3.7 方差

方差是各个数据与其算术平均数的离差平方和的平均数。

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

除以n-1而不是n，是因为这样能使我们以较小的样本集更好地逼近总体的标准差，即统计上所谓的“无偏估计”

分母为n造成的偏差

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[(X_i-\mu)+(\mu-\bar{X})\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+\frac{2}{n}\sum_{i=1}^{n}(X_i-\mu)(\mu-\bar{X})+\frac{1}{n}\sum_{i=1}^{n}(\mu-\bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+2(\bar{X}-\mu)(\mu-\bar{X})+(\mu-\bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\mu-\bar{X})^2
\end{aligned}
$$

换言之，除非正好$\bar{X}=\mu$，否则我们一定有

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2<\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2,$$

而不等式右边的那位才是的对方差的“正确”估计！

这个不等式说明了，为什么直接使用$\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$会导致对方差的低估。

那么，在不知道随机变量真实数学期望的前提下，如何“正确”的估计方差呢？答案是把上式中的分母n换成$n-1$，通过这种方法把原来的偏小的估计“放大”一点点，我们就能获得对方差的正确估计了：

$$\mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right]=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right]=\sigma^2.$$

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

但，這個 estimator 有 bias，因為:

$$\begin{aligned}
E(S_1^2) &= \frac{1}{n}\sum_{i=1}^{n} E\left((X_i - \overline{X})^2\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2\right) \\
&= \frac{1}{n}E\left(\sum_{i=1}^{n}\left((X_i - \mu)^2 - 2(X_i - \mu)(\overline{X} - \mu) + (\overline{X} - \mu)^2\right)\right) \\
&= \frac{1}{n}E\left(\sum_{i=1}^{n}(X_i - \mu)^2 - 2\sum_{i=1}^{n}(X_i - \mu)(\overline{X} - \mu) + n(\overline{X} - \mu)^2\right) \\
&= \frac{1}{n}E\left(\sum_{i=1}^{n}(X_i - \mu)^2 - 2n(\overline{X} - \mu)(\overline{X} - \mu) + n(\overline{X} - \mu)^2\right) \\
&= \frac{1}{n}E\left(\sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}E\left((X_i - \mu)^2\right) - nE\left((\overline{X} - \mu)^2\right)\right) \\
&= \frac{1}{n}\left(nVar(X) - nVar(\overline{X})\right) \\
&= Var(X) - Var(\overline{X}) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2,
\end{aligned}$$

而 (n-1)/n * σ^2 **!=** σ^2，所以，為了避免使用有 bias 的 estimator，我們通常使用它的修正值 S^2:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

方差的Python例子

```
import numpy as np

arr = [1,2,3,4,5,6]
                                   #求均值
arr_mean = np.mean(arr)
                                   #求方差
arr_var = np.var(arr)
                                   #求标准差
arr_std = np.std(arr,ddof=1)

print("平均值为：%f" % arr_mean)
print("方差为：%f" % arr_var)
print("标准差为:%f" % arr_std)
```

3.8 协方差

协方差用来刻画两个随机变量(X, Y)之间的相关性。

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

如果协方差为正，说明X，Y同向变化，协方差越大说明同向程度越高；如果协方差为负，说明X，Y反向运动，协方差越小说明反向程度越高。

协方差的python的例子

```python
import numpy as np

# 随机生成两个样本
x = np.random.randint(0, 9, 1000)
y = np.random.randint(0, 9, 1000)

# 计算平均值
mx = x.mean()
my = y.mean()

# 计算标准差
stdx = x.std()
stdy = y.std()

# 计算协方差矩阵
covxy = np.cov(x, y)
print(covxy)
```

```
[[6.83907508 0.10925926]
 [0.10925926 6.53390891]]
6.832236
6.527375
0.10914999999999989
[[1.         0.01634455]
 [0.01634455 1.        ]]
```