Math Review

This version: September 21, 2023

1 Data v.s. Matrix

In most cases of structured data, you can conceptualize it as similar to an Excel spreadsheet. Each row stands for an individual observation, while each column signifies a variable. Note that the term "observation" differs from "sample." An observation refers to a single record, whereas a sample encompasses the entire set of observations.

ID	age	gender	salary
1	18	male	20k
2	19	male	30k
3	20	female	40k
4	21	female	50k

We can use matrix to represent data:

$$\boldsymbol{X}_{n\times p} = \begin{pmatrix} & \vdots & \\ \cdots & x_{ij} & \cdots \\ & \vdots & \end{pmatrix}_{n\times p}$$

Unless otherwise specified, n denotes the sample size and p denotes the number of variables. The subscript $i = 1, 2, \ldots, n$ refers to the observation and $j = 1, 2, \ldots, p$ to the variable.

For every observation i, we will denote it as

$$X_i = \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1p}, \end{pmatrix}_{p\times 1}$$

i.e., a vector that contains the value of the p variables.

Remark 1. *In this course, every vector is a column vector.*

In addition to traditional data formats like Excel spreadsheets, many types of data can be represented as matrices. For example:

1. Image : x_{ij} = color at $(x, y) = (i, j)$.
2. Text data : x_{ij} = word count of word j of article i.

These matrix representations offer a way to handle different kinds of data in a more generalized mathematical framework.

2 Univariate distribution

For now, we assume $p = 1$, then we have

$$\boldsymbol{X}_{n\times 1} = \begin{pmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{n1} \end{pmatrix}_{n\times 1} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}_{n\times 1},$$

in which we can ignore the j subscript since there is only one variable. In statistics, it is generally assumed that observations are **identically and independently (i.i.d.)** drawn from a population. Mathematically, this is represented as

$$X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F_x(\cdot),$$

where $F_x(\cdot)$ is the cumulative distribution function (CDF). Since $X_1, X_2, \ldots, X_n$ all follow the same distribution, we use X as the random variable to describe their stochastic behavior.[1]

To fully characterize the stochastic behavior of random quantities, we use the CDF. The cumulative distribution function (CDF) of X, $F_X(x)$, is defined as:

$$F_X(x) = P(X \leq x),$$

i.e., the probability that X is below x for $x \in \mathbb{R}$.

The CDF fully characterizes the stochastic behavior in the sense that, once you know CDF, you know

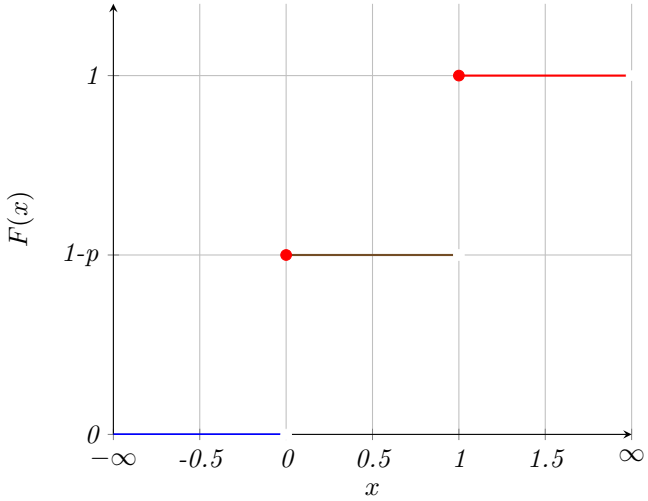
[1] X is different from $\boldsymbol{X}$. The former is a random variable (a scaler) that has the same distribution as $X_i \; \forall \; i = 1, 2, ..., n$, and the latter is the $n \times p$ matrix that contains the data.

the probability that any event would occur. For example, we know

$$\begin{aligned} P(X > a) &= 1 - F_X(a), \\ P(X = a) &= F_X(x) - \lim_{x \to a^-} F_X(x), \\ P(a < X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= F_X(b) - F_X(a). \end{aligned}$$

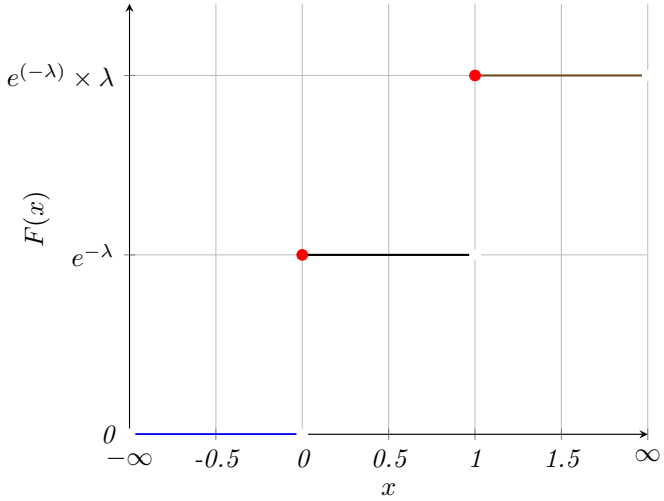
Example 1 (Bernoulli)**.**

$$\begin{aligned} P(X = 1) &= p \\ P(X = 0) &= 1 - p \\ P(X < 0) &= 0 \end{aligned}$$



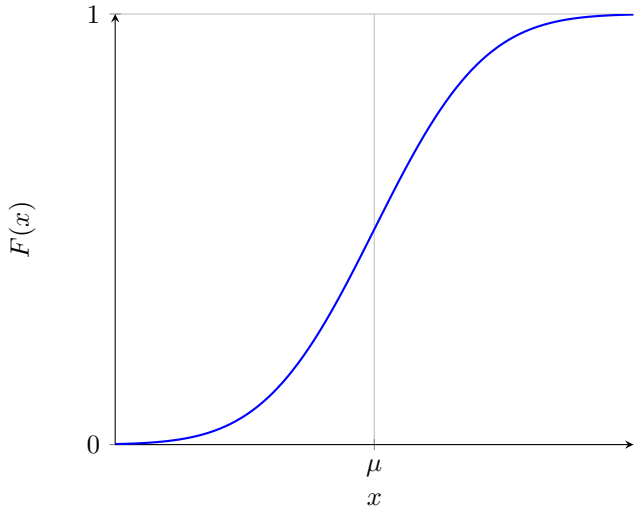
Example 2 (Poisson)**.**

$$P(X = x) = \frac{e^{-\lambda} \times \lambda^x}{x!}, \; x = 0, 1, 2, \cdots$$



Example 3 (Normal)**.**

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx$$



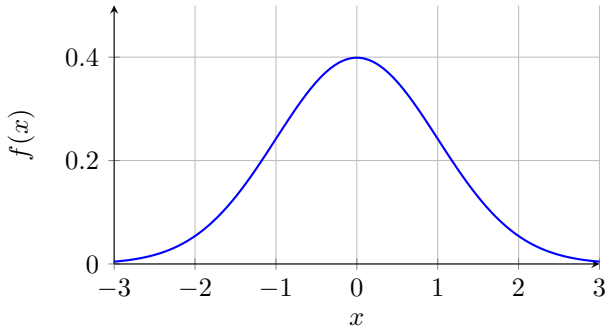
Remark 2. *If the distribution is known, it is often easier to work with the probability mass function (pmf) or the probability density function (pdf). However, pmf only exists for discrete random variables and pdf only exists for continuous random variables. But cdf is universal, i.e., every random variable has a cdf, but not necessarily pdf or pmf.*

Definition 1. *f(x) is said to be a p.d.f of $F_x(x)$ if*

$$P(X \leq x) = \int_{-\infty}^{x} f_x(x)\, dx, \forall x \in \mathbb{R}$$

Example 4 (density of normal)**.**

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}), \forall x \in \mathbb{R}$$



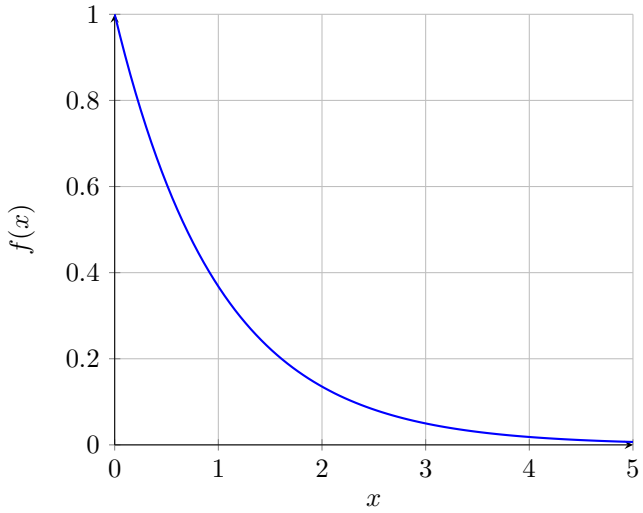
Notice that, pdf is not unique (but cdf is unique). For example, you can check that the function

$$\tilde{f}_X(x) = \begin{cases} f_X(x) & for \quad x \neq 0, \\ 10^{10} & for \quad x = 0. \end{cases}$$

is also a pdf of $\mathcal{N}(0,1)$.

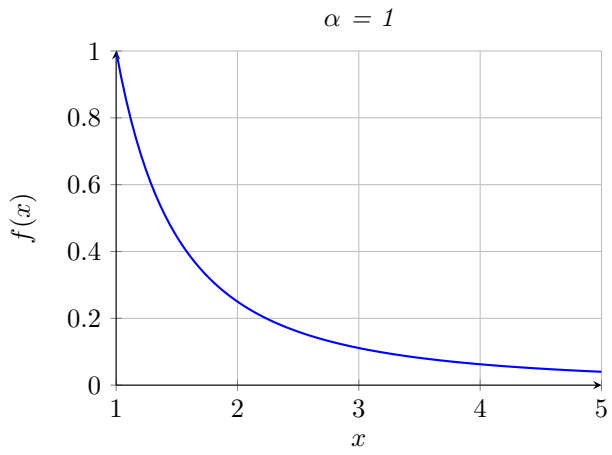
Example 5 (Exponential)**.** *X is a Exponential distribution if its pdf*

$$f_X(x) = \begin{cases} \lambda \times e^{-\lambda x} & for \quad x \geq 0 \\ 0 & for \quad x < 0 \end{cases}$$



Example 6 (Pareto)**.** *X is a Pareto distribution, if its pdf*

$$f_X(x) = \begin{cases} \frac{\alpha}{x^{\alpha+1}} & for \quad x > 0 \\ 0 & for \quad o.w. \end{cases}, \forall \alpha > 0.$$



At first glance, you might think Exponential and Pareto distribution look very similar. Let us compare Pareto distribution and Exponential distribution at the "tail":

	f(x)	f(1)	f(10)	f(100)
exp(1)	e^{-x}	e^{-1}	e^{-10}	e^{-100}
Pareto(1)	$\frac{1}{x^2}$	1	0.01	0.001

You can see that the Pareto distribution is "heavy tail", that is to say, extremely large number happen more often comparing to "thin tail" distribution.

Remark 3. *Distributions are not entirely human constructs. Some of them emerge naturally according to the physical mechanism, for example,*

1. *Normal: the distribution of height*
2. *Exponential: the distribution of lifetime of a light bulb*
3. *Pareto: the distribution of household wealth or city size*

For the examples above, you can find either biological, physical or economical reasons why these distributions emerge. For example, the central limit theorem implies (CLT) that the distribution of heights should be normal.

Example 7. *In data analysis, examining the distribution of variables can provide valuable insights. For instance, by analyzing the distribution of customer lifetime value (CLTV), you can make informed decisions about appropriate acquisition costs for a customer.*

3 Multivariate distribution ($p \geq 2$)

A random vector $\boldsymbol{X}$ is a vector of random variables

$$\boldsymbol{X_i} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}_{p\times 1} \in \mathbb{R}^k.$$

Similar to the univariate case where we use the cumulative distribution function (CDF), for random vectors we employ the joint CDF

$$F_X(x_1, x_2, \cdots, x_p) = P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_p \leq x_p).$$

Example 8 (Multivariate normal)**.** *The random vector* $\boldsymbol{X}$ *has a* multivariate normal distribution

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\mu}_{p\times 1} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{p\times p} = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{12} \\ & \ddots & \\ \sigma_{p1} & \cdots & \sigma_{p^2} \end{pmatrix}_{p\times p},$$

and

$$\sigma_{ij} = \begin{cases} Cov(X_i, Y_j) & for \quad i \neq j \\ Var(X_i) & for \quad i = j. \end{cases}$$

for some $p \times 1$ *vector* $\boldsymbol{\mu}$ *and some* $p \times p$ *positive definite matrix* $\boldsymbol{\Sigma}$*. For multivariate normal, the joint pdf is*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

Recall the univariate normal $X \sim N(\mu, \sigma^2)$ *has pdf*

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

Note that $\boldsymbol{\Sigma}$ *has to be a symmetric matrix since*

$$\begin{aligned} Cov(X_i, Y_i) &= \mathbb{E}[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))] \\ &= \mathbb{E}[(X_j - \mathbb{E}(X_j))(X_i - \mathbb{E}(X_i))] \\ &= Cov(X_j, X_i). \end{aligned}$$

Special case : When $\mathbf{\Sigma}$ is diagonal,

$$\mathbf{\Sigma}_{p\times p} = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \sigma_{p^2} \end{pmatrix}_{p\times p}$$

We can obtain that

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) &= (\boldsymbol{x}-\boldsymbol{\mu})^T \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \sigma_{p^2} \end{pmatrix}_{p\times p} \begin{pmatrix} x_1-\mu_1 \\ \vdots \\ x_p-\mu_p \end{pmatrix}_{p\times 1} \\ &= (\boldsymbol{x}-\boldsymbol{\mu})^T \begin{pmatrix} \sigma_1^2(x_1-\mu_1) \\ \vdots \\ \sigma_p^2(x_p-\mu_p) \end{pmatrix}_{p\times 1} \\ &= \Big((x_1-\mu_1),\cdots,(x_p-\mu_p)\Big) \begin{pmatrix} \sigma_1^2(x_1-\mu_1) \\ \vdots \\ \sigma_p^2(x_p-\mu_p) \end{pmatrix}_{p\times 1} \\ &= \frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} + \cdots + \frac{(x_p-\mu_p)^2}{\sigma_p^2}\end{aligned}$$

and

$$\det(\Sigma) = \sigma_1^2 \times \sigma_2^2 \times \cdots \times \sigma_2^p.$$

The pdf is therefore

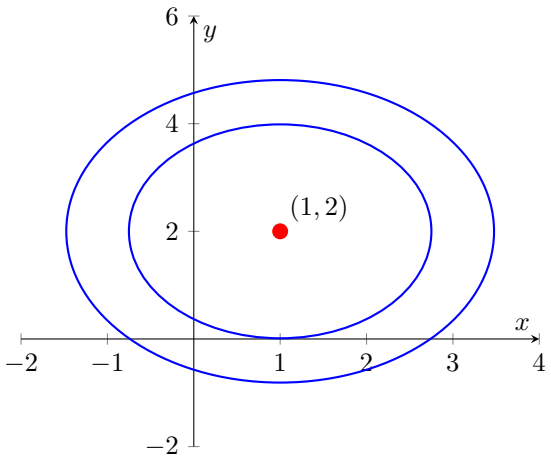
$$\begin{aligned}f_{\boldsymbol{X}}(\boldsymbol{x}) &= (2\pi)^{-\frac{p}{2}} \det(\mathbf{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\ &= (2\pi)^{-\frac{p}{2}} \left(\sigma_1^2\times\sigma_2^2\times\cdots\times\sigma_2^p\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\ &= (2\pi)^{-\frac{p}{2}} \left(\sigma_1^2\times\sigma_2^2\times\cdots\times\sigma_2^p\right)^{-\frac{1}{2}} \exp\left(\frac{-(x_1-\mu_1)^2}{2\sigma_1^2}\right)\times\cdots\times\exp\left(\frac{-(x_p-\mu_p)^2}{2\sigma_p^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right)\times\cdots\times\frac{1}{\sqrt{2\pi\sigma_p^2}}\exp\left(-\frac{(x_p-\mu_p)^2}{2\sigma_p^2}\right) \\ &= f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_p}(x_p) \\ &= \mathcal{N}(\mu_1,\sigma_1^2)\mathcal{N}(\mu_2,\sigma_2^2)\cdots\mathcal{N}(\mu_p,\sigma_p^2).\end{aligned}$$

So, Σ is diagonal, the pdf of the multivariate normal is just the product of the pdf of univariate normal.

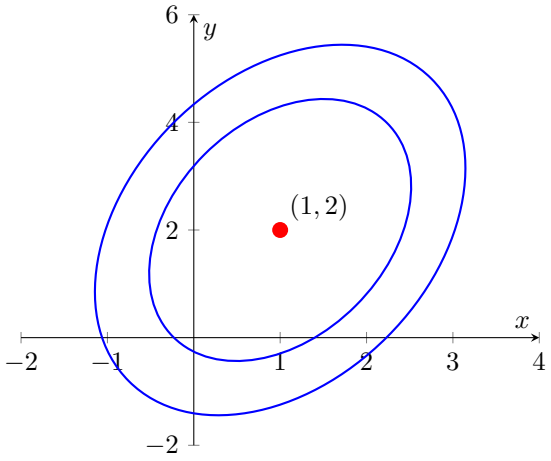
Specifically, when $p = 2$,

$$\begin{aligned} f_{\boldsymbol{X}}(x_1, x_2) =& \frac{1}{\sqrt{2\pi\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2} \right) \\ =& f_{X_1}(x_1) f_{X_2}(x_2). \end{aligned}$$

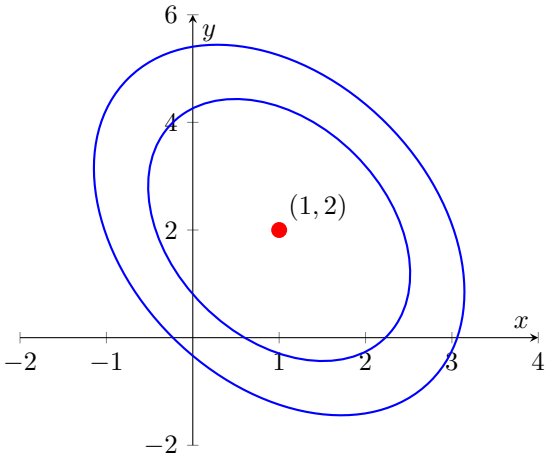
Example 9. *Suppose* $\mu = (1, 2)^T$, $p = 2$, Σ *is diagonal, and* $\sigma_1 > \sigma_2$, *the pdf will look like*



Example 10. *Suppose* Σ *is not diagonal and* $Cov(X_1, X_2) > 0$, *the pdf will look like*



Example 11. *For* $Cov(X_1, X_2) < 0$*:*



3.1 Properties of Multivariate Normal

Given a $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have the following properties:

1. If $Cov(X_i, X_j) = 0$ then $X_i \perp X_j$.[2]
2. Let $\boldsymbol{A}$ be a $k \times p$ real matrix, $\boldsymbol{B}$ be a $k \times 1$ real vector, then

$$\boldsymbol{AX} + \boldsymbol{B} \sim \mathcal{N}_k(\boldsymbol{A\mu} + \boldsymbol{B}, \boldsymbol{A\Sigma A'}).$$

3. $X_1 \mid X_2, \ldots, X_p$ is also normal.

Remark 4. *The second properties really means the following three parts:*

1. *The random vector $\boldsymbol{AX} + \boldsymbol{B}$ possesses a mean of $\boldsymbol{A\mu} + \boldsymbol{B}$.*
2. *The covariance matrix of the random vector $\boldsymbol{AX} + \boldsymbol{B}$ is given by $\boldsymbol{A\Sigma A'}$.*
3. *The distribution of $\boldsymbol{AX} + \boldsymbol{B}$ retains multivariate normal.*

While the first two statements apply to any distribution (you will check this in HW2), the uniqueness of the multivariate normal distribution is the third statement.

Quiz 1. *True or false:*

[2]The notation $\perp$ denotes "independence". It's worth noting that this property is specific to the multivariate normal distribution and does not generally hold for other distributions.

1. *Let*

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

If $X_1, X_2, \cdots, X_p$ are pairwise independent (i.e., $X_i \perp X_j, \forall i \neq j$). Are $X_1, X_2, \cdots, X_p$ independent? P.S. Pairwise independence is not the same as joint independence.

2. *Let X_1, X_2 be two normal random variables. Does $Cov(X_1, X_2) = 0$ imply $X_1 \perp X_2$?*

3. *X_1, X_2 are two normal random variable. Is $X_1 + X_2$ also normal?*

Example 12. *(Joint distribution of two normal random variables are not necessarily normal) Let W, Z be two independent variables that follow the standard normal distribution. Consider the following two random variables:*

$$X = W,$$
$$Y = sgn(W) \cdot |Z|,$$

where the "sign" function $sgn(\cdot)$ is defined as:

$$sgn(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0, \end{cases}$$

that is, the sign function extracts the sign of a number. You can show that X and Y both follow normal distribution but (X, Y) are not jointly normal.

4 Independence

[3]

The two events E, F are independent if

$$P(E \cap F) = P(E)P(F).$$

[3] This part was not taught in class. You can read it by your own.

Then we can obtain that

$$\begin{aligned} P(E \mid F) &= \frac{P(E \cap F)}{P(F)} \\ &= \frac{P(E)P(F)}{P(F)} \\ &= P(E) \end{aligned}$$

and

$$\begin{aligned} P(E \cap F) &= P(E \mid F)P(F) \\ &= P(E)P(F). \end{aligned}$$

Example 13. *Consider the outcome of rolling a dice, E = {odd number}, F = {multiples of 3}.*

$$P(E) = P(\{1,3,5\}) = \frac{1}{2}$$

$$P(F) = P(\{3,6\}) = \frac{1}{3}$$

$$P(E \cap F) = P(\{3\}) = \frac{1}{6} = P(E)P(F)$$

From calculation, we can obtain that E and F are independent.

Quiz 2. *Consider two event, E = {1,2,3}, F = {4,5,6}. Are these two events independent?*

Two random X, Y are independent if

$$P(X \in E, Y \in F) = P(X \in E)P(Y \in F)$$

for any event E, F $\subset \mathbb{R}$.
And we can obtain that

$$P(X \in E \mid Y \in F) = P(X \in E),$$

which means that "Knowing X tells you nothing about Y".
Notice that $X_1, X_2, X_3, \cdots, X_p$ are independent if

$$P(X_1 \in E_1, X_2 \in E_2, \cdots, X_p \in E_P) = P(X_1 \in E_1)P(X_2 \in E_2) \cdots P(X_p \in E_p)$$

5 Covariance

Covariance is an easy way to measure dependency of two random variables. Let X, Y be two random variables,

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

and the correlation coefficient is

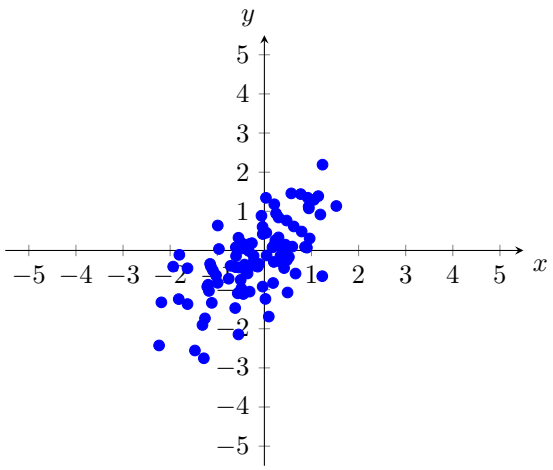
$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

What does it mean for two variables to have a positive correlation? To understand this, we can delve into the definition of covariance and correlation. A positive correlation indicates that:

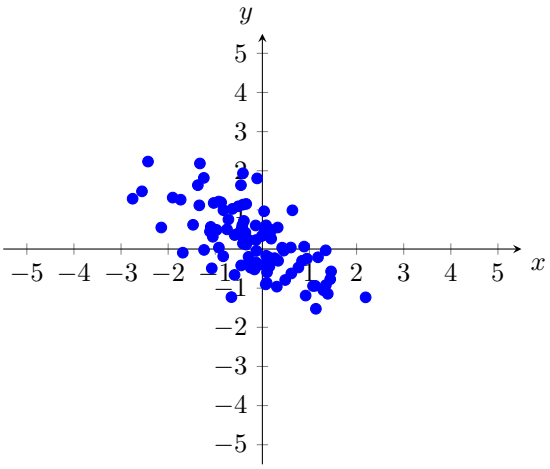
- When the value of X exceeds its expected value ($E[X]$), Y often also exceeds its expected value ($E[Y]$).
- Conversely, when the value of X is below its expected value ($E[X]$), Y frequently falls below its expected value ($E[Y]$) as well.

In essence, a positive correlation signifies that the deviations of X and Y from their respective expected values tend to occur in the same direction.

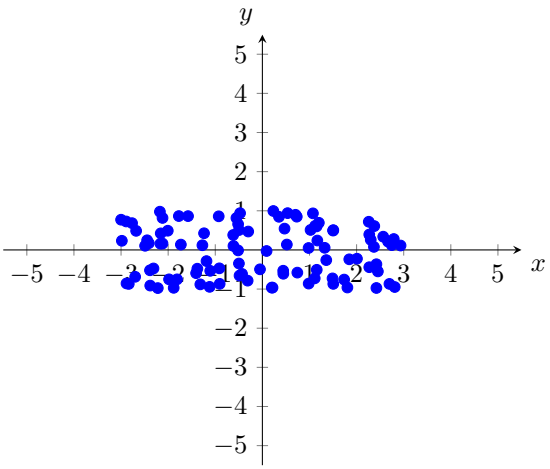
Example 14 (Cov(X,Y) > 0)**.**



Example 15 (Cov(X,Y) < 0)**.**



Example 16 (Cov(X,Y) $= 0$)**.**

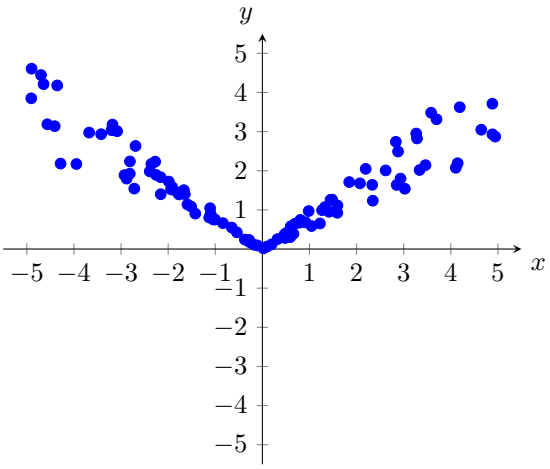


Covariance/ correlation captures "linear dependence". Consider the following example and quiz.

Example 17. *Consider two random variables X and Y. One can show that, if $Y = aX + b$, then $Corr(X, Y) = 1$.*

However, two random variables can be highly (nonlinear) dependent but has zero correlation. See the quiz below.

Quiz 3. *Is* $Cov(X,Y) > 0$*,* $Cov(X,Y) < 0$*, or* $Cov(X,Y) = 0$*?*



In the economics department, we are frequently reminded that correlation does not imply causality. Nevertheless, even if correlation is not causality, it remains useful for several objectives:

1. generating hypothesis
2. risk hedging
3. prediction
4. dimension reduction
5. uncovering hidden structure

Over the next two weeks, our focus will be on exploiting correlation for dimensionality reduction and uncovering hidden structures. Specifically, we will delve into principal component analysis for the former and explore factor analysis for the latter. We discuss the other usages of correlation in below.

5.1 Generating hypothesis

When you get new data, it's a good practice to make a "pair plot". A pair plot is a grid of scatterp lots used to visualize the pairwise relationships between multiple variables in a dataset. Specifically,

- Diagonal: The diagonal of the matrix typically displays histograms or kernel density estimates for each variable. These plots help in understanding the distribution of individual variables.
- Off-diagonal: The off-diagonal plots are scatterplots representing the relationship between two variables. For instance, the scatterplot at the ith row and jth column of the matrix represents the relationship between the ith and jth variables.

A pair plot allows for a quick identification of strong correlations between two variables. If a potential causal relationship between the variables is suspected, it warrants a deeper analysis or the design of a targeted experiment.

5.2 Risk hedging through diversification

Correlation can be leveraged to reduce investment risk. Let's consider two stocks with returns represented by R_1 and R_2. Assume both stocks have identical expected returns and risks:

$$E[R_1] = E[R_2] = r,$$
$$Var(R_1) = Var(R_2) = \sigma^2.$$

Further, suppose their returns are negatively correlated, i.e., $Cov(R_1, R_2) < 0$.

A mixed portfolio, represented as $\tilde{R} = \frac{1}{2}R_1 + \frac{1}{2}R_2$, would then have:

$$E[\tilde{R}] = r,$$

with

$$Var(\tilde{R}) < \sigma^2.$$

This implies that by diversifying investments across these two stocks, one can achieve the same expected return but with a reduced risk.

5.3 Prediction

In machine learning, prediction is the task of guessing an outcome y based on the observations of some other variables $\boldsymbol{x}$. Prediction is essentially exploiting correlation. For example, if it rained yesterday, we might predict rain for today, given the tendency for weather patterns on consecutive days to correlate.

Recommender system is another example of prediction. Our task is to predict what a user may like based on his/ her record. For example, if we find that individuals who favor "Movie X" also typically appreciate "Movie Y" (implying a positive correlation between the ratings of these films), then a user's preference for "Movie X" could prompt a recommendation of "Movie Y". This strategy is usually referred to as "item-to-item" recommendation.

We can also go another way around. If "User X" and "User Y" consistently exhibit analogous tastes (evidenced by a positive covariance in their movie ratings), a liking expressed by "User X" for "Movie Z" might lead to "Movie Z" being recommended to "User Y". This strategy is usually referred to as "user-to-user" recommendation.