Mathematical Expectation

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Mean of a Random Variable

Consider a university having 15,000 students and let X be a random variable representing the number of courses for which a randomly selected student is registered. We can calculate the pmf of X from gathered data a follows.

x	1	2	3	4	5	6	7
Number registered	150	450	1950	3750	5850	2550	300
$f(x)$	.01	.03	.13	.25	.39	.17	.02

The average number of courses per student, or the *average number of* X, results from computing the total number of courses registered by all students and dividing by the total number of students, which can be simplify to

$$\frac{1(150) + 2(450) + 3(1950) + 4(3750) + 5(5850) + 6(2550) + 7(300)}{15,000} = 4.57$$

Since $150/15,000 = f(1)$, $450/15,000 = f(2)$, and so on, an alternative expression for the average number of courses is

$$1 \cdot f(1) + 2 \cdot f(2) + \cdots + 7 \cdot f(7) = \sum_{i=1}^{7} i \cdot f(i)$$

The Expected Value of a Random Variable

> Let X be a random variable with probability distribution $f(x)$. The **mean** or **expected value** of X is
>
> $$E(X) = \mu = \sum_{\forall x} x f(x)$$

Processing math: 100%

> if X is a discrete rv, and
>
> $$E(X) = \mu = \int_{-\infty}^{\infty} x f(x)$$
>
> if X is a continuous rv.

Discrete Example Consider examining a lot of light bulbs from a production line which has 5% of producing a defective bulb, the examination is done until a defective bulb is found. Let X be the number of light bulbs one has to examine until the examination is terminated (including the defective one). The pmf of X is then

$$f(x) = .05 \cdot (.95)^{x-1}, \quad x = 1, 2, 3, \ldots$$

So the average number of bulbs one has to examine, or the expected value of X, is

$$E(X) = \sum_{x=1}^{\infty} x(.05)(.95)^{x-1}$$

To solve this summation, we need a little bit of calculus ticks: Consider

$$\begin{aligned}
\sum_{x=1}^{n} x r^{x-1} &= \frac{d}{dr} \sum_{x=1}^{n} r^x \\
&= \frac{d}{dr}\left(\frac{r^{n+1}-1}{r-1}\right) \\
&= \frac{(r-1)(n+1)r^n - r^{n+1} + 1}{(r-1)^2} \\
\sum_{x=1}^{n} x(.95)^{x-1} &= \frac{(-0.05)(n+1)(0.95)^n - (0.95)^{n+1} + 1}{(-0.05)^2}
\end{aligned}$$

Apply the formula to calculate $E(X)$:

Processing math: 100%

$$
\begin{aligned}
E(X) &= \sum_{x=1}^{\infty} x(.05)(.95)^{x-1} \\
&= (0.05) \lim_{n \to \infty} \frac{(-0.05)(n+1)(0.95)^n - (0.95)^{n+1} + 1}{(0.05)^2} \\
&= -\lim_{n \to \infty} \left[(n+1)(0.95)^n\right] + \frac{1}{0.05} \\
&= -\lim_{n \to \infty} \left[\frac{n+1}{(0.95)^{-n}}\right] + \frac{1}{0.05} \\
&\overset{\text{L'H}}{=} -\lim_{n \to \infty} \left[\frac{1}{-(0.95)^{-n}\ln(0.95)}\right] + \frac{1}{0.05} \\
&= 0 + \frac{1}{0.05} = 20
\end{aligned}
$$

Thus, the average number of bulbs one has to examine before a defective bulb is found is 20 bulbs.

Continuous Example Let X be the random variable that denotes the life in hours of a certain electronic device. The probability density function is

$$
f(x) = \begin{cases} \dfrac{20,000}{x^3} & \text{if } x > 100 \\ 0 & \text{elsewhere} \end{cases}
$$

The expected value of X is

$$
E(X) = \int_{100}^{\infty} x \frac{20,000}{x^3} dx = \int_{100}^{\infty} \frac{20,000}{x^2} dx = \frac{20,000}{100} = 200.
$$

Therefore, we can expect this device to last, *on average*, 200 hours.

The Expected Value of a Function

Sometimes interest will focus on the expected value of some function $g(X)$ rather than on just $E(X)$. Note that the function $g(X)$ is another random variable that is completely determined by the rv X, whenever X assumes the value x, then $g(X)$ assumes the value $g(x)$.

For example, suppose X is a discrete random variable with probability distribution $f(x)$, for $x = -1, 0, 1, 2$, and $g(X) = X^2$, then

$$P(g(X)=0)=P(X=0)=f(0)$$
$$P(g(X)=1)=P(X=-1)+P(X=1)=f(-1)+f(1)$$
$$P(g(X)=4)=P(X=2)=f(2)$$

and so the probability distribution of $g(X)$ may be written as

$$f_g(y)=\begin{cases} f(0) & \text{if } y=0 \\ f(-1)+f(1) & \text{if } y=1 \\ f(2) & \text{if } y=4 \end{cases}$$

By the definition of the expected value of a random variable, we obtain

$$E[g(X)]=0f(0)+1[f(-1)+f(1)]+4f(2).$$

This result is generalized in the following theorem for both discrete and continuous random variables.

> Let X be a random variable with probability distribution $f(x)$. The **expected value** of the random variable $g(X)$ is
>
> $$\mu_{g(X)}=E[g(X)]=\sum_{\forall x} g(x)f(x)$$
>
> if X is a discrete rv, and
>
> $$\mu_{g(X)}=E[g(X)]=\int_{-\infty}^{\infty} g(x)f(x)\,dx$$
>
> if X is a continuous rv.

Discrete Example A computer store has purchased three computers at $500 apiece. They will sell them at $1,000 apiece. The manufacturer has agreed to repurchase any computers still unsold after a 2 years at $200 apiece.

Let X denotes the number of computers sold within that 2 years period, and suppose that the pmf of X can be defined as follows.

# of computer sold (x)	**0**	**1**	**2**	**3**
probability [$f(x)$]	.1	.2	.3	.4

Let $h(X)$ denotes the profit associated with selling X computers, the given information implies

$$h(X)=\text{revenue}-\text{cost}=1000X+200(3-X)-1500=800X-900.$$

The expected profit is then

Processing math: 100%

$$
\begin{aligned}
E[h(X)] &= \sum_{x=0}^{3} h(x)f(x) \\
&= (-900)(.1) + (-100)(.2) + (700)(.3) + (1500)(.4) \\
&= \$700
\end{aligned}
$$

Continuous Example Two species are competing in a region for control of a limited amount of a certain resource. Let X represent the proportion of the resource controlled by species 1 and suppose X has pdf

$$
f(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}
$$

Then the species that controls the majority of this resource controls the amount

$$
g(X) = \max(X, 1 - X)
$$

The expected amount controlled by the species having majority control is then

$$
\begin{aligned}
E[g(X)] &= \int_{-\infty}^{\infty} g(x)f(x)\,dx \\
&= \int_0^1 \max(x, 1-x) \cdot 1\,dx \\
&= \int_0^{1/2} (1-x)\,dx + \int_{1/2}^{1} x\,dx \\
&= \frac{3}{8} + \frac{3}{8} = \frac{3}{4}
\end{aligned}
$$

Rule of Expected Value

> For a random variable X
>
> $$E(aX + b) = a \cdot E(X) + b.$$

Example From the previous example where we calculate the expected profit $E[h(x)]$ for selling computers, we can simplify the calculation as follows:

$$
E[h(x)] = E(800X - 900) = 800E(X) - 900 = 800(0.2 + 0.6 + 1.2) - 900 = 700.
$$

Jointly Expectation

> Let X and Y be random variables with joint probability distribution $f(x, y)$. The **expected value** of the random variable $g(X, Y)$ is
>
> $$\mu_{g(X,Y)} = E[g(X, Y)] = \sum_{\forall x} \sum_{\forall y} g(x, y)f(x, y)$$

Processing math: 100%

> if X and Y are discrete rv's, and
>
> $$\mu_{g(X,Y)} = E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,dxdy$$
>
> if X and Y are continuous rv's.

Discrete Example Suppose that X and Y have the following joint probability function

f(x,y)		x	
		2	**4**
	0	0.10	0.15
y	3	0.20	0.30
	5	0.10	0.15

Let $g(X,Y) = XY^2$. Then the expected value of g is

$$\begin{aligned} E[g(X,Y)] &= \sum_{x \in \set{2,4}} \sum_{y \in \set{1,3,5}} xy^2 f(x,y) \\ &= (2)(1)(0.1) + (2)(9)(0.2) + (2)(25)(0.1) + (4)(1)(0.15) + (4)(9)(0.3) + (4)(25)(0.15) \\ &= 35.2 \end{aligned}$$

Continuous Example Let X and Y be random variables with joint density function

$$f(x,y) = \begin{cases} 4xy & \text{if } 0 < x, y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

and let $Z = \sqrt{X^2 + Y^2}$. So the expected value of Z is

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sqrt{x^2 + y^2} f(x,y)\,dxdy \\ &= \int_0^1\int_0^1 4xy\sqrt{x^2 + y^2}\,dxdy \end{aligned}$$

To solve this, we do the substitution $u = x^2 + y^2$. So $du = 2x\,dx$ and when $x = 0 \rightarrow u = y^2$, $x = 1 \rightarrow u = 1 + y^2$:

$$\begin{aligned} E(Z) &= \int_0^1\int_{y^2}^{1+y^2} 2y\sqrt{u}\,dudy \\ &= \int_0^1 2y\left[\frac{2}{3}u^{3/2}\right]_{y^2}^{1+y^2} dy = \int_0^1 \frac{4}{3}y(1+y^2)^{3/2} - \frac{4}{3}y^4\,dy \\ &= \int_0^1 \frac{4}{3}y(1+y^2)^{3/2}\,dy - \frac{4}{15} \end{aligned}$$

Processing math: 100%

Now we make another substitution $u = 1 + y^2$. So $du = 2y\, dy$ and when $y = 0 \to u = 1$, $y = 1 \to u = 2$. So the expected value of Z determined to be

$$E(Z) = \frac{4}{3}\int_1^2 \frac{1}{2} u^{3/2}\, du - \frac{4}{15} = \frac{4}{15}(2^{5/2} - 1) - \frac{4}{15} \approx 0.975$$

Expected Value of Independent RV's

> If X and Y are two **independent** random variables, then
>
> $$E[XY] = E[X] \cdot E[Y].$$

Example A rectangular frame made by a machine have deviation in it sides length. The width and the height of a frame is represented as rv's W and H having the following joint pdf:

$$f(w, h) = \begin{cases} 1/300 & \text{if } 20 \le w \le 40,\ 15 \le h \le 30 \\ 0 & \text{elsewhere} \end{cases}$$

given that the marginal pdf $u(w)$ and $v(h)$ are greater than zero when $20 \le w \le 40$ and $15 \le h \le 30$ respectively. What is the expected value of the area of this frame WH?

First, let's find the marginal pdf $u(w)$ and $v(h)$:

$$\begin{aligned} u(w) &= \int_{-\infty}^{\infty} f(w, h)\, dh = \int_{15}^{30} 1/300\, dh \\ &= \frac{1}{300} 30 - 15 = 1/20. \\ v(h) &= \int_{-\infty}^{\infty} f(w, h)\, dw = \int_{20}^{40} 1/300\, dh \\ &= \frac{1}{300} 40 - 20 = 1/15. \end{aligned}$$

We can see that W and H are independent since $f(w, h) = u(w)v(h)$, so the expected value of WH is

$$\begin{aligned} E[WH] = E[W]E[H] &= \int_{20}^{40} w \cdot u(w)\, dw \cdot \int_{15}^{30} h \cdot v(h)\, dh \\ &= \frac{1}{20}\left[\frac{w^2}{2}\right]_{20}^{40} \cdot \frac{1}{15}\left[\frac{h^2}{2}\right]_{15}^{30} \\ &= 675. \end{aligned}$$

Therefore, on average, the expected area of the frame is 675 unit square.

Conditional Expectation

> Let X and Y be random variables, if the conditional distribution of X given that $Y = y$ if $f(x|y)$. Then the **conditional expectation** of X given $Y = y$ is
>
> $$\mu_{X|Y} = E[X|Y = y] = \sum_{\forall x} xf(x|y)$$
>
> if X and Y are discrete rv's, and
>
> $$\mu_{X|Y} = E[X|Y = y] = \int_{-\infty}^{\infty} xf(x|y)\, dx$$
>
> if X and Y are continuous rv's.

Example The joint probability distribution of X and Y is

$$f(x, y) = x + y, \quad x \in [0, 1], y \in [0, 1].$$

Find $E[X|Y = 1/4]$.

First, find the marginal distribution $h(y)$ of Y:

$$h(y) = \int_0^1 x + y\, dx = \left[\frac{x^2}{2} + xy\right]_0^1 = y + \frac{1}{2}.$$

Thus, the conditional pdf of X given $Y = y$ is

$$f(x|y) = \frac{f(x, y)}{h(y)} = \frac{x + y}{y + 1/2}, \quad x \in [0, 1], y \in [0, 1].$$

Therefore, the expected value of X given that $Y = 1/4$ is

$$\begin{aligned} E[X|Y = 1/4] &= \int_{-\infty}^{\infty} xf(x|\frac{1}{4})\, dx = \int_0^1 x\frac{x + 1/4}{1/4 + 1/2}\, dx \\ &= \frac{4}{3}\int_0^1 x^2 + \frac{x}{4}\, dx = \frac{4}{3}\left[\frac{x^3}{3} + \frac{x^2}{8}\right]_0^1 \\ &= \frac{11}{18}. \end{aligned}$$

.

Variance and Covariance

The most important measure of variability of a random variable X is obtained by finding the expected value of $g(X) = (X - \mu)^2$. The quantity is referred to as the **variance** of the random variable X or the **variance of the probability distribution of X**.

The Variance of a Random Variable

> Let X be a random variable with probability distribution $f(x)$ and mean μ. The **variance** of X is
>
> $$\mathrm{Var}(X) = \sigma^2 = E[(X-\mu)^2] = \sum_{\forall x} (x-\mu)^2 f(x)$$
>
> if X is a discrete rv, and
>
> $$\mathrm{Var}(X) = \sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$
>
> if X is a continuous rv.

> The positive square root of the variance, σ, is called the **standard deviation** of X.

The quantity $x - \mu$ is called the **deviation of an observation** from its mean. Since the deviations are squared and then averaged, σ^2 will be much smaller for a set of x values that are close to μ than it will be for a set of values that vary considerably from μ.

Discrete Example Let the random variable X represent the number of defective parts for a machine when 3 parts are sampled from a production line and tested. The following is the probability distribution of X.

x	0	1	2	3
$f(x)$	0.51	0.38	0.10	0.01

The expected value of X is

$$\mu = E(X) = (0)(0.51) + (1)(0.38) + (2)(0.10) + (3)(0.01) = 0.61.$$

And so the variance of X is

$$\mathrm{Var}(X) = E(X-\mu)^2 = (0-0.61)^2(0.51) + (1-0.61)^2(0.38) + (2-0.61)^2(0.10) + (3-0.61)^2(0.01) = 0.4$$

Continuous Example The distribution of the amount of gravel (in tons) sold by a particular construction supply company in a given week is a continuous rv X with pdf

$$f(x) = \begin{cases} \frac{3}{2}(1-x^2) & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

 the average amount of gravel sold in tons per week is

$$E(X) = \int_0^1 x \cdot \frac{3}{2}(1 - x^2)\,dx = \frac{3}{2}\int_0^1 x - x^3\,dx$$
$$= \frac{3}{2}\left(\frac{1}{2} - \frac{1}{4}\right) = \frac{3}{8}.$$

And so the variance is

$$\mathrm{Var}(X) = \int_0^1 \left(x - \frac{3}{8}\right)^2 \cdot \frac{3}{2}(1 - x^2)\,dx$$
$$= \frac{3}{2}\int_0^1 \left(x^2 - \frac{3}{4}x + \frac{9}{64}\right)(1 - x^2)\,dx$$
$$= \frac{3}{2}\int_0^1 \left(\frac{9}{64} - \frac{3x}{4} + \frac{55x^2}{64} + \frac{3x^3}{4} - x^4\right)dx$$
$$= \frac{3}{2}\left(\frac{9}{64} - \frac{3}{8} + \frac{55}{192} + \frac{3}{16} - \frac{1}{5}\right) = \frac{19}{320} \approx 0.059.$$

Furthermore, the standard deviation of X is $\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{0.059} = 0.243.$

Shortcut Formula for Variance

> The variance of a random variable X is
>
> $$\sigma^2 = E(X^2) - \left[E(X)\right]^2$$

Example From the previous example where X has the probability distribution defined by

x	0	1	2	3
$f(x)$	0.51	0.38	0.10	0.01

To calculate the variance, we first calculate

$$E(X) = (0)(0.51) + (1)(0.38) + (2)(0.10) + (3)(0.01) = 0.61$$

and

$$E(X^2) = (0)(0.51) + (1)(0.38) + (4)(0.10) + (9)(0.01) = 0.87$$

therefore

$$\sigma^2 = \mathrm{Var}(X) = 0.87 - (0.61)^2 = 0.4979.$$

Processing math: 100%

The Variance of a Function

> Let X be a random variable with probability distribution $f(x)$. The **variance** of the random variable $g(X)$ is
>
> $$\mathrm{Var}(g(X)) = \sigma^2_{g(X)} = E\left\{\left[g(X) - \mu_{g(X)}\right]^2\right\} = \sum_{\forall x}\left[g(x) - \mu_{g(X)}\right]^2 f(x)$$
>
> if X is a discrete rv, and
>
> $$\mathrm{Var}(g(X)) = \sigma^2_{g(X)} = E\left\{\left[g(X) - \mu_{g(X)}\right]^2\right\} = \int_{-\infty}^{\infty}\left[g(x) - \mu_{g(X)}\right]^2 f(x)\,dx$$
>
> if X is a continuous rv.

Example Calculate the variance of $g(X) = 2X + 3$, where X is a random variable with probability distribution

x	0	1	2	3
$f(x)$	$1/4$	$1/8$	$1/2$	$1/8$

First, we find the mean of $g(X)$:

$$\mu_{g(X)} = E(2X + 3) = 2E(X) + 3 = 2(1/8 + 1 + 3/8) + 3 = 6.$$

Now, we have

$$\begin{aligned}\sigma^2_{g(X)} &= E[(2X + 3 - \mu_{g(X)})^2] = E[(2X + 3 - 6)^2] = E(4X^2 - 12X + 9)\\ &= 4E(X^2) - 12E(X) + 9 = 4(13/4) - 12(3/2) + 9 = 4.\end{aligned}$$

Covariance

When two random variables X and Y are not independent, it is frequently of interest to assess **how strongly they are related to one another**. This measurement can be done using *covariance*.

> The **covariance** between two random variables X and Y with joint probability distribution $f(x, y)$ denoted by $\mathrm{Cov}(X, Y)$ or σ_{XY} is
>
> $$\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = \begin{cases} \sum_{\forall y}\sum_{\forall x}(x - \mu_X)(y - \mu_Y)f(x, y) & \text{if } X, Y \text{ are discrete}\\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - \mu_X)(y - \mu_y)f(x, y)\,dxdy & \text{if } X, Y \text{ are continuous}\end{cases}$$

> An alternative formula for calculating the covariance of rv's X and Y is
>
> $$\text{Cov}(X, Y) = E[XY] - E[X]E[Y].$$

The value of $\text{Cov}(X, Y)$ can be interpret as follows:

- **Positive covariance**:
  - If $\text{Cov}(X, Y) > 0$ with large magnitude, there is a **positive relationship** between X and Y, meaning that when one variable increases, the other tends to increase as well.
- **Negative covariance**:
  - If $\text{Cov}(X, Y) < 0$ with large magnitude, there is a **negative relationship** between X and Y, meaning that when one variable increases, the other tends to decrease.
- **Covariance near zero**:
  - If $\text{Cov}(X, Y) = 0$ or very close to 0, there is a **no linear relationship** between X and Y.

Do note that covariance is used to measure **linear relationships** between random variables, so covariance of value zero doesn't imply that the rv's are independent; non-linear relationships might still exist between them. But if two random variable are known to be independent, surely the covariance is zero.

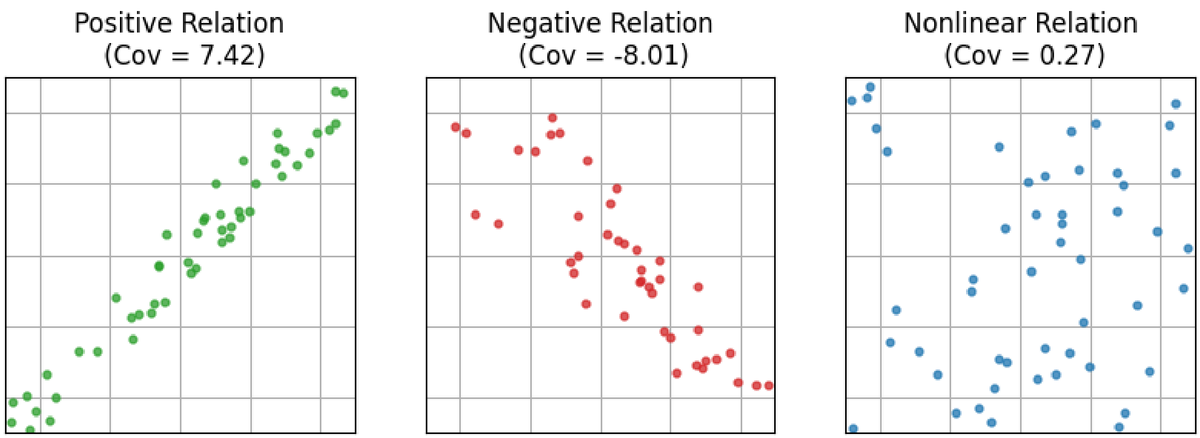
In [2]:
```
fig, axs = plt.subplots(1, 3, figsize=(10,3))

for ax in axs:
    ax.grid()
    ax.set(xlim=[-5,5], ylim=[-5,5], axisbelow=True)
    ax.tick_params(
        axis='both', labelbottom=False, labelleft=False,
        bottom=False, left=False
    )

np.random.seed(0)
X  = (np.random.rand(50) - 0.5) * 10
Y1 =  X + (np.random.rand(50) - 0.5) * 2
Y2 = -X + (np.random.rand(50) - 0.5) * 5
Y3 = (np.random.rand(50) - 0.5) * 10

axs[0].scatter(X, Y1, c='tab:green', s=12, alpha=0.75)
axs[1].scatter(X, Y2, c='tab:red'  , s=12, alpha=0.75)
axs[2].scatter(X, Y3, c='tab:blue' , s=12, alpha=0.75)

axs[0].set(title=f"Positive Relation\n (Cov = {np.cov(X, Y1)[0,1]:.2f})")
axs[1].set(title=f"Negative Relation\n (Cov = {np.cov(X, Y2)[0,1]:.2f})")
axs[2].set(title=f"Nonlinear Relation\n (Cov = {np.cov(X, Y3)[0,1]:.2f})");
```



Correlation

Although the covariance between two random variables does provide information regarding the nature of the relationship, the magnitude of $\mathrm{Cov}(X, Y)$ does not indicate anything regarding the strength of the relationship, since $\mathrm{Cov}(X, Y)$ is *not scale-free*. Its magnitude will depend on the units used to measure both X and Y. There is a scale-free version of the covariance called the **correlation coefficient** that is used widely in statistics.

> The **correlation coefficient** of random variables X and Y, denoted by $\mathrm{Corr}(X, Y)$ or ρ_{XY}, is defined by
>
> $$\mathrm{Corr}(X, Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

The interpretation of correlation is the same as covariance, but the value is scale down to the maximum magnitude of 1. So a magnitude closer to 1 imply a stronger linear relationship and a magnitude close to 0 imply that there is no linear relationship.

Example The following table provides the joint probability mass function of X and Y

f(x,y)		x	
		1	**2**
y	**0**	0.55	0.05
	1	0.15	0.25

What is the covariance and correlation between X and Y?

First, we have to calculate the expected values of X and Y. For that we need the marginal pmf of X and Y:

$$
\begin{aligned}
g(x) &= \sum_{\forall y} f(x,y) = f(x,0) + f(x,1) \\
&= \begin{cases} f(1,0) + f(1,1) = 0.55 + 0.15 = 0.70 & \text{if } x = 1 \\ f(2,0) + f(2,1) = 0.05 + 0.25 = 0.30 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
h(y) &= \sum_{\forall x} f(x,y) = f(1,y) + f(2,y) \\
&= \begin{cases} f(1,0) + f(2,0) = 0.55 + 0.05 = 0.60 & \text{if } y = 0 \\ f(1,1) + f(2,1) = 0.15 + 0.25 = 0.40 & \text{if } y = 1 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

So the expected values are

$$E[X] = (1)(0.70) + (2)(0.30) = 1.3 \quad \text{and} \quad E[Y] = (0)(0.60) + (1)(0.40) = 0.4$$

$$E[X^2] = (1)(0.70) + (4)(0.30) = 1.9 \quad \text{and} \quad E[Y^2] = (0)(0.60) + (1)(0.40) = 0.4$$

Thus, the covariance can be calculated as

$$
\begin{aligned}
\mathrm{Cov}(X,Y) &= \sum_{x \in \backslash\text{set}1,2} \sum_{y \in \backslash\text{set}0,1} (x - E[X])(y - E[Y]) f(x,y) \\
&= (1-1.3)(0-0.4)(0.55) + (1-1.3)(1-0.4)(0.15) + (2-1.3)(0-0.4)(0.05) + (2-1.3)(1- \\
&= 0.13
\end{aligned}
$$

Finally, we can calculate the correlation coefficient of X and Y:

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{0.13}{\sqrt{(1.9 - 1.3^2)(0.4 - 0.4^2)}} \approx 0.58$$

Properties of Variance & Covariance

To derive some of the important properties of variance, let's consider X and Y being random variables with joint probability distribution $f(x,y)$, so

$$\mathrm{Var}(aX + bY + c) = E[(aX + bY + c - \mu_{aX+bY+c})^2]$$

Now

$$
\begin{aligned}
\mu_{aX+bY+c} = E[aX + bY + c] &= aE[X] + bE[Y] + c \\
&= a\mu_X + b\mu_Y + c
\end{aligned}
$$

Substitute the value back in:

$$\begin{aligned}\operatorname{Var}(aX+bY+c) &= E[(aX+bY+c-a\mu_X-b\mu_Y-c)^2] \\ &= E\set[a(X-\mu_X)+b(Y-\mu_Y)]^2 \\ &= E[a^2(X-\mu_X)^2+2ab(X-\mu_X)(Y-\mu_Y)+b^2(Y-\mu_Y)^2] \\ &= a^2E[(X-\mu_X)^2]+b^2E[(Y-\mu_Y)^2]+2abE[(X-\mu_X)(Y-\mu_Y)] \\ &= a^2\operatorname{Var}(X)+b^2\operatorname{Var}(Y)+2ab\operatorname{Cov}(X,Y)\end{aligned}$$

From this result alone, we can derive many of the important properties of variances.

> If X and Y are two random variables with joint probability distribution $f(x,y)$ and a,b,c are constants, then
>
> $$\operatorname{Var}(aX+bY+c)=\sigma^2_{aX+bY+c}=a^2\sigma^2_X+b^2\sigma^2_Y+2ab\sigma_{XY}$$

> Corollaries from the above properties:
>
> - Setting $b=0$, we have
>
> $$\sigma^2_{aX+c}=a^2\sigma^2_X$$
>
> - Setting $b=c=0$ and $a=-a$, we have
>
> $$\sigma^2_{aX}=\sigma^2_{-aX}=a^2\sigma^2_X$$
>
> - If X and Y are independent ($\sigma_{XY}=0$), then
>
> $$\sigma^2_{aX+bY}=\sigma^2_{aX-bY}=a^2\sigma^2_X+b^2\sigma^2_Y$$
>
> - For random variables X and Y,
>
> $$\sigma_{XY}=\operatorname{Cov}(aX,bY)=ab\operatorname{Cov}(X,Y)$$

Chebyshev's Theorem (OPTIONAL)

The Russian mathematician P. L. Chebyshev (1821–1894) discovered that **the fraction of the area between any two values symmetric about the mean is related to the standard deviation**. Since the area under a probability distribution curve or in a probability histogram adds to 1, the area between any two numbers is the probability of the random variable assuming a value between these numbers.

The following theorem, due to Chebyshev, gives a **conservative estimate** of the probability that a random variable assumes a value within k standard deviations of its mean for any real number k.

> **(Chebyshev's Theorem)** The probability that any random variable X will assume a value within k standard deviations from the mean is **at least** $1 - 1/k^2$. That is
>
> $$P(\mu_X - k\sigma_X < X < \mu_X + k\sigma_X) \geq 1 - \frac{1}{k^2}.$$

Chebyshev's theorem holds for any distribution of observations, and for this reason the results are usually weak. The value given by the theorem is a lower bound only. That is, we know that the probability of a random variable falling within two standard deviations of the mean can be no less than 3/4, but we never know how much more it might actually be. Only when the probability distribution is known can we determine exact probabilities. For this reason we call the theorem a **distribution-free result**. When specific distributions are assumed, as in future chapters, the results will be less conservative. **The use of Chebyshev's theorem is relegated to situations where the form of the distribution is unknown**.

Moments of a Random Variable

In statistics, **moments** are numerical measures used to *describe the shape and characteristics of a distribution* or a set of data. They give insight into the properties of the distribution, such as its center, spread, and shape.

> The **k-th moment** about the **origin** of the random variable X with probability distribution $f(x)$ is given by
>
> $$\mu_k^{'} = E[X^k] = \begin{cases} \sum_{\forall x} x^k f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^k f(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

We can see that the first moment about the origin $\mu_1^{'} = E[X]$ which is simply just the expected value of X. We can write the formula for variance of X in terms of first and second moment about the origin as

$$\sigma_X^2 = \mu_2^{'} - (\mu_1^{'})^2 = \mu_2^{'} - \mu^2$$

Most of the time, we will also interest in the moments of a random variable about other values, especially the **central moment** which is the *moment about the mean*.

> The **k-th moment** about the **mean** of the random variable X with probability distribution $f(x)$ is given by
>
> $$E[(X-\mu)^k] = \begin{cases} \sum_{\forall x}(x-\mu)^k f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty}(x-\mu)^k f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

Notice that the first moment about the mean is $E[X-\mu] = E[X] - \mu = 0$, so the *first moment about the mean is always zero* for any random distribution and the second moment about the mean is $E[(X-\mu)^2]$ which is the *variance* of the random variable σ^2.

Moment-Generating Function (OPTIONAL)

Although the moments of a random variable can be determined directly from the definition given in the previous section, an alternative procedure exists. This procedure requires us to utilize a *moment-generating function*.

> The **moment-generating function (MGF)** of the random variable X with probability distribution $f(x)$ is given by $E[e^{tX}]$ and is denoted by $M_X(t)$. Hence
>
> $$M_X(t) = E[e^{tX}] = \begin{cases} \sum_{\forall x} e^{tx} f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{tx} f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

Moment-generating functions will exist only if the sum or integral of the definition converges. If a moment-generating function of a random variable X does exist, it can be used to generate all the moments of that variable.

> Let X be a random variable with moment-generating function $M_X(t)$. Then
>
> $$\left.\frac{d^k M_X(t)}{dr^k}\right|_{t=0} = \mu_k'.$$

To see why the statement above holds, consider the series expansion of e^{tX}:

$$e^{tX} = \sum_{n=0}^{\infty} \frac{t^n X^n}{n!} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^2}{3!} + \cdots + \frac{t^nX^n}{n!} + \cdots$$

Hence, the MGF of X can be expanded as

$$\begin{aligned}M_X(e^{tX}) = E(e^{tX}) &= \sum_{n=0}^{\infty} \frac{t^n E(X^n)}{n!} \\ &= 1 + tE(X) + \frac{t^2 E(X^2)}{2!} + \frac{t^3 E(X^3)}{3!} + \cdots + \frac{t^n E(X^n)}{n!} + \cdots\end{aligned}$$

And therefore,

$$\begin{aligned}\frac{d^k M_X(t)}{dt^k} &= \frac{d^k}{dt^k} \sum_{n=0}^{\infty} \frac{t^n E(X^n)}{n!} = \sum_{n=k}^{\infty} \frac{n!}{(n-k)!} \frac{t^{n-k} E(X^n)}{n!} \\ &= \sum_{n=k}^{\infty} \frac{t^{n-k} E(X^n)}{(n-k)!} \\ &= E(X^k) + tE(X^{k+1}) + \frac{t^2 E(X^{k+1})}{2!} + \frac{t^3 E(X^{k+3})}{3!} + \cdots\end{aligned}$$

so, when evaluating at $t = 0$:

$$\left.\frac{d^k M_X(t)}{dt^k}\right|_{t=0} = E(X^k) = \mu_k'.$$

Properties of Moment-Generating Functions

> Theorem: $M_{X+a}(t) = e^{at} M_X(t)$.
>
> Proof: $M_{X+a}(t) = E[e^{t(X+a)}] = E[e^{tX} e^{ta}] = e^{ta} E[e^{tX}] = e^{ta} M_X t$.

> Theorem: $M_{aX}(t) = M_X(at)$.
>
> Proof: $M_{aX}(t) = E[e^{t(aX)}] = E[e^{(at)X}] = M_X(at)$.

> Theorem: If $X_1, X_2, \ldots, X_n$ are independent random variables with moment-generating function $M_{X_1}(t), M_{X_2}(t), \ldots, M_{X_n}(t)$, respectively, and $Y = X_1 + X_2 + \cdots + X_n$, then
>
> $$M_Y(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t).$$

> Proof:
>
> $$M_Y(t) = E[e^{tY}] = E[\exp(tX_1 + tX_2 + \cdots + tX_n)]$$
> $$= E[e^{tX_1} e^{tX_2} \cdots e^{tX_n}].$$
>
> Since $X_1, X_2, \ldots, X_n$ are independent, then
>
> $$M_Y t = E[e^{tX_1}]E[e^{tX_2}] \cdots E[e^{tX_n}] = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t).$$

Uniqueness Theorem

> Let X and Y be two random variables with moment-generating functions $M_X(t)$ and $M_Y(t)$, respectively. If $M_X(t) = M_Y(t)$ for all values of t, then X and Y have the **same probability distribution**.

We will come back and discuss about moment-generating function again in the future chapter to show that how the concept and properties of moment-generating functions can be used to proof one of the most important theorem in statistics, the *Central Limit Theorem*.