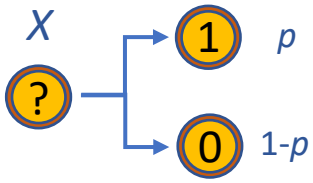
Probability Distributions Cheat Sheet

Prepared by: Reza Bagheri

Bernoulli distribution *(Univariate-Discrete)*

Parameters: p $(0 \le p \le 1)$ **Denoted by:** $X \sim \text{Bern}(p)$

Story: A random variable X with a Bernoulli distribution with parameter p has two possible outcomes labeled by 0 and 1 in which X=1 (success) occurs with probability p and X=0 (failure) occurs with probability 1-p. For example, X can represent the outcome of a coin toss where X=1 and X=0 represent obtaining a head and a tail respectively, and p would be the probability of the coin landing on heads.

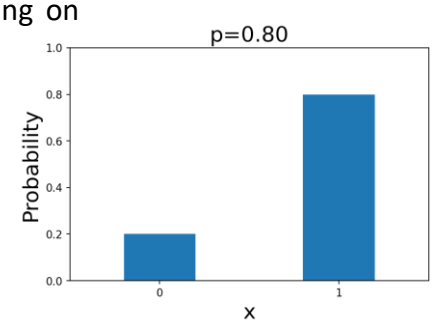


PMF: $$p_X(x) = \begin{cases} p^x(1-p)^{1-x} & for\ x = 0,1 \\ 0 & otherwise \end{cases}$$

Mean: p

Variance: $p(1-p)$

Related distributions: A Bernoulli distribution is a special case of the Binomial distribution when n=1.

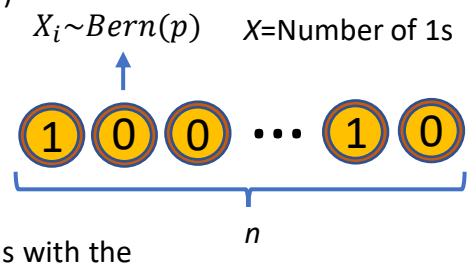


Binomial distribution *(Univariate-Discrete)*

Parameters: n, p $(0 \le p \le 1, n=1,2,\ldots)$ **Denoted by**: $X \sim \text{Bin}(n, p)$

Story: A random variable X with a binomial distribution with the parameters n and p is equal to the sum of n random variables that have a Bernoulli distribution with the parameter p.

$$X = X_1 + X_2 + \cdots + X_n \qquad X_i \sim Bern(p)$$



X represents the total number of successes in n Bernoulli trials with the parameter p. For example, it can represent the total number of heads in n tosses of a coin where the probability of getting heads is p.

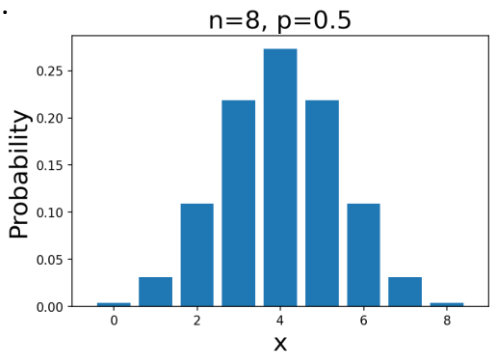
PMF:

$$p_X(x) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x} & for\ x = 0, 1 \ldots, n \\ 0 & otherwise \end{cases}$$

Mean: np

Variance: $np(1-p)$

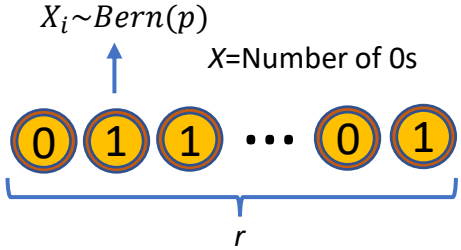
Related distributions: A random variable with a binomial distribution is the sum of n Bernoulli random variables.



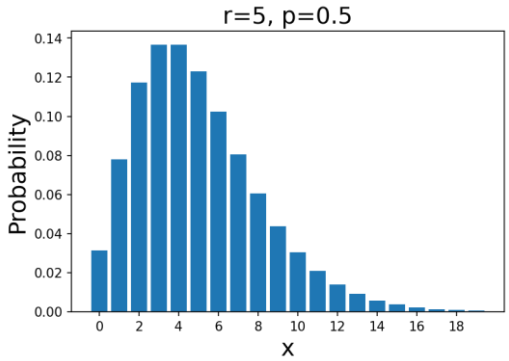
Negative binomial distribution *(Univariate-Discrete)*

Parameters: r, p $(0 \le p \le 1, r=1,2,\ldots)$ **Denoted by**: $X \sim \text{NBin}(r, p)$

Story: Suppose that we have a sequence of Bernoulli trials with the parameter p. A random variable X with a negative binomial distribution with the parameters r and p, represents the number of failures that occur before the rth success.



Negative binomial distribution (Cont'd)



PMF:

$$p_X(x) = \begin{cases} \binom{r+x-1}{x} p^r (1-p)^x & for\ x = 0,1,2,\dots \\ 0 & otherwise \end{cases}$$

Mean: $r(1-p)/p$

Variance: $r(1-p)/p^2$

Related distributions: A random variable with negative binomial distribution with parameters r and p is the sum of r random variable that have a geometric distribution with the parameter p.

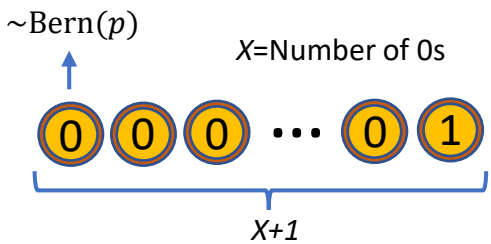
$$X = X_1 + X_2 + \cdots + X_r \quad where\ X_i \sim Geom(p) \rightarrow X \sim NBin(r,p)$$

Geometric distribution

(*Univariate-Discrete*)

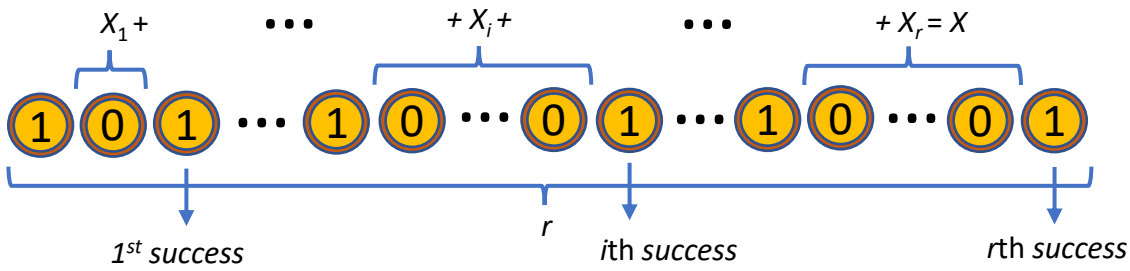
Parameters: p ($0 \le p \le 1$) **Denoted by**: $X \sim Geom(p)$

Story: Suppose that we have a sequence of Bernoulli trials with the parameter p. A random variable X with geometric distribution with the parameter p, represents the number of failures that occur before the 1st success.

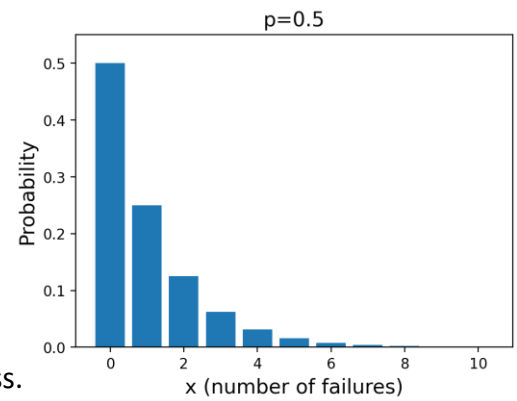


A random variable X which has a negative binomial distribution with parameters r and p can be written as the sum of r random variables that have a geometric distribution with parameter p:

$$X = X_1 + X_2 + \cdots + X_r \quad where\ X \sim NBin(r,p), X_i \sim Geom(p)$$



So, a random variable X with a geometric distribution with the parameter p gives the number of failures between two consecutive successful trials in a sequence of Bernoulli trials with the parameter p.



PMF:

$$p_X(x) = \begin{cases} p(1-p)^x & for\ x = 0,1,2,\dots \\ 0 & otherwise \end{cases}$$

Mean: $(1-p)/p$

Variance: $(1-p)/p^2$

Properties: The geometric distribution is memoryless:

$$P(X \ge n+m | X \ge m) = P(X \ge n)$$

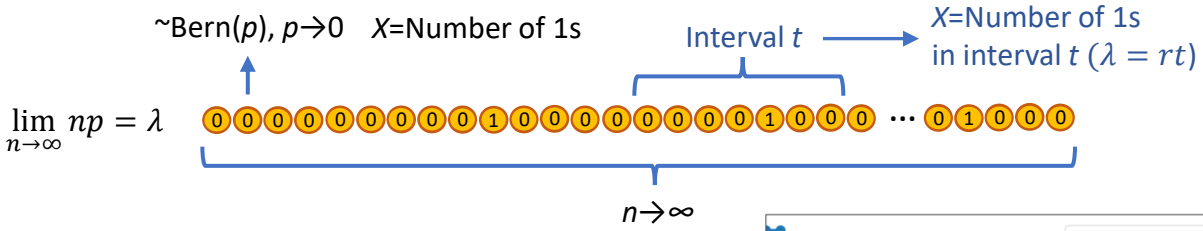
And it is the only discrete distribution which is memoryless.

Related distributions: Geometric distribution is a special case of a negative binomial distribution where r=1.

Poisson distribution (*Univariate-Discrete*)

Parameters: λ (λ>0) **Denoted by**: $X \sim Pois(\lambda)$

Story: The Poisson distribution is a limiting case of the binomial distribution when the number of trials n tends to infinity and p tends to zero while the product $np=\lambda$ remains constant. The parameter λ is also the mean of the distribution, so it gives the average number of the successful events. We can also write $\lambda=rt$ where r is the average rate and t is the time interval for that. Here λ gives the average number of events in the time interval t.

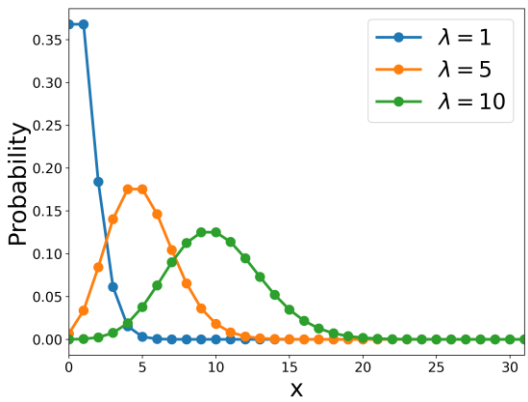


PMF: $p_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & for\ x = 0,1,2,\dots \\ 0 & otherwise \end{cases}$

Mean: λ

Variance: λ

Related distributions: The Poisson distribution is a limiting case of the binomial distribution when the number of trials n tends to infinity and p tends to zero while the product $np=\lambda$ remains constant.



Uniform distribution (*Univariate-Discrete*)

Parameters: a, b (a, $b \in Z$) **Denoted by**: $X \sim DU(a, b)$

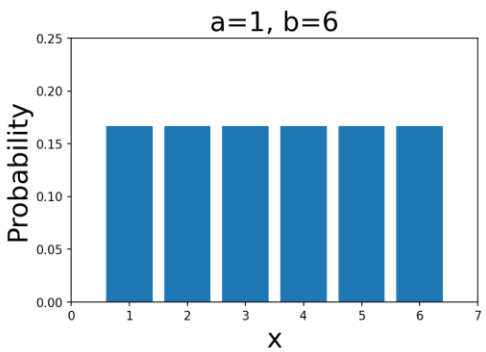
Story: A random variable X with a discrete uniform distribution with parameters a and b can take each of the integers a, a+1..., b with equal probability. So, it is a probability distribution where all outcomes have an equal chance of occurring.

PMF:

$$p_X(x) = \begin{cases} \frac{1}{b-a+1} & for\ x = a, a+1, \dots b \\ 0 & otherwise \end{cases}$$

Mean: $(a+b)/2$

Variance: $\frac{(b-a+1)^2-1}{12}$



Uniform distribution (*Univariate-Continuous*)

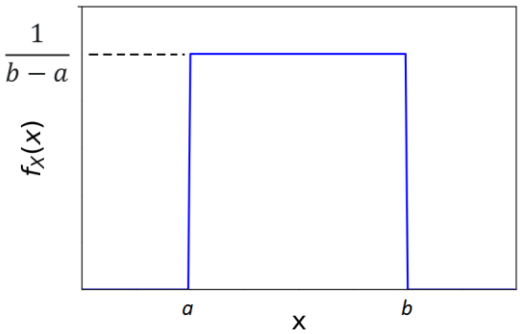
Parameters: a, b **Denoted by**: $X \sim U(a, b)$

Story: If the random variable X has a continuous unform distribution with the parameters a and b, then for every subinterval of [a, b], the probability that X belongs to that subinterval is proportional to the length of that subinterval, and all intervals of the same length are equally probable.

Uniform distribution (Cont'd)

PDF:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & for\ a \le x \le b \\ 0 & otherwise \end{cases}$$



Mean: $(a+b)/2$

Variance: $\frac{(b-a)^2}{12}$

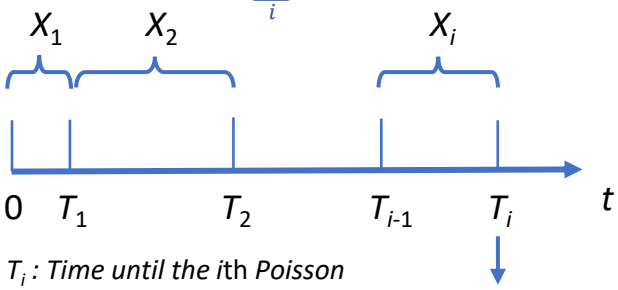
Related distributions : The continuous uniform distribution is a limiting case of the discrete uniform distribution when the number of values that the random variable *X* can take tends to infinity.

Exponential distribution (*Univariate-Continuous*)

Parameters: λ (λ>0) **Denoted by**: $X \sim \text{Exp}(\lambda)$

Story: A random variable *X* with exponential distribution with parameter λ represents the waiting time until the first occurrence of a Poisson event (an event in a Poisson process) with the average rate of λ. It can also represent the waiting time between any two successive events in a Poisson process with the average rate of λ.

$X_i \sim \text{Exp}(\lambda)$, $T_i = \sum_i X_i$, $T_i \sim \text{Gamma}(i, \lambda)$

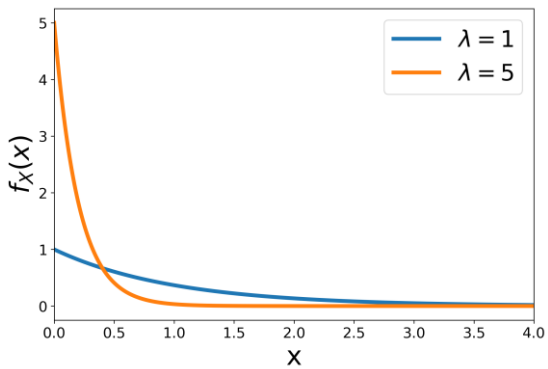


T_i : *Time until the* ith *Poisson event* (average rate of λ)

*i*th *Poisson event*

PDF:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & for\ x > 0 \\ 0 & otherwise \end{cases}$$



Mean: $1/\lambda$

Variance: $1/\lambda^2$

Properties: The geometric distribution is memoryless:

$$P(X > t+s | X > t) = P(x > s) \qquad for\ s,t > 0$$

And it is the only continuous distribution that is memoryless.

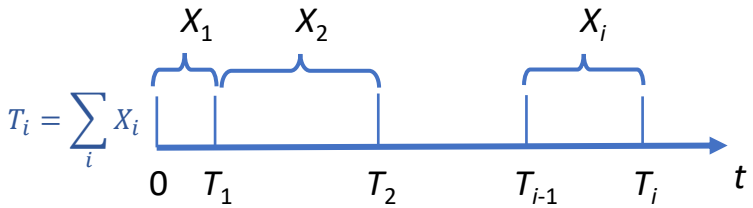
Related distributions : The exponential distribution is a limiting form of the Geometric distribution. Let *X* have a geometric distribution with parameter *p* and let $p=\lambda h$. If $h \to 0$, the random variable *hX* tends in distribution to an exponential random variable with parameter λ. The exponential distribution is a special case of the gamma distribution: Exp(λ)=Gamm(1, λ).

Gamma distribution (*Univariate-Continuous*)

Parameters: α, β (α, β>0) **Denoted by**: $X \sim \text{Gamma}(\alpha, \beta)$

$X_i \sim \text{Exp}(\beta)$, $T_i \sim \text{Gamma}(i, \beta)$

Story: A random variable *T* with gamma distribution with parameters α and β represents the waiting time until the αth occurrence of a Poisson event with an average rate of β (please note that this story is only valid when α is positive integer).



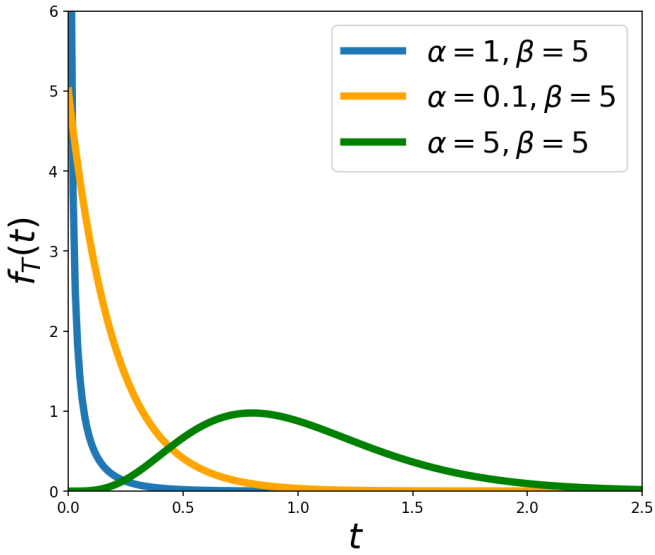
X_i : *Time until the* ith *Poisson event* (average rate of β)

*i*th *Poisson event*

Gamma distribution (Cont'd)

PDF:

$$f_T(t) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & for\ x > 0 \\ 0 & otherwise \end{cases}$$



Mean: α/β

Variance: α/β^2

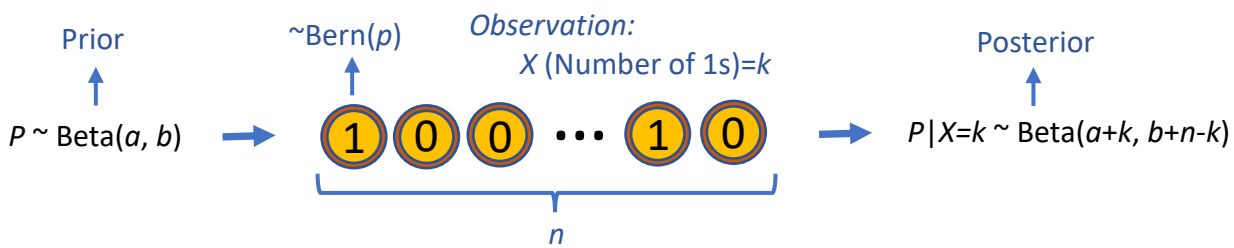
Properties: If X_i has the gamma distribution with parameters α_i and β then the sum $X_1+...+X_k$ has the gamma distribution with parameters $\alpha_1+...+\alpha_k$ and β. If $X_i \sim \text{Exp}(\lambda)$, then the sum $X_1+...+X_k$ has the gamma distribution with parameters k and β.

Related distributions: The gamma distribution is a generalization of the exponential distribution: Exp(λ)=Gamma(1, λ).
If α is a positive integer, the gamma distribution with this α is also called the *Erlang* distribution.
If a random variable X has the gamma distribution with parameters $\alpha = m/2$ and $\beta = ½$ where $m>0$, then X has a chi-square distribution with m degrees of freedom.

Beta distribution (*Univariate-Continuous*)

Parameters: α, β (α, $\beta > 0$) **Denoted by**: $X \sim \text{Beta}(\alpha, \beta)$

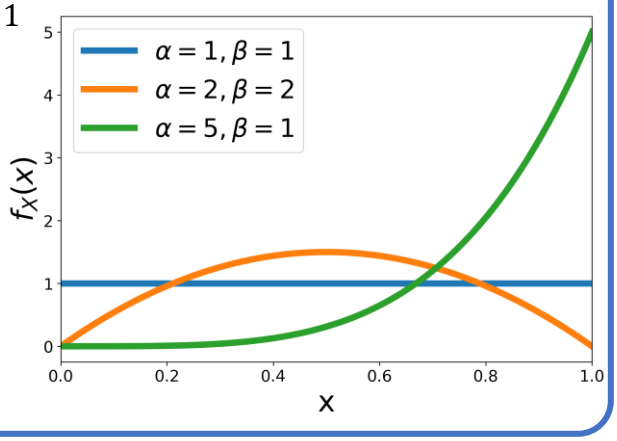
Story: Let X have a binomial distribution with parameters n and p where p is unknown and is represented by the random variable P. For example, let X give the number of head in n tosses of a coin. Before observing the value of X, we assume that P has a Beta(a, b) distribution (prior distribution). If we observe that $X=k$ (in the case of a coin, k is the number of heads and in n tosses), then the posterior distribution of P after this observation is Beta($a+k$, $b+n-k$).



PDF:

$$f_X(x) = \begin{cases} \frac{1}{\text{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & for\ 0 \le x \le 1 \\ 0 & otherwise \end{cases}$$

where $\text{B}(\alpha,\beta) = \Gamma(\alpha+\beta)/\left(\Gamma(\alpha)\,\Gamma(\beta)\right)$



Mean: $\alpha/(\alpha+\beta)$

Variance: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Related distributions: Beta(1, 1) is equal to the uniform distribution on the interval [0, 1].

Beta distribution (Cont'd)

If X ~ Gamma(α, λ) and Y ~ Gamma(β, λ) are independent random variables then $\frac{X}{X+Y} \sim Beta(\alpha, \beta)$. We also have $X+Y$ ~ Gamma(α+β, λ).

If α and β are integers, then X *and* $X+Y$ represent the waiting time until the αth and (α+β)th occurrences of a Poisson event with an average rate of λ, and their ratio has a beta distribution with parameters α and β.

$X \sim Gamma(\alpha, \lambda)$
$Y \sim Gamma(\beta, \lambda)$

$$\frac{X}{X+Y} \sim Beta(\alpha, \beta)$$

Poisson process (average rate of λ)

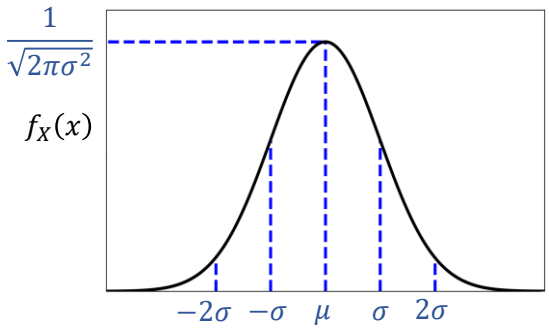
Normal distribution

(*Univariate-Continuous*)

Parameters: μ, σ^2 **Denoted by**: $X \sim N(\mu, \sigma^2)$

Story: The normal distribution is the only distribution that allows us to choose the mean and variance of the distribution as the parameters of the distribution. Hence, they do not depend on each other.

Mean (μ) and variance (σ^2) as independent parameters

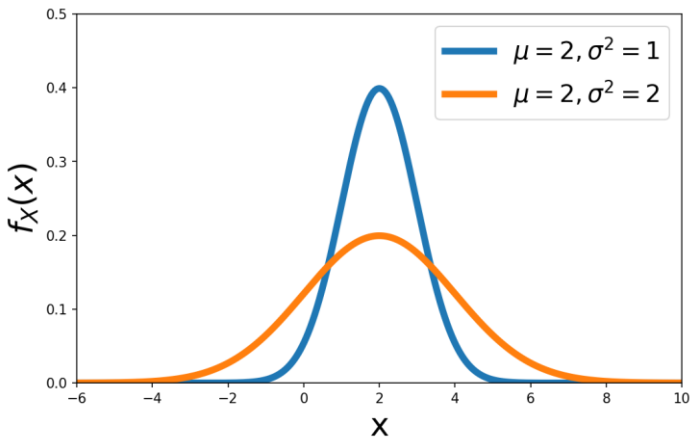


PDF:

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

Mean: μ

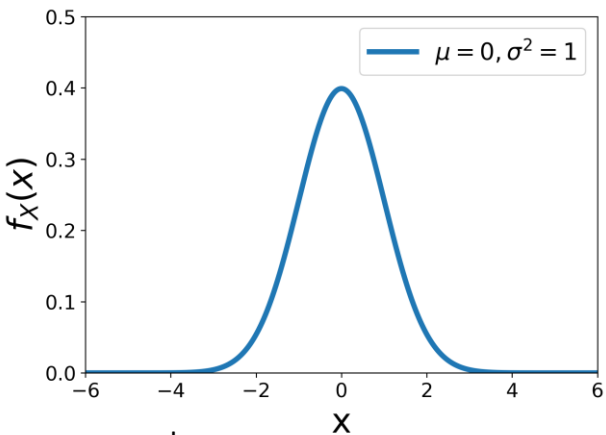
Variance: σ^2



Standard normal distribution: The normal distribution with μ=0 and σ^2 =1 is called the standard normal distribution.

if X has a normal distribution with mean μ and variance σ^2, then the random variable $(X-\mu)/\sigma$ has the standard normal distribution.

Standard normal distribution



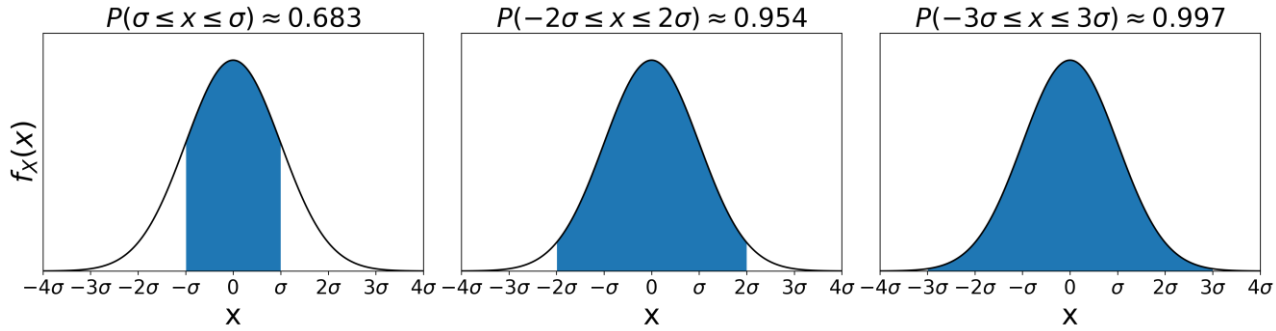
Properties: if X has a normal distribution with mean μ and variance σ^2, the probability that its value falls within one standard deviation of the mean is roughly 0.66

$$P(-\sigma \leq X \leq \sigma) \approx 0.66$$

$P(-\sigma \leq X \leq \sigma)$ is the area under the PDF curve between x=-σ and x=σ.

Similarly, we have: $P(-2\sigma \leq X \leq 2\sigma) \approx 0.95$, $P(-3\sigma \leq X \leq 3\sigma) \approx 0.997$

Normal distribution (Cont'd)

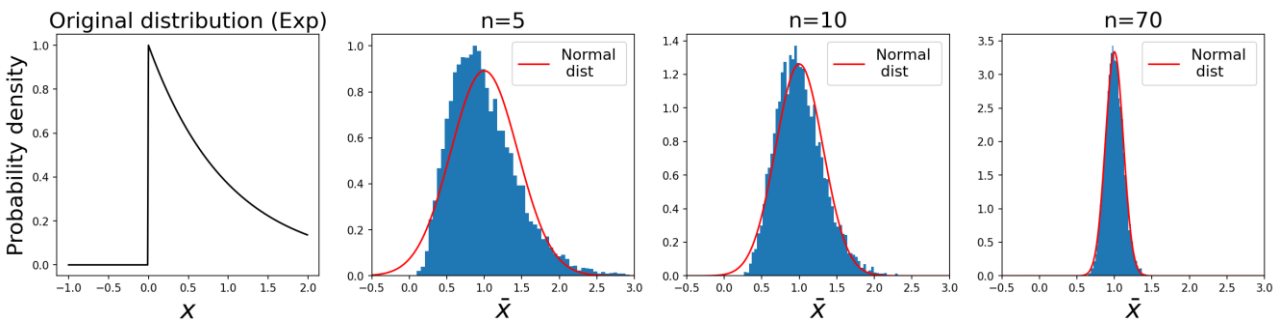


Linear Combinations of Normally Distributed Variables: If the random variables $X_1,\ldots,X_k$ are independent and each X_i has the normal distribution with mean μ_i and variance σ^2 then the sum $a_1X_1+\ldots+a_nX_n+b$ has the normal distribution with mean $a_1\mu_1 + \cdots + a_n\mu_n + b$ and variance $a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2$.

$$X = a_1X_1 + a_nX_n + b \quad X_i \sim N(\mu_i, \sigma_i^2) \longrightarrow X \sim N(a_1\mu_1 + \cdots + a_n\mu_n + b,\ a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2)$$

Central limit theorem (Lindeberg and Levy): If a sufficiently large random sample of size n is taken from any distribution (regardless of whether this distribution is discrete or continuous) with mean μ and variance σ^2, then the distribution of the sample mean ($\bar{X}$) will be approximately the normal distribution with mean μ and variance σ^2/n. In addition, the sum $\sum_{i=1}^{n} X_i$ will be approximately the normal distribution with mean $n\mu$ and variance $n\sigma^2$. As a rule, sample sizes equal to or greater than 30 are usually considered sufficient for the CLT to hold.

$$X_i \sim Any\ disitrubtion\ (\mu, \sigma^2),\ n \to \infty \longrightarrow \bar{X} \sim N(\mu, \sigma^2/n), \quad \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$



Central limit theorem (Lyapunov): This is a more general version of the CLT. Suppose that X_1, X_2, …, X_n, are independent but not necessarily identically distributed, so each of them can have a different distribution. We also assume that the mean and variance of each X_i are μ_i and σ_i^2 respectively. If these two equations are satisfied

$$E[|X_i - \mu_i|]^3 < \infty \quad for\ i = 1,2,\ldots \qquad \lim_{n\to\infty} \frac{\sum_{i=1}^{n} E[|X_i - \mu_i|]^3}{\left(\sum_{i=1}^{n} \sigma_i^2\right)^{3/2}} = 0$$

then for a sufficiently large value of n, the distribution of $X_1+X_2+\ldots+X_n$ will be approximately the normal distribution with mean $\mu_1+\mu_2+\ldots+\mu_n$ and variance $\sigma_1^2+\sigma_2^2+\ldots+\sigma_n^2$.

Related distributions : The normal distribution is related to all other distribution through the central limit theorem. The sum of the square of n independent random variables with a standard normal distribution has a chi-square distribution with n degrees of freedom. The t distribution and F distribution are also defined based on random variables that have the standard normal and chi-square distributions.

Chi-square distribution (*Univariate-Continuous*)

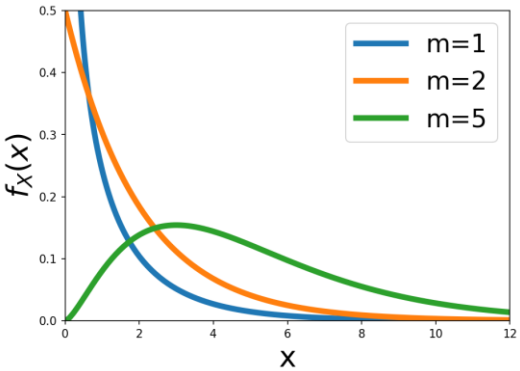
Parameters: *m* (*m*>0) **Denoted by**: $X \sim \chi^2(m)$

$$Z_i \sim N(0,1)$$
$$\uparrow$$
$$X = Z_1^2 + Z_2^2 + \cdots Z_m^2$$
$$\downarrow$$
$$X \sim \chi^2(m)$$

Story: Let the random variable *X* have a gamma distribution with parameters $\alpha = m/2$ and $\beta = ½$ where *m*>0, then *X* has a chi-square distribution with *m* degrees of freedom. In addition, If $Z_1, Z_2, \ldots, Z_m$ are independent, standard normal random variables, then $X = Z_1^2 + Z_2^2 + \cdots Z_m^2$ has a chi-square distribution with *m* degrees of freedom.

PDF:

$$f_X(x) = \begin{cases} \dfrac{1}{2^{m/2}\Gamma\left(\frac{m}{2}\right)} x^{\frac{m}{2}-1} e^{-\frac{x}{2}} & for\ x > 0 \\ 0 & otherwise \end{cases}$$



Mean: *m*

Variance: 2*m*

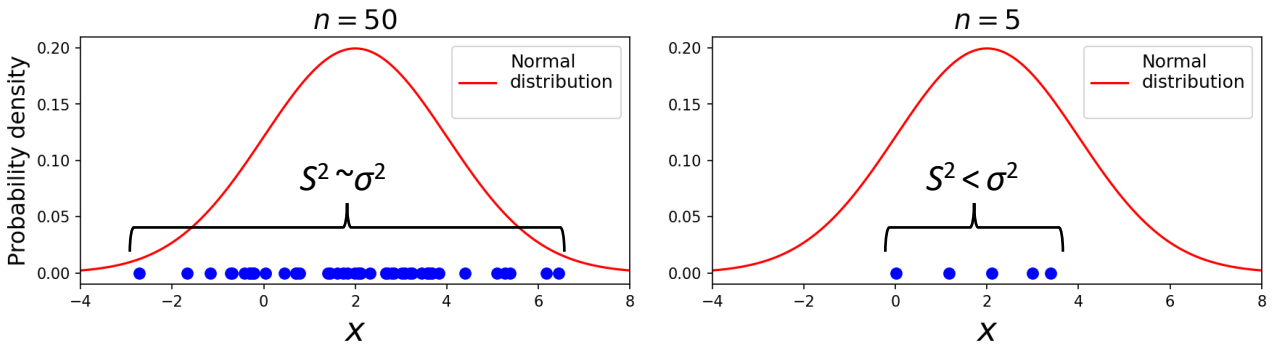
Properties: Let $X_1, X_2, \ldots, X_n$ be a random sample (i.i.d) from a normal distribution with mean μ and variance σ^2, and let S^2 be the sample variance of a sample of size *n* defined as: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)$. Then $\frac{n-1}{\sigma^2}S^2$ has chi-square distribution with *n*-1 degrees of freedom.

Related distributions : A random variable with chi-square distribution with *m* degrees of freedom is the sum of squares of *m* random variables with the standard normal distribution.
If *Z* ~ *N*(0,1) and V $\sim \chi^2(n)$, then the random variable $T=Z/\sqrt{V/n}$ has a *t* distribution with *n* degrees of freedom. We also have $\chi^2(2)$ = Exp(*1/2*).
If $Y_1 \sim \chi^2(d_1)$ and $Y_2 \sim \chi^2(d_2)$, then $(Y_1/d_1)/(Y_2/d_2)$ has the *F* distribution with d_1 and d_2 degrees of freedom.

Student's *t* distribution (*Univariate-Continuous*)

Parameters: *n* (*n*>0) **Denoted by**: $T \sim t_n$

Story: Let the random variable *Z* have the standard normal distribution and the random variable *V* have a chi-square distribution with *n* degrees of freedom, then the random variable $T=Z/\sqrt{V/n}$ has a *t* distribution with *n* degrees of freedom.
A random sample of size *n* from a normal distribution is not a good representative of the original distribution when the sample size is small. That is because the outliers have a smaller chance of occurrence when *n* is small. Hence, there is big chance that the sample variance (S^2) for one specific sample be smaller than the variance of the original distribution.



Student's *t* distribution (Cont'd)

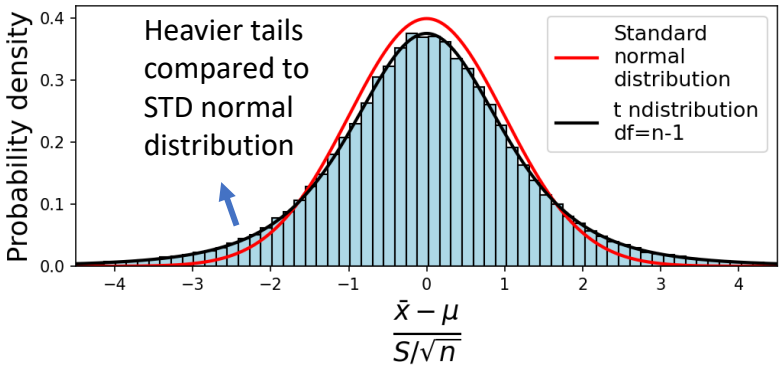
Based on the central limit theorem, the distribution of the sample mean ($\bar{X}$) will be approximately a normal distribution with mean μ and variance σ^2/n, *so* the ratio $(\bar{X}-\mu)/\sigma/\sqrt{n}$ has the standard normal distribution.

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

However, the ratio $(\bar{X}-\mu)/S/\sqrt{n}$ doesn't have the standard normal distribution. For each sample, there is a big chance that $S<\sigma$. So, in $(\bar{X}-\mu)/S/\sqrt{n}$ the extreme values (outliers) have a higher chance of occurrence. As a result, the distribution of $(\bar{X}-\mu)/S/\sqrt{n}$ has heavier tails compared to $(\bar{X}-\mu)/\sigma/\sqrt{n}$.

The ratio $(\bar{X}-\mu)/S/\sqrt{n}$ has a *t* distribution with *n*-1 degrees of freedom

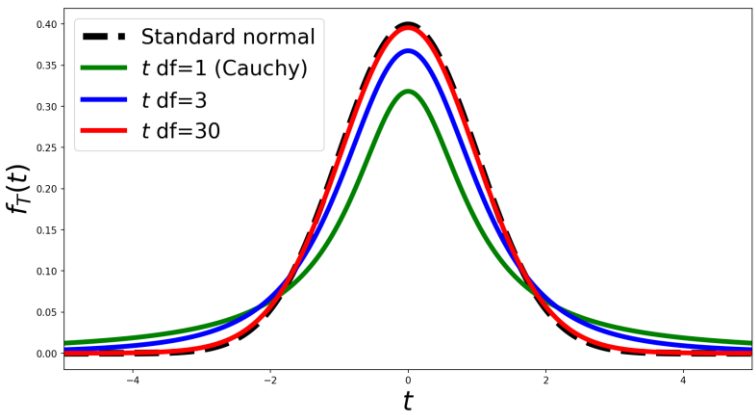
$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$



PDF:

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}(1+t^2/n)^{-(n+1)/2}$$

$$-\infty < t < \infty$$



Mean: 0 for $n>1$, otherwise undefined

Variance: $n/(n-2)$ for $n>2$, ∞ for $1<n\leq 2$, otherwise undefined

Related distributions: As $n\to\infty$, the PDF of the *t* distribution tends to the PDF of the standard normal distribution.
If *T* has the *t* distribution with *n* degrees of freedom, then T^2 has the *F* distribution with 1 and *n* degrees of freedom.

F distribution (*Univariate-Continuous*)

Parameters: d_1, d_2 (d_1, d_2 *are* positive integers)

Denoted by: $X \sim F(d_1, d_2)$

Story: Suppose that Y_1 and Y_2 are two independent random variables such that Y_1 has the chi-square distribution with d_1 degrees of freedom and Y_2 has the chi-square distribution with d_2 degrees of freedom (d_1 and d_2 are positive integers). The random variable $X=(Y_1/d_1)/(Y_2/d_2)$ has the *F* distribution with d_1 and d_2 degrees of freedom.

$$Y_1 \sim \chi^2(d_1)$$

$$X \sim F(d_1, d_2) \longleftarrow X = \frac{Y_1/d_1}{Y_2/d_2}$$

$$Y_2 \sim \chi^2(d_2)$$

F distribution (Cont'd)

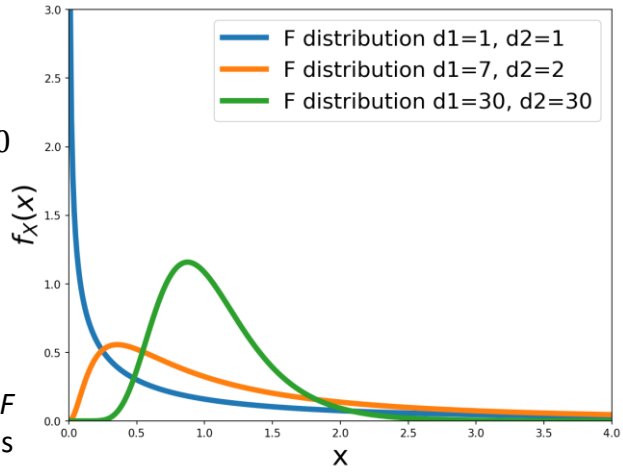
PDF:

$$f_X(x) = \frac{\Gamma\left(\frac{d_1+d_2}{2}\right) d_1^{d_1/2} d_2^{d_2/2} x^{\frac{d_1}{2}-1}}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)(d_1 x + d_2)^{(d_1+d_2)/2}} \quad x > 0$$

Mean: $d_2/(d_2-2)$ for $d_2>2$

Variance: $\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$ for $d_2>4$

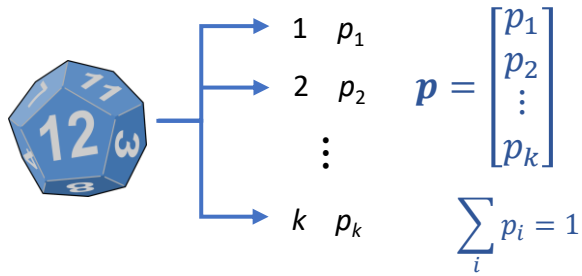
Related distributions: A random variable with *F* distribution is the ratio of two random variables with chi-square distribution.



Multinomial distribution (*Multivariate-Discrete*)

Parameters: *n*, ***p*** ($0 \le p_i \le 1$ (n=1,2,...), $\sum_i p_i = 1$) **Denoted by**: ***X*** ~ Mult(*n*, ***p***)

Story 1: Suppose that we have *n* independent trials. Each trial has *k* (*k*≥2) different outcomes, and the probability of the *i*th outcome is p_i ($\sum_i p_i = 1$). The vector ***p*** denotes these probabilities. Let the discrete random variable X_i represent the number of times the outcome number *i* is observed over the *n* trials. The random vector ***X*** is defined as

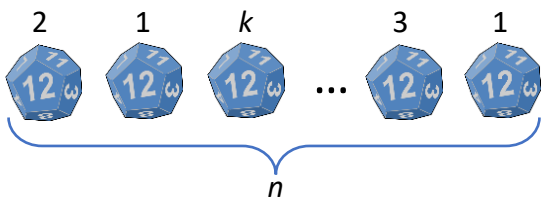


$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}$$

has the multinomial distribution with parameters *n* and ***p***.

X_1=Number of 1s
X_2=Number of 2s
⋮
X_k=Number of *k*s

X ~ Mult(*n*, ***p***)

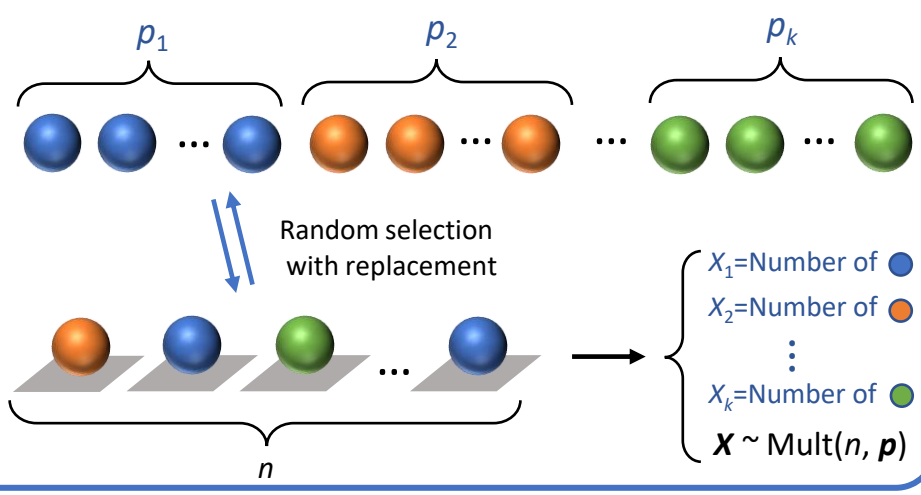


The multinomial distribution can be used to describe a *k*-sided die. Suppose that we have a *k*-sided die, and the probability of getting side *i* is p_i. If we roll it *n* times, and X_i represents the total number of times that side *i* is observed, then ***X*** has a multinomial distribution with parameters *n* and ***p***.

Story 2: We have a population of items of *k* different categories ($k \ge 2$). The proportion of the items in the population that are in category *i* is p_i, and $\sum_i p_i = 1$. Now we randomly select *n* items from the population with replacement, and we

Population

p_i: proportion of the items in category *i*



Multinomial distribution (Cont'd)

assume that the discrete random variable X_i represents the number of selected items that are in category *i*. If we assume that X_i and p_i are the *i*th elements of the vectors $\boldsymbol{X}$ and $\boldsymbol{p}$, then $\boldsymbol{X}$ has a multinomial distribution with parameters *n* and $\boldsymbol{p}$.

Joint PMF:

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = p_{X_1,X_2,\dots,X_k}(x_1, x_2, \dots, x_k)$$
$$= P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) =$$
$$\begin{cases} \frac{n!}{x_1!x_2!\dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k} & if\ x_1 + x_2 + \dots + x_k = n \\ 0 & otherwise \end{cases}$$

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim \text{Mult}(5, \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix})$$

Mean: $E[X_i]=np_i$

Variance: $Var(X_i) = np_i(1 - p_i)$

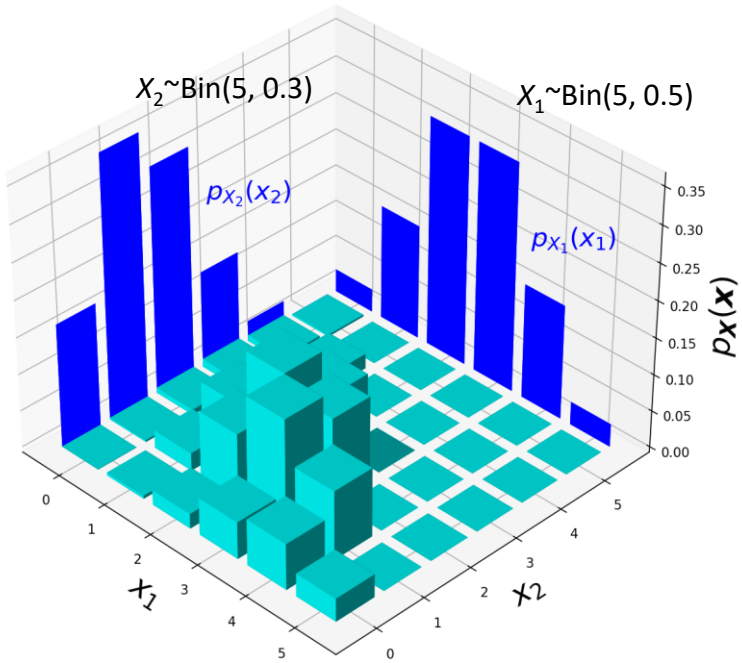
Covariance: $Cov(X_i, X_j) = -np_ip_j$

Properties: We can merge multiple elements in a multinomial random vector to get a new multinomial random vector. For example, if:

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} \sim \text{Mult}(n, \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix})$$

Then

$$\boldsymbol{X} = \begin{bmatrix} X_1 + X_2 \\ X_3 \\ X_4 \end{bmatrix} \sim \text{Mult}(n, \begin{bmatrix} p_1 + p_2 \\ p_3 \\ p_4 \end{bmatrix})$$



A random vector $\boldsymbol{X}$ that has a multinomial distribution with parameters *n* and $\boldsymbol{p}$ can be written as the sum of *n* random vectors that have a multinomial distribution with parameters 1 and $\boldsymbol{p}$:

$$\boldsymbol{X} \sim \text{Mult}(n, \boldsymbol{p}) \longrightarrow \boldsymbol{X} = \boldsymbol{Y}_1 + \boldsymbol{Y}_2 + \dots . + \boldsymbol{Y}_n \quad where \quad Y_i \sim \text{Mult}(1, \boldsymbol{p})$$

Related distributions: The multinomial distribution is a generalization of the binomial distribution. If *k*=2, the multinomial distribution reduces to a binomial distribution:

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad \boldsymbol{p} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} \longrightarrow \begin{matrix} X_1 \sim \text{Bin}(n, p_1) \\ X_2 \sim \text{Bin}(n, p_2) \end{matrix}$$

If $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} \sim \text{Mult}(n, \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{bmatrix})$ and *k*>2 then the marginal distribution of each X_i is a binomial distribution with parameters *n* and p_i: $X_i \sim \text{Bin}(n, p_i)$.

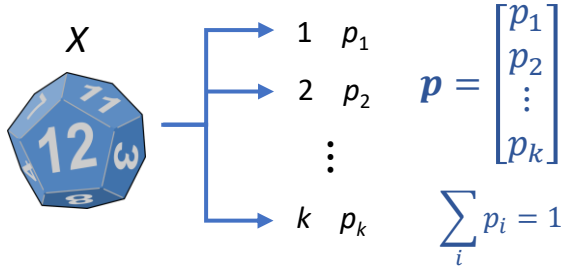
The sum of some of the elements of multinomial random vector $\boldsymbol{X}$ has a binomial distribution. If $X_{i,1}$, $X_{i,2}$, …, $X_{i,m}$ are *m* elements of the random vector $\boldsymbol{X}$ (*m*<*k*) and their corresponding probabilities in vector $\boldsymbol{p}$ are $p_{i,1}$, $p_{i,2}$, …, $p_{i,m}$, then the sum $X_{i,1} + X_{i,2} + \dots + X_{i,m}$ has a binomial distribution with parameters *n* and $p_{i,1} + p_{i,2} + \dots + p_{i,m}$. For example:

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} \sim \text{Mult}(5, \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix}) \longrightarrow X_1 + X_3 + X_4 \sim \text{Bin}(5, p_1 + p_3 + p_4)$$

Categorical or multinoulli distribution (*Discrete*)

Parameters: $\boldsymbol{p}$ ($0 \le p_i \le 1$ (n=1,2,...), $\sum_i p_i = 1$) **Denoted by**: $\boldsymbol{X} \sim \mathrm{Mu}(\boldsymbol{p})$ or $X \sim \mathrm{Cat}(\boldsymbol{p})$

Story 1: Suppose that the random variable X has k possible outcomes labeled by 1 to k, and the probability of the ith outcome is p_i. Then X is said to have a categorical distribution with the parameter $\boldsymbol{p}$, and it is univariate distribution. For example, X can represent the outcome of rolling a k-sided die where $X=i$ represents obtaining side i and p_i is the probability of getting that side.



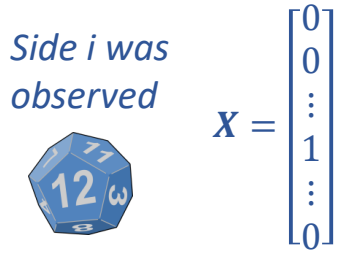
PMF:

$$\begin{cases} f_X(x = i|\boldsymbol{p}) = p_i & for\ i = 1 \dots k \\ f_X(x|\boldsymbol{p}) = 0 & otherwise \end{cases}$$

Story 2: The random vector $\boldsymbol{X}$ has a categorical or multinoulli distribution, if $\boldsymbol{X}$ has a multinomial distribution with n=1. Hence it is a special case of the multinomial distribution:

$$\boldsymbol{X} \sim \mathrm{Mult}(1, \boldsymbol{p}) \iff \boldsymbol{X} \sim \mathrm{Mu}(\boldsymbol{p})$$

And it is a multivariate distribution in this case. Here $\boldsymbol{X}$ is one-hot encoded vector in which only one element is one and the other elements are zero. If we observe the ith outcome then X_i=1 and the other elements of $\boldsymbol{X}$ are zero.



Joint PMF: $p_{\boldsymbol{X}}(\boldsymbol{x}) = \begin{cases} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k} & if\ x_i = 1, x_j = 0\ \ j = 1 \dots k, j \neq i \\ 0 & otherwise \end{cases}$

Mean: $E[X_i]=p_i$

Variance: $Var(X_i) = p_i(1 - p_i)$

Related distributions : Based on story 1, it is a generalization of Bernoulli distribution story 1, and based on story 2, it is a special case of multinomial distribution for n=1.

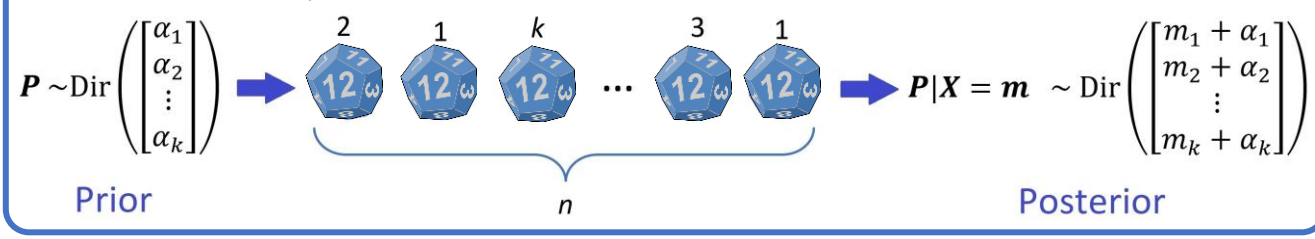
Dirichlet distribution (*Multivariate-Continuous*)

Parameters: $\boldsymbol{\alpha}$ (α_i>0, number of elements of $\boldsymbol{\alpha}$ ≥2) **Denoted by**: $\boldsymbol{X}$ ~ Dir($\boldsymbol{\alpha}$)

Story: Let $\boldsymbol{X}$ have a binomial distribution with parameters n and $\boldsymbol{p}$ where $\boldsymbol{p}$ is unknown and is represented by the random vector $\boldsymbol{P}$. For example, let each element of $\boldsymbol{X}$ (X_i) represent the number of times that the side i is observed in n rolls of a k-sided die. Before observing the value of $\boldsymbol{X}$, we assume that $\boldsymbol{P}$ has a Dir($\boldsymbol{\alpha}$) distribution (prior distribution). If we observe that $\boldsymbol{X}=\boldsymbol{m}$ (in the case of a die, m_i is the number of times that side i is observed) then the posterior distribution of $\boldsymbol{P}$ after this observation is Dir($\boldsymbol{\alpha}+\boldsymbol{m}$).

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} \qquad \boldsymbol{X}|\boldsymbol{p} \sim \mathrm{Mult}\left(n, \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{bmatrix}\right)$$

$$Observations: \begin{cases} m_1: Number\ of\ 1s \\ m_2: Number\ of\ 2s \\ \vdots \\ m_k: Number\ of\ ks \end{cases}$$



Dirichlet distribution (Cont'd)

Joint PDF: $p_{\boldsymbol{X}}(\boldsymbol{x}) = \begin{cases} \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} x_1^{\alpha_1-1} x_2^{\alpha_2-1} \dots x_k^{\alpha_k-1} & if\ 0 \le x_i \le 1\ and\ \sum_{i=1}^{k} x_i = 1\ (i = 1 \dots k) \\ 0 & otherwise \end{cases}$

where $\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)} = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\dots\Gamma(\alpha_k)}{\Gamma(\alpha_1+\alpha_2+\dots+\alpha_k)}$

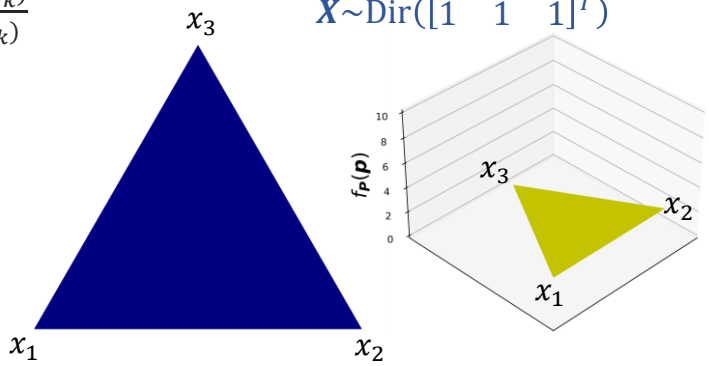
Mean: $E[\boldsymbol{X}] = \frac{1}{\alpha_0} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix}$

where $\alpha_0 = \sum_{i=1}^{k} \alpha_i$

Variance: $Var(X_i) = \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$

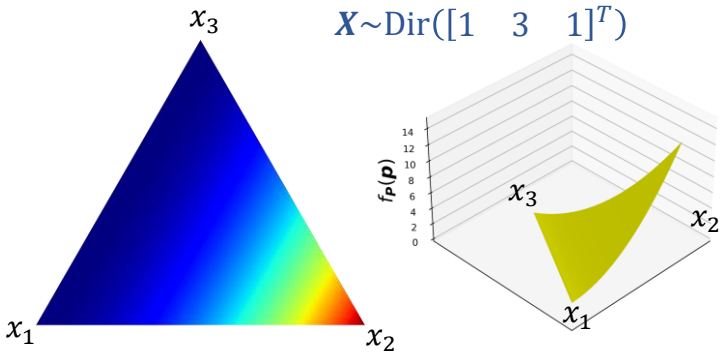
Covariance: $Cov(X_i, X_j) = \frac{-\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)} \quad i \ne j$

$\boldsymbol{X}\sim\mathrm{Dir}([1 \quad 1 \quad 1]^T)$



Properties: If the *k*-dimensional random vector ***X*** has a Dirichlet distribution then the support of this distribution is a *k*-1 dimensional simplex. The Dirichlet distribution $\mathrm{Dir}([1 \quad 1 \quad \dots \quad 1]^T)$ is the same as the uniform distribution over its *k*-1 dimensional simplex since the joint PDF has the same value over the simplex.

$\boldsymbol{X}\sim\mathrm{Dir}([1 \quad 3 \quad 1]^T)$



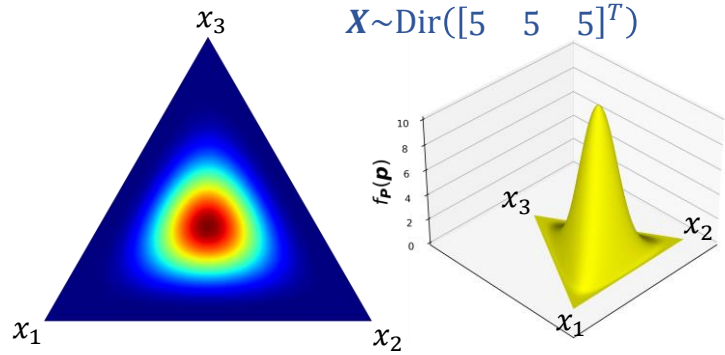
Aggregation property: Let the random vector ***X*** have the following Dirichlet distribution:

$$\boldsymbol{X} = [X_1 \quad \dots \quad X_i \quad \dots \quad X_j \quad \dots \quad X_k]^T \sim \mathrm{Dir}([\alpha_1 \quad \dots \quad \alpha_i \quad \dots \quad \alpha_j \quad \dots \quad \alpha_k]^T)$$

We drop the random variables X_i and X_j from ***X*** and add X_i+X_j to it at an arbitrary place and call the resulting random vector ***X'***. The random vector ***X'*** has the following Dirichlet distribution:

$$\boldsymbol{X}' = [X_1 \quad \dots \quad X_i + X_j \quad \dots \quad X_k]^T \sim \mathrm{Dir}([\alpha_1 \quad \dots \quad \alpha_i + \alpha_j \quad \dots \quad \alpha_k]^T)$$

$\boldsymbol{X}\sim\mathrm{Dir}([5 \quad 5 \quad 5]^T)$



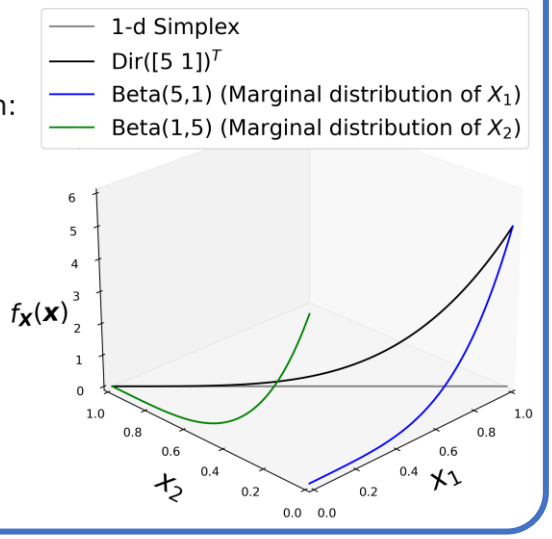
Marginal distributions: Let ***X*** have a Dirichlet distribution:

$$\boldsymbol{X} = [X_1 \quad X_2 \quad \dots \quad X_k]^T \sim \mathrm{Dir}([\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_k]^T)$$

Then the marginal distribution of each X_i is the following beta distribution:

$$X_i \sim \mathrm{Beta}(\alpha_i, \alpha_0 - \alpha_i)$$

where $\alpha_0 = \sum_{i=1}^{k} \alpha_i$



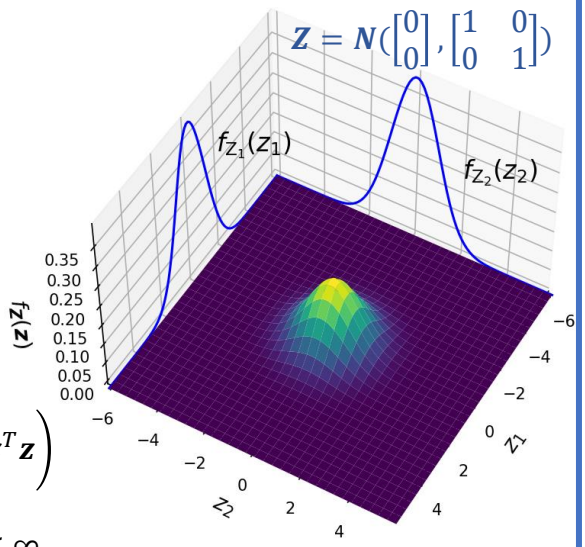
Related distributions: The Dirichlet distribution is a generalization of the beta distribution. If ***X*** has a Dirichlet distribution, the marginal distribution of each random variable in ***X*** is a beta distribution.

Standard multivariate normal distribution (*Multivariate-Continuous*)

Parameters: *NA* **Denoted by**: $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I})$

Story: Suppose that we have n independent random variables $Z_1, Z_2, \dots Z_n$, and each of them has a standard normal distribution, and the PDF of each Z_i is $f_{Z_i}(z_i)$. Then the random vector

$\boldsymbol{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$ has the standard multivariate normal (SMVN) distribution.

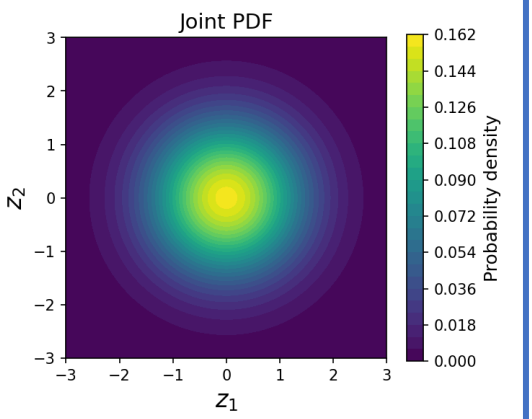


Joint PDF:

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = f_{Z_1,Z_2,\dots,Z_n}(z_1, z_2, \dots, z_n) = \frac{1}{(2\pi)^{\frac{n}{2}}} exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}\right) \qquad -\infty < z_i < \infty$$

Mean: $\boldsymbol{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

Covariance: $\boldsymbol{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$



Properties: SMVN distribution is a generalization of the standard normal distribution to a random vector.
The PDF of SMVN distribution can be also written as:

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = f_{Z_1}(z_1) f_{Z_2}(z_2) \dots f_{Z_n}(z_n)$$

Related distributions: SMVN is a special case of MVN distribution when $\boldsymbol{\mu}=\boldsymbol{0}$ and $\boldsymbol{\Sigma}=\boldsymbol{I}$.
If $\boldsymbol{Z}$ only has one element (n=1), then SMVN is equivalent to the standard normal distribution.

Multivariate normal distribution (*Multivariate-Continuous*)

Parameters: $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ ($\boldsymbol{\Sigma}$ is positive semidefinite) **Denoted by**: $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Story: Suppose the random vector

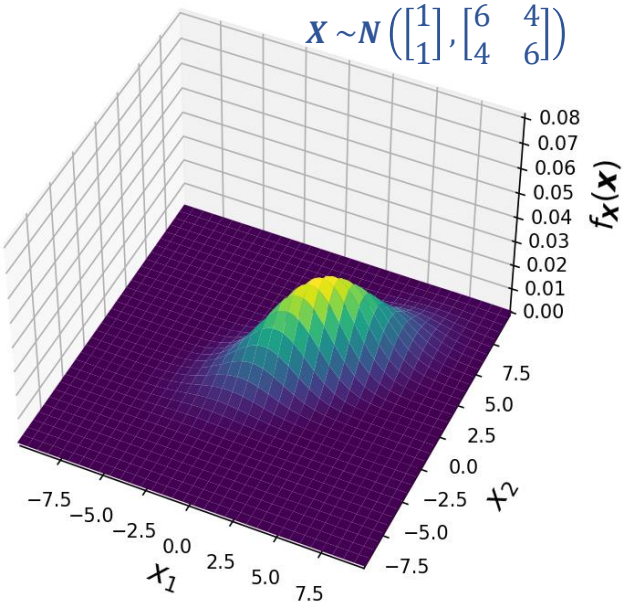
$\boldsymbol{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}$ has an SMVN distribution, and let $\boldsymbol{X}$ be a random vector with n elements defined as

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}$$

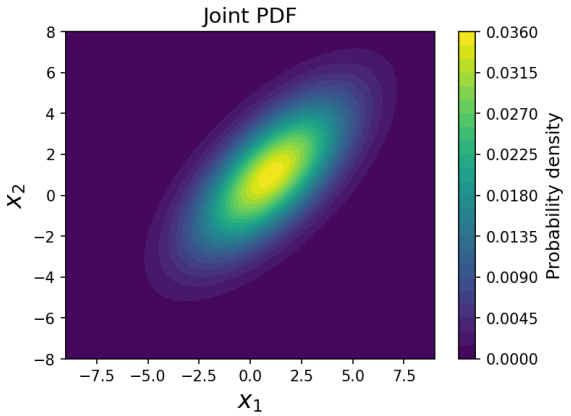
where $\boldsymbol{\mu}$ is the mean vector:

$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$ and $\boldsymbol{A}$ is a symmetric $n \times n$ matrix

($\boldsymbol{A}=\boldsymbol{A}^T$). In addition, we define the $n \times n$ matrix $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{A}^T\boldsymbol{A}$ and call it the covariance matrix. Then $\boldsymbol{X}$ is said to have a multivariate normal (MVN) distribution with the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.



Multivariate normal distribution (Cont'd)



MVN distribution is a generalization of the normal distribution to a random vector. The standard deviation (σ) transforms a random variable Z with the standard normal distribution to the random variable X with a normal distribution ($X = \mu + \sigma Z$).

Similarly, the matrix $\boldsymbol{A}$ transform the random vector $\boldsymbol{Z}$ with an SMVN distribution to the random vector $\boldsymbol{X}$ with an MVN distribution ($\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{AZ}$). Hence, it plays the same role of standard deviation for random vectors with the MVN distribution.

Joint PDF: If the random vector $\boldsymbol{X}$ (with n elements) has an MVN distribution with the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is a positive definite matrix then its joint PDF is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1,X_2,\dots,X_n}(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \quad -\infty < x_i < \infty$$

Mean: $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$

Covariance: $\boldsymbol{\Sigma} = \boldsymbol{AA}^T = \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}^2 = \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \dots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & Var(X_2) & \dots & Cov(X_2,X_n) \\ \vdots & \vdots & \dots & \vdots \\ Cov(X_n,X_1) & Cov(X_n,X_2) & \dots & Var(X_n) \end{bmatrix}$

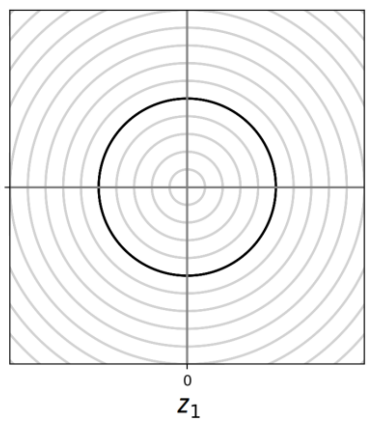
Precision matrix: The inverse of the covariance matrix is called the precision matric and is denoted by $\boldsymbol{\Lambda}$. So, we can also denote the MVN distribution by $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ where $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1}$.

Bivariate normal distribution: If $n=2$ (n is the number of the elements of $\boldsymbol{X}$), the MVN distribution is called a bivariate normal distribution.

Contours of joint PDF: The shape of the contours of an MVN distribution is determined by the covariance matrix.

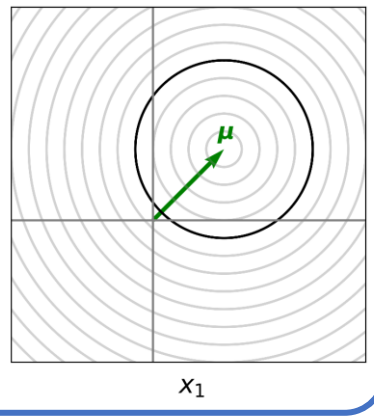
In the SMVN distribution, the mean vector is zero and the covariance matrix is the identity matrix, so for a 2-dimensional SMVN distribution, the PDF contours are circles centered at the origin. For an n-dimensional SMVN distribution, they are n-dimensional hyperspheres centered at the origin.

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



In a bivariate normal distribution, if the covariance matrix is a multiple of the identity matrix ($\boldsymbol{\Sigma} = c\boldsymbol{I}$), the joint PDF contours are circles centered at $\boldsymbol{\mu}$ (the mean vector). For an n-dimensional MVN distribution, if $\boldsymbol{\Sigma} = c\boldsymbol{I}$, the contours of the joint PDF are n-dimensional hyperspheres centered at $\boldsymbol{\mu}$.

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix}$$



Multivariate normal distribution (Cont'd)

In a bivariate normal distribution, the joint PDF contours are ellipses centered at $\boldsymbol{\mu}$. More generally, in an n-dimensional MVN distribution, the contours are n-dimensional hyper-ellipsoids centered at $\boldsymbol{\mu}$. The principal axes of these ellipsoids are along the eigenvectors of the covariance matrix ($\boldsymbol{v}_i$ i=1…n). The eigenvectors and eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ are given by the following equations:

$$\boldsymbol{\Sigma v}_i = \lambda_i^2 \boldsymbol{v}_i$$

$$\boldsymbol{A v}_i = \lambda_i \boldsymbol{v}_i$$

The semidiameter of each hyper-ellipsoid along each principal axis represented by $\boldsymbol{v}_i$ is equal to λ_i.

MVN with uncorrelated random variables: When the random variables X_1, X_2, …X_n are uncorrelated ($Cov(X_i, X_j)$=0), their covariance matrix becomes diagonal:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Each diagonal element (σ_i^2) is an eigenvalue of $\boldsymbol{\Sigma}$ and its corresponding eigenvector is $\boldsymbol{e}_i$ (the ith vector of the standard basis). So, each eigenvector is along one of the coordinate axes, and the principal axes of the hyper-ellipsoid are also the coordinate axes.

When $\boldsymbol{\Sigma}$ is diagonal, the marginal distribution of each X_i has a normal distribution with a mean of μ_i and variance of σ_i^2:

$$X_i \sim N(\mu_i, \sigma_i^2)$$

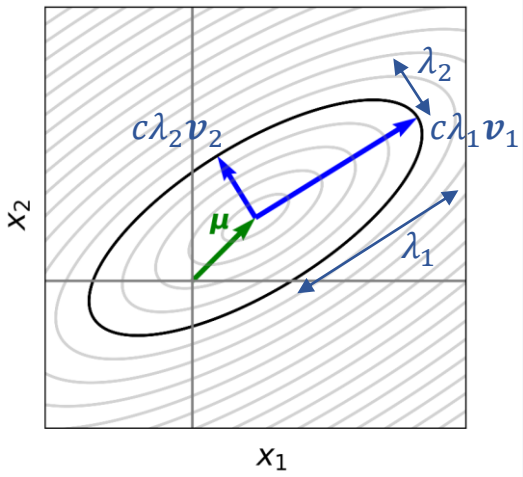
The PDF of such an MVN distribution is the product of the PDF of these marginal distributions:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1}(x_1) f_{X_2}(x_2) \dots f_{X_n}(x_n)$$

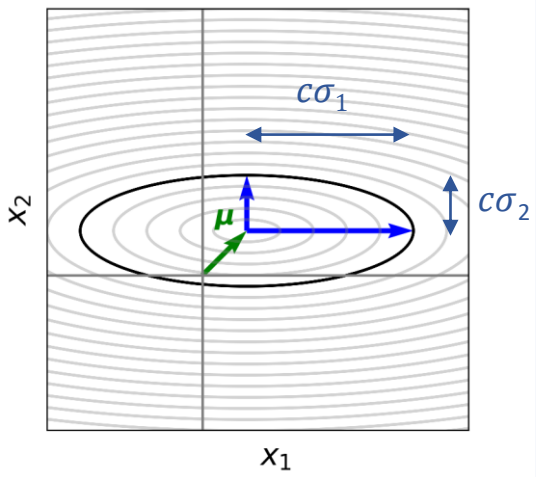
So, X_1,X_2,…X_n are also independent.

By changing the coordinate system for an n-d MVN distribution, we can convert the covariance matrix into a diagonal matrix. We can define a new coordinate system by rotating the axes of the original coordinate system to be along the eigenvectors of the covariance matrix and move the origin by $\boldsymbol{\mu}$.

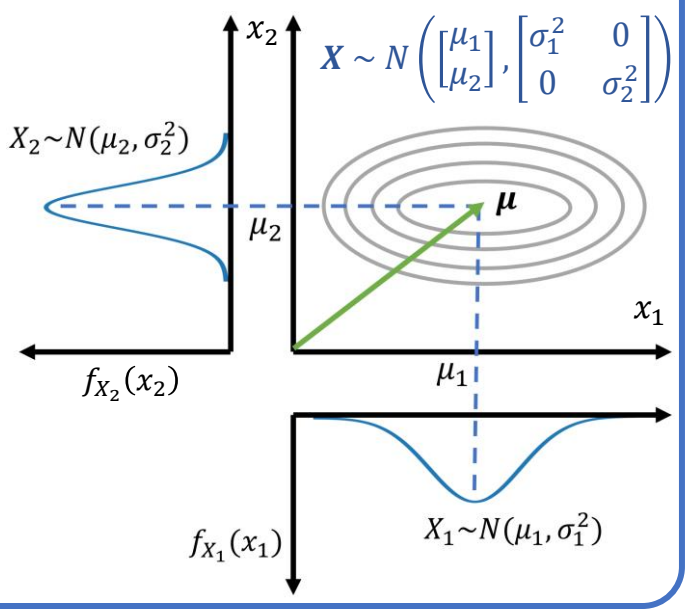
The eigenvectors of both $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ are $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. The eigenvalues of $\boldsymbol{A}$ are λ_1 and λ_2. The eigenvalues of $\boldsymbol{\Sigma}$ are λ_1^2 and λ_2^2



$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$



X_1 and X_2 are uncorrelated and independent



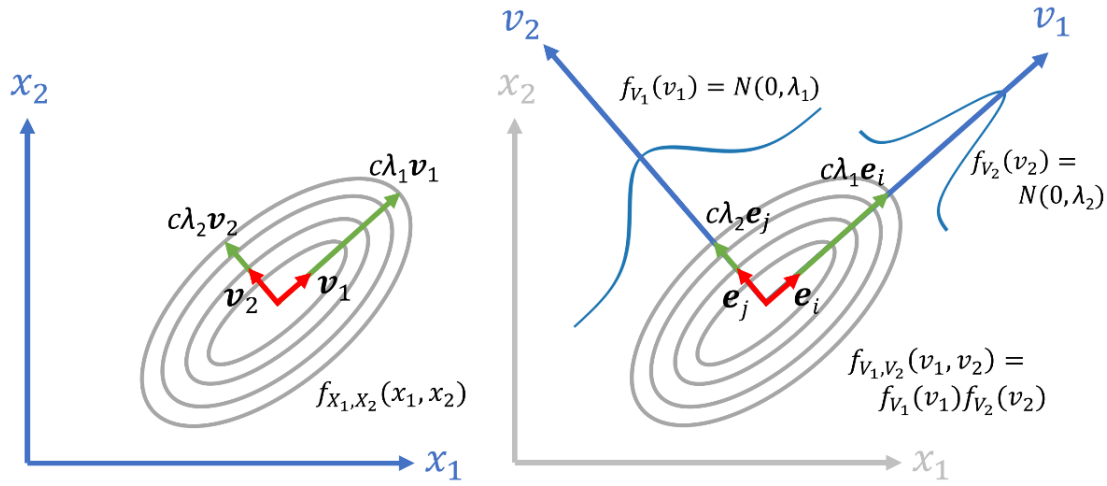
Multivariate normal distribution (Cont'd)

In this new coordinate system, the ellipsoids which are the contours of the joint PDF are centered at the origin and their principal axes are along the coordinate axes. The original correlated random variables $X_1,\dots X_n$ turn into the uncorrelated and independent random variables $V_1,\dots V_n$ where each of them has the following normal distribution:

$$V_i \sim N(0, \lambda_i)$$

The MVN distribution in this new coordinate has the following parameters:

$$\boldsymbol{V} = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix} \right)$$



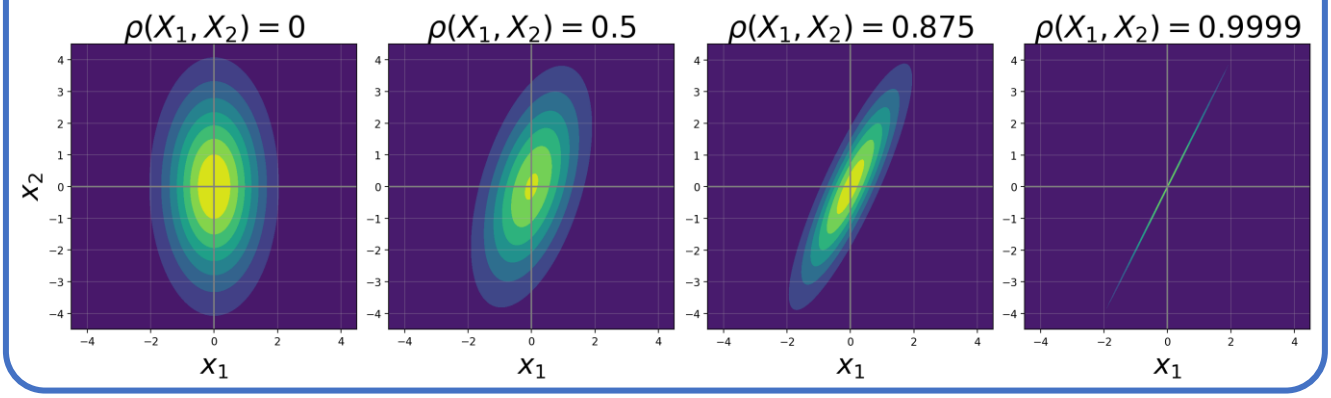
The effect of correlation coefficients on the joint PDF contours: In the MVN distribution, the covariance between each pair of random variables $Cov(X_i, X_j)$ is an indicator of the dependence between them. By changing $Cov(X_i, X_j)$, the covariance matrix and the contours of the joint PDF change. When $Cov(X_i, X_j)=0$, the random variables X_i and X_j are independent, and as it increases the dependence between them becomes stronger.

We can also write the covariance in terms of the correlation coefficient and standard deviations of X_i and X_j: $Cov(X_i, X_j) = \rho(X_i, X_j)\sigma_{X_i}\sigma_{X_j}$

So, in a bivariate normal distribution, the covariance matrix can be also written as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{X_1}^2 & \rho(X_1, X_2)\sigma_{X_1}\sigma_{X_2} \\ \rho(X_2, X_1)\sigma_{X_2}\sigma_{X_1} & \sigma_{X_2}^2 \end{bmatrix}$$

When the correlation coefficient is zero, X_1 and X_2 are independent, the principal axes of the ellipses are along the coordinate axis. As the correlation between X_1 and X_2 increases, the joint PDF in the plane of X_1 and X_2 tilts and becomes narrower which means that X_1 and X_2 are more dependent on each other.



Multivariate normal distribution (Cont'd)

When $|\rho(X_i, X_j)| = 1$, the covariance of X_i and X_j is maximized, and there is a linear relationship between them. Hence, knowing the value of one can determine the exact value of the other. In that case, one of the random variables is a deterministic function of the other, and we have a degenerate MVN distribution.

Degenerate MVN distribution: When the covariance matrix is singular, the MVN distribution is said to be degenerate. Such an MVN distribution has no joint PDF since the covariance matrix is not invertible. When the random variables X_1, X_2 …, X_n have a degenerate MVN distribution, then their values x_1, x_2 …, x_n lie in a space that has a dimension less than *n*.

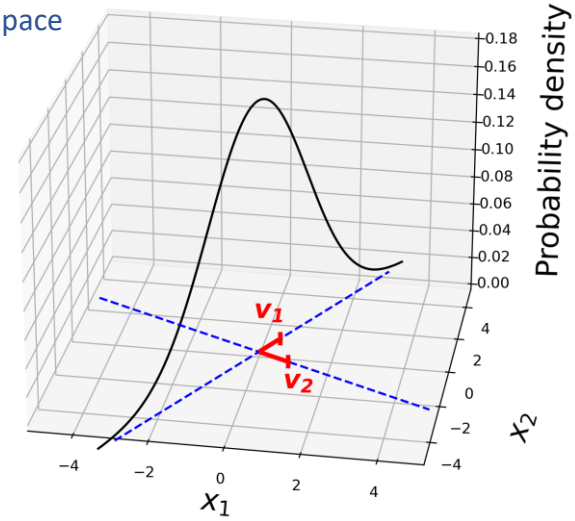
Degeneracy occurs when at least one of the random variables is a deterministic function of the others.

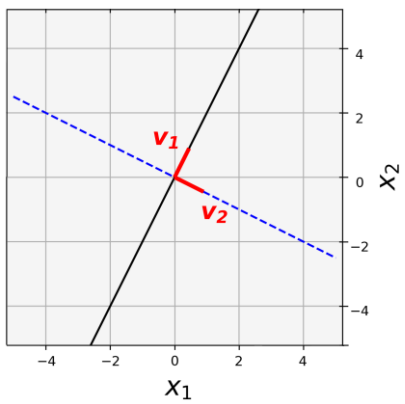
In a degenerate MVN distribution with *n* random variables, $\boldsymbol{\Sigma}$ (and $\boldsymbol{A}$) is not a full-rank matrix, so *rank* $\boldsymbol{\Sigma}$=*m* where *m*<*n* (also *rank* $\boldsymbol{A}$=*m*), and *m* random variables are a deterministic function of the others.

This also means that *n*-*m* eigenvalues of $\boldsymbol{\Sigma}$ (and $\boldsymbol{A}$) are zero. Hence, $\boldsymbol{\Sigma}$ (and $\boldsymbol{A}$) is not positive definite anymore (but it is still positive semidefinite). Finally, $\boldsymbol{\Sigma}$ (and $\boldsymbol{A}$) is a singular matrix that is not invertible, and its determinant is zero.

$$\boldsymbol{X} \sim \boldsymbol{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 2\\2 & 4\end{bmatrix}\right) \qquad \lambda_1 = 5,\ \lambda_2 = 0$$

In this degenerate MVN distribution, x_1 and x_2 lie in a 1-d space. You think of that as a univariate normal distribution which lies in a 2-dspace





Properties: The univariate normal distribution is a special case of an MVN distribution when the random vector $\boldsymbol{X}$ has only one element. If $\boldsymbol{X}$=[X], $\boldsymbol{\mu}$=[μ], , $\boldsymbol{\Sigma}$ =[σ^2] (all of them can be thought of as a matrix with only one element), then $X \sim N(\mu, \sigma^2)$.

Concatenating normal random variables: If the random variables $X_1, X_2, \ldots X_n$ are mutually independent and each X_i has a normal distribution with mean μ_i and variance σ_i^2 then concatenating them results in a in a random vector with an MVN distribution.

$$X_i \sim N(\mu_i, \sigma_i^2) \longrightarrow \begin{bmatrix}X_1\\X_2\\\vdots\\X_n\end{bmatrix} \sim N\left(\begin{bmatrix}\mu_1\\\mu_2\\\vdots\\\mu_n\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & 0 & \ldots & 0\\0 & \sigma_2^2 & \ldots & 0\\\vdots & \vdots & \ldots & \vdots\\0 & 0 & \ldots & \sigma_n^2\end{bmatrix}\right)$$

Linear transformation of an MVN random vector: Let $\boldsymbol{X}$ be a random vector with *n* elements that has an MVN distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Let $\boldsymbol{b}$ be a vector with *m* elements and $\boldsymbol{C}$ be an *m*×*n* matrix. Then the random vector $\boldsymbol{Y}$ (with *m* elements) defined as $\boldsymbol{Y}=\boldsymbol{b}+\boldsymbol{CX}$ has an MVN distribution with mean $\boldsymbol{b}+\boldsymbol{C\mu}$ and covariance matrix $\boldsymbol{C\Sigma C}^T$.

$$\boldsymbol{Y} = \boldsymbol{b} + \boldsymbol{CX} \qquad \boldsymbol{Y} \sim N(\boldsymbol{b} + \boldsymbol{C\mu}, \boldsymbol{C\Sigma C}^T)$$

Multivariate normal distribution (Cont'd)

Linear combinations of independent MVN random vectors: Suppose that $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots\boldsymbol{X}_m$ are m independent random vectors. Each $\boldsymbol{X}_i$ has n_i elements and an MVN distribution with parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$. Let $\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots\boldsymbol{C}_m$ be $p\times n_i$ (i=1...m) matrices. Then the random vector $\boldsymbol{Y}$ (with p elements) defined as

$$\boldsymbol{Y} = \boldsymbol{b} + \sum_{i=1}^{m} \boldsymbol{C}_i \boldsymbol{X}_i$$

has an MVN distribution with the following parameters

$$\boldsymbol{\mu} = \boldsymbol{b} + \sum_{i=1}^{m} \boldsymbol{C}_i \boldsymbol{\mu}_i \qquad \boldsymbol{\Sigma} = \sum_{i=1}^{m} \boldsymbol{C}_i \boldsymbol{\Sigma}_i \boldsymbol{C}_i^T$$

Concatenating MVN random vectors: Suppose that $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots\boldsymbol{X}_m$ are m independent random vectors, and

$$\boldsymbol{X}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \boldsymbol{X}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \ldots, \boldsymbol{X}_m \sim N(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

Then concatenating them results in a in a random vector with an MVN distribution.

$$\begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_m \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_1 & 0 & \ldots & 0 \\ 0 & \boldsymbol{\Sigma}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \boldsymbol{\Sigma}_m \end{bmatrix} \right)$$

Marginal distributions: Let $\boldsymbol{X}$ be a random vector with an MVN distribution:

$$\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and $\boldsymbol{X}_s$ be a subset vector of $\boldsymbol{X}$. Then $\boldsymbol{X}_s$ also has the following MVN distribution:

$$\boldsymbol{X}_S \sim N(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S)$$

which is called the marginal distribution of $\boldsymbol{X}_s$. Here the vector $\boldsymbol{\mu}_s$ is a subset of $\boldsymbol{\mu}$ that only contains the corresponding means of the random variables in $\boldsymbol{X}_s$, and the matrix $\boldsymbol{\Sigma}_s$ is the covariance matrix of the random variables in $\boldsymbol{X}_s$. For example, suppose that

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} & \Sigma_{14} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} & \Sigma_{34} \\ \Sigma_{41} & \Sigma_{42} & \Sigma_{43} & \Sigma_{44} \end{bmatrix} \right)$$

Then the marginal distribution of X_1, X_3 is:

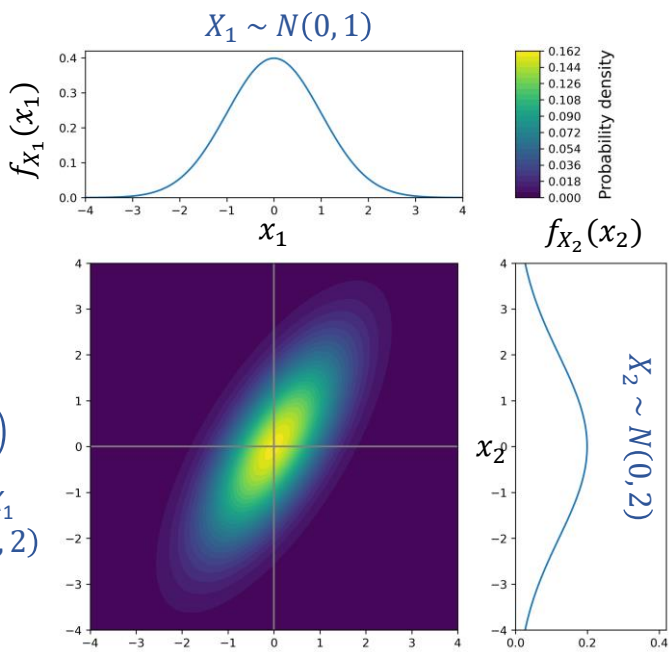
$$\begin{bmatrix} X_1 \\ X_3 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{13} \\ \Sigma_{31} & \Sigma_{33} \end{bmatrix} \right)$$

And the marginal distribution of X_1 is:

$$X_1 \sim N(\mu_1, \Sigma_{11})$$

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \right)$$

Marginal distribution of X_1 and X_1 are: $X_1 \sim N(0,1)$ and $X_2 \sim N(0,2)$



Multivariate normal distribution (Cont'd)

Partitioning an MVN random vector: Let $\boldsymbol{X}$ be a random vector with an MVN distribution:

$$\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

We can partition $\boldsymbol{X}$ into m MVN random sub-vectors $\boldsymbol{X}_1,\ldots\boldsymbol{X}_m$, and partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ accordingly:

$$\begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_m \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \ldots & \boldsymbol{\Sigma}_{1m} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \ldots & \boldsymbol{\Sigma}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{m1} & \boldsymbol{\Sigma}_{2m} & \ldots & \boldsymbol{\Sigma}_{mm} \end{bmatrix} \right)$$

Each $\boldsymbol{\mu}_i$ contains the corresponding means of the random variables in $\boldsymbol{X}_i$, and $\boldsymbol{\Sigma}_{ij}$ is the covariance matrix of the random variables in $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. Now each sub-vector $\boldsymbol{X}_i$ has an MVN distribution:

$$\boldsymbol{X}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii})$$

In addition, if in the partitioned covariance matrix, we have $\boldsymbol{\Sigma}_{ij}=\boldsymbol{\Sigma}_{ji}=\boldsymbol{0}$ then it follows that the random vectors $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are independent. For example, if we have:

$$\begin{matrix} & & & & \boldsymbol{\Sigma}_{11} & & \boldsymbol{\Sigma}_{12} & \\ \boldsymbol{X}_1 & \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} & & \begin{bmatrix} 0 \\ 0 \end{bmatrix} \boldsymbol{\mu}_1 & \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} & & \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} & \\ & & \sim N\left(\right. & & & , & & \\ \boldsymbol{X}_1 & \begin{bmatrix} X_3 \\ X_4 \\ X_4 \end{bmatrix} & & \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \boldsymbol{\mu}_2 & \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} & & \begin{bmatrix} 1 & 0 & 2 \\ 0 & 7 & 0 \\ 2 & 0 & 5 \end{bmatrix} & \left. \right) \\ & & & & \boldsymbol{\Sigma}_{21} & & \boldsymbol{\Sigma}_{22} & \end{matrix}$$

Then we conclude that:

$$\boldsymbol{X}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}),\ \boldsymbol{X}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

And since all the elements of $\boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_{21}$ are zero, we conclude that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent.

Related distributions: SMVN is a special case of MVN distribution when $\boldsymbol{\mu}=\boldsymbol{0}$ and $\boldsymbol{\Sigma}=\boldsymbol{I}$. If $\boldsymbol{X}$ has a MVN distribution, the marginal distribution of each random variable in $\boldsymbol{X}$ is a normal distribution. If $\boldsymbol{X}$ only has one element (n=1), then MVN is equivalent to the normal distribution.

This cheat sheet was prepared by Reza Bagheri (https://www.linkedin.com/in/reza-bagheri-71882a76/). It is a summary of the following Medium articles:

1. Understanding Probability Distributions using Python (https://medium.com/towards-data-science/understanding-probability-distributions-using-python-9eca9c1d9d38)
2. Understanding Multinomial Distribution using Python (https://medium.com/towards-data-science/understanding-multinomial-distribution-using-python-f48c89e1e29f)
3. Understanding Multivariate Normal Distribution (https://medium.com/@reza-bagheri79/understanding-multivariate-normal-distribution-54089b5b106c)
4. Dirichlet Distribution: The Underlying Intuition and Python Implementation (https://medium.com/towards-data-science/dirichlet-distribution-the-underlying-intuition-and-python-implementation-59af3c5d3ca2)