12

STATISTICAL DISTRIBUTIONS

12.1 Common distributions and properties

12.1.1 Overview

In this chapter we survey statistical distributions commonly used in real-world applications and in subsequent chapters of this book. For each distribution, we study their basic properties like its mean and variance, and many other useful properties allowing us to construct more complex distributions. Along the lines, we will also discuss the connections between different statistical distributions.

We do not pretend to give a comprehensive overview on all aspects of statistical distributions. For an extensive discussion on statistical distributions, see [1][2].

12.1.2 Bernoulli distribution

Definition 12.1.1 (Bernoulli distribution). *A random variable Y with sample space $\{0, 1\}$ is said to have Bernoulli distribution $Ber(\theta)$ with parameter θ if it has a pmf given as*

$$p(y) = \theta^y (1-\theta)^{1-y}, y \in \{0, 1\}.$$

Example 12.1.1. Consider the experiment of toss a biased coin. The probability of getting head is p and getting tail is $1 - p$. The outcome of coin toss can be modeled by a Bernoulli random variable.

Lemma 12.1.1 (basic properties). *Let X be a random variable with distribution $Ber(p)$. Then*

- $M_X(t) = (1 - p + pe^t)$.
- $E[X] = p, E[X^2] = p, Var[X^2] = p - p^2 = p(1-p)$.

Proof. Straight forward. □

12.1.3 Poisson distribution

Definition 12.1.2 (Poisson distribution). *A discrete random variable X is said to have a Poisson distribution $Poisson(\lambda)$ with parameter λ if it has pmf given as*

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

with $x \in \{0, 1, 2, ...\}$.

Lemma 12.1.2 (basic property of Poisson distribution). *[3, p. 154] Let X be a random variable with distribution $Poisson(\lambda)$. Then*

- $M(t) = \exp(\lambda(e^t - 1))$.
- $E[X] = \lambda, Var[X] = \lambda$.

Proof. (1)

$$\begin{aligned} M_X(t) &= E[e^{tX}] \\ &= \sum_{n=0}^{\infty} e^{tn} \frac{\lambda^n}{n!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} \\ &= e^{-\lambda} e^{\lambda e^t}. \end{aligned}$$

(2) $E[X] = M'_X(0) = \lambda, E[X^2] = M''_X(0) = \lambda^2 - \lambda.$ □

Lemma 12.1.3 (sum of Poisson distribution). *Assume $X_1, ..., X_n$ to be independent random variables, and $X_i \sim Poisson(\theta_i), i = 1, ..., n$. Then*

$$Y = \sum_{i=1}^{n} X_i \sim Poisson(\sum_{i=1}^{n} \theta_i).$$

Proof. Note that

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \exp \sum_{i=1}^{n} \theta_i (e^t - 1).$$

□

Lemma 12.1.4 (Normal approximate sum of Poisson). *Let $X_1, ..., X_n$ be independent iid random variable of Poisson(θ), then*

$$Y = \sum_{i=1}^{n} X_i$$

can be approximated by

$$\frac{Y - n\theta}{\sqrt{n\theta}} \sim N(0,1),$$

or equivalently

$$Y \sim N(n\theta, n\theta).$$

Proof. Directly from Central Limit Theorem [Theorem 11.11.3]. □

12.1.4 Geometric distribution

Definition 12.1.3 (geometric distribution). *A discrete random variable X is said to have a geometric distribution Geo(θ) with parameter θ if it has pmf given as*

$$p(X = k) = (1-\theta)^{k-1}\theta$$

with $k \in \{1, 2, ...\}$.

Example 12.1.2 (number of trials needed to succeed in Bernoulli trias). The geometric distribution The probability distribution of the number X of Bernoulli trials needed to get one success, supported on the set $\{1, 2, 3, ...\}$

Lemma 12.1.5 (basic statistics of geometric distribution). *The expected value of a geometrically distributed random variable X with parameter p is $1/p$ and the variance is $(1-p)/p^2$.*

Proof. (1)

$$E[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p$$

$$(1-p)E[X] = \sum_{k=1}^{\infty} k(1-p)^{k}p$$

subtract and get $pE[X] = 1$.
(2)

$$Var[X] = \sum_{k=1}^{\infty}(k - 1/p)^2(1-p)^{k-1}p$$

can be proved similarly. □

Example 12.1.3 (coupon collection problem, subsection 11.12.4). Consider the **coupon collection problem** where there is an urn of m different coupons. How many coupons do you expect you need to draw **with replacement** before having drawn each coupon at least once?

Let Z_i denote the number of additional samples needed to go from $i-1$ distinct coupons to i distinct coupons. Let W_k denote the number of samples needed to get k distinct coupons. Then $Z_j, j = 1, ..., m$ is a sequence of independent random variables has the geometric distribution with parameter $p_i = \frac{m-i+1}{m}$.

When $i = 1$, Z_1 has a geometric distribution with parameter $p_1 = 1$. Similarly, Z_2 has a geometric distribution with parameter $p_2 = (m-1)/m$; Z_3 has a geometric distribution with parameter $p_3 = (m-2)/m$. Then, we can generalize to Z_i has a geometric distribution with parameter $p_i = (m-(i-1))/m$.

Further, we have

- $W_k = \sum_{i=1}^{k} Z_i$.
- $E[W_k] = \sum_{i=1}^{k} \frac{m}{m-i+1}$.

Example 12.1.4 (number of visits in a Markov chain, Lemma 20.2.1). Consider a state i in a Markov chain. Let f_{ii} denote the probability that a trajectory starting from state i will *ever* revisit i.

Then the probability of of n visit is $f_{ii}^{n-1}(1-f_{ii})$, which is the product of the probability visiting state i $n-1$ times and then never visit again.

The expected total visit is

$$\sum_{n=1}^{\infty} nf_{ii}^{n-1}(1-f_{ii}) = \frac{1}{1-f_{ii}}.$$

12.1.5 Binomial distribution

Definition 12.1.4 (binomial distribution). *A discrete random variable X is said to have a Binomial distribution $Binomial(n, p)$ with parameter n, p if it has pmf given as*

$$f(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

with $x \in \{0, 1, 2, ..., n\}$.

Remark 12.1.1 (interpretation). Binomial distribution represents the probability distribution of the number of successes in a sequence of n independent binary experiments, each of which yields 1 with probability p.

Remark 12.1.2 (relation to Bernoulli distribution). Let X_i be iid random variables with Bernoulli distribution of parameter p, then

$$Y = \sum_{i=1}^{n} X_i$$

is a random variable of binomial distribution with parameter (n, p).

Lemma 12.1.6 (sum of independent binomial random variable). *Let $X_1, X_2, ..., X_K$ be the independent binomial random variables with parameter $(n_1, p), (n_2, p), ..., (n_K, p)$. Let $Y = \sum_{i=1}^{K} X_i$. Then*

- $M_{X_i}(t) = (1 - p + pe^t)^{n_i}, i = 1, ..., K.$
- $M_Y(t) = (1 - p + pe^t)^{\sum_{i=1}^{K} n_i}$
- $Y \sim Binomial(\sum_{i=1}^{K} n_i, p).$

Proof. (1) Use the mgf of Bernoulli distribution [Lemma 12.1.1]. (2)(3) Consider $X_1 \sim Binomial(n_1, p)$ and $X_2 \sim Binomial(n_2, p)$, each has momemt generating function of $(1 - p + pe^t)^{n_1}$ and $(1 - p + pe^t)^{n_2}$. $X_1 + X_2$ will have mgf of $(1 - p + pe^t)^{n_1 + n_2}$ [Theorem 11.6.2], corresponding to $Binomial(n_1 + n_2, p)$. It is straight forward to extend multiple cases. □

Lemma 12.1.7 (convergence of binomial distribution to Poisson distribution). *Suppose that $p_n \in (0, 1)$ for $n \in \mathcal{N}_+$ and $np_n \to \lambda$ as $n \to \infty$. Then the binomial distribution* *with parameters n and p_n converges to the Poisson distribution with parameter λ* ***in distribution*** *as $n \to \infty$. That is, for fixed $k \in \mathcal{N}$,*

$$\binom{n}{k} p_n^k (1-p_n)^{n-k} \to e^{-\lambda} \frac{\lambda^k}{k!}$$

as $n \to \infty$.

Proof. (direct method)Note that

$$\begin{aligned}
\binom{n}{k} p_n^k (1-p_n)^{n-k} &= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}(p_n)^k(1-p_n)^{n-k} \\
&= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k} \\
&\approx \frac{n^k}{k!}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k} \\
&= \frac{\lambda^k}{k!}(1-\frac{\lambda}{n})^{n-k} \\
&\to e^{-\lambda\frac{n-k}{n}}\frac{\lambda^k}{k!} \\
&\approx e^{-\lambda}\frac{\lambda^k}{k!}
\end{aligned}$$

(use generating function) Note that binomial distribution has probability generating function [**??**]

$$((1-p_n)+p_n s)^n = (1+(pns-p_n)n/n)^n \to e^{(s-1)a}, n \to \infty$$

where $e^{(s-1)a}$ is the generating function of Poisson distribution. □

Remark 12.1.3 (Poisson distribution as an approximate for large n and small k)**.** Note that the lemma requires that k fixed. In other words, when $n \gg k$, we can use Poisson distribution to approximate binomial distribution.

12.1.6 Normal distribution

Definition 12.1.5 (normal distribution). *A random variable X with normal distribution $N(\mu, \sigma^2)$, characterized by parameters μ and σ, has its pdf given by*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}(x-\mu)^2/\sigma^2), -\infty < x < \infty.$$

X is called normal random variable, or Gaussian random variable. If $\mu = 0, \sigma = 1$, X is also called standard normal random variable.

Lemma 12.1.8 (moment generating function). *Let X be a random variable with normal distribution $N(0,1)$, then the moment generating function is*

$$m_X(t) = \exp(\frac{1}{2}t^2).$$

If Y is a random variable with normal distribution $N(\mu,\sigma^2)$, then the moment generating function is

$$m_Y(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2).$$

Proof. (1) $m_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ complete the square and get the result. (2)Let $Y = \sigma X + \mu$ and use Theorem 11.6.2. Then $m_Y(t) = e^{\mu t} m_X(\sigma t)$ □

Lemma 12.1.9 (basic properties of normal random variable). *Consider $X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2)$.*

- *If X, Y **are independent**, then we have*
  -
    $$aX + b \sim N(a\mu + b, a^2\sigma_x^2)$$
  -
    $$X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$
    $$aX + bY \sim N(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$$
- *If X **and** Y **are not independent but jointly normal**, then $X + Z$ will be normal, and*
  $$aX + bZ \sim N(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_z^2 + 2abCov(X,Z)).$$
- *Assume X, Y are independent. Further let $W = \rho X + \sqrt{1-\rho^2}Y, \rho \in [-1,1]$. Then W is normal, correlated with X and Y, and the sum $X + W$ is also normal; that is,*
  $$W \sim N, aX + bW \sim N.$$
- ***In general, the sum of two dependent normal random variable is not necessarily normal.** See Corollary 12.1.1.1.*

Proof. (1) Directly from the properties of moment generating functions at Theorem 11.6.2. (2)The proof of two general jointly normal random variable will be showed in Corollary 12.1.1.1. (3) Note that

$$(X,W)^T = (X, \rho X + \sqrt{1-\rho^2}Y)^T = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} (X,Y)^T,$$

therefore (X,W) are jointly normal [Theorem 14.1.1]. Then we use (2). □

Lemma 12.1.10 (moments of standard normal distribution). *Let $X \sim N(0,1)$, then*

$$E[X] = 0, E[X^2] = 1, E[X^3] = 0, E[X^4] = 3$$

Moreover, all odd moments are 0.

Proof. The mgf is $m(t) = e^{t^2/2}$, then

$$m'(t) = te^{t^2/2}$$
$$m''(t) = e^{t^2/2} + t^2e^{t^2/2}$$
$$\cdots$$

For all odd moments

$$\int x^{2k+1} f(x)dx$$

has integrand as odd function. □

Corollary 12.1.0.1 (moments of normal distribution). *Let $X \sim N(0,\sigma^2)$, then*

$$E[X] = 0, E[X^2] = \sigma^2, E[X^3] = 0, E[X^4] = 3\sigma^4$$

Moreover, all odd moments are 0.

Proof. The mgf is $m(t) = e^{\sigma^2t^2/2}$, then

$$m'(t) = \sigma^2te^{\sigma^2t^2/2}$$
$$m''(t) = \sigma^2e^{\sigma^2t^2/2} + \sigma^4t^2e^{t^2/2}$$
$$\cdots$$

For all odd moments

$$\int x^{2k+1} f(x)dx$$

has integrand as odd function. □

12.1.7 Half-normal distribution

Definition 12.1.6 (half-normal distribution). *Let X follow an ordinary normal distribution $N(0,\sigma^2)$. Then $Y = |X|$ follows a* ***half-normal distribution*** *with parameter σ. It has probability density function*

$$f_Y(y;\sigma) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}}\exp(-\frac{y^2}{2\sigma^2}).$$

Lemma 12.1.11 (basic properties of half normal distribution). *Let Y follow a half-normal distribution with parameter σ. Then*

- $E[Y] = \frac{\sigma\sqrt{2}}{\sqrt{\pi}}$.
- $Var[Y] = \sigma^2(1-\frac{2}{\pi})$

12.1.8 Laplace distribution

Definition 12.1.7 (Laplace distribution). *A random variable X has a* ***Laplace distribution****, denoted by $Lap(\mu,b)$, if its probability density function is*

$$f(x|\mu,b) = \frac{1}{2b}\exp(-\frac{|x-\mu|}{b}) = \begin{cases} \frac{1}{2b}\exp(-\frac{\mu-x}{b}), if\ x<\mu \\ \frac{1}{2b}\exp(-\frac{x-\mu}{b}), if\ x\geq\mu \end{cases}.$$

Lemma 12.1.12 (properties of Laplace distribution). *Let X be a random variable with Laplace distribution with parameter μ, b. It follows that*

- *The mean and the median are μ.*
- *The variance is $2b^2$.*
- *The cdf is given by*

$$\begin{aligned} F(x) = \int_{-\infty}^{x} f(u)du &= \begin{cases} \frac{1}{2}\exp(-\frac{\mu-x}{b}), if\ x<\mu \\ 1-\frac{1}{2}\exp(-\frac{x-\mu}{b}), if\ x\geq\mu \end{cases} \\ &= \frac{1}{2}+\frac{1}{2}sgn(x-\mu)(1-\exp(-\frac{|x-\mu|}{b})). \end{aligned}$$

- *The inverse cdf is given by*

$$F^{-1}(p) = \mu - b \cdot sgn(p - 0.5)\ln(1 - 2|p - 0.5|).$$

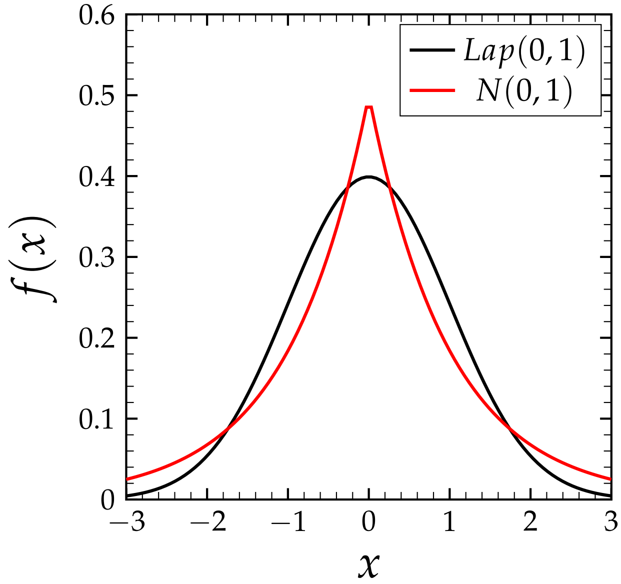

Figure 12.1.1: Comparison of Laplace distribution and normal distribution.

12.1.9 Multivariate Gaussian/normal distribution

12.1.9.1 *Basic definitions*

Definition 12.1.8 (multivariate Gaussian/normal distribution). *A random vector is said be multivariate Gaussian/normal random variable if it's pdf is multivariate Gaussian/normal distribution, whose support is* $\mathbb{R}^n$ *and its pdf is*

$$\rho(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|det\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

with mean $\mu \in \mathbb{R}^n$ *and covariance matrix* $\Sigma \in \mathbb{R}^{n\times n}$.

Lemma 12.1.13 (mgf of multivariate Gaussian/normal random variables). *[3, p. 181] An n-dimensional random vector $X \sim MN(\mu, \Sigma)$ has its mgf given by*

$$M_X(t) \triangleq E[\exp(t^T X)] = \exp(t^T \mu + \frac{1}{2} t^T \Sigma t)$$

for all $t \in \mathbb{R}^n$.

Proof.

$$\begin{aligned} M_X(t) &= E[\exp(t^T X)] \\ &= \exp(E[t^T X] + \frac{1}{2} Var[t^T X]) \\ &= \exp(t^T \mu + \frac{1}{2} t^T \Sigma t) \end{aligned}$$

where we use the fact that $t^T X \sim N(t^T \mu, t^T \Sigma t)$ from Theorem 14.1.1. □

Remark 12.1.4 (implication). Given a random vector X, if we want to check whether X is a multivariate Gaussian, we can check its mgf. If its mgf is the exponential of a linear form plus a quadratic form, then it is multivariate Gaussian.

Lemma 12.1.14 (criterion via linear combination). *A vector $X = (X_1, X_2, ..., X_n)^T$ is a multivariate Gaussian distribution if every linear combination*

$$S = a^T X, a \in \mathbb{R}^n$$

has a normal distribution.

Proof. Because $a^T X$ is normal, then it has characteristic function

$$E[\exp(ita^T X)] = \exp(ita^T \mu_X - \frac{1}{2} t^2 (a)^T \Sigma_X a).$$

Since a is arbitrary, we can say for any $t' \in \mathbb{R}^n$, we have

$$E[\exp(it' X)] = \exp(i[t']^T \mu_X - \frac{1}{2} [t']^T \Sigma_X t).$$

That is, X is multivariate Gaussian. □

Example 12.1.5 (bivariate Gaussian distribution). Let $f(x,y)$ be the density of a bivariate Gaussian distribution $MN(\mu,\Sigma)$, where

$$\mu = \begin{Bmatrix} \mu_X \\ \mu_Y \end{Bmatrix}, \Sigma = \begin{Bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{Bmatrix}.$$

Then,

$$f(x,y) = \frac{\exp(-\frac{1}{2(1-\rho^2)}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\frac{\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}]).$$

12.1.9.2 *Affine transformation and its consequences*

Theorem 12.1.1 (affine transformation for multivariate normal distribution). *Let X be an n-dimensional random vector with $MN(\mu,\Sigma)$ distribution. Let $Y = AX + b, A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$. Then Y is an m-dimensional random vector having a $MN(A\mu + b, A\Sigma A^T)$ distribution.*

Proof. Use moment generating function to prove. Let $Y = AX + b$, then from Lemma 11.6.5

$$M_Y(t) = e^{t^T b} M_X(A^T t) = e^{t^T(A\mu+b)+\frac{1}{2}t^T A\Sigma A^T t}$$

which indicates $Y \sim MN(A\mu + b, A\Sigma A^T)$. □

Corollary 12.1.1.1 (sum of two multivariate normal random vectors). *Let $X_1 \sim MN(\mu_1,\Sigma_1)$ and $X_2 \sim MN(\mu_1,\Sigma_1)$ be two n dimensional multivariate normal random variable. It follows that*

- *If X_1 and X_2 are independent, then $Y = X_1 + X_2$ is a multivariate normal random vector with $MN(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$.*
- *If X_1 and X_2 are dependent and (X_1, X_2) are* ***jointly normal*** [a] *with covariance matrix given by*

$$\Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12} & \Sigma_2 \end{bmatrix},$$

then $Y = X_1 + X_2$ is a multivariate normal random vector with $MN(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2 + 2\Sigma_{12})$.

[a] If X_1 and X_2 are not jointly normal, Y is not normal.

Proof. (1) Consider a $2n$-dimensional multivariate normal random variable Z with distribution $\mu = [\mu_1; \mu_2]$, $\Sigma = \Sigma_1 \oplus \Sigma_2$ (Σ is a diagonal block matrix with two blocks Σ_1 and Σ_2). Then we can construct a linear transformation matrix

$$A = \begin{bmatrix} I_n & I_n \end{bmatrix}$$

to construct $Y = AZ$. Apply affine transformation theorem [Theorem 14.1.1] to $Y = AZ$, we have $Y \sim MN(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$. (2) same as (1). □

Note 12.1.1. caution! The joint distribution of two Gaussian margins are not necessarily joint Gaussian:

- Two multivariate normal random variables are not necessarily joint normal.[a]. For example, consider two marginal distribution of Gaussian. For Gaussian copula, the joint distribution is multivariate Gaussian; however, for other copulas including Frank copula and Clayton copula, the joint distribution is not multivariate Gaussian.
- If two multivariate normal random variables are independent, then they are joint normal.

[a] link

Corollary 12.1.1.2 (orthonormal transformation maintains independence). *Let X be a n dimensional random vector with $MN(0, I)$. If C is an orthonormal matrix, then $Y = CX$ has distribution $MN(0, I)$. That is, orthonormal transformation will preserve independence.*

Proof. $Cov(Y) = C^T I C = I$. □

12.1.9.3 *Marginal and conditional distribution*

Lemma 12.1.15 (marginal distribution). *The multivariate Gaussian distribution $\rho(x;\mu,\Sigma)$ on $\mathbb{R}^n$ has marginal distribution on $\mathbb{R}^k, k \leq n$ given as $\rho(x_1;\mu_1,\Sigma_{11}), x_1 \in \mathbb{R}^k$ where we decompose*

$$\mu = [\mu_1,\mu_2]^T, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Proof. Use above Theorem 14.1.1. Let

$$A = \begin{bmatrix} I & 0 \end{bmatrix}.$$

Then $X_1 = AX$. □

Lemma 12.1.16 (full joint distribution can be constructed from pair joint distribution). *Let $X = (X_1, X_2, ..., X_n)^T$ be a random multivariate Gaussian vector with mean $\mu = (\mu_1, \mu_2, ..., \mu_n)^T$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$. The the pair $(X_i, X_j), i \neq j$ has joint distribution*

$$\hat{\mu} = (\mu_i, \mu_j), \hat{\Sigma} \in \mathbb{R}^{2\times 2}, \hat{\Sigma}_{11} = \Sigma_{ii}, \hat{\Sigma}_{12} = \Sigma_{ij}.$$

That is, all the pair joint distribution can construct the full joint distribution.

Proof. Directly from Lemma 14.1.3. □

Remark 12.1.5 (caution! not all the distribution has this property). If the full joint distribution is not Gaussian, then such property (reconstruct full distribution from pair distribution) will not generally hold.

Theorem 12.1.2 (conditional distribution). *The multivariate Gaussian distribution $\rho(x;\mu,\Sigma)$ on $\mathbb{R}^n$ has a conditional Gaussian distribution on $\mathbb{R}^k, k \leq n$ given by*

$$\frac{f(x_1,x_2)}{f(x_2)} = MN(x_1;\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

where we decompose

$$\mu = [\mu_1^T, \mu_2^T]^T, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

with $\mu_1 \in \mathbb{R}^k, \mu_2 \in \mathbb{R}^{n-k}$.

Proof. See link □

Remark 12.1.6 (gaining information). From the conditional distribution, we can see that given the information of x_2, the mean of x_1 will be corrected and the variance of x_1 will be reduced.

Example 12.1.6 (bivariate Gaussian distribution). Let $f(x,y)$ be the density of a bivariate Gaussian distribution $MN(\mu,\Sigma)$, where

$$\mu = \begin{Bmatrix} \mu_X \\ \mu_Y \end{Bmatrix}, \Sigma = \begin{Bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{Bmatrix}.$$

Then,

$$X|Y \sim N(\mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y-\mu_Y), (1-\rho^2)\sigma_X^2).$$

12.1.9.4 *Box Muller transformation*

Lemma 12.1.17 (Box Muller transformation). *Let $X,Y \sim N(0,1)$ and X,Y be independent. Let*

$$R = \sqrt{X^2+Y^2}, \Theta = \arctan(Y/X)$$

. Then

- *R and Θ are independent.*
- *$\Theta \sim U(0,2\pi)$ and $F_R(r) = 1-\exp(-r^2/2)$.*
- *Suppose we have U_1, U_2 being independent uniform on $[0,1]$. Then $2\pi U_1$ and $\sqrt{-2\ln(1-U_2)}$ are independent and have the same distribution of R and Θ.*
- *Further, $\sqrt{-2\ln(1-U_2)}\cos(2\pi U_1)$ and $\sqrt{-2\ln(1-U_2)}\sin(2\pi U_1)$ are independent and have the same distribution of X and Y.*

Proof. (1)Using polar transformation Lemma 11.4.9, we have

$$\begin{aligned} Pr(R<r,\Theta<\theta) &= \int_0^r\int_0^\theta \frac{1}{2\pi}\exp(-\frac{r^2}{2})rdrd\theta \\ &= \int_0^r\int_0^\theta \exp(-\frac{r^2}{2})rdr\frac{1}{2\pi}d\theta \\ &= F_R(R<r)F_\Theta(\Theta<\theta) \end{aligned}$$

Using independence condition Lemma 11.4.3, we know that R and Θ are independent. (2) Integrate directly in (1). (3) Let $U = 1-\exp(-R^2/2)$. Based on probability integral transform [Lemma 14.4.2], we know U is an uniform random variable. Or equivalently, $R = \sqrt{-2\ln(1-U)}$ has the same distribution of R. (4) Note that $X = R\cos(\Theta), Y = R\sin(\Theta)$. □

12.1.10 Lognormal distribution

12.1.10.1 *Univariate lognormal distribution*

Definition 12.1.9 (lognormal distribution). *A random variable Y has a lognormal distribution with parameters μ and σ^2, written as*

$$Y \sim LN(\mu, \sigma^2)$$

if $\log(Y)$ *is normally distributed as* $N(0, \sigma^2)$. *Several equivalent definitions are:*

- $Y \sim LN(\mu, \sigma^2)$ *if and only if* $\log(Y) \sim N(\mu, \sigma^2)$.
- $Y \sim LN(\mu, \sigma^2)$ *if and only if* $Y = e^X$ *with* $X \sim N(\mu, \sigma^2)$.
- *The distribution function is given as*

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp(-\frac{(\ln y - \mu)^2}{2\sigma^2}).$$

Lemma 12.1.18 (basic properties of lognormal distribution). *Let* $Y \sim LN(\mu, \sigma^2)$, *or equivalently* $Y = \exp(X), X \sim N(\mu, \sigma^2)$ *then*

- *The distribution function for Y is given as*

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp(-\frac{(\ln y - \mu)^2}{2\sigma^2}).$$

- $E[Y] = \exp(E[X] + \frac{1}{2}Var[X^2]) = \exp(\mu + \sigma^2/2)$.
- $E[Y^2] = \exp(2\mu + 2\sigma^2), E[Y^m] = \exp(m\mu + \frac{1}{2}m^2\sigma^2)$
- $Var[Y] = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$. *In particular* $\mu = 0$, *we have*

$$E[Y] = \exp(\frac{1}{2}\sigma^2), E[Y^m] = \exp(\frac{1}{2}m^2\sigma^2), Var[Y] = \exp(2\sigma^2) - \exp(\sigma^2).$$

- *If* $X_1 \in N(\mu_1, \sigma_1^2), X_2 \in N(\mu_2, \sigma_2^2)$, *then*

$$E[\exp(X_1 + X_2)] = \exp(E[X_1] + E[X_2] + \frac{1}{2}Var[X] + \frac{1}{2}Var[X_2] + Cov(X_1, X_2)).$$

- $$\mu = \log(\frac{E[Y]^2}{\sqrt{E[Y^2]}}), \sigma^2 = \ln(\frac{E[Y^2]}{E[Y]^2}).$$
- *The median of* Y *is* $\exp(\mu)$.
- *skewness*
$$(\exp(\sigma^2)+2)\sqrt{\exp(\sigma^2)-1} > 0.$$

Proof. (1) Note that
$$x = \ln y, f_Y(y) = f_X(\ln y)\left|\frac{d\ln y}{dy}\right|.$$

(2)(3) Note that for $X \sim N(\mu, \sigma^2)$, $M_X(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$. Then
$$E[Y] = E[\exp(X)] = M_X(1) = \exp(\mu + \frac{1}{2}\sigma^2),$$
and
$$E[Y^2] = E[\exp(2X)] = M_X(2) = \exp(2\mu + 2\sigma^2),$$
and
$$Var[Y] = E[Y^2] - (E[Y])^2 = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2).$$
(4) Note that $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2)$. Then we use (1). (5) Note that the exponential is a monotone function, the median of Y will be $\exp(median\ X) = \exp(\mu)$, where we used the fact that median of X is μ. □

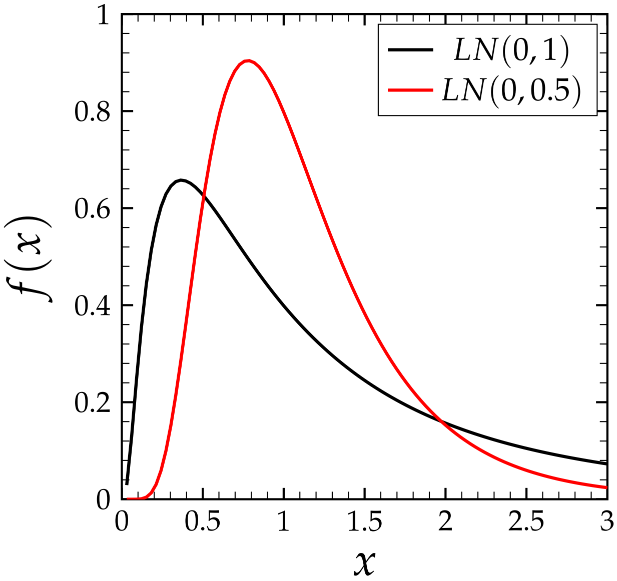

Figure 12.1.2: Density of $LN(0,1)$ and $LN(0,0.5)$. Note the positive skewness.

12.1.10.2 *Extension to univariate lognormal distribution*

Definition 12.1.10. [4]

- ***regular log-normal distribution*** *with parameter* (μ,σ^2) *is given by*

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}}\exp(-\frac{(\ln x-\mu)^2}{2\sigma^2}), x > 0.$$

- ***negative log-normal distribution*** *with parameter* (μ,σ)*, denoted by* $NLN(\mu,\sigma^2)$ *is given by*

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}}\exp(-\frac{(\ln -x-\mu)^2}{2\sigma^2}), x < 0.$$

- ***shifted log-normal distribution*** *with parameter* (μ,σ,τ)*,denoted by* $SLN(\mu,\sigma^2,\tau)$ *is given by*

$$f(x) = \frac{1}{(x-\tau)\sigma\sqrt{2\pi}}\exp(-\frac{(\ln x-\tau-\mu)^2}{2\sigma^2}), x > \tau.$$

- ***negative shifted log-normal distribution*** *with parameter* (μ,σ^2,τ) *is given by*

$$f(x) = \frac{1}{(-x-\tau)\sigma\sqrt{2\pi}}\exp(-\frac{(\ln(-x-\tau)-\mu)^2}{2\sigma^2}), x < -\tau.$$

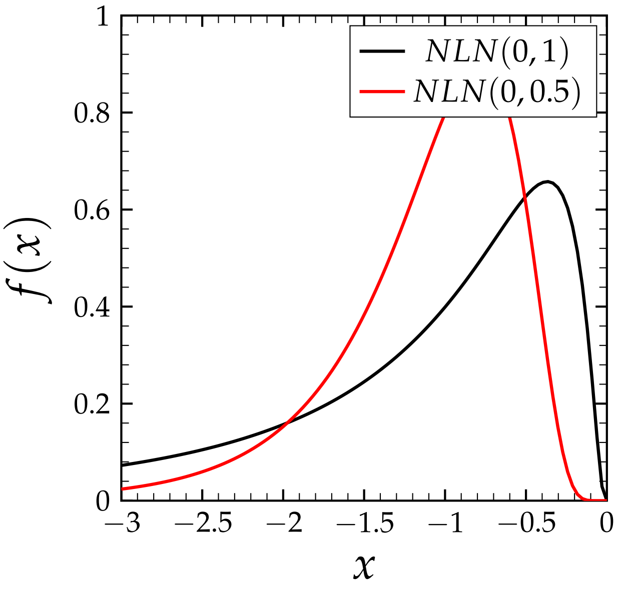

Figure 12.1.3: Density of $NLN(0,1)$ and $NLN(0,0.5)$. Note the negative skewness.

Lemma 12.1.19. *Let $X \sim LN(\mu, \sigma^2)$. It follows that*

- *Let $Y = -X$. Then $Y \sim NLN(\mu, \sigma^2)$.*
- *Let $Z = X + \tau$. Then $Y \sim NLN(\mu, \sigma^2, \tau)$.*
- *Let $W = -X - \tau$. Then $Y \sim NSLN(\mu, \sigma^2, \tau)$.*

Proof. Straight forward from definition and transformation. □

Lemma 12.1.20 (basic properties of shifted lognormal distribution). *Let $X \sim SLN(\mu, \sigma^2, \tau)$. Then*

- $$E[X] = \tau + \exp(\mu + \frac{1}{2}\sigma^2).$$
- $$E[X^2] = \tau^2 + 2\tau\exp(\mu + \frac{1}{2}\sigma^2) + \exp(2\mu + 2\sigma^2).$$
- $$E[X^3] = \tau^3 + 3\tau^2\exp(\mu + \frac{1}{2}\sigma^2) + 3\tau\exp(2\mu + 2\sigma^2) + \exp(3\mu + \frac{9}{2}\sigma^2).$$

Proof. Note that from Lemma 12.1.18, we have if $Y \sim LN(\mu, \sigma^2)$, then $E[Y^m] = \exp(m\mu + \frac{1}{2}m^2\sigma^2)$. Then, we use

$$E[X] = E[Y + \tau] = E[Y] + \tau,$$
$$E[X^2] = E[(Y + \tau)^2] = E[Y^2] + 2\tau E[Y] + \tau^2,$$
$$E[X^3] = E[(Y + \tau)^3] = E[Y^3] + 3\tau E[Y^2] + 3\tau^2 E[Y] + \tau^3.$$

□

12.1.10.3 *Multivariate lognormal distribution*

Definition 12.1.11 (multivariate lognormal distribution). *If $X = (X_1, X_2, ..., X_n) \sim MN(\mu, \Sigma)$, then $Y = \exp(X) = (\exp(X_1), \exp(X_2), ..., \exp(X_n)) \sim MLN(\mu, \Sigma)$, i.e., Y has multivariate lognormal distribution*

Lemma 12.1.21 (basic properties of multivariate lognormal distribution). *Let $X = (X_1, X_2, ..., X_n) \sim MN(\mu, \Sigma)$ and $Y = \exp(X) = (\exp(X_1), \exp(X_2), ..., \exp(X_n))$. Then*

- $E[Y_i] = \exp(\mu_i + \frac{1}{2}\Sigma_{ii})$.

- $E[Y_iY_j] = \exp(\mu_i + \mu_j + \frac{1}{2}(\Sigma_{ii} + \Sigma_{jj} + 2\Sigma_{ij})) = E[Y_i][Y_i]\exp(\Sigma_{ij})$.
- $Var[Y_i] = \exp(2\mu_i + \Sigma_{ii})(\exp(\Sigma_{ii}) - 1)$.
- $Cov[Y_iY_j] = \exp(\mu_i + \mu_j + \frac{1}{2}(\Sigma_{ii} + \Sigma_{jj}))(\exp(\Sigma_{ij}) - 1)$.

Proof. (1)Note that $M_X(t) = \exp(t^T\mu + \frac{1}{2}t^T\Sigma t), t \in \mathbb{R}^n$, and $E[Y_i] = M_X(e_i)$. (2) Let $t = e_i + e_j$. Then

$$E[Y_iY_j] = M_X(t) = \exp(\mu_i + \mu_j + \frac{1}{2}(\Sigma_{ii} + \Sigma_{jj} + 2\Sigma_{ij})).$$

(3)

$$Var[Y_i] = E[Y_iY_i] - E[Y_i]E[Y_i].$$

(4)

$$Cov[Y_i, Y_j] = E[Y_iY_j] - E[Y_i]E[Y_j].$$

□

12.1.11 Exponential distribution

Definition 12.1.12 (exponential distribution). *A random variable X is said to have an exponential distribution $Exp(\lambda)$ with parameter λ if it has pdf given as*

$$p(x|\lambda) = \lambda\exp(-\lambda x)$$

with $x \in [0, \infty)$.

Lemma 12.1.22 (basic properites). *Let X be a random variable with exponential distribution with parameter λ, then we have*

- $E[X] = 1/\lambda$
- $Var[X] = 1/\lambda^2$
- *memoryless:*

$$P(X > s + t|X > s) = P(X > t)$$

(even though $P(X > s + t) < P(X > t)$)

Proof. (1)(2) are straightforward. (3) The cmf is given as

$$F(t) = \int_0^t \lambda\exp(-\lambda\tau)d\tau = 1 - \exp(-\lambda t)$$

$$P(X > s+t | X > s) = \frac{P(X > s+t \cap X > s)}{P(X > s)} = \frac{P(X > s+t)}{P(X > s)} = \frac{\exp(-\lambda(s+t))}{\exp(-\lambda t)} = \exp(-\lambda t)$$

□

Remark 12.1.7 (interpretation of memorylessness). Supppose we are waiting for an event to occur, and we model the waiting time as a random variable X with $Exp(\lambda)$. If we already wait for s time, the distribution that we need to wait an extra of t time is the same as the distribution of the waiting time at time 0. **Exponential distribution is the only memoryless continuous distribution**[5].

Lemma 12.1.23 (Normal approximate sum of Exponential). *Let $X_1, ..., X_n$ be independent iid random variable of $Exp(\lambda)$, then*

$$Y = \sum_{i=1}^{n} X_i$$

can be approximated(when $n \to \infty$) by

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

where $\mu = n/\lambda$,and $\sigma = n/\lambda^2$.

Proof. Directly from Central Limit Theorem [Theorem 11.11.3]. Also see Gamma distribution properties, since exponential distribution is a special case of Gamma distribution. □

12.1.12 Gamma distribution

Definition 12.1.13 (Gamma distribution). *[6, p. 42] A random variable X is said to have a Gamma distribution $Gamma(a,b)$ with parameter a, b if it has pdf given as*

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

with support $x \in (0, \infty)$.

Remark 12.1.8 (exponential distribution is a special case). An exponential distribution with parameter b is a Gamma distribution $Gamma(1,b)$with

$$f(x) = be^{-bx}.$$

Remark 12.1.9 (Application in arrival times of Poisson process). If $N(t)$ is a Poisson process with rate λ, then the arrival time $T_1, T_2, ...$ have $T_n \sim Gamma(n, \lambda)$ distribution.(See Lemma 18.5.4)

Caution! Gamma distribution is different from Gamma function $\Gamma(t)$, which is given as

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$$

Remark 12.1.10 (conjugate prior for Poisson distribution). Gammad distribution conjugate prior for the parameter of Poisson distribution. When integrate out x in $\Gamma(t)$, we have

$$\int_0^\infty x^{a-1} e^{-bx} dx = \Gamma(a)/b^2$$

Lemma 12.1.24 (mean and variance). *The Gamma distribution $Gamma(a, b)$ has mean a/b and variance a/b^2.*

Proof. Using the property of

$$\int_0^\infty x^{a-1} e^{-bx} dx = \Gamma(a)/b^a,$$

we can show the result. □

Theorem 12.1.3 (sum of Gamma random variables). *[3, p. 163] Let $X_1, ..., X_n$ be independent random variables. Suppose $X_i \sim Gamma(a_i, b), \forall i = 1, ..., n$. Then*

$$Y = \sum_{i=1}^n X_i \sim Gamma(\sum_{i=1}^n a_i, b)$$

Proof. This can be proved using moment generating functions. □

Lemma 12.1.25 (Normal approximate sum of Gamma). *Let $X_1, ..., X_n$ be independent iid random variables of $Gamma(a, b)$, then*

$$Y = \sum_{i=1}^n X_i$$

can be approximated(when $n \to \infty$) by

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

where $\mu = na/b$,and $\sigma = na/b^2$.

Proof. Directly from Central Limit Theorem [Theorem 11.11.3]. □

12.1.13 Hypergeometric distribution

Definition 12.1.14 (hypergeometric distribution distribution). *[3, p. 148] A random variable X is said to have a hypergeometric distribution $HG(N,K,n)$ with parameter N,K,n if it has pmf given as*

$$p(x=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

with support $x \in \{0,1,...,\min(n,K)\}$. Note that the parameters should be non-negative integers and satisfying

$$N \geq K, N \geq n.$$

Remark 12.1.11 (interpretation). $p(x=k)$ describes the probability of k successes in n draws, without replacement, from a finite population of size N that contains exactly K successes.

Lemma 12.1.26 (combinatorial identities). *Assuming $K \geq n$, we have*

$$\sum_{0\leq k \leq n} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} = 1$$

Lemma 12.1.27 (mean of a hypergeometric distribution). *[3, p. 148] Let X be a random variable with $HG(N,K,n)$, then its mean is*

$$E[X] = n\frac{K}{N}$$

12.1.14 Beta distribution

Definition 12.1.15 (Beta distribution). *A random variable X is said to have a Beta distribution $B(a,b)$ with parameter a,b if it has a pdf given as*

$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

with support $x \in [0,1]$.

Remark 12.1.12.

- $$\int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$
- Beta distribution is commonly **used as the conjugate prior for binomial distribution**, where
$$p(y_1,...,y_n|\theta) = \theta^{\sum_i y_i}(1-\theta)^{n-\sum_i y_i}, y_i \in \{0,1\}$$
then the posterior distribution will also be Beta.

Lemma 12.1.28 (basic property). *Let X be a random variable with distribution $B(a,b)$.*

- $$E[X] = \frac{a}{a+b}.$$
- $$E[X^2] = \frac{a(a+1)}{(a+b)(a+b+1)}.$$
- $$E[X^r] = \frac{a(a+1)\cdots(a+r-1)}{(a+b)(a+b+1)\cdots(a+b+r-1)}.$$
- $$Var[X] = \frac{ab}{(a+b)^2(a+b+1)}.$$
- *The mode of X, i.e., the value x that has the maximum probability is*
$$x^* = \frac{a-1}{a+b-2}.$$

Proof. (1)This can be proved using properties of Gamma distribution.

$$\begin{aligned}
E[X] &= \int_0^1 xf(x)dx \\
&= \int_0^1 \frac{x^a(1-x)^{b-1}}{B(a,b)} \\
&= \frac{B(a+1,b)}{B(a,b)} \\
&= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} / \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \\
&= \frac{\Gamma(a+1)\Gamma(a+b)}{\Gamma(a+b+1)\Gamma(a)} \\
&= \frac{a}{a+b}
\end{aligned}$$

(2)

$$\begin{aligned}
E[X^2] &= \int_0^1 x^2 f(x)dx \\
&= \int_0^1 \frac{x^{a+1}(1-x)^{b-1}}{B(a,b)} \\
&= \frac{B(a+2,b)}{B(a,b)} \\
&= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} / \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \\
&= \frac{\Gamma(a+2)\Gamma(a+b)}{\Gamma(a+b+2)\Gamma(a)} \\
&= \frac{a(a+1)}{(a+b)(a+b+1)}
\end{aligned}$$

(3) Use $Var[X] = E[X^2] - E[X]^2$. (4) To find the maximizer for $x^{a-1}(1-x)^{b-1}$, we take the log and maximize it. We have

$$\ln f(x) = (a-1)\ln x + (b-1)\ln(1-x).$$

Take the derivative with respect to x and set to 0, we have

$$\begin{aligned}
\frac{a-1}{x} &= \frac{b-1}{1-x} \\
(a-1)(1-x) &= x(b-1) \\
\implies x^* &= \frac{a-1}{a+b-2}.
\end{aligned}$$

□

12.1.15 Multinomial distribution

Definition 12.1.16. *[6, p. 35] A discrete random vector $X = (X_1, ..., X_n)$ is said to have multinomial distribution with parameters $(p_1, ..., p_n)$ and m if its pmf is given as*

$$f(x_1, x_2, ..., x_n) = \frac{m!}{x_1!...x_n!} p_1^{x_1} ... p_n^{x_n}$$

where we require $x_i \in \{0, ..., m\}, \sum x_i = m, \sum p_i = 1$.

Remark 12.1.13. Consider m independent experiment, each has n outcomes with probability p_i to occur. The outcome distribution is given as[7]

$$f(x_1, x_2, ..., x_n) = \frac{m!}{x_1!...x_n!} p_1^{x_1} ... p_n^{x_n}$$

where $\sum x_i = m, \sum p_i = 1$.

Lemma 12.1.29 (basic properties of Multinomial Distribution). *Let $X = (X_1, ..., X_n)$ discrete random vector with multinomial distribution with parameters $p = (p_1, ..., p_n)$ and m.*

-
- $$E[X_i] = np_i.$$
- $$Var[X_i] = np_i(1 - p_i), Cov(X_i, X_j) = np_i(1 - p_i),$$
  or in vector form
  $$Var[X] = n(diag(p) = pp^T).$$

Proof. (1)This can be proved using properties of Gamma distribution.

$$\begin{aligned} E[X_i] &= \int_0^1 x_i f(x) dx \\ &= \frac{\prod_{k=1}^K \Gamma(a_k + \delta_{ik})}{\Gamma(a_0 + 1)} / \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(a_0)} \\ &= \frac{a_i}{a_0} \end{aligned}$$

(2)

$$\begin{aligned} E[X_i^2] &= \int_0^1 x_i^2 f(x)dx \\ &= \frac{\prod_{k=1}^K \Gamma(a_k + 2\delta_{ik})}{\Gamma(a_0+2)} / \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(a_0)} \\ &= \frac{a_i(a_i+1)}{(a_0+1)a_0} \end{aligned}$$

(3) Use $Var[X] = E[X^2] - E[X]^2$. (4) To find the maximizer for $f(x)$, we take the log and maximize it. The optimality condition requires that $x_i^* \propto a_i - 1$ and $\sum_{i=1}^K a_i = 1$. □

12.1.16 Dirichlet distribution

Definition 12.1.17. *[6, p. 49] A random vector $X = (X_1, ..., X_K)$ is said to have a Dirichlet distribution distribution with parameter $a = (a_1, ..., a_K)$ if it has pdf given as*

$$f(x_1, ..., x_K) = \frac{1}{B(a)} \prod_{k=1}^K x_k^{a_k - 1}$$

with support $x \in \{x : 0 \leq x_k \leq 1, \sum_k x_k = 1, \forall k = 1, 2..., K\}$, and $B(a)$ is a normalization constant given as

$$B(a) = \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(\sum_k a_k)},$$

where $\Gamma(\cdot)$ is the Gamma function.

Remark 12.1.14.

- Dirichlet distribution can be viewed as multivariate generalization of Beta distribution.
- Dirichlet distribution is usually **used as the conjugate prior** for multinomial distribution.

Lemma 12.1.30 (basic properties of Dirichlet Distribution). *Let $X = (X_1, X_2, ..., X_K), x_i \in (0,1), \sum_{i=1}^K x_i = 1$, be a random vector with distribution $B(a), a \in \mathbb{R}^K$. Let $a_0 = \sum_{i=1}^K a_i$.*

- $$E[X_i] = \frac{a_i}{\sum_{i=1}^K a_i}.$$

- $$E[X_i^2] = \frac{a_i(a_i+1)}{(a_0)(a_0+1)}.$$
- $$Var[X_i] = \frac{a_i(a_0-a_i)}{a_0^2(a_0+1)}.$$
- *The mode of X, i.e., the value x that has the maximum probability is*
  $$x_i^* = \frac{a_i-1}{a_0-K}.$$

Proof. (1)This can be proved using properties of Gamma distribution.

$$\begin{aligned} E[X_i] &= \int_0^1 x_i f(x)dx \\ &= \frac{\prod_{k=1}^K \Gamma(a_k+\delta_{ik})}{\Gamma(a_0+1)} / \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(a_0)} \\ &= \frac{a_i}{a_0} \end{aligned}$$

(2)

$$\begin{aligned} E[X_i^2] &= \int_0^1 x_i^2 f(x)dx \\ &= \frac{\prod_{k=1}^K \Gamma(a_k+2\delta_{ik})}{\Gamma(a_0+2)} / \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(a_0)} \\ &= \frac{a_i(a_i+1)}{(a_0+1)a_0} \end{aligned}$$

(3) Use $Var[X] = E[X^2] - E[X]^2$. (4) To find the maximizer for $f(x)$, we take the log and maximize it. The optimality condition requires that $x_i^* \propto a_i - 1$ and $\sum_{i=1}^K a_i = 1$. □

12.1.17 χ^2-distribution

12.1.17.1 *Basic properties*

Definition 12.1.18. *A random variable X is said to have a $\chi^2(n)$ distribution with parameter $n \in \mathbb{Z}_+$ if it has pdf given as*

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-y/2}$$

with $x \in (0, +\infty)$

Remark 12.1.15 (special case of Gamma distribution). $\chi^2(n)$ has the same distribution of Gamma(n/2,2).

Definition 12.1.19 (alternative). *The* χ^2*-distribution with k degrees of freedom is the distribution of a sum of squares of k independent standard normal random variables. Mathematically, if* $X_1, X_2, ..., X_k$ *are iid random variable with* $X_i \sim N(0,1)$*, the random variable*

$$Q = \sum_{i=1}^{k} X_i^2$$

is distributed according to the χ^2 *distribution with k degrees of freedom, writen as* $Q \sim \chi^2(k)$*.*

Lemma 12.1.31 (basic property). *[3, pp. 161–163] Let* X_1, X_2 *be independent random variables. Suppose* $X_1 \sim \chi^2(a_1), X_2 \sim \chi^2(a_2)$*. Then*

- $Y = X_1 + X_2 \sim \chi^2(a_1 + a_2)$
- $\lambda X_1 \sim \lambda^2 \chi^2(a_1)$
- *The moment generating function is given by*

$$M(t) = (1-2t)^{-r/2}.$$

Proof. (1)This can be proved using properties of Gamma distribution. (2) λX_1 can be viewed as the sum of squares of normal random variables Y_i with $N(0, \lambda^2)$. Then $\sum_{i=1}^n (Y_i/\lambda)^2 \sim \chi^2(n)$. □

Lemma 12.1.32 (expectation and variance). *Let random variable X has distribution of* $\chi^2(n)$*, then*

$$E[X] = n, Var[X] = 2n$$

In particular,

$$E[X/n] = 1, Var[X/n] = 0 \text{ as } n \to \infty.$$

that is the random variable X/n *becomes deterministic constant as* $n \to \infty$*.*

Proof. (1)Let $Z \sim \chi^2(1), Z = Y^2, Y \sim N(0,1)$, then $E[Z] = Var[Y] + (E[Y])^2 = 1$. $Var[Z] = E[Z^2] - (E[Z])^2 = E[Y^4] - 1 = 3 - 1 = 2$. (2) Use linearity of expectation that $E[X/n] = E[X]/n = 1$. Use $Var[X/n] = Var[X]/n^2 = 2/n$. □

12.1.17.2 *Noncentral chi-squared distribution*

Definition 12.1.20 (noncentral chi-squared distribution). *Let $(X_1, X_2, ..., X_k)$ be k independent, normally distributed random variables with mean μ_i and unit variances. Then the random variable*

$$Y = \sum_{i=1}^{k} X_i^2$$

is distributed according to the ***noncentral chi-squared distribution*** *with parameter k specifying the degree of freedom and λ, known as the* ***noncentrality parameter****, given by*

$$\lambda = \sum_{i=1}^{k} \mu_i^2.$$

12.1.18 Wishart distribution

Definition 12.1.21 (Wishart distribution). *Let $X_1, ..., X_n$ be independent p dimensional multivariate normal random vector with distribution $MN(0, V)$. Let $X = [X_1, ..., X_n]$. Then $M = XX^T$ is said to have Wishart distribution with parameter (n, p, V).*

Definition 12.1.22 (Wishart distribution). *A random matrix $M \in \mathbb{R}^{p\times p}$ is said to have the Wishart distribution with parameters $W_p(n, V)$ if it has pdf*

$$f(M) = \frac{1}{2^{np/2}\Gamma_p(\frac{n}{2}|V|^{n/2})}|M|^{n-p-1/2}\exp(\frac{1}{2}Tr[V^{-1}M]),$$

with the support M be the set of all symmetric positive definite matrices. Here $\Gamma_p(\alpha)$ is the multivariate gamma function.

Lemma 12.1.33 (basic properties).

- *(reduction to χ^2) If $M \in \mathbb{R}^{1\times 1}$, then*

$$M \sim W_1(n, \sigma^2) = \sigma^2\chi^2(n).$$

- *For $M \sim W_p(n, V)$, then $B^T MB \sim W_m(n, B^T VB)$, where $B \in \mathbb{R}^{p\times m}$.*
- *For $M \sim W_p(n, V)$, then $V^{-1/2}MV^{-1/2} \sim W_m(n, I)$.*
- *If M_i are independent $W_p(n_i, V)$, then $\sum_{i=1}^{k} \sim W_p(\sum_{i=1}^{k} n_i, V)$.*

- *If $M \sim W_p(n, V)$, then $E[M] = nV$.*
- *If M_1, M_2 are independent and $M_1 + M_2 = M \sim W_p(n, V)$. Further if $M_1 \sim W_p(n_1, V)$, then $M_2 \sim W_p(n - n_1, V)$.*

Lemma 12.1.34 (sample covariance). *The sample covariance*

$$\hat{Cov} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^T$$

where X_i are iid $MN(0, V)$, and $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, has the property of

$$E[\hat{Cov}] = V.$$

12.1.19 t-distribution

12.1.19.1 *Standard t distribution*

Definition 12.1.23 (t distribution). *[3, p. 192] A random variable X is said to have a $t(n)$ distribution with parameter $n \in \mathbb{Z}_+$ if it has pdf given as*

$$f(x) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}}(1 + \frac{t^2}{n})^{-(n+1)/2}$$

with $x \in (-\infty, +\infty)$

Definition 12.1.24 (alternative). *Let random variable $W \sim N(0, 1)$, Let random variable $V \sim \chi^2(n)$* ***independent of*** *V. Define a new random variable T as*

$$T = \frac{W}{\sqrt{V/n}}$$

Then T has a t-distribution with degree of freedom n, denoted by T_n or t_n.

Remark 12.1.16 (comparison with normal distribution).

- t distribution generally have shorter peak and fatter tails than normal distribution.
- $t_n \to N(0, 1)$ as $n \to \infty$.

Lemma 12.1.35 (mean and variance of t-distribution). *The mean for a t-distribution with degree of n is given by*

$$E[t_n] = \begin{cases} 0, n > 1 \\ \infty(undefined), n = 1 \end{cases}.$$

The variance for a t-distribution with degree of n is given by

$$Var[t_n] = \begin{cases} \frac{n}{n-2}, n > 2 \\ \infty, n = 1, 2 \end{cases}.$$

12.1.19.2 *classical t distribution*

Definition 12.1.25. *[8, p. 95] If Y has a standard t_n distribution, then*

$$Z = \mu + \lambda Y$$

is said to have a $t_n(\mu, \lambda^2)$ distribution.

Lemma 12.1.36 (mean and variance of classical t-distribution). *Let Z be a random variable $t_n(\mu, \lambda^2)$. Then*

$$E[Z] = \begin{cases} \mu, n > 1 \\ \infty(undefined), n = 1 \end{cases},$$

and

$$Var[Z] = \begin{cases} \lambda^2 \frac{n}{n-2}, n > 2 \\ \infty, n = 1, 2 \end{cases}.$$

12.1.19.3 *Multivariate t distribution*

Definition 12.1.26 (multivariate t distribution). *[8]*

- *Let Z be a d dimensional multivariate Gaussian $MN(0, \Sigma)$, and $\mu \in \mathbb{R}^d$. The d dimensional random vector Y, defined as,*

$$X = \mu + \sqrt{\frac{n}{W}} Z,$$

where $W \sim \chi^2(n)$ and W is ***independent*** *of Z, has a $t_n(\mu, \Sigma)$ multivariate distribution.*

- *Let $X \sim t_n(\mu, \Sigma)$. Then X has the density given by*

$$f(x) = \frac{\Gamma((n+d)/2)}{\Gamma(n/2)n^{d/2}\pi^{d/2}|\Sigma|^{1/2}}(1 + \frac{1}{n}(x-\mu)^T\Sigma^{-1}(x-\mu))^{-(n+d)/2}.$$

Lemma 12.1.37 (mean and variance of multivariate t-distribution). *Let Z be a random variable $t_n(\mu, \Sigma)$. Then*

$$E[Z] = \begin{cases} \mu, n > 1 \\ \infty(undefined), n = 1 \end{cases},$$

and

$$Cov[Z] = \begin{cases} \Sigma\frac{n}{n-2}, n > 2 \\ \infty, n = 1, 2 \end{cases}.$$

12.1.20 F-distribution

Definition 12.1.27 (F distribution). *[3, p. 192] A random variable X is said to have a $F(n_1, n_2)$ distribution with parameter $n_1, n_2 \in \mathbb{Z}_+$ if it has pdf given as*

$$f(x) = \frac{\Gamma((n_1+n)2)/2)(n_1/n_2)^{n_1/2}y^{n_1/2-1}}{\Gamma(n_1/2)\Gamma(n_2/2)[1 + (n_1x/n_2)]^{(n_1+n_2)/2}}$$

with $x \in (0, +\infty)$

Definition 12.1.28 (alternative). *Given two **independent** chi-squared random variables W and V having r_1 and r_2 degrees of freedom. We define a new random variable*

$$W = \frac{U/r_1}{V/r_2}$$

Then W has a F-distribution with parameter (r_1, r_2).

Lemma 12.1.38 (inverse relationship). *Let X be a random variable with distribution $F(n_1, n_2)$, then $1/X$ is a random variable with distribution $F(n_2, n_1)$.*

Proof. Directly from definition. □

Lemma 12.1.39 (relationship to t distribution). *Let X be a random variable with standard t distribution with n degrees of freedom. Then*

$$X^2 \sim F(1, n).$$

That is, X^2 has the distribution of $F(1, n)$.

Proof. Directly from definition. □

Definition 12.1.29 (noncentral F distribution). *Given two chi-squared random variables W and V such that V is a noncentral chi-squared random variable with non-centrality parameter λ and degree of freedom r_1, and W is a chi-squared random variable having r_1 r_2 degrees of freedom. We define a new random variable*

$$W = \frac{U/r_1}{V/r_2}$$

Then W has a noncentral F-distribution with parameter (λ, r_1, r_2). .

12.1.21 Empirical distributions

Definition 12.1.30 (empirical cumulative distribution function(CDF)). *Given N iid random variables $Y_1, Y_2, ..., Y_N$ with common cdf $F(t)$, the empirical CDF is defined by*

$$\hat{F}_N(t) = \frac{\text{number of elements in the sample} \leq t}{N} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{Y_i \leq t}$$

.

Lemma 12.1.40 (basic statistic properties). *Let $\hat{F}_N(t)$ be the empirical cdf of a random sample of size N. For a fixed t, we have*

- *$N\hat{F}_N(t)$ is a binomial random variable with parameter (N, p),where $p = F(t)$.*
- *$N\hat{F}_N(t)$ is an unbiased estimator for $NF(t)$.*
- *$N\hat{F}_N(t)$ has variance $NF(t)(1 - F(t))$.*

Proof. (1) Note that based on the definition of $\hat{F}_N(t)\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{Y_i \leq t}$, $\mathbf{1}_{Y_i \leq t}$ is a Bernoulli random variable with parameter $p = F(t)$. Therefore, $N\hat{F}_N(t) = \sum_{i=1}^{N}\mathbf{1}_{Y_i \leq t}$ will follow a binomial distribution of parameter (N, p). (2)

$$E[N\hat{F}_N(t)] = Np = NF(t).$$

(3)

$$Var[N\hat{F}_N(t)] = Np(1 - p) = NF(t)(1 - F(t)).$$

□

12.1.22 Heavy-tailed distributions

12.1.22.1 *Basic characterization*

Definition 12.1.31 (Heavy-tailed distribution). *The distribution of a random variable X with distribution function F is said to have a heavy right tail if*

$$\lim_{x\to\infty} e^{\lambda x} Pr(X > x) = \infty, \forall \lambda > 0.$$

Remark 12.1.17 (interpretation). Heavy-tailed distributions have densities decaying slower in the tails than the normal.

12.1.22.2 *Pareto and power distribution*

Definition 12.1.32 (Pareto distribution). *A random variable X is said to have Pareto distribution with scale parameter $x_m > 0$ and shape parameter $\alpha > 0$ if its has pdf*

$$f_X(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, x \geq x_m, \\ 0, x < x_m. \end{cases};$$

or cdf

$$f_X(x) = \begin{cases} 1 - (\frac{\alpha x_m}{x})^\alpha, x \geq x_m, \\ 0, x < x_m. \end{cases}.$$

X has support $[x_m, \infty)$.

Definition 12.1.33 (power law distribution). *A random variable X is said to have power law distribution with parameters* K, α *if its has probability characterization on its tail given by*

$$Pr(X > x) = Kx^{-\alpha}.$$

Remark 12.1.18 (Pareto distribution and power law distribution are heavy-tailed distribution)**.** Note that since power grows much slower than the exponentialA.2.1, therefore

$$\lim_{x\to\infty} e^{\lambda x} Pr(X > x) = \infty, \forall \lambda > 0.$$

12.1.22.3 *Student t distribution family*

Definition 12.1.34 (Student's t-Distribution family). *The t distribution has a single parameter,* $\nu > 0$*, known as* degrees of freedom*. The density function is given as*

$$f_\nu(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1 + \frac{x^2}{\nu})^{-\frac{1}{2}(\nu+1)}$$

The first two members of family are

1. $f_1(x) = \frac{1}{\pi(1+x^2)}$
2. $f_2(x) = \frac{1}{2\sqrt{2}}(1 + x^2/2)^{-3/2}$

The $\nu = 1$ *density is known as* Cauchy's density*. As* $\nu \to \infty$*, the density distribution tends to the standard normal density.*

Definition 12.1.35 (Cauchy distribution). *The Cauchy distribution with parameter* (x_0, γ) *has the probability density function*

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma}(\frac{\gamma^2}{(x - x_0)^2 + \gamma^2}),$$

where x_0 is the location parameter, specifying the location of the peak of the distribution, and γ is the scale parameter which specifies the half-width at half-maximum. ***Standard Cauchy distribution*** *is Cauchy distribution with parameter $(0,1)$.*

Remark 12.1.19 (nonexistence of moments)**.**

- The Cauchy distribution is an example of a distribution which has no mean, variance or higher moments. And therefore the moment generating function does not exist. However, the **mode and median** are well defined and both equal to x_0.
- The nonexistence of expectation is because of the $E[\![X|] < \infty$.

Lemma 12.1.41 (sum of Cauchy distribution). *If $X_1, ..., X_n$ are independent and identically distributed random variables, each with a standard Cauchy distribution, then the sample mean*

$$\overline{X} = (X_1 + ... + X_n)/n$$

has the same standard Cauchy distribution.

Proof. Note that we need to use characteristic function to prove, since the moment generating function does not exist. □

12.1.22.4 *Gaussian mixture distributions*

Definition 12.1.36 (normal scale mixture distribution). *[8, p. 99] The normal scale mixture distribution is the distribution of the random variable*

$$Y = \mu + \sqrt{U}Z,$$

where μ is constant equal to the mean, and $Z \sim N(0,1)$, U is a positive random variable giving the variance of each component, and Z and U are independent.

If U can assume only a finite number of values, then Y has a ***discrete scale mixture distribution****. If U is continuously distributed, then Y has a* ***continuous scale mixture distribution****.*

Example 12.1.7 (discrete Gaussian mixture distribution). Let $\mu = 0$, and U have the following distribution

$$P(U = 25) = 0.1, P(U = 1) = 0.9.$$

Then

$$Y = \mu + \sqrt{U}Z,$$

is the mixture of 10% of $N(0,25)$ and 90% of $N(0,1)$.

Example 12.1.8 (t distribution). The t_n distribution with n degrees of freedom is a continuous Gaussian mixture with

$$\mu = 0, U = \frac{n}{W},$$

where $W \sim \chi^2(n)$.

Definition 12.1.37 (multivariate normal variance mixtures). *The random vector X has a multivariate normal variance mixture distribution if*

$$X \triangleq \mu + \sqrt{W} A Z$$

where

- $Z \sim MN(0, I_k)$
- *W is a* ***positive*** *scalar random variable which is independent of Z*
- $A \in \mathbb{R}^{d\times k}$ *and* $\mu \in \mathbb{R}^d$ *are a matrix and a vector of constants*

Example 12.1.9 (special case: multivariate t distribution). The t_n distribution with n degrees of freedom is a continuous Gaussian mixture with

$$\mu = 0, U = \frac{n}{W},$$

where $W \sim \chi^2(n)$.

12.2 Characterizing distributions

12.2.1 Skewness and kurtosis

Skewness is a measure of symmetry of a statistical distribution [Figure 12.2.1].There are two types of skewness.

- **Negative skewness** indicates that the mean of the data values is less than the median, and the data distribution is **left-skewed**.
- **Positive skewness** indicates that the mean of the data values is greater than the median, and the data distribution is **right-skewed**.

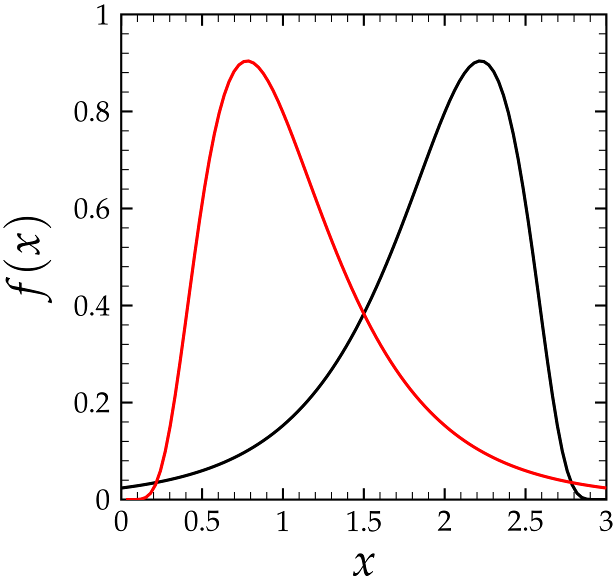

Figure 12.2.1: Distributions with left-skewness (black) and right-skewness (red).

Skewness can be computed quantitatively via the following definition.

Definition 12.2.1 (skewness). *The skewness of an univariate population for random variable X is defined by*

$$\gamma_1 = E[(\frac{X-\mu}{\sigma})^3] = \frac{E[(X-\mu)^3]}{(E[(X-\mu)^2])^{3/2}} = \frac{\mu_3}{\mu_2^{3/2}}$$

where μ_2 and μ_3 are the second and the third ***central moments****.*

Example 12.2.1. Let $X \sim N(\mu, \sigma^2)$. Then the skewness of X distribution is $\gamma_1 = E[(\frac{X-\mu}{\sigma})^3] = E[Z^3] = 0, Z \sim N(0,1)$, where we use the fact the third moment for a standard normal is zero [Lemma 12.1.10].

Kurtosis is a measure of tail shape of a distribution. There are three types of kurtosis [Figure 12.2.2]:

- **Mesokurtic** distributions have zero excess kurtosis. Normal distribution is mesokurtic.
- **Leptokurtic** distributions have excess kurtosis greater than 0. This type of distribution is one with extremely thick tails and a very thin and tall peak. t-distribution and Laplace distribution are leptokurtic.
- **Platykurtic** distributions have excess kurtosis smaller than 0. This type of distribution has a short and broad-looking peak. Uniform distribution is platykurtic.

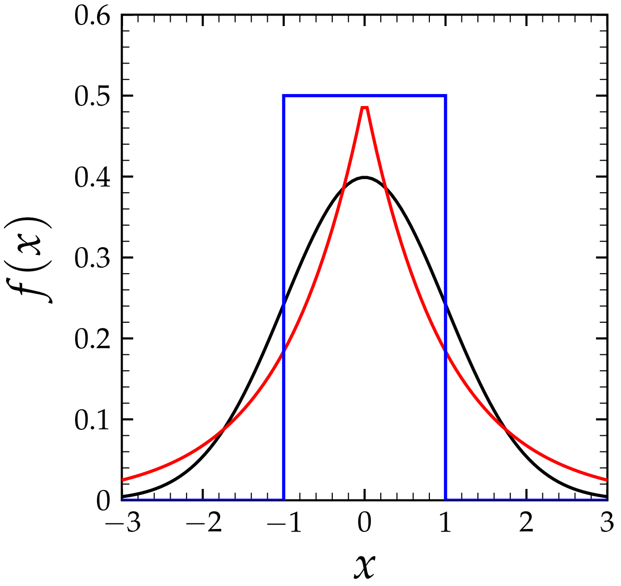

Figure 12.2.2: Distributions with zero excess Kurtosis (Normal distribution, black), positive excess kurtosis (Laplace distribution, red), and negative excess Kurtosis (Uniform distribution, blue).

Kurtosis can be computed quantitatively via the following definition.

Definition 12.2.2 (kurtosis, excess kurtosis).

- *The **kurtosis** of a univariate population is defined by*

$$\gamma_2 = E[(\frac{X-\mu}{\sigma})^4] = \frac{E[(X-\mu)^4]}{(E[(X-\mu)^2])^2} = \frac{\mu_4}{\mu_2^2},$$

where μ_2 and μ_4 are the second and the fourth central moments.

- *The **excess kurtosis** of a univariate population is defined by*

$$\gamma_2^{ex} = \gamma_2 - 3.$$

Example 12.2.2. Let $X \sim N(\mu, \sigma^2)$. Then the Kurtosis of X distribution is $\gamma_2 = E[(\frac{X-\mu}{\sigma})^4] = E[Z^4] = 3, Z \sim N(0,1)$, where we use the fact the fourth moment for a standard normal is 3 [Lemma 12.1.10].

12.2.2 Percentiles and quantiles

12.2.2.1 *Basics*

Percentiles are cut points in the support that divide the distribution into different parts, with each part occupying different probability mass. Compared to skewness and kurtosis, percentiles offers a more comprehensive characterization on the shape of the distribution.

Definition 12.2.3 (percentile of a distribution). *The α **percentile**($\alpha \in [0,1]$) of a probability distribution of random number X is a number p in the support D of the support such that*

$$Pr(x < p) = \alpha, Pr(x > p) = 1 - \alpha.$$

Or equivalently, the α percentile is given by the inverse of cdf

$$p = F_X^{-1}(\alpha).$$

Example 12.2.3 (percentiles of a standard normal distribution). In Figure 12.2.3, we plot percentiles at $\alpha = 0.1, 0.2, ..., 0.9$ for a standard normal distribution.

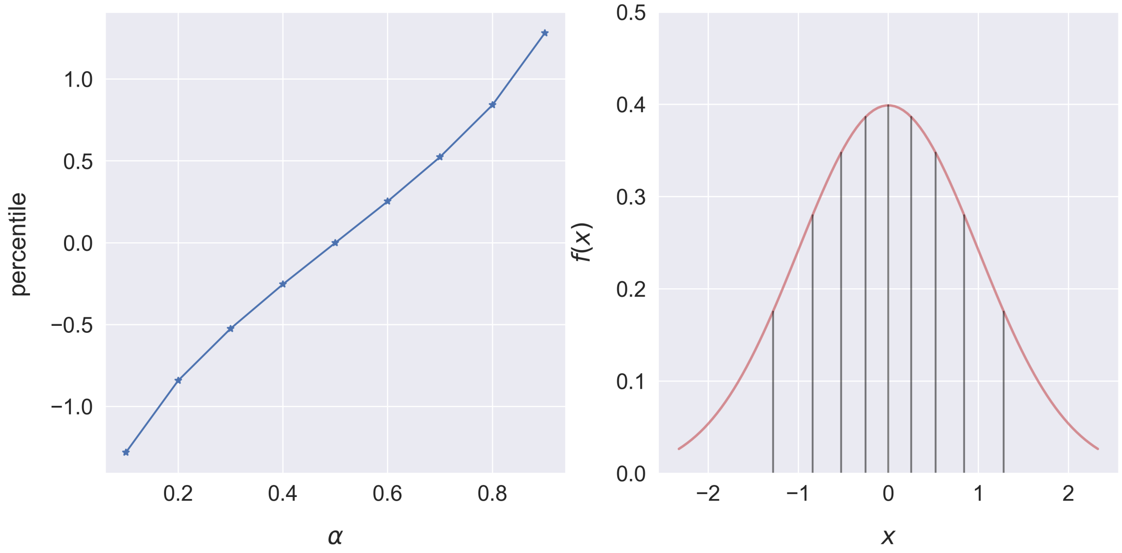

Figure 12.2.3: Percentile points at $\alpha = 0.1, 0.2, ..., 0.9$ for a standard normal distribution.

Definition 12.2.4 (percentile in a set of sample values). *The α **percentile** ($\alpha \in [0,1]$) of a set of values is a value in $\mathbb{R}$ that divides them so that $100\alpha\%$ of values lie below and $100(1-\alpha)\%$ of the values lie above.*

The calculation of a α percentile in N sample values is quite straight forward:

- Sort the N values in ascending order.
- The number located at $N\alpha$ (rounded integer) is the percentile.

Quantiles are the cut points dividing the range of a probability distribution into contiguous intervals with equal probabilities. We can visually compare if two distributions are similar by plotting the quantiles of two distributions against each other, known as **QQ plot**. In particular, if one random variable is the affine transformation of the other random variable, then the percentiles of the former is the affine transformation of the latter, as showed in the following.

Lemma 12.2.1 (linear relationship between percentiles from two distributions). *Let X and Y be two random variables with cdf F_X and F_Y. Let $p_X = F_X^{-1}(\alpha)$ and $p_Y = F_Y^{-1}(\alpha)$ for $\alpha \in [0,1]$. It follows that*

- *If $Y = aX + b$, then*
$$p_Y = ap_X + b$$
- *If $Y = \alpha X^\beta$, then*
$$p_{\ln Y} = \beta p_{\ln X} + \ln \alpha,$$

where $p_{\ln Y} = F^{-1}_{\ln Y}(\alpha), p_{\ln X} = F^{-1}_{\ln X}(\alpha),$

Proof. (1) We know that

$$\alpha = F_Y(p_Y) = F_X(p_X).$$

From scale-location transformation [Lemma 11.4.7], we have

$$p_X = (p_Y - b)/a.$$

(2) From $Y = \alpha X^\beta$, we have $\ln Y = \beta \ln X + \ln \alpha$. □

In Figure 12.2.4, we demonstrate the usage of QQ plot to compare different sample distributions as standard normal distribution. As expected, for samples drawn from a normal distribution or a shifted-scaled normal distribution, the QQ plot is approximately a perfect straight line; For Student's t distribution, which has a heavier tails than normal distribution, we see the deviation from the straight line at the two ends; For a log-normal distribution, which is vastly different from the normal distribution, we see strong deviations accordingly.

12.2.2.2 *Cornish-Fisher expansion*

With the knowledge of skewness and kurtosis, we can also approximate quantiles of a random variable using following **Cornish-Fisher expansion**.

Theorem 12.2.1 (Cornish-Fisher expansion). *Consider a distribution with mean μ and variance σ^2. Then its α quantile can be approximate by*

$$\mu + \sigma z^{cf}_\alpha$$

where

$$z^{cf}_\alpha = q_\alpha + \frac{(q^2_\alpha - 1)S(X)}{6} + \frac{(q^3_\alpha - 3q_\alpha)K(X)}{24} - \frac{(2q^3_\alpha - 5q_\alpha)S^2(X)}{36},$$

where $S(X)$ is skewness, $K(X)$ is kurtosis, z^{cf}_α is the Cornish-Fisher approximate quantile value for the confidence level α, and q_α is the quantile value for the standard normal distribution with confidence level α.

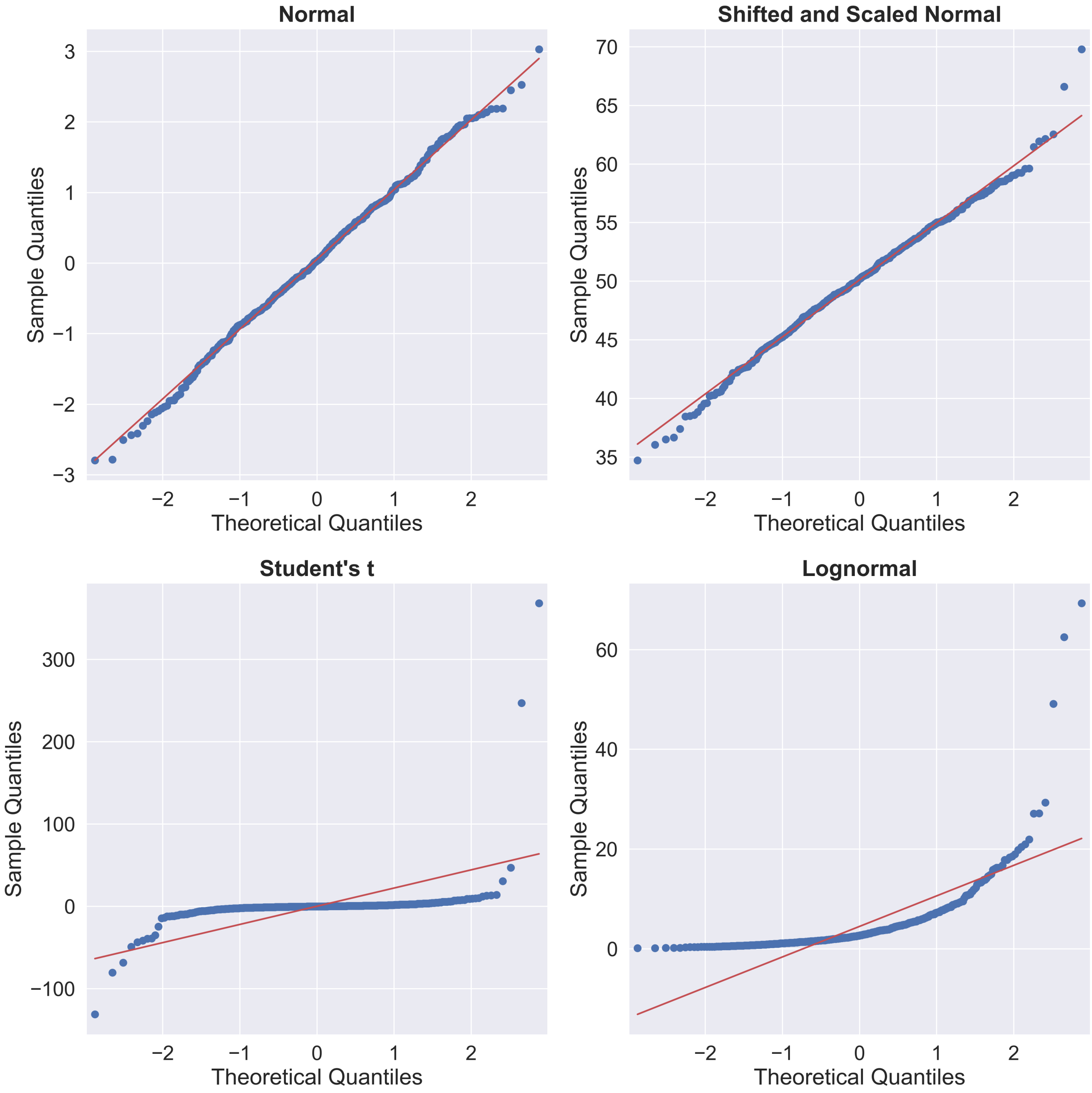

Figure 12.2.4: QQ plot of different sample distributions against standard normal distribution, including standard normal $N(0,1)$, shifted-scaled normal $N(50,5)$, Student's t with degree 1, and lognormal $LN(0,1)$. Red solid lines are the fitted linear lines.

12.3 Moment matching approximation methods

In our previous sections, we have covered the mostly widely used statistical distributions. In real-world applications, the statistical distribution of a random quantity is in general unknown. A practical modeling approach is to approximate the distribution by these discrete or continuous parametric distributions. Arguably the most common used, and perhaps simplest, approximate distribution is normal distribution, which is largely justified by the Central Limit Theorem [Theorem 11.11.3].

However, normal distributions also have their limitations: first, normal distributions do not have a bounded support, which is not ideal in modeling distribution with bounds (e.g., range greater than 0); second, normal distributions have zero skewness, making it a bad modeling choice for skewed distributions.

In this section, we consider the log-normal distribution family as an alternative to normal distribution. In subsection 12.1.10, we see that log-normal distributions have different extensions and mulitple parameters to control the boundedness and skewness; further, log-normal distributions have excellent analytical tractability.

Using log-normal distribution as an example, here we discuss the moment matching method to determine distribution parameters.

Lemma 12.3.1 (2 parameter log-normal approximation via moment matching). *Suppose we have a random variable X with range $X > 0$. Suppose X has moments given by*

$$E[X] = M_1, E[X^2] = M_2.$$

Let Y be a log-normal random variable defined by

$$Y = M_1 \exp(-\frac{1}{2}v^2 + vZ), Z \in N(0,1),$$

where

$$v^2 = \log(M_2/M_1^2)$$

Then Y has the same first two moments as X; that is

$$E[Y] = M_1, E[Y^2] = M_2.$$

Proof. Using moment generating function of Z, we know that

$$E[Y] = M_Z(v)M_1 \exp(-\frac{1}{2}v^2) = M_1.$$

and

$$E[Y^2] = M_Z(2v)M_1^2 \exp(-v^2) = \exp(v^2)M_1^2 = \frac{M_2}{M_1^2}M_1^2 = M_2.$$

□

Similarly, this moment matching method can be generalized to there-parameter log-normal distributions.

Lemma 12.3.2 (3 parameter shifted lognormal approximation via moment matching). *Suppose we have a random variable X having moments given by*

$$E[X] = M_1, E[X^2] = M_2, E[X^3] = M_3.$$

Let Y be a shifted log-normal random variable with parameter $SLN(\mu, \sigma^2, \tau)$ such that

$$E[Y] = \tau + \exp(\mu + \frac{1}{2}\sigma^2),$$

$$E[Y^2] = \tau^2 + 2\tau\exp(\mu + \frac{1}{2}\sigma^2) + \exp(2\mu + 2\sigma^2),$$

$$E[Y^3] = \tau^3 + 3\tau^2\exp(\mu + \frac{1}{2}\sigma^2) + 3\tau\exp(2\mu + 2\sigma^2) + \exp(3\mu + \frac{9}{2}\sigma^2).$$

If we can find (μ, σ, τ) such that

$$E[X] = E[Y], E[X^2] = E[Y^2], E[X^3] = E[Y^3],$$

then X and Y have matched moments.

Proof. For moments of Y, see Lemma 12.1.20. □

Based on the target distribution's location and skewness, we can choose the type of lognormal distribution we want to use. The table below is a good summary[4].

skewness	$\gamma > 0$	$\gamma > 0$	$\gamma < 0$	$\gamma < 0$
location	$\tau \geq 0$	$\tau < 0$	$\tau \geq 0$	$\tau < 0$
choice of approximation	regular	shifted	negative	negative shifted

12.4 Gaussian quadratic forms

12.4.0.1 *Quadratic forms and chi-square distribution*

Let $X = (X_1, X_2, ..., X_n)^T$ be a random vector, we called

$$Q = X^T \Sigma X, \Sigma \in \mathbb{R}^{n\times n},$$

a **quadratic form of random vector** X. Note that Q is also a random variable. We are particularly interested in the case where X follows a multivariate Gaussian distribution. In this case, Q is known as **Gaussian quadratic form**.

Gaussian quadratic forms are widely used in characterizing residual or error distributions in linear regression applications. We start with its basic property.

Lemma 12.4.1. *Let X be a m-dimensional random vector with multivariate Gaussian distribution, i.e., $X \sim N(\mu, \Sigma)$. It follows that*

- $$\Sigma^{1/2}(x-\mu) \sim N(0, I).$$
- $$(x-\mu)\Sigma^{-1}(x-\mu) \sim \chi^2(m).$$

Proof. (1) Directly from affine transformation property of multivariate Gaussian random variable [Theorem 14.1.1]. (2) Use the definition that sum of iid normal random variable square is chi-square random variable. □

Theorem 12.4.1 (chi-square orthogonal decomposition). *Let $X_1, X_2, ..., X_n$ be independent standard normal variables such that*

$$\sum_{i=1}^n X_i^2 \sim \chi^2(n).$$

Denote $X = (X_1, ..., X_n)^T$. If there exists an orthogonal projector $P \in \mathbb{R}^{n\times n}$ such that $Y = PX, Z = (I-P)X$, then

- *$Y \sim MN(0, P), Z \sim MN(0, I-P)$, and Y, Z are independent of each other.*
- *$Y^T Y \sim \chi^2(r), r = rank(P)$; or equivalently, the quadratic form $Q = X^T P X \sim \chi^2(r)$.*
- *$Z^T Z \sim \chi^2(n-r)$;or equivalently, the quadratic form $Q = X^T(I-P)X \sim \chi^2(n-r)$*

In summary, for a quadratic form $Q = X^T\Sigma X$, if Σ is idempotent and symmetric, then $Q \sim \chi^2(rank(\Sigma))$.

Proof. (1) From affine transform of multivariate normal [Theorem 14.1.1],

$$Y \sim MN(0, P\Sigma_X P^T) = MN(0, P^2) = MN(0, P).$$

To show independence, we have $E[YZ^T] = E[PXX^T(I-P)^T] = E[P(I-P)] = 0$.

(2) Let U be the eigen-decomposition of P such that $P = UU^T$. Let $Z = U^T X, Z \in \mathbb{R}^r, Z \sim MN(0, I_r)$. Let V be the eigen-decomposition of $I - P$ such that $I - P = VV^T$. Let $W = V^T X, W \in \mathbb{R}^{n-r}, W \sim MN(0, I_{n-r})$. We want to show that the characteristic function of the random quantity $Y^T Y$ is the same as the characteristic function of $\chi^2(r)$.

$$\begin{aligned}
&E[\exp(itY^TY)] \\
=&\frac{1}{(2\pi)^{n/2}}\int\int\cdots\int \exp(it(U^TX)^T(U^TX)\exp(-\frac{1}{2}X^T(I-P+P)X))dx_1dx_2\cdots dx_n \\
=&\frac{1}{(2\pi)^{n/2}}\int\int\cdots\int \exp(itZ^TZ)\exp(-\frac{1}{2}(Z^TZ+W^TW))dz_1\cdots dz_r dw_{r+1}\cdots dw_n \\
=&\frac{1}{(2\pi)^{r/2}}\int\int\cdots\int \exp(itZ^TZ)\exp(-\frac{1}{2}(Z^TZ))dz_1\cdots dz_r
\end{aligned}$$

where we change the integral variable such that

$$[dz_1\cdots dz_r dw_{r+1}\cdots dw_n]^T = [U\ V](dx_1 dx_2\cdots dx_n)^T$$

. The last line is the characteristic function of $\chi^2(r)$. (3) similar to (2). □

Lemma 12.4.2 (moment generating functions for Gaussian quadratic forms). *[3, p. 523] Let $X = (X_1, X_2, ..., X_n)^T$ where $X_1, X_2, ..., X_n$ are iid $N(0,1)$. Consider the quadratic form $Q = X^T A X$ for a symmetric matrix A of rank $r \leq n$. It follows that*

- *Q has the moment generating function $M(t) = \prod_{i=1}^r (1-2t\lambda_i)^{-1/2} = |I - 2tA|^{-1/2}$, where $\lambda_1, \lambda_2, ..., \lambda_r$ are the nonzero eigenvalues of A, $|t| < \frac{1}{\max|\lambda_i|}$.*
- *If A is an orthogonal projector such that $\lambda_1 = \lambda_2 = \cdots = \lambda_r = 1$, then*

$$M(t) = M_{\chi^2(r)}.$$

Proof. (1)Let the eigen-decomposition of A be

$$A = U\Lambda U^T, U \in \mathbb{R}^{n\times r}, \Lambda \in \mathbb{R}^{r\times r}.$$

Then

$$Q = X^T AX = X^T U\Lambda U^T X = X^T(\sum_{i=1}^r \lambda_i u_i u_i^T)X = \sum_{i=1}^r \lambda_i (u_i^T X)^T.$$

Let $Y_i = u_i^T X, i = 1, 2, ..., r$. It can be shown that $Y_i \sim N(0,1), E[Y_iY_j] = u_i^T E[XX^T]u_j^T = \delta_{ij}$; that is $Y_1, Y_2, ..., Y_r \sim MN(0, I_r)$. Therefore, $Y_i^2 \sim \chi^2(1)$.

The moment generating function is given by

$$\begin{aligned} M(t) &= E[\exp(tQ)] \\ &= E[\exp(t\sum_{i=1}^r \lambda_i Y_i^2)] \\ &= \prod_{i=1}^r E[\exp(t\lambda_i Y_i^2)] \\ &= \prod_{i=1}^r M_{\chi^2(1)}(\lambda_i t) \\ &= \prod_{i=1}^r (1 - 2\lambda_i t)^{-1/2} \end{aligned}$$

where we use the moment generating function of $\chi^2(1)$ from Lemma 12.1.31. (2) straight forward. □

Lemma 12.4.3 (independence of quadratic forms). *[3, p. 528] Let $X = (X_1, X_2, ..., X_n)$ be a random vector where $X_1, X_2, ..., X_n$ are iid $N(0,1)$. For real symmetric matrices $A, B \in \mathbb{R}^{n\times n}$, let $Q_1 = X^T AX$ and $Q_2 = X^T BX$. Then Q_1 and Q_2 are independent if and only if $AB = 0$.*

Proof. Let $rank(A) = r, rank(B) = s$. Let the eigendecomposition of A, B be such that

$$A = \sum_{i=1}^r \lambda_i u_i u_i^T, A = \sum_{i=1}^s \beta_i v_i v_i^T.$$

If $AB = 0$, then $u_1, ..., u_r, v_1, ..., v_r$ will be orthogonal to each other. Then

$$Q_1 + Q_2 = \sum_{i=1}^{r+s} \lambda_i u_i u_i^T,$$

where $u_{r+i} = v_i, \lambda_{r+i} = \beta_i$.

It is easy to see that [Lemma 12.4.2]

$$M_{Q_1,Q_2}(t_1, t_2) = M_{Q_1}(t_1)M_{Q_2}(t_2).$$

Then from independence-from-mgf [Lemma 11.6.3], we can prove Q_1 and Q_2 are independent. □

12.4.0.2 *Applications*

The first application is to prove Student's theorem, which specifies the distribution of sample variance.

Theorem 12.4.2 (Student's Theorem). *Let $X_1, X_2, ..., X_n$ be iid random variables each having a normal distribution with mean μ and variance σ^2. Define random variables as:*

$$\bar{X} = \frac{1}{n}\sum_{i=0}^{n} X_i, S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

1. *$\bar{X}$ has a $N(\mu, \sigma^2/n)$ distribution*
2. *$\bar{X}$ and S^2 are independent.*
3. *$(n-1)S^2/\sigma^2$ has a $\chi^2(n-1)$ distribution*
4. *The random variable*

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has t-distribution with $n-1$ degrees of freedom.

Proof. (1)From Lemma 12.1.9. (2) We can prove $\overline{X}$ and the random vector $Y = (X_1 - \overline{X}, ..., X_n - \overline{X})$ are independent. Note that

$$\overline{X} = \frac{1}{n}\mathbf{1}^T X, Y = (I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)X,$$

and hence $\overline{X}$ and Y are both normal.

$$\begin{aligned} Cov(\overline{X}, Y) &= X^T(\frac{1}{n}\mathbf{1}^T(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T))X \\ &= X^T\frac{1}{n}(\mathbf{1}^T - \frac{1}{n}\mathbf{1}^T\mathbf{1}\mathbf{1}^T)X \\ &= X^T\frac{1}{n}(\mathbf{1}^T - \mathbf{1}^T))X = 0 \end{aligned}$$

where we use the fact that $\mathbf{1}^T\mathbf{1} = n$.

Then $S^2 = \frac{1}{n-1}Y^TY$ will be independent of $\overline{X}$ because S^2 is a function of Y [Lemma 11.3.2]. (3) See reference and Corollary 12.4.3.1. (4) From the definition of the t distribution, we have

$$Y = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is the $N(0,1)$. $W = (n-1)S^2/\sigma^2$ has a $\chi^2(n-1)$ distribution. Then

$$\frac{Y}{\sqrt{W/n-1}}$$

has $t(n-1)$ distribution. □

The second application is to prove Cochran's theorem.

Lemma 12.4.4. *Let $X_1, X_2, ..., X_n$ be real numbers. Suppose that $\sum_{i=1}^n X_i^2$ can be decomposed into a sum of positive semi-definite quadratic forms, that is*

$$\sum_{i=1}^n X_i^2 = Q_1 + ... + Q_k$$

where $Q_i = X^T A_i X$ with $rank(A_i) = r_i$. If $\sum_{i=1}^k r_i = n$, then there exists an orthonormal matrix C such that $X = CY$ and

$$\begin{aligned} Q_1 &= Y_1^2 + ... + Y_{r_1}^2 \\ Q_2 &= Y_{r_1+1}^2 + ... + Y_{r_1+r_2}^2 \\ &\cdots \end{aligned}$$

Proof. (informal)Note that when we decompose a matrix, its sum of rank of the decomposed matrix will increase [Theorem 4.4.1], i.e.,

$$rank(A+B) \leq rank(A) + rank(B)$$

and the equality only holds when $\mathcal{R}(A) \cap \mathcal{R}(B) = \emptyset$.
Since in our case $rank(\sum A_i) = \sum rank(A_i)$, then we must have $\mathbb{R}^n = \mathcal{R}(A_1) \oplus \mathcal{R}(A_2) ... \oplus \mathcal{R}(A_k)$. Take the basis of each $\mathcal{R}(A_i)$ and make it to be orthonormal matrix C. Then Y_i are just the orthonormal projection to subspace $\mathcal{R}(A_j)$. An accessible proof is at Theorem 12.4.1 □

Theorem 12.4.3 (Cochran's theorem). *Let $X_1, X_2, ..., X_n$ be iid $N(0, \sigma^2)$ random variables. Suppose that $\sum_{i=1}^n X_i^2$ can be decomposed into a sum of positive semi-definite quadratic forms, that is*

$$\sum_{i=1}^n X_i^2 = Q_1 + ... + Q_k$$

where $Q_i = X^T A_i X$ with $rank(A_i) = r_i$. If $\sum_{i=1}^k r_i = n$, then there exists an orthonormal matrix C such that $X = CY, Y = C^T X$ ($Y_1, Y_2, ..., Y_n$ are independent random variables with $N(0, \sigma^2)$)and

$$Q_1 = Y_1^2 + ... + Y_{r_1}^2$$
$$Q_2 = Y_{r_1+1}^2 + ... + Y_{r_1+r_2}^2$$
$$\cdots$$

Moreover, we have

- $Q_1, Q_2, ..., Q_k$ *are independent*
- $Q_i \sim \sigma^2 \chi^2(r_i)$.

Proof. (1) Use above lemma. Note that $Y_1, Y_2, ..., Y_n$ are still independent normal because of Lemma 14.1.2. (2) Since Q_i and Q_j have non-overlapping Y_is, they are independent to each other. (3) From properties of χ^2 distribution [Lemma 12.1.31]. □

Corollary 12.4.3.1 (distribution of sample variance). *Let $Y_1, ..., Y_n$ be iid random variable with $N(\mu, \sigma^2)$, then*

$$\sum_{i=1}^n (Y_i - \overline{Y})^2 \sim \sigma^2 \chi^2(n-1)$$

and

$$\sum_{i=1}^n (Y_i - \mu)^2 / n \sim \sigma^2 \chi^2(1)$$

Proof.

$$\sum_{i=1}^n (Y_i - \mu)^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2 + \sum_{i=1}^n (Y_i - \mu)^2 / n$$

And

$$(Y - \mu)^T (Y - \mu) = (Y - \mu)^T (I - \frac{1}{n} J)(Y - \mu) + (Y^T - \mu)^T (\frac{1}{n} J)(Y - \mu)$$

and $rank(\frac{1}{n}J)$ has rank 1 and $rank(I - \frac{1}{n}J) = n - 1$. □

Remark 12.4.1.

- The matrix $\frac{1}{n}J$ has rank 1 is because it only has one linearly independent column.
- The matrix $I - \frac{1}{n}J$ is because $rank(I - \frac{1}{n}J) \geq rank(I) - rank(\frac{1}{n}J) = n - 1$ [Theorem 4.4.1]. Also $I - \frac{1}{n}J$ has eigenvector 1 associated with eigenvalue 0. Therefore, $rank(I - \frac{1}{n}J) < n$. In summary, we have $rank(I - \frac{1}{n}J) = n - 1$.

- The matrix $I - \frac{1}{n}J$ has rank $n-1$ because it is orthogonal projector($P^T = P, P^2 = P$) and $rank(I - \frac{1}{n}J) = Tr(I - \frac{1}{n}J) = n-1$. [Theorem 4.5.7]

12.5 Notes on bibliography

For an extensive discussion on statistical distributions, see [1][2].

BIBLIOGRAPHY

1. Forbes, C., Evans, M., Hastings, N. & Peacock, B. *Statistical distributions* (John Wiley & Sons, 2011).
2. Krishnamoorthy, K. *Handbook of statistical distributions with applications* (CRC Press, 2016).
3. Hogg, R. V., McKean, J. & Craig, A. T. Introduction to Mathematical Statistics, 7 ed (2012).
4. Borovkova, S., Permana, F. J. & Weide, H. V. A closed form approach to the valuation and hedging of basket and spread options. *Journal of Derivatives* **14,** 8 (2007).
5. Brzezniak, Z. & Zastawniak, T. *Basic stochastic processes: a course through exercises* (Springer Science & Business Media, 1999).
6. Murphy, K. P. *Machine learning: a probabilistic perspective* (MIT press, 2012).
7. Casella, G. & Berger, R. L. *Statistical inference* (Duxbury Pacific Grove, CA, 2002).
8. Ruppert, D. *Statistics and data analysis for financial engineering, 2ed* (Springer, 2015).