Descriptive Statistics and Graphs

BRIEFS AND COMMENTS

Summarize Data by Numbers

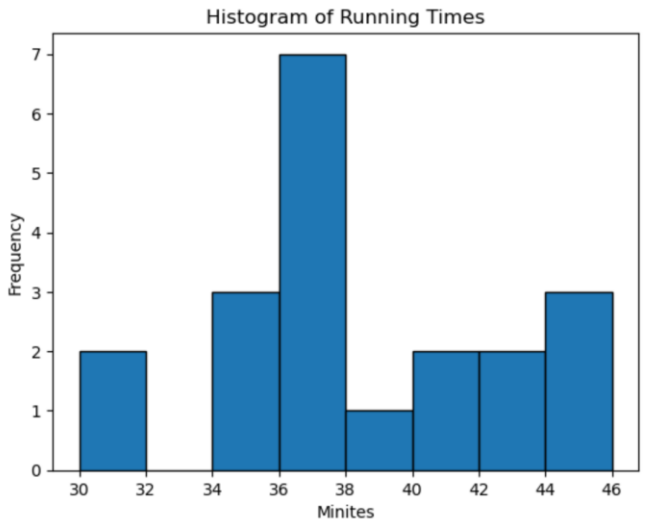
- Quantitative Variables
  - Mean and Median (Center of a Distribution)
  - Standard Deviation and Quartiles with Extremes (Spread of a Distribution)
  - Percentiles (If you need a more thorough examination of a distribution)
- Categorical Variables
  - Frequency and Relative Frequency

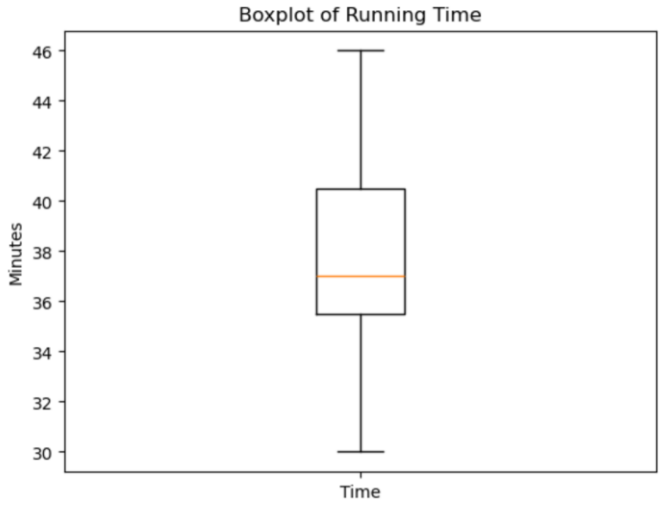
Remarks

- Mean and standard deviation are arithmetic-based statistics.
- Mean and standard deviation are often used to summarize symmetric distributions.
- Mean and standard deviation are often used in parametric methods.

- Median, quartiles, and extremes are location-based (rank-based) statistics.
- Median and quartiles are often used to summarize asymmetric distributions.
- Median is often used in non-parametric methods.

Example: 5K Race





	Time
count	20.000000
mean	37.850000
std	4.295346
min	30.000000
25%	35.500000
50%	37.000000
75%	40.500000
max	46.000000

	Time
1	44.0
2	37.0
3	40.0
4	46.0
5	44.0
6	30.0
7	40.0
8	34.0
9	34.0
10	37.0
11	36.0
12	42.0
13	39.0
14	36.0
15	37.0
16	37.0
17	42.0
18	34.0
19	37.0
20	31.0

The **sample variance** is the quantity:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Sample Standard Deviation

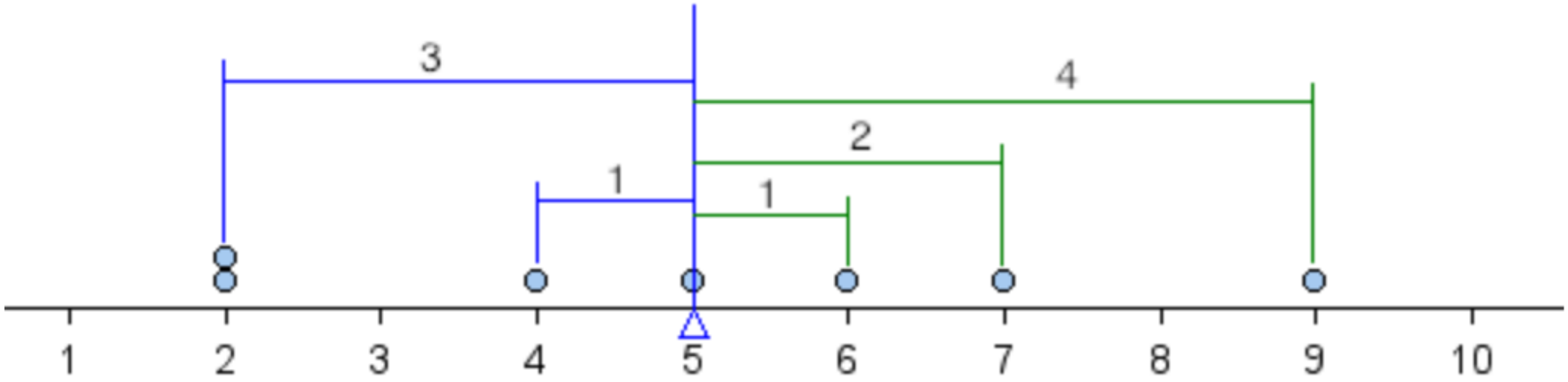
The **sample standard deviation** is the *positive* square root of the variance:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} = \sqrt{s^2}$$

Understand Standard Deviation

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

```
> x = c(2,2,4,5,6,7,9)
> xbar = mean(x)
> xbar
[1] 5
>
> x-xbar
[1] -3 -3 -1  0  1  2  4
> sd(x)
[1] 2.581989
```



https://pressbooks.cuny.edu/conceptsinstatistics/chapter/standard-deviation-1-of-4-concepts-in-statistics/

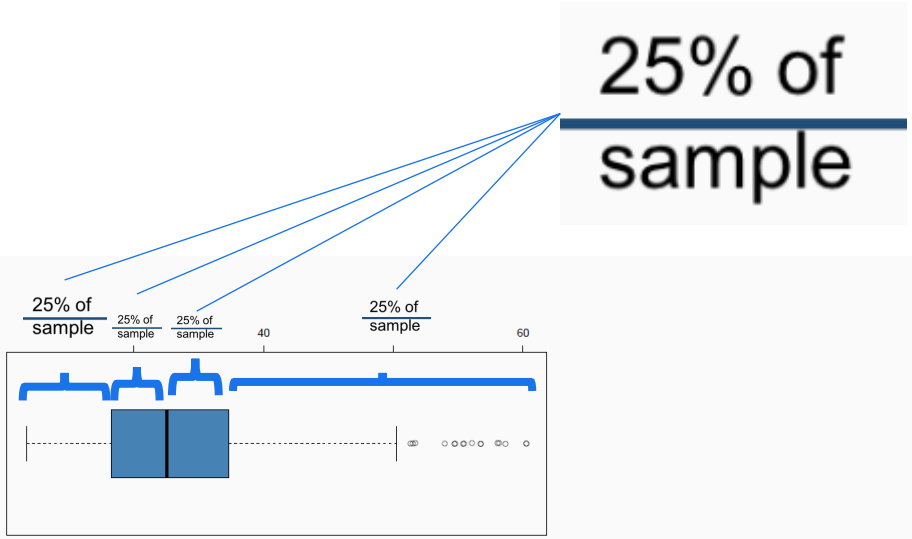
Remarks

- The sample standard deviation provides an assessment of roughly the average deviation from observed values to the mean (center of the distribution).
- The population standard deviation has a slightly different formula.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

- The sample standard deviation is an unbiased estimate of the population standard deviation if the sample represents the population.

Making a Boxplot



The **interquartile range (IQR)** is the range of the middle 50% of the data

$$\text{IQR} = \text{Q3} - \text{Q1}$$

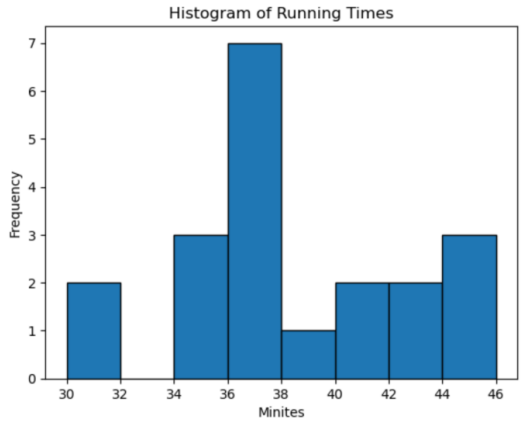
Definition of **outliers** for boxplots:

- Any point larger than Q3 + 1.5IQR
- Any point smaller than Q1 - 1.5IQR

Recall that Q1=1st Quartile, Q3=3rd Quartile, IQR=Q3-Q1

- **Center line** at median
- **Box** defined by 1st and 3rd quartile
- *IQR* (interquartile range) is the difference between third and first quartile
- *Whiskers* at largest/smallest data value within 1.5*IQR* of third/first quartile
- Outliers plotted individually for data exceeding whiskers
- Can be vertical or horizontal

Making a Histogram



Python

bins : ***int or sequence or str, default:*** `rcParams["hist.bins"]` ***(default:*** `10`***)***

If *bins* is an integer, it defines the number of equal-width bins in the range.

If *bins* is a sequence, it defines the bin edges, including the left edge of the first bin and the right edge of the last bin; in this case, bins may be unequally spaced. All but the last (righthand-most) bin is half-open. In other words, if *bins* is:

```
[1, 2, 3, 4]
```

then the first bin is `[1, 2)` (including 1, but excluding 2) and the second `[2, 3)`. The last bin, however, is `[3, 4]`, which *includes* 4.

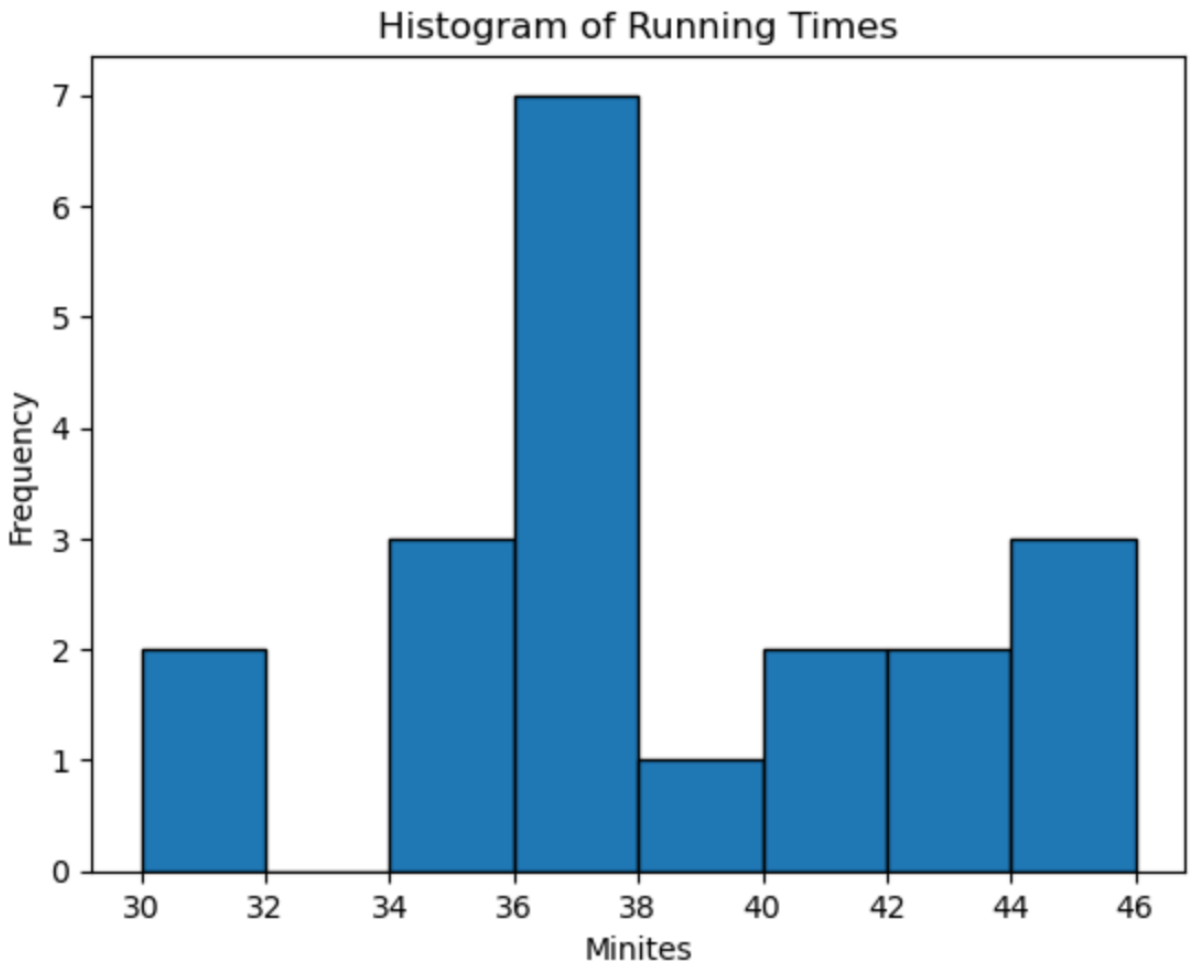
If *bins* is a string, it is one of the binning strategies supported by `numpy.histogram_bin_edges`: 'auto', 'fd', 'doane', 'scott', 'stone', 'rice', 'sturges', or 'sqrt'.

R

right logical; if `TRUE`, the histogram cells are right-closed (left open) intervals.

I know it might sound confusing, but let's pretend for a moment that we don't know the actual values of running times; what we have is only the histogram.

Additional Note for Doing HW



- Q1 occurs around 36.
- Median occurs in the interval [36, 38).
- Q3 occurs around 42.

The bins here are [30,32), [32,34), [34,36), [36,38), [38,40), [40,42), [42, 44), [44,46].

Let's see where the quartiles are

```
> sort(Time)
 [1] 30 31 34 34 34 36 36 37 37 37 37 37 39 40 40 42 42 44 44 46
```

Remarks

Please note that there are several methods to compute the quartiles, specifically the first and third quartiles. For example, the following shows how the argument "type=" from R in the function quantile() controls the choice of different methods.

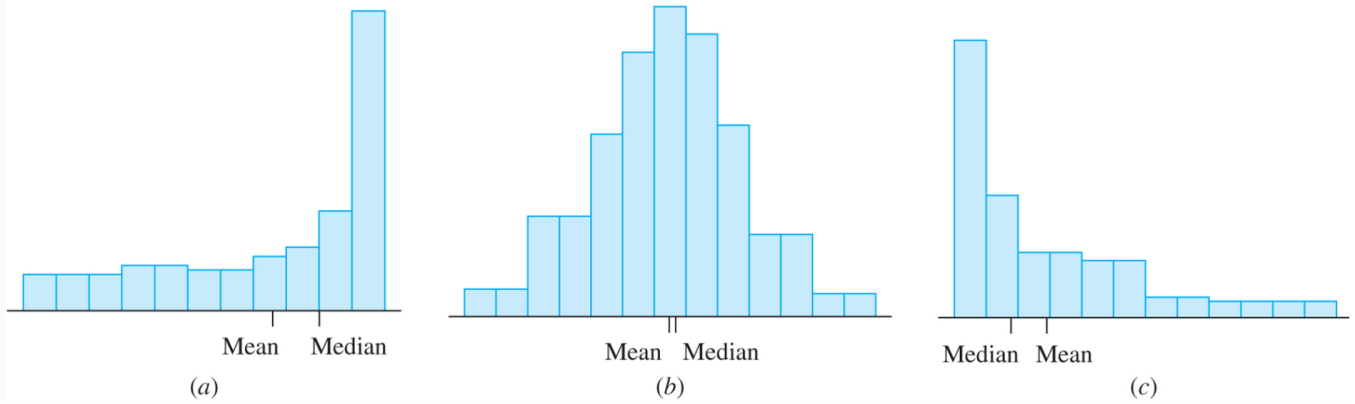
type	an integer between 1 and 9 selecting one of the nine quantile algorithms detailed below to be used.

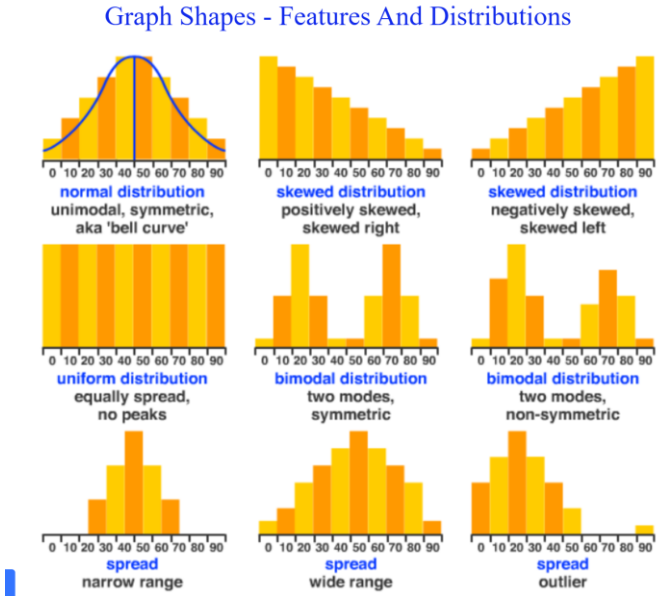
https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/quantile

When calculating the quartiles in Python or R, please use the default setting unless you have a special reason to change it.

Shapes of Distributions - Histogram

- For symmetric data, the sample mean and median are approximately equal
- For Right skewed data, the mean is greater than the median
- For Left skewed data, the mean is less than the median



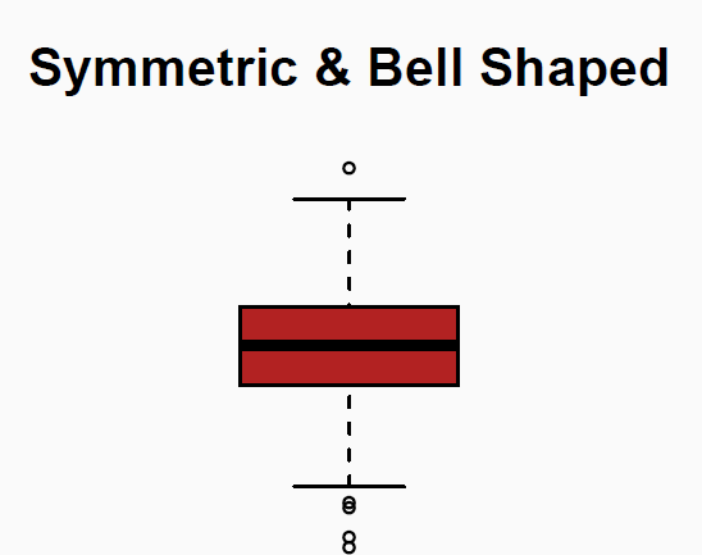


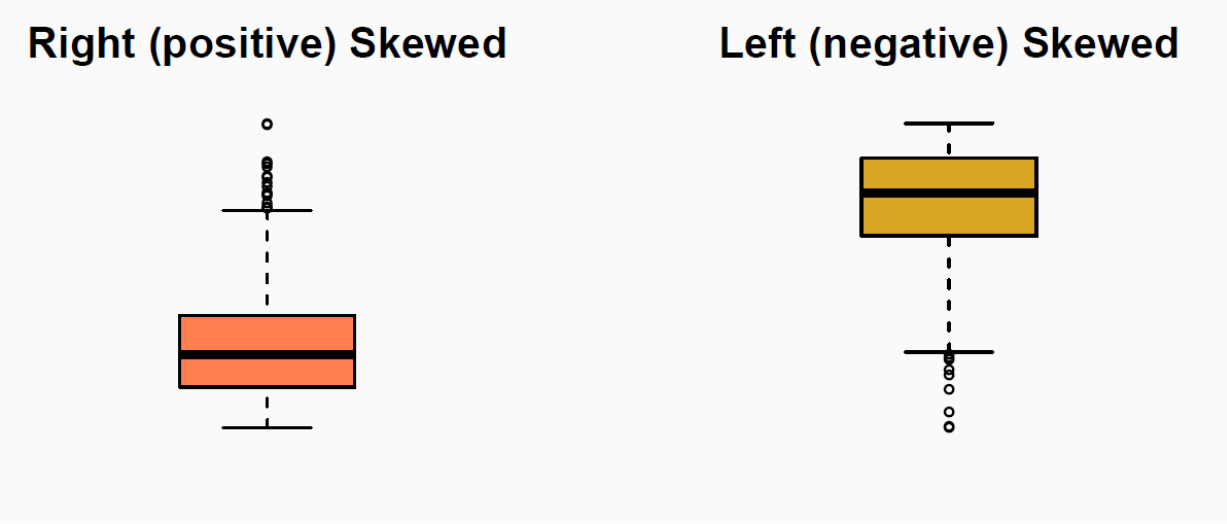
If you want to learn more, you can refer to this post:

https://mathtec.weebly.com/graph-shapes.html

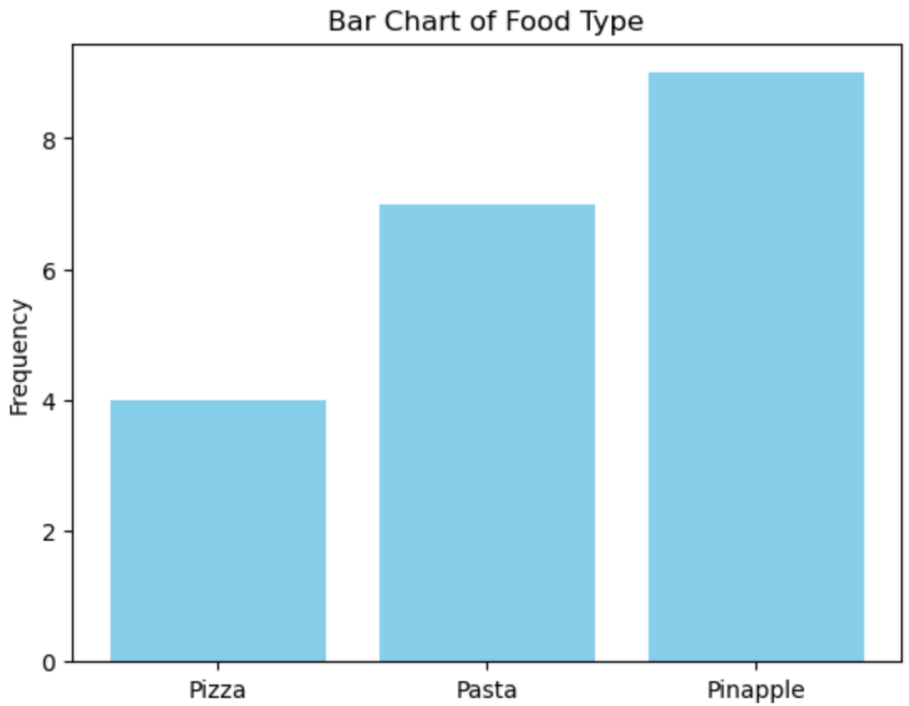
This is not required materials.

Shapes of Distributions - Boxplot





Example: Food Choice After Running




```
Food
Pinapple    9
Pasta       7
Pizza       4
Name: count, dtype: int64
Food
Pinapple    0.45
Pasta       0.35
Pizza       0.20
```

	Food
1	Pinapple
2	Pinapple
3	Pasta
4	Pizza
5	Pasta
6	Pasta
7	Pinapple
8	Pizza
9	Pasta
10	Pasta
11	Pinapple
12	Pasta
13	Pizza
14	Pinapple
15	Pasta
16	Pizza
17	Pinapple
18	Pinapple
19	Pinapple
20	Pinapple

Remarks

- You cannot summarise the shape of the distribution of a categorical variable using its bar chart. Because the categories can be rearranged. Then, the overall shape is changed.
- The best way to see the pattern from a bar chart is to arrange the bars (categories) using the ascending or descending order of their frequencies or relative frequencies.

Summary

- Commonly Used Descriptive Statistics
- Commonly Used Graphs.

Quote of the Day

Doctrine and Covenants 45:62

For verily I say unto you, that great things await you;

Typos?

- Please email me any typos and errors you found in this lesson.
- Thank you!