Chapter 1: Reviews of Probability & Sampling Distributions

1 Random Variables

Recall the "fundamental principle": Data is a realization of a random process. Throughout this course, we will model the data using **random variables**. The goal of this lecture is to review, with a statistical focus, relevant concepts concerning random variables and their distributions.

Definition 1.1. *A probability space is the triplet $(\Omega, \mathcal{F}, P)$, where Ω is the sample space, $\mathcal{F}$ is the collection of events (σ-algebra), and P is a probability measure defined over $\mathcal{F}$, with $P(\Omega) = 1$.*

Definition 1.2. *(Measurable function). Let f be a function from a measurable space $(\Omega, \mathcal{F})$ into the real numbers. We say that the function is measurable if for each Borel set $B \in \mathcal{B}$, the set $\{\omega \in \Omega : f(\omega) \in \mathcal{B}\} \in \mathcal{F}$.*

Definition 1.3. *(random variable). A random variable X is a measurable function from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into the real numbers $\mathbb{R}$ (or a subset). Example: Number of heads turning up (X) when tossing a coin thrice.*

A **discrete** random variable X can take a finite or countably infinite number of possible values. We use discrete random variables to model categorical data (for example, which presidential candidate a voter supports) and count data (for example, how many cups of coffee a graduate student drinks in a day). The distribution of X is specified by its **probability mass function** (**PMF**):

$$f_X(x) = \mathbb{P}[X = x].$$

Then for any set A of values that X can take,

$$\mathbb{P}[X \in A] = \sum_{x \in A} f_X(x).$$

A **continuous** random variable X takes values in $\mathbb{R}$ and models continuous real-valued data (for example, the height of a person). For any single value $x \in \mathbb{R}, \mathbb{P}[X = x] = 0$.

Instead, the distribution of X is specified by its **probability density function** (**PDF**) $f_X(x)$, which satisfies for any set $A \subseteq \mathbb{R}$

$$\mathbb{P}[X \in A] = \int_A f_X(x)dx.$$

In both cases, when it is clear which random variable is being referred to, we will simply write $f(x)$ for $f_X(x)$.

For any random variable X and real-valued function g, the **expectation** or mean of $g(X)$ is its "average value". If X is discrete with PMF $f_X(x)$, then

$$\mathbb{E}[g(X)] = \sum_x g(x)f_X(x)$$

where the sum is over all possible values of X. If X is continuous with PDF $f_X(x)$, then

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)f_X(x)dx$$

The expectation is *linear*: For any random variables $X_1, \ldots, X_n$ (not necessarily independent) and any $c \in \mathbb{R}$

$$\mathbb{E}\left[X_1 + \ldots + X_n\right] = \mathbb{E}\left[X_1\right] + \ldots + \mathbb{E}\left[X_n\right], \quad \mathbb{E}[cX] = c\mathbb{E}[X]$$

If $X_1, \ldots, X_n$ are independent, then

$$\mathbb{E}\left[X_1 \ldots X_n\right] = \mathbb{E}\left[X_1\right] \ldots \mathbb{E}\left[X_n\right]$$

The **variance** of X is defined by the two equivalent expressions

$$\operatorname{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$$

For any $c \in \mathbb{R}, \operatorname{Var}[cX] = c^2 \operatorname{Var}[X]$. If $X_1, \ldots, X_n$ are independent, then

$$\operatorname{Var}\left[X_1 + \ldots + X_n\right] = \operatorname{Var}\left[X_1\right] + \ldots + \operatorname{Var}\left[X_n\right]$$

If $X_1, \ldots, X_n$ are not independent, then this is not true $-\operatorname{Var}\left[X_1 + \ldots + X_n\right]$ will depend on the covariance between each pair of variables. The **standard deviation** of X is $\sqrt{\operatorname{Var}[X]}$.

The distribution of X can also be specified by its **cumulative distribution function** (**CDF**)$F_X(x) = \mathbb{P}[X \leq x]$. In the discrete and continuous cases, respectively, this is given by

$$F_X(x) = \sum_{y: y \leq x} f_X(y), \quad F_X(x) = \int_{-\infty}^{x} f_X(y)dy.$$

In the continuous case, the fundamental theorem of calculus implies

$$f_X(x) = \frac{d}{dx} F_X(x).$$

By definition, F_X is monotonically increasing: $F_X(x) \leq F_X(y)$ if $x < y$. If F_X is continuous and strictly increasing, meaning $F_X(x) < F_X(y)$ for all $x < y$, then F_X has an inverse function $F_X^{-1} : (0,1) \to \mathbb{R}$ called the **quantile function**: For any $t \in (0,1), F_X^{-1}(t)$ is the t^{th} quantile of the distribution of X. I.e. the probability that X is less than this value is exactly t.

1.1 Conditional Probability

Definition 1.4. *Two events A and B are independent if and only if the probability of their intersection equals the product of their individual probabilities, that is*

$$P(A \cap B) = P(A)P(B).$$

Definition 1.5. *Given two events A and B, with $P(B) > 0$, the conditional probability of A given B, denoted $P(A \mid B)$, is defined by the relation*

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)}.$$

In connection with these definitions, the following result holds. Let $\{C_j : j = 1, \ldots, n\}$ be a partition of Ω, this is, $\Omega = \cup_{j=1}^{n} C_j$ and $C_i \cap C_k = \emptyset$ for $i \neq k$. Let also A be an event. The *Law of Total Probability* states that

$$P(A) = \sum_{j=1}^{n} P\left(A \mid C_j\right) P\left(C_j\right).$$

Definition 1.6. *Let X and Y be discrete, jointly distributed random variables. For $P(X = x\ ) > 0$ the conditional probability function of Y given that $X = x$ is*

$$p_{Y|X=x}(y) = P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)},$$

and the conditional cumulative distribution function of Y given that $X = x$ is

$$F_{Y|X=x}(y) = P(Y \leq y \mid X = x) = \sum_{z \leq y} p_{Y|X=x}(z).$$

Definition 1.7. *Let X and Y have a joint continuous distribution. For $f_X(x) > 0$, the conditional density function of Y given that $X = x$ is*

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)},$$

where $f_{X,Y}$ is the joint probability density function of X and Y, and f_X is the marginal probability density function of X. The conditional cumulative distribution function of Y given that $X = x$ is

$$F_{Y|X=x}(y) = \int_{-\infty}^{y} f_{Y|X=x}(z)dz.$$

Remark 1. *The law of total probability. Let X and Y have a joint continuous distribution. Suppose that $f_X(x) > 0$, and let $f_{Y|X=x}(y)$ be the conditional density function of Y given that $X = x$ The law of total probability states that*

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X=x}(y) f_X(x)dx.$$

1.2 Families of Distributions

The distributions presented here are *parametric distributions*. A parametric distribution is a distribution that has one or more *parameters* (also known as *statistical parameters*). Finally, a parameter (or statistical parameter) is a numerical characteristic that indexes a family of probability distributions.

Discrete distributions: A random variable X is said to be discrete if the range of X is countable. Some examples of discrete variables and their corresponding *probability mass functions* are presented below.

Example 1.1. *The Bernoulli distribution. The simplest example of a discrete random variable corresponds to the case where the range of X is the set $\{0, 1\}$. The distribution of X is:*

$$\mathbb{P}(X = x) = \begin{cases} p, & \text{for } x = 1 \\ 1 - p, & \text{for } x = 0. \end{cases}$$

This distribution is known as the ***Bernoulli*** *distribution, and it is often denoted as $X \sim \text{Bernoulli}(p)$. Often, the event $\{x = 1\}$ is called a "success", and the event $\{x = 0\}$ is called a "failure". Thus, the parameter p is known as "the probability of success". It follows that:*

$$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p,$$
$$\text{Var}[X] = (1 - p)^2 \cdot p + (0 - p)^2(1 - p) = p(1 - p).$$

A more particular example is the case where X is the outcome observed from tossing a fair coin once. In this case $\mathbb{P}(X = \text{heads}) = \mathbb{P}(X = \text{tails}) = \frac{1}{2}$.

Bernoulli random variables are used in many contexts, and they are often referred to as *Bernoulli trials*. A Bernoulli trial is an experiment with two, and only two, possible outcomes. **Parameters:** $0 \leq p \leq 1$.

Example 1.2. *The Binomial distribution. A Binomial random variable X is the total number of successes in n Bernoulli trials. Consequently, the range of X is the set $\{0, 1, 2, \ldots, n\}$. The probability of each outcome is given by:*

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the Binomial coefficient (also known as combination). If a random variable X has Binomial distribution, it is denoted as $X \sim \text{Binomial}(n, p)$, where n is the number of trials and p is the probability of success.

The mean and variance of a Binomial random variable are $\mathbb{E}[X] = np$ and $\text{Var}[X] = np(1-p)$.

A more particular example is the case where X is the number of heads in $n = 10$ fair coin tosses. Then, for $x = 0, 1, \ldots, 10$:

$$\mathbb{P}(X = x) = \binom{10}{x} 0.5^x (0.5)^{n-x} = \binom{10}{x} (0.5)^n.$$

Parameters: *$n \in \mathbb{Z}_+$ and $0 \leq p \leq 1$.*

Example 1.3. *The Poisson distribution. A random variable X has a Poisson distribution if it takes values in the non-negative integers and its distribution is given by:*

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

It is possible to show that $\mathbb{E}[X] = \text{Var}[X] = \lambda$. ***Parameters:*** *$\lambda > 0$.*

Example 1.4. *The Multinomial distribution. The Multinomial distribution is a generalization of the binomial distribution. For n independent trials, each of which produces and outcome (success) in one of $k \geq 2$ categories, where each category has a given fixed success probability $\theta_i, i = 1, \ldots, k$, the Multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories. Thus, the pmf is*

$$p\left(x_1, \ldots, x_k; \theta_1, \ldots, \theta_k\right) = \frac{n!}{x_1! \cdots x_k!} \theta_1^{x_1} \cdots \theta_k^{x_k},$$

*where **parameters** $\theta_i \geq 0$ for $i = 1, \ldots, k$ and $\sum_{i=1}^k \theta_i = 1$.*

There are many other discrete probability distributions of practical interest, for example, the hypergeometric distribution, and the discrete uniform distribution, among others.

Continuous distributions: A random variable X is said to be continuous if its range is uncountable and its distribution is continuous everywhere. Moreover, a random variable X is said to be *absolutely continuous* if there exists a nonnegative function f such that for any open set B:

$$\mathbb{P}(X \in B) = \int_B f(x)dx$$

The function f is called the *probability density function* of X. This definition can be used to link the probability density function and the cumulative distribution F as follows:

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

Definition 1.8. *Suppose that $f : \mathcal{D} \to \mathbb{R}_+$is a density function. Then, the support of $f, \operatorname{supp}(f)$, is the set of points where f is positive:*

$$\operatorname{supp}(f) = \{x \in \mathcal{D} : f(x) > 0\}.$$

Definition 1.9. *A random variable X, with* cdf *F has the characteristic function*

$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right] = \int_{-\infty}^{\infty} e^{itx} dF(x).$$

If the pdf f exists, then

$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right] = \int_{-\infty}^{\infty} e^{itx} f(x)dx.$$

Another feature of continuous distributions is that $\mathbb{P}(X = x) = 0$, for all x in the range of X. Some examples of continuous distributions are presented below.

Example 1.5. *The Beta Distribution. The probability density function of a Beta random variable $X \in (0, 1)$ is:*

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

where $a, b > 0$ are shape parameters, $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the Beta function. It is important to distinguish the Beta distribution from the Beta special function.

Definition 1.10. *The Beta function and Gamma function are special functions defined as follows*

$$\mathrm{B}(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$$

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$$

The mean of the Beta distribution is $\mathbb{E}[X] = \frac{a}{a+b}$*, and the mode (maximum) is* $\mathrm{Mode}(X) = \frac{a-1}{a+b-2}$ *for* $b > 1$*. The variance is* $\mathrm{Var}[X] = \frac{ab}{(a+b)^2(a+b+1)}$*.* ***Parameters:*** $a > 0$ *and* $b > 0$*.*

The uniform distribution is a special case of the Beta distribution for the case $a = b = 1$.

Example 1.6. *Normal or Gaussian Distribution. The probability density function of a Gaussian random variable* $X \in \mathbb{R}$ *is:*

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

where $-\infty < \mu < \infty$ *and* $\sigma > 0$ *are parameters of this density function. In fact,* $\mathbb{E}[X] = \mu$ *and* $\mathrm{Var}[X] = \sigma^2$*. If a random variable* X *has normal distribution with mean* μ *and variance* σ^2*, we will denote it* $X \sim N(\mu,\sigma^2)$*. This is one of the most popular distributions in applications and it appears in a number of statistical and probability models.* ***Parameters:*** $-\infty < \mu < \infty$ *and* $\sigma > 0$*.*

The Normal distribution has many interesting properties. One of them is that it is closed under summation, meaning that the sum of normal random variables is normally distributed. That is, let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution $N(\mu,\sigma^2)$, and $Y = \sum_{j=1}^n X_j$, $Z = \frac{1}{n}\sum_{j=1}^n X_j$. Then, $Y \sim N(n\mu, n\sigma^2), Z \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Example 1.7. *The Logistic distribution. The probability density function of a logistic random variable* $X \in \mathbb{R}$ *is:*

$$f(x;\mu,\sigma) = \frac{\exp\left\{-\frac{x-\mu}{\sigma}\right\}}{\sigma\left(1+\exp\left\{-\frac{x-\mu}{\sigma}\right\}\right)^2},$$

where $-\infty < \mu < \infty$ *and* $\sigma > 0$ *are location and scale parameters, respectively. The mean and variance of* X *are given by* $\mathbb{E}[X] = \mu$ *and* $\mathrm{Var}[X] = \frac{\pi^2\sigma^2}{3}$*. The cdf of a logistic random variable is*

$$F(x;\mu,\sigma) = \frac{\exp\left\{\frac{x-\mu}{\sigma}\right\}}{1+\exp\left\{\frac{x-\mu}{\sigma}\right\}}.$$

Parameters: *$-\infty < \mu < \infty$ and $\sigma > 0$. This distribution is very popular in practice as well. In particular, the so-called "logistic regression model" is based on this distribution, as well as some Machine Learning algorithms (logistic sigmoidal activation function in neural networks).*

Example 1.8. *The Exponential distribution. The probability density function of an Exponential random variable $X > 0$ is:*

$$f(x;\lambda) = \lambda \exp\{-\lambda x\}$$

where $\lambda > 0$ is a rate parameter. The mean and variance are given by $\mathbb{E}[X] = \frac{1}{\lambda}$ and $\mathrm{Var}[X] = \frac{1}{\lambda^2}$. ***Parameters:*** *$\lambda > 0$. This distribution is widely used in engineering and the analysis of survival times.*

Example 1.9. *The Gamma distribution. The probability density function of a Gamma random variable $X > 0$ is:*

$$f(x;\kappa,\theta) = \frac{1}{\Gamma(\kappa)\theta^\kappa} x^{\kappa-1} \exp\left\{-\frac{x}{\theta}\right\},$$

where $\kappa > 0$ is a shape parameter, $\theta > 0$ is a scale parameter, and $\Gamma(z) = \int_0^\infty s^{z-1}e^{-s}ds$ is the Gamma function (for positive integers n, $\Gamma(n) = (n-1)!$). ***Parameters:*** *$\kappa > 0$ and $\theta > 0$. The mean and variance of X are given by $\mathbb{E}[X] = \kappa\theta$ and $\mathrm{Var}[X] = \kappa\theta^2$. This distribution is widely used in engineering and the analysis of survival times.*

1.3 Joint distributions

If random variables $X_1, \ldots, X_k$ are independent, then their distribution may be specified by specifying the individual distribution of each variable. If they are not independent, then we need to specify their **joint distribution**. In the discrete case, the joint distribution is specified by a **joint PMF**

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \mathbb{P}\left[X_1 = x_1, \ldots, X_k = x_k\right].$$

In the continuous case, it is specified by a **joint PDF** $f_{X_1,\ldots,X_k}(x_1,\ldots,x_k)$, which satisfies for any set $A \subseteq \mathbb{R}^k$,

$$\mathbb{P}\left[(X_1,\ldots,X_k) \in A\right] = \int_A f_{X_1,\ldots,X_k}(x_1,\ldots,x_k)\, dx_1 \ldots dx_k.$$

When it is clear which random variables are being referred to, we will simply write $f(x_1,\ldots,x_k)$ for $f_{X_1,\ldots,X_k}(x_1,\ldots,x_k)$.

Example 1.10. $(X_1, \ldots, X_k)$ *have a multinomial distribution,*

$$(X_1, \ldots, X_k) \sim \text{Multinomial}\left(n, (p_1, \ldots, p_k)\right),$$

if these random variables take nonnegative integer values summing to n, with joint PMF

$$f\left(x_1, \ldots, x_k\right) = \begin{pmatrix} n \\ x_1, \ldots, x_n \end{pmatrix} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}.$$

Here, $p_1, \ldots, p_k$ are values in $[0, 1]$ that satisfy $p_1 + \ldots + p_k = 1$ (representing the probabilities of k different mutually exclusive outcomes), and $\begin{pmatrix} n \\ x_1, \ldots, x_n \end{pmatrix}$ is the multinomial coefficient $\begin{pmatrix} n \\ x_1, \ldots, x_n \end{pmatrix} = \frac{n!}{x_1! x_2! \ldots x_n!}$. (It is understood that the above formula is only for $x_1, \ldots, x_k \geq 0$ such that $x_1 + \ldots + x_k = n$; otherwise $f\left(x_1, \ldots, x_k\right) = 0$.) $X_1, \ldots, X_k$ describe the number of samples belonging to each of k different outcomes, if there are n total samples each independently belonging to outcomes $1, \ldots, k$ with probabilities $p_1, \ldots, p_k$. For example, if I roll a standard six-sided die 100 times and let $X_1, \ldots, X_6$ denote the numbers of 1's to 6's obtained, then $(X_1, \ldots, X_6) \sim \text{Multinomial}\left(100, \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)\right)$.

A second example of a joint distribution is the Multivariate Normal distribution (to be discussed later).

A third example is: The Dirichlet distribution is a distribution of continuous random variables relevant to the Multinomial distribution. Sampling from a Dirichlet distribution leads to a random vector with length k and each element of this vector is non-negative and the summation of elements is 1, meaning that it generates a random probability vector.

Example 1.11. *The Dirichlet distribution is a multivariate distribution over the simplex $\sum_{i=1}^{k} x_i = 1$ and $x_i \geq 0$. Its probability density function is*

$$p\left(x_1, \cdots, x_k; \alpha_1, \cdots, \alpha_k\right) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1},$$

where $B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha)}{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}$ with $\Gamma(a)$ being the Gamma function and $\alpha = (\alpha_1, \cdots, \alpha_K)$ are the parameters of this distribution.

You can view it as a generalization of the Beta distribution. For $Z = (Z_1, \cdots, Z_k) \sim \text{Dirch}\left(\alpha_1, \cdots, \alpha_k\right)$, $\mathbb{E}\left(Z_i\right) = \frac{\alpha_i}{\sum_{j=1}^{k} \alpha_j}$ and the mode of Z_i is $\frac{\alpha_i - 1}{\sum_{j=1}^{k} \alpha_j - k}$ so each parameter α_i determines the relative importance of category (state) i. Because it is a distribution putting probability over K categories, Dirichlet distribution is very popular in social sciences and linguistics analysis.

The Dirichlet distribution is often used as a prior distribution for the multinomial parameter $p_1, \cdots, p_k$ in Bayesian inference.

The **covariance** between two random variables X and Y is defined by the two equivalent expressions

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

So $\operatorname{Cov}[X, X] = \operatorname{Var}[X]$, and $\operatorname{Cov}[X, Y] = 0$ if X and Y are independent. The covariance is *bilinear*: For any constants $a_1, \ldots, a_k, b_1, \ldots, b_m \in \mathbb{R}$ and any random variables $X_1, \ldots, X_k$ and $Y_1, \ldots, Y_m$ (not necessarily independent),

$$\operatorname{Cov}\left[a_1X_1 + \ldots + a_kX_k, b_1Y_1 + \ldots + b_mY_m\right] = \sum_{i=1}^{k}\sum_{j=1}^{m} a_ib_j \operatorname{Cov}\left[X_i, Y_j\right].$$

The **correlation** between X and Y is their covariance normalized by the product of their standard deviations:

$$\operatorname{corr}(X, Y) = \frac{\operatorname{Cov}[X, Y]}{\sqrt{\operatorname{Var}[X]}\sqrt{\operatorname{Var}[Y]}}.$$

For any $a, b > 0$, we have $\operatorname{Cov}[aX, bY] = ab \operatorname{Cov}[X, Y]$. On the other hand, the correlation is invariant to rescaling: $\operatorname{corr}(aX, bY) = \operatorname{corr}(X, Y)$, and satisfies always $-1 \leq \operatorname{corr}(X, Y) \leq 1$.

1.4 Moment generating functions

A tool that will be particularly useful for us is the moment generating function (MGF) of a random variable X. This is a function of a single argument $t \in \mathbb{R}$, defined as

$$M_X(t) = \mathbb{E}\left[e^{tX}\right]$$

Depending on the random variable $X, M_X(t)$ might be infinite for some values of t. Here are two examples:

Example 1.12. *(Normal MGF). Suppose $X \sim \mathcal{N}(0, 1)$. Then*

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \int e^{tx}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx = \int \frac{1}{\sqrt{2\pi}}e^{\frac{-x^2+2tx}{2}}\,dx.$$

To compute this integral, we complete the square:

$$\int \frac{1}{\sqrt{2\pi}}e^{\frac{-x^2+2tx}{2}}\,dx = \int \frac{1}{\sqrt{2\pi}}e^{\frac{-x^2+2tx-t^2}{2}+\frac{t^2}{2}}\,dx = e^{\frac{t^2}{2}}\int \frac{1}{\sqrt{2\pi}}e^{-\frac{(x-t)^2}{2}}\,dx.$$

The quantity inside the last integral above is the PDF of the $\mathcal{N}(t, 1)$ distribution-hence it must integrate to 1. Then $M_X(t) = e^{t^2/2}$.

Now suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. Then $X = \mu + \sigma Z$, where $Z \sim \mathcal{N}(0, 1)$. So

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[e^{\mu t + \sigma t Z}\right] = e^{\mu t}\mathbb{E}\left[e^{\sigma t Z}\right] = e^{\mu t} M_Z(\sigma t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

For a normal random variable $X, M_X(t)$ is finite for all $t \in \mathbb{R}$.

Example 1.13. *(Gamma MGF). Suppose $X \sim \text{Gamma}(\alpha, \beta)$, for $\alpha, \beta > 0$. Then*

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \int_0^\infty e^{tx} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{(t-\beta)x} dx.$$

If $t > \beta$, then $\lim_{x\to\infty} x^{\alpha-1}e^{(t-\beta)x} = \infty$, so certainly the integral above is infinite. If $t = \beta$, note that $\int_0^\infty x^{\alpha-1}dx = \frac{1}{\alpha}x^\alpha\big|_0^\infty = \infty$, since $\alpha > 0$. Hence $M_X(t) = \infty$ for any $t \geq \beta$. For $t < \beta$, let us rewrite the above to isolate the PDF *of the* $\text{Gamma}(\alpha, \beta - t)$ *distribution:*

$$M_X(t) = \frac{\beta^\alpha}{(\beta - t)^\alpha} \int_0^\infty \frac{(\beta - t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta-t)x} dx.$$

As the PDF of the $\text{Gamma}(\alpha, \beta - t)$ *distribution integrates to 1, we obtain finally*

$$\begin{aligned} M_X(t) &= \begin{cases} \infty & t \geq \beta \\ \frac{\beta^\alpha}{(\beta-t)^\alpha} & t < \beta \end{cases} \\ &= \begin{cases} \infty & t \geq \beta \\ (1 - \beta^{-1}t)^{-\alpha} & t < \beta. \end{cases} \end{aligned}$$

If the MGF of a random variable X is finite in any interval that contains 0 as an interior point, as in the above two examples, then (like the PDF or CDF) it also completely specifies the distribution of X. This is the content of the following theorem (which we will not prove in this class):

Theorem 1.1. *Let X and Y be two random variables such that, for some $h > 0$ and every $t \in (-h, h)$, both $M_X(t)$ and $M_Y(t)$ are finite and $M_X(t) = M_Y(t)$. Then X and Y have the same distribution.*

The reason why the MGF will be useful for us is because if $X_1, \ldots, X_n$ are independent, then the MGF of their sum satisfies

$$M_{X_1+\ldots+X_n}(t) = \mathbb{E}\left[e^{t(X_1+\ldots+X_n)}\right] = \mathbb{E}\left[e^{tX_1}\right] \times \ldots \times \mathbb{E}\left[e^{tX_n}\right] = M_{X_1}(t)\ldots M_{X_n}(t)$$

This gives us a very simple tool to understand the distributions of sums of independent random variables.

2 Sampling Distributions

For data $X_1, \ldots, X_n$, a statistic $T(X_1, \ldots, X_n)$ is any real-valued function of the data. In other words, it is any number that you can compute from the data. For example, the sample mean

$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$$

the sample variance

$$S^2 = \frac{1}{n-1}\left(\left(X_1 - \bar{X}\right)^2 + \ldots + \left(X_n - \bar{X}\right)^2\right)$$

and the range

$$R = \max(X_1, \ldots, X_n) - \min(X_1, \ldots, X_n)$$

are all statistics. Since the data $X_1, \ldots, X_n$ are realizations of random variables, a statistic is also a (realization of a) random variable. A major use of probability in this course will be to understand the distribution of a statistic, called its **sampling distribution**, based on the distribution of the original data $X_1, \ldots, X_n$. Let's work through some examples:

Example 2.1. *(Sample mean of IID normals). Suppose* $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$. *The sample mean* $\bar{X}$ *is actually a special case of the quantity* $a_1X_1 + \ldots + a_nX_n$ *from Example 5.1, where* $a_i = \frac{1}{n}$, $\mu_i = \mu$, *and* $\sigma_i^2 = \sigma^2$ *for all* $i = 1, \ldots, n$. *Then from that Example,*

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Example 2.2. *(Chi-squared distribution). Suppose* $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(0,1)$. *Let's derive the distribution of the statistic* $X_1^2 + \ldots + X_n^2$.

By independence of $X_1^2, \ldots, X_n^2$,

$$M_{X_1^2+\ldots+X_n^2}(t) = M_{X_1^2}(t) \times \ldots \times M_{X_n^2}(t).$$

We may compute, for each X_i, its MGF

$$M_{X_i^2}(t) = \mathbb{E}\left[e^{tX_i^2}\right] = \int e^{tx^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx = \int \frac{1}{\sqrt{2\pi}}e^{\left(t-\frac{1}{2}\right)x^2}dx.$$

If $t \geq \frac{1}{2}$, then $M_{X_i^2}(t) = \infty$. Otherwise,

$$M_{X_i^2}(t) = \frac{1}{\sqrt{1-2t}}\int \sqrt{\frac{1-2t}{2\pi}}e^{-\frac{1}{2}(1-2t)x^2}dx.$$

We recognize the quantity inside this integral as the PDF of the $\mathcal{N}\left(0, \frac{1}{1-2t}\right)$ distribution, and hence the integral equals 1. Then

$$M_{X_i^2}(t) = \begin{cases} \infty & t \geq \frac{1}{2} \\ (1-2t)^{-1/2} & t < \frac{1}{2} \end{cases}$$

This is the MGF of the Gamma $\left(\frac{1}{2}, \frac{1}{2}\right)$ distribution, so $X_i^2 \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$. This is also called the **chi-squared distribution with 1 degree of freedom**, denoted χ_1^2.

Going back to the sum,

$$M_{X_1^2+\ldots+X_n^2}(t) = M_{X_1^2}(t) \times \ldots \times M_{X_n^2}(t) = \begin{cases} \infty & t \geq \frac{1}{2} \\ (1-2t)^{-n/2} & t < \frac{1}{2} \end{cases}$$

This is the MGF of the Gamma $\left(\frac{n}{2}, \frac{1}{2}\right)$ distribution, so $X_1^2 + \ldots + X_n^2 \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$. This is called the **chi-squared distribution with n degrees of freedom**, denoted χ_n^2.

The mean and variance of the χ_n^2 can be obtained from the MGF:

$$\frac{d}{dt}\left[M_{\chi^2}(t)\right]\Big|_{t=0} = \frac{d}{dt}\left[(1-2t)^{-n/2}\right]\Big|_{t=0} = \mathbb{E}\left(\chi^2\right) = n.$$

$$\frac{d^2}{dt^2}\left[M_{\chi^2}(t)\right]\Big|_{t=0} = \frac{d^2}{dt^2}\left[(1-2t)^{-n/2}\right]\Big|_{t=0} = n(n+2);\ \therefore \text{Var}\left(\chi^2\right) = 2n.$$

Example 2.3. *The Chi-square distribution. The probability density function of the chi-square (χ_n^2) distribution with n degrees of freedom is*

$$f(x;n) = \frac{x^{\frac{n}{2}-1}e^{-\frac{n}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}, \quad x > 0.$$

The mean of the chi-square distribution is n and the variance is $2n$. ***Parameters:*** *$n > 0$.*

Example 2.4. *Student's t-distribution. Student's t-distribution has pdf*

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}},$$

where $x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0$, and $\nu > 0$. The mean of the Student's t-distribution is μ for degree of freedom $\nu > 0$, and undefined for $\nu \leq 1$. The variance of this distribution is $\frac{\nu}{\nu-2}$ for $\nu > 2$, and undefined or infinite for $\nu \leq 2$. Moreover, the Student's t-distribution converges to the normal distribution (pointwise) as $\nu \to \infty$. ***Parameters:*** *$\mu \in \mathbb{R}, \sigma > 0$, and $\nu > 0$.*

Relationship with other distributions:

- Let $X_1, \ldots, X_n$ be $i.i.d$ random variables with distribution $N(0,1)$. Let $Y = \sum_{j=1}^{n} X_j^2$, then $Y \sim \chi_n^2$.
- Let $X \sim N(0,1)$ and $W \sim \chi_n^2$. Then, $Y = \frac{X}{\sqrt{\frac{W}{n}}}$ has Student's t-distribution $(\mu = 0, \sigma = 1)$ with n degrees of freedom.

Example 2.5. *F distribution. Suppose X and Y are independently distributed as chi-squared with m and n degrees of freedom, respectively. Define a statistic $F_{m,n}$ as the ratio of the dispersion of the two distributions*

$$F_{m,n} = \frac{X/m}{Y/n}$$

is said to have F distribution with (m,n) degrees of freedom.

Remark 2. *"Degrees of freedom" in the expression of χ^2, t, and F distribution is the number of unrestricted variables in the expression. Suppose $\sum_{i=1}^{n}(X_i - \bar{X})^2$ has $(n-1)$ degrees of freedom since out of n variables one restriction $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$ is imposed. As a rule of thumb, degrees of freedom = (the number of quantities involved in the expression) - (the number of linear restrictions).*

3 Kernels and Parameters

Definition 3.1. *The qth quantile of the distribution of a random variable X, is that value x such that $P(X < x) = q$. If $q = 0.5$, the value is called the median. The cases $q = 0.25$ and $q = 0.75$ correspond to the lower quartile and upper quartile, respectively.*

Definition 3.2. *The kernel of a probability density function (pdf) or probability mass function (pmf) is the factor of the pdf or pmf in which any factors that are not functions of any of the variables in the domain are omitted.*

For example, the kernel of the Beta distribution is:

$$K(x; a, b) = x^{a-1}(1-x)^{b-1}.$$

The kernel of the Gaussian (Normal) distribution is:

$$K(x; \mu, \sigma) = \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Types of parameters

The parameters of a distribution are classified into three types: location parameters, scale parameters, and shape parameters.

Definition 3.3. *A parameter μ of a distribution function $F(x;\mu)$ is called a location parameter if $F(x;\mu) = F(x-\mu;0)$. An equivalent definition can be made on the probability density function. This is, a parameter μ of a density function $f(x;\mu)$ is called a location parameter if $f(x;\mu) = f(x-\mu;0)$.*

An example of a location parameter is the parameter μ in the Gaussian distribution.

Definition 3.4. *A parameter $\sigma > 0$ of a distribution function $F(x;\sigma)$ is called a scale parameter if $F(x;\sigma) = F\left(\frac{x}{\sigma};1\right)$. An equivalent definition can be made on the probability density function. That is, a parameter σ of a density function $f(x;\sigma)$ is called a scale parameter if $f(x;\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma};1\right)$.*

An example of a scale parameter is the parameter σ in the Gaussian distribution.

Definition 3.5. *A shape parameter is a parameter that is neither a location nor a scale parameter. This kind of parameter controls the shape of a distribution (equivalently, density) function; e.g., Lehmann alternatives.*

An example of a shape parameter is the parameter κ in the Gamma distribution.

Definition 3.6. *A distribution is said to belong to the **Location-Scale family** of distributions if it is parameterized in terms of a location and a scale parameter. Examples of members of the location-scale family are the Gaussian and Logistic distributions. The Gamma distribution is not a member of this family (why?).*

Definition 3.7. *A distribution $F(x;\theta)$ is said to be **Identifiable** if $F(x;\theta_1) = F(x;\theta_2)$, for all x, implies that $\theta_1 = \theta_2$ for all possible values of θ_1, θ_2.*

4 Bivariate Normal Distribution

Definition 4.1. *A bivariate random variable (X, Y) is said to have a bivariate normal distribution if the pdf of (X, Y) is of the following form:*

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\}}, \quad (x,y) \in \mathbb{R}^2$$

where $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 > 0$, $|\rho| < 1$. Then, we write $(X, Y) \sim$ $\mathrm{BN}\left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho\right)$.

4.1 Marginal Distribution:

Note that,

$$\frac{1}{1-\rho_2}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\}$$

$$= \left[\frac{\left\{y-\mu_2-\rho\frac{\sigma_2}{\sigma_1}\left(x-\mu_1\right\}^2\right.}{\sigma_2^2\left(1-\rho^2\right)} + \left(\frac{x-\mu_1}{\sigma_1}\right)^2\right] = \frac{(y-\beta x)^2}{\sigma_{2.1}^2} + \frac{\left(x-\mu_1\right)^2}{\sigma_1^2},$$

where $\beta_x = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}\left(x-\mu_1\right)$ $\sigma_{2\cdot 1}^2 = \sigma_2^2\left(1-p^2\right)$. Marginal pdf of X is

$$f_X(x) = \frac{\exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right]}{\sigma_1\sqrt{2\pi}} \times \int_{-\infty}^{\infty} \frac{e^{-\frac{(y-\beta x)^2}{2\sigma_{2\cdot 1}{}^2}}}{(\sqrt{2\pi})\sigma_{2\cdot 1}} dy = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}, x \in \mathbb{R}$$

Here, $X \sim N\left(\mu_1, \sigma_1^2\right)$. Similarly, it can be shown that $Y \sim N\left(\mu_2, \sigma_2^2\right)$.

4.2 Conditional Distribution:

The PDF of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\ e^{-\frac{1}{2}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\beta_x}{\sigma_{2\cdot 1}}\right)^2\right\}}}{\frac{1}{\sigma_1\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}} = \frac{1}{\sigma_{2\cdot 1}\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(\frac{y-\beta x}{\sigma_{2\cdot 1}}\right)^2},\ y \in \mathbb{R}.$$

Hence, $Y|X = x \sim N\left(\beta x, \sigma_{2.1}^2\right)$. $\Leftrightarrow Y|X = x \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}\left(x-\mu_1\right), \sigma_2^2\left(1-\rho^2\right)\right)$. Similarly, it can be shown that, $X|Y = y \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}\left(y-\mu_2\right), \sigma_1^2\left(1-p^2\right)\right)$.

Remark 3. *1. Note that* $E(Y|X=x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)$ *and* $\operatorname{var}(Y|X=x) = \sigma_2^2(1-\rho^2)$. *Hence, the regression of* Y *on* X *is linear and the conditional distribution is homoscedastic.*

2.

$$
\begin{aligned}
E(XY) &= E[E(XY|X)] = E[X \cdot E(Y|X)] \\
&= E\left[X\left\{\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)\right\}\right] \\
&= \mu_1\mu_2 + \rho\frac{\sigma_2}{\sigma_1}\cdot\sigma_1^2 \quad \left[\because E\left\{X(X-\mu_1)\right\} = E(X-\mu_1)^2 = \sigma_1^2\right]
\end{aligned}
$$

$$
\Rightarrow E(XY) = \mu_1\mu_2 + \rho\sigma_1\sigma_2 \Rightarrow \frac{E(XY)-\mu_1\mu_2}{\sigma_1\sigma_2} = \rho \Rightarrow \rho_{XY} = \rho.
$$

3. If $\rho^2 = 1$, *then the PDF becomes undefined. But* $\rho = \pm 1$, *then* $P[\alpha X + \beta Y + \gamma = 0] = 1$ *for some non-null* (α, β), *which it known as singular on degenerate Bivariate distribution,*

5 Multivariate Normal Distribution

Definition 5.1. *The multivariate normal distribution of a* p*-dimensional random vector* $\mathbf{X} = (X_1, \ldots, X_p)$ *is said to be distributed as a multivariate Normal if and only if its probability density function is:*

$$
\phi_{\mathbf{X}}(x_1, \ldots, x_p; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}}\exp\left(-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu})\right),
$$

where $\mu = \mathbb{E}[\mathbf{X}]$ *is the location parameter and* $\Sigma = \mathbb{E}\left[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^\top\right]$ *is the covariance matrix. We denote it as* $\mathbf{X} \sim N_p(\mu, \Sigma)$.

The Multivariate Normal distribution of dimension p is a distribution for p random variables $X_1, \ldots, X_p$ which generalizes the normal distribution for a single variable. It is parametrized by a mean vector $\mu \in \mathbb{R}^p$ and a symmetric covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$, and we write

$$
(X_1, \ldots, X_p) \sim \mathcal{N}(\mu, \Sigma).
$$

Rather than writing down the general formula for its joint PDF (which we will not use in this course), let's define this distribution by the following properties:

Definition 5.2. *$(X_1, \ldots, X_p)$ have a multivariate normal distribution if, for every choice of constants $a_1, \ldots, a_p \in \mathbb{R}$, the linear combination $a_1X_1 + \ldots + a_pX_p$ has a (univariate) normal distribution. $(X_1, \ldots, X_p)$ have the specific multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ when, in addition,*

1. *$\mathbb{E}[X_i] = \mu_i$ and $\mathrm{Var}[X_i] = \Sigma_{ii}$ for every $i = 1, \ldots, p$, and*
2. *$\mathrm{Cov}[X_i, X_j] = \Sigma_{ij}$ for every pair $i \neq j$.*

When $(X_1, \ldots, X_p)$ are multivariate normal, each X_i has a (univariate) normal distribution, as may be seen by taking $a_i = 1$ and all other $a_j = 0$ in the above definition. The vector μ specifies the means of these individual normal variables, the diagonal elements of Σ specify their variances, and the off-diagonal elements of Σ specify their pairwise covariances.

Example 5.1. *If $X_1, \ldots, X_p$ are normal and independent, then $a_1X_1 + \ldots + a_pX_p$ has a normal distribution for any $a_1, \ldots, a_p \in \mathbb{R}$. To show this, we can use the MGF: Suppose $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Then $a_iX_i \sim \mathcal{N}(a_i\mu_i, a_i^2\sigma_i^2)$, so (from Example 1.12) a_iX_i has MGF*

$$M_{a_iX_i}(t) = e^{a_i\mu_i t + \frac{a_i^2\sigma_i^2t^2}{2}}.$$

As $a_1X_1, \ldots, a_pX_p$ are independent, the MGF of their sum is the product of their MGFs:

$$\begin{aligned} M_{a_1X_1+\ldots+a_pX_p}(t) &= M_{a_1X_1}(t) \times \ldots \times M_{a_pX_p}(t) \\ &= e^{a_1\mu_1t+\frac{a_1^2\sigma_1^2t^2}{2}} \times \ldots \times e^{a_p\mu_pt+\frac{a_p^2\sigma_p^2t^2}{2}} \\ &= e^{(a_1\mu_1+\ldots+a_p\mu_p)t+\frac{\left(a_1^2\sigma_1^2+\ldots+a_p^2\sigma_p^2\right)t^2}{2}}. \end{aligned}$$

But this is the MGF of a $\mathcal{N}\left(a_1\mu_1 + \ldots + a_p\mu_p, a_1^2\sigma_1^2 + \ldots + a_p^2\sigma_p^2\right)$ random variable! As the MGF uniquely determines the distribution, this implies $a_1X_1 + \ldots + a_pX_p$ has this normal distribution.

Then by definition, $(X_1, \ldots, X_p)$ are multivariate normal. More specifically, in this case we must have $(X_1, \ldots, X_p) \sim \mathcal{N}(\mu, \Sigma)$ where $\mu_i = \mathbb{E}[X_i], \Sigma_{ii} = \mathrm{Var}[X_i]$, and $\Sigma_{ij} = 0$ for all $i \neq j$.

Example 5.2. *Suppose $(X_1, \ldots, X_p)$ have a multivariate normal distribution, and $(Y_1, \ldots, Y_m)$ are such that each Y_j $(j = 1, \ldots, m)$ is a linear combination of $X_1, \ldots, X_p$:*

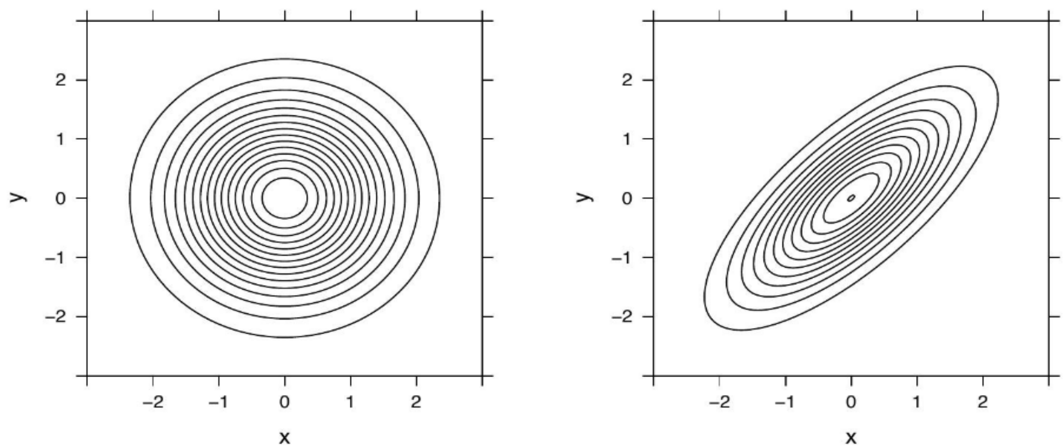
$$Y_j = a_{j1}X_1 + \ldots + a_{jp}X_p$$

for some constants $a_{j1}, \ldots, a_{jp} \in \mathbb{R}$. Then any linear combination of $(Y_1, \ldots, Y_m)$ is also a linear combination of $(X_1, \ldots, X_p)$, and hence is normally distributed. So $(Y_1, \ldots, Y_m)$ also have a multivariate normal distribution.

For two arbitrary random variables X and Y, if they are independent, then $\text{corr}(X, Y) = 0$. The converse is in general not true: X and Y can be uncorrelated without being independent. But this converse is true in the special case of the multivariate normal distribution; more generally, we have the following:

Theorem 5.1. *Suppose* $\mathbf{X}$ *is multivariate normal and can be written as* $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, *where* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are subvectors of* $\mathbf{X}$ *such that each entry of* $\mathbf{X}_1$ *is uncorrelated with each entry of* $\mathbf{X}_2$. *Then* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are independent.*

To visualize what the joint PDF of the multivariate normal distribution looks like, let's just consider the two-dimensional setting $k = 2$, where we obtain the special case of a **Bivariate Normal** distribution for two random variables X, Y. In this case, the distribution is specified by the means μ_1 and μ_2 of X and Y, the variances σ_1^2 and σ_2^2 of X and Y, and the correlation ρ between X and Y. When $\sigma_1^2 = \sigma_2^2 = 1$ and $\mu_1 = \mu_2 = 0$, the contours of the joint PDF of X and Y are shown below, for $\rho = 0$ on the left and $\rho = 0.75$ on the right:



When $\rho = 0$, X and Y are independent standard normal variables, and these contours are circular; the joint PDF has a peak at 0 and decays radially away from 0. When $\rho = 0.7$, the contours are ellipses. As ρ increases to 1, the contours concentrate more and more around the line $y = x$. (In the general k-dimensional setting and for general μ and Σ, the joint PDF has a single peak at the mean $\mu \in \mathbb{R}^k$, and it decays away from μ with contours that are ellipsoids around μ, with their shape depending on Σ.)

Example 5.3. *Multivariate t Distribution. The probability density function of the d-dimensional multivariate Student's t distribution is given by*

$$f(x, \Sigma, \nu) = \frac{1}{|\Sigma|^{1/2}} \frac{1}{\sqrt{(\nu\pi)^d}} \frac{\Gamma((\nu + d)/2)}{\Gamma(\nu/2)} \left(1 + \frac{x'\Sigma^{-1}x}{\nu}\right)^{-(\nu+d)/2},$$

where x is a $1 \times d$ vector, Σ is a $d \times d$ symmetric, positive definite matrix, and ν is a positive scalar. While it is possible to define the multivariate Student's t for singular Σ, the density cannot be written as above.

The multivariate Student's t distribution is a generalization of the univariate Student's t to two or more variables. It is a distribution for random vectors of correlated variables, each element of which has a univariate Student's t distribution. In the same way, as the univariate Student's t distribution can be constructed by dividing a standard univariate normal random variable by the square root of a univariate chi-square random variable, the multivariate Student's t distribution can be constructed by dividing a multivariate normal random vector having zero mean and unit variances by a univariate chi-square random variable. The multivariate Student's t distribution is parameterized with a correlation matrix, Σ, and a positive scalar degree of freedom parameter, ν. ν is analogous to the degrees of freedom parameter of a univariate Student's t distribution. The off-diagonal elements of Σ contain the correlations between variables. Note that when Σ is the identity matrix, variables are uncorrelated; however, they are not independent.

The multivariate Student's t distribution is often used as a substitute for the multivariate normal distribution in situations where it is known that the marginal distributions of the individual variables have fatter tails than the normal.

6 Modes of Convergence

Definition 6.1. *Let $X_1, X_2, \ldots$ be a sequence of independent random variables. X_n converges almost surely (a.s.) to the random variable X, as $n \to \infty$, if and only if*

$$P\left(\{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}\right) = 1.$$

Notation: $X_n \stackrel{a.s}{\to} X$ as $n \to \infty$. Almost sure convergence is often referred to as strong convergence.

Definition 6.2. *Let $X_1, X_2, \ldots$ be a sequence of independent random variables. X_n converges in probability to the random variable X, as $n \to \infty$, if and only if, for all $\varepsilon > 0$:*

$$P\left(|X_n - X| > \varepsilon\right) \to 0 \text{ as } n \to \infty.$$

Notation: $X_n \stackrel{P}{\to} X$ as $n \to \infty$.

Recall now that the expectation (or the mean) of a continuous random variable X with probability density function f is defined as:

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx,$$

the n-th moment of the random variable X is defined as:

$$E\left[X^n\right] = \int_{-\infty}^{\infty} x^n f(x) dx,$$

and the n-th absolute moment of the random variable X is defined as:

$$E|X|^n = \int_{-\infty}^{\infty} |x|^n f(x) dx.$$

Definition 6.3. *Let $X_1, X_2, \ldots$ be a sequence of independent random variables. X_n converges in r-mean to the random variable X, as $n \to \infty$, if and only if, for all $\varepsilon > 0$:*

$$E\left|X_n - X\right|^r \to 0 \text{ as } n \to \infty.$$

Notation: $X_n \xrightarrow{r} X$ as $n \to \infty$.

Definition 6.4. *Let $X_1, X_2, \ldots$ be a sequence of independent random variables. X_n converges in distribution to the random variable X, as $n \to \infty$, if and only if:*

$$F_{X_n}(x) \to F_X(x) \text{ as } n \to \infty \text{ for all } x \in C\left(F_X\right),$$

where F_{X_n} and F_X are the cumulative distribution functions of X_n and X, respectively, and $C\left(F_X\right)$ is the continuity set of F_X (that is, the points where F_X is continuous). Notation: $X_n \xrightarrow{d} X$ as $n \to \infty$.

Theorem 6.1. *Slutsky's theorem. Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be sequences of random variables. Suppose that*

$$X_n \xrightarrow{d} X, \quad \text{ and } Y_n \xrightarrow{P} a, \text{ as } n \to \infty,$$

where a is some constant. Then, as $n \to \infty$

$$\begin{aligned} X_n + Y_n &\xrightarrow{d} X + a, \\ X_n - Y_n &\xrightarrow{d} X - a, \\ X_n \cdot Y_n &\xrightarrow{d} X \cdot a, \\ \frac{X_n}{Y_n} &\xrightarrow{d} \frac{X}{a}, \quad \text{ for } a \neq 0. \end{aligned}$$

Theorem 6.2. *Convergence of sums of sequences of random variables*

1. *Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be sequences of random variables. Suppose that*

$$X_n \stackrel{ass}{\to} X, \text{ and } Y_n \stackrel{as.}{\to} Y, \text{ as } n \to \infty.$$

   Then,

$$X_n + Y_n \stackrel{a.s}{\to} X + Y, \text{ a.s } n \to \infty.$$

2. *Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be sequences of random variables. Suppose that*

$$X_n \xrightarrow{P} X, \quad \text{and} \quad Y_n \xrightarrow{P} Y, \quad \text{as } n \to \infty.$$

   Then,

$$X_n + Y_n \xrightarrow{P} X + Y, \text{ as } n \to \infty$$

6.1 The Law of Large Numbers and the Central Limit Theorem

Definition 6.5. *We say that two random variables X and Y are identically distributed if and only if $P(X \leq x) = P(Y \leq x)$, for all x. If two variables are independent and identically distributed, we say that they are "i.i.d.".*

Theorem 6.3. *The weak law of large numbers. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite mean μ, and set $S_n = X_1 + X_2 + \cdots + X_n, n \geq 1$. Then*

$$\bar{X}_n = \frac{S_n}{n} \xrightarrow{P} \mu \quad \text{as} \quad n \to \infty.$$

Alternatively, for any fixed $\varepsilon > 0$, as $n \to \infty$,

$$\mathbb{P}\left[\left|\bar{X}_n - \mu\right| > \varepsilon\right] \to 0.$$

Corollary 6.1. *Let h be a measurable function and $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables with distribution F. Suppose that $E[h(X)] < \infty$, for $X \sim F$. Then, by the law of large numbers*

$$\frac{1}{n} \sum_{i=1}^{n} h\left(X_i\right) \xrightarrow{P} E[h(X)] \quad \text{as} \quad n \to \infty.$$

Theorem 6.4. *The strong law of large numbers. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite mean μ and finite variance, and set $S_n = X_1 + X_2 + \cdots + X_n$, $n \geq 1$. Then*

$$\bar{X}_n = \frac{S_n}{n} \xrightarrow{a.s} \mu \quad a.s\ n \to \infty.$$

Theorem 6.5. *The central limit theorem (univariate case). Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite mean μ and finite variance σ^2, and set $S_n = X_1 + X_2 + \cdots + X_n$, $n \geq 1$. Then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty.$$

Alternatively, for any fixed $x \in \mathbb{R}$, as $n \to \infty$,

$$\mathbb{P}\left[\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \leq x\right] \to \Phi(x),$$

where Φ is the CDF of the $\mathcal{N}(0,1)$ distribution.

Theorem 6.6. *The central limit theorem (multivariate case) Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ be a sequence of i.i.d. p-dimensional random vectors, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$. Suppose that*

$$E\left\|\mathbf{X}_1\right\|^2 = E\left(X_{11}^2 + \cdots + X_{1p}^2\right) < \infty,$$

and set $\mathbf{S}_n = \mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n$. The central limit theorem asserts that

$$\sqrt{n}\left(\mathbf{S}_n - E\left[\mathbf{X}_1\right]\right) \xrightarrow{d} N\left(0, \operatorname{Cov}\left[\mathbf{X}_1\right]\right) \quad \text{as} \quad n \to \infty,$$

where $\operatorname{Cov}\left[\mathbf{X}_1\right]$ is the $p \times p$ covariance matrix of the random vector $\mathbf{X}_1$.

The LLN and CLT can be used as building blocks to understand other statistics, via the **Continuous Mapping Theorem**:

Theorem 6.7. *(Continuous mapping theorem). Let $X_1, X_2, \ldots$ be a sequence of random variables on $\mathbb{R}$. Suppose that $g : \mathbb{R} \to \mathbb{R}$ is a continuous function (almost surely). Then,*

$$\begin{aligned} X_n &\xrightarrow{d} X \quad \text{implies} \quad g\left(X_n\right) \xrightarrow{d} g(X), \\ X_n &\xrightarrow{P} X \quad \text{implies} \quad g\left(X_n\right) \xrightarrow{P} g(X), \\ X_n &\xrightarrow{a.s.} X \quad \text{implies} \quad g\left(X_n\right) \xrightarrow{a.s.} g(X). \end{aligned}$$

7 Order Statistics

Introduction: Let $X_1, X_2, \ldots, X_n$ be a random sample of size n drawn from a population with distribution function F. If the observations $X_1, X_2, \ldots, X_n$ are arranged in increasing order of magnitude then the rearranged random variables $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ are called the order statistics of the sample. In case of sampling from continuous population we have $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ with probability 1. It is clear that the order statistic $X_{(i)}^{\prime} s$ is dependent though the original observation $X_1, X_2, \ldots, X_n$ are independent. Thus the r^{th} order statistic of the sample of size r is simply the r^{th} order statistic (r^{th} smallest observation in the sample) and is denoted by $X_{(r)}$.

7.1 Exact sampling distribution of order statistic

(a) **Distribution of $X_{(1)}$:** Let us consider a random sample $X_1, X_2, \ldots, X_n$ drawn from a population having distribution function $F(\cdot)$. $X_{(1)}$ is the first order statistic, the distribution function of $X_{(1)}$ is given by

$$
\begin{aligned}
F_{X_{(1)}}(x) & =1-P\left[X_{(1)}>x\right] \\
& =1-P\left[X_1>x, X_2>x, \ldots ., X_n>x\right] \\
& =1-\prod_{i=1}^n P\left[X_i>x\right] \quad\left[\because X_1, \ldots, X_n \text{ are independent }\right] \\
& =1-\left\{P\left[X_1>x\right]\right\}^n \\
& =1-(1-F(x))^n .
\end{aligned}
$$

$\therefore$ The pdf of $X_{(1)}$ is given by.

$$
f_{X_{(1)}}(x)=n(1-F(x))^{n-1} f(x) .
$$

(b) **Distribution of $X_{(n)}$** : The distribution function of $X_{(n)}$ is given by,

$$
\begin{aligned}
F_{X_{(n)}}(x) & =P\left[X_{(n)} \leq x\right] \\
& =P\left[X_1 \leqslant x, X_2 \leq x, \ldots, X_n \leqslant x\right] \\
& =\left(P\left[X_1 \leqslant x\right]\right)^n \quad\left[\because X_i^{\prime} s \text{ are i.i.d. }\right] \\
& =[F(x)]^n
\end{aligned}
$$

$\therefore$ The pdf of $X_{(n)}$ is given by,

$$
f_{X_{(n)}}(x)=n\{F(x)\}^{n-1} f(x)
$$

(c) **Distribution of $X_{(r)}$, the general case:** Let $\left(X_{(1)}, X_{(2)}, \cdots, X_{(n)}\right)$ order statistics corresponding to the sample observation $\left(X_1, X_2, \ldots, X_n\right)$ having joint p.d.f.

$$
f_\theta\left(x_1, x_2, \ldots, x_n\right)=\prod_{i=1}^n f_\theta\left(x_i\right)
$$

if $g(y_1, y_2, \ldots, y_n)$ be the p.d.f. of $\left(X_{(1)}, X_{(2)}, \cdots, X_{(n)}\right)$ then

$$g(y_1, y_2, \ldots, y_n) = n! f(y_1, y_2, \ldots, y_n), -\infty < y_1 < y_2 < \cdots < y_n < \infty.$$

Let, $F_r(y)$ be the distribution function of r^{th} order statistic $X_{(r)}$ Then,

$$\begin{aligned} F_r(y) &= P[X_{(r)} \leq y] \\ &= P[\text{at least } r \text{ of the } n \text{ sample observation are } \leq y] \\ &= \sum_{t=r}^{n} \binom{n}{t} [F(y)]^t [1-F(y)]^{n-t}, \text{ where } F \text{ is the d.f. of } X \\ &= 1 - \sum_{t=0}^{r-1} \binom{n}{t} [F(y)]^t [1-F(y)]^{n-t} \\ &= 1 - \frac{1}{\beta(n-r+1, r)} \int_0^{1-F(y)} z^{n-r}(1-z)^{r-1} dz. \end{aligned}$$

So that the pdf of $X_{(r)}$ is

$$\begin{aligned} f_r(y) = \frac{d}{dy} F_r(y) &= [1-F(y)]^{n-r}[F(y)]^{n-1} \cdot f(y) \cdot \frac{1}{\beta(n-n+1, n)} \\ &= \frac{n!}{(r-1)!(n-r)!} [F(y)]^{r-1}[1-F(y)]^{n-r} \cdot f(y) \end{aligned}$$

Particular case: Let, $r = 1$, then the p.d.f. of minimum order statistic is, $f_1(y) = n[1-F(y)]^{n-1} f(y)$.

Let, $r = n$, then the p.d.f. of maximum order statistic is, $f_n(y) = n[F(y)]^{n-1} f(y)$.

Example 7.1. *Let $X \sim R(0, \theta)$ with pdf* $f_\theta(x) = \begin{cases} \frac{1}{\theta}; 0 < x < \theta \\ 0, \;\; ow \end{cases}$.

$\therefore$ *The pdf of $X_{(n)}$ is*

$$f_n(y) = \begin{cases} n\left(\frac{y}{\theta}\right)^{n-1} \cdot \frac{1}{\theta} & , 0 < y < \theta \\ 0 & , otherwise \end{cases}$$

and the pdf of $X_{(1)}$ is

$$f_1(y) = \begin{cases} n\left[1 - \frac{y}{\theta}\right]^{n-1} \cdot \frac{1}{\theta} & , 0 < y < \theta \\ 0 & , otherwise \end{cases}$$

Example 7.2. *Let $X \sim \mathrm{Exp}(\theta, 1)$*

$$\therefore f_\theta(x) = \begin{cases} e^{-(x-\theta)} & if\ x > \theta \\ 0 & otherwise \end{cases}$$

$\therefore$ *PDF of* $X_{(1)}$ *is*

$$f_1(y) = \begin{cases} n\left[1 - \int_\theta^y e^{-(x-\theta)}dx\right]^{n-1} \cdot e^{-(y-\theta)} & \text{if } \theta < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

(d) **Joint distribution of $X_{(1)}$ and $X_{(n)}$:**
The joint distribution function of $X_{(1)}$ and $X_{(n)}$ is given by

$$\begin{aligned} F_{X_{(1)},X_{(n)}}(x,y) &= P\left[X_{(1)} \leq x, X_{(n)} \leq y\right] \\ &= P\left[X_{(n)} \leq y\right] - P\left[X_{(1)} > x, X_{(n)} < y\right] \\ &= P\left[X_1, X_2, \ldots, X_n \leq y\right] - P\left[x < X_1, X_2, \ldots, X_n < y\right] \\ &= [F(y)]^n - [F(y) - F(x)]^n. \end{aligned}$$

$\therefore$ The joint pdf of $X_{(1)}$ and $X_{(n)}$ is given by,

$$\begin{aligned} f_{X_{(1)},X_{(n)}}(x,y) &= \frac{\partial^2}{\partial x \partial y} F_{X_{(1)},X_{(n)}}(x,y) \\ &= n(n-1)[F(y) - F(x)]^{n-2} f(x) f(y) \end{aligned}$$

(e) **Joint pdf of $X_{(r)}$ and $X_{(s)}$, the general case:**
The joint pdf of $X_{(r)}$ and $X_{(s)}$ is given by,

$$\begin{aligned} f_{X_{(r)},X_{(s)}}(x,y) &= \lim_{h\downarrow 0\, k\downarrow 0} \frac{1}{hk} P\left[x - \frac{h}{2} < X_{(r)} < x + \frac{h}{2}, y - \frac{k}{2} < X_{(s)} < y + \frac{k}{2}\right],\ if r < s \\ &= \lim_{h\downarrow 0\, k\downarrow 0} \frac{1}{hk} P\left[(r-1) \text{ obs.} < x - \frac{h}{2},\ \text{one obs. } \in \left(x - \frac{h}{2}, x + \frac{h}{2}\right),\right. \\ &\quad (s-r-1) \text{ obs.} \in \left(x + \frac{h}{2}, y - \frac{k}{2}\right),\ \text{one obs.} \in \left(y - \frac{k}{2}, y + \frac{k}{2}\right), \\ &\quad \left.(n-s) \text{ obs.} > y + \frac{k}{2}\right] \\ &= \lim_{h\downarrow 0 k\downarrow 0} \frac{n!}{(r-1)!(n-s)!(s-r-1)!} P\left[\frac{\text{a particular case}}{hk}\right] \\ &= \lim_{h\downarrow 0\, k\downarrow 0} \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \left\{F\left(x - \frac{h}{2}\right)\right\}^{r-1} \frac{hf(x)}{h} \cdot \left\{F\left(y - \frac{k}{2}\right)\right. \\ &\quad \left.-F\left(x + \frac{h}{2}\right)\right\}^{s-r-1} \frac{k \cdot f(y)}{k} \left\{1 - F\left(y + \frac{k}{2}\right)\right\}^{n-s} \\ &= \frac{n!}{(r-1)!(n-s)!(s-r-1)!} \{F(x)\}^{r-1} \{F(y) - F(x)\}^{s-r-1} \\ &\quad \{1 - F(y)\}^{n-s} f(x) f(y) \end{aligned}$$

Sample Median & Sample Range: Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ denote the order statistics of a random sample $X_1, X_2, \ldots, X_n$ from a density $f(\cdot)$. The sample median is defined to be the middle order statistic if n is odd and the average of the middle two order statistics if n is even. The sample range is defined to be $X_{(n)} - X_{(1)}$, sample mid-range is defined to be $\{X(n) + X(1)\}/2$.

Example 7.3. *Let $X_1, X_2, \ldots, X_n$ be iid RV's with common pdf*

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

Find the distribution of the sample range.

Solution: *The joint PDF of $X_{(1)}$ and $X_{(n)}$ is given by*

$$f_{X_{(1)},X_{(n)}}(x, y) = n(n-1)(F(y) - F(x))^{n-2}, 0 < x < y < 1.$$

Now, $F(y) = \int_0^y 1dx = y$ similarly,

$$\therefore \quad f(x) = x$$
$$f_{X_{(1)},X_{(n)}}(x, y) = n(n-1)(y-x)^{n-2}, 0 < x < y < 1.$$

Let us consider the following transformation. $\left(X_{(1)}, X_{(n)}\right) \longrightarrow \left(X_{(1)}, R\right)$ such that

$$R = X_{(n)} - X_{(1)}$$

$$|J| = \left|\frac{\partial X_{(n)}}{\partial R}\right| = 1.$$

Here, $r = y - x \Rightarrow y = x + r;\ 0 < y < 1 \Rightarrow 0 < x < 1 - r,\ 0 < r < 1.$
The joint PDF of $X_{(1)}$ and R is given by,

$$f_{X_{(1)},R}(x, r) = n(n-1)r^{n-2,}, 0 < x < 1 - r;\ 0 < r < 1$$

$\therefore$ *PDF of R is given by,*

$$\begin{aligned} f_R(r) &= n(n-1)r^{n-2} \int_0^{1-r} dx, 0 < r < 1 \\ &= n(n-1)r^{n-2}(1-r), 0 < r < 1 \end{aligned}$$

Hence the answer.

8 Additional tools

Higher moments can help us understand tail behavior, as seen in Markov and Chebyshev's Inequalities. This bound gets better as we take higher-order moments.

- Markov's inequality: $P[|X| \geq \alpha] \leq \frac{1}{\alpha^k} E\left[|X|^k\right]$ where $\alpha > 0$.
- Chebyshev inequality. Let $m = E[X]$ and $\alpha > 0$

$$P[|X - m| \geq \alpha] \leq \frac{1}{\alpha^2} \operatorname{Var}[X]$$

- Chernoff Bound. Suppose X is a random variable whose moment generating function is $M_X(t)$ and $a \in \mathbb{R}$

$$P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq e^{-ta} M_X(t) \quad \text{for some t } > \mathbf{0}$$

- Jensen's inequality. Let φ be a convex function on an interval containing the range of X. Then,

$$\varphi(E[X]) \leq E[\varphi(X)]$$

  The opposite is true for concave distributions.

- Holder's inequality. Let $p > 1, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$

$$E\|XY\| \leq E\left[|X|^p\right]^{\frac{1}{p}} \cdot E\left[|Y|^q\right]^{\frac{1}{q}}.$$

  The case $p = q = 2$ is known as the Cauchy-Schwarz inequality (show $|r_{XY}| \leq 1$!).

- The rank of a matrix $\mathbf{M}$ is the dimension of the vector space generated by its columns. This corresponds to the maximal number of linearly independent columns of $\mathbf{M}$. The rank is commonly denoted rank($\mathbf{M}$) or Rank($\mathbf{M}$). The square root of a non-singular matrix $\mathbf{M}$, denoted $\mathbf{M}^{\frac{1}{2}}$, is the matrix that satisfies $\mathbf{M} = \mathrm{M}^{\frac{1}{2}}\mathrm{M}^{\frac{1}{2}}$.

- Reparameterisation.

  Definition 8.1. *Let $f(\mathbf{x};\theta)$ be a pdf with parameters $\theta = (\theta_1, \ldots, \theta_p)^\top \in \Theta \subset \mathbb{R}^p, \mathbf{x} \in \mathcal{D} \subset \mathbb{R}^n$. A reparameterisation $\boldsymbol{\eta} = \varphi(\boldsymbol{\theta})$ is a change of variables $\theta_j \mapsto \eta_j, j = 1, \ldots, p$, via a one-to-one function φ such that, for each $\theta \in \Theta$, there exists $\eta \in \varphi(\Theta)$ such that $f(x;\theta) = f\left(x; \varphi^{-1}(\boldsymbol{\eta})\right)$. Analogously, for each $\eta \in \varphi(\Theta)$, there exists $\theta \in \Theta$ such that $f(x;\boldsymbol{\eta}) = f(x;\varphi(\boldsymbol{\theta}))$.*

  The use of reparameterization is very common in statistics. For instance, the Exponential distribution is often parameterized in terms of the rate parameter λ

or in terms of the mean $\beta = \frac{1}{\lambda}$. Another example is the Normal distribution, which is often parameterized in terms of the mean μ and the standard deviation σ; or in terms of the mean μ and the variance $\sigma_2 = \sigma^2$; or in terms of the mean μ and the precision $\tau = \frac{1}{\sigma^2}$. They are all equivalent as there exists a one-to-one function between the different parameterizations.

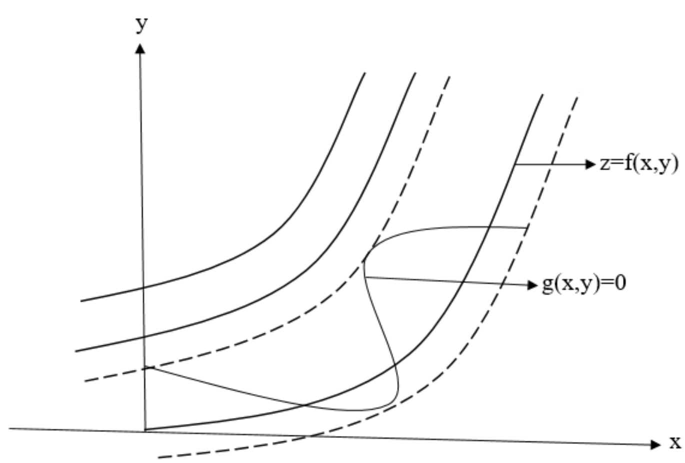
- Indicator function.

Definition 8.2. *The indicator function of the set A is a function*

$$\mathbf{I}_A : \mathbf{X} \to \{0,1\} \text{ defined as:}$$

$$\mathbf{I}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

- **Method of Lagrange's Multipliers:** Suppose we wish to minimize or maximize a function of two variables $z = f(x, y)$ where (x, y) is constrained to satisfy $g(x, y) = 0$. Assuming that these functions have continuous derivatives, we can visualize $g(x, y) = 0$ as a curve along with the level curve of $z = f(x, y)$.



Intuitively, if we move the level curve in the direction of increasing z, the largest or smallest z occurs at a point where a level curve touches $g(x, y) = 0$. The quadrants of 'f' and 'g ' should be in the same or opposite direction. Then $\nabla f = -\lambda \nabla g$ for some constant $\lambda \ni \nabla\{f + \lambda g\} = 0$.
Proof. $g(x, y) = 0$, $-\frac{dy}{dx} = -\frac{g_x}{g_y}$ and for $f(x, y) = c$, $\frac{dy}{dx} = -\frac{f_x}{f_y}$.

$$\text{At the point of tangency, } -\frac{f_x}{f_y} = \frac{dy}{dx} = -\frac{g_x}{g_y} \Rightarrow \quad \frac{f_x}{f_x} = \frac{f_y}{g_y} = -\lambda \text{ (say)}$$

$$\therefore (f_x, f_y) = -\lambda (g_x, g_y).$$

Hence, to find the maximum on minimum of $f(x,y)$ subject to $g(x,y)=0$, we find all the solution of equation,

$$\begin{aligned} &\nabla\{f+\lambda g\}=0 \text{ and } g(x,y)=0 \\ \Rightarrow &\frac{\partial F}{\partial x}=0=\frac{\partial F}{\partial y},\ g(x,y)=0;\ \text{where } F(x,y)=f(x,y)+\lambda g(x,y). \end{aligned}$$

Local maxima and minima will be among the solutions. If the curve $g(x,y)=0$ is closed and bounded, then the absolute maxima and minima of $f(x,y)$ exist and are among these solutions.

General Case: To maximize or minimizes $z=f\left(x_1,x_2,\ldots,x_n\right)$ subject to the constraints $g_i\left(x_1,x_2,\ldots,x_n\right)=0; i=1(1)k$, solve the following equations simultaneously,

$$\nabla\left\{f+\sum_{i=1}^{k}\lambda_i g_i\right\}=0 \text{ and } g_i\left(x_1,x_2,\ldots,x_n\right)=0, i=1(1)k.$$

The numbers $\lambda_1,\lambda_2,\ldots,\lambda_k$ are called the Lagrange's multipliers. The method for finding the extrema of a function subject to some constraints is called the "method of Lagrange's Multipliers".

Example 8.1. *Maximize $f(x,y)=x^2y$ subject to $x^2+xy=12$.*

We let $F(x,y)=x^2y+\lambda\left(x^2+xy-12\right)$

$$0=\frac{\partial F}{\partial x}=2xy+\lambda(2x+y)\rightarrow (i),\ 0=\frac{\partial F}{\partial y}=x^2+\lambda x\rightarrow (ii),\ x^2+xy=12\rightarrow (iii)$$

From (ii) $\rightarrow x(x+\lambda)=0\Rightarrow x=-\lambda$ as $x=0$ is not a solution of $x^2+xy=12$.

From (i) $\rightarrow -2\lambda y+\lambda 2(-\lambda)+\lambda y=0\Rightarrow -\lambda y=2\lambda^2\Rightarrow y=-2\lambda$

From (iii) $\rightarrow x=-\lambda, y=-2\lambda$, then $x^2+xy=12$ gives $\lambda=\pm 2$.

$\therefore (x,y)=(-2,-4)$ *or* $(2,4)$. *Hence* $\max\{xy\}=16, \min\{xy\}=-16.$