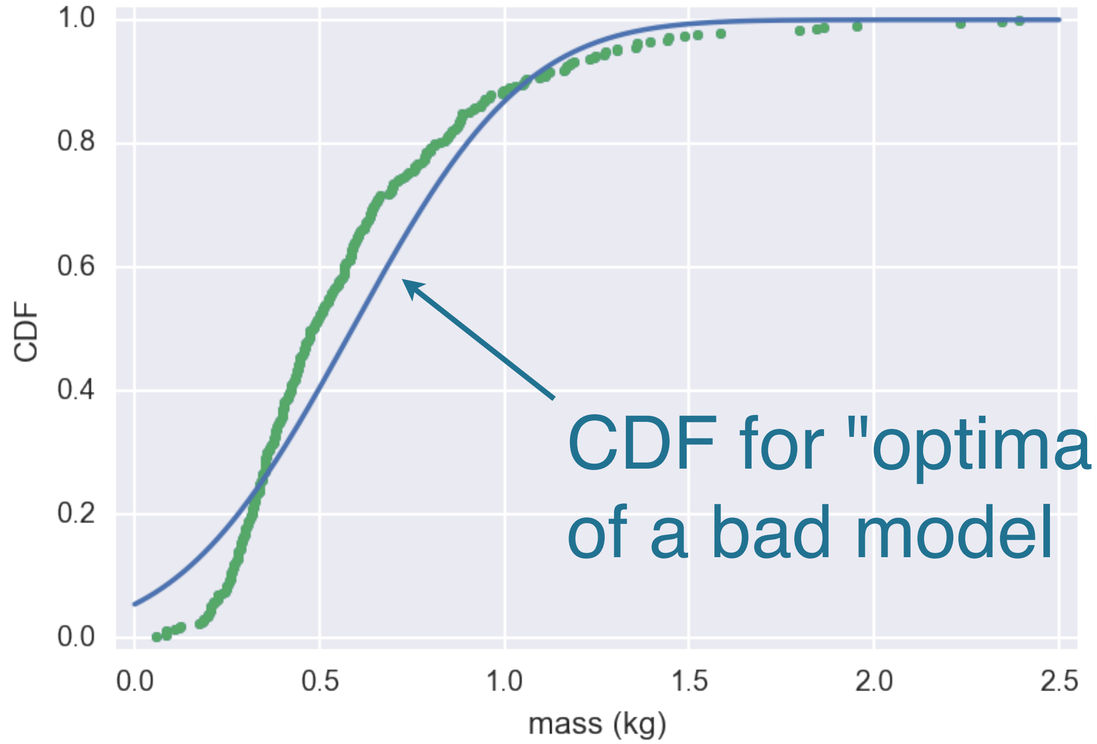


Data: Michelson, 1880

Optimal parameters

- Parameter values that bring the model in closest agreement with the data

Mass of MA large mouth bass



Source: Mass. Dept. of Environmental Protection

Packages to do statistical inference

`scipy.stats`

`statsmodels`

`hacker stats with numpy`

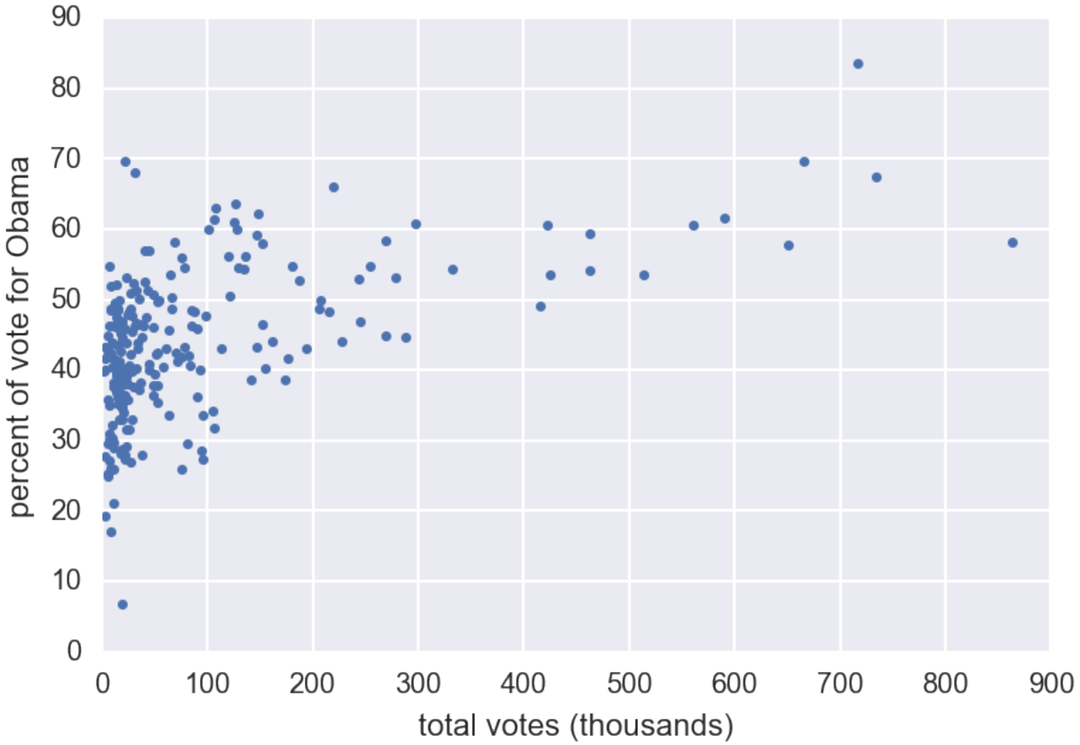
STATISTICAL THINKING IN PYTHON II

Let’s practice!

STATISTICAL THINKING IN PYTHON II

Linear regression by least squares

2008 US swing state election results

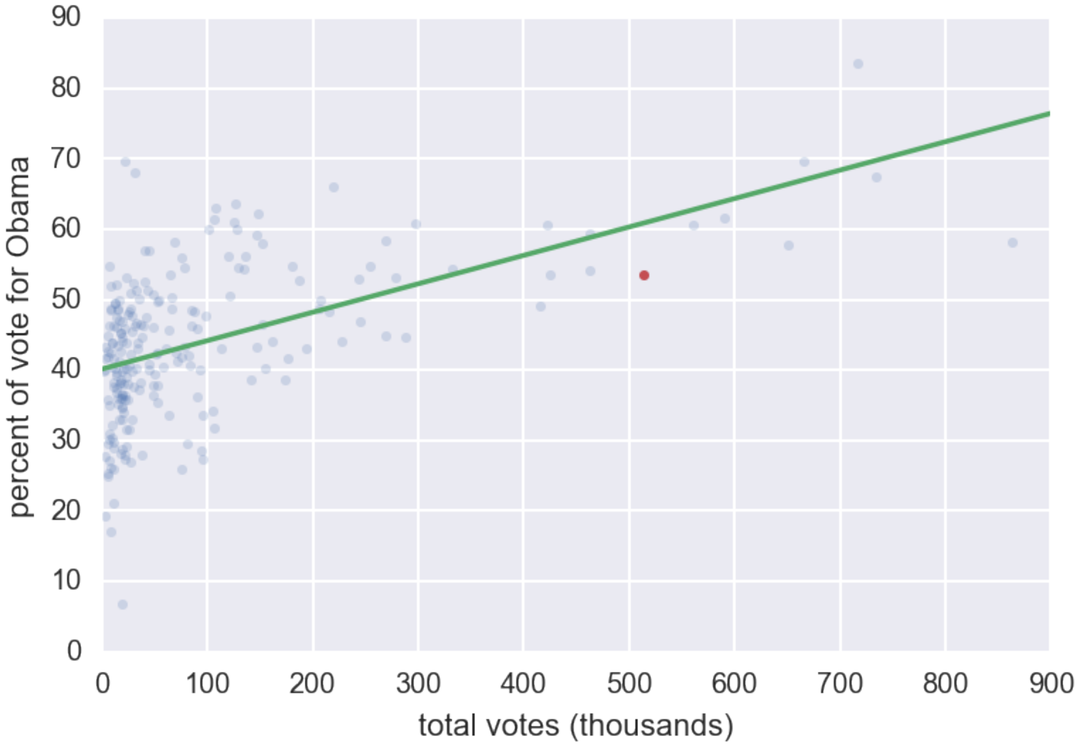


Data retrieved from Data.gov (https://www.data.gov/)

2008 US swing state election results

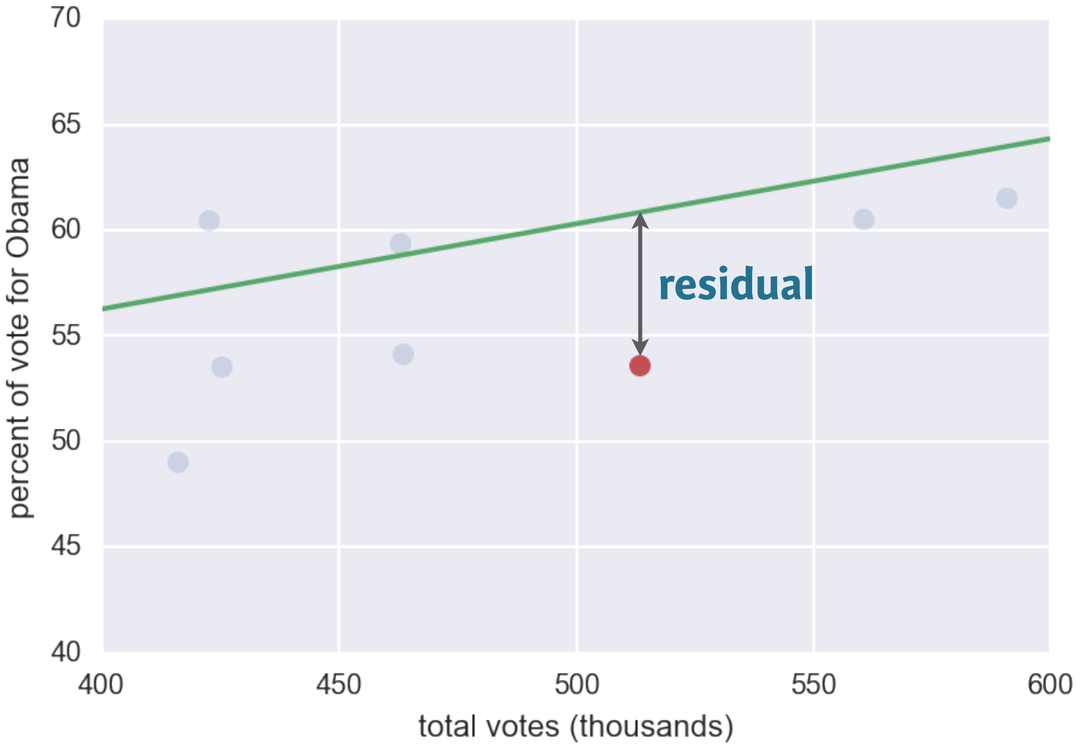
Data retrieved from Data.gov (https://www.data.gov/)

2008 US swing state election results



Data retrieved from Data.gov (https://www.data.gov/)

Residuals



Data retrieved from Data.gov (https://www.data.gov/)

Least squares

- The process of finding the parameters for which the sum of the squares of the residuals is minimal

Least squares with `np.polyfit()`

```
In [1]: slope, intercept = np.polyfit(total_votes,
   ...:                               dem_share, 1)

In [2]: slope
Out[2]: 4.037071700946555e-05

In [3]: intercept
Out[3]: 40.113911968641744
```

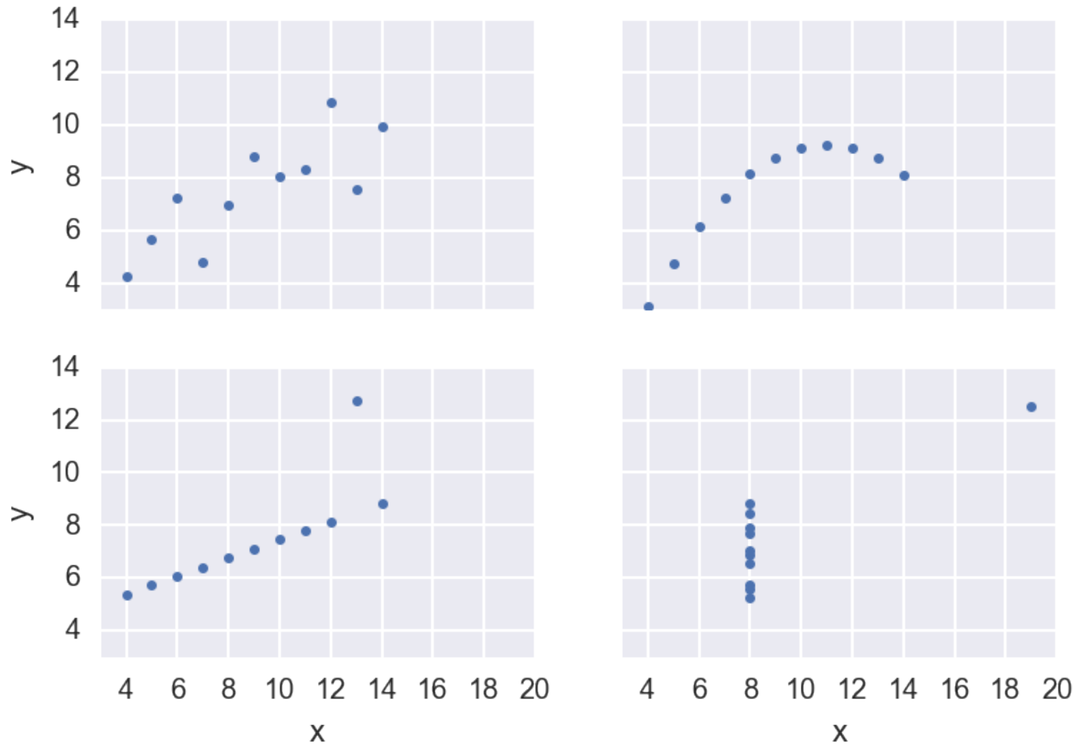
STATISTICAL THINKING IN PYTHON II

Let’s practice!

STATISTICAL THINKING IN PYTHON II

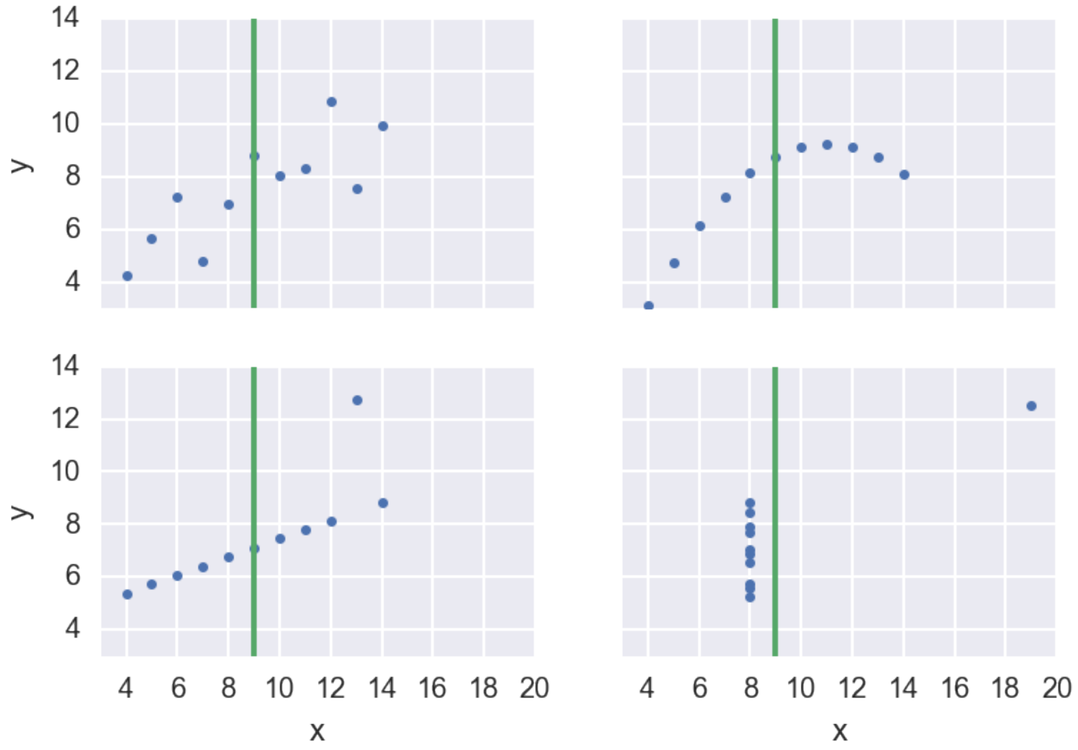
The importance of EDA: Anscombe's quartet

Anscombe's quartet



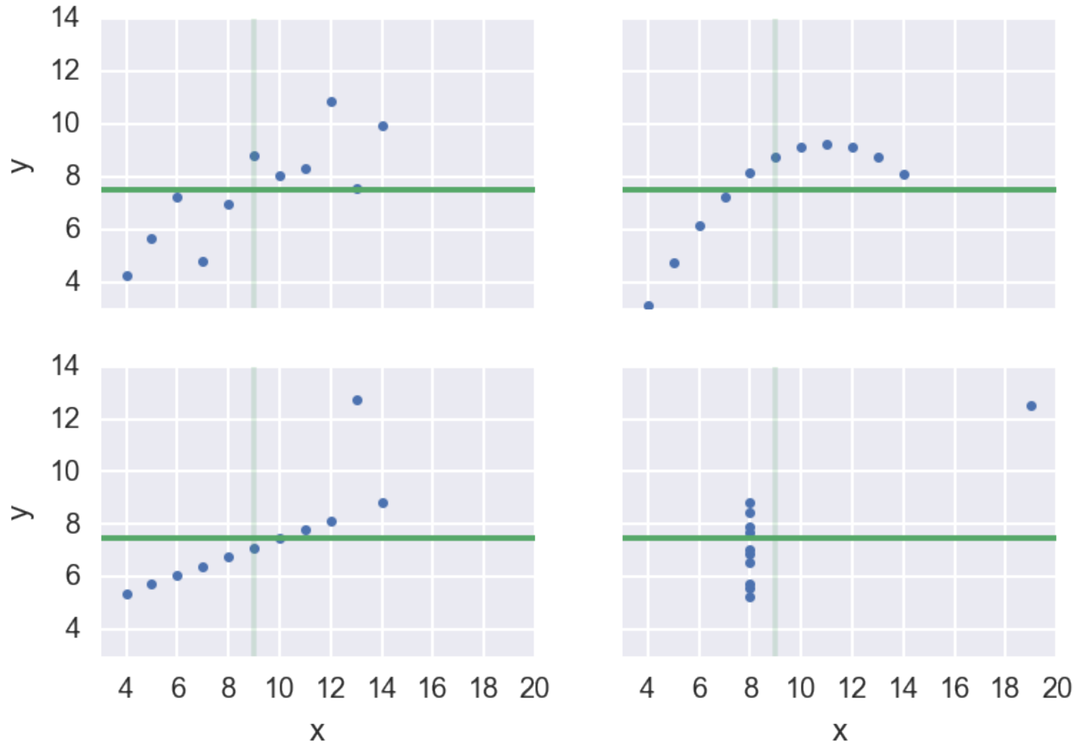
Data: Anscombe, The American Statistician, *1973*

Anscombe's quartet



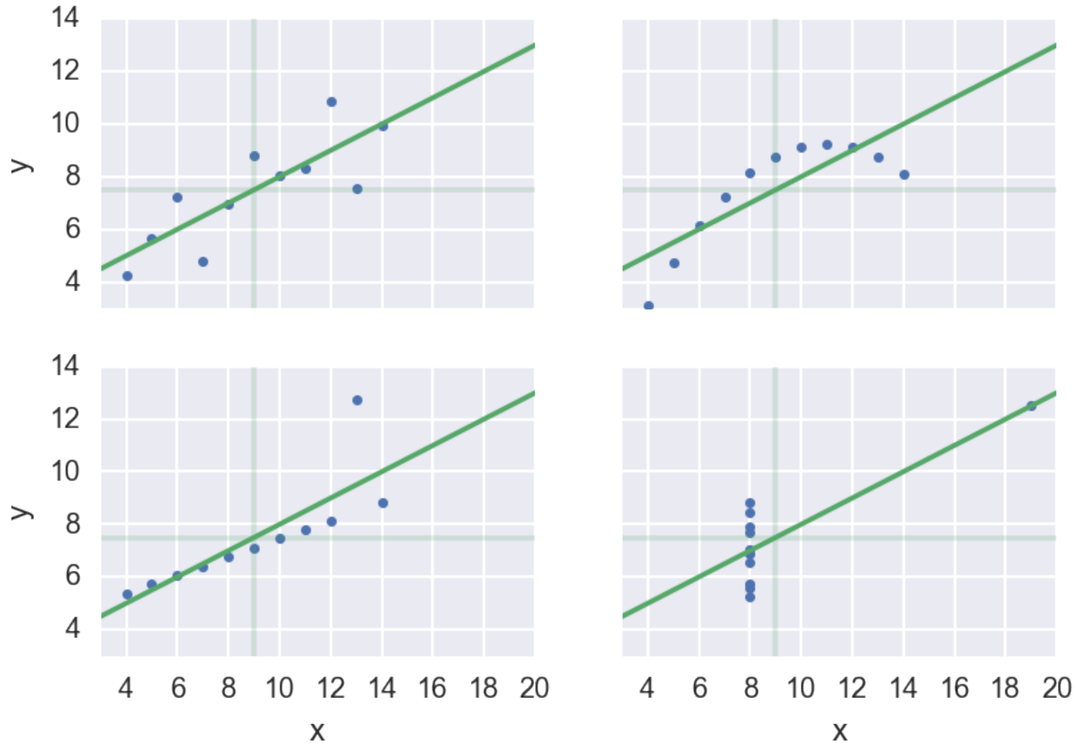
Data: Anscombe, The American Statistician, *1973*

Anscombe's quartet



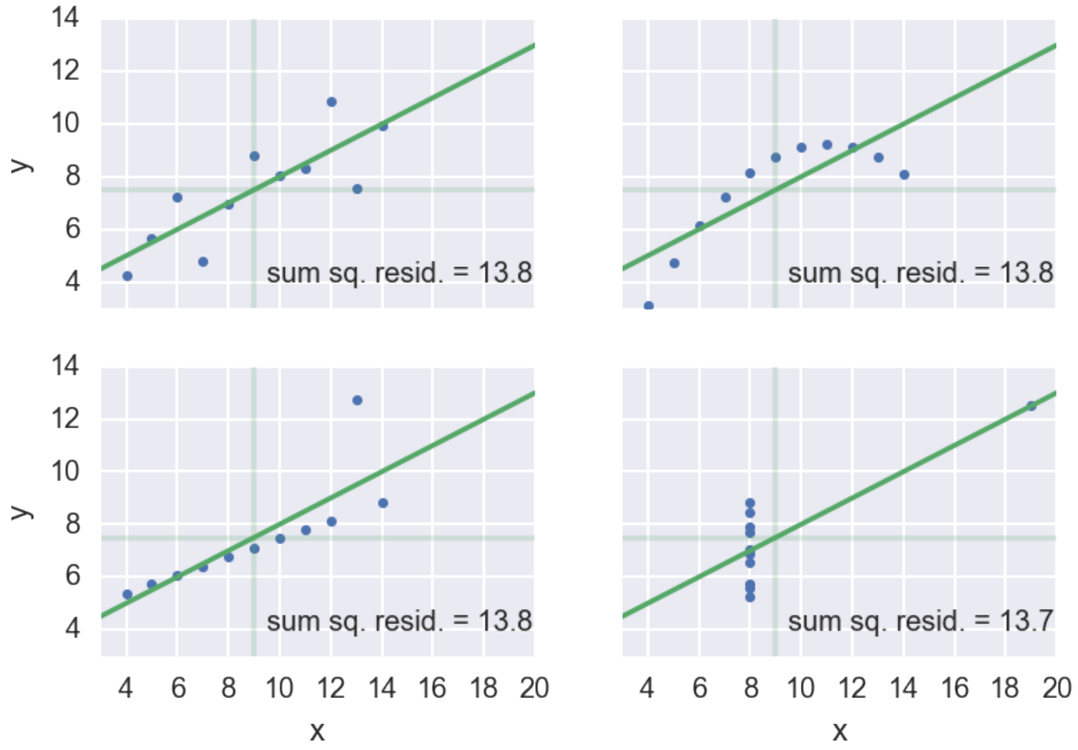
Data: Anscombe, The American Statistician, *1973*

Anscombe's quartet



Data: Anscombe, The American Statistician, *1973*

Anscombe's quartet

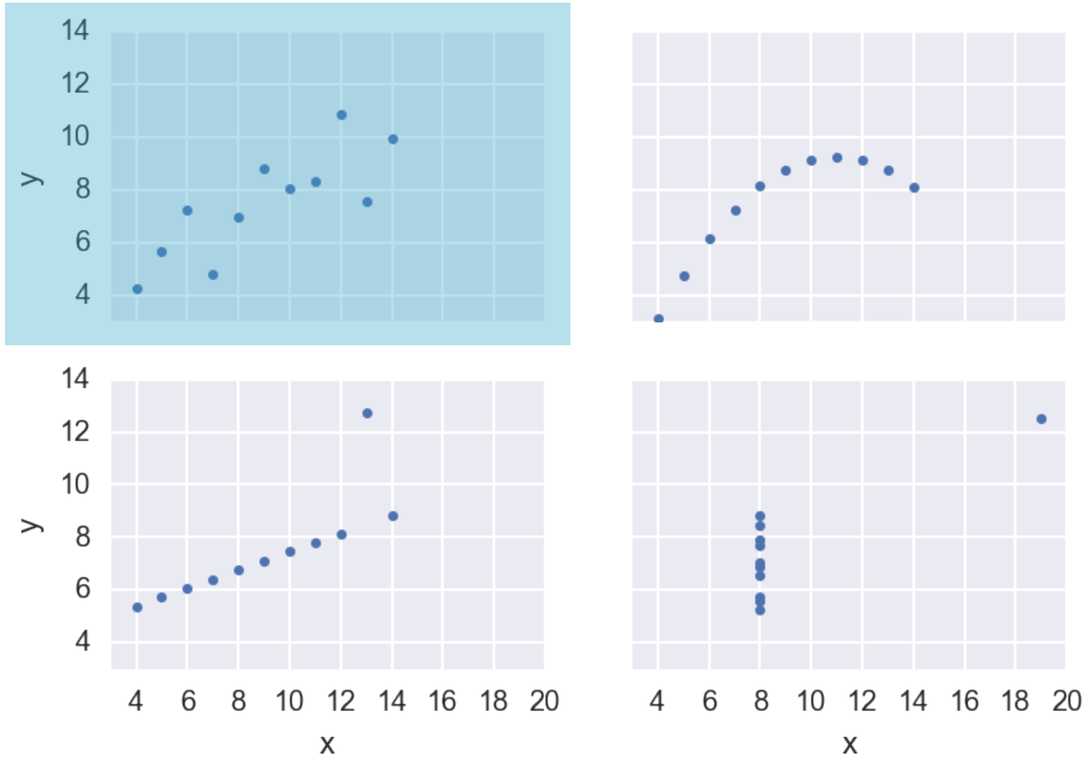


Data: Anscombe, The American Statistician, *1973*

Look before you leap!

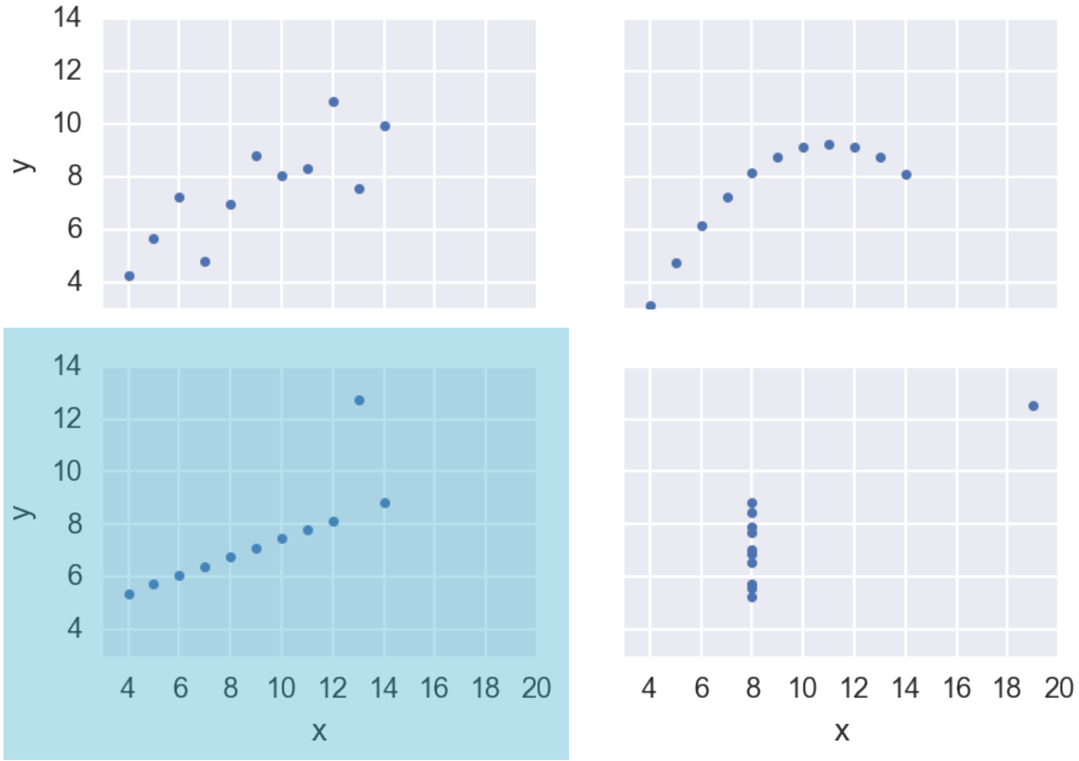
- Do graphical EDA first

Anscombe's quartet



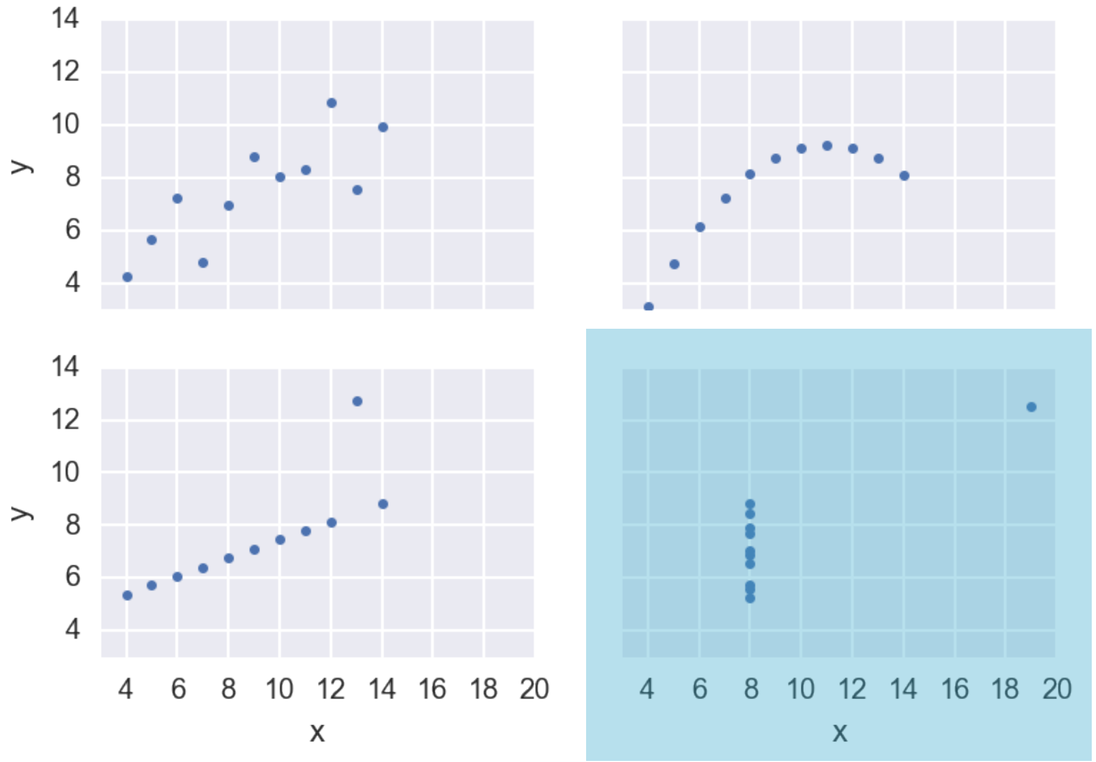
Data: Anscombe, The American Statistician, *1973*

Anscombe's quartet



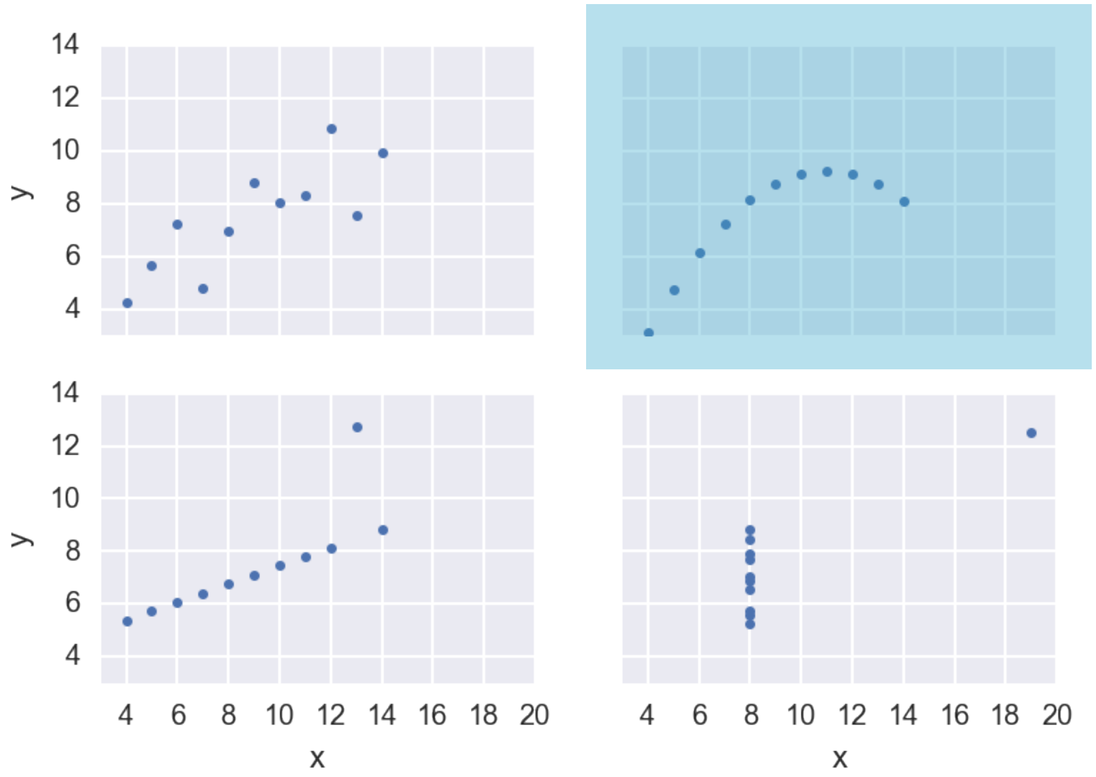
Data: Anscombe, The American Statistician, *1973*

Anscombe's quartet



Data: Anscombe, The American Statistician, *1973*

Anscombe's quartet



Data: Anscombe, The American Statistician, *1973*

STATISTICAL THINKING IN PYTHON II

Let’s practice!