13.1) Variability

Vitor Kamada

December 2019

Reference

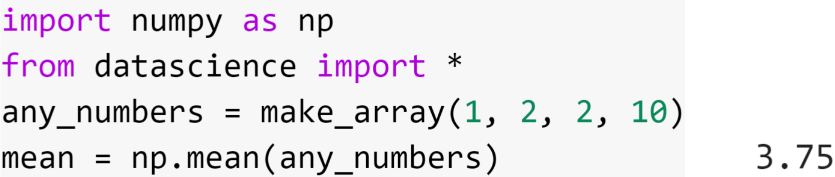
Tables, Graphics, and Figures from

Computational and Inferential Thinking: The Foundations of Data Science

Adhikari & DeNero (2019): Ch 14.2 Variability

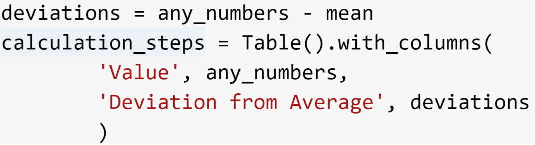
https://www.inferentialthinking.com/

Deviations from Average

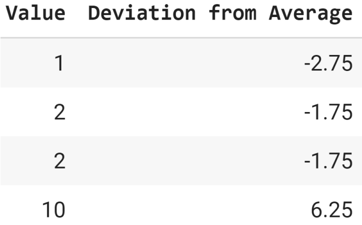


```
import numpy as np
from datascience import *
any_numbers = make_array(1, 2, 2, 10)
mean = np.mean(any_numbers)
```

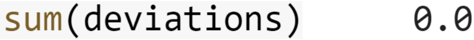
```
3.75
```



```
deviations = any_numbers - mean
calculation_steps = Table().with_columns(
        'Value', any_numbers,
        'Deviation from Average', deviations
        )
```



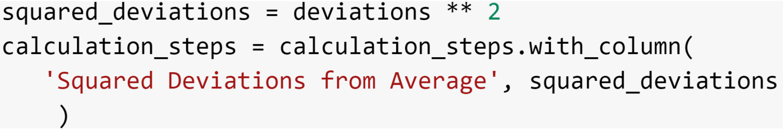
Value	Deviation from Average
1	-2.75
2	-1.75
2	-1.75
10	6.25



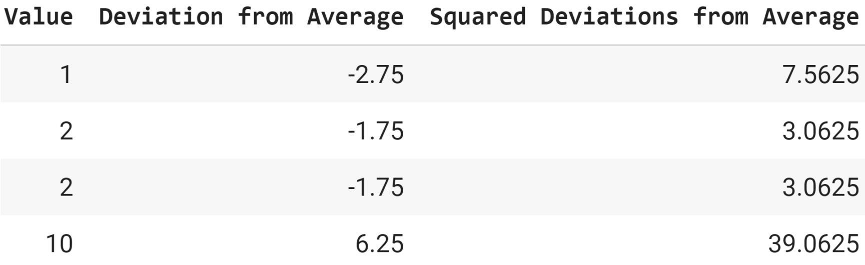
```
sum(deviations)
```

```
0.0
```

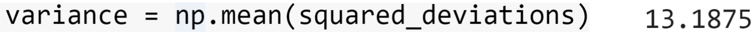
Variance and Standard Deviation (SD)



```python
squared_deviations = deviations ** 2
calculation_steps = calculation_steps.with_column(
   'Squared Deviations from Average', squared_deviations
    )
```

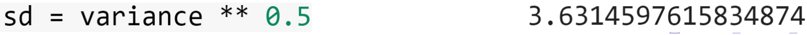


Value	Deviation from Average	Squared Deviations from Average
1	-2.75	7.5625
2	-1.75	3.0625
2	-1.75	3.0625
10	6.25	39.0625



```python
variance = np.mean(squared_deviations)
```

```
13.1875
```

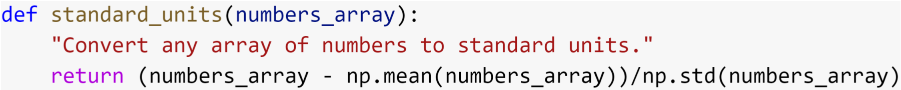


```python
sd = variance ** 0.5
```

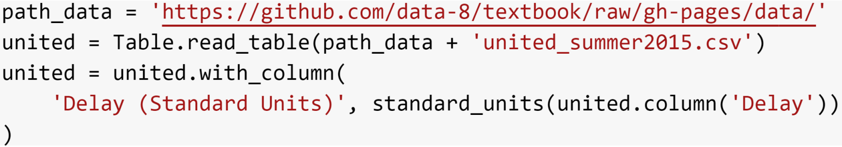
```
3.6314597615834874
```

Standard Units

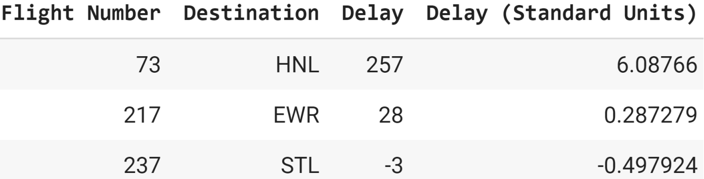
$$z = \frac{value - average}{SD}$$



```python
def standard_units(numbers_array):
    "Convert any array of numbers to standard units."
    return (numbers_array - np.mean(numbers_array))/np.std(numbers_array)
```



```python
path_data = 'https://github.com/data-8/textbook/raw/gh-pages/data/'
united = Table.read_table(path_data + 'united_summer2015.csv')
united = united.with_column(
    'Delay (Standard Units)', standard_units(united.column('Delay'))
)
```



Flight Number	Destination	Delay	Delay (Standard Units)
73	HNL	257	6.08766
217	EWR	28	0.287279
237	STL	-3	-0.497924

Chebychev's Bounds

For all numbers z, the proportion of entries that are in the range "average $\pm z$ SDs" is at least $1 - \frac{1}{z^2}$

"average $\pm$ 2 SDs" is at least 1 - $1/4 = 0.75$

"average $\pm$ 3 SDs" is at least 1 - $1/9 \approx 0.89$

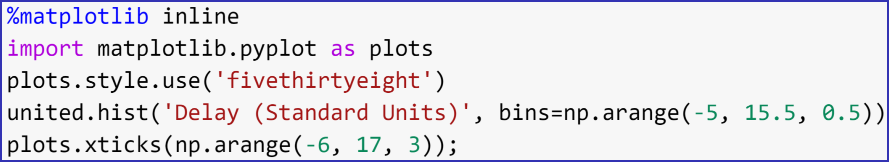
"average $\pm$ 4.5 SDs" is at least 1 - $1/$ $4.52 \approx 0.95$

```
united.sort('Delay', descending=True)
```

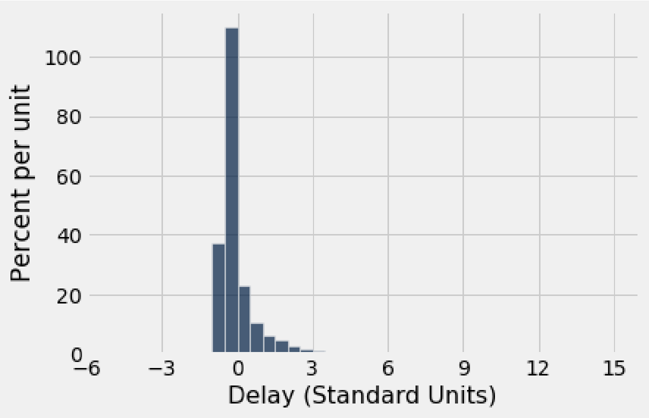
Flight Number	Destination	Delay	Delay (Standard Units)
1964	SEA	580	14.269
300	HNL	537	13.1798
1149	IAD	508	12.4453

```
within_3_sd = united.where('Delay (Standard Units)',
                           are.between(-3, 3))
within_3_sd.num_rows/united.num_rows
```

```
0.9790235081374322
```



```
%matplotlib inline
import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
united.hist('Delay (Standard Units)', bins=np.arange(-5, 15.5, 0.5))
plots.xticks(np.arange(-6, 17, 3));
```


Percent per unit
Delay (Standard Units)