解 设该超市每周进货 n 件水果 ($1 \leqslant n \leqslant 10$), 则每周的收益为

$$Y = \begin{cases} 10n & X \geqslant n \\ 10X - 4(n - X) & X < n. \end{cases}$$

由于收益 Y 是关于 X 的随机变量, 考虑在期望下的收益

$$\begin{aligned} E[Y] &= \sum_{i=1}^{n-1}(10i - 4(n-i))P(X=i) + \sum_{i=n}^{10} 10nP(X=i) \\ &= \sum_{i=1}^{n-1} \frac{14i - 4n}{10} + \sum_{i=n}^{10} n = \frac{-7n^2 + 107n}{10}. \end{aligned}$$

在上式中对 n 求一阶导数并令其等于零, 求解可得 $n = 7.64$, 最后取 $n = 8$.

6.2 协方差

随机变量的期望和方差仅针对单个随机变量, 无法刻画变量之间的统计信息. 本节研究新的数字特征: 协方差, 用于描述随机变量 X 和 Y 之间的相互关系.

定义 6.1 若二维随机向量 (X, Y) 的期望$E[(X - E[X])(Y - E[Y])]$ 存在, 则称其为 X 和 Y 的**协方差**, 记为

$$\mathrm{Cov}(X, Y) = E\left[(X - E[X])(Y - E[Y])\right] = E[XY] - E[X]E[Y].$$

协方差是两个随机变量与它们各自期望的偏差之积的期望, 可正可负. 根据协方差的定义有

$$\mathrm{Cov}(X, X) = \mathrm{Var}(X) \quad \text{和} \quad \mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X, Y).$$

关于协方差的性质有

性质 6.5 对任意随机变量 X, Y 和常数 c 有

$$\mathrm{Cov}(X, c) = 0 \qquad \text{和} \qquad \mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X).$$

性质 6.6 对任意随机变量 X, Y 和常数 a, b 有

$$\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X, Y) \quad \text{和} \quad \mathrm{Cov}(X + a, Y + b) = \mathrm{Cov}(X, Y).$$

证明 根据协方差的定义有

$$\mathrm{Cov}(aX, bY) = E[(aX - E[aX])(bY - E[bY])] = abE[(X - E[X])(Y - E[Y])],$$

以及

$$\mathrm{Cov}(X+a, Y+b) = E[(X+a-E[X+a])(Y+b-E[Y+b])] = E[(X-E[X])(Y-E[Y])].$$

性质 6.7 对任意随机变量 X_1, X_2, Y 有

$$\mathrm{Cov}(X_1+X_2, Y) = \mathrm{Cov}(X_1, Y) + \mathrm{Cov}(X_2, Y).$$

证明 根据协方差的定义有

$$\begin{aligned}\mathrm{Cov}(X_1+X_2, Y) &= E[(X_1+X_2-E[X_1+X_2])(Y-E(Y))] \\ &= E[(X_1-E[X_1])(Y-E[Y])] + E[(X_2-E[X_2])(Y-E[Y])] \\ &= \mathrm{Cov}(X_1, Y) + \mathrm{Cov}(X_2, Y).\end{aligned}$$

可将性质 6.7 推广到多个随机变量: 对随机变量$X_1, X_2, \ldots, X_n$和$Y_1, Y_2, \ldots, Y_m$ 有

$$\mathrm{Cov}\left(\sum_i^n X_i, \sum_j^m Y_j\right) = \sum_i^n \sum_j^m \mathrm{Cov}(X_i, Y_j),$$

以及进一步有

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \mathrm{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j\right) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j).$$

性质 6.8 若随机变量 X 与 Y 相互独立, 则有 $\mathrm{Cov}(X, Y) = 0$, 但反之不成立.

证明 若 X 与 Y 相互独立, 则有 $E[XY] = E[X]E[Y]$, 于是得到

$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y] = 0.$$

但反之并不成立, 即若有 $\mathrm{Cov}(X, Y) = 0$ 成立, 不能得出 X 与 Y 相互独立. 下面给出一个例子: 设随机变量 X 的分布列为

$$P(X=-1) = P(X=0) = P(X=1) = 1/3,$$

很容易得到 $E[X] = 0$. 设随机变量 Y 为

$$Y = \begin{cases} 0 & X \neq 0 \\ 1 & X = 0, \end{cases}$$

可以发现 $E(XY)=0$, 最后得到

$$\mathrm{Cov}(X,Y)=E[XY]-E[X]E[Y]=0.$$

但随机变量 X 与 Y 显然不是相互独立的, 因为

$$P(X=0,Y=0)=0 \quad \text{而} \quad P(X=0)P(Y=0)=2/9.$$

由此完成证明.

性质 6.9　对任意随机变量 X 与 Y 有

$$(\mathrm{Cov}(X,Y))^2 \leqslant \mathrm{Var}(X)\mathrm{Var}(Y),$$

等式成立的充要条件是 $Y=aX+b$ 几乎处处成立, 即 X 与 Y 几乎处处存在线性关系.

证明　根据 Cauchy-Schwartz 不等式有

$$\begin{aligned}|\mathrm{Cov}(X,Y)| &= |E[(X-E[X])(Y-E[Y])]| \\ &\leqslant \sqrt{E[(X-E[X])^2]E[(Y-E[Y])^2]}=\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}.\end{aligned}$$

下面证明等式成立的充要条件. 若 $Y=aX+b$ 几乎处处成立, 则有

$$\mathrm{Var}(Y)=a^2\mathrm{Var}(X) \qquad \text{和} \qquad \mathrm{Cov}(X,Y)=\mathrm{Cov}(X,aX+b)=a\mathrm{Var}(X),$$

由此直接验证 $(\mathrm{Cov}(X,Y))^2=\mathrm{Var}(X)\mathrm{Var}(Y)$. 另一方面, 可以考虑函数

$$\begin{aligned}f(t) &= E[t(X-E[X])-(Y-E[Y])]^2 \\ &= t^2E[X-E[X]]^2-2tE[(X-E[X])(Y-E[Y])]+E[Y-E[Y]]^2,\end{aligned}$$

根据二次方程的性质和条件 $(\mathrm{Cov}(X,Y))^2=\mathrm{Var}(X)\mathrm{Var}(Y)$ 有

$$\Delta=4(E[(X-E[X])(Y-E[Y])])^2-4E(X-E[X])^2E(Y-E[Y])^2=0.$$

由此存在一个根 t_0 使得 $f(t_0)=0$ 成立, 即

$$f(t_0)=E[(t_0(X-E[X])-(Y-E[Y]))^2]=0,$$

由此可得 $Y=t_0X-t_0E[X]+E[Y]$ 几乎处处成立, 由此完成证明.

定理 6.2 若随机向量 $(X,Y)\sim\mathcal{N}(\mu_x,\mu_y,\sigma_x^2,\sigma_y^2,\rho)$, 则有 $\mathrm{Cov}(X,Y)=\rho\sigma_x\sigma_y$.

证明 根据协方差的定义有

$$\mathrm{Cov}(X,Y)=E[(X-E[X])(Y-E[Y])]=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_x)(y-\mu_y)f(x,y)dxdy,$$

其中

$$f(x,y)=\frac{1}{2\pi\sqrt{1-\rho^2}\sigma_x\sigma_y}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right).$$

考虑变量变换

$$\begin{cases}u=(x-\mu_x)/\sigma_x\\ v=(y-\mu_y)/\sigma_y\end{cases}\qquad\text{于是有}\qquad\begin{cases}x=u\sigma_x+\mu_x\\ y=v\sigma_y+\mu_y.\end{cases}$$

积分变量变换的雅可比行列式为 $J=\sigma_x\sigma_y$, 于是有

$$\begin{aligned}\mathrm{Cov}(X,Y)&=\frac{\sigma_x\sigma_y}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}uv\exp\left(-\frac{u^2+v^2-2\rho uv}{2(1-\rho^2)}\right)dudv\\&=\frac{\sigma_x\sigma_y}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{+\infty}v\exp(-v^2/2)dv\int_{-\infty}^{+\infty}u\exp\left(-\frac{(u-\rho v)^2}{2(1-\rho^2)}\right)du\\&=\frac{\sigma_x\sigma_y}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\rho v^2\exp(-v^2/2)dv=\rho\sigma_x\sigma_y.\end{aligned}$$

其中第三个等式成立是因为利用正态分布 $\mathcal{N}(\rho v,1-\rho^2)$ 的期望, 而最后一个不等式成立是因为利用标准正态分布的方差. 定理得证.

结合定理 5.3 和定理 6.2 可得

推论 6.1 若随机向量 $(X,Y)\sim\mathcal{N}(\mu_x,\mu_y,\sigma_x^2,\sigma_y^2,\rho)$, 则 X 和 Y 相互独立的充要条件是协方差 $\mathrm{Cov}(X,Y)=0$.

例 6.3 若随机变量 $X_1,X_2,\cdots,X_n$ 相互独立、且服从方差为 σ^2 的正态分布, 讨论 $\bar{X}=\sum_{i=1}^n X_i/n$ 和 $\bar{X}-X_1$ 的独立性.

解 根据正态分布的性质可以知道 $\bar{X}$ 和 $\bar{X}-X_1$ 都服从正态分布, 而正态分布的独立性可通过协方差来研究. 根据协方差的性质有

$$\mathrm{Cov}(\bar{X},\bar{X}-X_1)=\mathrm{Cov}(\bar{X},\bar{X})-\mathrm{Cov}(\bar{X},X_1).$$

根据 $X_1, X_2, \cdots, X_n$ 的相互独立性有

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n}\\ \mathrm{Cov}(\bar{X}, X_1) &= \mathrm{Cov}\left(\sum_{i=1}^{n}\frac{X_i}{n}, X_1\right) = \frac{1}{n}\mathrm{Cov}(X_1, X_1) = \frac{\sigma^2}{n},\end{aligned}$$

于是得到 $\mathrm{Cov}(\bar{X}, \bar{X} - X_i) = 0$. 根据推论 6.1 得到 $\bar{X}$ 和 $\bar{X} - X_i$ 是相互独立的.

例 6.4 随机变量 (X, Y) 联合概率密度为

$$f(x,y) = \begin{cases}(x+y)/8 & 0 \leqslant x \leqslant 2, 0 \leqslant y \leqslant 2\\ 0 & \text{其它},\end{cases}$$

求 $\mathrm{Cov}(X, Y)$ 和 $\mathrm{Var}(X + Y)$.

解 根据协方差的定义有 $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$, 需计算

$$E[X] = E[Y] = \int_0^2\int_0^2 \frac{1}{8}x(x+y)dxdy = \frac{7}{6} \quad \text{和} \quad E[XY] = \int_0^2\int_0^2 \frac{1}{8}xy(x+y)dxdy = \frac{4}{3},$$

由此可得 $\mathrm{Cov}(X,Y) = -1/36$. 进一步计算

$$E[X^2] = E[Y^2] = \int_0^2\int_0^2 \frac{1}{8}x^2(x+y)dxdy = \frac{5}{3},$$

由此可得 $\mathrm{Var}(X) = \mathrm{Var}(Y) = 5/3 - (7/6)^2 = 11/36$. 最后得到

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y) = 11/18 - 1/18 = 5/9.$$

例 6.5 (匹配问题) 有 n 对夫妻参加一次聚会, 将所有人员随机分成 n 组, 每组一男一女, 用 X 表示夫妻两人被分到同一组的对数, 求 X 的期望和方差.

解 用随机变量 X_i 表示第 i 对夫妻是否被分到一组, 即

$$X_i = \begin{cases}1 & \text{若第}\, i\, \text{对夫妻被分到一组}\\ 0 & \text{否则},\end{cases}$$

则有 $X = X_1 + X_2 + \cdots + X_n$. 随机变量 X_i 的分布列为

$$P(X_i = 1) = (n-1)!/n! = 1/n \qquad \text{和} \qquad P(X_i = 0) = 1 - 1/n.$$

于是得到期望

$$E[X] = E[X_1 + X_2 + \cdots, X_n] = E[X_1] + E[X_2] + \cdots + E[X_1] = 1.$$

对任意 $i \neq j$ 有

$$P(X_i = 1, X_j = 1) = (n-2)!/n! = 1/n(n-1),$$

由此得到

$$\mathrm{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] = 1/n^2(n-1),$$

最后根据协方差的性质有

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j) = 1.$$

6.3 相关系数

两个随机变量之间的关系可分为独立与非独立, 在非独立中又可以分为线性关系和非线性关系. 非线性关系较为复杂, 目前尚无好的办法来处理. 但线性相关程度可以通过线性相关系数来刻画, 下面给出具体的定义:

定义 6.2 设 (X, Y) 为二维随机向量, 若方差 $\mathrm{Var}(X)$ 和 $\mathrm{Var}(Y)$ 存在且都不为零, 则称

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

为 X 与 Y 的 **线性相关系数**, 简称 **相关系数** (correlation coefficient). 若 $\rho_{XY} > 0$, 称 X 与 Y **正相关**; 若 $\rho_{XY} < 0$, 称 X 与 Y **负相关**; 若 $\rho_{XY} = 0$, 称 X 与 Y **不相关**.

相关系数是根据协方差和方差所定义, 与协方差同号, 可以看作是对协方差的一种规范化. 相关系数的很多性质可以通过协方差获得:

- 根据协方差性质 6.9 可知
$$|\rho_{XY}| \leqslant 1,$$
$|\rho_{XY}| = 1$ 的充要条件为 X 与 Y 几乎处处有线性关系 $Y = aX + b$.
- 根据协方差性质 6.8 可知, 若 X 与 Y 相互独立, 则 X 与 Y 不相关 ($\rho_{XY} = 0$), 但反之不成立;
- 随机变量 X 与 Y 不相关, 仅仅表示 X 与 Y 之间不存在线性关系, 可能存在其他关系. 例如, 设随机变量 $X \sim U(-1/2, 1/2)$ 和 $Y = \cos(X)$, 则有

$$\mathrm{Cov}(X, Y) = E(X\cos(X)) - E(X)E(\cos(X)) = E[X\cos(X)] = \int_{-1/2}^{1/2} x\cos(x)dx = 0.$$