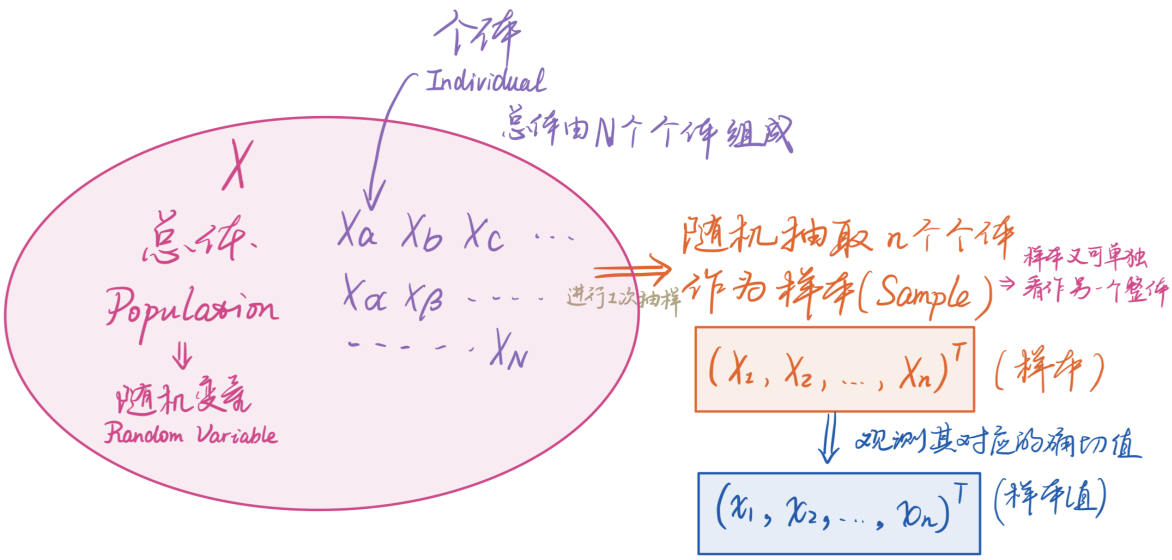
Python Statistical Analysis

There are some fundamental symbol and concept need to know first.

$$\text{Population} \quad \underset{estimate}{\overset{sampling}{\rightleftharpoons}} \quad \text{Sample}$$



Statistic (统计量)	Population (总体) - N 个	Sample (样本) - n 个
Mean	μ	$\bar{X} = \hat{\mu}$
Variance	σ^2	$s^2 = \hat{\sigma}^2$
Proportion	π	p
Correlative Coefficience	ρ	r

```
import numpy as np

# Convert multiple-rows data to the MATRIX of COLUMN VECTORS
np.matrix([<data_1>, <data_2>,...]).T

> input:  data_1 = [0, 1, 2]
          data_2 = [9, 8, 7]
          print(np.matrix([data_1, data_2]))  # Left
          print(np.matrix([data_1, data_2]).T)  # Right
> output: +---+---+---+         +---+---+
          | 0 | 1 | 2 |         | 0 | 9 |
          +---+---+---+         +---+---+
          | 9 | 8 | 7 |         | 1 | 8 |
```

```
          +---+---+---+         +---+---+
                                | 2 | 7 |
                                +---+---+


# Ravel the 2-D data to 1-D data
<data>.ravel

> input:  m = np.array([[1, 2, 3],
                        [0, 1, 2]])
          print(m.mean(axis=0))
          print(m.mean(axis=0).ravel())
> output: [[0.5, 1.5, 2.5]]
          [0.5, 1.5, 2.5]
```

1. Data Descriptive Analysis · 描述性统计分析

1.1. Descriptive Statistics · 描述性统计量

1.1.1. Measures of Location and Dispersion · 位置和分散程度的度量

There are 8 important statistics uesd to measure the location and dispersion, which are `Mean`, `Median`, `Percentile`, `Variation`, `Standard Deviataion`, `Standard Error of Mean`, `Coefficience of Variation`, and `Range`.

1. **Mean (均值, $\mathbb{E}[X] = \bar{X} = \langle X \rangle$)**

   The average of all entries.

$$
\begin{aligned}
\mu &= \frac{1}{N}\sum_{i=1}^{N} X_i \\
\hat{\mu} &= \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i
\end{aligned} \tag{1.1}
$$

> *e.g.*
>
> $$
> \begin{aligned}
> X &= \{1, 2, 3, 4, 5\} \\
> \bar{X} &= \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3
> \end{aligned}
> $$

```
import scipt.stats as st

# 1. Use numpy
#   axis = 0    → column-wise
#        = 1    → row-wise
#        = none → all entries
np.mean(<data>, axis=)
<ndarray-like data>.mean()

# 2. Use scipy.stats
#   Compute the mean of the closed interval
st.tmean(<data>, <interval as (a, b)>)
```

2. **Median (中位数)**

   The middle number of all entries. Data must be sorted in **ascending** order *(no need to sort first with Python)*.

   In `Symmetric Distribution` (like t-distribution and normal-distribution, 对称分布), the median is very close to the mean; while in `Skewed Distribution` (like F-distribution, 偏态分布), the difference between median and mean is relatively large.

$$m_e = \begin{cases} x_{\frac{N+1}{2}} & \text{, if } N \text{ is odd} \\ \frac{1}{2}\left(x_{\frac{N}{2}} + x_{\frac{N+1}{2}}\right) & \text{, if } N \text{ is even} \end{cases} \tag{1.2}$$

> *e.g.*
>
> $$\begin{aligned} X &= \{1,2,3,4,5,6\} \\ m_e &= \frac{1}{2}(x_3 + x_4) = \frac{3+4}{2} = 3.5 \end{aligned}$$

```python
# reverse: False by default (desending)
sorted(<data>, reverse=True)  # Order data by ascending

np.median(<data>)
```

3. **Percentile (百分位数)**

   Represent specific-location data. Data must be sorted in **ascending** order *(no need to sort first with Python)*. The symbol $\lceil\rceil$ means rounded up (*e.g.*, $\lceil 2.1 \rceil = \lceil 2.9 \rceil = 3$).

$$m_p = \begin{cases} x_{\lceil N \times perc \rceil} & \text{, if } N \times perc \notin \mathbb{N}^+ \\ \frac{1}{2}(x_{N \times perc} + x_{N \times perc + 1}) & \text{, if } N \times perc \in \mathbb{N}^+ \end{cases} \tag{1.3}$$

> *e.g.1*
>
> $$\begin{aligned} X &= \{1,2,3,4,5\} \\ m_{50\%} &= x_{\lceil 5 \times 50\% \rceil} = x_{\lceil 2.5 \rceil} = x_3 = 3 \end{aligned}$$
>
> *e.g.2*
>
> $$\begin{aligned} X &= \{1,2,3,4,5,6\} \\ m_{50\%} &= \frac{(x_{50\% \times 6} + x_{50\% \times 6 + 1})}{2} = \frac{(x_3 + x_4)}{2} = \frac{3+4}{2} = 3.5 \end{aligned}$$

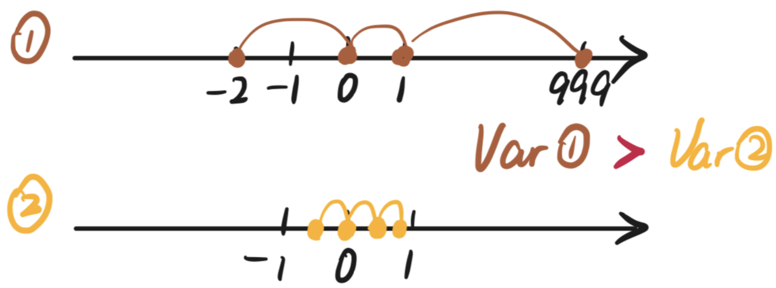
```python
np.quantile(<data>, <percentile>)
```

4. **Variance (Var, 方差, $\mathrm{Var}[X]$)**

   Measure the dispersion of data, namely the degree of the samples dispersion from the mean. **The more concentrated the data, the smaller the variance, vice versa.**

$$
\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2
$$

$$
\begin{cases}
\hat{\sigma}^2_{biased} = s^2_{biased} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 \\
\hat{\sigma}^2_{unbiased} = s^2_{unbiased} = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2
\end{cases} \tag{1.4}
$$

> *e.g.*, The following picture can easily figure the relationship between *variance* and *dispersion* of data.
>
> 
>

```
# Biased Estimator of variance
np.var(<data>)

# Unbiased Estimator of variance
st.tvar(<data>)
```

There are two kinds of estimation: `Unbiased Estimation` (无偏估计) and `Biased Estimation` (有偏估计).

If we want to discern an unknown distribution (*i.e.*, μ, σ^2 are all unknown), we need to take some samples to estimate its $\{\mu, \sigma^2\}$. Mean can be easily compute by (1.1), but variance is a little complex for which refers the conception of the `Freedom Degree` (自由度).

Consider an extreme case that we **take only one sample** $X = \{X_1\}$, so we can easily estimate that

$$
\text{the population mean estimator } \hat{\mu} = \text{the sample mean } \bar{X} = \frac{X_1}{1}
$$

> `Estimator`: an estimation **function** of samples (*i.e.*, $\bar{X}$).
>
> `Estimate`: a **certain value** compute by substitute one sample to the estimator (*i.e.*, $\frac{1+2}{2} = 1.5$).

But when we want to estimate the population variance σ^2 by the sample variance $s^2 = \hat{\sigma}^2$ with $\hat{\mu}$, we will find something counterintuitive:

- if we compute $\hat{\sigma}^2$ by divide n intuitively, we will get $\hat{\sigma}^2 = \frac{(X_1-X_1)^2}{1} = 0$ which is counterintuitively, because the only one element does not exist $\hat{\sigma}^2$, so we cannot compute $\hat{\sigma}^2$ by only one element;

- while if we compute $\hat{\sigma}^2$ by divide $(n-1)$ counterintuitively, we will get $\hat{\sigma}^2 = \frac{(X_1-X_1)^2}{1-1} = \frac{0}{0}$ which is cannot be computed but intuitively. This means that although the numerator looks like having an item $(X_1 - X_1)^2$, the information it contains is actually 0 (*i.e.*, freedom degree $df = 0$).

Expand this case to two samples: when we have two samples $X = \{X_1, X_2\}$, we can get $\hat{\mu} = \bar{X} = \frac{X_1+X_2}{2}$, $\hat{\sigma}^2 = s^2 = \frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2}{2}$. However actually, we can find that $(X_1 - \bar{X}) = -(X_2 - \bar{X})$, namely one of them **is not free**, so the information it contains is actually 1 (*i.e.*, freedom degree $df = 1$). **By analogy, when we have n samples, we get its$df = n - 1$**, which means we only calculated $(n-1)$ deviances from the population, so that:

- To **estimate the population variance σ^2 closest**, we **calculate the sample variance s^2 by divide** $(n-1)$, which named **unbiased estimator of the population variance $\hat{\sigma}^2$**, where $\mathbb{E}\left(s^2\right) = \sigma^2$;
- And if we calculate s^2 by n, we will get **biased estimator of the population variance $\hat{\sigma}^2$**, where $\mathbb{E}\left(s^2\right) \neq \sigma^2$.

5. **Standard Deviation (SD, 标准差)**

   The square root of variation.

$$
\begin{gathered}
\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2} \\
\begin{cases}
\hat{\sigma}_{biased} = s_{biased} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \\
\hat{\sigma}_{unbiased} = s_{unbiased} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}
\end{cases}
\end{gathered}
\tag{1.5}
$$

```
# Biased Estimator of standard deviation
np.std(<data>)

# Unbiased Estimator of standard deviation
st.tstd(<data>)
```

6. **Standard Error of the Mean (SEM, 标准误)**

   Measure the dispersion of the sample mean for every sample, and the deviation of *the population mean* μ and *the sample mean* $\bar{X}$. **The smaller the SEM, the smaller the deviation between μ and $\bar{X}$, vice versa.**

$$
\hat{\sigma}_{SEM} = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}\left(\bar{X}_j - \bar{\bar{X}}\right)} \quad \cong \quad \frac{s_{j\_unbiased}}{\sqrt{n}} \tag{1.6}
$$

   where $i = \{1, \ldots, n\}$ means the number of entries for every sample, $j = \{1, \ldots, m\}$ means the number of samples. In general, we sample one time and use *this sample SD* divided by $\sqrt{n}$, as shown on the right of (1.6).

> *e.g.*, Suppose we sample three times, every sample have four entries. The data are shown in the table below.
>
> | **sample** j **entry** i | **A** | **B** | **C** |
> |---|---|---|---|
> | **1** | 0 | 0 | 0 |
> | **2** | 1 | 2 | 4 |
> | **3** | 2 | 4 | 8 |
> | **4** | 3 | 6 | 12 |
> | $\bar{X}_j$ | **1.5** | **3** | **6** |
> | $s_{j\_unbiased}$ | **1.29** | **2.58** | **5.16** |
>
> $$\bar{\bar{X}} = \frac{1.5+3+6}{3} = 3.5$$
>
> $$\hat{\sigma}_m = \sqrt{\frac{(1.5-3.5)^2+(3-3.5)^2+(6-3.5)^2}{2}} \approx 2.29 \quad \cong \quad \frac{1.29}{\sqrt{4}} = 0.645$$

```
# Calculate the SEM by hand
sem = st.tstd(<data>) / np.sqrt(<the number of entries for this sample>)
```

7. **Coefficience of Variation (CV, 变异系数)**

   Measure the dispersion of data by eliminate the effects of *dimension (尺度)* and *scale (量纲)*.

$$CV = \frac{\sigma}{\mu} \tag{1.7}$$

```
# Calculate the CV by hand
cv = st.tstd(<data>) / np.mean(<data>)
```

8. **Range (极差, 全距)**

   The longest length of data.

$$\text{Range} = \text{max - min} \tag{1.8}$$

```
# Calculate the Range by hand
Range = np.max(<data>) - np.min(<data>)
```

1.1.2. Measures of Relationships · 关系度量

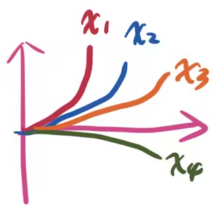
1. **Variance-Covariance Matrix (方差-协方差矩阵) & Correlation Coefficient Matrix (相关系数矩阵)**

   `Covariance`: Measure the changing trends between the 2 or more variables. This matrix is symmetric along the diagonal.

$$
\begin{cases}
\hat{\sigma}_{xx}^2 = s_{xx}^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \mathrm{Var}(X) \\
\hat{\sigma}_{yy}^2 = s_{yy}^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \mathrm{Var}(Y) \\
\hat{\sigma}_{xy}^2 = s_{xy}^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = s_{yx}^2 = \hat{\sigma}_{yx}^2
\end{cases} \tag{1.9}
$$

   `Correlation Coefficient`: Normalize (归一化) the covariance to measure the degree of linear correlation between 2 or more variables.

$$
\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \in [-1, 1] \tag{1.10}
$$

> The meaning of ρ is as follows
>
> 
>
>
> and the visualization is as follows
>
> | | x_1 | x_2 | x_3 | x_4 |
> |---|---|---|---|---|
> | x_1 | 1 | 0.8 | 0.6 | -0.2 |
> | x_2 | 0.8 | 1 | 0.4 | -0.4 |
> | x_3 | 0.6 | 0.4 | 1 | -0.6 |
> | x_4 | -0.2 | -0.4 | -0.6 | 1 |
>
> 
>

```
# Variance-Covariance Matrix
np.cov(<data>)  # The data form must be a ROW VECTOR


# Correlation Coefficient Matrix
# 1. Use Numpy
np.corrcoef(<data>)  # The data form must be a ROW VECTOR

# 2. Use Pandas
pd.DataFrame.corr(<data>)  # The data form must be a ROW VECTOR
```

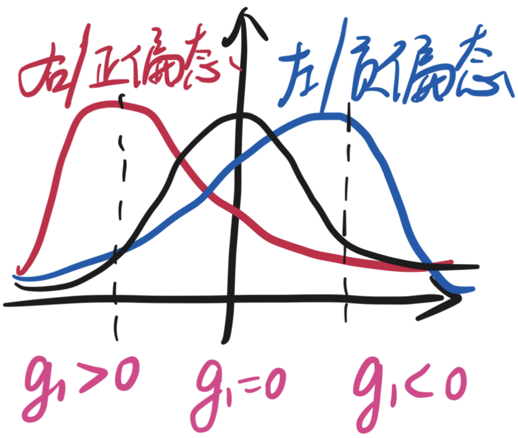
1.1.3. Measures of Distribution Shapes · 分布形状的度量

1. **Skewness (偏度)**

   Measure the degree of asymmetry in the data. The more the SK is to 0, the more the curve fits the normal distribution.

$$g_1 = SK(X) = \frac{n\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3}{(n-1)(n-2)s^3} = \frac{n^2\mu_3}{(n-1)(n-2)s^3}$$

   where s means the `Standard Deviatation`, μ_3 means `Third Order Central Moment` (三阶中心矩), $\mu_3 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3$.

> The distribution shapes under different skewness are as follows
>
> 
>

```python
# 1. Use Pandas, calculate the modified skewness (Unbiased Estimation)
data_s = pd.Series(<data>)
data_s.skew()

# 2. Use Scipy
#   bias = False —> Modified (Unbiased)
#        = True  —> Unmodified (Biased) (Default)
st.skew(<data>, bias=)
```

2. **Kurtosis (峰度)**

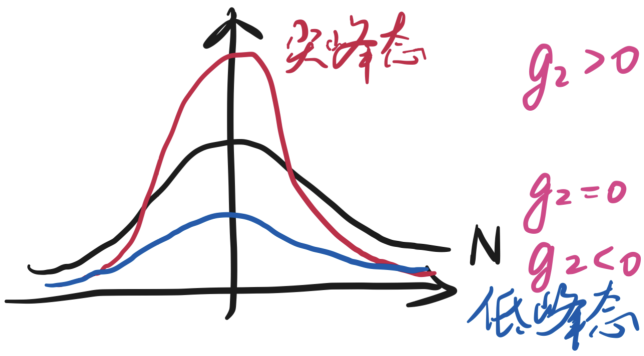
   Measure the sharpness of the `Probability Density Function` (PDF, 概率密度函数) at μ.

$$g_2 = K(X) = \frac{n(n+1)\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4}{(n-1)(n-2)(n-3)s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

$$= \frac{n^2(n+1)\mu^4}{(n-1)(n-2)(n-3)s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

   where s means the `Standard Deviatation`, μ_3 means `Fourth Order Central Moment` (四阶中心矩), $\mu_4 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4$.

> The distribution shapes under different kurtosis are as follows




```python
# 1. Use Pandas, calculate the modified kurtosis (Unbiased Estimation)
data_s = pd.Series(<data>)
data_s.kurt()

# 2. Use Scipy
#   bias = False → Modified (Unbiased)
#        = True  → Unmodified (Biased) (Default)
st.kurtosis(<data>, bias=)
```

1.1.4. Summary of Data Characteristics · 数据特性的总括

Data characteristics usually means the `Global Extremum`, `Mean`, `Unbiased Variance`, `Modified Skewness`, `Modified Kurtosis`, and `Distribution`. The *descriptive statistics* can be calculated by the above content, and the *distribution* can be tested by `Normality Test` (like `Shapiro Test`) or `Distribution Fit Test` (like `Kolmogorov-Smirnov Test`)

```python
# Check the decriptive statistics
st.describe(<data>, bias=False)

# Test the distribution
# 1. Normality Test
st.shapiro(<data>)
# 2. Distribution Fit Test
st.kstest(<data>, <name of the dist as "t">, <df of the dist as (3,)>)

# p.s. Generate an F-distribution with df=(2, 9)
st.f.rvs(size=, dfn=2, dfd=9)
```

1.2. Data Distribution · 数据分布

1.2.1. Fundamental Concepts · 基础概念

1. **Probability Space** (($\Omega, \mathcal{F}, P$),概率空间),
   Sample Space (Ω, 样本空间),
   Random Events ($A \in \mathcal{F}$, 随机事件),
   Random Variable (RV, 随机变量)

   asd

2. **Population (总体),**
   Individual (个体),
   Sample (样本),
   Parameter Space (参数空间),
   Distribution Family (分布族),
   Statistics (统计量),
   Sampling Distribution (抽样分布)

   asd

3. **Probability Distribution Function (Culmulative Distribution Function, CDF, 累积分布函数)**

   asd

4. **Discrete Random Variable (离散型随机变量)**

   asd

5. **Continuous Random Variable (连续型随机变量)**

   asd

6. **Mathematic Expectation ($\mathbb{E}(\cdot)$, 数学期望)**

   asd

7. **Moment (矩，可理解为一种距离)**

   asd

1.2.2. Discrete Probability Distribution · 离散型概率分布

1. **Binomial Distribution (二项分布)**

   asd

2. **Poisson Distribution (泊松分布)**

   asd

1.2.3. Continuous Probability Distribution · 连续型概率分布

1. **Normal Distribution (正态分布)**

   asd

2. **t-distribution (Student’s t distribution, t 分布)**

   asd

3. **Gamma Distribution (伽马分布)**

   asd

4. **Chi-Square Distribution (卡方分布)**

   asd

5. **F-distribution (F 分布)**

   asd

1.3. Histogram, Experience Distribution and QQ Graph · 直方图，经验分布函数与 QQ 图

1.3.1. Histogram and Kernel Density Estimation · 直方图与核密度估计

1.3.2. Experience Distribution · 经验分布函数

1.3.3. QQ Graph and Stem-and-Leaf Display · QQ 图与茎叶图

1.4. Multivariate Data · 多元数据

1.4.1. Numerical Features of Multivariate Data · 多元数据的数字特征

asd

1.4.2. Graphical Representation of Multivariate Data · 多元数据的图形表示

asd