第二章--数值型属性

任务目标

1、一元变量分析

2、二元变量分析

3、多元变量分析

4、数据规范化

5、正态分布

相关知识

本章讨论对数值型属性进行探索性数据分析的基本统计方法。衡量数据居中性（位置）的方法、数据离散度的衡量、线性相关的衡量以及属性之间的关联关系。

2.1 一元变量分析

一元变量分析每次聚焦一个属性。因此原数据矩阵D可以作为一个$n \times 1$的矩阵或一个简单的列向量。

$$D = \begin{pmatrix} \frac{X}{x_1} \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

其中X是数值属性，且$x_i \in \mathbb{R}$。假设X是一个随机变量，其中每一个$x_i (1 \leqslant i \leqslant n)$都当作一个恒等随机变量。假设观察到的数据是从$X$得到一个随机抽样，即每一个变量$x_i$都是和$X$独立同分布。可以将抽样数据看作一个$n$维的矩阵，写作$X \in \mathbb{R}^n$。

属性X的概率密度或概率质量函数$f(x)$与累积分布函数$F(x)$都是未知的。然而，可以直接从数据样本中估计出相关分布，可以通过计算统计量估计处若干重要的总体参数。

1. 经验累积分布函数

X的经验累积分布函数（empirical cumulative distribution function，经验CDF）如下式子：

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leqslant x)$$

其中，

$$I(x_i \leqslant x) = \begin{cases} 1 & x_i \leqslant x \\ 0 & x_i > x \end{cases}$$

是一个二值指示变量（binary indicator variable），判断给定的条件是否满足。为了计算经验CDF，计算对于每个$x \in R$，样本中有多少个点小于等于x。经验CDF给每个点x_i赋一个概率质量$\frac{1}{n}$。注意这里用$\hat{F}$来表示经验CDF事实上是对未知总体累积分布函数F的估计。

2. 逆累积分布函数

随机变量X的逆累积分布函数（inverse cumulative distribution function）或分位函数（quantile function）定义如下：

$$F^{-1}(q) = min\{x|F(x) \geqslant q\}, \quad q \in [0,1]$$

即逆CDF给出了X的最小值，其中比最小值更小的值比例为q，比最小值更大的值比例为$1-q$。经验逆累积分布函数（empirical inverse cumulative distribution function）$\hat{F}^{-1}$的公式。

3. 经验概率质量函数

X的经验概率质量函数（empirical probability mass function 经验PMF）如下式：

$$\hat{f}(x) = P(X = x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i = x)$$

其中，

$$I(x_i = x) = \begin{cases} 1 & x_i = x \\ 0 & x_i \neq x \end{cases}$$

经验PMF同样给每个点x_i赋一个概率质量$\frac{1}{n}$。

2.1.1 数据居中度度量

以下度量标识了概率质量的居中度、“中间”值等。

1. 均值

随机变量X的均值（mean）,也称作期望值（expected value），是X所有值的算术平均值。是一个表示X的分布所处的位置或居中度（central tendency）的值。

离散随机变量$X的均值或期望值定义如下：

$$\mu = E[X] = \sum_{x} xf(x)$$

其中$f(x)$是$X的概率质量函数。

连续随机变量X的均值或期望值定义如下：

$$\mu = E[X] = \int_{\infty}^{\infty} xf(x)dx$$

其中$f(x)$是X的概率密度函数。

样本均值 样本均值（sample mean）是一个统计量，即函数$\hat{\mu}:\{x_1, x_2, \cdots, x_n\} \to \mathbb{R}$可按下式定义为$x_i$的平均值：

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

样本均值是对未知的X均值μ的一个估计值。可以通过公式（2.4）中的经验概率质量函数$\hat{f}(x)$推导出来：

$$\hat{\mu} = \sum_{x} x\hat{f}(x) = \sum_{x}\left(\frac{1}{n}\sum_{i=1}^{n} I(x_i = x)\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

样本均值是无偏的若$E[\hat{\theta}]=\theta$，则称估计量$\hat{\theta}$是参数θ的一个无偏估计量（unbiased estimator）。样本均值$\hat{\mu}$是对总体均值μ的一个无偏估计，因为：

$$E[\hat{\mu}]=E\left[\frac{1}{n}\sum_{i=1}^{n}x_i\right]=\frac{1}{n}\sum_{i=1}^{n}E[x_i]=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu$$

利用一个条件是各随机变量x_i和X是独立同分布的，它们与X有着同样的均值μ，即对于所有的x_i都有$E[x_i]=\mu$。另一个条件是：期望函数E是一个线性算子，即对于任意两个随机变量X和Y及两个实数a和b，有$E[aX+bY]=aE[X]+bE[Y]$。

健壮性 如果一个统计量不受数据中极端值（如异常值）的影响，这个统计量是健壮的。样本均值是非健壮的，因为一个较大的值（异常值）就能够使均值偏移。一个更健壮的度量为切尾均值（trimmed mean）。该值是通过去掉两端的一小部分极端值后计算出来的。均值有时候会误导人们，因为均值不一定是在样本中出现的值，也不一定是随机变量能够取到的值（对于离散随机变量）。

2. 中位数

一个随机变量的中位数（median）可以定义为m，使得：

$$P(X\leqslant m)\geqslant\frac{1}{2}\text{且}P(X\geqslant m)\geqslant\frac{1}{2}$$

换句话说，中位数是"最中间"的值：X的一半取值大于m，另一半取值小于m。如果考虑（逆）累积分布函数，则中位数m满足：

$$F(m)=0.5\quad\text{或}\quad m=F^{-1}(0.5)$$

样本中位数（sample median）可以用经验累积分布函数[公式（2.1）]或者经验逆累积分布函数[公式（2.2）]计算得出：

$$\hat{F}(m)=0.5\quad\text{或}\quad m=\hat{F}^{-1}(0.5)$$

一种计算样本中位数的更简单的方法是先将所有值$x_i(i\in[1,n])$递增排序。若n是奇数，则中位数是$\frac{n+1}{2}$位置上的值；若n是偶数，则$\frac{n}{2}$和$\frac{n}{2}+1$位置上的值都是中位数。

与均值不同，中位数是健壮的，因此它不受极端值的影响。同时，它是样本中出现的值，也是随机变量能够实际取到的值。

3. 众数

随机变量X的众数（mode）是对应概率质量函数或概率密度函数（取决于X是离散随机变量还是连续随机变量）最大值的X取值。

样本众数(sample mode)是经验概率质量函数[见公式（2.3）]取最大值时的X取值，可以按下式定义：

$$\text{mode(X)}=\arg\max_x\hat{f}(x)$$

众数可能不是一个很好的表征数据居中性的度量，因为一个不具有代表性的元素可能是最频繁出现的元素。此外，若样本中的所有值都是独一无二的，则所有的值都是众数。

例2.1 （样本均值、中位数和众数）考虑鸢尾花数据集中的萼片长度属性（X_1），参见表1-2。样本均值按下式给出：

$$\hat{\mu}=\frac{1}{150}(5.9+6.9+\cdots+7.7+5.1)=\frac{876.5}{150}=5.843$$

图2-1 展示了所有150个不同的萼片长度值及样本均值。图2-2a给出了萼片长度的经验累积分布函数，图2-2b给出了萼片长度的经验逆累积分布函数。

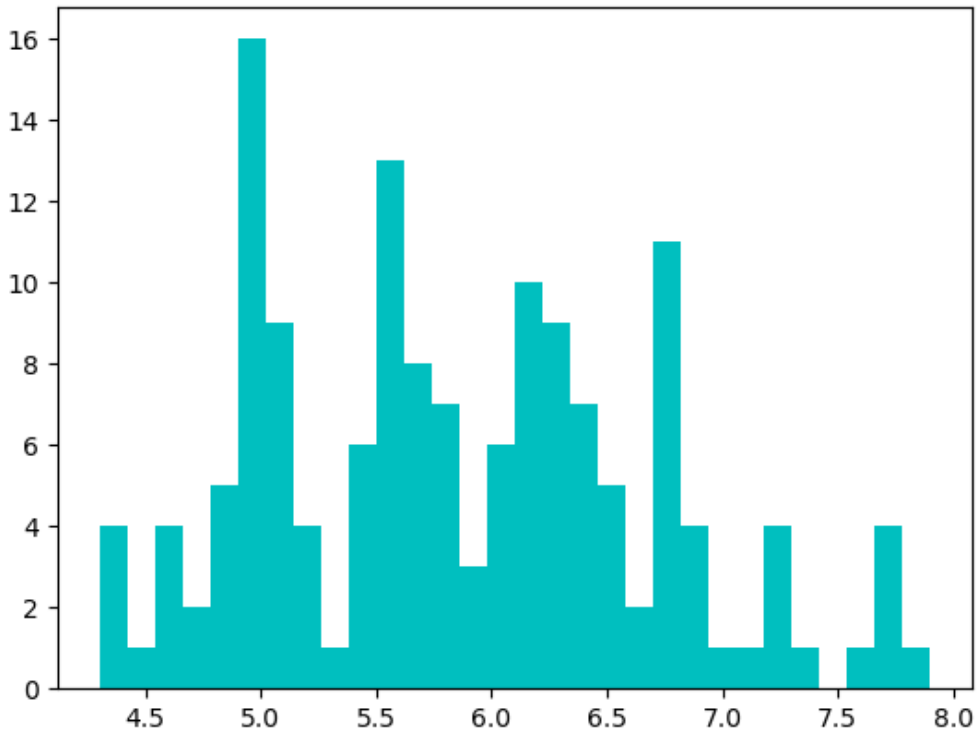
由于$n=150$是偶数，样本中位数是排序后在$\frac{n}{2}=75$和$\frac{n}{2}+1=76$处的值。萼片长度在这两个位置上的值为5.8，因此样本中位数是5.8。根据图2-2b中的逆累积分布函数，可以看到：

$$\hat{F}(5.8)=0.5 \text{或} 5.8=\hat{F}^{-1}(0.5)$$

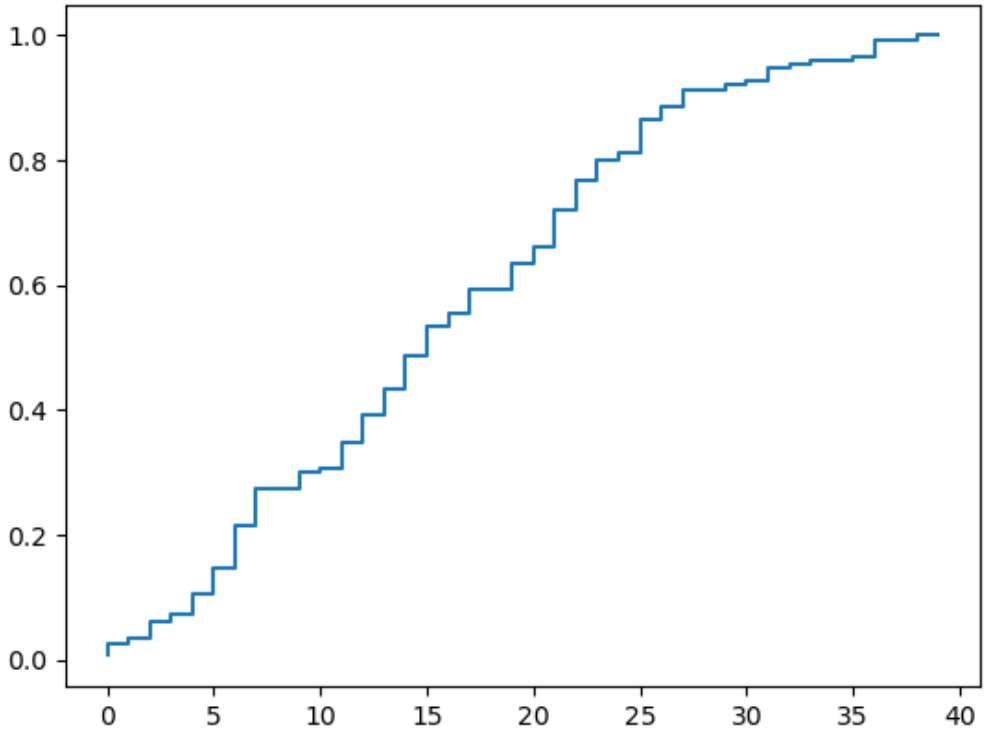
萼片长度的样本众数是5，从图2-1中5的频率可以看出来，在$x=5$处的经验概率质量为：

$$\hat{f}(5)=\frac{10}{150}=0.067$$

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
x = pd.read_csv('iris.csv',sep=',')
x1 = x['Sepal.Length'];
plt.hist(x1,bins=30,color='c')
plt.show()
```

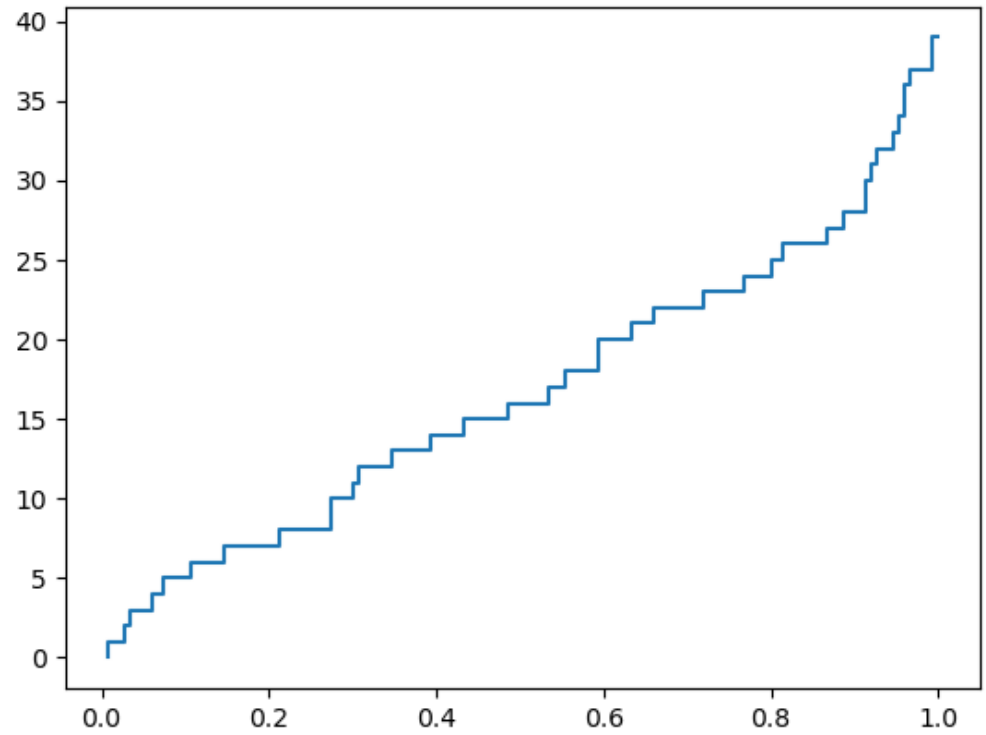



```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
x = pd.read_csv('iris.csv',sep=',')
x1 = x['Sepal.Length'];
y = plt.hist(x1,bins=40);
y1 = np.cumsum(y[0])/np.sum(y[0]);
plt.step(np.arange(0,len(y1)),y1);
plt.show();
```



反函数

```
plt.step(y1,np.arange(0,len(y1)));
plt.show();
```



2.1.2 数据离散度度量

离散度量表征了一个随机变量的值的分散或变化情况。

1. 极差

随机变量X的极差（value range 或 range）是X的最大值和最小值的差，即

$$r = \max\{X\} - \min\{X\}$$

X的极差是一个很流行的参数，但这个概念要和函数X的值域（代表X能够取到的所有值）区分开来。具体区分需要根据上下文来判定。

样本极差是一个统计量，如下所示：

$$\hat{r} = \max_{i=1}^{n}\{x_i\} - \min_{i=1}^{n}\{x_i\}$$

根据定义，极差对极差值很敏感，因此非健壮的。

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
x = pd.read_csv('iris.csv',sep=',');
x1 = x['Sepal.Length'];
max(x1)-min(x1);
```

2. 四分位差

四分位（quartile）是由四分函数[公式（2.2）]产生的将数据分成四等份的特殊值。即四分位与四分值0.25、0.5、0.75和1.0对应。第一个四分位是$q_1 = F^{-1}(0.25)$，其左边包含了25%的数据点；第二个四分位和中位值一样是$q_2 = F^{-1}(0.5)$，其左边包含了50%的数据点，第三个四分位是$q_3 = F^{-1}(0.75)$的数据点；第四个四分位是X的最大值，其左边包含了100%的数据点。

一个更健壮的X的离散度度量是四分位差（interquartile range，IQR）,定义如下：

$$IQR = q_3 - q_1 = F^{-1}(0.75) - F^{-1}(0.25)$$

IQR可以看作切边极差（trimmed range），其中丢弃X的25%的较小值和25%的较大值。换言之，是X中间50%值的极差，根据IQR的定义，它是健壮的。

样本IQR可以通过将经验CDF代入公式（2.7）：

$$\widehat{IQR} = \hat{q}_3 - \hat{q}_1 = \hat{F}^{-1}(0.75) - \hat{F}^{-1}(0.25)$$

```
import math
x2 = list(x1);
x2.sort()   #排序
x2[math.floor(len(x2)/4*1)]  #四分之一位
x2[math.floor(len(x2)/4*3)]  #四分之三位
```

3. 方差和标准差

随机变量X的方差（variance）衡量X的不同取值偏离X的均值或期望值的程度。方差事实上是X所有取值与均值之差的平方的期望值，

$$\sigma^2 = \mathrm{var}(X) = E[(X-\mu)]^2 = \begin{cases} \sum_x (x-\mu)^2 f(x) & X\text{是离散} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx & X\text{是连续} \end{cases}$$

标准差（Standard variation）σ定义为方差σ^2的平方根。

可以将方差表示为X^2的期望值与X期望值的平方的差。

$$\begin{aligned}\sigma^2 = var(X) =& E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ =& E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 \\ =& E[X^2] - (E[X])^2\end{aligned}$$

值得注意，方差事实上是均值的二阶中心矩（second moment about the mean）。随机变量X的r阶中心矩定义为：$E[(x-\mu)^r]$。

```
np.mean(x2)
np.var(x2)
np.std(x2)
```

样本方差 样本方差（sample variance）定义为：

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

上式是数据值x_i与样本均值$\hat{\mu}$的差的平方的均值，可以通过将公式（2.3）中的经验概率函数$\hat{f}$代入公式（2.8）来得到，因为：

$$\hat{\sigma}^2 = \sum_{i=1}(x-\hat{\mu})^2\hat{f}(x) = \sum_{x}(x-\hat{\mu})^2\left(\frac{1}{n}\sum_{i=1}^{n}I(x_i = x)\right) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

样本标准差（sample standard deviation）是样本方差的正平方根

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2}$$

样本值x_i的标准分数（standard score），又称为z分数（z-score），是其与均值距离与标准差的比值：

$$z_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

换言之，x_i的z分数度量了x_i偏离均值$\hat{\mu}$的程度（以$\hat{\sigma}$为单位）。

样本方差的几何意义 可以将X的数据样本当作n维空间中的一个向量，其中n是样本集的大小，即$X = (x_1, x_2, \cdots, x_n)^T \in \mathbb{R}^n$。进一步

$$Z = X - 1\cdot\hat{\mu} = \begin{pmatrix} x_1 - \hat{\mu} \\ x_2 - \hat{\mu} \\ \vdots \\ x_n - \hat{\mu} \end{pmatrix}$$

代表减去均值的属性向量，其中$1 \in \mathbb{R}^n$是所有元素的值均为1的n维向量。可以用Z的大小（magnitude）或其与自身的点乘重写公式（2.10）：

$$\hat{\sigma}^2 = \frac{1}{n}\parallel Z \parallel^2 = \frac{1}{n}Z^TZ = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

样本方差是居中属性向量的大小的平方或其与自身的点差，并按照样本大小标准化。

例2.2 考虑图2-1所示的萼片长度数据样本。样本位距：

$$\max_i\{x_i\} - \min_i\{x_i\} = 7.9 - 4.3 = 3.6$$

从图2-2b中萼片长度的逆累积分布函数，样本的IQR:

$$\hat{q_1} = \hat{F}^{-1}(0.25) = 5.1$$
$$\hat{q_3} = \hat{F}^{-1}(0.75) = 6.4$$
$$\widehat{IQR} = \hat{q_3} - \hat{q_1} = 6.4 - 5.1 = 1.3$$

样本均值方差 由于样本均值$\hat{\mu}$本身也是一个统计量，可以计算其均值和方差样本均值的期望值是μ，如公式（2.6）所示。所有的随机变量x_i都是独立的，因此：

$$\mathrm{var}\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} \mathrm{var}(x_i)$$

更进一步，由于所有的x_i和X是同分布的，因此它们的方差与X相同，即对所有的i有：

$$\mathrm{var}(x_i) = \sigma^2$$

将以上两点结合到一起，可以得到：

$$\mathrm{var}\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} \mathrm{var}(x_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$$

此外，还有：

$$E\left[\sum_{i=1}^{n} x_i\right] = n\mu$$

利用公式(2.9)、公式(2.12)和公式(2.13)，样本均值$\hat{\mu}$的方差可以计算如下：

$$\begin{aligned}\mathrm{var}(\hat{\mu}) =& E[(\hat{\mu} - \mu)^2] = E[\hat{\mu}^2] - \mu^2 = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] - \frac{1}{n}E\left[\sum_{i=1}^{n} x_i\right]^2 \\ =& \frac{1}{n^2}\left(E\left[\left(\sum_{i=1}^{n} x_i\right)^2\right] - E\left[\sum_{i=1}^{n} x_i\right]^2\right) = \frac{1}{n^2}\mathrm{var}\left(\sum_{i=1}^{n} x_i\right) \\ =& \frac{\sigma^2}{n}\end{aligned}$$

换言之，样本均值$\hat{\mu}$偏离均值μ的程度与总体方差σ^2成正比。该值通过增加样本数量而变小。

样本方差是有偏的，但又是渐进无偏的。

公式(2.10)中的样本方差是对真实总体方差σ^2的有偏估计（biased estimator），即$E[\hat{\sigma}^2] \neq \sigma^2$。为说明这一点，等式如下：

$$\sum_{i=1}^{n}(x_i - \mu)^2 = n(\hat{\mu} - \mu)^2 + \sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

第一步先利用公式(2.15)计算$\hat{\sigma}^2$。

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}^2)\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right] - E[(\hat{\mu} - \mu)^2]$$

由于所有的随机变量x_i和X都是独立同分布的，它们和X有着一样的均值μ和一样的方差σ^2。

$$E[(x_i - \mu)^2] = \sigma^2$$

再者，根据公式（2.14）,样本均值$\hat{\mu}$的方差是$E[(\hat{\mu}-\mu)^2]=\frac{\sigma^2}{n}$，代入公式（2.16），

$$E[\hat{\sigma}^2]=\frac{1}{n}n\sigma^2-\frac{\sigma^2}{n}=\left(\frac{n-1}{n}\sigma^2\right)$$

样本方差$\hat{\sigma}^2$是对σ^2的一个有偏估计，因为其期望值是总体方差和$\frac{n-1}{n}$的乘积。然而，它是渐进无偏的（asymptotically unbiased），即当$n\to\infty$时：

$$\lim_{n\to\infty}\frac{n-1}{n}=\lim_{n\to\infty}1-\frac{1}{n}=1$$

换言之，随着样本数量的增加，有

$$E[\hat{\sigma}^2]\to\sigma^2 \quad \text{随着} \quad n\to\infty$$

2.2 二元变量分析

在二元分析中，考虑两个属性，尤其对它们之间的关联或相关性感兴趣（如果存在的话）。因此将注意力集中在两个数值型属性X_1和X_2上，数据D表示为一个$n\times 2$的矩阵：

$$D=\begin{pmatrix} X_1 & X_2 \\ \hline x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

在几何层面，可以从两个角度来看待D:可以看作二维空间中的n个点（或向量），即$x_i=(x_{i1},x_{i2})^T\in\mathbb{R}^2$；也可以看作$n$维空间中的两个点（或向量），即矩阵中的每一列都是$\mathbb{R}^n$中的一个向量。

$$X_1=(x_{11},x_{21},\cdots,x_{n1})^T$$
$$X_2=(x_{12},x_{22},\cdots,x_{n2})^T$$

从概率角度看，列向量$\boldsymbol{X}=(X_1,X_2)^T$被看作一个二元向量随机变量，$x_i(1\leqslant i\leqslant n)$被看作从$X$中随机抽样得到的$n$个点，即$x_i$和$X$是独立同分布的。

```
x1 = x.iloc[:,1:3]
x1.mean()
x1.median()
x1.quantile([0.25,0.75])
x.describe()
```

	Unnamed: 0	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	75.500000	5.843333	3.057333	3.758000	1.199333
std	43.445368	0.828066	0.435866	1.765298	0.762238
min	1.000000	4.300000	2.000000	1.000000	0.100000
25%	38.250000	5.100000	2.800000	1.600000	0.300000
50%	75.500000	5.800000	3.000000	4.350000	1.300000
75%	112.750000	6.400000	3.300000	5.100000	1.800000
max	150.000000	7.900000	4.400000	6.900000	2.500000

```
np.corrcoef(x3,x4)  #相关系数
```

经验联合概率质量函数

X的经验联合概率质量函数（empirical joint probability mass function）

$$\hat{f}(x) = P(X = x) = \frac{1}{n} I(x_i = x)$$
$$\hat{f}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = \frac{1}{n} \sum_{i=1}^{n} I(x_{i1} = x_1, x_{i2} = x_2)$$

其中，$x = (x_1, x_2)^T$，I是一个指示变量，当其参数为真的时候，I的值为1：

$$I(x_i = x) = \begin{cases} 1 & x_{i1} = x_1 \text{且} x_{i2} = x_2 \\ 0 & \text{其他情况} \end{cases}$$

如同一元变量中的情况，概率函数给数据样本中的每个点赋予概率质量$\frac{1}{n}$。

2.2.1 位置和离散度的度量

1. 均值

二元变量均值定义为向量随机变量X的期望值，定义如下：

$$\mu = E[X] = E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

二元均值向量是由每个属性的期望值构成的向量。

样本均值向量可以从$\hat{f}_{X_1}$和$\hat{f}_{X_2}$（分别为X_1和X_2的经验概率质量函数）得出，同样也可以从公式（2.17）所示的联合经验PMF计算出来：

$$\hat{\mu} = \sum_x x\hat{f}(x) = \sum_x x\left(\frac{1}{n}\sum_{i=1}^{n} I(x_i = x)\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2. 方差

利用公式分别计算两个属性X_1和X_2的方差，即σ_1^2和σ_2^2。总方差（total variance）如下：

$$var(D) = \sigma_1^2 + \sigma_2^2$$

2.2.2 相关性度量

1. 协方差

两个属性X_1和X_2的协方差（covariance）提供衡量它们之间的线性相关度的方法，如下：

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)]$$

根据期望的线性性，可以得到：

$$\begin{aligned} \sigma_{12} &= E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= E[X_1X_2 - X_1\mu_2 - X_2\mu_1 + \mu_1\mu_2] \\ &= E[X_1X_2] - \mu_2E[X_1] - \mu_1E[X_2] + \mu_1\mu_2] \\ &= E[X_1X_2] - \mu_1\mu_2 \\ &= E[X_1X_2] - E[X_1][X_2] \end{aligned}$$

公式（2.21）看作将一元变量方差[公式（2.9）]泛化到二元的情况。

若X_1和X_2是独立随机变量，则它们的协方差为0。这是因为，若X_1和X_2是独立的，则有：

$$E[X_1 X_2] = E[X_1] \cdot E[X_2]$$

这意味着

$$\sigma_{12} = 0$$

然而这反过来是不成立的。即，若$\sigma_{12} = 0$，则不能得出X_1和X_2是相互独立的。只能说它们没有线性相关性，因为不能排除这两个属性之间可能存在高阶关系或相关性。

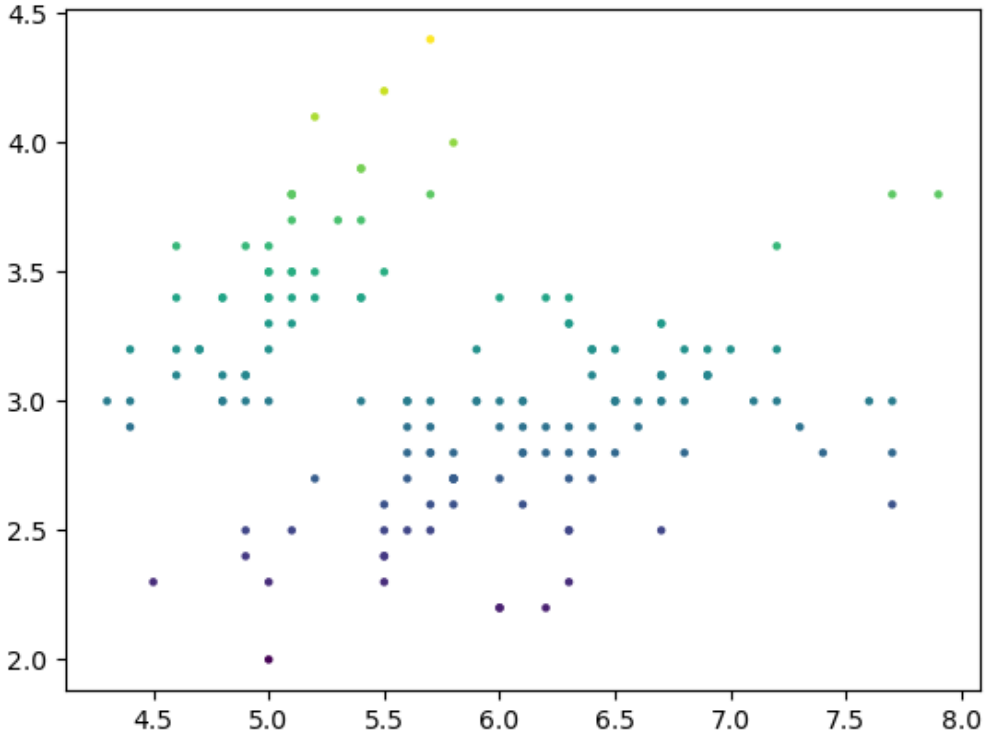
X_1和X_2的样本协方差（sample covariance）可以按下式给出

$$\hat{\sigma}_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \hat{\mu})(x_{i2} - \hat{\mu})$$

将公式（2.17）中的经验联合概率质量函数$\hat{f}(x_1, x_2)$代入公式（2.20），可以得到：

$$\begin{aligned}
\hat{\sigma}_{12} &= E[(X_1 - \mu_1)(X_2 - \mu_2)] \\
&= \sum_{x=(x_1,x_2)^T} (x_1 - \hat{\mu}_1)(x_2 - \hat{\mu}_2)\hat{f}(x_1, x_2) \\
&= \frac{1}{n}\sum_{x=(x_1,x_2)^T}\sum_{i=1}^{n}(x_1 - \hat{\mu}_1)\cdot(x_2 - \hat{\mu}_2)\cdot I(x_{i1} = x_1, x_{i2} = x_2) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_1 - \hat{\mu}_1)(x_2 - \hat{\mu})
\end{aligned}$$

```
x1 = x['Sepal.Length'];
x2 = x['Sepal.Width'];
plt.scatter(x1,x2,s=5,c=x2);
plt.show();
```



2.3 多元变量分析

$$D = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \\ \hline x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

按照行来看，以上数据可以看成d维属性空间中的n个点或者向量：

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{id})^T \in \mathbb{R}^d$$

按照列来看，以上数据可以看成n维空间中的d个点或者向量

$$X_j = (x_{1j}, x_{2j}, \cdots, x_{nj})^T \in \mathbb{R}^n$$

从概率的角度，d个属性可以建模为一个向量随机变量$\boldsymbol{X} = (X_1, X_2, \cdots, X_d)^T$，而点$x_i$可以看成从$X$中得到的随机样本，它们和$X$是独立同分布的。

1. 均值

   推广公式（2.18）,多元变量均值向量（multivariate mean vector）可以通过对每个属性取均值得到，如下式：

$$\mu = E[X] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_d] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix}$$

   推广公式（2.19）,样本均值可以按下式计算：

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

2.

对于公式（2.26）推广到d维的情况，多元变量的协方差信息可以由$d \times d$的对称协方差矩阵（方阵）来表示，该矩阵给出属性对之间的协方差

$$\mathbf{\Sigma} = E[(X-\mu)(X-\mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

对角线元素σ_i^2表示X_i的属性方差，而反对角元素$\sigma_{ij} = \sigma_{ji}$代表了属性对$X_i$和$X_j$之间的协方差。

3. 协方差矩阵是半正定的

Σ是一个半正定（positive semidefinite）矩阵

$$\begin{aligned} a^T \mathbf{\Sigma} a =& a^T E[(X-\mu)(X-\mu)^T]a \\ =& E[a^T(X-\mu)(X-\mu)^T a] \\ =& E[Y^2] \\ \geqslant& 0 \end{aligned}$$

其中Y是随机变量$Y = a^T(X-\mu) = \Sigma_{i=1}^d a_i(X_i - \mu_i)$，利用随机变量平方的期望非负的性质。

由于Σ同时也是对称的，说明Σ的特征值都是非负实数。Σ的d个特征值可以按如下方式从大到小排列：$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d \geqslant 0$。

$$det(\Sigma) = \prod_{i=1}^d \lambda_i \geqslant 0$$

4. 总方差和广义方差

$$var(D) = tr(\Sigma) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_d^2$$

由于一组平方数的和，总方差自然是非负的。

5. 样本协方差矩阵

样本协方差矩阵（sample covariance matrix）可按下式

$$\widehat{\Sigma} = E[(X-\mu)(X-\hat{\mu})^T] = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1d} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots & \hat{\sigma}_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{\sigma}_{d1} & \hat{\sigma}_{d2} & \cdots & \hat{\sigma}_d^2 \end{pmatrix}$$

为了计算上述矩阵，不采用逐个元素计算的方法，而是使用矩阵操作。令Z为居中数据矩阵，其中$Z_i = X_i - 1 \cdot \hat{\mu}_i$，$1 \in \mathbb{R}^n$。

$$Z = D - 1 \cdot \hat{\mu}^T = \begin{pmatrix} | & | & & | \\ Z_1 & Z_2 & \cdots & Z_d \\ | & | & & | \end{pmatrix}$$

居中数据矩阵可以用居中数据点$z_i = x_i - \hat{\mu}$来表示：

$$Z = D - 1 \cdot \hat{\mu}^T = \begin{pmatrix} x_1^T - \hat{\mu}^T \\ x_2^T - \hat{\mu}^T \\ \vdots \\ x_n^T - \hat{\mu}^T \end{pmatrix} = \begin{pmatrix} - & z_1^T & - \\ - & z_2^T & - \\ & \vdots & \\ - & z_n^T & - \end{pmatrix}$$

样本协方差矩阵利用矩阵表示：

$$\widehat{\Sigma}=\frac{1}{n}(Z^TZ)=\frac{1}{n}\begin{pmatrix}Z_1^TZ_1 & Z_1^TZ_2 & \cdots & Z_1^TZ_d\\ Z_2^TZ_1 & Z_2^TZ_2 & \cdots & Z_2^TZ_d\\ \vdots & \vdots & \ddots & \vdots\\ Z_d^TZ_1 & Z_d^TZ_2 & \cdots & Z_d^TZ_d\end{pmatrix}$$

根据上式，样本协方差是由居中属性向量的两两内积或点乘得出并按样本大小进行归一化。

利用居中z_i，样本协方差矩阵可以用外积如下：

$$\widehat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}z_i\cdot z_i^T$$

例2.4 （样本均值和协方差矩阵）

考虑鸢尾花数据集中所有的四个数值型属性，即萼片长度、萼片宽度、花瓣长度和花瓣宽度。对应的多元变量均值向量为：

$$\hat{\mu}=(5.843\quad 3.054\quad 3.759\quad 1.199)^T$$

样本协方差矩阵：

$$\widehat{\Sigma}=\begin{pmatrix}0.681 & -0.039 & 1.265 & 0.513\\ -0.039 & 0.187 & -0.320 & -0.117\\ 1.265 & -0.320 & 3.092 & 1.288\\ 0.513 & -0.117 & 1.288 & 0.579\end{pmatrix}$$

样本总体方差为：

$$var(D)=tr(\widehat{\Sigma})=0.681+0.187+3.092+0.579=4.539$$

广义方差为：

$$det(\widehat{\Sigma})=1.853\times 10^{-3}$$

例2.5 （内积和外积） 通过内积和外积的计算来得到样本协方差矩阵，考虑如下二维数据集：

$$D=\begin{pmatrix}\begin{array}{cc}A_1 & A_2\\ \hline 1 & 0.8\\ 5 & 2.4\\ 9 & 5.5\end{array}\end{pmatrix}$$

均值向量：

$$\hat{\mu}=\begin{pmatrix}\hat{\mu}_1\\ \hat{\mu}_2\end{pmatrix}=\begin{pmatrix}15/3\\ 8.7/3\end{pmatrix}=\begin{pmatrix}5\\ 2.9\end{pmatrix}$$

居中数据矩阵：

$$Z=D-1\cdot\hat{\mu}^T=\begin{pmatrix}1 & 0.8\\ 5 & 2.4\\ 9 & 5.5\end{pmatrix}-\begin{pmatrix}1\\ 1\\ 1\end{pmatrix}\begin{pmatrix}5 & 2.9\end{pmatrix}=\begin{pmatrix}-4 & -2.1\\ 0 & -0.5\\ 4 & 2.6\end{pmatrix}$$

用基于内积的方法[公式（2.30）]来计算样本协方差矩阵：

$$\hat{\Sigma} = \frac{1}{n} Z^T Z = \frac{1}{3} \begin{pmatrix} -4 & 0 & 4 \\ -2.1 & -0.5 & 2.6 \end{pmatrix} \begin{pmatrix} -4 & -2.1 \\ 0 & -0.5 \\ 4 & 2.6 \end{pmatrix}$$
$$= \frac{1}{3} \begin{pmatrix} 32 & 18.8 \\ 18.8 & 11.42 \end{pmatrix} = \begin{pmatrix} 10.67 & 6.27 \\ 6.27 & 3.81 \end{pmatrix}$$

用基于外积的方法[公式（2.31）]来计算样本协方差矩阵有：

$$\begin{aligned} \hat{\Sigma} =& \frac{1}{n} \sum_{i=1}^{n} z_i \cdot z_i^T \\ =& \frac{1}{3} \left[\begin{pmatrix} -4 \\ -2.1 \end{pmatrix} \cdot \begin{pmatrix} -4 & -2.1 \end{pmatrix} + \begin{pmatrix} 0 \\ -0.5 \end{pmatrix} \cdot \begin{pmatrix} 0 & -0.5 \end{pmatrix} + \begin{pmatrix} 4 \\ 2.6 \end{pmatrix} \cdot \begin{pmatrix} 4 & 2.6 \end{pmatrix} \right] \\ =& \frac{1}{3} \left[\begin{pmatrix} 16.0 & 8.4 \\ 8.4 & 4.4 \end{pmatrix} + \begin{pmatrix} 0.0 & 0.0 \\ 0.0 & 0.25 \end{pmatrix} + \begin{pmatrix} 16.0 & 10.4 \\ 10.4 & 6.76 \end{pmatrix} \right] \\ =& \frac{1}{3} \begin{pmatrix} 32.0 & 18.8 \\ 18.8 & 11.42 \end{pmatrix} = \begin{pmatrix} 10.67 & 6.27 \\ 6.27 & 3.81 \end{pmatrix} \end{aligned}$$

2.4 数据规范化

当分析两个或两个以上的属性时，需要对属性值进行规范化，在数据值的规模相差很大的情况下：

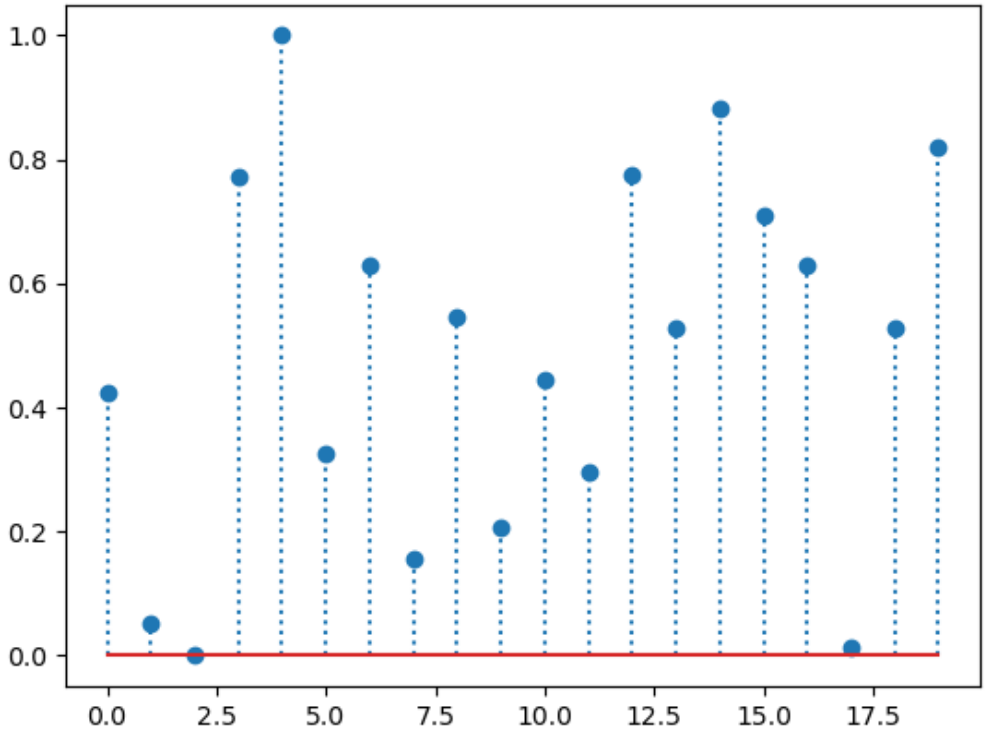
1. 极差归一化

X为一个属性，$x_1, x_2, \cdots, x_n$是对X的一次随机抽样。进行极差归一化（range normalization），每个值都按X的样本极差$\hat{r}$处理如下：

$$x_i' = \frac{x_i - min_i\{x_i\}}{\hat{r}} = \frac{x_i - min_i\{x_i\}}{max_i\{x_i\} - min\{x_i\}}$$

变换之后，新属性的值域为$[0, 1]$。

```
import numpy as np;
import matplotlib.pyplot as plt;
x = np.random.rand(20)*25;
minx = min(x);
maxx = max(x);
hatx = (x -minx)/(maxx-minx);
plt.stem(np.arange(0,len(x)),hatx,linefmt=':',markerfmt='o');
plt.show();
```



2. 标准差归一化

标准差归一化（standard score normalization）称为z归一化，每个值都由它的z分数替换：

$$x_i' = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

其中$\hat{\mu}$是X的样本均值，而$\hat{\sigma}^2$是X的样本方差。变幻之后，新的属性均值$\hat{\mu}' = 0$，标准差$\hat{\sigma}' = 1$。

例2.6 考虑表2-1中所示的样例数据集。属性Age（年龄）和Income（收入）的取值范围差别很大，后者的取值要远远大于前者。考虑x_1和x_2之间的距离：

$$\| x_1 - x_2 \| = \| (2, 200)^T \| = \sqrt{2^2 + 200^2} = \sqrt{40004} = 200.01$$

可以看出，Age的贡献被Income的值所掩盖。

Age的样本极差为$\hat{r} = 40 - 12 = 28$，最小值为12。极差归一化之后，新属性为：

$$Age' = (0, 0.071, 0.214, 0.393, 0.536, 0.571, 0.786, 0.893, 0.964, 1)^T$$

例如，对于点$x_2 = (x_{21}, x_{22}) = (14, 500)$，$x_{21} = 14$被转换：

$$x_{21}' = \frac{14 - 12}{28} = \frac{2}{28} = 0.071$$

同样，Income的样本极差为$6000 - 300 = 5700$，最小值为300，因此income 可以转换为

$$Income' = (0, 0.035, 0.123, 0.298, 0.561,)$$

因此x_{22}为0.035。极差归一化之后的x_1和x_2距离为：

$$\| x_1' - x_2' \| = \| (0, 0)^T - (0.071, 0.035)^T \| = \| (-0.071, -0.035) \| = 0.079$$

可以看出，变换后极差不再受Income的值的影响。

对于z归一化，计算出两个属性的均值和标准差。

	Age	Income
$\hat{\mu}$	27.2	2689
$\hat{\sigma}$	9.77	1726.15

Age可转换为：

$$Age' = (-1.56, -1.35, -0.94, -0.43, -0.02, 0.08, 0.70, 1.0, 1.21, 1.31)^T$$

例如，对于点$x_2 = (x_{21}, x_{22}) = (14, 500)$，值$x_{21} = 14$可以转换为：

$$x'_{21} = \frac{14 - 27.2}{9.77} = -1.35$$

同样，Income可以转换为

$$Income' = (-1.38, -1.26, -0.97, -0.39, 0.48, 0.77, 0.94, 1.92, -0.10, 0.01)^T$$

因此，$x_{22} = -1.26$。x_1和x_2经过z归一化之后的距离为：

$$\| x'_1 - x'_2 \| = \| (-1.56, -1.38)^T - (1.35, -1.26)^T \| = \| (-0.18, -0.12)^T \| = 0.216$$

表2-1待归一化的数据集

x_i	Age(X_1)	Income(X_2)
x_1	12	300
x_2	14	500
x_3	18	1000
x_4	23	2000
x_5	27	3500
x_6	28	4000
x_7	34	4300
x_8	37	6000
x_9	39	2500
x_{10}	40	2700

2.5 正态分布

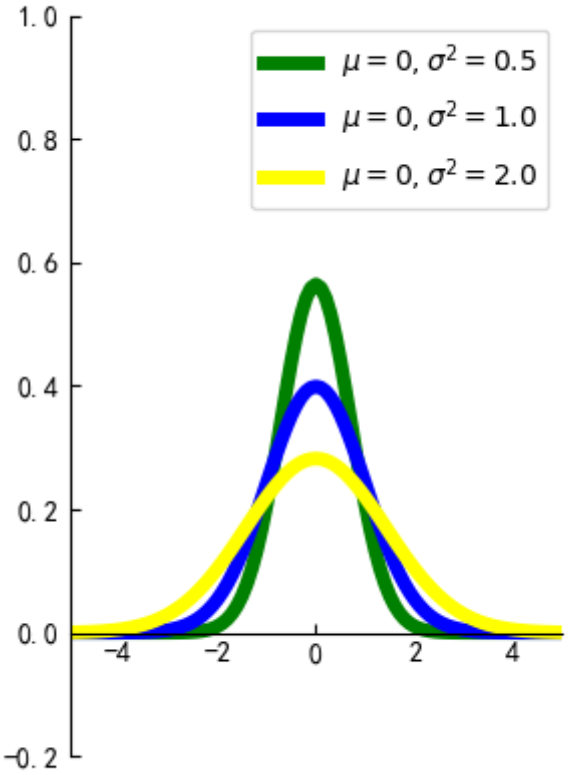
2.5.1 一元正态分布

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{ - \frac{(x-\mu)^2}{2\sigma^2} \right\}$$

$(x-\mu)^2$衡量了x与分布均值μ的距离，因此概率密度随着与均值距离的增加而呈指数式下降。概率密度在$x=\mu$的时候取到最大值$f(\mu)=\frac{1}{\sqrt{2\pi\sigma^2}}$，与标准差$\sigma$成反比。

例2.7 图2-5给出了标准正态分布的图像($\mu=0,\sigma^2=1$)。正态分布呈现典型的钟形，并关于均值对称。图像同时展示了不同标准差对分布形状的影响。较小的值（例如$\sigma=0.5$）对应的分布更“尖”，概率密度在两边衰减较快，而更大的值（例如$\sigma=2$）对应的分布更“扁平”,概率密度在两边衰减较慢。由于正态分布是对称的，分布的均值μ与中位数及众数相等。

```
from mpl_toolkits.mplot3d import axes3d;
import matplotlib.pyplot as plt;
import numpy as np;
import math

plt.rcParams['font.sans-serif'] = ['SimHei']  # 用来正常显示中文标签
plt.rcParams['axes.unicode_minus'] = False  # 用来正常显示负号
plt.rcParams['xtick.direction'] = 'in' #x的刻度向内
plt.rcParams['ytick.direction'] = 'in' #y的刻度向内
plt.rcParams['lines.linewidth'] = 5
plt.rcParams['lines.color'] ='y'


def gd(x, mu=0, sigma=1):
    left = 1 / (np.sqrt(2 * math.pi) * np.sqrt(sigma))
    right = np.exp(-(x - mu)**2 / (2 * sigma))
    return left * right

if __name__ == '__main__':
    x = np.arange(-6, 6, 0.1)
    y_1 = gd(x, 0, 0.5)
    y_2 = gd(x, 0, 1.0)
    y_3 = gd(x, 0, 2.0)

    plt.plot(x, y_1, color='green')
    plt.plot(x, y_2, color='blue')
    plt.plot(x, y_3, color='yellow')
    plt.xlim(-5.0, 5.0)
    plt.ylim(-0.2, 1)

    ax = plt.gca()
    ax.spines['right'].set_color('none')
    ax.spines['top'].set_color('none')
    ax.xaxis.set_ticks_position('bottom')
    ax.spines['bottom'].set_position(('data', 0))
    ax.yaxis.set_ticks_position('left')
    # ax.spines['left'].set_position(('data', 0))

    plt.legend(labels=['$\mu = 0, \sigma^2=0.5$', '$\mu = 0, \sigma^2=1.0$',
'$\mu = 0, \sigma^2=2.0$'])
    plt.show()
```



概率质量

给定区间$[a, b]$，在该区间上正态分布

$$P(a \leqslant x \leqslant b) = \int_a^b f(x|\mu, \sigma^2)dx$$

经常会对距离均值k个标准差内的概率质量感兴趣，即对于区间$[\mu - k\sigma, \mu + k\sigma]$可以计算如下：

$$P(\mu - k\sigma \leqslant x \leqslant \mu + k\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mu-k\sigma}^{\mu+k\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} dx$$

通过变量替换$z = \dfrac{x-\mu}{\sigma}$，可以等价得到以标准正态分布表示的公式：

$$\begin{aligned} P(-k \leqslant z \leqslant k) =& \frac{1}{\sqrt{2\pi}} \int_{-k}^{k} e^{-\frac{1}{2}z^2} dz \\ =& \frac{2}{\sqrt{2\pi}} \int_{0}^{k} e^{-\frac{1}{2}z^2} dz \end{aligned}$$

以上公式利用$e^{-\frac{1}{2}z^2}$是对称的特点，因此在$[-k, k]$区间上的积分为$[0, k]$区间上积分的两倍。最后，再通过一次变量变换$t = \frac{z}{\sqrt{2}}$，得到

$$P(-k \leqslant z \leqslant k) = 2 \cdot P(0 \leqslant t \leqslant k/\sqrt{2}) = \frac{2}{\sqrt{\pi}} \int_0^{k/\sqrt{2}} e^{-t^2} dt = \mathrm{erf}(k/\sqrt{2})$$

其中erf是高斯误差函数（Gauss error function），定义为：

$$\mathrm{erf(x)} = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

利用公式（2.32）可以计算出距均值k个标准差的概率质量。对于$k=1$，有

$$P(\mu-\sigma \leqslant x \leqslant \mu+\sigma)=\operatorname{erf}(1/\sqrt{2})=0.6827$$

这意味着68.27%的点都处在距均值1个标准差的范围内；对于$k=2$，有$\operatorname{erf}(2/\sqrt{2})=0.9545$；对于$k=3$，有$\operatorname{erf}(3/\sqrt{2})=0.9973$。因此，几乎正态分布的整个概率质量（也就是$99.73\%$）都在距均值$\pm 3$个标准差的范围内。

2.5.2 多元正态分布

给定d维空间中的向量随机变量$X=(X_1,X_2,\cdots,X_d)^T$，若X服从多元正态分布（univariate normal distribution），均值为μ，协方差矩阵为Σ，则联合多元概率密度函数如下式子：

$$f(x|\mu,\Sigma)=\frac{1}{(\sqrt{2\pi}^d\sqrt{|\Sigma|})}exp\left\{-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right\}$$

其中$|\Sigma|$是协方差矩阵的行列式。同一元的情况类似，下式为

$$(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

计算点x和分布均值μ的距离，称为马氏距离

$$(x_i-\mu)^TI^{-1}(x_i-\mu)=\| x_i-\mu\|^2$$

标准多元正态分布的参数为$\mu=0$和$\Sigma=I$。图2-6a画出了标准二元正态分布的概率密度，其中：

$$\mu=0=\begin{pmatrix}0\\0\end{pmatrix}$$

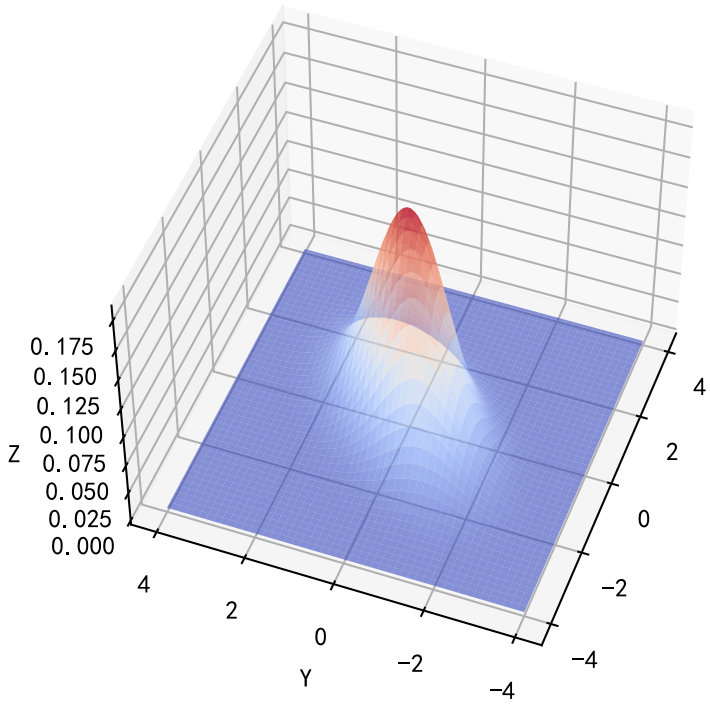
且

$$\Sigma=I=\begin{pmatrix}1&0\\0&1\end{pmatrix}$$

这事实上代表了两个属性相互独立的情况，且每个属性本身都服从标准正态分布。标准正态分布的对称性可以从图2-6b所示的等值线图（contour plot）看出来，每一条等高标准代表了一组等概率密度的点x，密度为$f(x)$。

```
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import axes3d
from matplotlib import cm
import matplotlib as mpl
import math;

mpl.rcParams['font.sans-serif'] = ['SimHei'];
mpl.rcParams['axes.unicode_minus'] = False;

if __name__ == '__main__':
    num =100;
    l = np.linspace(-4,4,num);
    X,Y = np.meshgrid(l,l);
    u = np.array([0,0]);
    o = np.array([[1,0.5],[0.5,1]]);
    pos =
np.concatenate((np.expand_dims(X,axis=2),np.expand_dims(Y,axis=2)),axis=2);
    a = np.dot((pos-u),np.linalg.inv(o));
    b = np.expand_dims(pos-u,axis=3);
```

```
    Z = np.zeros((num,num),dtype=np.float32);
    for i in range(num):
        Z[i] = [np.dot(a[i,j],b[i,j]) for j in range(num)];
    Z = np.exp(Z*(-0.5))/(2*np.pi*math.sqrt(np.linalg.det(o)));
    fig = plt.figure();
    ax = fig.add_subplot(111,projection='3d');
    ax.plot_surface(X,Y,Z,rstride=2,cstride=2,alpha=0.6,cmap=cm.coolwarm);
    ax.set_ylabel('Y');
    ax.set_zlabel('Z');
    plt.show()
```



1. 多元正态分布的几何结构

$$\Sigma\mu_i = \lambda_i u_i$$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix}$$

此外，特征向量都是单位向量且两两正交，因此它们是标准正交的：

$$\mu_i^T \mu_i = 1$$
$$\mu_i^T \mu_j = 0, \quad i \neq j$$

特征向量可以放在一起组成正交矩阵U，其中任意两列都是标准正交的：

$$U=\begin{pmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_d \\ | & | & & | \end{pmatrix}$$

Σ的特征值分解可以用下式简洁地表示：

$$\Sigma = U\Lambda U^T$$

2. 总方差和广义方差

协方差矩阵的行列式为$det(\Sigma)=\prod_{i=1}^{d}\lambda_i$，因此Σ的广义方差为其特征值的乘积。

$$var(D)=tr(\Sigma)=\sum_{i=1}^{d}\sigma_i^2=\sum_{i=1}^{d}\lambda_i=tr(\Lambda)$$

亦即$\sigma_1^2+\cdots+\sigma_d^2=\lambda_1+\cdots+\lambda_d$。

例2.8 （二元正态密度）将鸢尾花数据集（见表1-1）的萼片长度（X_1）和萼片宽度（X_2）当作连续型随机变量，

$$\hat{\mu}=(5.843,3.054)^T$$

样本的协方差矩阵为：

$$\hat{\Sigma}=\begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$$

两个属性的二元正态概率密度可见图2-7。该图同时展示了等高线和数据点。

考虑$x_2=(6.9,3.1)^T$，有

$$x_2-\hat{\mu}=\begin{pmatrix} 6.9 \\ 3.1 \end{pmatrix}-\begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}=\begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix}$$

x_2和$\hat{\mu}$之间的马氏距离为：

$$\begin{aligned}(x_i-\hat{\mu})^T\hat{\Sigma}^{-1}(x_i-\hat{\mu})=&\begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix}\begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}^{-1}\begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix} \\ =&\begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix}\begin{pmatrix} 1.486 & 0.31 \\ 0.31 & 5.42 \end{pmatrix}\begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix} \\ =&1.701\end{aligned}$$

它们之间的欧几里得距离的平方为：

$$\|(x_2-\hat{\mu})\|^2=\begin{pmatrix} 1.057 & 0.046 \end{pmatrix}\begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix}=1.119$$

$\hat{\Sigma}$的特征值和对应的特征向量如下：

$$\begin{aligned}\lambda_1=0.684 \quad u_1=(-0.997,0.078)^T \\ \lambda_2=0.184 \quad u_1=(-0.078,0.997)^T\end{aligned}$$

两个特征向量定义了新的坐标轴。它们的协方差矩阵是：

$$\Lambda=\begin{pmatrix} 0.684 & 0 \\ 0 & 0.184 \end{pmatrix}$$

原来的坐标轴$e_1=(1,0)^T$和u_1之间的角度决定了多元正态的旋转角度：

$$cos\theta = e_1^T u_1 = -0.997$$
$$\theta = cos^{-1}(-0.997) = 175.5^\circ$$

图2-7展示了新的坐标轴和新的方差。可以看到在原来的坐标轴上，等高线仅仅旋转了175.5°（或-4.5°。

2.7 习题

Q1. 判断下列句子的真假。

(a) 均值对于孤立点是健壮的。（不健壮）

(b) 中位数对于孤立点是健壮的。（健壮）

(c) 标准差对于孤立点是健壮的。（不健壮）

Q2. 令X和Y为两个随机变量，分别代表年龄和体重。考虑$n = 20$的随机样本：

$$X = (69, 74, 68, 70, 72, 67, 66, 70, 76, 68, 72, 79, 74, 67, 66, 71, 74, 75, 75, 76)$$
$$Y = (153, 175, 155, 135, 172, 150, 115, 137, 200, 130, 140, 265, 185, 112, 140, 150, 165, 185, 210, 220)$$

(a) 找出X的均值、中位数和众数。

(b) 计算Y的方差。

(c) 画出X的正态分布。

(d) 计算观察到年龄大于等于80的概率。

(e) 找出这两个变量的二维均数$\hat{\mu}$和协方差矩阵$\hat{\Sigma}$

(f) 说出年龄和体重之间的相关性

(g) 画出展示年龄和体重之间的关系的散点图

Q3. 证明公式（2.15）的等式是成立的，即

$$\sum_{i=1}^{n}(x_i - \mu)^2 = n(\hat{\mu} - \mu)^2 + \sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

证明：

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \mu)^2 &= \sum_{i=1}^{n}(x_i - \hat{\mu} + \hat{\mu} - \mu)^2 \\
&= \sum_{i=1}^{n}(x_i - \hat{\mu})^2 + 2\sum_{i=1}^{n}[(x_i - \hat{\mu})(\hat{\mu} - \mu)] + \sum_{i=1}^{n}(\hat{\mu} - \mu)^2 \\
&= n(\hat{\mu} - \mu)^2 + \sum_{i=1}^{n}(x_i - \hat{\mu})^2 + 2\sum_{i=1}^{n}[(x_i - \frac{1}{n}\sum_{i=1}^{n}x_i)(\hat{\mu} - \mu)] \\
&= n(\hat{\mu} - \mu)^2 + \sum_{i=1}^{n}(x_i - \hat{\mu})^2 + 2(\hat{\mu} - \mu)\underbrace{\left[\sum_{i=1}^{n}x_i - \frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{n}x_i\right]}_{0} \\
&= n(\hat{\mu} - \mu)^2 + \sum_{i=1}^{n}(x_i - \hat{\mu})^2
\end{aligned}$$

Q4. 证明若x_i是独立的随机变量，则

$$var\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} var(x_i)$$

$$\begin{aligned} var(X) =& E[(X-\mu)]^2 \\ var\left(\sum_{i=1}^{n} x_i\right) =& E[(\sum_{i=1}^{n} x_i - \mu)]^2 = \sum_{x}(\sum_{i=1}^{n} x_i - \mu)^2 f(x) \end{aligned}$$

Q5. 对一个随机变量X，定义均值绝对偏差（mean absolute deviation）如下：

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \mu|$$

这一度量是否是健壮的？为什么？

Q6. 证明向量随机变量$X = (X_1, X_2)^T$的期望值就是由随机变量X_1和X_2的期望值构成的向量[如公式（2.18）所示]。

Q7. 证明公式（2.23）所示的任意两个变量X_1和X_2的关联度处于期间[-1,1]。

Q8. 给定表2-2中的数据，计算协方差矩阵和广义方差。

表2-2 Q8的数据集

	X_1	X_2	X_3
x_1	17	17	12
x_2	11	9	13
x_3	11	8	19

Q9. 证明公式（2.31）中关于样本协方差矩阵的外积和公式（2.29）等价。

Q10. 假设给定两个一元正态分布N_A和N_B，令其均值和标准差为：$\mu_A = 4$、$\sigma_A = 1$、$\mu_B = 8$、$\sigma_B = 2$。

(a) 对任意的$x_i \in \{5, 6, 7\}$，这个样本集更可能由哪一个分布产生?

(b) 推导一个点的表达式，满足两个正态分布产生该点的概率相同的条件。

Q11. 考虑表2-3，假设属性X和Y都是数值型的，且该表代表了整个总体，若已知X和Y之间的相关性为0，如果推断出Y的值?

表2-3 Q11的数据集

X	Y
1	a
0	b
1	c
0	a
0	c

Q12. 在什么条件下协方差矩阵Σ会与相关矩阵相等（其中项(i, j)）给出属性X_i和X_j之间的关联关系）？对于这两个属性你能得出什么结论?