MATH 556 - Mathematical Statistics I

Pr. Johanna Nešlehová

Course notes by
Léo Raymond-Belzile
Leo.Raymond-Belzile@mail.mcgill.ca

The current version is that of September 6, 2018
Fall 2012, McGill University

Please signal the author by email if you find any typo.

These notes have not been revised and should be read carefully.

Contents

Chapter 1
Univariate random variables

An univariate random variable is a mapping $X : \Omega \to \mathbb{R}$, or more specifically $(\Omega, \mathcal{A}, \mathsf{P}) \to (\mathbb{R}, \mathbb{B})$ which is **measurable**. The distribution (or probabilistic behavior) of X is a probability measure on $(\mathbb{R}, \mathbb{B})$ which is induced by X:

$$\mathsf{P}^X(A) = \mathsf{P}\Big(\{\omega \in \Omega : X(\omega) \in A\}\Big)$$

Theorem 1.1

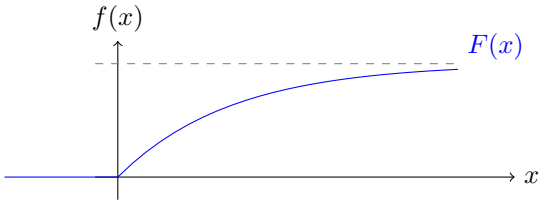
The distribution P^X of X is uniquely determined by the (cumulative) distribution function of X, defined for all $x \in \mathbb{R}$ by

$$F^X(x) = \mathsf{P}(X \leq x) = \mathsf{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathsf{P}^X((-\infty, x]). \tag{1.1}$$

Example 1.1

The unit exponential distribution has a CDF given by

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-x}, & x \geq 0. \end{cases}$$



Fact 1.2

- F describes P^X uniquely and

$$\mathsf{P}(X \in (a, b]) = F(b) - F(a) = \mathsf{P}(X \in (-\infty, b]) - \mathsf{P}(X \in (-\infty, a]);$$

- $\mathsf{P}(X < x) = F(x-) = \lim_{u \uparrow x} F(u)$;
- Every CDF is non-decreasing, right-continuous and such that $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.
- Whenever F has the above properties, then F is a CDF of a random variable X.

Be warned that the random variable corresponding to $F(x)$ may not be unique, even though X has a unique CDF, the opposite implication does not hold.

Example 1.2

Let X be Bernoulli, denote $\mathcal{B}(0.5)$. Notice that $\mathsf{P}(X = 0) = \mathsf{P}(X = 1) = 0.5$ and that

$$F(x) = \mathsf{P}(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 0.5 & \text{if } x \in [0, 1] \\ 1 & \text{if } x \geq 1. \end{cases}.$$

Looking at $Y = 1 - X$, we see that Y is clearly not the same as X, but has the same distribution as $\mathsf{P}(Y = 0) = \mathsf{P}(Y = 1) = 0.5$.

Remark

IF X and Y are random variables, then

- $X = Y$ means that $\forall\, \omega \in \Omega : X(\omega) = Y(\omega)$.
- $X = Y$ almost surely, sometimes denoted $X \overset{\text{a.s.}}{=} Y$ means that

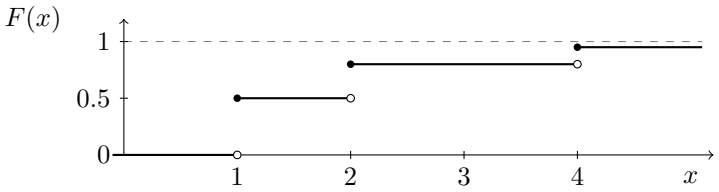
$$\mathsf{P}(X = Y) = \mathsf{P}(\omega : X(\omega) = Y(\omega)) = 1.$$

- $X \overset{d}{=} Y$, equality in distribution, means that the distributions of X and Y are equal, that is their CDFs satisfy $\forall\, x \in \mathbb{R}, F^X(x) = F^Y(x)$.

Equality implies equality almost surely, which then in turns imply equality in distribution. Each condition is weaker than the previous one.

Section 1.1. Discrete random variables

X is discrete if its CDF is a pure jump function (piecewise constant). It can be shown that F has at most countably many jumps and $S = \{x_i, i \in I\}$. A discrete distribution is uniquely given by its probability mass function (PMF), which is also called the counting density, and $f(x) = \mathsf{P}(X = x) = F(x) - F(x-)$, which corresponds to the jump height.



Usually, we only consider the points with a mass, $f(x_i) = \mathsf{P}(X = x_i)$ for $x_i \in S$. S is known as the support of X.

Proposition 1.3 (Examples of univariate discrete distributions)

1. **Bernoulli**(p), $S = \{0, 1\}$, denoted $X \sim \mathcal{B}(p)$. This correspond to a binary variable, which could be a coin toss. $f(0) = 1 - p$ and $f(1) = p$, so that p is the probability of success.

2. **Discrete uniform**(n), $S = \{1, \dots, n\}$, denoted $X \sim \mathcal{U}(n)$. The probability at every point of the support is equal and as such $f(k) = \frac{1}{n}, k \in S$. An example of application is a die toss.

3. **Binomial**(n, p), $S = \{0, \dots, n\}$, denoted $X \sim \mathcal{B}(n, p)$. The PMF of the Binomial distribution is given by $f(k) = \binom{n}{k} p^k (1-p)^{n-k}, k \in S$. If one consider a sequence of n independent Bernoulli, then the sum follows a Binomial distribution.

4. **Hypergeometric distribution**, denoted $\mathcal{H}(n, K, N - K)$. This distribution counts N items in total and K will denote for example defective items. Draw a sample of small n without distribution (contrary to binomial). The probability of drawing x defectives is given by

$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, \qquad x \leq n, x \leq K, n - x \leq N - K.$$

   A classical example of this is the Fisher exact test, which is used to determine whether there is any link between category, for example in assignment of two drugs. One might want to look at whether you have complete remission in the sick mice population. The

	Y	N	
Drug A	7	3	10
Drug B	2	7	9
	9	10	19

   probability of such table existing is simply governed by the Hypergeometric distribution. Hence $\mathsf{P}(X = 7) = \binom{10}{7}\binom{9}{2}/\binom{19}{9}$.

   A classical property of this distribution is the following. Suppose that both $N, K \to \infty$ in a way such that $K/N \to p \in [0, 1]$. If n is fixed, then the support of the Hypergeometric will become simply $\{0, \dots, n\}$ with $f(x) = \binom{n}{k} p^x (1-p)^{n-x}$.

5. **Poisson distribution**, denoted $\mathcal{P}(\lambda)$, where $\lambda > 0$. The probability mass function is given by

$$\mathsf{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

with $k = 0, 1, \ldots$. The number k could represent for example the number of costumers per hour at a store, or the number of mutations in a given DNA stretch.

Similarly to the Hypergeometric, the Poisson is related to the Binomial in the following way. The Poisson describes the number of rare events. Let $p_n \in (0,1)$ be such that $np_n \to \lambda$ as $n \to \infty$.. Then $X_n \sim \mathcal{B}(n, p_n)$. Then

$$\mathsf{P}(X_n = k) = \binom{n}{k} p_n^k (1-p_n)^{n-k} \xrightarrow{n\to\infty} e^{-\lambda} \frac{\lambda^k}{k!}.$$

6. **Geometric distribution**, denoted $\mathcal{G}(p)$ with $p \in (0,1)$. Its support is given by $\{0, 1, \ldots\}$ with PMF
$$f(x) = p^x(1-p)$$
where p is the probability of failure. In a sequence $X_1, X_2, \ldots$ of Bernoulli $\mathcal{B}(p)$ trials, X describes the waiting time until the first zero (or success).

7. **Negative binomial** is a more interesting distribution, that can be viewed as a generalization of the Geometric distribution. It is denoted $\mathcal{NB}(r, p)$ where $r \in \mathbb{N}, p \in (0,1)$. A special case of this is for $r = 1$, in which case you recover the $\mathcal{G}(p)$ distribution. The PMF has the form
$$f(k) = \binom{r+k-1}{r-1} p^k (1-p)^r,$$
k counting the number of failures, r, p are fixed. Another way of writing this, which will count the waiting time until the r^{th} success, is by taking $s = k + r$; then the probability of having the r^{th} success in the s^{th} trial
$$\mathsf{P}(r^{\text{th}} \text{ success in } s^{\text{th}} \text{ trial}) = \binom{s-1}{r-1} p^{s-r}(1-p)^r.$$
The latter form is more widely used.

We now move on to continuous random variables.

Section 1.2. Continuous random variables

First a remark. One must be careful as X can be called continuous if F is a continuous function (in which case it has no jump). Another used definition is that X is **absolutely continuous** if for every $x \in R$, the distribution function can be written as

$$F(x) = \int_{-\infty}^{x} f(t)dt \tag{1.2}$$

for some non-negative function f, which will be called the **density** of F (or X). These definitions are obviously not the same: continuity doesn't imply absolute continuity, but the converse implication is true. The latter is assured to have a derivative. For example, Cassela and Berger describe a continuous random variable as one assumed to have a density.

In practice, if there are no ties in the data, we can safely assume that X has a density (as long as the context does not tell us otherwise).

Some notes on densities.

Note

1. A density f needs not be unique, but $f(x) = F'(x)$ for almost all x (meaning that, for a given f, $\{x : f(x) \neq F'(x)\}$ has Lebesgue measure 0).

2. If $f : \mathbb{R} \to \mathbb{R}$ is such that

   - f is nonnegative
   - $\int_{-\infty}^{\infty} f(t)dt = 1$.

   then f is a density of a continuous random variable X and vice-versa. If $h \geq 0$ and $\int_{-\infty}^{\infty} h(t)dt = k \neq 1$, then setting $f = h/k$ will always yield a density.

3. The density in the absolutely continuous case and the probability mass function in the case discrete case play similar roles, hence the notation .

Example 1.3

The exponential distribution $\mathcal{E}(1)$ has a density given by $f(x) = F'(x) = e^{-x}$ if $x > 0$ and 0 otherwise.

Definition 1.4 (Support)

The support of a random variable X is the set

$$S = \{x \in \mathbb{R} : P(x - \varepsilon < X < x + \varepsilon) > 0 \; \forall \, \varepsilon > 0\}$$

Remark

The support is always a closed set. In most cases we will discuss, S will be of the form $S = \overline{\{x \in \mathbb{R} : f(x) > 0\}}$. This is possible since we are dealing with density, there is no mass point and $\forall \, x$, $\mathsf{P}(X = x) = \int_x^x f(t)dt = 0$.

Proposition 1.5 (Examples of univariate continuous distributions)

1. **Uniform distribution**, denoted $\mathcal{U}(a, b)$ with density given by

$$f(x) = \frac{1}{b - a}, \quad x \in [a, b].$$

If $X \sim \mathcal{U}(0,1)$ and we set

$$Y = \begin{cases} 1 & \text{if } X \leq p \in (0,1) \\ 0 & \text{if } X > p \end{cases},$$

then $\mathsf{P}(Y=1) = \mathsf{P}(X \leq p) = p$ and $Y \sim \mathcal{B}(p)$.

2. **Exponential distribution**, denoted $\mathcal{E}(\alpha)$ for $\alpha > 0$. The density is given by

$$f(x) = \alpha e^{-\alpha x}, \quad x \in (0, \infty).$$

This distribution is well-known for its memoryless property. Suppose we are interested in

$$\begin{aligned} \mathsf{P}(X \in (x, x+\varepsilon] | X > x) &= \frac{1 - e^{-\alpha(x+\varepsilon)} - 1 + e^{-\alpha x}}{e^{-\alpha x}} \\ &= 1 - e^{-\alpha \varepsilon} \\ &= \mathsf{P}(X \leq \varepsilon). \end{aligned}$$

3. **Gamma distribution**, $\mathcal{G}(\alpha), \alpha > 0$. Recall the Gamma function given by

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt. \tag{1.3}$$

We have that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ and $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$. The Gamma distribution has density

$$f(x) = \frac{x^{\alpha-1} e^{-x}}{\Gamma(\alpha)}, \quad x > 0.$$

A special case of the Gamma is when $\alpha = 1$, in which case we recover the Exponential distribution with parameter 1. The Gamma distribution can be extended to a more general form. If we have $\mathcal{G}(\alpha, \beta)$, then α is the shape parameter, $\beta > 0$ is the scale parameter and the distribution is in this case is

$$f(x) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^\alpha \Gamma(\alpha)}, \quad x > 0.$$

When $\alpha = 1$, we get $\mathcal{E}(1/\beta)$ and when $\alpha = 1/2, \beta = 2$ we get $\chi^2(\nu)$ distribution.

4. **Normal distribution**, denoted $\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0$. The density is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

For any μ, σ^2 then the probability of being one, two or three standard deviations away from the mean is

$$\begin{aligned}\mathsf{P}(X \in [\mu - \sigma, \mu + \sigma]) &\approx 68\% \\ \mathsf{P}(X \in [\mu - 2\sigma, \mu + 2\sigma]) &\approx 95\% \\ \mathsf{P}(X \in [\mu - 3\sigma, \mu + 3\sigma]) &\approx 98\%.\end{aligned}$$

5. **Student-t distribution**, denoted $t(\nu), \nu > 0$, with density

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{(\nu+1)}{2}}, \qquad t \in \mathbb{R}.$$

An interesting special case of the Student-t distribution is the Cauchy distribution, which arises when $\nu = 1$, yielding

$$f(x) = \frac{1}{\sqrt{\pi}\Gamma\left(\frac{1}{2}\right)} \frac{1}{1 + x^2} = \frac{1}{\pi(1 + x^2)}.$$

Remark (Other univariate distributions)

Beware that not all univariate distributions have a PMF or a density (PDF). If F is a pure jump function, then X is discrete. If X has a PDF, then F is continuous and in particular $\mathsf{P}(X = x) = F(x) - F(x-) = 0$.

Sometimes the random variable falls in neither of these categories.

Example 1.4

Let X be the lifetime of a lightbulb and $X \sim \mathcal{E}(1)$. Suppose we observe the lightbulb for only 1 year, so that we actually see $X^\star = \min(X, 1)$. The CDF of $X^\star$ will be

$$F^\star(x) = \mathsf{P}(X^\star \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ \mathsf{P}(X \leq x) = 1 - e^{-x} & \text{if } x \in [0, 1) \\ 1 & \text{if } x \geq 1. \end{cases}$$

Here, $F^\star$ is neither continuous nor is a pure jump function. Observe that $\mathsf{P}(X^\star = 1) = e^{-1}$. In this case, we say that the point 1 is an **atom** of the distribution $F^\star$. In other words, $X^\star$ is neither discrete nor does it have a density. In this case, we have to work directly with the CDF. A naive idea would be to define a density like

$$(F^\star)(x) = \begin{cases} 0 & \text{if } x \leq 0, x \geq 1 \\ e^{-x}, & \text{if } x \in (0, 1). \end{cases}$$

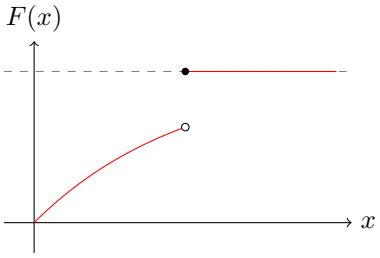


Figure 1: CDF of $X^\star$

But this doesn't work as $\int_{-\infty}^{\infty}(F^\star)(x)dx < 1$.

Any CDF F can be rewritten in the form

$$F(x) = p_1F_1(x) + p_2F_2(x) + p_3F_3(x)$$

where $p_i \in [0,1]$ such that $p_1 + p_2 + p_3 = 1$ and F_1, F_2, F_3 are themselves CDFs such that F_1 is discrete (pure jump function with countably many discontinuities), F_2 is absolutely continuous (having a density), F_3 is a singular CDF (is continuous, but has no density). One such example of CDF corresponding to F_3 is the Devil's staircase, which is continuous but has derivative equal to zero almost everywhere. F_3 is less rare in the case of random vectors. This decomposition is the so-called Lebesgue decomposition.

Example 1.5
Coming back to our example with a rounded exponential, we could set

- $F_1(x) = \mathbf{1}(x \geq 1)$ and $p_1 = e^{-1}$
- $F_2(x)$ has density

$$f_2(x) = \begin{cases} 0, & \text{if } x < 0 \\ e^{-x}/(1-e^{-1}), & \text{if } x \in (0,1) \end{cases}$$

  with $p_2 = 1 - e^{-1}$.

Thus, we can write $F(x) = p_1F_1(x) + p_2\int_{-\infty}^{x} f_2(t)dt$.

Section 1.3. Random vectors

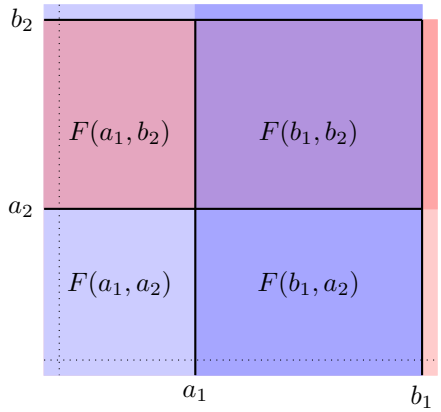
A random vector is a mapping $\boldsymbol{X} : (\Omega, \mathcal{A}, \mathsf{P}) \to (\mathbb{R}^d, \mathbb{B}^d)$, where $\boldsymbol{X} = (X_1, \ldots, X_d)$ where each X_i are random variables.

Definition 1.6 (Cumulative distribution function)
A cumulative distribution function $F_{\boldsymbol{X}}$ of $\boldsymbol{X}$ is a mapping $F : \mathbb{R}^d \to [0,1]$ given by

$$F(x_1, \ldots, x_d) = \mathsf{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_d \leq x_d) = \mathsf{P}\left(\bigcap_{j=1}^{d} \{X_j \leq x_j\}\right).$$

Working with F in $d = 2$, then $\mathsf{P}(X_1 \in (a_1, b_1], X_2 \in (a_2, b_2])$, which using the inclusion-exclusion principle is equal to $F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2)$. These half-open intersections are intersection-stable and generate the Borel σ-field. This motivates the following

Figure 2: Viewing 2−monotonicity with sets



The above express $\mathsf{P}(A \cup B \cup C \cup D) - \mathsf{P}(A \cup C) - \mathsf{P}(A \cup B) + \mathsf{P}(A)$ since they are disjoint sets. In terms of areas, this is $(A+B+C+D)-(A+C)-(A+B)+A = D$ hence $\mathsf{P}(x_1 < X < x_1^*, x_2 < X_2 < x_2^*)$. Without (4), we could build F and get a defective d.

Theorem 1.7
A CDF determines the **distribution** of a random vector $\boldsymbol{X}$ uniquely.

As in the univariate case, we will need to impose some conditions for a function F to be a CDF

Theorem 1.8
A function F is a CDF of some random vector $\boldsymbol{X}$ if and only if

1. $F(x_1, \ldots, x_d) \to 0$ if at least one $x_i \to -\infty$.
2. $F(x_1, \ldots, x_d) \to 1$ if $x_i \to \infty$ for all $i \in \{1, \ldots, d\}$.

3. Right-continuity: F is "right-continuous" in the sense that

$$\lim_{\substack{t_i \to x_i \\ t_i \geq x_i}} F(t_1, \ldots, t_d) = F(x_1, \ldots, x_d) \qquad \forall\, i \in \{1, \ldots, d\}.$$

4. All the open boxes for the open intervals is non-decreasing; in the case $d = 2$, $\forall\, a_1 \leq b_1, a_2 \leq b_2$, we need

$$F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2) \geq 0.$$

For generally, for arbitrary d $\forall\, a_1 \leq b_1, \ldots, a_d \leq b_d$

$$\sum_{\text{vertices of } (a_1, b_1 \times (a_d, b_d]} F(c_1, \ldots, c_d)(-1)^{v(c)} \geq 0$$

where $c_i \in \{a_1, b_i\}$ and where $v(c) = |\{\#i : c_i = a_i\}|$. [1]

Remark

The above has the same interpretation as

$$\sum_{\substack{c_i \in \{a_i, b_i\} \\ i = 1, \ldots, d}} (-1)^{v(c)} F(c_1, \ldots, c_d) = \mathsf{P}(X_1 \in (a_1, b_1], \ldots, X_d \in (a_d, b_d])$$

Property (4) is called delta-monotonicity or quasi-monotonicity.

The last condition may fail much more often than in the univariate case.

Case 1: $\boldsymbol{X}$ is discrete

We have $S = \{(x_1, \ldots, x_d) : \mathsf{P}(X_1 = x_1, \ldots, X_d = x_d) > 0\}$ is at most countable and

$$\sum_{(x_1, \ldots x_d) \in S} \mathsf{P}(X_1 = x_1, \ldots, X_d = x_d) = 1.$$

The distribution of $\boldsymbol{X}$ is completely characterized by its "density" (probability mass function), which is a function $f : \mathbb{R}^d : [0, \infty)$ and

$$f(x_1, \ldots, x_d) = \mathsf{P}(X_1 = x_1, \ldots, X_d = x_d).$$

Example 1.6 (Multinomial distribution)

We have n objects, $d + 1$ categories, each object falls into category j with probability

[1] In dimension one, we needed that F be non-decreasing, we needed $\mathsf{P}(X \in (a, b]) \geq 0$ as to compute this as $F(b) - F(a)$.

$p_j \in [0,1]$ with $p_1 + \dots p_{d+1} = 1$. This comes up often in biostatistics. We can count the number of objects falling in categories $1, \dots, d$, so that $\boldsymbol{X} = (X_1, \dots, X_d)$ has PMF given by

$$\mathsf{P}(X_1 = n_1, \dots, X_d = n_d) = \frac{n!}{n_1! \cdots n_d! n_{d-1}!} p_1^{n_1} p_d^{n_d} p_{d+1}^{n_{d+1}}$$

where $n_1, \dots, n_d \leq n$ and $n_1, \dots, n_d \geq 0$.

Multinomial distribution ($d = 2$) has three parameters $\mathcal{M}(n, p_1, p_2)$ and the probability mass function is

$$f(x,y) = \frac{n!}{x!y!(n-x-y)!} p_1^x p_2^y (1 - p_1 - p_2)^{n-x-y}$$

where $x, y \in \{0, \dots, n\}, x + y \leq n$.

We next show a construction due to Albert Marshall and Ingram Olkin, who wrote a book about stochastic ordering and wrote in 1983 a paper on "A Family of Bivariate Distributions Generated by the Bivariate Bernoulli Distribution."

Proposition 1.9 (Construction of bivariate distributions)
Start with a Bivariate Bernoulli with a random pair (X, Y), with probabilities

$$\mathsf{P}(X = 0, Y = 0) = p_{00} \qquad \mathsf{P}(X = 1, Y = 0) = p_{10}$$
$$\mathsf{P}(X = 0, Y = 1) = p_{01} \qquad \mathsf{P}(X = 1, Y = 1) = p_{11}$$

where $p_{00}, p_{01}, p_{10}, p_{11} \in (0,1)$ and $p_{00} + p_{10} + p_{01} + p_{11} = 1$. Clearly, $X \sim \mathcal{B}(p_{10} + p_{11})$ and $Y \sim \mathcal{B}(p_{01} + p_{11})$. Look at the bivariate binomial we could take a sequence of bivariate Bernoulli, that is $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent $\mathcal{B}_2(p_{00}, p_{01}, p_{10}, p_{11})$. One can set $X = \sum_{i=1}^n X_i, Y = \sum_{i=1}^n Y_i$. For example, if we have

$$\binom{0}{1} \quad \binom{1}{1} \quad \binom{0}{0} \quad \binom{0}{0} \quad \binom{0}{1} \quad \binom{1}{0}$$

and taking the numerator for X and Y for the denominator, we get $X = 2, Y = 3$. (X, Y) has a bivariate Binomial distribution. We have

$$\begin{aligned} f(k,l) &= \mathsf{P}(X = k, Y = l) \\ &= \sum_{a=\max(0,k+l-n)}^{\min(k,l)} \frac{n!}{a!(k-a)!(l-a)!(n-k-l+a)!} p_{11}^a p_{10}^{k-a} p_{01}^{l-a} p_{00}^{n-k-l+a} \end{aligned}$$

with $a = \#\binom{1}{1}$ and $k - a$ such $\binom{1}{0}$, $l - a$ of $\binom{0}{1}$ and the remaining number as $\binom{0}{0}$ and where $k, l \in \{0, \dots, n\}$. For higher dimensions, we need to specify 2^d probabilities and it

rapidly becomes intractable. One can also verify that the distribution of X or Y alone are binomial.

For the bivariate Poisson, we can view the Poisson as the limit of a Binomial, having $np_{01} \to \lambda_{01}, np_{11} \to \lambda_{11}, np_{10} \to \lambda_{10}$ and $\lambda_{01}, \lambda_{10}, \lambda_{11} \in (0, \infty)$. and we can show that as $n \to \infty$, one obtains

$$f(k,l) = \sum_{a=0}^{\min(k,l)} \frac{\lambda_{11}^a \lambda_{10}^{k-a} \lambda_{01}^{l-a}}{a!(k-a)!(l-a)!} e^{-\lambda_{11}-\lambda_{01}-\lambda_{10}} \qquad k, l \in \mathbb{N}$$

which is a Bivariate Poisson distribution, it is possible to show that $Z_1 \sim \mathcal{P}(\lambda_{10}), Z_2 \sim \mathcal{P}(\lambda_{01}), Z_3 \sim \mathcal{P}(\lambda_{10})$ and $X = Z_1 + Z_3, Y = Z_2 + Z_3$ and we take the pair (X, Y). The thunderstorm in Quebec in Montreal, some could be jointly happening in either or both.

As an exercise, try to construct the Bivariate Geometric, where X denote the waiting time for "1" in the first component and Y for "1" in the second component.

Just getting a multivariate distribution with constraints by just writing down the PMF, this may not be easy to do. We will see other techniques later in the course.

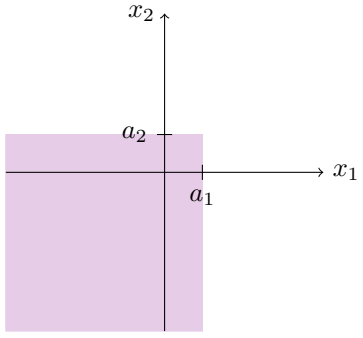
We will now conclude our survey of random vectors with cases where we have continuous random vectors.

Case 2: $\boldsymbol{X}$ is continuous (that is F, the CDF of $\boldsymbol{X}$ is a continuous function. $\boldsymbol{X}$ is continuous if and only if $\mathsf{P}(X_1 = x_1, \ldots, X_d = x_d) = 0$ for all $x_1, \ldots, x_d \in \mathbb{R}$. We say that $\boldsymbol{X}$ has density if there exists $f : \mathbb{R}^d \to [0, \infty)$ such that f is measurable and such that $\forall\, x_1, \ldots, x_d \in \mathbb{R}$,

$$F(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(t_1, \ldots, t_d) d_{t_d} \ldots d_{t_1}$$

as a d-fold integral.

Figure 3: CDF $F(a_1, a_2)$ for a distribution



If F has a density, then

$$f(x_1, \ldots, x_d) = \frac{\partial F}{\partial x_1 \ldots \partial x_d} \text{ almost surely}$$

If F has density f, then F is continuous but **not** conversely. [2]

Example 1.7

Let

$$F(x_1, x_2) = 1 - e^{-2x_1} - e^{-2x_2} + e^{-2x_1 - 2x_2} = (1 - e^{-2x_1})(1 - e^{-2x_2})$$

when $x_1, x_2 > 0$. What about the density? If we look at

$$\frac{\partial F}{\partial x_1 \partial x_2} = 4e^{-2x_1 - 2x_2} = f(x_1, x_2)$$

for $x_1, x_2 \geq 0$ and indeed,

$$F(x_1, x_2) = \int_0^{x_1} \int_0^{x_2} 4e^{-2t_1} e^{-2t_2} dt_2 dt_1.$$

Let us now take some more interesting cases

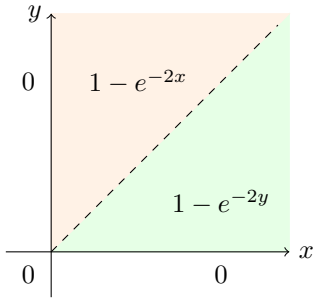
Example 1.8

Take $Z \sim \mathcal{E}(2)$ and set $X = Z, Y = Z$. If we want to obtain the

$$\begin{aligned} \mathsf{P}(X \leq x, Y \leq y) &= \mathsf{P}(Z \leq x, Z \leq y) = \mathsf{P}(Z \leq \min(x, y)) \\ &= 1 - e^{-2\min(x,y)}, \qquad x, y \geq 0 \end{aligned}$$

and this is a valid distribution function, we could also sketch the distribution.

Figure 4: Illustration of $\mathsf{P}(Z \leq \min(X, Y))$



It has a kink at this line, but we are interested in its existence only almost everywhere. It

[2] The order of partials does not matter in this case. Note that this is **only** an implication.

does not exist along the line $x = y$. Our only candidate is 0 everywhere, so this distribution function has no density.

Example 1.9 (Marshall-Olkin distribution)
This distribution describes the lifetime of some component, e.g. two engine airplane. We take $Z_1 \sim \mathcal{E}(1)$, which is the failure time of the first engine only, and Z_2 independent of Z_1, which is again exponential with parameter 1 for only the second engine fails. $Z_3 \sim \mathcal{E}(1)$ is independent of Z_1 and Z_2, which denotes the event for failure of both engines.

Take $X = \min(Z_1, Z_3)$ to be the lifetime of the first engine, $Y = \min(Z_2, Z_3)$ the lifetime of the second engine. If we want to describe the joint distribution of X, Y and

$$\begin{aligned}
&\mathsf{P}(X \le x, Y \le y) = 1 - \mathsf{P}(X > x) - \mathsf{P}(Y > y) + \mathsf{P}(X > x, Y > y) \\
&= 1 - \mathsf{P}(\min(Z_1, Z_3) > x) - \mathsf{P}(\min(Z_2, Z_3) > y) + \mathsf{P}(\min(Z_1, Z_3) > x, \min(Z_2, Z_3) > y) \\
&= 1 - \mathsf{P}(Z_1 > x)\mathsf{P}(Z_3 > x) - \mathsf{P}(Z_2 > y)\mathsf{P}(Z_3 > y) + \mathsf{P}(Z_1 > x, Z_3 > x, Z_2 > y, Z_3 > y) \\
&= 1 - e^{-2x} - e^{-2y} - e^{-x-y-\max(x,y)}
\end{aligned}$$

only if $x, y > 0$.

Figure 5: Marshall-Olkin distribution: $\mathsf{P}(X \le x, Y \le y)$

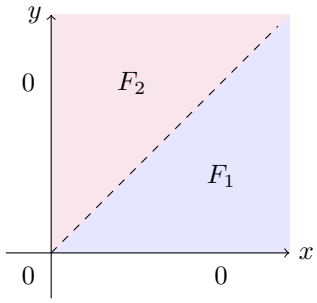


Illustration of values of the CDF for

$$\begin{aligned}
F_1 &= 1 - e^{-2x} - e^{-2y} - e^{-2x-y} \\
F_2 &= 1 - e^{-2x} - e^{-2y} - e^{-x-2y}.
\end{aligned}$$

Let's differentiate and see what happens. The distribution is continuous

$$\begin{aligned}
\frac{\partial F_1}{\partial x \partial y} &= 2e^{-2x-y} \\
\frac{\partial F_2}{\partial x \partial y} &= 2e^{-x-2y}
\end{aligned}$$

If we integrate it out,

$$\int_0^\infty \int_0^\infty f(t_1,t_2)dt_1dt_2 = \int_0^\infty \int_0^\infty 2e^{-2t_1-t_2}dt_1dt_2 + \int_0^\infty \int_0^\infty 2e^{-t_1-2t_2}dt_1dt_2 = \ldots = \frac{2}{3}$$

This is an example of a CDF F which is continuous, but does not have a density! This is because there is probability that the first event to happen is the joint failure

R illustration of the three cases, as a line has measure zero for the second case, the distribution is singular and we only see realization concentrated on the line (no area). This is also the set of points for which the derivative did not exist.

The third example looks like a blend on the two. There is a clustering of points along this line, where the common shock happens first. The first case is absolutely continuous, the second is singular and the third is neither.

Proposition 1.10 (Multivariate Normal)
This distribution is denoted $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = \mathbf{0}$. and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is the variance-covariance matrix which has to be positive definite (for the density to exist) and symmetric. In the standard case, $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$, namely the identity matrix and the density is given by

$$f(x_1,\ldots,x_d) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}(x_1^2+\cdots+x_d^2)} = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{x}\right)$$

and in the more general case

$$f(x_1,\ldots,x_d) = \frac{1}{2\pi^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

Note that the variance-covariance matrix has the form for $d = 2$ given by $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

Proposition 1.11 (Multivariate t)
Standard case is where $\forall\, x_1,\ldots,x_d \in \mathbb{R}, \nu > 0$, we have

$$f_{\boldsymbol{X}}(x_1,\ldots,x_d) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{\frac{d}{2}}\pi^{\frac{d}{2}}} \left[1+\frac{1}{\nu}\boldsymbol{x}^\top\boldsymbol{x}\right]^{-\frac{\nu+d}{2}}$$

In the more general case, we have $\boldsymbol{\mu} = (\mu_1 \cdots \mu_d)^\top$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is a positive definite and symmetric with joint density given by

$$f_{X_1,\ldots,X_d}(x_1,\ldots,x_d) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{\frac{d}{2}}\pi^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \left[1+\frac{1}{\nu}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{-\frac{d+\nu}{2}}$$

Chapter 2
Marginals, conditional distributions and dependence

Section 2.1. Marginal distributions

The univariate margins of $\boldsymbol{X}$ (or F) are the CDF of the individual random variables $X_1, \ldots, X_d$. The CDF of X_i can be obtained from

$$F_i(x) = \lim_{\substack{x_j \to \infty \\ i \neq j}} F(x_1, \ldots, x_{i-1}, x, , x_{i+1}, \ldots, x_d)$$

Using the previous examples,

Example 2.1

$$\begin{aligned} F(x_1, x_2) &= 1 - e^{-2x_1} - e^{-2x_2} + e^{-2x_1 - 2x_2} \\ \tilde{F}(x_1, x_2) &= 1 - e^{-2\min(x_1, x_2)} \\ F^*(x_1, x_2) &= 1 - e^{-2x_1} - e^{-2x_2} + e^{-x_1 - x_2 - \max(x_1, x_2)} \end{aligned}$$

where $(X_1, X_2) \in [0,1]^2$. Let $(X_1, X_2) \sim F, (X_q^*, X_2^*) \sim F^*, (\tilde{X}_1, \tilde{X}_2) \sim \tilde{F}$ and $X_1 \overset{d}{=} X_2$, $\tilde{X}_1 \overset{d}{=} \tilde{X}_2$ and $X_1^* \overset{d}{=} X_2^*$

Now we have

$$F_1(x_1) = \lim_{x_2 \to \infty} F(x_1, x_2) = \begin{cases} 1 - e^{-2x_1} & \text{if } x_1 > 0 \\ 0 & \text{if } x_1 < 0 \end{cases}$$

and notice that we have $F_1(x_1) = \lim_{x_2 \to \infty} F(x_1, x_2) = F_1^*(x_1^*) = \lim_{x_2^* \to \infty} F^*(x_1^*, x_2^*)$ and equal to $\tilde{F}_1(\tilde{x}_1) = \lim_{\tilde{x}_2 \to \infty} \tilde{F}(\tilde{x}_1, \tilde{x}_2)$; thus $X_1, X_1^*, \tilde{X}_1 \sim \mathcal{E}(2)$.

The **morale** is that F specifies the marginal CDFs $F_1, \ldots, F_d$ uniquely, but **not vice versa**.

If we were given the density rather than the CDF, we could still work our way through to obtain the margins.

Lemma 2.1

If F has density (PMF) f, then for all $i \in \{1, \ldots, d\}$, we have X_i has density (PMF) f_i

given by for all $x \in \mathbb{R}$,

$$\begin{aligned} f_i(x) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_d) dx_1 \ldots dx_d \\ &= \sum_{x_1} \cdots \sum_{x_d} f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_d) \end{aligned}$$

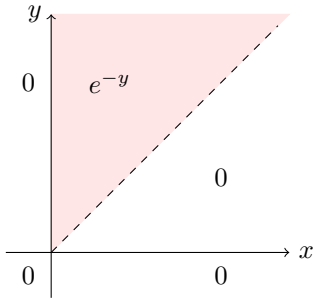
in the first case for densities, which is $d-1$ integrals, while the other case is for PMF, where the sum is over all variables x_j but not x. Note that if the joint distribution has a density, then any margin has a density but not conversely.

Example 2.2

Let

$$f(x, y) = e^{-y} \mathbf{1}_{(0<x<y<\infty)}$$

Figure 6: Illustration of the density $f(x, y) = e^{-y} \mathbf{1}_{(0<x<y<\infty)}$



We are interested in f_1, the density of X. Then

$$f_1(x) = \int_{-\infty}^{\infty} e^{-y} \mathbf{1}_{(0<x<y<\infty)} dy = \int_{x}^{\infty} e^{-y} dy = e^{-x}$$

for $x \geq 0$ so $X \sim \mathcal{E}(1)$.

For the second parameter, the density f_2 of Y is

$$f_2(y) = \int_{0}^{y} e^{-y} dx = y e^{-y}$$

for $y > 0$ and we conclude that $Y \sim \mathcal{G}(2)$, which is the Gamma distribution.

Example 2.3
Let $(X, Y) \sim \mathcal{M}(n, p_1, p_2)$ with PMF given by

$$\frac{n!}{x!y!(n-x-y)!} p_1^x p_d^y (1-p_1-p_2)^{n-x-y}$$

if $x, y \in \{0, \dots, n\}$ and $x + y \leq n$.

For the margin of X,

$$\begin{aligned} f_1(x) &= \sum_{y=0}^{n-x} \frac{n!}{x!y!(n-x-y)!} p_1^x p_2^y (1-p_1-p_2)^{n-x-y} \\ &= \frac{n!}{x!(n-x)!} p_1^x \sum_{y=0}^{n-x} \binom{n-x}{y} p_2^y (1-p_1-p_2)^{n-x-y} \\ &= \frac{n!}{x!(n-x)!} p_1^x (p_2 + 1 - p_1 - p_2)^{n-x} \\ &= \binom{n}{x} p_1^x (1-p_1)^x \end{aligned}$$

and so $X \sim \mathcal{B}(n, p_1)$ and $Y \sim \mathcal{B}(n, p_2)$.

Exercise 2.1

- Compute the margins of the bivariate Normal (can try for an easier time with $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_2)$) or the bivariate Student t
- Compute the margins of the bivariate binomial, Poisson and geometric.

Remark
We could also look at the multivariate margins. If we have a random vector $(X_1, \dots, X_d)$, suppose $1 \leq p < d$ and define $\boldsymbol{X}_p = (X_1, \dots, X_p)$ and $\boldsymbol{X}_{d-p} = (X_{p+1}, \dots, X_d)$. Then $\boldsymbol{X}_p$ has CDF

$$F_{\boldsymbol{X}_p}(x_1, \dots, x_p) = \lim_{\substack{x_i \to \infty \\ i > p}} F(x_1, \dots, x_p, x_{p+1}, \dots, x_d)$$

Section 2.2. Independence

Coming back to our morale, we look at independence. Recall that two events A and B are independent if $\mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B)$.

Definition 2.2 (Independence)
Let (X, Y) be a random vector with joint distribution function F and margins F_1, F_2. Then

X and Y are **independent** ($X \perp\!\!\!\perp Y$) if and only if $\forall\, x, y \in \mathbb{R}$,

$$F(x, y) = F_1(x)F_2(y)$$

Remark
Independence of X and Y means

$$\mathsf{P}(X \leq x, Y \leq y) = \mathsf{P}(X \leq x)\mathsf{P}(Y \leq y)$$

where the left hand side is also $\mathsf{P}(\{X \leq x\} \cap \{Y \leq y\})$. Suppose

$$\begin{aligned}\mathsf{P}(X \in [a, b], Y \in [c, d]) &= F(b, d) - F(b, c) - F(a, d) + F(a, c) \\ &= F_1(b)F_2(d) - F_1(b)F_2(c) - F_1(a)F_2(d) + F_1(a)F_2(c) \\ &= (F_1(b) - F_1(a))(F_2(d) - F_2(c))\end{aligned}$$

if $X \perp\!\!\!\perp Y$. Hence $X \perp\!\!\!\perp Y$ implies that

$$\mathsf{P}(X \in (a, b], Y \in (c, d]) = \mathsf{P}(X \in (a, b]) \times \mathsf{P}(Y \in (c, d])$$

and because $(a, b] \times (c, d]$ generated $\mathbb{B}^2$, the Borel $\sigma-$field, we can deduce that

$$X \perp\!\!\!\perp Y \to \mathsf{P}(X \in A, Y \in B) = \mathsf{P}(X \in A)\mathsf{P}(Y \in B) \quad \forall\, A, B \in \mathbb{B}^2$$

Theorem 2.3
If (X, Y) has density (or PMF) f, then $X \perp\!\!\!\perp Y$ if and only if

$$f(x, y) = f_1(x)f_2(y)$$

for almost all $x, y \in \mathbb{R}$.

Proof $X \perp\!\!\!\perp Y$ implies that $F(x, y) = F_1(x)F_2(y)$ and for almost all $x, y \in \mathbb{R}$,

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F = \left(\frac{\partial F_1}{\partial x}\right)\left(\frac{\partial F_2}{\partial y}\right)$$

which are almost surely equal to f_1, f_2. Conversely,

$$\begin{aligned}F(x, y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_1(t)f_2(s)dsdt \\ &= \int_{-\infty}^{x} f_1(t)\int_{-\infty}^{y} f_2(s)dsdt \\ &= F_1(x)F_2(y)\end{aligned}$$

■

Example 2.4

Take the uniform density on a circle

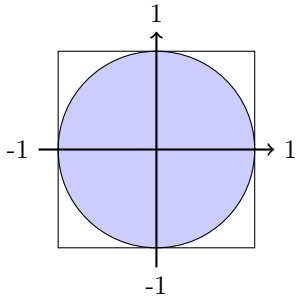
$$f(x,y) = \frac{1}{\pi}\mathbf{1}_{(x^2+y^2\leq 1)}$$

Suppose that (X,Y) has density f, then X and Y is not independent. If one takes a small neighborhood of a point in the unit square, not in the circle,

$$f_1(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} dy = \frac{2}{\pi}\sqrt{1-x^2}$$

with $x \in [-1,1]$. [3]

Figure 7: Unit circle and unit square



Using symmetry, we have that $f_2(y) = 2\pi^{-1}\sqrt{1-y^2}$ for $y \in [-1,1]$. Observe that if the

$$\text{Support}(X,Y) \neq \text{Support}(X) \times \text{Support}(Y),$$

then X and Y cannot be independent.

Example 2.5

Consider

$$f(x,y) = e^{-y}\mathbf{1}_{(0<x<y<\infty)}$$

with $f_1(x) = e^{-x}$ for $x > 0$ and $f_2(y) = ye^{-y}$ for $y > 0$.

Definition 2.4

$X_1, \ldots, X_d$ are independent if and only if $\forall\ n \in \{1,\ldots,d\}$, $\forall\ i_1,\ldots,i_n \in \{1,\ldots,d\}$ distinct, then $X_{i_1}, X_{i_2}, \ldots, X_{i_n}$ are independent. This is the case if and only if the CDF of

[3] Recall the support is the range where the density is positive or a small neighborhood of a point will have positive probability.

$(X_1, \ldots, X_d)$ satisfies for $x_1, \ldots, x_d \in \mathbb{R}$, then

$$F(x_1, \ldots, x_d) = \prod_{i=1}^{d} F_i(x_i)$$

Example 2.6
In the trivariate case, we have that (X, Y, Z) has independent component iff $X \perp\!\!\!\perp Y, Y \perp\!\!\!\perp Z, Z \perp\!\!\!\perp X$ and $X \perp\!\!\!\perp Y \perp\!\!\!\perp Z$

Example 2.7 (Trivariate FGM distribution)
Consider a Farlie-Gumbel-Morgenstern distribution of the form

$$F(x, y, z) = xyz + \theta x(1-x)y(1-y)z(1-z)$$

where $x, y, z \in [0, 1]$ and $\theta \in [-1, 1]$ Then $F_1(x) = x$ if $x \in [0, 1]$, $F_2(y) = y$ if $y \in [0, 1]$ and $F_3(z) = z$ if $z \in [0, 1]$. and $F_{X,Y}(x, y) = xy$ for $x, y \in [0, 1]$. But (X, Y, Z) are not independent.

Definition 2.5
$X_1, X_2, \ldots$ is an independent sequence if $\forall\, n \in \mathbb{N}$, for $i_1 \neq \cdots \neq i_n \in \{1, 2, \ldots\}$, $X_{i_1}, \ldots, X_{i_n}$ are independent.

Section 2.3. Conditional distributions

One question that arises is when X and Y are not independent, what then to do. One has two options, namely

- conditional distributions
- copulas

As motivation, consider height and ages. These variables are clearly related (childs are smallers, adults shrink). One could decide to look at height for any given age, which brings us to conditional probabilities.

Definition 2.6 (Conditional probability)
If A, B are such that $\mathsf{P}(B) > 0$, then the conditional probability of A given B is

$$\mathsf{P}(A|B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}$$

If we measure age in years and height in centimeters, we could look at X, Y (discrete) and calculate

$$\mathsf{P}(X = x, Y = y) = \frac{\mathsf{P}(X = x, Y = y)}{\mathsf{P}(Y = y)} = \frac{f(x, y)}{f_2(y)},$$

also denoted $f(x|y)$ for all y such that $\mathsf{P}(Y = y) > 0$.

Observation

For any fixed $y \in \mathbb{R}$ so that $\mathsf{P}(Y = y) > 0$, then $f(x|y)$ for $x \in \mathbb{R}$ is a PMF as

1. $f(x|y) \geq 0$
2. $$\sum_{x \in \mathbb{R}} \frac{f(x|y)}{f_2(y)} = \frac{1}{f_2(y)} \sum_x f(x, y) = 1.$$

$f(x|y)$, for $f_2(y) > 0$ is the PMF of the conditional distribution of X given $Y = y$.

Also remark that $X \perp\!\!\!\perp Y$ imply $f(x|y) = f_1(x)$, hence the orthogonality notation.

If X, Y are discrete, then

$$f_{X|Y=y}(x) = \mathsf{P}(X = x|Y = y) = \frac{f(x, y)}{f(y)} \quad \forall\, y, f_Y(y) > 0$$

and if $X \perp\!\!\!\perp Y$ are independent, this $f_{X|Y=y}(x) = f_X(x)\ \forall\ x$.

If X, Y are not discrete, then the conditional distribution of X given $Y = y$ is not so easily characterized except where (X, Y) has joint density. The **conditional CDF of** X given $Y = y$.

$$\mathsf{P}(X \leq x|Y = y) = \lim_{\varepsilon \downarrow 0} \mathsf{P}(X \leq x|Y \in (y - \varepsilon, y + \varepsilon])$$

and $\forall\, y \in \mathbb{R}$, $\mathsf{P}(Y = y) = 0$

$$= \lim_{\varepsilon \downarrow 0} \frac{\mathsf{P}(X \leq x, Y \in (y - \varepsilon, y + \varepsilon])}{\mathsf{P}(Y \in (y - \varepsilon, y + \varepsilon])}$$
$$= \lim_{\varepsilon \downarrow 0} \frac{\int_{-\infty}^{x} \int_{y-\varepsilon}^{y+\varepsilon} f(t, s) ds dt}{\int_{y-\varepsilon}^{y+\varepsilon} f_Y(t) dt}$$

which is nothing but $\frac{F(x,y+\varepsilon) - F(x,y-\varepsilon)}{F_Y(y+\varepsilon) - F_Y(y-\varepsilon)}$.

This is equal to

$$\lim_{\varepsilon \downarrow 0} \frac{\int_{-\infty}^{x} 2\varepsilon f(t, s_{\varepsilon,t}) dt}{2\varepsilon f_Y(t_\varepsilon)}$$

Now f_Y is continuous on $[y-\varepsilon, y+\varepsilon]$ and f is also continuous and as such $t_\varepsilon \in [y-\varepsilon, y+\varepsilon] \to y$ as $\varepsilon \downarrow 0$ and $s_{\varepsilon,t} \in [y-\varepsilon, y+\varepsilon] \to y$. We thus get from this heuristic calculation that

$$\frac{\int_{-\infty}^{x} f(t,y)dt}{f_Y(y)} = \int_{-\infty}^{x} \frac{f(t,y)}{f_Y(y)}dt.$$

The conditional distribution of X given $Y = y$ has density

$$f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}, \quad y \in \text{ Support}(Y)$$

where as a reminder $\text{Support}(Y) = \{y \in \mathbb{R} \text{ such that } \forall\, \varepsilon > 0, \mathsf{P}(Y \in (y-\varepsilon, y+\varepsilon)) > 0\}$.

Definition 2.7 (Conditional distribution)

Let (X, Y) be a random vector with PMF (density) f and marginal PMF (density) f_X and f_Y. Then, the conditional distribution of X given $Y = y$ has PMF (density) given by

$$f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}, \quad y \in \text{ Support}(Y)$$

and is undefined otherwise.

Example 2.8

Let $f(x,y) = e^{-y}\mathbf{1}_{(0<x<y<\infty)}$. The margins are given by

$$f_X(x) = e^{-x}\mathbf{1}_{(x>0)}$$
$$f_Y(y) = ye^{-y}\mathbf{1}_{(y>0)}$$

and

$$\begin{aligned} f_{X|Y=y}(x) &= \frac{e^{-y}\mathbf{1}_{(0<x<y<\infty)}}{ye^{-y}}, \quad y > 0 \\ &= \frac{1}{y}\mathbf{1}_{(0<x<y)} \end{aligned}$$

hence we conclude that $X|Y = y \sim \mathcal{U}(0, y)$. For the other margin,

$$\begin{aligned} f_{Y|X=x}(y) &= \frac{e^{-y}\mathbf{1}_{(0<x<y<\infty)}}{e^{-x}}, \quad x > 0 \\ &= e^{x-y}\mathbf{1}_{(y>x)} \end{aligned}$$

Exercise 2.2

Bivariate Normal determine the conditional distribution if $\boldsymbol{\mu} = (\mathbf{0}\ \mathbf{0})^\top$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

Remark

We have that $X \perp\!\!\!\perp Y \Leftrightarrow f_{X|Y=y}(x) = f_X(x)\ \forall\, y : f_Y(y) > 0$.

Section 2.4. Copula

We present next another characterization of dependence which has become popular in recent years. Here is a motivating question: given two variables, X, Y, does there exists a joint distribution such that, for example, $X \sim \mathcal{P}(\lambda)$ and $Y \sim \mathcal{G}(\alpha)$? That is, how to construct a distribution with given marginals. This is motivated by observations: in insurance, claim size is Poisson distributed, while the number of claims could be modelled using a Gamma distribution. If we are given marginal CDFs, $F_1, \ldots, F_d$, can we characterize all joint CDFs that have precisely the margins $F_1, \ldots, F_d$. This was answered following a correspondence between Kendall and Sklar and published in a subsequent paper by Sklar in 1959.

Take $F(x, y) = F_1(x)F_2(y) = C(F_1(x), F_2(y))$ and define a function $C : [0, 1]^2 \to [0, 1]$ and $(u, v) \mapsto (u, v)$. C happens to be a CDF on $[0, 1]^2$ whose margins are uniform on $[0, 1]$.

Definition 2.8 (Bivariate copula)
A bivariate **copula** is a bivariate CDF whose margins are standard uniform.

Example 2.9

- Independence copula: $C(u, v) = uv$
- Farlie-Gumbel-Morgenstern copula $C(u, v) = uv + \theta(uv)(1 - u)(1 - v)$
- Clayton copula $C(u, v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^1 1/\theta, \quad \theta \geq 0.$

Lemma 2.9
Let C be an **arbitrary copula** and F_X and F_Y arbitrary univariate CDFs. Then F given for all $x, y \in \mathbb{R}$ by

$$F(x, y) = C(F_X(x), F_Y(y))$$

is a bivariate CDF with margins F_X and F_Y

Proof

$$\begin{aligned} \lim_{x \to -\infty} F(x, y) &= \lim_{x \to -\infty} C(F_X(x), F_Y(y)) \\ &= C(0, F_Y(y)) = 0 \end{aligned}$$

and symmetrically, since copulas are always continuous functions and they are themselves distribution functions (uniform) Now take

$$\lim_{\substack{x \to \infty \\ y \to \infty}} C(F_X(x), F_Y(y)) = C(1, 1) = 1$$

and

$$\lim_{x\to\infty} C(F_X(x), F_Y(y)) = C(1, F_Y(y)) = F_Y(y)$$
$$\lim_{y\to\infty} C(F_X(x), F_Y(y)) = C(F_X(x), 1) = F_X(x).$$

Now for quasi-monotonicity: if we let $x_1 \leq x_2, y_1, \leq y_2$, which corresponds to the the set $(x_1, x_2] \times (y_1, y_2]$ and

$$\begin{aligned}
&F(x_1, y_1 - F(x_1, y_2) - F(x_2, y_1) + F(x_2, y_2) \\
&= C(F_X(x_1), F_Y(y_1)) - C(F_X(x_1), F_Y(y_2)) - C(F_X(x_2), F_Y(y_1)) + C(F_X(x_2), F_Y(y_2)) \\
&= C(u_1, v_1) - C(u_1, v_2) - C(u_2, v_1) + C(u_2, v_2)
\end{aligned}$$

because F_X, F_Y are non-decreasing, we have that $F_X(x_1) = u_1 \leq F_X(x_2) = u_2$ and $F_Y(y_1) = v_1 \leq F_Y(y_2) = v_2$ and thus ■

Theorem 2.10 (Sklar's representation)
Let F be a bivariate CDF with margins F_X, F_Y. Then, there exists at least one copula C such that for all $x, y \in \mathbb{R}$

$$F(x, y) = C(F_X(x), F_Y(y)).$$

Moreover, C is unique on Range$(F_X)\times$Range(F_Y). In other words, C is unique if F_X and F_Y are continuous.

Algorithm 2.1 (Generation of sample from copula)

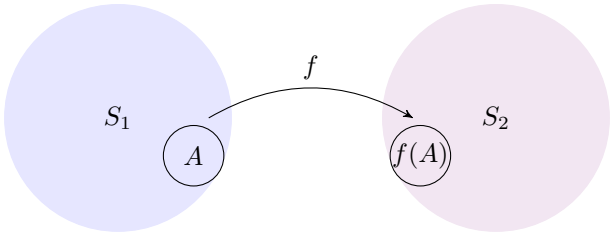
1. To generate (X, Y) from $F(x, y) = C(F_X, F_Y)$
   (a) (U, V) with CDF C
   (b) Set $X = F_X^{-1}(U)$ and $Y = F_Y^{-1}(V)$.
2. To show Sklar's representation
   (a) $(X, Y) \sim F$
   (b) C is a CDF of $(F_X(X), F_Y(Y))$

$(X, Y) \mapsto (F_X(x), F_Y(y))$. We will show why $F_X(x)$ and $F_Y(y)$ lives on the unit square and is uniform using probability integral transform. Similarly, we can take uniform margins and apply the quantile function to those margins to get the desired CDFs, namely $(U, V) \mapsto (F_X^{-1}(U), F_Y^{-1}(V))$.

Chapter 3
Univariate Transformations

Let X be a random variable and $g(X)$ be a function, applying to each realization this mapping. We want to compute the distribution of $g(X)$, which is a new random variable.

Consider the mapping $f : S_1 \to S_2$, where S_1, S_2 are two spaces.



We could look at the image set; for $A \subseteq S_1$, we can define the **image set** of A, $f(A)$ by

$$f(A) = \{f(a), a \in A\}$$

For example, taking $S_1, S_2 = \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ and $x \mapsto e^x$. Then $f(0,1) = (e^0, e^1) = (1, e)$. For $B \subseteq S_2$, we can define the so-called **pre-image set** $f^{-1}(B)$

$$f^{-1}(B) = \{s_1 \in S_1 : f(s_1) \in B\}.$$

One must be careful, as f, which is arbitrary, does not need to have an inverse. In our example, we have $f^{-1}((-4, 0)) = \emptyset$. Calculating the distribution of $g(X)$ boils down to calculating the pre-image of g.

Now, one may ask if $g(X)$ is a valid random variable (that is if $g(X)$: $\Omega \to \mathbb{R}$ is measurable). This is true if and only if $g : \mathbb{R} \to \mathbb{R}$ is measurable.

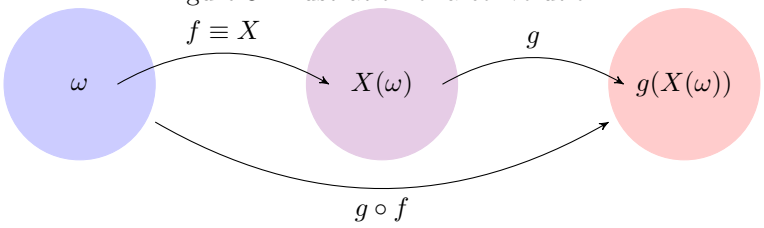
If we want to compute

$$\begin{aligned} \mathsf{P}(g(X) \leq y) &= \mathsf{P}(\omega : g(X(\omega)) \leq y) \\ &= \mathsf{P}(\omega : g(X(\omega)) \in (-\infty, y]) \\ &= \mathsf{P}(\omega : X(\omega) \in g^{-1}((-\infty, y]) \end{aligned}$$

where $(-\infty, y] \equiv B$. We can compute this only if $g^{-1}((-\infty, y]) \in \mathbb{B}$ if and only if g is measurable. [4]

[4] The mapping needs not be monotone or nice; the pre-image could be an interval, but it could also be an ugly set and one might need to perform case-by-case analysis. If g is monotone or piecewise monotone, we will have an easy time, while if g is arbitrary, this might be less the case.

Figure 8: Illustration of a convolution



Example 3.1

Let $X \sim \mathcal{B}(n,p)$ and consider the random variable $Y = n - X$, which corresponds to the number of failures. We will show that $Y \sim \mathcal{B}(n, 1-p)$. The support of Y will be the set $\{0,\dots,n\}$, hence Y is discrete. If we compute $\mathsf{P}(Y=k) = \mathsf{P}(\omega : X(\omega) \in g^{-1}\{y\})$

$$\begin{aligned}\mathsf{P}(Y=k) &= \mathsf{P}(n-X=k) = \mathsf{P}(X=n-k)\\ &= \binom{n}{n-k} p^{n-k}(1-p)^k\\ &= \binom{n}{k} p^{n-k}(1-p)^k\end{aligned}$$

so that $g : k \mapsto n-k$ and $Y \sim \mathcal{B}(n,1-p)$.

Example 3.2

Let $X \sim \mathcal{U}(0,1)$ and $g(x) = -\log(x)$ so $Y = g(X) = -\log(X)$. Now,

$$\begin{aligned}\mathsf{P}(Y \le y) &= \mathsf{P}(-\log(X) \le y)\\ &= \mathsf{P}(\log X \ge -y)\\ &= \mathsf{P}(X \ge e^{-y})\\ &= 1 - \mathsf{P}(X \le e^{-y})\end{aligned}$$

with $g^{-1}((-\infty, y]) = [e^{-y}, \infty)$. Thus, we have

$$\mathsf{P}(Y \le y) = \begin{cases} 1 - e^{-y} & \text{if } y > 0\\ 0 & \text{if } y \le 0.\end{cases}$$

as $\mathsf{P}(X \le x) = x$ if $x \in [0,1)$ and 1 if $x \ge 1$. We conclude that $Y \sim \mathcal{E}(1)$

Example 3.3

Take $X \sim \mathcal{E}(1)$ and $g : [0,\infty) \to \mathbb{Z}^+$ so that $x \mapsto \lfloor x \rfloor$.

If we are interested in $Y = g(X)$, we see that $\text{Support}(Y) = \{0,1,\dots\}$ hence Y is discrete and we can spare our life by computing the PMF, $\mathsf{P}(Y=k)$, for $k \in \mathbb{Z}^+$. This happens

when $\mathsf{P}(X \in [k, k+1))$, where $[k, k+1) = g^{-1}(\{k\})$. Notice that X is continuous, so we can compute

$$\begin{aligned}\mathsf{P}(k < X \leq k+1) &= \mathsf{P}(X \leq k+1) - \mathsf{P}(X \leq k) \\ &= F(k+1) - F(k) \\ &= 1 - e^{-(k+1)} - 1 - e^{-k} \\ &= e^{-k} - e^{-k-1} \\ &= e^{-k}(1 - e^{-1})\end{aligned}$$

and $Y \sim \mathcal{G}(1 - e^{-1})$, that is Y has a geometric distribution. We can also generalize the example by considering $\mathcal{E}(\lambda)$, which would yield $Y \sim \mathcal{G}(1 - e^{-\lambda})$. Thus, the geometric distribution can be viewed as a rounding of the exponential distribution. This is also why we saw the exponential distribution shape on the histogram and why both distributions share the memoryless property.

Section 3.1. Probability integral and quantile transformation

We look at $X \sim F$ and apply the transformation to get $F(X)$ and take $U \sim \mathcal{U}(0, 1)$ and consider $F^{-1}(U)$. For a CDF F, the so-called **quantile function** F^{-1} is given by[5]

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}, \quad u \in [0, 1].$$

Recall the definition of infimum, if we have $A \subseteq \mathbb{R}$. Then $\inf(A)$ is a real number, including $\pm\infty$, such that

- $\forall\, a \in A : \inf(A) \leq a$;
- if $x > \inf(A)$, $\exists a \in A$ such that $a < x$. This is equivalent to having for all $x \in \mathbb{R}$ such that $x \leq a\ \forall\, a \in A \Rightarrow x \leq \inf(A)$.

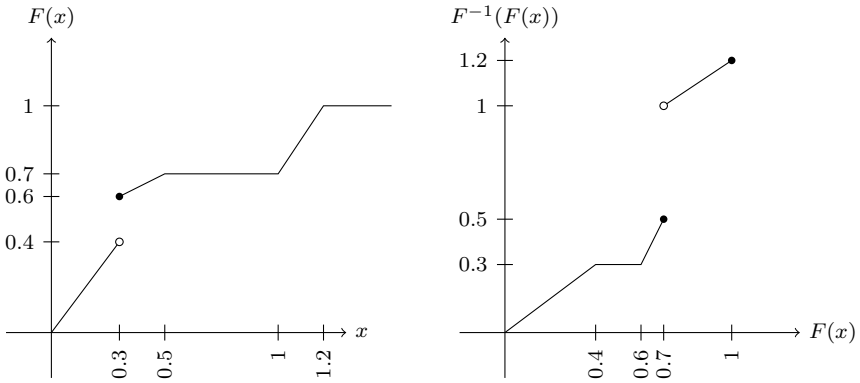
For example, if we take $[0, 1]$, then $\inf(A) = \min(A) = 0$ while if $A = (0, 1]$, then $\inf(A) = 0$. By convention, we have $\inf(\mathbb{R}) = -\infty$ and $\inf(\emptyset) = \infty$.

The function is non-decreasing, will be continuous and strictly increasing if the function is continuous, a plateau of F will cause a jump in F^{-1} and a jump in F will cause a plateau for F^{-1}.

Lemma 3.1

Let F be an arbitrary univariate CDF and F^{-1}, its quantile function. Then

[5] Remark that $F^{-1}(0) = -\infty$



1. F^{-1} is non-decreasing and left-continuous. If F is continuous, then F^{-1} is strictly increasing (not necessarily continuous though).

2. $F^{-1} \circ F(x) \leq x, \ \forall \ x \in \mathbb{R}$. If F is strictly increasing, then equality holds.

3. $F \circ F^{-1}(u) \geq u, \ \forall \ u \in [0,1]$. If F is continuous, then we have equality.

Proof

1. Left as an exercise on Assignment 2

2. $F^{-1} \circ F(x) = \inf\{y \in \mathbb{R} : F(y) \geq F(x)\} \equiv A$. Surely, $x \in A$ and $\inf(A) \leq x$, and $\inf(A) = F^{-1} \circ F(x)$. Suppose now that F is strictly increasing and $F^{-1} \circ F(x) < x$ for some $x \in \mathbb{R}$. Hence $\exists a \in A$ so that $a < x$. But F is strictly increasing, so $F(a) < F(x)$. Contradiction to the fact $a \in A$, which entails $F(a) \geq F(x)$.

3. F is by definition continuous from the right. We have equality at 0, so assume without loss of generality that $y \in (0,1]$ and $F^{-1}(1) < \infty$, or $u \in (0,1)$. Then $F^{-1} \in (-\infty, \infty)$ and

$$F(F^{-1}(u)) = \lim_{\substack{x_n \to F^{-1}(u) \\ x_n > F^{-1}(u)}} F(x_n).$$

   Now $x_n > F^{-1}(u)$, so $\exists x_n^* \in A = \{x : F(x) \geq u\}$ so that $x_n^* < x_n$. F is non-decreasing, so $u \leq F(x_n^*) \leq F(x_n)$ and hence $x_n \in A$ *i.e.* $F(x_n) \geq u$. Then $F \circ F^{-1}(u) = \lim_{n \to \infty} F(x_n) \geq u$. If F is continuous, then

$$F(F^{-1}(u)) = \lim_{\substack{n \to \infty \\ y_n \to F^{-1}(u) \\ y_n < F^{-1}(u)}} F(y_n) \quad \Leftrightarrow \qquad y_n < \inf(A)$$

which imply $\Rightarrow y_n \notin A \Rightarrow F(y_n) < u$.

■

Theorem 3.2 (Probability integral transform and quantile transform)
Suppose that X has CDF F and $U \sim \mathcal{U}(0,1)$. Then

1. $F^{-1}(U)$ has CDF F (quantile transform)
2. If F **is continuous**, then $F(X) \sim \mathcal{U}(0,1)$. (probability integral transform).

Remark
The second result is a core result in non-parametric statistics, while the first statement is used for random number generation.

Proof

1. Observe that $\forall\, u \in (0,1), x \in \mathbb{R}$,

$$F^{-1}(u) \leq x \quad \Leftrightarrow \quad u \leq F(x)$$

Indeed, F is non-decreasing and hence if $F^{-1}(u) \leq x$, we can argue that $F(F^{-1}(u)) \leq F(x)$ and by our lemma, $u \leq F(F^{-1}(u))$, which proves necessity.

For sufficiency, if $u \leq F(x)$, then $x \in \{y : F(y) \geq u\}$ and so $\inf\{y : F(y) \geq u\} \leq x \Rightarrow F^{-1}(u) \leq x$. Now let $U \sim \mathcal{U}(0,1)$ and set $X = F^{-1}(U)$.

$$F(x) = \mathsf{P}(X \leq x) = \mathsf{P}(F^{-1}(U) \leq x) = \mathsf{P}(U \leq F(x)) = F(x)$$

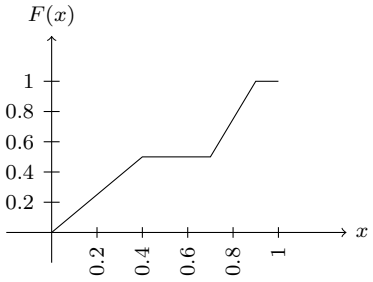
as $F(x)$ takes its values in the interval $[0,1]$.

2. First, observe that $F(X)$ has continuous CDF. Indeed, $\forall\ u \in [0,1]$, $\mathsf{P}(F(X) = u) \stackrel{?}{=} 0$. This is $\mathsf{P}(X = F^{-1}(u)) = 0$ as there are no atoms, since X is absolutely continuous random variable. If there is a plateau, we could have some point and $\mathsf{P}(X \in [F^{-1}(u), F^{+}(u)]) = F(F^{-1^{+}}(u)) - F(F^{-1}(u)) = u - u = 0$. The third case is

$$\mathsf{P}(F(X) \leq u) = 1 - \mathsf{P}(F(X) > u) = 1 - \mathsf{P}(F(X) \geq u)$$

using continuity, this happens if and only if $F^{-1}(u) \leq X$ and thus

$$\begin{aligned} &= 1 - \mathsf{P}(X \geq F^{-1}(u)) \\ &= \mathsf{P}(X < F^{-1}(u)) \\ &= F(F^{-1}(u)) = u. \end{aligned}$$



using the third part of the lemma.

■

Example 3.4
Let $X \sim \mathcal{E}(1)$, then $F(x) = 1 - e^{-x}$, if $x \geq 0$. Then

$$\begin{aligned} & 1 - e^{-x} = u \\ \Leftrightarrow \quad & 1 - u = e^{-x} \\ \Leftrightarrow \quad & -\ln(1-u) = x \end{aligned}$$

and as a result

$$F^{-1}(u) = \begin{cases} -\ln(1-u), & \text{if } u \in (0,1) \\ -\infty, & \text{if } u = 0 \\ \infty, & \text{if } u = 1 \end{cases}$$

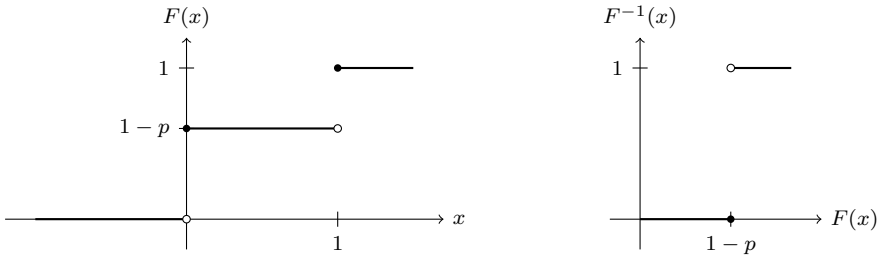
Example 3.5
Let $U \sim \mathcal{U}(0,1)$. Then

$$F^{-1}(U) = -\ln(1-U) \stackrel{\mathrm{d}}{=} -\ln(u) \sim \mathcal{E}(1).$$

Example 3.6
Let $X \sim \mathcal{B}(p)$ with $\mathsf{P}(X=1) = p$ and $\mathsf{P}(X=0) = 1-p$.

F^{-1} can only take two values, either 0 or 1 and

$$\begin{aligned} P(F^{-1}(U) = 0) &= \mathsf{P}(0 \leq U \leq 1-p) = 1-p \\ P(F^{-1}(U) = 1) &= \mathsf{P}(1-p < U \leq 1) = p \end{aligned}$$



and so $F^{-1}(U) \sim \mathcal{B}(p)$, that is a Bernoulli random variable.

If we now look at the probability integral transform and consider $F(X) \in \{F(0), F(1)\} = \{1-p, 1\}$ and in particular, $F(X)$ is a discrete random variable and in particular, it cannot be uniform on (0,1).

Remark

If F is arbitrary, then to transform F onto a uniform, we can take $F(X\_ + \mathcal{U}(F(X) - F(X\_))$ is uniform on $(0, 1)$. This is called randomization of the ties.

Section 3.2. Monotone transformations

Suppose X has CDF F and density f and consider a function $g(X)$, assuming for now that $g : \mathbb{R} \to \mathbb{R}$ is strictly increasing and "smooth". Then

$$\mathsf{P}(g(X) \leq x) = \mathsf{P}(X \leq g^{-1}x) = F(g^{-1}(x))$$

Then the density of $g(X)$ is

$$f_{g(X)}(x) = f_X(g^{-1}(x)) \cdot \frac{\partial}{\partial x} g^{-1}(x)$$

If g is decreasing, then

$$\begin{aligned} F_{g(X)}(x) &= \mathsf{P}(g(X) \leq x) \\ &= \mathsf{P}(X \geq g{-}1(x)) \\ &= 1 - \mathsf{P}(X < g^{-1}(x)) \\ &= 1 - \mathsf{P}(X \leq g^{-1}(x)) \\ &= 1 - F(g^{-1}(x)) \end{aligned}$$

and the density is in this case

$$f_{g(X)}(x) = -f_X(g^{-1}(x)) \cdot \frac{\partial}{\partial x} g^{-1}(x)$$

Together, we have

$$f_{g(X)}(x) = f_X(g^{-1}(x)) \cdot \left| \frac{\partial}{\partial x} g^{-1}(x) \right| \tag{3.4}$$

Theorem 3.3

Let X be a random variable with density f_X. Suppose that $\mathbb{R} = A_0 \cup A_1$ where $A_1 = (a, b)$ is an **open interval**, $\mathsf{P}(X \in A_0) = 0$,[6] f_X is continuous on A_1. Let g be a strictly monotone function on A_1 and its inverse g^{-1} is continuously differentiable on $g(A_1)$. Then $Y = g(X)$ has density

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{\partial}{\partial y} g^{-1}(y) \right| \mathbf{1}_{(y \in g(A_1))}. \tag{3.5}$$

Example 3.7

Take again $X \sim \mathcal{E}(1)$ and look at $F^{-1}(u) = -\ln(1-u)$. Then $\mathbb{R} = A_0 \cup A_1$ and $A_1 = (0, 1)$. Then $g(0, 1) \to (0, \infty)$ and $u \mapsto -\ln(1-u)$. Then $g^{-1}(x)$ is

$$-\ln(1-u) = x \quad \Leftrightarrow \quad e^{-x} = 1 - u \quad \Leftrightarrow \quad u = 1 - e^{-x}$$

so $g^{-1} : (0, \infty) \to (0, 1)$ and $x \mapsto 1 - e^{-x}$. Also $\frac{\partial}{\partial x} g^{-1}(x) = e^{-x}$.

Then $F^{-1}(U)$ as density

$$\begin{aligned} f_{F^{-1}(U)}(x) &= f_U(1 - e^{-x} e^{-x} \mathbf{1}_{(x \in (0,\infty))} \\ &= e^{-x} \mathbf{1}_{(x \in (0,\infty))} \end{aligned}$$

and we conclude again that $F^{-1}(U) \sim \mathcal{E}(1)$. Now take $X \sim \mathcal{E}(1)$ so that $F(X) = 1 - e^{-X} \sim \mathcal{U}(0, 1)$. Then $A_1 : (0, \infty)$ and $g(0, \infty) \to (0, 1), x \mapsto 1 - e^{-x}$. Thus $g^{-1}(0, 1) \to (0, \infty)$ and $u \mapsto -\ln(1-u)$ and the derivative is $\frac{\partial}{\partial u} g^{-1}(u) = -(1-u)^{-1} \cdot (-1) = (1-u)^{-1}$. Thus

$$f_{F(X)}(u) = e^{-\ln(1-u)} \left| \frac{1}{1-u} \right| \mathbf{1}_{(u \in (0,1))}$$

and $F(X) \sim \mathcal{U}(0, 1)$.

Theorem 3.4 (Monotone transformation theorem)

Let X be a random variable with density f_X. Suppose that $\mathbb{R} = A_0 \cup A_1 \cup \cdots \cup A_k$ such that

1. $\mathsf{P}(X \in A_0) = 0$

[6] A_0 is a trash can for all discontinuity, points where the derivative does not exist, etc.

2. $A_1, \dots, A_k$ are open intervals, where $A_i = (a_i, b_i)$

3. f_X is continuous on A_i for $i \in \{1, \dots, k\}$.

Suppose that $g : A_1 \cup \cdots \cup A_k \to \mathbb{R}$ is a mapping such that for any $i \in \{1, \dots, k\}$.

Also assume that $g_i : A_i \to \mathbb{R}, x \mapsto g(x)$ is strictly monotone and its inverse $g_i^{-1} : g(A_i) \to \mathbb{R}$ is continuously differentiable.

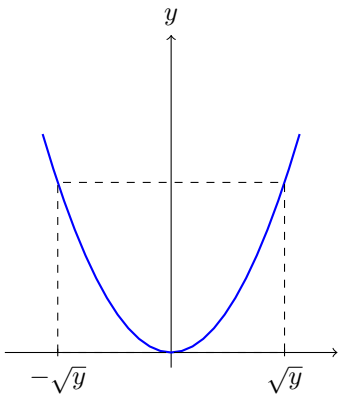
Then $Y = g(X)$ has density

$$f_Y(y) = \sum_{i=1}^{k} f_X(g_i^{-1}(y)) \left| \frac{\partial}{\partial y} g_i^{-1}(y) \right| \mathbf{1}_{(y \in g(A_i))}$$

Example 3.8

Let $X \sim \mathcal{N}(0,1)$ where $f_X(x) = (2\pi)^{-\frac{1}{2}} \exp(-x^2/2)\mathbf{1}_{(x \in (-\infty, \infty))}$ and consider the transformation $g(x) = x^2$ and $Y = X^2$. Then $A_1 = (-\infty, 0), A_2 = (0, \infty)$ and $A_0 = \{0\}$. Then we have

$$\begin{aligned}
g_1 : (-\infty, 0) &\to (0, \infty) & g_2 : (0, \infty) &\to (0, \infty) \\
x &\mapsto x^2 & x &\mapsto x^2 \\
g^{-1} : (0, \infty) &\to (-\infty, 0) & g_2^{-1} : (0, \infty) &\to (0, \infty) \\
y &\mapsto -\sqrt{y} & y &\mapsto \sqrt{y} \\
\frac{\partial}{\partial y} g_1^{-1} &= -\frac{1}{2\sqrt{y}} & \frac{\partial}{\partial y} g_2^{-1} &= -\frac{1}{2\sqrt{y}}
\end{aligned}$$



and we have

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y))|g^{-1}(y)'|\mathbf{1}_{(y\in(0,\infty))} \\ &= 2\frac{1}{\sqrt{2\pi}}e^{-\frac{y}{2}}\frac{1}{2\sqrt{y}}\mathbf{1}_{(y\in(0,\infty))} \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{y}{2}}\frac{1}{\sqrt{y}}\mathbf{1}_{(y\in(0,\infty))} \end{aligned}$$

and $Y \sim \chi^2(1)$. As an exercise, try $|X|$.

We proceed with the proof of the monotone transformation theorem.

Proof We have to show that $\forall\, y \in \mathbb{R}$

$$\mathsf{P}(Y \leq y) = \int_{-\infty}^{y} f_Y(t)dt.$$

We have

$$\begin{aligned} \mathsf{P}(Y \leq y) &= \mathsf{P}(g(X) \leq y) \\ &= \sum_{i=1}^{k} \mathsf{P}(g(X) \leq y, X \in A_i) + \mathsf{P}(g(X) \leq y, X \in A_0) \\ &= \sum_{i=1}^{k} \mathsf{P}(g(X) \leq y, X \in A_i) \end{aligned}$$

as $\mathsf{P}(X \in A_0) = 0$.

1. If g_i is increasing on $A_i = (a_i, b_i)$, then $g(A_i) = (g(a_i), g(b_i))$ Now if $y \in (g(a_i), g(b_i))$, then

$$\begin{aligned} \mathsf{P}(g_i(X)) \leq y, X \in A_i) &= \mathsf{P}(a_i < X \leq g_i^{-1}(y)) \\ &= \int_{a_i}^{g^{-1}(y)} f_X(s)ds \\ &= \int_{g(a_i)}^{s} g_i^{-1}(t)g_i^{-1}(t)'dt \\ &= \int_{-\infty}^{y} f_X(g_i^{-1}(t))(g_i^{-1}(t))'\mathbf{1}_{(t\in(g_i(a_i),g_i(b_i))=g(A_i))} \end{aligned}$$

   using the transformation $s = g_i^{-1}(t)$.

2. If $y \leq g(a_i)$ then the probability of being on the interval is zero, as

$$0 = \int_{-\infty}^{y} f_X(g_i^{-1}(t))(g_i^{-1}(t))' \mathbf{1}_{(y \in (-\infty, g(a_i))} dt$$

3. If $y \geq g(b_i)$, then the probability is the same as $\mathsf{P}(X \in A_i)$.

■

Example 3.9

Let X be a random variable with $f_X(x) = \frac{2}{9}(x+1)$, where $x \in (-1, 2)$ and take $g : \mathbb{R} \to \mathbb{R}$ and $x \mapsto x^2$. This is not strictly monotone, but we can partition the space to make it monotone on the intervals and take $A_1 = (-1, 0)$, $A_2 = (0, 2)$ and $A_0 = (-\infty, 1] \cup \{0\} \cup [2, \infty)$. Checking the conditions for the monotone transformation theorem, $\mathsf{P}(X \in A_0) = 0$ is fulfilled, f_X is continuous on $A_1 \cup A_2$, and $g_1 : A_1 \to (0, 1)$ so $x \mapsto x^2$ with $g_1^{-1} : (0, 1) \to (-1, 0)$ and $x \mapsto -\sqrt{x}$ with $(g_1^{-1})'(\mathrm{x}) = -\frac{1}{2\sqrt{x}}$. Similarly, we have $g_2 : (0, 2) \to (0, 4)$, $x \mapsto x^2$, $g_2^{-1} : (0, 4) \to (0, 2), x \mapsto \sqrt{x}$ with $(g_2^{-1})'(x) = \frac{1}{2\sqrt{x}}$

Now $Y = g(X) = X^2$; we have

$$\begin{aligned} f_Y(y) &= f_X(-\sqrt{y}) \frac{1}{2\sqrt{y}} \mathbf{1}_{(y \in (0,1))} + f_X(\sqrt{y}) \frac{1}{2\sqrt{y}} \mathbf{1}_{(y \in (0,4))} \\ &= \frac{2}{9}(-\sqrt{y} + 1) \frac{1}{2\sqrt{y}} \mathbf{1}_{(y \in (0,1))} + \frac{2}{9}(\sqrt{y} + 1) \frac{1}{2\sqrt{y}} \mathbf{1}_{(y \in (0,4))} \\ &= \frac{2}{9} \frac{1}{\sqrt{y}} \mathbf{1}_{(y \in (0,1))} + \frac{2}{9}(\sqrt{y} + 1) \frac{1}{2\sqrt{y}} \mathbf{1}_{(y \in (1,4))} \end{aligned}$$

Section 3.3. Multivariate transformations

Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a random vector with $g(\boldsymbol{X}); g : \mathbb{R}^d \to \mathbb{R}^m$ where d, m are arbitrary. If $g(\boldsymbol{X}) = \boldsymbol{Y} = (Y_1, \ldots, Y_m)$, then $Y_j = g_j(X_1, \ldots, X_d)$ with $g = (g_1, \ldots, g_m)$ and $g_j : \mathbb{R}^d \to \mathbb{R}$.

Our goal is to determine the distribution of $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$. In principle, we can compute the CDF;

$$\begin{aligned} \mathsf{P}(Y_1 \leq y_1, \ldots, Y_m \leq y_m) &= \mathsf{P}(\boldsymbol{Y} \in (-\infty, y_1] \times \cdots \times (-\infty, y_m]) \\ &= \mathsf{P}(g(\boldsymbol{X}) \in (-\infty, y_1] \times \cdots \times (-\infty, y_m])) \\ &= \mathsf{P}(\boldsymbol{X} \in g^{-1}((-\infty, y_1] \times \cdots \times (-\infty, y_m]) \end{aligned}$$

where $g^{-1}(\cdot)$ is the preimage set.

First, we do some **by hand transformations**.

Example 3.10

Let X, Y be independent random variables, with $X \sim \mathcal{P}(\lambda), Y \sim \mathcal{P}(\mu)$ with $\lambda, \mu > 0$. We are interested in $Z = g(X,Y) = X + Y$. What is the distribution of Z? We have $g : \mathbb{R}^2 \to \mathbb{R}, (x,y) \mapsto (x+y)$. Remark first that Z is discrete, so we need to think of the support of Z, which will be $\mathbb{N}_0$, the positive integers. Now, the preimage set is

$$g^{-1}(\{k\}) = \{(i,j) : i + j = k\} = \{i \in \{0, \dots, k\}, j = k - i\}$$

so that the PMF is

$$\begin{aligned}
f_Z(k) &= \mathsf{P}(X + Y = k) = \mathsf{P}(g(X,Y) \in \{k\}) \\
&= \sum_{i=0}^{k} \mathsf{P}(X = i, Y = k - i) \\
&= \sum_{i=0}^{k} \mathsf{P}(X = i)\mathsf{P}(Y = k - i) \\
&= \sum_{i=0}^{k} e^{-\lambda} \frac{\lambda^i}{i!} e^{-\mu} \frac{\mu^{k-i}}{(k-i)!} \\
&= e^{-\lambda-\mu} \sum_{i=0}^{k} \binom{k}{i} \lambda^i \mu^{k-i} \\
&= e^{-\lambda-\mu} \frac{1}{k!} (\lambda + \mu)^k
\end{aligned}$$

and we conclude that $Z \sim \mathcal{P}(\lambda + \mu)$.

Exercise 3.1

If $Z \sim \mathcal{P}(\lambda_{11})$, $X \sim \mathcal{P}(\lambda_{01})$ and $Y \sim \mathcal{P}(\lambda_{01})$ independent random variables. If one considers $(X + Z, Y + Z)$, then get a Bivariate Poisson.

Example 3.11

Let X be Beta $\mathcal{B}(\alpha, \beta)$ and $Y \sim \mathcal{B}(\alpha + \beta, \gamma)$ where $X \perp\!\!\!\perp Y$. We are looking at the transformation $Z = XY$. The transformation of interest is $g : \mathbb{R}^2 \to \mathbb{R}, (x,y) \mapsto xy$. Recall the density of the bivariate Beta with independent parameters is

$$f_X(x) f_Y(y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \frac{\Gamma(\alpha + \beta + \gamma)}{\Gamma(\alpha + \beta)\Gamma(\gamma)} y^{\alpha+\beta-1}(1-y)^{\gamma-1} \mathbf{1}_{(x,y \in (0,1))}$$

Z is continuous and takes values in $(0,1)$, thus we have

$$\begin{aligned}\mathsf{P}(Z \leq z) = \mathsf{P}(XY \leq z) &= \mathsf{P}((X,Y) \in g^{-1}((-\infty, z]) \\ &= \int_{g^{-1}((-\infty,z])} f_X(x) f_Y(y) dx dy\end{aligned}$$

where

$$\begin{aligned}g^{-1}((-\infty, z]) &= \{(x,y) : xy \leq z\} \\ &= \left\{ y \in (0,1), x \in \left(0, \frac{z}{y}\right) \right\}\end{aligned}$$

The integral becomes

$$\begin{aligned}\mathsf{P}(Z \leq z) &= \int_0^1 \int_0^{\frac{z}{y}} f_X(x) f_Y(y) dx dy \\ &= \int_0^1 \int_0^z f_X\left(\frac{t}{y}\right) f_Y(y) \frac{1}{y} dt dy \\ &= \int_0^z \int_0^1 f_X\left(\frac{t}{y}\right) f_Y(y) \frac{1}{y} dy dt \\ &= \int_0^z f_Z(t) dt\end{aligned}$$

using the transformation $x = t/y$, $dx = dt/y$ where $t \in (0, z)$ in the first step and applying Fubini's theorem in the second step. The expression $f_Z(t)$ will be the density of Z as

$$\begin{aligned}f_Z(t) &= \mathbf{1}_{(t \in (0,1))} \int_0^1 f_X\left(\frac{t}{y}\right) f_Y(y) \frac{1}{y} dy \\ &= C \int_t^1 \frac{t}{y} \left(1 - \frac{t}{y}\right)^{\beta-1} y^{\alpha+\beta-1} (1-y)^{\gamma-1} \frac{1}{y} dy \\ &= C \int_t^1 \left(\frac{t}{y}\right)^{\alpha-1} \left(1 - \frac{t}{y}\right)^{\beta-1} y^{\alpha+\beta-2} (1-y)^{\gamma-1} dy \\ &= C t^{\alpha-1} \int_t^1 \left(\frac{y-t}{1-t}\right)^{\beta-1} \left(\frac{1-(y-t)-t}{1-t}\right)^{\gamma-1} dy\end{aligned}$$

Regrouping the y terms. Make now the change of variable $x = (y-t)/(1-t) \in (0,1)$ with

$dx = (1-t)^{-1}dy$. The function then becomes

$$\begin{aligned}
&= C(1-t)^{\beta+\gamma-2}t^{\alpha-1}\int_0^1 x^{\beta-1}(1-x)^{\gamma=1}dx \\
&= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)}\frac{\Gamma(\beta)\Gamma(\gamma)}{\Gamma(\beta+\gamma)}(1-t)^{\beta+\gamma-1}t^{\alpha-1} \\
&= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta+\gamma)}(1-t)^{\beta+\gamma-1}t^{\alpha-1}
\end{aligned}$$

since the integral is the Beta function $\mathcal{B}(\beta,\gamma)$ We conclude that $Z = XY \sim \mathcal{B}(\alpha, \beta+\gamma)$.

Example 3.12

Take

$$f_{(X,Y)}(x,y) = e^{-y}\mathbf{1}_{(0<x<y<\infty)}$$

and consider $Z = X + Y$. Then

$$\mathsf{P}(Z \leq z) = \int_0^z \int_0^{\min(z-y,y)} e^{-y}dxdy$$

where $\int_0^{\min(z-y,y)} e^{-y}dx$ is the density of Z. Computing the integral, we get

$$f_Z(z) = \left(e^{-\frac{z}{2}} - e^{-z}\right)\mathbf{1}_{(z>0)}$$

Note that the two previous examples could have been solved using the multivariate transformation theorem, which we introduce next.

Theorem 3.5

Suppose that $\boldsymbol{X} = (X_1, \ldots, X_d)$ has a density f_X, denote the interior of the support of $\boldsymbol{X}$ by $A \in \mathbb{R}^d$. Suppose $g : A \to g(A) \subseteq \mathbb{R}^d$ (the same dimension) is one-to-one and assume that its inverse $h : g(A) \to A, h = (h_1, \ldots h_d)$ is such that $\forall\ i \in \{1, \ldots, d\}$, h_i is continuously differentiable in each argument and the Jacobi matrix

$$\mathrm{J} = \begin{pmatrix} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_1}{\partial y_d} \\ \vdots & \ddots & \cdots \\ \frac{\partial h_d}{\partial y_1} & \cdots & \frac{\partial h_d}{\partial y_1} \end{pmatrix}$$

is not identically 0 on A. Then $g(\boldsymbol{X})$ has density

$$f_{g(\boldsymbol{X})}(\boldsymbol{y}) = f_{\boldsymbol{X}}(h_1(\boldsymbol{y}), \ldots h_d(\boldsymbol{y}))|\det \mathrm{J}|\mathbf{1}_{(\boldsymbol{y}\in g(A))} \tag{3.6}$$

for $\boldsymbol{y} = (y_1, \ldots, y_d)$.

Example 3.13
Again take $X \sim \mathcal{B}(\alpha, \beta)$ and $Y \sim \mathcal{B}(\alpha+\beta, \gamma)$ with $X \perp\!\!\!\perp Y$. Again, we are interested in $Z_1 = XY$. We cannot apply the theorem, however we can consider a dummy random variable $Z_2 = Y$ and marginalize over Y. The function $g : (0,1)^2 \to ((u,v) : 0 < u < v < 1)$ and $(x,y) \mapsto (xy, y)$. If we want to compute the inverse, we must have $g : \{(u,v) : 0 < u < v < 1\} \to (0,1)^2$ and $(u,v) \mapsto \left(\frac{u}{v}, v\right)$, with $h_1(u,v) = \frac{u}{v}$ and $h_2(u,v) = u$. The Jacobian is found using the Jacobi matrix

$$\mathrm{J} = \begin{pmatrix} \frac{1}{v} & -\frac{u}{v^2} \\ 0 & 1 \end{pmatrix}$$

which has determinant $1/v$. The joint density of Z_1, Z_2 is

$$\begin{aligned} f_{(Z_1,Z_2)}(u,v) &= f_X(h(u,v)) f_Y(h_2(u,v)) \frac{1}{v} \mathbf{1}_{(0<u<v<1)} \\ &= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} \left(\frac{u}{v}\right)^{\alpha-1} \left(1-\frac{u}{v}\right)^{\beta-1} v^{\alpha+\beta-2}(1-v)^{\gamma-1} \mathbf{1}_{(0<u<v<1)} \end{aligned}$$

where C, which is the front term, is constant. We have Z_1 has density

$$f_{Z_1}(u) = \int_u^1 C \left(\frac{u}{v}\right)^{\alpha-1} \left(1-\frac{u}{v}\right)^{\beta-1} v^{\alpha+\beta-2}(1-v)^{\gamma-1} dv$$

which is exactly the same integral calculated previously, but with a different parametrization.

Section 3.4. Linear transformations

Consider $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$, a $d \times 1$ (column) vector and a matrix $\boldsymbol{C} \in \mathbb{R}^{d \times d}$, $\boldsymbol{C}$ invertible. Consider $Y = \boldsymbol{C} \cdot \boldsymbol{X}$, so that $\boldsymbol{X} = \boldsymbol{C}^{-1}\boldsymbol{Y}$, $h = \boldsymbol{C}^{-1}\boldsymbol{y}$. Take $\boldsymbol{J} = \boldsymbol{C}^{-1}$ such that

$$|\det \boldsymbol{J}| = \frac{1}{|\det \boldsymbol{C}|} = |\det \boldsymbol{C}^{-1}|.$$

Then, the density of $\boldsymbol{Y}$ is given by

$$f_{\boldsymbol{Y}}((y_1, \ldots y_d) = \frac{1}{|\det \boldsymbol{C}|} \cdot f_{\boldsymbol{X}} \left(\left(\boldsymbol{C}^{-1}\boldsymbol{Y}\right)^\top \right)$$

Example 3.14
Let $(X, Y) \sim \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{I})$. Then

$$\boldsymbol{Z} = \begin{pmatrix} X+Y \\ X-Y \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

where $\det \boldsymbol{C} = -1 - 1 = -2$, so $|\det \boldsymbol{C}|^{-1} = 1/2$. Then

$$\begin{aligned} f_{bsZ}(u,v) &= \frac{1}{2} f_{(X,Y)} \left(\left(\boldsymbol{C}^{-1} \left(\begin{smallmatrix} u \\ v \end{smallmatrix} \right) \right)^\top \right) \\ &= \frac{1}{2} f_{(X,Y)} \left(\frac{u+v}{2}, \frac{u-v}{2} \right) \\ &= \frac{1}{2} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(u+v)^2}{2 \cdot 4} \right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(u-v)^2}{2 \cdot 4} \right) \\ &= \frac{1}{2} \frac{1}{2\pi} \exp\left(-\frac{-2u^2 + 2v^2}{8} \right) \\ &= \frac{1}{2} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{4} \right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{4} \right) \end{aligned}$$

so

$$\begin{pmatrix} X+Y \\ X-Y \end{pmatrix} \sim \mathcal{N}_2(\boldsymbol{0}, 2\boldsymbol{I}_2)$$

where $\boldsymbol{0} = (0\ 0)^\top$ and $2\boldsymbol{I}_2 = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix} \right)$. Remark that $(X+Y) \perp\!\!\!\perp (X-Y)$ is a coincidence, but the fact that $((X+Y),(X-Y))$ is Bivariate Normal is not a coincidence. If we further generalize this example.

Example 3.15

Let $(X_1, \ldots, X_d)^\top \sim \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{I}_d)$ with

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{x_1^2 + \cdots + x_d^2}{2} \right) \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\boldsymbol{x}^\top \boldsymbol{x}}{2} \right) \end{aligned}$$

Now if $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, for $\boldsymbol{A}$ invertible, with $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^\top$, $\sqrt{\det \boldsymbol{\Sigma}} = |\det \boldsymbol{A}|$. Thus

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= \frac{1}{\sqrt{\det \boldsymbol{\Sigma}}} f_{\boldsymbol{X}}(\boldsymbol{A}^{-1}\boldsymbol{y}) \\ &= \frac{1}{\sqrt{\det \boldsymbol{\Sigma}} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{(\boldsymbol{A}^{-1}\boldsymbol{y})^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{2} \right) \\ &= \frac{1}{\sqrt{\det \boldsymbol{\Sigma}} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\boldsymbol{y}^\top (\boldsymbol{A}^{-1})^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{2} \right) \end{aligned}$$

But notice that $(\boldsymbol{A}^{-1})^\top \boldsymbol{A}^{-1} = \boldsymbol{\Sigma}^{-1}$, as $\boldsymbol{A}^{-1}$ is a Choleski decomposition. Thus

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\sqrt{\det \boldsymbol{\Sigma}}(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\boldsymbol{y}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{y}}{2}\right)$$

and $\boldsymbol{Y} \sim \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma})$

In particular, if $\boldsymbol{A}$ is orthonormal, then $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I}_d$ and $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{I}_d)$. So the distribution of $\boldsymbol{X}$ is not affected by rotations. This is so as the function depends only on the arguments through the the Euclidean distance.

Example 3.16
Suppose $\boldsymbol{X}$ has a density of the form

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = h(x_1^2 + \cdots + x_d^2)$$

of which examples are Normal or Student$-t$ distributions. Then if $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$ and $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}_d$, then

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = h((\boldsymbol{A}^{-1}\boldsymbol{y})^\top(\boldsymbol{A}^{-1}\boldsymbol{y})) = h(\boldsymbol{y}^\top\boldsymbol{y})$$

These are examples of **spherical distributions.**

Definition 3.6 (Spherical distributions)
If $\boldsymbol{X}$ has a density of the form $h(\boldsymbol{X}^\top\boldsymbol{X})$, then $\boldsymbol{X}$ is said to have a **spherical distribution**.

Example 3.17
Let $\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{I}_d)$ and $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\mu}$ where $\boldsymbol{\mu} = (\mu_1 \dots \mu_d)^\top \in \mathbb{R}^{d\times 1}$ and $\boldsymbol{A}$ is invertible, $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Sigma}$. Consider $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma})$. Then $\boldsymbol{Y} = \boldsymbol{Z} + \boldsymbol{\mu}$. Thus $g(\boldsymbol{z}) = (z_1 + \mu_1, \dots, z_d + \mu_d)$. Now $\boldsymbol{J} = \boldsymbol{I}$, the identity matrix and as such $\det(\boldsymbol{J}) = 1$. Therefore

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= f_{\boldsymbol{Z}}(\boldsymbol{y} - \boldsymbol{\mu}) \\ &= \frac{1}{\det(\boldsymbol{\Sigma})^{1/2}(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2}\right) \end{aligned}$$

and hence $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Example 3.18
Let $\boldsymbol{X}$ be a spherical distributions, with $f_X(\boldsymbol{x}) = h(\boldsymbol{x}^\top\boldsymbol{x})$. Then, for $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\mu}$, for $\boldsymbol{A}$ invertible $d \times d$ matrix with $\boldsymbol{\mu} \in \mathbb{R}^{d\times 1}$. In such case,

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\det(\boldsymbol{\Sigma})^{1/2}} h\left((\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)$$

where $\|\boldsymbol{y}-\boldsymbol{\mu}\|_{\mathrm{M}} \coloneqq (\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$ is the **Mahalanobis distance**.

This can be used for financial return models as they retain essential properties of the Multivariate Normal, as tractability. This can also serve in multivariate regression for the specification of the error terms.

Section 3.5. Convolutions

Let $X \sim F_X$ and $Y \sim F_Y$, with $X \perp\!\!\!\perp Y$. We are interested in the distribution of $X+Y$. For example, insurance companies are interested in the sum of claims, rather than the individual claims.

We will look first at the discrete case: let X, Y be discrete, the sum will be again a discrete random variable. How can we describe the PMF?

$$\begin{aligned} f_{X+Y}(z) = \mathsf{P}(X+Y=z) &= \sum_{y \in \mathrm{Supp}(Y)} \mathsf{P}(X+Y=z|Y=y)\mathsf{P}(Y=y) \\ &= \sum_{y \in \mathrm{Supp}(Y)} \mathsf{P}(X=z-y)\mathsf{P}(Y=y) \\ &= \sum_{y \in \mathrm{Supp}(Y)} f_X(z-y) f_Y(y) \\ &= f_X * f_Y(z) \end{aligned}$$

In some cases, such as sum of Poisson random variables, the sum is again Poisson as seen earlier in the examples.

In the continuous case, f_X, f_Y, the densities of X, Y exist. We could invent a transformation $g : (x, y) \mapsto (x+y, y)$ is one-to-one and $h : (u, v) \mapsto (u-v, v)$. The Jacobian is

$$\boldsymbol{J} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \Rightarrow |\det \boldsymbol{J}| = 1.$$

The density of the transformed variables $(X+Y, Y)$ is then

$$f_{X+Y,Y}(u, v) = f_X(u-v) f_Y(v).$$

The quantity of interest is then

$$f_{X+Y}(u, v) = \int_{-\infty}^{\infty} f_X(u-v) f_Y(v) dv = f_X * f_Y(u)$$

and the formulas (for the discrete and the continuous case) share the same form. Convoluting two Normal variables yields a Normal. Here is another example

Example 3.19

Let $X \sim \mathcal{E}(1) \equiv \mathcal{G}(1)$ and $Y \sim \mathcal{G}(1)$, as the Exponential distribution is a special case of the Gamma distribution. If $f_X(x) = e^{-x}\mathbf{1}_{(x>0)}$, then $X+Y$ has density

$$\begin{aligned} f_{(X+Y}(u) &= \int_{-\infty}^{\infty} e^{-u+v}\mathbf{1}_{(u-v>0)}e^{-v}\mathbf{1}_{(v>0)}dv \\ &= \int_0^u e^{-u}dv \\ &= e^{-u}u\mathbf{1}_{(u>0)} \end{aligned}$$

hence $X+Y \sim \mathcal{G}(2)$. This is valid for any Gamma distribution, if $X \sim \mathcal{G}(\alpha)$ and $Y \sim \mathcal{G}(\beta)$, then $X+Y \sim \mathcal{G}(\alpha+\beta)$.

Exercise 3.2

Try the case $X_1 + \cdots + X_n$, where $X_i \sim \mathcal{E}(1)\ \forall\ i$ and $X_i \perp\!\!\!\perp X_j\ \forall\ i \neq j$.

$X+Y$ has CDF given (in the discrete case) by

$$\begin{aligned} \sum_{t\leq z} f_{X+Y}(t) &= \sum_{t\leq z}\sum_{y\in \text{Supp}(Y)} f_X(t-y)f_Y(y) \\ &= \sum_{y\in \text{Supp}(Y)}\sum_{t\leq z} f_X(t-y)f_Y(y) \\ &= \sum_{y\in \text{Supp}(Y)} F_X(z-y)f_Y(y), \end{aligned}$$

while in the continuous case, we have

$$\begin{aligned} \mathsf{P}(X+Y\leq z) &= \int_{-\infty}^{z}\int_{-\infty}^{\infty} f_X(t-y)f_Y(y)dydt \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{z-y} f_X(s)ds f_Y(y)dy \\ &= \int_{-\infty}^{\infty} F_X(z-y)f_Y(y)dy, \end{aligned}$$

interchanging the integrals as the terms are positive and bounded between 0 and 1, making the change of variable $t-y=s, dt=ds$.

Remark

If $X \perp\!\!\!\perp Y$ and F_X, F_Y are arbitrary, then $X+Y$ has CDF

$$\begin{aligned} F_{X+Y}(z) &= \int_{-\infty}^{\infty} F_X(z-y)dF_X(y) \\ &= \mathsf{E}\left(F_X(z-Y)\right) \end{aligned}$$

the so-called **Lebesgue-Stieltjes integral**.

Chapter 4
Expectations and moments

Suppose X is a random variable. Consider the case of a univariate uniform discrete random variables, $\mathsf{P}(X = x_i) = \frac{1}{n}$. Then $\mathsf{E}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$, which is nothing but the average. We formalize and generalize this notion.

Definition 4.1 (Expectation)
The expected (mean) value of a random variable $g(X)$ where X is a random variable with CDF F is given by

$$\mathsf{E}(g(X)) = \sum_{x \in \text{Supp}(X)} g(x) f_X(x)$$

if X is discrete and

$$\int_{-\infty}^{\infty} f(x) f) X(x) dx$$

if X has a density f_X. or in a unified more general form in terms of the Lebesgue-Stieltjes integral, defined for F arbitrary as

$$\int_{-\infty}^{\infty} g(x) dF(x).$$

Careful. If $\mathsf{E}|g(X)| = \int_{-\infty}^{\infty} |g(X)| dF(x)$ is infinite, then $\mathsf{E}(g(X))$ **does not exist**.

Example 4.1
Let $X \sim \mathcal{B}(n, p)$, and

$$\begin{aligned}
\mathsf{E}(X) &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(n-k)!(k-1)!} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= np \sum_{l=0}^{n-1} \binom{n-1}{l} p^l (1-p)^{n-1-l} \\
&= np(p + (1-p)^{n-1}) \\
&= np
\end{aligned}$$

and in the last step, one could have used instead of the binomial formula the fact that we have the PMF of a binomial with parameters $(n-1, l)$, which sums to 1.

Example 4.2 (Expectation of a Normal random variable)
Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\mathsf{E}(X) = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

First, we verify that $\mathsf{E}|X|$ is finite, that is

$$\begin{aligned}
\mathsf{E}(X) &= \int_{-\infty}^{\infty} \frac{|x|}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
&= \int_{-\infty}^{\infty} |x-\mu+\mu| \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
&\leq \frac{|x-\mu|}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) + |\mu|
\end{aligned}$$

which we can write using the fact that the function is even and the density integrates to one. Thus

$$\begin{aligned}
&|\mu| + \int_{-\infty}^{\infty} |t| \exp\left(-\frac{t^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi}\sigma} dt \\
&= |\mu| + 2\int_{0}^{\infty} t \exp\left(-\frac{t^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi}\sigma} dt \\
&= |\mu| + \int_{0}^{\infty} e^{-y} \frac{1}{\sqrt{2\pi}} dy < \infty.
\end{aligned}$$

using a change of variable $\left(\frac{t}{\sigma}\right)^2 \frac{1}{2} = y$, $\frac{t}{\sigma} dt$.

Thus, since the expectation exists, we can use

$$\begin{aligned}
\mathsf{E}(X) &= \int_{-\infty}^{\infty} (x-\mu+\mu) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
&= \mu + \int_{-\infty}^{\infty} t \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right) \\
&= \mu + \int_{0}^{\infty} t \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right) - \int_{0}^{\infty} t \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right) \\
&= \mu
\end{aligned}$$

Example 4.3 (Expectation of a Cauchy random variable)
Let $X \sim t(1)$, a Cauchy random variable. Then

$$f(x) = \frac{1}{1+x^2} \frac{1}{\pi}, \quad x \in \mathbb{R}$$

Then

$$\mathsf{E}(X) = \int_{-\infty}^{\infty} \frac{x}{1+x^2}\frac{1}{\pi}dx$$

but before proceeding, we verify that $\mathsf{E}(X)$ exists.

$$\begin{aligned}
\mathsf{E}|X| &= \int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)}dx \\
&= 2\int_{0}^{\infty} \frac{x}{\pi(1+x^2)}dx \\
&\geq \frac{2}{\pi}\int_{1}^{\infty} \frac{x}{\pi(1+x^2)}dx \\
&= \frac{2}{\pi}\int_{1}^{\infty} \frac{1}{\pi(1+1/x^2)}dx \\
&\geq \frac{1}{\pi}\int_{1}^{\infty} \frac{1}{x}dx
\end{aligned}$$

using the fact that $\frac{1}{\pi(1+1/x^2)}$ is increasing on $\left(\frac{1}{2}, 1\right)$. Hence $\mathsf{E}|X| = \infty$ and hence $\mathsf{E}(X)$ does not exist.

The Student-t and the Pareto distribution have expectation that do not exist for some values of ν, α

Example 4.4
Let $Y \sim \mathcal{E}(1)$. We are interested in $\mathsf{E}(\min(Y, 1))$

$$\begin{aligned}
\mathsf{E}(\min(Y,1)) &= \int_0^{\infty} \min(y,1)e^{-y}dy \\
&= \int_0^1 ye^{-y}dy + \int_1^{\infty} e^{-y}dy - 1 - e^{-1} \\
&= 1 - 2(e^{-1} + e^{-1})
\end{aligned}$$

If you are not clever and try $X = \min(Y, 1)$ for some $X \sim F$, and try to compute X, then by the Lebesgue decomposition theorem, you can write $F(x) = p_1F_1(x) + p_2F_2(x) + p_3F_3(x)$ with $p_i \in [0, 1]$ and $p_1 + p_2 + p_3 = 1$, where $F_1(x)$ is discrete, $F_2(x)$ is a density and $F_3(x)$ doesn't have a density, yet is continuous. We can write

$$F(x) = e^{-1}\mathbf{1}_{(x\geq 1)} + (1 - e^{-1}))\left[\frac{1-e^{-x}}{1-e^{-1}}\mathbf{1}_{(x\in(0,1))} + \mathbf{1}_{(x\geq 1)}\right]$$

and could use

$$\mathsf{E}(X) = \int_{-\infty}^{\infty} x dF(x) = p_1 \int_{-\infty}^{\infty} x dF_1(x) + p_2 \int_{-\infty}^{\infty} x dF_2(x)$$

This can happen in insurance, where companies pay max of claim amount and a maximal claim size (the remaining part is paid by a reinsurance company) or when dealing with truncations/censoring phenomenon in biostatistics.

Theorem 4.2

Let X be a random variable, $a, b, c \in \mathbb{R}$ and let g_1, g_2 be measurable functions for which $\mathsf{E}(g_i(X))$ exists, $i = \{1, 2\}$. Then

a) $\mathsf{E}(ag_1(X) + bg_2(X)) = a\mathsf{E}(g_1(X)) + b\mathsf{E}(g_2(X))$.

b) If $g_1(x) \geq 0 \; \forall \; x \in \mathbb{R}$, then $\mathsf{E}(g_1(X)) \geq 0$.

c) If $g_1(x) \geq g_2(x) \; \forall \; x \in \mathbb{R}$, then $\mathsf{E}(g_1(X)) \geq \mathsf{E}(g_2(X))$.

d) If $a \leq g_1(x) \leq b \; \forall \; x \in \mathbb{R}$, then $a \leq \mathsf{E}(g_1(X)) \leq b$.

Proof

a)

$$\begin{aligned}
\mathsf{E}(|ag_1(X) + bg_2(X)|) &= \int_{-\infty}^{\infty} |ag_1(x) + bg_2(x)| dF(x) \\
&=\leq \int_{-\infty}^{\infty} |a||g_1(x)| + |b||g_2(x)| \\
&\leq |a| \int_{-\infty}^{\infty} |g_1(x)| dF(x) + |b| \int_{-\infty}^{\infty} |g_2(x)| dF(x) \\
&< \infty
\end{aligned}$$

■

E will be a monotone operator.

The above conditions are unnecessarily stringent, as we are only concerned about the above conditions holding almost surely. Otherwise, the density or PMF is zero. We will make this more precise in the following remarks

Remark

- $\mathsf{P}(X \in A) = \mathsf{E}\left(\mathbf{1}_{(X \in A)}\right)$, thus

$$\int_A f(x)dx = \int_{-\infty}^{\infty} \mathbf{1}_{(x \in A)} f(x)dx = \mathsf{E}\left(\mathbf{1}_{(X \in A)}\right)$$

- Suppose that $\mathsf{P}(X \in A) = 0$, then $\mathsf{E}\left(\mathbf{1}_{(X \in A)} g(X)\right) = 0$. This is immediate as the intersection in the support with $x \in A$ is empty, so we are summing over no term. This may not be that obvious in the continuous case.

Lemma 4.3

If A is such that $\mathsf{P}(X \in A) = 0$, then for arbitrary (measurable) g, $\mathsf{E}\left(g(X)\mathbf{1}_{(X \in A)}\right) = 0$

Proof We will show that $\mathsf{E}\left(|g(X)|\mathbf{1}_{(X \in A)}\right) = 0$, using the fact that $|g(X)|$ is a non-negative function. Suppose first that g is of the form

$$g(x) = \sum_{k=1}^{m} a_k \mathbf{1}_{(x \in A_k)}$$

such that g is **simple**. If we are interested in

$$\begin{aligned}
\mathsf{E}\left(g(X)\mathbf{1}_{(X \in A)}\right) &= \mathsf{E}\left(\sum_{k=1}^{m} a_k \mathbf{1}_{(X \in A_k)} \mathbf{1}_{(X \in A)}\right) \\
&= \mathsf{E}\left(\sum_{k=1}^{m} a_k \mathbf{1}_{(X \in A_k \cap A)}\right) \\
&= \sum_{k=1}^{m} a_k \mathsf{E}\left(\mathbf{1}_{(X \in A_k \cap A)}\right) \\
&= \sum_{k=1}^{m} a_k \mathsf{P}(X \in A_k \cap A) \\
&\leq \mathsf{P}(X \in A) = 0
\end{aligned}$$

■

Proposition 4.4

For any non-negative measurable g, there exists a sequence $\{g_n\}$ of simple functions such that $g_n(x) \to g(x)$ as $n \to \infty$. The simple functions can be taken increasing and $\mathsf{E}\left(g_n(X)\right) \to \mathsf{E}\left(g(X)\right)$. Since each of these are zero, then $\mathsf{E}\left(g(X)\right) = 0$. This uses monotone convergence theorem (Lévy's theorem).

Theorem 4.5
The previous theorem can be generalized as follows:

b*) If $\mathsf{P}(g_1(X) \geq 0) = 1$, then $\mathsf{E}(g_1(X)) \geq 0$ (almost surely)

c*) If $\mathsf{P}(g_1(X) \geq g_2(X) = 1$, then $\mathsf{E}(g_1(X)) \geq \mathsf{E}(g_2(X))$.

d*) $\mathsf{P}(a \leq g_1(X) \leq b) = 1$, then $a \leq \mathsf{E}(g_1(X)) \leq b$.

Proof

b*) We have

$$\begin{aligned}\mathsf{E}(g_1(X)) &= \mathsf{E}\left(g_1(X)\left\{\mathbf{1}_{(x\in A)} + \mathbf{1}_{(X\in A^{\complement})}\right\}\right)\\ &= \mathsf{E}(g_1(X))\,\mathbf{1}_{(X\in A)} + \mathsf{E}(g_1(X))\,\mathbf{1}_{(X\notin A)}\\ &= \mathsf{E}(g_1(X))\,\mathbf{1}_{(g_1(X)\geq 0)} \geq 0\end{aligned}$$

where $A = \{x : g_1(x) \geq 0\}$ using the statement b) of the previous theorem.

■

Lemma 4.6
Suppose that $g(x) \geq 0\ \forall\ x \in \mathbb{R}$ and $\mathsf{E}(g(X)) = 0$. Then $\mathsf{P}(g(X) > 0) = 0$

Proof Let $A = \{g(X) > 0\}$. Then

$$A = \bigcup_{n=1}^{\infty} A_n, \quad A_n = \left\{g(X) > \frac{1}{n}\right\}$$

which for a sequence of nested intervals $A_1 \subset A_2 \subset \ldots \subset A$. We can compute $\mathsf{P}(A) = \lim_{n\to\infty} \mathsf{P}(A_n)$. Then,we need $\mathsf{P}(A_n) = \mathsf{P}\left(g(X) > \frac{1}{n}\right)$. We could write

$$\begin{aligned}\mathsf{E}(g(X)) &= \mathsf{E}\left(g(X)\mathbf{1}_{(g(X)>\frac{1}{n})}\right) + \mathsf{E}\left(g(X)\mathbf{1}_{(g(X)\leq\frac{1}{n})}\right)\\ &\geq \mathsf{E}\left(g(X)\mathbf{1}_{(g(X)>\frac{1}{n})}\right)\\ &\geq \frac{1}{n}\mathsf{E}\left(\mathbf{1}_{(g(X)>\frac{1}{n})}\right)\end{aligned}$$

Thus, we conclude $\mathsf{P}(A_n) \geq 0$ for all n and $\mathsf{P}(a_n) = 0\ \forall\ n$. ■

This can be used to show that if $\mathsf{Var}(X) = 0$, then the random variable is almost surely a constant.

Suppose that $\mathsf{E}\left((X-b)^2\right) < \infty$. We look at

$$\begin{aligned}
\min_b \mathsf{E}\left((X-b)^2\right) &= \min_b \left\{\mathsf{E}\left(X \pm \mathsf{E}\left(X\right) - b\right)^2\right\} \\
&= \min_b \left\{\mathsf{E}\left(X - \mathsf{E}\left(X\right)\right)^2 - 2\mathsf{E}\left(X - \mathsf{E}\left(X\right)\right)\left(b - \mathsf{E}\left(X\right)\right) + 2\left(b - \mathsf{E}\left(X\right)\right)^2\right\} \\
&= \min_b \left\{\mathsf{E}\left(X - \mathsf{E}\left(X\right)\right)^2 + \left(b - \mathsf{E}\left(X\right)\right)^2\right\} \\
&= \mathsf{E}\left(X - \mathsf{E}\left(X\right)\right)^2
\end{aligned}$$

when $b = \mathsf{E}\left(X\right)$.

Definition 4.7 (Moments and central moments)
Let X be a random variable, $r > 0$. Then

- the r^{th} moment of X is given by $\mathsf{E}\left(X^r\right)$, provided $\mathsf{E}\left(X^r\right) < \infty$.
- the r^{th} central moment of X is given by $\mathsf{E}\left(X - \mathsf{E}\left(X\right)\right)^r$.
- When $r = 2$, then $\mathsf{E}\left(X - \mathsf{E}\left(X\right)\right)^2 \equiv \mathsf{Var}\left(X\right)$, the variance, which is a measure of dispersion.

Remark
If $\mathsf{E}|X|^r < \infty$, then $\mathsf{E}|X|^s < \infty \; \forall \; s \in [0, r]$.

Lemma 4.8 (Properties of the variance)
Let X be a random variable such that $\mathsf{E}\left(X\right) < \infty$.

(a) $\mathsf{Var}\left(X\right)$ exists if and only if $\mathsf{E}\left(X^2\right) < \infty$.

(b) $\mathsf{Var}\left(X\right) = \mathsf{E}\left(X^2\right) - \left(\mathsf{E}\left(X\right)\right)^2$.

(c) $\mathsf{Var}\left(aX + b\right) = a^2\mathsf{Var}\left(X\right)$, for $a, b \in \mathbb{R}$.

(d) $\mathsf{Var}\left(X\right) = 0$ if and only if $X = b$ almost surely for some $b \in \mathbb{R}$.

(e) $\mathsf{E}\left(X - b\right)^2 \geq \mathsf{E}\left(X - \mathsf{E}\left(X\right)\right)^2$ for all $b \in \mathbb{R}$, provided $\mathsf{E}\left(X\right)^2 < \infty$.

Proof

(a) Looking at

$$\left(X - \mathsf{E}\left(X\right)\right)^2 + \left(X + \mathsf{E}\left(X\right)\right)^2 = 2X^2 + 2\left(\mathsf{E}\left(X\right)\right)^2$$

and taking expectations, we have

$$\mathsf{E}\left(X-\mathsf{E}\left(X\right)\right)^2+\mathsf{E}\left(X+\mathsf{E}\left(X\right)\right)^2=2\mathsf{E}\left(X^2\right)+2(\mathsf{E}\left(X\right))^2$$

and since the second term on the left hand side is positive, then $\mathsf{Var}\left(X\right)\leq 2(\mathsf{E}\left(X^2\right)+(\mathsf{E}\left(X\right))^2)$ and as such

$$\mathsf{E}\left(X^2\right)\leq 2\mathsf{E}\left(X-\mathsf{E}\left(X\right)\right)^2+2(\mathsf{E}\left(X\right))^2\Leftrightarrow\mathsf{E}\left(X^2\right)\leq 2\mathsf{Var}\left(X\right)+2(\mathsf{E}\left(X\right))^2$$

(d) $\mathsf{Var}\left(X\right)=0\Leftrightarrow\mathsf{E}\left(X-\mathsf{E}\left(X\right)\right)^2=0\Rightarrow X=\mathsf{E}\left(X\right)$ almost surely.

■

Let $\boldsymbol{X}=(X_1,\ldots,X_d)^\top$, for example take $\boldsymbol{X}\sim\mathcal{N}_d(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Now $X_i\sim\mathcal{N}(\mu_i,\sigma_{ii})$ and remark that $\boldsymbol{\mu}=\left(\mathsf{E}\left(X_1\right),\ldots,\mathsf{E}\left(X_d\right)\right)^\top$.

Definition 4.9 (Moments of random vectors)
Let $\boldsymbol{X}$ be a random vector such that $\mathsf{E}\left(X_i\right)<\infty$ for $i\in\{1,\ldots,d\}$. Then $\mathsf{E}\left(X\right)=\left(\mathsf{E}\left(X_1\right),\ldots,\mathsf{E}\left(X_d\right)\right)^\top$.

Remark
Taking $g(\boldsymbol{X})=(g_1(X_1,\ldots,X_d),\ldots,g_n(X_1,\ldots,X_d))^\top$. Then

$$\mathsf{E}\left(g(\boldsymbol{X})\right)=\begin{pmatrix}\mathsf{E}\left(g_1(X_1,\ldots,X_d)\right)\\ \cdots\\ \mathsf{E}\left(g_n(X_1,\ldots,X_d)\right)\end{pmatrix}$$

and we can write each term as

$$\mathsf{E}\left(g_i(X_1,\ldots,X_d)\right)=\int_{-\infty}^{\infty}ydG_i(y)$$

, where G_i is the CDF of $g_i(X_1,\ldots,X_d)$. If you are a little more astute, you can write this as

$$=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}g(x_1,\ldots,x_d)dF(x_1,\ldots,x_d)$$

and where F is the CDF of $\boldsymbol{X}=(X_1,\ldots,X_d)$, which is either the sum over the support of

each X_i if $\boldsymbol{X}$ is discrete, and the integral if it has a density f, namely

$$\begin{aligned} &= \sum g_i(x_1, \ldots, x_d) f(x_1, \ldots, x_d) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_d) f(x_1, \ldots, x_d) dx_1 \ldots dx_d \end{aligned}$$

Definition 4.10 (Covariance and correlation)
Let (X, Y) be a random vector such that $\mathsf{Var}(X)$ and $\mathsf{Var}(Y)$ both exist. Then

(a) The covariance is defined as $\mathsf{Cov}(X, Y) = \mathsf{E}((X - \mathsf{E}(X))(Y - \mathsf{E}(Y)))$

(b) $\mathsf{cor}(X, Y)$ is given by

$$\mathsf{cor}(X, Y) \equiv \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{Var}(X)\,\mathsf{Var}(Y)}},$$

also known as linear correlation or Pearson's correlation.

Lemma 4.11 (Properties of covariance and correlation)
Suppose that the following moments are well-defined. Then

(a) $\mathsf{Cov}(X, Y) = \mathsf{E}(XY) - (\mathsf{E}(X))(\mathsf{E}(Y))$

(b) $\mathsf{Cov}(X, Y) = \mathsf{Cov}(Y, X)$

(b*) $\mathsf{cor}(X, Y) = \mathsf{cor}(Y, X)$

(c) $\mathsf{Cov}(aX + b, cY + d) = ac\mathsf{Cov}(X, Y)$ for $a, b, c, d \in \mathbb{R}$

(c*) $\mathsf{cor}(aX + b, cY + d) = \frac{ac}{|a||c|}\mathsf{cor}(X, Y)$ provided that $a, c \neq 0$.

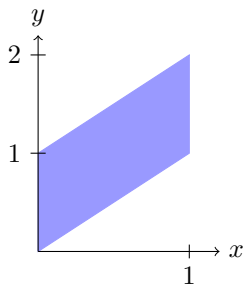
(d) $\mathsf{Cov}(X, X) = \mathsf{Var}(X)$

(d*) $\mathsf{cor}(X, X) = 1$

Example 4.5
Take $f(x, y) = \mathbf{1}_{(0<x<1, x<y<x+1)}$, a triangular density

Then (X, Y) with density $f$illustrated, what is $\mathsf{Cov}(X, Y)$? Note that X has density $f_1(x) = \int_x^{x+1} \mathbf{1} dy = \mathbf{1}_{(x \in (0,1))}$ so $X \sim \mathcal{U}(0, 1)$ and $\mathsf{E}(X) = \frac{1}{2}$ while $\mathsf{Var}(X) = \frac{1}{12}$. Now Y has density

$$f_2(y) = \begin{cases} \int_0^y \mathbf{1} dx = y \\ \int_{y-1}^1 \mathbf{1} dx = 2 - y \end{cases}$$



Thus

$$\begin{aligned}
\mathsf{E}(Y) &= \int_0^2 y f_2(y) dy \\
&= \int_0^1 y^2 dy + \int_1^2 y(2-y) dy \\
&= \frac{1}{3} + (4-1) - \left(\frac{8}{3} - \frac{1}{3}\right) \\
&= 3 - 2 = 1
\end{aligned}$$

so $\mathsf{Var}(Y) = \mathsf{E}\left(Y^2\right) - 1$ and

$$\begin{aligned}
\mathsf{E}\left(Y^2\right) &= \int_0^1 y^3 dy + \int_1^2 y^2(2-y) dy \\
&= \frac{1}{4} + \frac{2}{3}(8-1) = \left(\frac{16}{4} - \frac{1}{4}\right) \\
&= -\frac{14}{4} + \frac{14}{3} = \frac{14}{12}
\end{aligned}$$

thus $\mathsf{Var}(Y) = \frac{14}{12} - 1 = \frac{1}{6}$. Now

$$\begin{aligned}
\mathsf{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy \\
&= \int_0^1 \int_x^{x+1} \frac{x}{2} dydx \\
&= \int_0^1 \frac{x}{2}\left((x+1)^2 - x^2\right) dx \\
&= \int_0^1 \left(x^2 + \frac{x}{2}\right) dx \\
&= \frac{1}{3} + \frac{1}{4} \\
&= \frac{7}{12}
\end{aligned}$$

so $\mathsf{Cov}(X,Y) = \frac{7}{12} - \frac{1}{2} = \frac{1}{12}$ therefore

$$\mathsf{cor}(X,Y) = \frac{\frac{1}{12}}{\sqrt{\frac{1}{12}\cdot\frac{2}{12}}} = \frac{1}{\sqrt{2}}.$$

Theorem 4.12 (Independence and correlation)
Let X, Y be random variable such that $\mathsf{Var}(X) \neq 0$ and $\mathsf{Var}(Y) \neq 0$ both exist. Then

(a) $X \perp\!\!\!\perp Y$ imply that $\mathsf{Cov}(X,Y) = \mathsf{cor}(X,Y) = 0$

(b) If $\mathsf{cor}(X,Y) = 0$ (or $\mathsf{Cov}(X,Y) = 0$). then X and Y are not necessarily independent.

Proof

(a)

$$\mathsf{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_1(x)f_2(y)dydx = \mathsf{E}(X)\,\mathsf{E}(Y)$$

and hence $\mathsf{Cov}(X,Y) = 0$

(b) Take $Z \sim \mathcal{N}(0,1)$ with $X = Z, Y = Z^2$, then $\mathsf{E}(XY) = \mathsf{E}\left(Z^3\right) = 0$ and $\mathsf{E}(X) = 0$ imply that $\mathsf{Cov}(X,Y) = 0$.

■

Remark

Note that $\mathsf{E}(X+Y) = \mathsf{E}(X) + \mathsf{E}(Y)$; this is always the case using linearity and $\mathsf{E}(XY) = \mathsf{E}(X)\,\mathsf{E}(Y)$ when $X \perp\!\!\!\perp Y$ (this can also happen otherwise, but is then a coincidence.)

Theorem 4.13 (Properties of the variance)

Consider X, Y random variables with finite variances.

(a) $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\mathsf{Cov}(X,Y)$

(b) When $X \perp\!\!\!\perp Y$, then $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$. (or more generally, when $\mathsf{Cov}(X,Y) = 0$, that is when X and Y are **uncorrelated**).

Example 4.6

Taking $U \sim \mathcal{U}(0,2)$ and $V \sim \mathcal{U}(0,1)$ and $U \perp\!\!\!\perp V$, then $X = U$ and $Y = U + V$ has precisely the triangular density we worked with earlier with the transformation

$$f(x,y) = \mathbf{1}_{(0<x<1)}\mathbf{1}_{(y-x\in(0,1))} = \mathbf{1}_{(0<x<1)}\mathbf{1}_{(x<y<x+1)}$$

Then $\mathsf{E}(Y) = \mathsf{E}(Y) + \mathsf{E}(V) = \frac{1}{2} + \frac{1}{2} = 1$ and $\mathsf{Var}(Y) = \mathsf{Var}(U+V) = \mathsf{Var}(U) + \mathsf{Var}(V) = \frac{2}{12}$ and $\mathsf{Cov}(X,Y) = \mathsf{Cov}(U, U-V) = \mathsf{Var}(Y) + \mathsf{Cov}(U,V) = \frac{1}{12}$.

Remark

Think about it first before calculation, there may be a easier way to do it

We may wonder whether $\mathsf{cor}(X,Y)$ is a good measure of the dependence between X and Y.
Pros:

- It is easy to compute, as the empirical estimate is

$$\widehat{\mathsf{cor}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

- It is bounded and standardized quantity, as $|\mathsf{cor}(X,Y)| \leq 1$
- If $X \perp\!\!\!\perp Y$ imply $\mathsf{cor}(X,Y) = 0$
- $\mathsf{cor}(X,Y) = 1$ if and only if $Y = aX + b$ almost surely where $a > 0$. This is used in regression, in the R^2 statistic.
- Also, $\mathsf{cor}(X,Y) = -1$ if and only if $Y = aX + b$ almost surely where $a < 0$.

if $(X,Y) \sim \mathcal{N}_2\left((\mu_1,\mu_2)^\top, \left(\begin{smallmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{smallmatrix}\right)\right)$, then $\mathsf{E}(X) = \mu_1, \mathsf{E}(Y) = \mu_2$ and $\mathsf{Var}(X) = \sigma_{11}$, similarly $\mathsf{Var}(Y) = \sigma_{22}$. Then $\mathsf{Cov}(X,Y) = \sigma_{12}$, $\mathsf{cor}(X,Y) = 0$ imply $X \perp\!\!\!\perp Y$ and all values

in the interval $[-1, 1]$ are attained. If we are not however in this specific framework, this statement is grossly false, which we will show shortly.

Also, in the case where (X, Y) are bivariate Normal

1. ρ is always well-defined.
2. $\rho(X, Y) = 0 \Leftrightarrow X \perp Y$
3. $|\rho| \leq 1$: we have $\rho = 1$, if and only if $Y{=}aX + b, b \in \mathbb{R}, a > 0$ almost surely. Similarly, $\rho = -1$ if and only if $Y{=}aX + b$, for $a < 0, b \in \mathbb{R}$ almost surely.
4. ρ appears directly in the model, as $\rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$. Values of ρ are interpretable, ρ close to zero implies that X and Y are close to be independent, and ρ close to ± 1 implies that X, Y are close to perfect dependence.

While if X, Y are arbitrary

1. ρ is not necessarily well-defined (we need $\mathsf{Var}(X) < \infty, \mathsf{Var}(Y) < \infty$)
2. If $X \perp\!\!\!\perp Y$, then $\rho(X, Y) = 0$, but not conversely.
3. If $|\rho| \leq 1$.

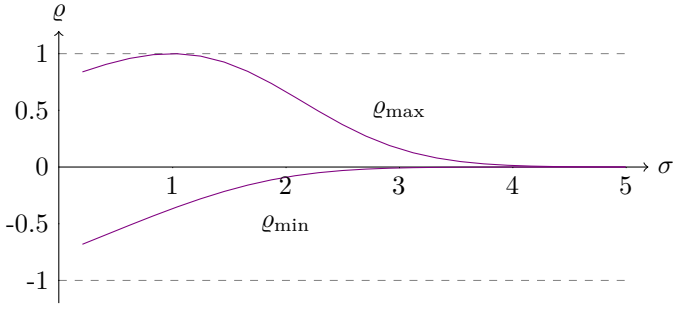
   Consider the function $h(t) = \mathsf{E}\left\{((X - \mathsf{E}(X))t + (Y - \mathsf{E}(Y)))^2\right\}$ for $h : \mathbb{R} \to \mathbb{R}$, $h(t) \geq 0$ and $h(t) = t^2\mathsf{Var}(X) + \mathsf{Var}(Y) + 2t\mathsf{Cov}(X, Y)$, which is a quadratic function in t. If we look at the roots of the function, namely $h(t)$ has at most one root. The discriminant is

   $$4(\mathsf{Cov}(X, Y))^2 - 4\mathsf{Var}X\mathsf{Var}Y \leq 0 \quad \Leftrightarrow \quad |\mathsf{Cov}(X, Y)| \leq \sqrt{\mathsf{Var}(X)\,\mathsf{Var}(Y)}$$

   If follows that $\rho = 1$ if and only if $Y{=}aX + b$ for $a > 0, b \in \mathbb{R}$ almost surely and that $\rho = -1$ if and only if $Y{=}aX + b$, for $a < 0, b \in \mathbb{R}$ almost surely. This is since if $|\rho| = 1$, then $h(t_0) = 0$ for some t_0 implies $\mathsf{E}\left(((X - \mathsf{E}(X))t_0 + (Y - \mathsf{E}(Y)))^2\right) = 0$, then $((X - \mathsf{E}(X))t + (Y - \mathsf{E}(Y))) = 0$ almost surely, and we conclude $Y = \mathsf{E}(Y) + t_0\mathsf{E}(X) - t_0X$ almost surely. Careful: for given marginal distributions of X and Y, the bounds ± 1 are not necessarily attained.
4. ρ is not necessarily interpretable. If $\rho \approx 0$, this does not imply that X and Y are close to being independent. Also, if $|\rho| \approx 1$, this also does not entail that X and Y are close to be perfectly dependent. For example, $Y = g(x)$ where g is monotone. This may impose stringent restrictions on the values obtainable on the correlation.

   For example, the third point, in insurance, we know that claims are non-negative. A common distribution is lognormal: suppose that X, Y are both lognormal distributed.

Figure 9: Bounds for ρ as a function of σ, Lognormal distribution.



Another measure of dependence, proposed in 1903 by Spearman, is the Spearman's correlation coefficient ρ_s. Suppose that $X \sim F, Y \sim G$ both continuous distributions. Then

$$\rho_s = \mathsf{cor}(F(x), G(y))$$

This will be no problem with (1), since the variance of a uniform random variable is finite, similarly for (3). The only problematic remains (2).

In the multivariate case, $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$, then $\mathsf{E}(X) = (\mathsf{E}(X_1), \ldots, \mathsf{E}(X_d))^\top$ and for the variance, we look at the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ and $\Sigma_{ij} = \mathsf{Cov}(X_i, X_j)$. You only get the dependence of the pairs.

Provided that $\boldsymbol{\Sigma}$ exists, (that is $\mathsf{E}\left(X_i^2\right) < \infty$) $\forall\, i \in \{1, \ldots, d\}$. Then

- $\boldsymbol{\Sigma}$ is symmetric, since $\mathsf{Cov}(X_i, X_j) = \mathsf{E}(X_i - \mathsf{E}X_i)(X_j - \mathsf{E}X_j) = \mathsf{Cov}(X_j, X_i)$
- $\boldsymbol{\Sigma}$ is positive semi-definite, if $\boldsymbol{h} \in \mathbb{R}^d$, then

$$\begin{aligned}
\boldsymbol{h}^\top \boldsymbol{\Sigma} \boldsymbol{h} &= \sum_{i,j=1}^{d} h_i \mathsf{Cov}(X_i, X_j)\, h_j \\
&= \sum_{i,j=1}^{d} \mathsf{E}\left(h_i (X_i - \mathsf{E}X_i)(h_j (X_j - \mathsf{E}X_j)\right) \\
&= \mathsf{E}\left(\sum_{i,j=1}^{d} h_i (X_i - \mathsf{E}X_i)(h_j (X_j - \mathsf{E}X_j)\right) \\
&= \mathsf{E}\left(\left\{\sum_{i=1}^{d} h_i (X_i - \mathsf{E}X_i)\right\}^2\right)
\end{aligned}$$

- If $\boldsymbol{A} \in \mathbb{R}^{m \times d}$, then $\mathsf{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$

Chapter 5
Moment generating function

Let X be a random variable. As we have seen in some of the examples, it can be tricky and tedious to calculate say the r^{th} moment of a distribution.

Definition 5.1 (Moment generating function)
Let X be a random variable with CDF F. The moment generating function M_X of X is given by

$$M_X(t) = \mathsf{E}\left(e^{tX}\right)$$

provided M_X exists for all $t \in (-\varepsilon, \varepsilon)$ for $\varepsilon > 0$. Then,

$$M_X(t) = \begin{cases} \int_{-\infty}^{\infty} e^{tx} f(x) dx & \text{if } X \text{ has density } f \\ \sum_x e^{tx} f(x) & \text{if } X \text{ is discrete .} \end{cases}$$

Example 5.1
Let $X \sim \mathcal{B}(n, p)$, then

$$\begin{aligned} Ee^{tX} &= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k} \\ &= (pe^t + 1 - p)^n \quad \forall\, t \in \mathbb{R} \end{aligned}$$

Example 5.2
Let $X \sim \mathcal{G}(\alpha, \beta)$ with $\alpha > 0, \beta > 0$ with density given by

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, x > 0$$

Then

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} dx \\ &= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x\left(\frac{1}{\beta} - t\right)} dx \end{aligned}$$

If $\frac{1}{\beta} - t > 0$ if and only if $t < \frac{1}{\beta}$, integral will be finite. By a change of variable with

$y = x\left(\frac{1}{\beta} - t\right), dy = dx = \left(\frac{1}{\beta} - t\right)$, we can take the constants outside the integral to get

$$= \frac{1}{\Gamma(\alpha)\beta^\alpha} \frac{1}{\left(\frac{1}{\beta} - t\right)^\alpha} \int_0^\infty (x\left(\frac{1}{\beta} - t\right)^{\alpha-1} e^{-x\left(\frac{1}{\beta}-t\right)}$$
$$= \frac{1}{\Gamma(\alpha)\beta^\alpha} \frac{1}{\left(\frac{1}{\beta} - t\right)^\alpha} \int_0^\infty y^{\alpha-1} e^{-y} dy$$

using the fact that the integral is that of the Gamma function, equal to $\Gamma(\alpha)$. and so we conclude the moment generating function is $(1 - \beta t)^{-\alpha}$.

Example 5.3

Let X be Pareto distributed $\mathcal{Pa}(\alpha)$ with distribution given by $f(x) = \alpha x^{-\alpha-1}, x \geq 1$. Then the MGF is given by

$$M_X(t) = \int_1^\infty e^{tx} \alpha x^{-\alpha-1} dx$$

which diverges as $x \to \infty$ when $t > 0$. Thus M_X does not exist.

If we look at a dirty calculation, we can express the exponential as a Taylor series

$$\mathsf{E}\left(e^{tX}\right) = \mathsf{E}\left(\sum_{k=0}^\infty \left(\frac{(tX)^k}{k!}\right)\right)$$
$$\stackrel{?}{=} \sum_{k=0}^\infty \frac{(t^k \mathsf{E}\left(X\right)^k)}{k!}$$

Then looking at the derivative,

$$M_X^{(l)}(t) \stackrel{?}{=} \sum_{k=l}^\infty \frac{\mathsf{E}\left(X^k\right)}{k!} k(k-1)\cdots(k-l+1)t^{k-l}$$

and setting $t = 0$, we have

$$M_X^{(l)}(t)\Big|_{t=0} = \mathsf{E}\left(X^l\right)\frac{l!}{l!} = \mathsf{E}\left(X^l\right)$$

Let X be a random variable. Then $M_X(t) = \mathsf{E}\left(e\right)^{tX}$, for $t \in (-\varepsilon, \varepsilon)$.

Exercise 5.1

· If $X \sim \mathcal{P}(\lambda)$, then $M_X(t) = e^{\lambda(e^t-1)}, \quad t \in \mathbb{R}$.

· If $X \sim \mathcal{NB}(r,p)$ then

$$M_X(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r, \qquad \text{for } p \in (0,1),\ t < -\log(1-p)$$

· If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right), \quad t \in \mathbb{R}$$

Theorem 5.2

Suppose that the moment generating function (MGF) exists for all $t \in (-\varepsilon, \varepsilon)$ for $\varepsilon > 0$. Then, $\mathsf{E}\left(X^k\right)$ exists for every $k \in \{1, 2, \ldots\}$ and

$$\mathsf{E}\left(X^k\right) = M_X^{(k)}(0)$$

Remark

If $\mathsf{E}\left(X^k\right)$ does not exist for some $k \in \{1, 2, \ldots\}$, then M_X does not exist. (Student-t, Cauchy, Pareto are among classical examples of distributions who do not have a MGF).

Proof We can write for fixed $t \in (-\varepsilon, \varepsilon)$

$$\begin{aligned}\mathsf{E}\left(e^{tX}\right) &= \mathsf{E}\left(\sum_{l=0}^{\infty} \frac{t^l X^l}{l!}\right) \\ &= \mathsf{E}\left(\lim_{n\to\infty} \sum_{l=0}^{n} \frac{t^l X^l}{l!}\right)\end{aligned}$$

and define f_n by $f_n(x) = \sum_{l=0}^{n} \frac{t^l x^l}{l!}$.

Then

$$\mathsf{E}\left(e^{tX}\right) = \mathsf{E}\left(\lim_{n\to\infty} f_n(X)\right)$$

The question now is whether we could interchange expectations and limit. If it is the case that we have a sequence of non-negative increasing functions, we could do so. However, t is arbitrary. There is nevertheless another way to do this, using the **Dominated Conver-**

gence theorem, namely if $\forall\, n, |f_n(x)| \leq g(x)$ and $\mathsf{E}\left(g(X)\right) < \infty$. In our case

$$\begin{aligned}
|f_n(x)| &\leq \sum_{l=0}^{n} \frac{|t|^l |x|^p}{l!} \\
&\leq \sum_{l=0}^{\infty} \frac{|t|^l |x|^p}{l!} \\
&= e^{|t||x|} \\
&= e^{tX} + e^{-tX} = g(x)
\end{aligned}$$

and it suffices to verify that

$$\begin{aligned}
\mathsf{E}\left(g(X)\right) &= \mathsf{E}\left(e^{tX} + e^{-tX}\right) \\
&= \mathsf{E}\left(e^{tX}\right) + \mathsf{E}\left(e^{-tX}\right) \\
&= M_X(t) + M_X(-t) < \infty
\end{aligned}$$

Hence

$$\begin{aligned}
M_X(t) & \lim_{n\to\infty} \mathsf{E}\left(\sum_{l=0}^{n} \frac{t^l X^l}{l!}\right) \\
&= \lim_{n\to\infty} \sum_{l=0}^{n} \frac{t^l \mathsf{E}\left(X^l\right)}{l!} \\
&= \sum_{l=0}^{\infty} \frac{t^l \mathsf{E}\left(X^l\right)}{l!}
\end{aligned}$$

Hence $M_X(t)$ is a power series, it converges for $t \in (-\varepsilon, \varepsilon)$ and hence it can be differentiated term by term of any order at 0. ■

Example 5.4

If $X \sim \mathcal{G}(\alpha, \beta)$, then the MGF exists and is equal to

$$M_X(t) = (1 - t\beta)^{-\alpha}, \quad t < \frac{1}{\beta}$$

Then

$$\begin{aligned}
\mathsf{E}\left(X\right) &= M_X'(t)|_{t=0} \\
&= (-\alpha)(1 - t\beta)^{-\alpha-1}(-\beta) \\
&= \alpha\beta
\end{aligned}$$

Example 5.5
If $X \sim \mathcal{P}(\lambda)$, then $M_X(t) = e^{\lambda(e^t-1)}$. We can compute the first moment using the MGF as $M'_X(t) = e^{\lambda(e^t-1)}e^t\lambda$ and so $\mathsf{E}(X) = M'_X(0) = \lambda$.

We have seen that if some higher moments do not exist for a distribution, then we do not have a MGF. Is it the case that a MGF necessarily exists for a distribution having all finite moments. Unfortunately not; here is a canonical example.

Example 5.6 (Non-existence of MGF for Lognormal distribution)
Consider $X \sim \mathcal{LN}(0,1)$, $X \stackrel{d}{=} e^Y$ for $Y \sim \mathcal{N}(0,1)$. Then

$$f_X(x) = \frac{1}{\sqrt{2\pi}x}\exp\left(-\frac{\log(x)^2}{2}\right), \quad 0 \le x < \infty$$

and then

$$\mathsf{E}\left(X^k\right) = \int_0^\infty \frac{x^{k-1}}{\sqrt{2\pi}}\exp\left(-\frac{\log(x)^2}{2}\right)dx < \infty$$

But

$$\begin{aligned} M_X(t) &= \int_0^\infty \frac{e^{tx}}{\sqrt{2\pi}x}\exp\left(-\frac{\log(x)^2}{2}\right)dx \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}}\exp\left(tx - \log(x) - \frac{\log(x)^2}{2}\right)dx \end{aligned}$$

goes to ∞ if $t > 0$.

Remark
If $\mathsf{E}\left(X^k\right) < \infty$ for all $k \in \{1,2,\ldots\}$, then $M_X(t)$ does not necessarily exist.

Theorem 5.3
Let X be a random variable so that $\mathsf{E}\left(X^k\right)$ exists for $k \in \{1,2,\ldots\}$. Then

(a) If $\sum_{l=0}^\infty \frac{t^l\mathsf{E}\left(|X|^l\right)}{l!} < \infty$ for $t \in (-\varepsilon,\varepsilon)$ for some $\varepsilon > 0$, then M_X exists on $(-\varepsilon,\varepsilon)$

(b) If X has a bounded support, then M_X exists

Proof Taking

$$\mathsf{E}\left(|X|^l\right) = \int_{-\infty}^\infty |x|^l f(x)dx \le c^l$$

for c some constant. ■

The moment problem: suppose that $\mathsf{E}\left(X^k\right), \mathsf{E}\left(Y^l\right)$ exist for all $k \in \{1, 2, \ldots\}$ and $\mathsf{E}\left(X^k\right) = \mathsf{E}\left(Y^k\right)\ \forall\, k \in \{1, 2, \ldots\}$. Can we conclude that $X \stackrel{d}{=} Y$? Again, not.

Example 5.7

Let $X \sim \mathcal{LN}(0, 1)$ and Y has density

$$\begin{aligned} f_Y(y) &= f_X(y) + f_X(y)\sin(2\pi\log(y)) \\ &= \frac{1}{\sqrt{2\pi}y} e^{-\frac{\ln(y)^2}{2}} (1 + \sin(2\pi\log(y)) \mathbf{1}_{(y>0)} \end{aligned}$$

and $\mathsf{E}\left(X^k\right) = \mathsf{E}\left(Y^k\right)$ for all $k \in \{1, 2, \ldots\}$. One could use a transformation with $t = \log(y) - k$.

Theorem 5.4

Let X and Y be random variables

(a) If M_X and M_Y exist and are equal, *i.e.* $M_X(t) = M_Y(t)\ \forall\, t \in (-\varepsilon, \varepsilon)$, then $X \stackrel{d}{=} Y$.

(b) If X and Y have bounded supports and $\mathsf{E}\left(X^k\right) = \mathsf{E}\left(Y^k\right)$ for all $k \in \{1, 2, \ldots\}$, then $X \stackrel{d}{=} Y$

An interesting property of the MGF M_X is that it characterizes the distribution of X uniquely.

Theorem 5.5

Let X and Y be random variables with MGFs M_X, M_Y. Then

(a) For all $a, b \in R$ and t such that $M_X(at)$ exists,

$$M_{aX+b}(t) = e^{bt} M_X(at)$$

(b) If $X \perp\!\!\!\perp Y$, then for all t for which $M_X(t)$ and $M_Y(t)$ exist,

$$M_{X+Y}(t) = M_X(t) M_Y(t)$$

Proof

(a) $M_{aX+b}(t) = \mathsf{E}\left(e^{(aX+b)t}\right) = \mathsf{E}\left(e^{aXt}e^{b}t\right) = e^{bt}\mathsf{E}\left(e^{atX}\right) = e^{bt}M_X(at)$

(b) If one look at the moment generating function of $X + Y$ at t

$$M_{X+Y}(t) = \mathsf{E}\left(e^{tX+tY}\right) = \mathsf{E}\left(e^{tX}e^{tY}\right) = \mathsf{E}\left(e^{tX}\right)\mathsf{E}\left(e^{tY}\right) = M_X(t)M_Y(t)$$

using independence in the last step, and taking t in the smallest of the two neighborhoods.

■

Note that if e^{bt}, $t \in \mathbb{R}$ is a MGF of a degenerate random variable almost surely equal to b, a constant.

Example 5.8 (Convolution of Poisson using MGF)

Let $X \sim \mathcal{P}(\lambda), Y \sim \mathcal{P}(\mu)$ independent, then $X + Y$ we have shown will be Poisson with parameter $\lambda + \mu$. Then

$$M_{XY}(t) = e^{\lambda(e^t-1)}e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)}$$

and since the MGF determines the distribution function uniquely, then we conclude $X + Y \sim \mathcal{P}(\lambda + \mu)$. This result works well for distributions exhibiting infinite divisibility.

Example 5.9

Let $X \sim \mathcal{G}(\alpha, \beta)$ and $Y \sim \mathcal{G}(\alpha^*, \beta)$, which share the same scale parameter, but can have different shape and such that $X \perp\!\!\!\perp Y$. Then

$$M_{X+Y}(t) = (1 - t\beta)^{-\alpha}(1 - t\beta)^{-\alpha^*} = (1 - t\beta)^{-(\alpha+\alpha^*)}$$

for $t < \frac{1}{\beta}$ so that $X + Y \sim \mathcal{G}(\alpha + \alpha^*, \beta)$.

Suppose now that $Y \sim \mathcal{G}(\alpha^*, \beta^*)$. Then

$$M_{X+Y}(t) = (1 - t\beta)^{-\alpha}(1 - t\beta^*)^{-\alpha^*}$$

for $t < \min(\beta^{-1}, \beta^{*-1})$ which is not so helpful.

More generally, $X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_n$, then $M_{\sum X_i} = \prod_{i=1}^n M_{X_i}(t)$

Exercise 5.2

If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{G}(p)$, then one can check that $\sum -i = 1^n X_i \sim \mathcal{NB}(n, p)$.

We can also define the MGF for random vectors, which will now be a d-placed function, for $\boldsymbol{X} = (X_1, \ldots, X_d)$ and

$$\begin{aligned} M_{\boldsymbol{X}}(t_1, \ldots, t_\alpha) &= \mathsf{E}\left(e^{t_1X_1+\cdots+t_dX_d}\right) \\ &= \mathsf{E}\left(e^{\boldsymbol{t}^\top \boldsymbol{X}}\right). \end{aligned}$$

The MGF of $\boldsymbol{X}$ exists if $\mathsf{E}\left(e^{\boldsymbol{t}^\top \boldsymbol{X}}\right)$ is finite for $t_i \in (-\varepsilon, \varepsilon)$ for all $i \in \{1, \ldots, d\}$ and $\varepsilon = \min\{\varepsilon_i\}$. As in the univariate case, we have

- MGF does not need to exist (for example, the multivariate Student t
- If MGF exists, then it characterizes the distribution of $\boldsymbol{X}$
- Moments can be recovered easily. In the simple case where $d = 2$

$$\left.\frac{\partial^{k+l} M_{\boldsymbol{X}}(t_1, t_2)}{\partial^k t_1 \partial^l t_2}\right|_{t_1=0, t_2=0} = \mathsf{E}\left(X_1^k X_2^l\right)$$

  for $l, k \in \{0, 1, \ldots\}$.
- The MGF of X_i is

$$M_{X_i}(t_i) = \mathsf{E}\left(e^{t_i X_i}\right) = M_{\boldsymbol{X}}(0, \ldots t_i, 0, \ldots, 0)$$

- $X_1, \ldots, X_d$ are independent if and only if $M_{\boldsymbol{X}}(t_1, \ldots, t_d) = \prod_{i=1}^d M_{X_i}(t_i)$. Indeed, we can factor the margins if we have interrelatedness, and since the MGF characterizes the distribution uniquely, then the other side follows.

While the MGF is handy, it is not always defined and doesn't exist for certain distributions. There is something that should remind you of Fourier transforms.

Section 5.1. Characteristic function

Definition 5.6 (Univariate characteristic function)
Let X be a random variable. The the CF of X is given by

$$X_X(t) = \mathsf{E}\left(e^{itX}\right), \quad t \in \mathbb{R}$$

is always defined for any t as $e^{itx} = \cos(tx) + i\sin(tx)$ and $\mathsf{E}\left(e^{itX}\right) = \mathsf{E}\left(\cos(tX)\right) + i\mathsf{E}\left(\sin(tX)\right)$ and since both $\sin, \cos$ are bounded in absolute value by 1, and the norm of e^{itx} is less than one. Indeed, $|\mathsf{E}\left(cos(tX)\right)| \leq 1$ because $|\cos(tX)| \leq 1$ and $\mathsf{E}\left(\sin(tX)\right) \leq 1$. Hence $C_X(t)$ exist $\forall\, t \in \mathbb{R}$, whatever X

Example 5.10
Let X be double exponential (Laplace distribution) with density

$$f_X(x) = \frac{1}{2} e^{-|x|} \mathbf{1}_{(x \in \mathbb{R})}$$

Then

$$\mathsf{E}\left(\sin(tX)\right) = \int_{-\infty}^{\infty} \sin(tx) e^{|x|} \frac{1}{2} dx = 0$$

since $\sin(\cdot)$ is an odd function, there is no contribution from the imaginary part. Now

$$\begin{aligned}\mathsf{E}\left(\cos(tX)\right) &= 2\int_0^\infty \cos(tx)e^{-x}\frac{1}{2}dx \\ &= \frac{1}{1+t^2}\end{aligned}$$

using partial integration twice, therefore $C_X(t) = (1+t^2)^{-1}$.

Example 5.11
Taking $X \sim \mathcal{N}(0,1)$, we have

$$\begin{aligned}C_X(t) &= \int_{-\infty}^\infty e^{itx}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \\ \int_{-\infty}^\infty e^{-\frac{t^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-it)^2}{2}}dx & \\ &= e^{-\frac{t^2}{2}}\end{aligned}$$

completing the square and using integrating the density of a Normal random variable with mean parameter it.

Example 5.12
Let $X \sim \mathcal{C}(0,1)$, the Cauchy distribution, with $C_X(t) = e^{-|t|}$

Proposition 5.7 (Properties of the characteristic functions)

1. The characteristic function (CF) characterizes the distribution of X uniquely.
2. $|C_X(t)| \leq 1$
3. If X has density f, then

$$C_X(t) = \int_{-\infty}^\infty e^{itx}f(x)dx$$

The Fourier transform of f is

$$\int_{-\infty}^\infty e^{-2\pi itx}f(x)dx$$

f can be reconstructed from C_X by means of the inverse Fourier transform. The characteristic function may not be differentiated at 0.

If X has moments up to the order $k \in \{1,2,\ldots\}$ (provided $\mathsf{E}\left(X^k\right)$ exist), if and only if, C_X

is $k-$times differentiable at 0 and

$$\mathsf{E}\left(X^k\right) = i^k C_X^{(k)}(0)$$

If $X \perp\!\!\!\perp Y$, then $C_{X+Y}(t) = C_X(t)C_Y(t)$ and this can be generalized for n independent variables, that is if $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_n$, then

$$C_{X_1+X_2+\cdots+X_n}(t) = \prod_{i=1}^{n} C_{X_i}(t)$$

Considering a sum of Cauchy random variables, then $C_{X_1+X_2}$ will be Cauchy with a different scale parameters.

Melian, Stieltjes and Laplace transforms are other example of transformations that are used frequently in probability theory.

Chapter 6
Exponential families

Definition 6.1
Let $\mathcal{P}^X = \{\mathcal{P}^X_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ be a family of probability distributions on $(\mathbb{R}, \mathbb{B})$. Then $\mathcal{P}^X$ is called a k-parameter **exponential** family if the PDFs/PMFs can be expressed in the form

$$f(x|\boldsymbol{\theta}) = h(x) \cdot c(\boldsymbol{\theta}) \exp\left\{\sum_{j=1}^{k} \omega_j(\boldsymbol{\theta}) \cdot t_j(x)\right\}, \quad \forall\, x \in \mathbb{R}$$

where[7]

1. $\boldsymbol{\theta} \in \Theta$ is a d-dimensional parameter
2. $h(x)$ is a real-valued (measurable) function; $h \geq 0$ and **does not depend on** $\boldsymbol{\theta}$.
3. c is **strictly positive**, $c > 0$ is a real-valued function that **does not depend on** x.
4. $\boldsymbol{t}(x) = (t_1(x), \ldots, t_k(x))$ is a vector of real-valued measurable functions that do not depend on $\boldsymbol{\theta}$.
5. $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_k)$ is a vector of **real-valued** functions that do not depend on x.

Example 6.1
Consider $\mathcal{P}^X = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in (0, \infty)\}$ Then

$$\begin{aligned} f(x|\mu, \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathbf{1}_{(x \in \mathbb{R})} \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{2x\mu}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \mathbf{1}_{(x \in \mathbb{R})} \end{aligned}$$

[7] The condition $\forall\, x \in \mathbb{R}$ is crucial in this definition

Then taking

$$\begin{aligned} h(x) &= \frac{1}{\sqrt{2\pi}}\mathbf{1}_{(x\in R)} \\ c(\mu,\sigma^2) &= \frac{1}{\sigma}\exp\left(-\frac{\mu^2}{2\sigma^2}\right) \\ t_1(x) &= -x^2 \\ t_2(x) &= x \\ \omega_1(\mu,\sigma) &= \frac{1}{2\sigma^2} \\ \omega_2(\mu,\sigma^2) &= \frac{\mu}{\sigma^2} \end{aligned}$$

therefore the Normal random variables belong to an Exponential family.

Example 6.2

Consider $\mathcal{P}^X = \{\mathcal{B}(n,p), p \in (0,1), n \in \{1,2,\ldots\}\}$, but also $\mathcal{P}_n^X = \{\mathcal{B}(n,p), p \in (0,1)\}$ Then

$$\begin{aligned} f(x|n,p) &= \binom{n}{x} p^x (1-p)^{n-x} \mathbf{1}_{(x\in\{0,1,\ldots,n\})} \\ &= \frac{n!}{x!(n-x)!} e^{x\log(p)}(1-p)^n e^{-x\log(n-p)} \end{aligned}$$

We clearly see that $\mathcal{P}_X$ is not an Exponential family if n is a parameter, as we have $\mathbf{1}_{(x\in(0,\ldots,n))}$ depends on x and n. For n fixed, we however have

$$\begin{aligned} h(x) &= \binom{n}{x}\mathbf{1}_{(x\in\{0,1,\ldots,n\})} \\ c(p) &= (1-p)^n \\ t(x) &= x \\ \omega(p) &= \log(p) - \log(1-p) = \log\left(\frac{p}{1-p}\right) \end{aligned}$$

and so $\mathcal{P}_n^X$ is Exponential family.

Example 6.3

Taking $\mathcal{P}^X = \{f(x|\theta), \theta \in (0,\infty)\}$ Then

$$\begin{aligned} f(x|\theta) &= \frac{1}{\theta}\exp\left(1-\frac{x}{\theta}\right)\mathbf{1}_{(x>\theta)} \\ &= \frac{1}{\theta}e^1 e^{-\frac{x}{\theta}}\mathbf{1}_{(x>\theta)} \end{aligned}$$

Whenever the domain depend on θ, there is no chance that the family be exponential.

Recall we had

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} \omega_j(\boldsymbol{\theta}) t_j(\boldsymbol{x})\right)$$

where $c(\boldsymbol{\theta}) > 0$ is strictly positive.

Example 6.4

Consider the case of $\mathcal{N}(\mu, \sigma^2)$, then we can take

$$\begin{aligned} h(x) &= \frac{1}{\sqrt{2\pi}} \\ c(\mu, \sigma^2) &= \frac{1}{\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \\ \omega_1(\mu, \sigma^2) &= \frac{1}{\sigma^2} \\ \omega_2(\mu, \sigma^2) &= \frac{\mu}{\sigma^2} \\ t_1(x) &= -\frac{x^2}{2} \\ t_2(x) &= x \end{aligned}$$

Note

- $h, , c, w, t$ are not unique. We could for example have taken $h(x) = 1, c(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp(-\mu^2/(2\sigma^2))$, $t_1(x) = x^2$) and $\omega_1(\mu, \sigma^2) = -(2\sigma^2)^{-1}$.
- Note that k is not unique. Indeed, consider

$$\begin{aligned} \omega_1(\mu, \sigma^2)t_1(x) + \omega_2(\mu, \sigma^2)t_2(x) &= -\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} \\ &= -\frac{x^2}{2\sigma^2} + \frac{\mu x}{l\sigma^2} + \cdots + \frac{\mu x}{l\sigma^2} \end{aligned}$$

  for some $l \in \mathbb{N}$, the second component appears l times.

- If $x_1, \ldots, x_n$ is an IID sample from $f(\boldsymbol{x}|\boldsymbol{\theta})$, then the likelihood function is

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} h(x_i) \left(c(\boldsymbol{\theta})\right)^n \exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{k} \omega_i(\boldsymbol{\theta}) t_j(x_i)\right\} \end{aligned}$$

and recall from statistics that we can gather all informations regarding the unknown parameters from the sufficient statistic, which in this case, we have

$$\left(\sum_{i=1}^{n} t_i(x_i), \ldots, \sum_{i=1}^{n} t_k(x_i)^l\right) \text{ is sufficient for } \boldsymbol{\theta}$$

Remark

Rather than using the Lebesgue or the counting measure, we could have Exponential family can be defined on $(\mathcal{X}, \mathcal{B})$ and $f(\boldsymbol{x}|\boldsymbol{\theta})$ is a density on $\mathcal{B}_{\boldsymbol{\theta}}^X$ with respect to some measure μ on $(\mathcal{X}, \mathcal{B})$.

Remark

We have

$$\begin{aligned}\int_{-\infty}^{\infty} f(\boldsymbol{x}|\boldsymbol{\theta})dx = 1 \quad &\Leftrightarrow \int_{-\infty}^{\infty} h(x)c(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{k} \omega_i(\boldsymbol{\theta})t_j(x_i)\right\} dx = 1 \\ &\Leftrightarrow c(\theta) = \frac{1}{\int_{-\infty}^{\infty} h(x) \exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{k} \omega_i(\boldsymbol{\theta})t_j(x_i)\right\} dx}\end{aligned}$$

Hence $c(\boldsymbol{\theta})$ depends on θ only through $\omega_1(\boldsymbol{\theta}), \ldots, \omega_k(\boldsymbol{\theta})$. We could consider a new parametrization with new parameters $(\eta_1, \ldots, \eta_k)$ where $\eta_j = \omega_j(\boldsymbol{\theta})$. These parameters are called **canonical** or **natural parameters.** In our previous example, $\eta_1 = \frac{1}{\sigma^2}$ and $\eta_2 = \frac{\mu}{\sigma^2}$ (where we chose $t_1(x) = -\frac{x^2}{2}$, $t_2(x) = x$.

In the canonical parametrization, we have

$$F(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})c^*(\boldsymbol{\eta}) \exp\left(\sum_{j=1}^{k} \eta_j t_j(x)\right)$$

We see that for $(\eta_1, \ldots, \eta_k) \in \{(\omega_1(\boldsymbol{\theta}), \ldots, \omega_k(\boldsymbol{\theta})|\boldsymbol{\theta} \in \Theta\}$. Then $f(\boldsymbol{x}|\boldsymbol{\eta})$ is a valid PDF (or PMF). But there can be more η's; we need

$$\mathcal{H} = \left\{(\eta_1, \ldots, \eta_k) : \int_{-\infty}^{\infty} h(x) \exp\left(\sum_{j=1}^{k} \eta_j t_j(x)\right) dx < \infty\right\}$$

Example 6.5

Again, consider the Normal family $\mathcal{N}(\mu, \sigma^2)$. Our $\mathcal{H}$ corresponds to

$$\begin{aligned}\mathcal{H} &= \left\{(\eta_1, \eta_2) : \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\eta_1\left(-\frac{x^2}{2}\right) + \eta_2 x\right) dx\right\} = (0, \infty) \times \mathbb{R} \\ &= \left\{\left(\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}\right), \sigma^2 > 0, \mu \in \mathbb{R}\right\} = \omega(\Theta)\end{aligned}$$

Example 6.6

Consider the family $\mathcal{N}(\mu, \mu^2)$ with

$$\begin{aligned}f(\boldsymbol{x}|\mu) &= \frac{1}{\sqrt{2\pi}|\mu|} \exp\left(-\frac{x^2}{2\mu^2} + \frac{x\mu}{\mu^2} - \frac{1}{2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{|\mu|} e^{-\frac{1}{2}} e^{-\frac{x^2}{2\mu^2} + \frac{x}{2}}\end{aligned}$$

Cleverly choose

$$\begin{aligned}c(\mu) &= \frac{1}{\sqrt{2\pi}} \frac{1}{|\mu|} e^{-\frac{1}{2}} \\ \omega_1(\mu) &= \frac{1}{\mu^2} \\ \theta_1(x) &= -\frac{x^2}{2} \\ \omega_2 &= \frac{1}{\mu} \\ t_2(x) &= x\end{aligned}$$

and $h(x) = 1$. The natural parameters are $\eta_1 = \frac{1}{\mu^2}, \eta_2 = \frac{1}{\mu}$ and hence

$$\begin{aligned}\mathcal{H} &= \left\{(\eta_1, \eta_2) : \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}\eta_1 + x\eta_2} dx < \infty\right\} \\ &= (0, \infty) \times \mathbb{R} \neq \left\{\left(\frac{1}{\mu^2}, \frac{1}{\mu}\right), \mu \in \mathbb{R}\right\}\end{aligned}$$

and hence conclude that $\omega(\Theta) \subseteq \mathcal{H}$.

Example 6.7

Consider the family of Binomial distributions $\mathcal{B}(n, p)$, with n fixed and $p \in (0, 1)$. Then

$$f(x|p) = \mathbf{1}_{(x \in \{0, \ldots, n\})} \binom{n}{x} (1-p)^n e^{x \ln \frac{p}{1-p}}$$

with $h(x) = \binom{n}{x}$, $c(p) = (1-p)^n$, $t_1(x) = x$, $\omega_1(p) = \ln\left(\frac{p}{1-p}\right)$. The canonical parameter is $\eta = \ln\left(\frac{p}{1-p}\right)$ and

$$\mathcal{H} = \left\{\eta : \sum_{x=0}^{n} \binom{n}{x} e^{x\eta} < \infty\right\} = \{(1+e^{\eta})^n < \infty \ \forall\ \eta \in \mathbb{R}\} = \mathbb{R}$$

Here

$$\mathcal{H} = \left\{\ln\left(\frac{p}{1-p}\right), p \in (0,1)\right\}$$

Theorem 6.2
Let $\mathcal{P}^X$ be a k parametric Exponential family (EF). Then the following are equivalent

1. k is minimal $\{\mathcal{P}^X$ is called a **strictly** k-parametric EF$\}$
2. $\sum_{j=1}^{k} \omega_j(\boldsymbol{\theta}) a_j + a_0 = 0 \ \forall\ \boldsymbol{\theta} \in \Theta$ if and only if $a_0 = a_1 = \cdots = a_k = 0$ (namely $\omega_1, \ldots, \omega_k$ are affine independent).
3. $\sum_{j=1}^{k} t_j(x) a_j = a_0 = 0$ for almost all x if and only if $a_0 = a_1 = \ldots, a_k = 0$ ($t_1, \ldots, t_k$ are affine independent.)

Proof [Sketch] (1) $\Rightarrow$ (2): Suppose that $\sum_{j=1}^{k} \omega_j(\boldsymbol{\theta}) a_j + a_0 = 0 \ \forall\ \theta \in \Theta$, but that $\exists a_j, j \in \{1, \ldots, k\}$ so that $a_j = 0$. WLOG, say $a_k \neq 0$. Then

$$\omega_k(\boldsymbol{\theta}) = -\frac{a_0}{a_k} - \sum_{j=1} \frac{a_j}{a_k} \omega_j(\boldsymbol{\theta})$$

and remark that we could work with $k-1$ parameters. Indeed,

$$\begin{aligned} f(\boldsymbol{x}|\boldsymbol{\theta}) &= h(\boldsymbol{x}) c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k-1} \omega_j(\boldsymbol{\theta}) t_j(\boldsymbol{x}) - \frac{a_0}{a_k} t_k(\boldsymbol{x}) - \sum_{j=1}^{k-1} \frac{a_j}{a_k} \omega_j(\boldsymbol{\theta}) t_k(\boldsymbol{x})\right) \\ &= h(x) \exp\left(-\frac{a_0}{a_k} t_k(\boldsymbol{x})(\boldsymbol{\theta})\right) \exp\left(\sum_{j=1}^{k-1} \omega_j(\boldsymbol{\theta}) t_j(\boldsymbol{x}) - \frac{a_j}{a_k} t_k(\boldsymbol{x})\right) \\ &= h^*(\boldsymbol{x}) c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k-1} \omega_j(\boldsymbol{\theta}) t_j(\boldsymbol{x}) - \frac{a_j}{a_k} t_k(\boldsymbol{x})\right). \end{aligned}$$

This is a $k-1$ parametric Exponential family and so this is a contradiction to (1). ■

Lemma 6.3
The natural (canonical) parameter space $\mathcal{H}$ of a strictly k-parametric EF is **convex** and has a non-empty interior.

Proof We sketch the result for the case $k = 2$. The proof uses convexity and hence Hölder's inequality to show non-empty interior. Then $a_1\eta_1 + a_2\eta_2 = a_0 = 0 \; \forall \; (\eta_1, \eta_2) \in \mathcal{H}$ implying that $a_1 = a_2 = a_0 = 0$, so $\mathcal{H}$ must contain at least three points that do not line on a line. Because the canonical parameter space is convex, the triangle formed by the three points yields a non-empty interior. $k-1$ hyperplane, we have k points which are vertex of a simplex points, so the interior is non-empty. ■

Note

· An Exponential family is called full if

(1) k is minimal (*i.e.* ω's and the t's are affine independent)

(2) $\mathcal{H}$ is an open set in $\mathbb{R}^k$[8]

· An Exponential family is called **regular**, if the above two conditions and furthermore

(3) $\mathcal{H} = \{(\omega_1(\boldsymbol{\theta}), \ldots, \omega_k(\boldsymbol{\theta})\theta \in \Theta\}$

while we say that an Exponential family is **curved** if

$$\{(\omega_1(\boldsymbol{\theta}), \ldots, \omega_k(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta\} \subset \mathcal{H}$$

and example of which is our previous example with $\mathcal{N}(\mu, \mu^2)$.

Section 6.1. Properties of the Exponential family

Definition 6.4
Let $f(\boldsymbol{x}|\boldsymbol{\theta})$ be a PDF/PMF with some d-dimensional parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$. If Θ is open and $\frac{\partial f}{\partial \theta_l}$ exists on Θ for all $l \in \{1, \ldots, d\}$. Then

$$S(\boldsymbol{x}|\boldsymbol{\theta}) = \begin{pmatrix} S_1(\boldsymbol{x}|\boldsymbol{\theta}) \\ \vdots \\ S_d(\boldsymbol{x}|\boldsymbol{\theta}) \end{pmatrix}$$

and $S_l(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{\partial}{\partial \theta^l} \log f(\boldsymbol{x}|\boldsymbol{\theta})$ is called a **score function**.

Note
$S(\boldsymbol{X}|\boldsymbol{\theta})$ is a random vector in $\mathbb{R}^d$.

[8] We thus avoid estimation problems on the boundary of the set.

Lemma 6.5

Under suitable regularity conditions,

$$\mathsf{E}_\theta(S_l(\boldsymbol{X}|\boldsymbol{\theta})) = 0 \quad \forall\, l \in \{1, \ldots, d\}$$

Proof

$$\begin{aligned}
\mathsf{E}_\theta(S_l(\boldsymbol{X}|\boldsymbol{\theta})) &= \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial \theta_l} \log f(\boldsymbol{x}|\boldsymbol{\theta}) \right) f(\boldsymbol{x}|\boldsymbol{\theta}) dx \\
&= \int_{-\infty}^{\infty} \frac{\frac{\partial}{\partial \theta_l} f(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{x}|\boldsymbol{\theta})} f(\boldsymbol{x}|\boldsymbol{\theta}) dx \\
&= \frac{\partial}{\partial \theta_l} \int_{-\infty}^{\infty} f(\boldsymbol{x}|\boldsymbol{\theta}) dx \\
&= \frac{\partial}{\partial \theta_l} 1 = 0
\end{aligned}$$

under the regularity condition, interchanging the integration and differentiation, conditions which can be found in the Casella and Berger book in a section devoted to this matter. ■

If $f(\boldsymbol{x}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$ is an Exponential family, then

$$\log f(\boldsymbol{x}|\boldsymbol{\theta}) = \log(h(x)) + \log(c(\theta)) + \sum_{j=1}^{k} \omega_j(\boldsymbol{\theta}) t_j(\boldsymbol{x})$$

Then by the lemma,

$$\frac{\partial \log(f(\boldsymbol{x}|\boldsymbol{\theta}))}{\partial \theta_l} = \frac{\partial}{\partial \theta_l} \log(c(\boldsymbol{\theta})) + \sum_{j=1}^{k} \frac{\partial \omega_j(\boldsymbol{\theta})}{\partial \theta_l} t_j(\boldsymbol{x})$$

Hence

$$\sum_{j=1}^{k} \frac{\partial \omega_j(\boldsymbol{\theta})}{\partial \theta_l} \mathsf{E}\left(t_j(\boldsymbol{X})\right) = -\frac{\partial}{\partial \theta_k} \log(c(\theta))$$

for $l \in \{1, \ldots, d\}$

$$\mathsf{E}\left(S_l(X|\boldsymbol{\theta})\right) = 0 \; \forall\, l \in \{1, \ldots, d\}$$

For the Exponential family, we have however stronger results, namely

$$\log(f(x|\boldsymbol{\theta})) = \log h(x) + \log(c(\boldsymbol{\theta})) + \sum_{j=1}^{k} \omega_j(\boldsymbol{\theta}) t_j(x)$$

and thus

$$\frac{\partial \log(f(x|\boldsymbol{\theta}))}{\partial \theta_l} = \frac{\partial \log(c(\boldsymbol{\theta})}{\partial \theta_l} + \sum_{j=1}^{k} \frac{\partial \omega_j(\boldsymbol{\theta})}{\partial \theta_l} t_j(x)$$

and taking expectations, we have

$$\mathsf{E}\left(S_l(X|\boldsymbol{\theta})\right) = 0 \Leftrightarrow \sum_{j=1}^{k} \mathsf{E}\left(\frac{\partial \omega_j(\boldsymbol{\theta})}{\partial \theta_l} t_j(X)\right) = -\frac{\partial \log c(\boldsymbol{\theta})}{\partial \theta_l}$$

Example 6.8
Let $X \sim \mathcal{B}(n, p), p \in (0, 1)$. Then

$$\begin{aligned} c(p) &= (1-p)^n \\ \omega(p) &= \log \frac{p}{1-p} \\ t(x) &= x \end{aligned}$$

and thus the expectation of the score function

$$\mathsf{E}\left(\frac{1-p}{p}\frac{1-p+p}{(1-p)^2}X\right) = -\frac{\partial}{\partial p}\left(\log(1-p)^n\right)$$

and upon simplification, we have

$$\begin{aligned} & \mathsf{E}\left(\frac{1}{p(1-p)}X\right) = -n\frac{1}{1-p}(-1) \\ \Leftrightarrow \quad & \frac{1}{p(1-p)}\mathsf{E}\left(X\right) = \frac{n}{1-p} \\ & \Leftrightarrow \mathsf{E}\left(X\right) = np \end{aligned}$$

Example 6.9

Working again with $\mathcal{N}(\mu, \sigma^2)$, we can parametrize the exponential family as

$$\begin{aligned} c(\mu, \sigma^2) &= \frac{1}{\sigma} e^{-\frac{\mu^2}{2\sigma^2}} \\ \omega_1(\mu, \sigma) &= \frac{1}{\sigma^2} \\ \omega_2(\mu, \sigma) &= \frac{\mu}{\sigma^2} \\ t_1(x) &= -\frac{x^2}{2} \\ t_2(x) &= x \end{aligned}$$

and we would get two equations, differentiating with respect to respectively μ and σ

$$\mathsf{E}\left(\frac{1}{\sigma^2} X\right) = \frac{\partial}{\partial \mu}\left(\log(\sigma) + \frac{\mu^2}{2\sigma^2}\right) = \frac{\mu}{\sigma^2}$$

and therefore we conclude that $\mathsf{E}(X) = \mu$. Differentiating with respect this time to σ, one obtains

$$\begin{aligned} & \mathsf{E}\left(\frac{-2}{\sigma^3}\left(\frac{-X^2}{2}\right) - \frac{2}{\sigma^3}\mu X\right) = \frac{1}{\sigma} + \frac{\mu^2}{2}\left(-\frac{2}{\sigma^3}\right) \\ \Leftrightarrow \quad & \frac{1}{\sigma^3}\mathsf{E}\left(X^2\right) - 2\frac{\mu}{\sigma^3}\mathsf{E}(X) = \frac{1}{\sigma} - \frac{\mu^2}{\sigma^2} \end{aligned}$$

and replacing $\mathsf{E}(X)$ by μ, we get

$$\frac{1}{\sigma^3}\mathsf{E}\left(X^2\right) = \frac{1}{\sigma} + \frac{\mu^2}{\sigma^3}$$

and hence multiplying by σ^3, the final result for the second moment is

$$\mathsf{E}\left(X^2\right) = \sigma^2 + \mu^2.$$

Inn the canonical form, we have

$$\log(f(x|\boldsymbol{\eta})) = \log h(x) + \log(c^*(\boldsymbol{\eta})) + \sum_{j=1}^{k} \eta_j t_{ij}(x)$$

and again

$$\frac{\partial \log(f(x|\boldsymbol{\eta}))}{\partial \eta_l} = \frac{\partial \log c^*(\boldsymbol{\eta})}{\partial \eta_l} + t_l(x)$$

and

$$\mathsf{E}\left(t_l(X)\right) = -\frac{\partial}{\partial \eta_l} \log c^*(\boldsymbol{\eta})$$

Example 6.10

Coming back to our example with the $\mathcal{B}(n,p)$, we have

$$\eta = \log \frac{p}{1-p} \Leftrightarrow p = \frac{e^\eta}{1+e^\eta}$$
$$c^*(\eta) = \left(\frac{1}{1+e^\eta}\right)^\eta$$

and

$$\mathsf{E}\left(X\right) = \frac{\partial}{\partial \eta}\left(n \log(1+e^\eta)\right) = n\frac{e^\eta}{1+e^\eta} = np$$

Example 6.11

In the Normal case, we have $\eta_1 = \frac{1}{\sigma^2}, \eta_2 = \frac{\mu}{\sigma^2}$ and

$$c^*(\eta_1, \eta_2) = \eta_1^{-\frac{1}{2}} \exp\left(-\frac{\eta_2^2}{2\eta_1}\right)$$

and taking natural logarithm on both sides, we obtain

$$-\log(c^*(\eta_1,\eta_2)) = -\frac{1}{2}\log(\eta_1) + \frac{\eta^2}{2\eta_1}$$

and the first moment of X is then

$$\mathsf{E}\left(X\right) = \frac{\partial}{\partial \eta_2}\left(-\frac{1}{2}\log(\eta_1)\frac{\eta^2}{2\eta_1}\right) = \frac{\eta_2}{\eta_1}$$

and

$$\mathsf{E}\left(-\frac{X^2}{2}\right) = \frac{\partial}{\partial \eta_1}(\cdots) = -\frac{1}{2\eta_1} - \frac{\eta_2^2}{2\eta_1^2}$$

and

$$\mathsf{E}\left(X^2\right) = \frac{1}{\eta_1} + \frac{\eta^2}{\eta_1^2} = \sigma^2 + \frac{\mu^2}{\sigma^4}\sigma^4$$

Definition 6.6

Let $f(x|\boldsymbol{\theta})$ be a PDF/PMF with a d dimentional parameter $\boldsymbol{\theta}$. the **Fisher-Information**

matrix is a $d \times d$ matrix given by

$$\mathcal{I}(\boldsymbol{\theta}) = e_{\boldsymbol{\theta}}\left(S(X|\boldsymbol{\theta})^{\top} S(X|\boldsymbol{\theta})\right) \mathcal{I}_{il}(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}}\left(S_i(X|\boldsymbol{\theta})^{\top} S_l(X|\boldsymbol{\theta})\right)$$

Lemma 6.7
Under regularity conditions,

$$\mathcal{I}_{i,l}(\boldsymbol{\theta}) = -\mathsf{E}_{\boldsymbol{\theta}}\left(\left[\frac{\partial^2}{\partial\theta_i \partial\theta_l} \log f(X|\boldsymbol{\theta})\right]\right)$$

Proof We know

$$\mathsf{E}\left(S_i(X|\boldsymbol{\theta})\right) = \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta_i} \log(f(x|\boldsymbol{\theta})) \times f(x|\boldsymbol{\theta}) dx$$

Hence

$$\frac{\partial}{\partial\theta_l} \int_{-\infty}^{\infty} \left\{\frac{\partial}{\partial\theta_i} \log f(x|\boldsymbol{\theta})\right\} f(x|\boldsymbol{\theta}) dx = 0$$

which implies in turn, modulo the regularity conditions (interchanging differentiation and integration)

$$\int_{-\infty}^{\infty} \left\{\frac{\partial^2}{\partial\theta_i \partial\theta_l} \log f(X|\boldsymbol{\theta})\right\} f(x|\boldsymbol{\theta}) dx - \left\{\frac{\partial}{\partial\theta_i} \log(f(x|\boldsymbol{\theta})\right\} \frac{\partial}{\partial\theta_l} f(x|\boldsymbol{\theta}) dx = 0$$

and we can rewrite this as

$$-\mathsf{E}\left(\frac{\partial^2}{\partial\theta_i \partial\theta_l} \log f(X|\boldsymbol{\theta})\right) = \int_{-\infty}^{\infty} \left\{\frac{\partial}{\partial\theta_i} \log(f(x|\boldsymbol{\theta})\right\} \frac{\frac{\partial}{\partial\theta_l} f(x|\boldsymbol{\theta})}{f(x|\boldsymbol{\theta})} f(x|\boldsymbol{\theta}) dx$$

multiplying/dividing by $f(x|\boldsymbol{\theta})$ on the right hand side. The right hand side term

$$\frac{\frac{\partial}{\partial\theta_l} f(x|\boldsymbol{\theta})}{f(x|\boldsymbol{\theta})} = \frac{\partial}{\partial\theta_l} \log(f(x|\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}}\left(S_i(X|\boldsymbol{\theta})^{\top} S_l(X|\boldsymbol{\theta})\right)$$

as requested. ■

If $f(x|\boldsymbol{\theta})$ is an Exponential family,

$$\log f(x|\boldsymbol{\theta}) = \log h(x) + \log c(\boldsymbol{\theta}) + \sum_{j=1}^{k} \omega_j(\boldsymbol{\theta}) t_j(x)$$

and so the mixed partials with respect to θ_l, θ_i yield

$$\frac{\partial^2 \log f(x|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_l} = \frac{\partial^2 \log c(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_l} + \sum_{j=1}^{k} \frac{\partial^2}{\partial\theta_i\partial\theta_l}\omega_j(\boldsymbol{\theta})t_j(x)$$

hence

$$-\frac{\partial^2 \log c(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_l} - \sum_{j=1}^{k} \frac{\partial^2}{\partial\theta_i\partial\theta_l}\omega_j(\boldsymbol{\theta})\mathsf{E}\left(t_j(X)\right) = \mathsf{Cov}\left(\sum_{j=1}^{k} \frac{\partial}{\partial\theta_i}\omega_j(\boldsymbol{\theta})t_j(X), \sum_{j=1}^{k} \frac{\partial}{\partial\theta_l}\omega_j(\boldsymbol{\theta})t_j(X)\right)$$

In the canonical parametrization

$$-\frac{\partial^2}{\partial\eta_i\partial\eta_l} \log c^*(\boldsymbol{\eta}) = \mathsf{Cov}\left(t_i(X), t_l(X)\right)$$

and

$$-\frac{\partial^2}{\partial^2\eta_i} \log c^*(\boldsymbol{\eta}) = \mathsf{Var}\left(t_i(X)\right)$$

Example 6.12

For the Binomial example,

$$\begin{aligned}
\mathsf{Var}X &= -\frac{\partial^2}{\partial^2\eta}\left(\log\left(\frac{1}{1+e^\eta}\right)^n\right) \\
&= \frac{\partial^2}{\partial^2\eta}(n\log(1+e^\eta)) \\
&= n\frac{\partial}{\partial\eta}\frac{e^\eta}{1+e^\eta} \\
&= n\frac{e^\eta + e^{2\eta} - e^{2\eta}}{(1+e^\eta)^2} \\
&= n\frac{e^\eta}{1+e^\eta}\frac{1}{1+e^\eta} \\
&= np(1-p)
\end{aligned}$$

Example 6.13

Let $X \sim \mathcal{P}(\lambda)$; the Poisson is a member of the Exponential family. We have

$$f(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}\mathbf{1}_{(x\in\{0,1,\ldots\})}$$

choosing $h(x) = \frac{1}{x!}\mathbf{1}_{(x\in\{0,1,\ldots\})}$, $c(\lambda) = e^{-\lambda}$ and $\log(\lambda) = \omega(\lambda)$ and so $\eta = \log\lambda \Leftrightarrow \lambda = e^\eta$

and $c^*(\eta) = e^{-e^\eta}$ and so

$$\mathsf{E}(X) = -\frac{\partial}{\partial \eta}(\log c^*(\eta))$$
$$\frac{\partial}{\partial \eta} e^\eta = \lambda$$

and $\mathsf{Var}(X) = (e^\eta)'' = e^\eta = \lambda$

Notice that we have $\eta = \log \frac{p}{1-p}$ for the binomial and $\eta = \log \lambda$ will come again in generalized linear model course; they are the link functions.

Recall that if $f(x|\boldsymbol{\theta})$ is an Exponential family and $\boldsymbol{X}$ is a random sample from $f(x|\boldsymbol{\theta})$, then

$$\sum_{i=1}^n t_i(X_i, \ldots, \sum_{i=1}^n t_k(X_i)$$

is sufficient for $\boldsymbol{\theta}$. You will see that efficient estimators will be functions of the sufficient statistics.

But for $j \in \{1, \ldots, d\}$,

$$\begin{aligned}\mathsf{E}\left(\sum_{i=1}^n t_j(X_i)\right) &= n\mathsf{E}(t_j(X_1)) \\ &= n\left(-\frac{\partial}{\partial \eta_j}\log c^*(\eta)\right)\end{aligned}$$

and

$$\mathsf{Var}\left(\sum_{i=1}^n t_j(X_i)\right) = n\mathsf{Var}(t_j(X_1))$$
$$\mathsf{Cov}\left(\sum_{i=1}^n t_j(X_i), \sum_{i=1}^n t_l(X_i)\right) = n\mathsf{Cov}(t_j(X_1), t_l(X_1))$$

Section 6.2. Exponential tilting

Suppose that $f(x)$ is a PMF/PDF whose MGF exists in some neighborhood of O, say N. Then, $\forall\, t \in N$,

$$M(t) = E^{e^{tX}} = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

We can introduce the so-called cumulant generating function,

$$K(t) = \log(M(t))$$

one recovers by differentiating and evaluating at zero, we recover the cumulants, namely the central moments for the given distribution. We can write

$$e^{K(t)} = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

and we get

$$1 = \int_{-\infty}^{\infty} e^{tx - K(t)} f(x) dx$$

and the integrand is non-negative, therefore

$$f^*(x|t) = e^{tx - K(t)} f(x)$$

for $x \in \mathbb{R}, t \in N$ is a family of PDFs/PMFs. The family $f^*(x|t)$ is called the **exponential tilting of** f

Setting $t = 0$, we get $f^*(x|t) = f(x)$; the exponential tilting of f is an EF. and

$$f^*(x|t) = f(x) e^{-K(t)} e^{tx}$$

is already in canonical parametrization ,with $f(x) = h(x)$ and $e^{-K(t)} = c(t)$.

Let $f(x)$ be a density or PMF. This density (PMF) may not have a closed form, but it has a MGF which is available in closed form. This approximation is not a one-to-one correspondence, but may work well in certain cases.

Suppose M_X exists in a neighborhood N of O; $\forall\ t \in N : K(t) = \log(M(t))$. If we take $f(x)e^{tx - K(t)} = f(x|t)\ \forall\ x \in \mathbb{R}, t \in N$. $f(x|t)$ is thus a valid PMF/PDF.

- The value of the parameter $t = 0$ gives us $f(x)$.
- $\{f(x|t), t \in N\}$ is a one-parameter Exponential family in a canonical form.

If X_t has density/PMF $f(x|t)$ then

$$\begin{aligned} \mathsf{E}\,(X_t) &= \frac{\partial}{\partial t} \left\{ -\log(e^{-K(t)} \right\} \\ &= \frac{\partial}{\partial t} K(t) = \dot{K}(t) \end{aligned}$$

If we wanted to compute the variance,

$$\mathsf{E}\left(\mathsf{Var}\left(X_t\right)\right) = \frac{\partial^2}{\partial t^2}\left\{K(t)\right\} = \ddot{K}(t)$$

Let M_{X_t} denote the MGF of X_t. Then

$$\begin{aligned} M_{X_t}(s) = \mathsf{E}\left(e^{X_t s}\right) &= \int_{-\infty}^{\infty} e^{xs} f(x|t) dx \\ &= \int_{-\infty}^{\infty} e^{xs} f(x) e^{tx - K(t)} dx \\ &= e^{-K(t)} \int_{-\infty}^{\infty} e^{x(s+t)} f(x) dx \\ &= \frac{M(s+t)}{e^{K(t)}} \\ &= \frac{M(s+t)}{M(t)} \end{aligned}$$

for s sufficiently small so that $s + t \in N$.

Saddle point approximation

If we are in the presence of a unimodal absolutely continuous distribution, then we can use the saddle point approximation, and use the density of the Normal to approximate the density around the mode of the distribution.

1. Fix $x_0 \in \mathbb{R}$
2. Compute t_{x_0} so that $\dot{K}(t_{x_0}) = x_0$.
3. Approximate $f(x_0|t)$ by a Normal density at x_0

$$\begin{aligned} f(x_0|t) &\approx \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\ddot{K}(t_{x_0})}} \exp\left(-\frac{(x_0 - x_0)^2}{2\ddot{K}(t_{x_0})}\right) \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\ddot{K}(t_{x_0})}} \end{aligned}$$

4. This gives the approximation

$$f(x_0) \approx \exp(K(t_{x_0}) - t_{x_0} x_0) \frac{1}{\sqrt{2\pi \ddot{K}(t_{x_0})}}$$

This approximation works well for sums of random variables, especially in the IID case, of the form $S_n = X_1 + \cdots + X_n$.

Section 6.3. Exponential dispersion models

Observe that $\{f(x|\boldsymbol{\eta}), \eta \in \mathcal{H}\}$ is a family of PDF/PMFs member of the Exponential family canonical form. If we look at $\mathsf{E}\left(t_j(X)\right) = \frac{\partial}{\partial \eta_j}(-\log c^*(\boldsymbol{\eta}))$ and

$$\mathsf{Cov}\left(t_j(X), t_l(X)\right) = \frac{\partial^2}{\partial \eta_j \partial \eta_l}(-\log(c^*(\boldsymbol{\eta})))$$

For example, in the case of a Poisson distribution, we may see overdispersion, yet getting an estimate of λ, we know the underlying variance parameter which doesn't necessarily match the observed sample variance.

We can generalize the form of the PMF/PDF to

$$\begin{aligned} f(x|\boldsymbol{\theta}, \boldsymbol{\phi}) &= \exp\left\{ d(x, \boldsymbol{\phi}) + \frac{\log c(\boldsymbol{\theta})}{r(\boldsymbol{\phi})} + \frac{1}{r(\boldsymbol{\phi})} \sum_{j=1}^{k} \omega_j(\boldsymbol{\theta}) t_j(x) \right\} \\ &= e^{d(x|\boldsymbol{\phi})} \left\{c(\boldsymbol{\theta})\right\}^{\frac{1}{r(\boldsymbol{\phi})}} \exp\left(\frac{1}{r(\boldsymbol{\phi})} \sum_{j=1}^{k} \omega_(\boldsymbol{\theta}) t_j(x) \right) \end{aligned}$$

where $\boldsymbol{\phi}$ is the dispersion parameter; in canonical parametrization, this is of the form

$$f(x|\boldsymbol{\eta}, \boldsymbol{\phi}) = e^{d(x|\boldsymbol{\phi})} \left\{c^*(\boldsymbol{\theta})\right\}^{\frac{1}{r(\boldsymbol{\phi})}} \exp\left(\frac{1}{r(\boldsymbol{\phi})} \sum_{j=1}^{k} \boldsymbol{\eta}_j t_j(x) \right)$$

and taking logarithms, we get

$$\log f(x|\boldsymbol{\eta}, \boldsymbol{\phi}) = d(x|\boldsymbol{\phi}) + \frac{1}{r(\boldsymbol{\phi})} \log(c^*(\boldsymbol{\eta})) + \frac{1}{r(\boldsymbol{\phi})} \sum_{j=1}^{k} \boldsymbol{\eta}_j t_j(x)$$

and

$$\frac{\partial \log f(x|\boldsymbol{\eta}, \boldsymbol{\phi})}{\partial \eta_j} = \frac{1}{r(\boldsymbol{\phi})} \frac{\partial}{\partial \eta_j} \log(c^*(\boldsymbol{\eta}) + t_j(x) \frac{1}{r(\boldsymbol{\phi})}$$

and

$$\mathsf{E}\left(t_j(X)\right) = -\frac{\partial}{\partial \eta_j} \log(c^*(\boldsymbol{\eta}))$$

but more interestingly now

$$\mathsf{Cov}\left(t_j(X), t_l(X)\right) = r(\boldsymbol{\phi}) \frac{\partial^2}{\partial \eta_j \partial \eta_l} (-\log(c^*(\boldsymbol{\eta})))$$

Chapter 7
Location and scale families

Lemma 7.1
Let f be a PDF, with $\mu \in \mathbb{R}, \sigma > 0$. Then

$$g(x|\mu, \sigma) = f\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}$$

is a valid PDF.

Proof Clearly, g is non-negative. It remains to show that it integrates to 1. Indeed,

$$\int_{-\infty}^{\infty} f\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}dx = \int_{-\infty}^{\infty} f(t)dt = 1$$

making the change of variable $t = (x-\mu)/\sigma$. ■

Remark
If X has PDF f, then $\sigma X + \mu$ has PDF g (this is a linear transformation). If X is discrete, with support S, then $\sigma X + \mu$ is also discrete, with support $\{\sigma s = \mu, s \in S\}$ and [9]

$$\mathsf{P}(\sigma X + \mu) = \mathsf{P}(\sigma s + \mu) = \mathsf{P}(X = s) = f(s) = f\left(\frac{x-\mu}{\sigma}\right)$$

We call μ the **location parameter** and σ the **scale parameter.**

Example 7.1 (Family of Normal distributions)
Consider the family of $\mathcal{N}(0,1)$ variables; then $\{g(x|\mu,\sigma), \mu \in \mathbb{R}, \sigma > 0\}$ gives the $\mathcal{N}(\mu, \sigma^2)$ random variables with PDF

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

Example 7.2 (Fréchet distribution)
This distribution has no MGF and has PDF given by

$$f(x) = \frac{\alpha}{x^{\alpha+1}}\exp\left(-\left(\frac{1}{x}\right)^{\alpha}\right)\mathbf{1}_{(x>0)}$$

[9] In this case, there is no σ^{-1} parameter arising from the Jacobian of the transformation.

for $\alpha > 0$. The corresponding location and scale form is

$$g(x|\mu,\sigma,\alpha) = \frac{1}{\sigma}\frac{\alpha}{\left(\frac{x-\mu}{\sigma}\right)^{\alpha+1}}\exp\left(-\left(\frac{\sigma}{x-\mu}\right)^{\alpha}\right)$$

This is an example of heavy-tailed distribution.

Definition 7.2 (Location family)
Let f be any PDF. The family of distributions with densities $f(x-\mu), \mu \in \mathbb{R}$ is called a location family with parameter μ.

Observation
The mean of $X+\mu$) is $\mathsf{E}(X+\mu) = \mathsf{E}(X) + \mu$ if $\mathsf{E}(X)$ exists. In particular, if $\mathsf{E}(X) = 0$, then μ is the expectation of $X+\mu$ with density $f(\cdot - \mu)$.

Example 7.3
The family $\mathcal{N}(0,1)$ yields $\mathcal{N}(\mu,\sigma^2)$. However, we loose this interpretation when the expectation does not exist. For, given $f(x) = \frac{1}{\pi(1+x^2)}$, the Cauchy distribution, then $\mathsf{E}(X)$ doesn't exist, by the distribution $f(x-\mu) = \frac{1}{\pi(1+(x-\mu)^2)}$ for $\mu \in \mathbb{R}$. However, if X is a random variable with PDF f has median x_{med}

$$\mathsf{P}(X \leq x_{\text{med}}) = \mathsf{P}(X \geq x_{\text{med}}) = \frac{1}{2}$$

If Y has PDF $f(\cdot - \mu)$, then its CDF is

$$\mathsf{P}(Y \leq x) = F(x-\mu)$$

where $F()$ is the CDF of X, thus $x_{\text{med}} + \mu$ is the median of Y. If $x_{\text{med}} = 0$, then μ is the median of Y.

Definition 7.3 (Scale family)
Let f be an arbitrary PDF. The family of densities

$$f\left(\frac{x}{\sigma}\right)\frac{1}{\sigma}, \quad x \in \mathbb{R}, \sigma > 0$$

is called the scale family with scale parameter σ. If X has PDF f , then $Y = X\sigma$ has PDf $f\left(\frac{\cdot}{\sigma}\right)\frac{1}{\sigma}$.

Example 7.4
If $f : \mathcal{N}(0,1), \rightsquigarrow \mathcal{N}(0,\sigma^2)$. If $f : \mathcal{E}(1), \rightsquigarrow e^{-\frac{x}{\sigma}}\frac{1}{\sigma} \sim \mathcal{E}\left(\frac{1}{\sigma}\right)$. We can generalize also to

$$f : \mathcal{G}(\alpha) \rightsquigarrow \frac{1}{\Gamma(\alpha)}\frac{1}{\sigma}\frac{x^{\alpha-1}}{\sigma^{\alpha-1}}e^{-\frac{x}{\sigma}}$$

which is equivalent to writing

$$\frac{1}{\Gamma(\alpha)}\frac{x^{\alpha-1}}{\sigma^{\alpha}}e^{-\frac{x}{\sigma}}$$

which is the distribution of a $\mathcal{G}\left(\alpha, \frac{1}{\sigma}\right)$.

Observation (Interpretation of σ)

$$\mathsf{Var}\left(Y\right) = \mathsf{Var}\left(\sigma X\right) = \sigma^2 \mathsf{Var}\left(X\right)$$

provided that $\mathsf{Var}\left(X\right)$ exists. If $\mathsf{Var}\left(X\right)$=1, then $\sigma^2 = \mathsf{Var}\left(Y\right)$. Again, consider central moment $\mathsf{E}\left(|Y - y_{\text{med}}|\right)$ and observe first that Y has CDF $F_Y(y) = F_X\left(\frac{y}{\sigma}\right)$. If x_{med} is the median of X then $\sigma x_{\text{med}} = y_{\text{med}}$ so that

$$\mathsf{E}\left(|\sigma X - \sigma x_{\text{med}}|\right) = \sigma \mathsf{E}\left(|X - x_{\text{med}}|\right)$$

and the latter may or not exist.

Chapter 8
Hierarchical models

Let (X, Y) be a random vector. Observe that

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

We could thus construct the joint model from the conditional distribution, specify the distribution for y and get the joint distribution. We start with an example to illustrate this fact.

Example 8.1

Suppose we have a fruit fly. How many eggs will hatch out of the laid eggs?

Level 2: Take $N \sim \mathcal{P}(\lambda)$

Level 1: Given the number n of eggs laid, the number of eggs hatched, X can be modeled using $\mathcal{B}(n,p)$. Yet the number of eggs laid is unknown.

First, remark that X is discrete, with support $\{0, 1, \ldots\}$ and we could try to compute $\Pr(X = k)$, knowing that $k < n$

$$\begin{aligned}
\mathsf{P}(X=k) &= \sum_{n=k}^{\infty} \mathsf{P}(X=k, N=n) \\
&= \sum_{n=k}^{\infty} \mathsf{P}(X=k|N=n)\mathsf{P}(N=n) \\
&= \sum_{n=0}^{\infty} e^{-\lambda}\frac{\lambda^n}{n!}\binom{n}{k}p^k(1-p)^{n-k} \\
&= \sum_{n=k}^{\infty} e^{-\lambda}\frac{\lambda^n}{n!}\frac{n!}{k!(n-k)!}\left(\frac{p}{1-p}\right)^k(1-p)^n \\
&= \frac{e^{-\lambda}}{k!}p^k\lambda^k\sum_{n=k}^{\infty}\frac{1}{(n-k)!}\{\lambda(1-p)\}^{n-k} \\
&= \frac{e^{-\lambda}}{k!}p^k\lambda^k e^{\lambda(1-p)} \\
&= \frac{e^{-\lambda p}}{k!}(p\lambda)^k
\end{aligned}$$

and so $X \sim \mathcal{P}(\lambda p)$.

Remark

Recall if X, Y are random variables, $\mathsf{E}(X|Y=y) = \int_{-\infty}^{\infty} x f_{X|Y=y}(x)dx$ or in the discrete

case $\sum x f_{X|Y=y}(x)$, which are both functions of y, say $h(y)$. One can then show that $h(y)$ is a measurable mapping.

$h(Y) = \mathsf{E}(X|Y)$ is then a new random variable, which is a function of y. One could think of this as a projection onto the space of functions of $g(y)$ for g measurable.

Lemma 8.1 (Law of iterated expectation)
$\mathsf{E}(h(Y)) = \mathsf{E}(\mathsf{E}(X|Y)) = \mathsf{E}(X)$

Proof

$$\begin{aligned}
\mathsf{E}(\mathsf{E}(X|Y)) &= \mathsf{E}\left(\int_{-\infty}^{\infty} x \frac{f(x,Y)}{f_Y(Y)} dx\right) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \frac{f(x,y)}{f_Y(y)} dx f_Y(y) dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y) dx dy \\
&= \int_{-\infty}^{\infty} x f_X(x) dx \\
&= \mathsf{E}(X)
\end{aligned}$$

swapping the integrals ■

Remark
Note that

$$\mathsf{E}(g(X,Y)) = \mathsf{E}(\mathsf{E}(g(X,Y)|Y))$$

Remark
We also have

$$\min_g \mathsf{E}(X - g(Y))^2 = \mathsf{E}(X - \mathsf{E}(X|Y))^2$$

Example 8.2
We can use the above remarks to get the first moment in the previous example.

$$\mathsf{E}(X) = \mathsf{E}(\mathsf{E}(X|N)) = \mathsf{E}(pN) = p\mathsf{E}(N) = p\lambda.$$

We could also if one is as lucky to get the moment generating function provided it exists.

$$M_X(t) = \mathsf{E}\left(e^{tX}\right) = \mathsf{E}\left(\mathsf{E}\left(e^{tX|N}\right)\right)$$

and the inner expectation is the MGF of the Binomial distribution

$$\mathsf{E}\left(e^{tX|N=n}\right) = (1-p+pe^t)^n$$

thus taking N random

$$\begin{aligned}\mathsf{E}\left((1-p+pe^t)^N\right) &= \mathsf{E}\left(e^{N\log(1-p+pe^t)}\right)\\ &= e^{\lambda e^{\log(1-p+pe^t)}-1}\\ &= e^{\lambda(1-p+pe^t-1)}\\ &= e^{p\lambda(e^t-1)}\end{aligned}$$

the MGF of a Poisson distribution with parameter $p\lambda$.

In general, a hierarchical model is specified in terms of a hierarchy. At level k, $\boldsymbol{X}_k(= X_{k_1},\ldots,X_{k_{d_k}})$ has some specified distribution (may have parameters). At level $k-1$, we have $\boldsymbol{X}_{k-1}|\boldsymbol{X}_k$ has some distribution. The dimension of the vector needs not be the same at every level. Similarly, at level $k-2$, we have $\boldsymbol{X}_{k-2}|\boldsymbol{X}_k,\boldsymbol{X}_{k-1}$. Repeating until level 1, where we have $\boldsymbol{X}_1|\boldsymbol{X}_2,\ldots,\boldsymbol{X}_k$.

We are thus writing this

$$f_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k) = f_{\boldsymbol{X}_k}(\boldsymbol{x}_k)f_{\boldsymbol{X}_k|\boldsymbol{X}_{k-1}}(\boldsymbol{x}_{k-1})\cdots f_{\boldsymbol{X}_1|\boldsymbol{X}_2,\ldots,\boldsymbol{X}_k}(\boldsymbol{x}_1)$$

Note that we could have mixture of densities and counting densities, so that the above is a mix of Lebesgue measure and counting measure.

Example 8.3

We could use the example of fruit flies and take it a little further.
Level 3: $\Lambda \sim \mathcal{G}(\alpha,\beta)$,
Level 2: $N|\Lambda=\lambda \sim \mathcal{P}(\lambda)$
Level 1: $X|N=n \sim \mathcal{B}(n,p)$
For this hierarchy, we would have in Level 3 a random probability of hatching, specified as $P \sim \mathcal{B}(\alpha^*,\beta^*)$ so that we actually have $X|N=n,P=p$ is $\mathcal{B}(n,p)$ in the first level.

X is discrete $\{0,1,\ldots\}$ and

$$f_X(x) = \sum_{n=0}^{\infty}\int_0^{\infty} f_{X|N=n}(x)f_{N|\Lambda=\lambda}(n)f_\Lambda(\lambda)d\lambda$$

which we will not attempt to compute. If we instead look at the distribution of N, with

support is $\{0, 1, \ldots\}$

$$\begin{aligned} f_N(n) &= \int_0^\infty f_{N|\Lambda=\lambda}(n) f_\Lambda(\lambda) d\lambda \\ &= \int_0^\infty e^{-\lambda} \frac{\lambda^n}{n!} \frac{1}{\Gamma(\alpha)\beta^\alpha} \lambda^{\alpha-1} e^{-\frac{\lambda}{\beta}} d\lambda \\ &= \frac{1}{n!\Gamma(\alpha)\beta^\alpha} \int_0^\infty \lambda^{n+\alpha-1} e^{-\lambda(1+\frac{1}{\beta})} d\lambda \end{aligned}$$

and upon making the change of variable $t = \lambda\left(1 + \frac{1}{\beta}\right)$ yields the Gamma function

$$\frac{1}{n!\Gamma(\alpha)\beta^\alpha} \frac{1}{\left(1+\frac{1}{\beta}\right)^{n+\alpha-1}} \int_0^\infty t^{n+\alpha-1} e^{-t} dt$$

$$= \frac{\Gamma(n+\alpha)}{n!\Gamma(\alpha)} \left(\frac{\beta}{\beta+1}\right)^n \left(\frac{1}{\beta+1}\right)^\alpha$$

and in the particular case where $\alpha \in \mathbb{N}, \alpha = r$

$$\begin{aligned} \mathsf{P}(N=n) &= \frac{(n+r-1)!}{n!(r-1)!} \left(\frac{\beta}{\beta+1}\right)^n \left(\frac{1}{\beta+1}\right)^r \\ &= \frac{(n+r-1)!}{n!(r-1)!} p^r (1-p)^n \end{aligned}$$

which is a Negative binomial. The one in the case where $\alpha > 0$ is not an integer yields that N is distributed according to the Extended negative binomial $(\alpha, (1+\beta)^{-1})$. It turns out that $X \sim E\mathcal{NB}(\alpha, (1+\beta p)^{-1})$ (using the moment generating function). Now we can compute the expectation

$$\begin{aligned} \mathsf{E}(X) &= \mathsf{E}(\mathsf{E}(X|N)) \\ &= \mathsf{E}(Np) = p\mathsf{E}(N) \\ &= p\mathsf{E}(\mathsf{E}(N|\Lambda)) \\ &= p\mathsf{E}(\Lambda) = p\alpha\beta \end{aligned}$$

Compare this to $\mathsf{E}(X) = p\lambda$ in the previous example. Comparing the hierarchy $N \sim \mathcal{P}(\lambda), X|N = n \sim \mathcal{B}(n,p)$ to $\Gamma \sim \mathcal{G}(\alpha,\beta)$, $N|\Lambda = \lambda \sim \mathcal{P}(\lambda)$ and $X|N = n \sim \mathcal{B}(n,p)$.

We had in the first case $X_1 \sim \mathcal{P}(\lambda p)$ versus $X_2 \sim E\mathcal{NB}(\alpha, (1+\beta p)^{-1})$, with $\mathsf{E}(X_1) = p\lambda$ and $\mathsf{E}(X_2) = p\alpha\beta$ with equality of the first moment if $\lambda = \lambda\beta$. For X_1, clearly $\mathsf{Var}(X_1) = p\lambda$, but the variance of X_2 will be bigger. This comes from the following observation.

Lemma 8.2 (Iterated variance formula)
For any two random variables X and Y,

$$\mathsf{Var}\left(X\right) = \mathsf{E}\left(\mathsf{Var}\left(X|Y\right)\right) + \mathsf{Var}\left(\mathsf{E}\left(X|Y\right)\right)$$

Proof

$$\begin{aligned}\mathsf{Var}\left(X\right) &= \mathsf{E}\left(X - \mathsf{E}\left(X\right)\right)^2 \\ &= \mathsf{E}\left(X \pm \mathsf{E}\left(X|Y\right) - \mathsf{E}\left(X\right)\right)^2 \\ &= \mathsf{E}\left(X - \mathsf{E}\left(X|Y\right)\right)^2 + \mathsf{E}\left(\mathsf{E}\left(X|Y\right) - \mathsf{E}\left(X\right)\right)^2 + 2\mathsf{E}\left((X - \mathsf{E}\left(X|Y\right))(\mathsf{E}\left(X|Y\right) - \mathsf{E}\left(X\right)\right)\end{aligned}$$

where the second term has $\mathsf{E}\left(X|Y\right)$ which we can write as $h(Y)$. Thus the second term is equal to $\mathsf{Var}\left(h(Y)\right) = \mathsf{Var}\left(\mathsf{E}\left(X|Y\right)\right)$, while the third term is equal to

$$2\mathsf{E}\left((X - h(Y))(h(Y) - \mathsf{E}\left(X\right))\right) = 2\mathsf{E}\left(\mathsf{E}\left(X|Y\right)\right) = 0$$

and so

$$\begin{aligned}\mathsf{E}\left((X - h(Y))(h(Y) - \mathsf{E}\left(X\right)|Y = y\right) &= (h(y) - \mathsf{E}\left(X\right))\mathsf{E}\left(X - h(y)|Y = y\right) \\ &= (h(y) - \mathsf{E}\left(X\right))\mathsf{E}\left(X|Y = y\right) - h(y) = h(y) - h(y) = 0\end{aligned}$$

and the first term is, conditioning,

$$\mathsf{E}\left(X - h(Y)\right)^2 = \mathsf{E}\left(\mathsf{E}\left((X - h(Y))^2|Y\right)\right)$$

and

$$\mathsf{E}\left((X - h(Y))^2|Y = y\right) = \mathsf{Var}\left(X|Y = y\right)$$

and so the third term is $\mathsf{E}\left(\mathsf{Var}\left(X|Y\right)\right)$. ■

In the case of the Poisson-Binomial hierarchy,

$$\begin{aligned}\mathsf{Var}\left(X\right) &= \mathsf{E}\left(\mathsf{Var}\left(X|N\right)\right) + \mathsf{Var}\left(\mathsf{E}\left(X|N\right)\right) \\ &= \mathsf{E}\left(Np(1-p)\right) + \mathsf{Var}\left(Np\right) \\ &= p(1-p)\mathsf{E}\left(N\right) + p^2\mathsf{Var}\left(N\right) \\ &= \lambda p\end{aligned}$$

and for the Gamma-Poisson-Binomial hierarchy,

$$\begin{aligned}\mathsf{Var}(X) &= p(1-p)\mathsf{E}(N) + p^2\mathsf{Var}(N)\\ &= p(1-p)\mathsf{E}(\mathsf{E}(N|\Lambda)) + p^2\left[\mathsf{E}(\mathsf{Var}(N|\Lambda)) + \mathsf{Var}(\mathsf{E}(N|\Lambda))\right]\\ &= p(1-p)\mathsf{E}(\Lambda) + p^2[\mathsf{E}(\Lambda) + \mathsf{Var}(\Lambda)]\\ &= \mathsf{E}(\Lambda)\,p + p^2\mathsf{Var}(\Lambda)\\ &= p\alpha\beta + p^2\alpha\beta^2\end{aligned}$$

so completing our example, $\mathsf{Var}(X_2) = p\alpha\beta + p^2\alpha\beta^2$.

We discuss two more examples of hierarchical models before continuing with inequalities.

Example 8.4

Consider an insurance company who incur claims, we are interested in the total amount of claim size. Assuming that the pool of people is homogeneous, let X be the total loss at the end of the year. If we knew the total number of claims, we could write this as $\sum_{i=1}^{n} X_i$. However, the number of claims is unknown, so we are really interested in

$$S_N = \sum_{i=1}^{N} X_i.$$

Suppose that we model this with a two-level model hierarchy:
At level 2, say N follows a discrete distribution on $\{0, 1, \ldots\}$, such as the Poisson with parameter λ and the negative binomial with parameters (r, p).

At the first level, given $N = n$,

$$S_n \overset{\mathrm{d}}{=} \sum_{i=1}^{n} X_i, \qquad X_i \overset{\mathrm{iid}}{\sim} F,\ X_i \geq 0$$

If $N \sim \mathcal{P}(\lambda)$, this is called a **compound Poisson** distribution.

If we are interested in

$$\mathsf{E}(S_N) = \mathsf{E}(\mathsf{E}(S_N|N)) = \mathsf{E}(N\mathsf{E}(X_i)) = \mathsf{E}(N)\,\mathsf{E}(X_i)$$

We could also easily calculate the variance

$$\begin{aligned}\mathsf{Var}(S_N) &= \mathsf{E}(\mathsf{Var}(S_N|N)) + \mathsf{Var}(\mathsf{E}(N|S_N))\\ &= \mathsf{E}(N\mathsf{Var}(X_i)) + \mathsf{Var}(N\mathsf{E}(X_i))\\ &= \mathsf{Var}(X_i)\,\mathsf{E}(N) + (\mathsf{E}(X_i))^2\mathsf{Var}(N)\end{aligned}$$

In particular, if $N \sim \mathcal{P}(\lambda)$, then this reduces to

$$\mathsf{Var}\left(S_N\right) = \lambda(\mathsf{Var}\left(X_i\right) + \left(\mathsf{E}\left(X_i\right)^2\right) = \lambda \mathsf{E}\left(X_i^2\right)$$

We could also look at the MGF of S_N, since expectations are relatively easy in these models. Assuming that the MGF exist (otherwise we could do the calculation with the characteristic function).

$$\begin{aligned} M_{S_N}(t) &= \mathsf{E}\left(e^{tS_N}\right) \\ &= \mathsf{E}\left(\mathsf{E}\left(e^{tS_N}|N\right)\right) \\ &= \mathsf{E}\left(\mathsf{E}\left(e^{tS_N}|N=n\right)\right) \\ &= \mathsf{E}\left((M_X(t))^n|N=n\right) \\ &= \mathsf{E}\left(M_X(t))^N\right) \\ &= \mathsf{E}\left(e^{N\log(M_X(t))}\right) \\ &= M_N(\log M_X(t)) \end{aligned}$$

conditioning and then unconditioning, valid only for t which are valid for the neighborhood. In the particular case where N is Poisson, this is

$$e^{\lambda(M_X(t)=1)}, \quad t \in \mathbb{R}$$

If we were asked to find an approximation to the density of S_N. One could use the saddle point approximation, given the MGF, which could be remarkably accurate.

Example 8.5 (Location and scale mixture - Financial returns)
Financial return is the price of today relative to yesterday, given by

$$\frac{S_t - S_{t-1}}{S_{t-1}} = R_t$$

If we look at the histogram of these return (which are not quite IID, there is some dependence), they almost look like Normally distributed. The density is however a little more heavy tailed than the normal, or more skewed. How can we tweak the model (using hierarchies) to account for these factors. We aim at making the variance random (larger kurtosis), and the mean random (will inject skewness in the model). In general, we can consider location-scale mixture. Say $f(m, \nu)$ is the density of the original variable.

At level 2: $(M, V) \sim g$

At level 1: $X|M = m, V = \nu$ has density

$$f\left(\frac{x-m}{\nu}\right)\frac{1}{\nu}.$$

for some suitable density f.

Two suitable examples are the following:

At level 2, $V \sim \mathcal{G}\left(\frac{\nu}{2}, 2\right), \nu > 0$.
At level 1, $X|V = v$ is Normal $\left(0, \frac{v}{\nu}\right)$ and interestingly enough $X \sim t(v)$. This will be explored on Assignment 4.

At level 2, $C \sim \mathcal{G}\left(\frac{\nu}{2}, 2\right), \nu > 0$.
At level 1, $X|V = v$ is Normal $\left(\frac{\gamma v}{\nu}, \frac{v}{\nu}\right)$ and X will be distributed as a skewed t_v distribution.

Chapter 9
Inequalities

Section 9.1. Concentration inequalities

We are interested in $\mathsf{P}(X \geq x)$ and $\mathsf{P}(X \leq x)$, which say are not tractable analytically. This could be used to fix the financial reserves, for example (Value at Risk, or VaR).

Theorem 9.1 (Markov inequality)
Let X be an arbitrary random variable and g a measurable function, with $\mathsf{P}(g(X) \geq 0) = 1$. Then $\forall\, \epsilon > 0$,

$$\mathsf{P}(g(X) \geq \epsilon) \leq \frac{\mathsf{E}\,(g(X))}{\epsilon}$$

and we will be interested in ϵ large ...

Proof Let f be the density of X, then assuming wlog that $g(X)$ is non-negative

$$\begin{aligned}\mathsf{E}\,(g(X)) &= \int_{-\infty}^{\infty} g(x)f(x)dx \geq \int_{\{x:g(x)\geq\epsilon\}} g(x)f(x)dx \\ &\geq \int_{\{x:g(x)\geq\epsilon\}} \epsilon f(x)dx \\ &= \epsilon \mathsf{P}(g(X) \geq \epsilon)\end{aligned}$$

■

Lemma 9.2 (Chebychev's inequality)
Let X be a random variable with expectation $\mathsf{E}\,(X)$ and variance $\mathsf{Var}\,(X)$ (assumed to exist). Then

$$\mathsf{P}(|X - \mathsf{E}\,(X)\,| \geq \epsilon\sqrt{\mathsf{Var}X}) \leq \frac{1}{\epsilon^2}$$

Proof We can write the left hand side as

$$\begin{aligned}&\mathsf{P}\left((X - \mathsf{E}\,(X))^2 \geq \epsilon^2 \mathsf{Var}\,(X)\right) \\ &\qquad\qquad \leq \frac{\mathsf{Var}\,(X)}{\epsilon^2 \mathsf{Var}\,(X)}\end{aligned}$$

■

	Chebychev	True	Approx.
$\epsilon = 1$	0	0.68	0.52
$\epsilon = 2$	0.75	0.95	0.946
$\epsilon = 3$	0.88	0.997	0.997

Remark

One could also have used the Markov inequality, namely

$$\mathsf{P}\left(|X - \mathsf{E}(X)| \geq \epsilon\sqrt{\mathsf{Var}(X)}\right) \leq \frac{\mathsf{E}(|)X - \mathsf{E}(X)|}{\epsilon\sqrt{\mathsf{Var}(X)}}$$

which is less tractable.

The Chebychev inequality (also sometimes spelled Tchebychev) is rather liberal and not so helpful.

For $X \sim \mathcal{N}(0,1)$, which has no explicit CDF. Using the symmetry of the normal distribution, if $x \geq 0$,

$$\begin{aligned}\mathsf{P}(|X| \geq x) &= \mathsf{P}(X \geq x) + \mathsf{P}(X \leq -x)\\ &= 2\mathsf{P}(X \geq x)\\ &= 2\int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt\end{aligned}$$

and we cannot evaluate this using a change of variable; we would need to have a t. Suppose that we put it there, and bound by adding $\frac{t}{x}$ which is greater or equal than 1. This would then become

$$\begin{aligned}&\leq 2\int_x^\infty \frac{t}{x}\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt\\ &= \frac{2}{\sqrt{2\pi}}\frac{1}{x}\int_{x^2/2}^\infty e^{-z} dz\\ &= \frac{\sqrt{2}}{\sqrt{\pi}}\frac{1}{x} e^{-\frac{x^2}{2}}\end{aligned}$$

and using this "Normal approximation", giving the approximation $\mathsf{P}(|X| \geq x) \leq \frac{1}{x^2}$

Indeed, if we look at the standard Normal $\mathcal{N}(0,1)$ random variable, The heavier tail, the x for a given level, say 95%, would be smaller for distributions with light tail, which decay quickly.

Theorem 9.3 (Chernoff bounds)
Let X be a random variable with MGF M, defined on $(-h, h)$. Then for $x > 0$,

$$\begin{aligned} \mathsf{P}(X \geq x) &\leq e^{-tx} M(t), \quad \forall\, t \in (0, h) \\ \mathsf{P}(X \leq x) &\leq e^{-tx} M(t). \quad \forall\, t \in (-h, 0) \end{aligned}$$

Proof The proof follows from Markov's inequality. Fix some $t \in (0, h)$ arbitrary. Then

$$\begin{aligned} \mathsf{P}(X \geq x) &= \mathsf{P}(tX \geq tx) \\ &= \mathsf{P}\left(e^{tX} \geq e^{tx}\right) \\ &\leq \frac{\mathsf{E}\left(e^{tX}\right)}{e^{tx}} \\ &= M(t) e^{-xt} \end{aligned}$$

and we can reverse the inequality for t negative. ■

Example 9.1
Let $X \sim \mathcal{P}(\lambda)$, then $M(t) = e^{\lambda(e^t - 1)}, t \in \mathbb{R}$ and we could use Chernoff bounds to get

$$\mathsf{P}(X \geq x) \leq e^{\lambda(e^t - 1)} e^{-xt}, t > 0.$$

Denote $g(t) = e^{\lambda(e^t - 1)} e^{-xt}$, and differentiating with respect to t, we have $g'(t) = 0$ for $t = \log(x/\lambda)$ and for $x > \lambda$. Computing the second derivative, we find out that $g''(t) \geq 0$, so that for $x > \lambda$,

$$\begin{aligned} \mathsf{P}(X \geq x) &\leq e^{\lambda} e^{\lambda\left(\log\left(\frac{x}{\lambda}\right) - 1\right) - x\left(\log\left(\frac{x}{\lambda}\right)\right)} \\ &= e^{(x-\lambda) - x\log\left(\frac{x}{\lambda}\right)} \\ &= e^{-\lambda} \left(\frac{ex}{\lambda}\right)^x \end{aligned}$$

Section 9.2. Triangle inequalities

Lemma 9.4
Let $p, q > 1$ conjugate exponents (such that $\frac{1}{p} + \frac{1}{q} = 1$), then for all $a, b > 0$

$$\frac{a^p}{p} + \frac{b^q}{q} \geq ab$$

Proof Fix b and consider $g(a) = \frac{a^p}{p} + \frac{b}{q} - ab$ and $g''(a) > 0$, then one finds that

$$g'(a) = 0 \Leftrightarrow a^{p-1} = b \Leftrightarrow a_0 = b^{\frac{1}{p-1}}$$

where a_0 is the minimum. Then $g(a_0) = 0$ and hence $g(a) \geq 0 \; \forall \; a > 0$. Filling in the gaps is left as an exercise. ■

Theorem 9.5 (Hölder inequality)
Let X, Y be random variable, with p, q conjugate exponents. Then if $\mathsf{E}|X|^p < \infty$ and $\mathsf{E}|Y|^q < \infty$, then

$$|\mathsf{E}(XY)| \leq \mathsf{E}(|XY|) \leq (\mathsf{E}|X|^p)^{\frac{1}{p}} \cdot (\mathsf{E}|Y|^q)^{\frac{1}{q}}$$

and as a special case, if $p = q = 2$, then we have the

$$|\mathsf{E}(XY)| \leq \sqrt{\mathsf{E}(X^2)\,\mathsf{E}(Y^2)}$$

and in particular, we can use this

$$\begin{aligned} |\mathsf{E}((X - \mathsf{E}(X))(Y - \mathsf{E}(Y)))| &= |\mathsf{Cov}X, Y| \\ &\leq \sqrt{\mathsf{Var}(X)\,\mathsf{Var}(Y)} \\ &= |\rho(X < Y)| \leq 1 \end{aligned}$$

Proof The first part follows from the triangle inequality.

$$\left| \int xyf(x,y)dxdy \right| \leq \int |xy| f(x,y)dxdy$$

Using the previous inequality, $a^p/p + b^q/q \geq ab$, and the fact that random variables are mapping, $\forall\, \omega \in \Omega$, take

$$a = \frac{X(\omega)}{(\mathsf{E}|X|^p)^{\frac{1}{p}}} \qquad b = \frac{Y(\omega)}{(\mathsf{E}|Y|^q)^{\frac{1}{q}}}$$

thus

$$\frac{|X(\omega)|^p}{p\,(\mathsf{E}|X|^p)^{\frac{1}{p}}} + \frac{|Y(\omega)|^q}{q\,(\mathsf{E}|Y|^q)^{\frac{1}{q}}} \geq \frac{|X(\omega)Y(\omega)|}{(\mathsf{E}|X|^p)^{\frac{1}{p}}\,(\mathsf{E}|Y|^q)^{\frac{1}{q}}}.$$

Now for $Z \leq W$, we know that $\mathsf{E}(Z) \leq \mathsf{E}(W)$, the RHS becomes

$$\frac{\mathsf{E}|XY|}{(\mathsf{E}|X|^p)^{\frac{1}{p}}(\mathsf{E}|Y|^q)^{\frac{1}{q}}} \leq \frac{\mathsf{E}|X|^p}{p(\mathsf{E}|X|^p)^{\frac{1}{p}}} + \frac{\mathsf{E}|Y|^q}{q(\mathsf{E}|Y|^q)^{\frac{1}{q}}} = \frac{1}{p} + \frac{1}{q} = 1$$

■

An interesting case is the case where $p = q = 2$, then

$$\mathsf{E}|XY| \leq \sqrt{\mathsf{E}(X^2)\,\mathsf{E}(Y^2)} \quad \Leftrightarrow \quad |\mathsf{Cov}(X,Y)| \leq \sqrt{\mathsf{Var}(X)\,\mathsf{Var}(Y)}$$

and we can also take

$$\mathsf{E}|X| \leq (\mathsf{E}|X|^p)^{\frac{1}{p}}$$

$\forall\, p > 1$, then taking $Y \equiv 1$, we get

$$\mathsf{E}|X|^r \leq (\mathsf{E}|X|^{rp})^{\frac{1}{p}} \equiv (\mathsf{E}|X|^s)^{\frac{r}{s}}$$

where $s = rp > r$, which can be used to show existence of moments. This inequality we derived from Hölder's inequality, the so-called **Lyapunov inequality** , which can be termed as

$$(\mathsf{E}|X|^r)^{\frac{1}{r}} \leq (\mathsf{E}|X|^s)^{\frac{1}{s}}$$

whenever $1 \leq r < s$.

If we are interested in $g(X)$ and $g(EX)$ for some random variable X with finite first moment.

Theorem 9.6 (Jensen inequality)
Let X be a random variable and $\mathsf{E}(X)$ exist. Let g be a function such that $\mathsf{E}(g(X))$

1. If f is convex,

$$\mathsf{E}(g(X)) \geq g(\mathsf{E}(X))$$

2. If g is concave,

$$\mathsf{E}g(X) \leq g(\mathsf{E}(X))$$

Equality will hold for linear functions.

Proof For the second part, if g is concave, then $-g$ is convex, so we will prove the first

part. Suppose g is convex, then

$$g'_+(x) = \lim_{\substack{h \downarrow 0 \\ h>0}} \frac{g(x+h) - g(x)}{h}$$

exists for all $x \in \mathbb{R}$. If we consider the graph of a continuous convex function and a tangent, then the graph will lie above the tangent. Tangent in x is given by the expression: $g(x) + g'_+(x)(z-x)$ for $z \in \mathbb{R}$. Thus,

$$g(x) + g'_+(x)(z-x) \leq g(z), \qquad \forall\, z \in \mathbb{R},\ \forall\, x \in \mathbb{R}$$

For all $\omega \in \Omega$, $g(x) + g'_+(x)(X(\omega) - x) \leq g(X(\omega))$, therefore cleverly chose $x = \mathsf{E}(X)$, then

$$g(\mathsf{E}(X)) + g'_+(\mathsf{E}(X))(X(\omega) - \mathsf{E}(X)) \leq g(X(\omega))$$

and taking expectation on both sides, on the LHS we get

$$g(\mathsf{E}(X)) + g'_+(\mathsf{E}(X))\mathsf{E}(X - \mathsf{E}(X)) \leq \mathsf{E}(g(X))$$

and since $\mathsf{E}(X - \mathsf{E}(X)) = 0$, we recover the Jensen's inequality. To remember it, notice that $\mathsf{E}\left(X^2\right) \geq (\mathsf{E}(X))^2$, since $\mathsf{Var}(X) \geq 0$ and it can be composed into $\mathsf{E}\left(X^2\right) - (\mathsf{E}(X))^2 \geq 0$. ■

This inequality is particularly useful to asses whether they are biased estimators.

Example 9.2

Suppose we have a random sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{E}(\lambda)$, where the PDF is given by $f(x) = \lambda e^{-\lambda x}, x > 0$. Knowing $\mathsf{E}(X) = \lambda^{-1}$ and we are interested in estimating λ. Then, a natural estimate could be the sample mean,

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n) \quad \mathsf{E}\left(\overline{X}_n\right) = \frac{1}{\lambda}$$

but we are interested in λ. Then, taking $\lambda_n = 1/\overline{X}_n$, we have $\mathsf{E}(\lambda_n) = \mathsf{E}\left(1/\overline{X}_n\right)$ will be biased, and always in the same direction. Since the function $g(x) = \frac{1}{x}$ is convex, then

$$\mathsf{E}(\lambda_n) \geq g(\mathsf{E}\left(\overline{X}_n\right)) = g\left(\frac{1}{\lambda}\right) = \lambda$$

so λ_n is overestimating the true value and the estimator is biased.

Exercise 9.1

If $a_1, \ldots, a_n$ be positive real numbers. We could look at the arithmetic mean, which is given

by

$$a_{\mathrm{A}} = \frac{a_1 + \cdots + a_n}{n}$$

or the geometric mean, given by

$$a_{\mathrm{G}} = \sqrt[n]{a_1 \cdots a_n}$$

and the harmonic mean

$$a_{\mathrm{H}} = \frac{1}{\frac{1}{n}\left(\frac{1}{a_1} + \cdots + \frac{1}{a_n}\right)}$$

and one can show that $a_{\mathrm{H}} \leq a_{\mathrm{G}} \leq a_{\mathrm{A}}$. Trick: if X is a discrete random variable with support $\{a_1, \ldots, a_n\}$, then $\mathsf{P}(X = a_i) = \frac{1}{n}$ and $\mathsf{E}(X) = a_{\mathrm{A}}$. Take log and exp to get the result

Chapter 10
Properties of random samples

Definition 10.1
A random sample from either CDF F, PDF/PMF f or random variable X with CDF F is a collection $X_1, \ldots, X_n$ of IID random variables with CDF F or PMF/PDF or distribution function. For now, we are interested in the distribution of statistics.

In practice, we can have different problems linked to dependence, failure of independence assumption or other issues. This will be discussed in the second part of the course.

Definition 10.2 (Statistic)
$T(X_1, \ldots, X_n)$ is a measurable function of $X_1, \ldots, X_n$, where n is called the sample size.

Example 10.1

1. The sample mean
$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$
2. The sample variance
$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$
3. The maximum $X_{(n)} = \max(X_1, \ldots, X_n)$
4. The range $X_{(n)} - X_{(1)} = \max(X_1, \ldots, X_n) - \min(X_1, \ldots, X_n)$.
5. The r^{th} order statistic $X_{(r)}$, the r^{th} smallest observation among $X_1, \ldots, X_n$.

Observation

- $T(X_1, \ldots, X_n)$ must not depend on any unknown parameters. Taking a sample from $\mathcal{N}(\mu, \sigma^2)$, where μ, σ^2 are unknown, the function $n^{-1}\sum_{i=1}^{n}(X_i - \mu)^2$ is not a statistic.
- The distribution of $T(X_1, \ldots, X_n)$ is called a **sample distribution** and it may well depend on unknown parameters.
- The sample distribution often simplifies greatly if $n \to \infty$ (the asymptotic distribution, on which frequentist approach to statistics is based).

Section 10.1. Sample mean

Lemma 10.3
Suppose that $\mathsf{E}(X)$ and $\mathsf{Var}(X)$ exist. Then

1. $\mathsf{E}\left(\overline{X}_n\right) = \mathsf{E}(X)$
2. $\mathsf{Var}\left(\overline{X}_n\right) = \frac{\mathsf{Var}(X)}{n}$

so the estimate of the expectation of X gets more precise as the sample size goes to infinity.

Proof By linearity,

$$\mathsf{E}\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n} n \mathsf{E}(X)$$

and

$$\mathsf{Var}\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n^2} n \mathsf{Var}(X)$$

■

Lemma 10.4
If X has MGF, then the MGF of $\overline{X}_n$ is given by

$$M_{\overline{X}_n}(t) = \left(M_X\left(\frac{t}{n}\right)\right)^n, \quad t \in (-\varepsilon \cdot n, \varepsilon \cdot n)$$

Proof

$$\begin{aligned}\mathsf{E}\left(e^{t(n^{-1}(X_1+\cdots+X_n))}\right) &= \prod_{i=1}^{n} \mathsf{E}\left(e^{tn^{-1}X_i}\right) \\ &= \left(M_X\left(\frac{t}{n}\right)\right)^n\end{aligned}$$

and t/n goes smaller as $n \to \infty$ ■

Example 10.2
Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then the MGF of X is

$$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}, \quad t \in \mathbb{R}$$

and therefore

$$\begin{aligned} M_{\overline{X}_n}(t) &= \left(e^{\mu\frac{t}{n}+\frac{\sigma^2 t^2}{2n^2}}\right)^n \\ &= e^{\mu t+\frac{\sigma^2 t^2}{2n}} \end{aligned}$$

so we conclude that $\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

Example 10.3

Consider a random sample from $\mathcal{G}(\alpha, \beta)$ then

$$M_X(t) = (1-\beta t)^{-\alpha}, \quad t < \frac{1}{\beta}$$

$$M_{\overline{X}_n}(t) = \left(1-\frac{\beta t}{n}\right)^{-\alpha n}, \quad t < \frac{n}{\beta}$$

then $\overline{X}_n \sim \mathcal{G}(\alpha n, \beta/n)$. In the case where $\alpha = 1$, this is an exponential distribution.

Lemma 10.5

The characteristic function (CF) of the sample mean is given by

$$C_{\overline{X}_n}(t) = \left(C_X\left(\frac{t}{n}\right)\right)^n, \quad t \in \mathbb{R}$$

and we can work with this if the MGF doesn't exist.

Example 10.4

Consider $X_1, \ldots, X_n$, which is a random sample from Cauchy with parameters $(0, 1)$. We could look at

$$C_X(t) = e^{-|t|} C\overline{X}_n(t) = \left(e^{-\frac{|t|}{n}}\right)^n = e^{-|t|}$$

and this doesn't give us any information, since the second moment not existing.

Example 10.5

Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, 1)$, then by previous results, $\overline{X}_n \overset{\mathrm{d}}{=} \mathcal{N}\left(\mu, \frac{1}{n}\right)$ and Normal random variables belong to Exponential family, where the density is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} e\mu x e^{-\frac{\mu^2}{2}} \mathbf{1}_{(x\in\mathbb{R})}$$

and can be termed as $(2\pi)^{-\frac{1}{2}} e^{-\frac{x^2}{2}}$ and $c(\mu) = e^{-\frac{\mu^2}{2}}$, $t_1(x) = x, \omega_1(\mu) = \mu$.

Set $T_1 = X_1 + \cdots + X_n \sim \mathcal{N}(n\mu, n)$ and

$$f_{T_1}(t) = \frac{1}{\sqrt{2\pi n}} e^{-\frac{t^2}{2n}} e^{-\frac{\mu^2}{2}n} e^{\mu t}$$

with now $h^*(t) = (2\pi n)^{-\frac{1}{2}} e^{-\frac{t^2}{2n}}$, $c(\mu)^n = e^{-\frac{\mu^2}{2}n}$ and $t_1(t) = t, \omega_1(\mu) = \mu$

We could look at another example with the Exponential family, that is $X \sim \mathcal{B}(p)$, a Bernoulli random variable.

$$f_X(x) = (1-p)e^{x \log\left(\frac{p}{1-p}\right)} \mathbf{1}_{(x \in \{0,1\})}$$

and using the sufficient statistic for p, where $T_1 = X_1 + \cdots + X_n \sim \mathcal{B}(n, p)$ and the PMF is given by

$$f_T(t) = \binom{n}{t} \mathbf{1}_{(t \in \{0,\ldots,n\})} (1-p)^n e^{t \log\left(\frac{p}{1-p}\right)}$$

and we remark again that $C^*(p) = (C(p))^n$, which is a result of the following result, namely

Theorem 10.6
Let $X_1, \ldots, X_n$ be a random sample from f which is from an Exponential family.

$$f(x) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} t_j(x)\omega_j(\boldsymbol{\theta})\right)$$

Then, the PDF(PMF) of $(T_1, \ldots T_k)$, that is $T_j = \sum_{i=1}^{n} t_j(X_i)$ is

$$f_{T_1,\ldots T_k}(t_1, \ldots, t_k) = h^*(t_1, \ldots, t_d) \left(c(\boldsymbol{\theta})\right)^n \exp\left(\sum_{j=1}^{k} t_j(x)\omega_j(\boldsymbol{\theta})\right)$$

Section 10.2. Sample variance

Let $X_1, \ldots, X_n$ be a random sample from F, and we write $X \sim F$ for any generic X. The **sample variance** is given by

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

Lemma 10.7
Let $x_1, \ldots, x_n$ be any real numbers (fixed). Then

1. $\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \sum_{i=1}^{n} x_i^2 - n(\overline{x}_n)^2$;
2. $\sum_{i=1}^{n}(x_i - a)^2$ is minimized if $a = \overline{x}_n$.

Proof Recall that Y is a random variable $\mathsf{Var}(Y) = \mathsf{E}\left(Y^2\right) - (\mathsf{E}(Y))^2$ and $\mathsf{E}(Y-a)^2$ is minimized of $a = \mathsf{E}(Y)$. Now choose Y to be a discrete random variable with support $\{X_1, \ldots, X_n\}$ and $\mathsf{P}(Y = x_i) = \frac{1}{n}$. Then

$$\mathsf{E}(Y) = \overline{x}_n$$

the arithmetic mean and

$$\begin{aligned}\mathsf{Var}(Y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x}_n)^2\end{aligned}$$

and also

$$\mathsf{E}(Y-a)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2$$

is minimized if $a = \overline{x}_n$. ■

Lemma 10.8
If $\mathsf{Var}(X)$ exists, then $\mathsf{E}\left(S_n^2\right) = \mathsf{Var}(X)$.

Proof

$$\begin{aligned}\mathsf{E}\left(S_n^2\right) &= \frac{1}{n-1}\mathsf{E}\left(\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right) \\ &= \frac{1}{n-1}\mathsf{E}\left(\sum_{i=1}^{n} x_i^2 - n(\overline{X}_n)^2\right) \\ &= \frac{n}{n-1}\mathsf{E}\left(X^2\right) = \frac{n}{n-1}\mathsf{E}\left(\overline{X}_n^2\right)\end{aligned}$$

Now $\mathsf{E}\left(X^2\right) = \mathsf{Var}(X) + (\mathsf{E}(X))^2$ and

$$\mathsf{E}\left(\overline{X}_n^2\right) = \mathsf{Var}\left(\overline{X}_n\right) + \left(\mathsf{E}\left(\overline{X}_n\right)\right)^2 = \frac{\mathsf{Var}(X)}{n} + (\mathsf{E}(X))^2$$

and

$$\begin{aligned}\mathsf{E}\left(S_n^2\right) &= \frac{n}{n-1}\left(\mathsf{Var}\left(X\right) + \left(\mathsf{E}\left(X\right)\right)^2 - \frac{\mathsf{Var}\left(X\right)}{n} - \left(\mathsf{E}\left(X\right)\right)^2\right) \\ &= \frac{n}{n-1}\mathsf{Var}\left(X\right)\left(\frac{n-1}{n}\right) \\ &= \mathsf{Var}\left(X\right)\end{aligned}$$

and in the case where the variance does not exist, the sample variance can be computed, but does not estimate anything meaningful. ■

In the special case where $X \sim \mathcal{N}(\mu, \sigma^2)$, then $X_1, \ldots, X_n$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$.

Theorem 10.9

In this special case

1. $\overline{X}_n$ and S_n^2 are independent
2. $(n-1)S_n^2/\sigma^2 \sim \chi^2(n-1)$.

Proof Observe that

$$\sum_{i=1}^{n}(X_i - \overline{X}_n) = 0 \Leftrightarrow (X_1 - \overline{X}_n) = -\sum_{i=2}^{n}(X_i - \overline{X}_n)$$

Therefore,

$$S_n^2 = \frac{1}{n-1}\left(\sum_{i=2}^{n}(X_i - \overline{X}_n)^2 + \left(\sum_{i=2}^{n}(X_i - \overline{X}_n)\right)^2\right)$$

S_n^2 is thus a function of $X_2 - \overline{X}_n, \ldots, X_n - \overline{X}_n$. Now $\overline{X}_n$ happens to be independent from $(X_2 - \overline{X}_n, \ldots, X_n - \overline{X}_n)$ and the first statement follow. We prove it by brute force, considering the transformation

$$\begin{aligned} g &: (X_1, \ldots, X_n) \mapsto (\overline{x}_n, x_2 - \overline{x}_n, \ldots, x_n - \overline{x}_n) \\ g^{-1} &: (y_1, \ldots, y_n) \mapsto (y_1 - y_2 - y_3 - \cdots - y_n, y_2 + y_1, \ldots y_n + y_1)\end{aligned}$$

and the Jacobian is

$$J = J = \begin{pmatrix} 1 & -1 & \cdots & -1 \\ 1 & 1 & 0 & \cdots \\ \vdots & \ddots & \ddots & 0 \\ 1 & \cdots & 0 & 1 \end{pmatrix}, \quad |\det(J)| = n$$

In the simple case where we assume that $\mu = 0, \sigma^2 = 1$,

$$\begin{aligned} f_{(\overline{X}_n, X_2 - \overline{X}_n, \ldots, X_n - \overline{X}_n)}(y_1, \ldots, y_n &= n \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2} \left(\left(y_1 - \sum_{i=2}^{n} y_i \right)^2 + (y_2 + y_1)^2 + \cdots + (y_n + y_1)^2 \right) \right) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2} n y_1^2 + \left(\sum_{i=2}^{n} y_i \right)^2 + \sum_{i=2}^{n} y_i^2 \right) \\ &= \frac{\sqrt{n}}{\sqrt{2\pi}} e^{-\frac{y_1^2}{2/n}} \cdot \frac{\sqrt{n}}{(2\pi)^{\frac{n}{2} - \frac{1}{2}}} \exp\left(-\frac{1}{2} \left(\sum_{i=2}^{n} y_i \right)^2 + \sum_{i=2}^{n} y_i^2 \right) \end{aligned}$$

and we can also do the same thing for the general case.

For (2), recall that

- If $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi^2(1)$
- $W_1, \ldots, W_n$ independent and $W \sim \chi^2(1)$, then

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

- The MGF of $\chi^2(n)$ is $(1-2t)^{-\frac{\nu}{2}}, t < \frac{1}{2}$.

First,

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$$

and thus

$$\begin{aligned} \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i \pm \overline{X}_n - \mu)^2 &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 + \frac{n}{\sigma^2} (\overline{X}_n - \mu)^2 \\ &= \frac{(n-1)}{\sigma^2 S_n^2} + \left(\frac{X_i - \mu}{\sigma} / \sqrt{n} \right)^2 \end{aligned}$$

and the MGF of the :LHS is

$$(1-2t)^{-\frac{n}{2}} = M_{\frac{(n-1)}{\sigma^2} S_n^2}(t)(1-2t)^{-\frac{1}{2}}$$

and the MGF of $\frac{(n-1)}{\sigma^2} S_n^2$ is $(1-2t)^{-\frac{n}{2}+\frac{1}{2}}$ which is the MGF of a $\chi^2(n-1)$ random variable.

■

Theorem 10.10

If $Z \sim \mathcal{N}(0,1)$, and $W \sim \chi^2(\nu)$, $Z \perp\!\!\!\perp W$, then

$$\frac{Z}{\sqrt{W/\nu}} \sim t(\nu)$$

Corollary 10.11

If $X_1, \ldots, X_n$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$, then

$$\frac{\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}}{\sqrt{\frac{(n-1)S_n^2}{\sigma^2(n-1)}}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{S_n^2}} \sim t(n-1)$$

Theorem 10.12

If $W_1 \sim \chi^2(\nu_1)$ and $W_2 \sim \chi^2(\nu_2)$ and $W_1 \perp\!\!\!\perp W_2$, then

$$\frac{W_1/\nu_1}{W_2/\nu_2} \sim F(\nu_1, \nu_2)$$

which is the Fisher-Snedecor distribution.

Corollary 10.13

If $X_1, \ldots, X_n$ are random variable from $\mathcal{N}(\mu, \sigma^2)$ and $Y_1, , \ldots, , Y_m$ are random variables from $\mathcal{N}(\mu^*, \sigma^2)$ (that is, they share the same variance) are independent. Then

$$\frac{\frac{(n-1)S_n^2}{\sigma^2(n-1)}}{\frac{(m-1)T_m^2}{\sigma^2(m-1)}}$$

where T_m^2 is the sample variance of $Y_1, \ldots, Y_m$. Simplifying, we get

$$\frac{S_n^2}{T_m^2} \sim F(n-1, m-1)$$

Exercise 10.1

- If $X \sim F(\nu_1, \nu_2)$, then $\frac{1}{X} \sim F(\nu_2, \nu_1)$.
- If $T \sim t(\nu)$, then $T^2 \sim F(1, \nu)$
- If $X \sim F(\nu_1, \nu_2)$, then

$$\frac{\frac{\nu_1}{\nu_2}X}{1 + \frac{\nu_1}{\nu_2}X} \sim \mathcal{B}e\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)$$

The following is not to be disregarded for the final exam preparation. There are other measures of central tendency

Section 10.3. Order statistics

Definition 10.14
Let $X_1, \ldots, X_n$ be a random sample from F, then

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

the X_i's are ordered in ascending order) is called an order statistic.

Example 10.6
if we have $X_1 = 1.12, X_2 = 2.04, X_3 = -0.5$, then $X_{(1)} = -0.5, X_{(2)} = 1.12,\ X_{(3)} = 2.04$.

In the above, the random variable $X_{(1)}$ is the sample minimum, $X_{(n)}$ is the sample maximum, $X_{(n)} - X_{(1)}$ is the range, and the sample median is

$$X_{(\text{med})} = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd;} \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+1}{2}\right)}\right) & \text{if } n \text{ is even} \end{cases}$$

by convention, although any value between those could be chosen

Example 10.7
Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{E}(1)$, then

$$\begin{aligned} \mathsf{P}(X_{(1)} \leq x) &= 1 - \mathsf{P}(X_{(1)} > x) \\ &= 1 - \prod_{i=1}^{n} P(X_i > x) \\ &= 1 - e^{-nx} \end{aligned}$$

and we conclude that $X_{(1)} \sim \mathcal{E}(n)$.

For the sample maximum,

$$\mathsf{P}(X_{(n)} \leq x) = (\mathsf{P}(X_i \leq x))^n = (1 - e^{-x})^n$$

which is no longer exponentially distributed.

Theorem 10.15

Let $X_1, \ldots, X_n$ be a random sample from a distribution F. Then, for the j^{th} order statistic,

$$\mathsf{P}(X_{(j)} \leq x) = \sum_{k=j}^{n} \binom{n}{k} \{F(x)\}^k \{1 - F(x)\}^{n-k}$$

Proof

$$\mathsf{P}(X_{(j)} \leq x) = \mathsf{P}(\text{at least } j \text{ observations are less than or equal to } x)$$

and so $Y_X \approxeq \#X_i's$ that are $\leq x$, so $Y_X \sim \mathcal{B}(n, F(x))$ which is $P(Y_x \geq j)$.

If F is continuous with density f, then

$$\begin{aligned} f_{X_{(j)}}(x) &= \sum_{k=j}^{n} \binom{n}{k} k \{F(x)\}^{k-1} f(x) \{1 - F(x)\}^{n-k} \\ &\quad - \sum_{k=j}^{n-1} \binom{n}{k} \{F(x)\}^k f(x) \{1 - F(x)\}^{n-k-1} (n-k) \\ &= \binom{n}{j} j \{F(x)\}^{j-1} f(x) \{1 - F(x)\}^{n-j} \end{aligned}$$

which we can simplify big time in the last line, since the last term, upon making a simple change of variable $k^* = k - 1$

$$\begin{aligned} &\sum_{k^*=j-1}^{n-1} \frac{n!}{(n-k^*-1)!(k^*+1)!} (k^*+1)^{k^*} \{F(x)\}^{k^*} f(x) \{1 - F(x)\}^{n-k^*-1} \\ &\quad - \sum_{k=j}^{n-1} \frac{n!(n-k)}{k!(n-k)!} \{F(x)\}^k f(x) \{1 - F(x)\}^{n-k-1} \end{aligned}$$

which are equal. Here is a trick to remember it.

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

■

This trick applies in general; as for the joint distribution, if $j < k$

$$f_{(X_{(j)}, X_{(k)})}(x, y) = \mathbf{1}_{(x<y)} \frac{n!}{(j-1)!(k-1-j)!(n-k)!} F(x)^{j-1} f(x) (F(y) - F(x))^{k-j} f(y) (1 - F(y))^{n-k}$$

Figure 10: Distribution of j^{th} order statistics at x: multinomial interpretation

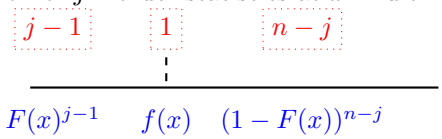
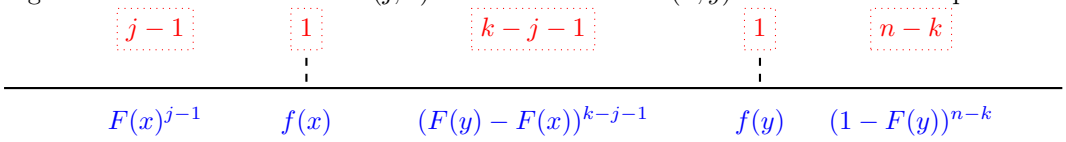


Figure 11: Joint distribution of $(j,k)^{\text{th}}$ order statistics at (x,y): multinomial interpretation



and for all order statistic,

$$f_{(X_{(1)},\dots,X_{(n)})}(X_1,\dots,X_n)=\mathbf{1}_{(x_1<\cdots<x_n)}n!f(x_1)\cdots f(x_n)$$

Exercise 10.2
If $X_1,\dots,X_n\sim\mathcal{U}(0,1)$, then

$$f_{X_{(j)}}(x)=\frac{n!}{(j-1)!(n-j)!}x^{j-1}(1-x)^{n-j}\sim\mathcal{B}(j,n-j+1)$$

If F is discrete, there is no simple formula. Indeed, we have to work with

$$\begin{aligned}\mathsf{P}(X_{(j)}=x)&=\mathsf{P}(X_{(j)}\leq x)-\mathsf{P}(X_{(j)}<x)\\&=\sum_{k=j}^{n}\binom{n}{k}\left(F(x)^k(1-F(x))^{n-k}-F(x-)^k(1-F(x-))^{n-k}\right)\end{aligned}$$

Chapter 11
Convergence concepts

Recall we had some intuition of convergence and consistency of estimators. For a frequentist, one would want to know about the distribution of the estimator or the statistic, expecting that as the sample size increase to the population size, as $n \to \infty$, we get asymptotic distribution which gives approximate P values and confidence interval.

Recall that $X_1, \ldots, X_n$ is a random sequence of mapping, from $\Omega \Rightarrow \mathbb{R}^d$. Characterizing the convergence of the sequence, as in analysis with pointwise convergence or uniform convergence. However, we have random samples, and we cannot talk about the above. Instead, we define probabilistic concepts.

Definition 11.1 (Convergence in probability)
A sequence of random variables, $X_1, X_2, \ldots$ is said to **converge in probability** to a random variable X, denoted $X_n \xrightarrow{\mathsf{p}} X$ if

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

Probability is more general because every number can be regarded as the realization of a random variable.

Remark

- $X_1, X_2, \ldots$ are arbitrary; in particular, they are most of the time not IID, nor are they independent.
- Often X is a constant, say c and we have in that case

$$X_n \xrightarrow{\mathsf{p}} c \Leftrightarrow \mathsf{P}(|X_n - c| > \varepsilon) \text{ as } n \to \infty$$

Example 11.1 (Convergence in probability of the sample variance)
$\mathsf{E}\left(S_n^2\right) = \sigma^2$ if $\mathsf{Var}\left(X\right) = \sigma^2$. If $X_1, \ldots, X_n$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$, we have

$$\begin{aligned}
\mathsf{E}\left(S_n^2 - \sigma^2\right) &= \mathsf{Var}\left(S_n^2\right) \\
&= \mathsf{Var}\left(\frac{(n-1)S_n^2}{\sigma^2}\frac{\sigma^2}{n-1}\right) \\
&= \frac{\sigma^4}{(n-1)^2}\mathsf{Var}\left(\frac{(n-1)S_n^2}{\sigma^2}\right) \\
&= \frac{\sigma^4}{(n-1)^2}2(n-1) \\
&= \frac{2\sigma^4}{(n-1)}
\end{aligned}$$

as $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$ and this gives us a reason to believe our estimate is getting more and more precise. What does this have to do with convergence in probability? If we can get an upper bound that goes to zero, this will imply convergence. Fix $\varepsilon > 0$ arbitrary, then

$$\begin{aligned} 0 \leq \mathsf{P}(|S_n^2 - \sigma^2| > \varepsilon) &= \mathsf{P}((S_n^2 - \sigma^2)^2 > \varepsilon^2) \\ &\leq \frac{\mathsf{Var}\left(S_n^2\right)}{\varepsilon^2} \\ &= \frac{2\sigma^4}{(n-1)\varepsilon^2} \to 0 \text{ as} n \to \infty \end{aligned}$$

using Chebychev's inequality.

Theorem 11.2 (Weak law of large number)
Let X_1, X_2 be an IID sequence such that $\mathsf{E}(X_1) = \nu$ and $\mathsf{Var}(X_1) < \infty$. Then

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{\mathsf{P}} \mu.$$

This result is the so-called Weak law of large numbers (WLLN)

Proof Using Chebychev's inequality,

$$\mathsf{P}(|\overline{X}_n - \mu| > \varepsilon) \leq \frac{\mathsf{Var}\left(\overline{X}_n\right)}{\varepsilon^2} = \frac{\mathsf{Var}(X_1)}{n\varepsilon^2} \xrightarrow{n\to\infty} 0$$

■

Remark
The WLLN holds if $\mathsf{E}(X_1)$ exists (meaning that $\mathsf{E}|X_1| < \infty$)

Remark
A stronger concept of stochastic convergence is convergence **almost surely** , which means that $X_1, X_2, \ldots$ is a sequence of random variables and X another random variable and

$$X_n \xrightarrow{\text{a.s.}} X(\text{a.s.}) \Leftrightarrow \mathsf{P}\left(\lim_{n\to\infty} |X_n - X| = 0\right) = 1$$

Alternatively, we can say that $X_n \to X$ almost surely if and only if

$$\forall\, \varepsilon > 0, \mathsf{P}\left(\lim_{n\to\infty} |X_n - X| > \varepsilon\right) = 0$$

The take home messages are

- $X_n X \Rightarrow X_n \xrightarrow{\mathsf{P}} X$, but **not conversely**.

- If $X_1, X_2, \ldots$ is an IID sequence with $\mathsf{E}|X_1| < \infty$, then $\overline{X}_n \to \mathsf{E}(X_1)$ almost surely (this is the Strong law of large number, SLLN)
- Almost sure convergence is often (much) harder to establish than convergence in probability and it is often too strong and not very useful (in statistics)
- There is an interesting partial converse, if $X_n \xrightarrow{\mathsf{P}} X$ if and only if for any subsequence $\{X_{n_k}\}$, there exists a subsubsequence $X_{n_{k_j}}$ such that $X_{n_{k_j}} \to X$ almost surely.[10]

Theorem 11.3 (Continuous mapping)
If $X_n \xrightarrow{\mathsf{P}} X$ and g is continuous, then $g(X_n) \xrightarrow{\mathsf{P}} g(X)$.

Proof If $Z_n \to Z$ almost surely, then $g(Z_n) \to g(Z)$ almost surely. Looking at the sequence $g(X_n)$, then the subsequence X_{n_k} has a subsequence $X_{n_{k_j}}$ converges to X almost surely, and so the subsequence $g(X_{n_k})$, then for the subsubsequence $g(X_{n_{k_j}}) \xrightarrow{j\to\infty} g(X)$ almost surely. ■

This could be used for the sample variance.

Example 11.2

$$\begin{aligned} S_n^2 &= \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n(\overline{X}_n)^2\right) \\ &= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right) - \frac{n}{n-1}\left(\overline{X}_n\right)^2 \end{aligned}$$

and by the continuous mapping theorem, we have

$$\left(\overline{X}_n\right)^2 \xrightarrow{\mathsf{P}} (\mathsf{E}(X))^2$$

and if $\mathsf{E}(X_i)^4 < \infty$, then by the WLLN,

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{\mathsf{P}} \mathsf{E}\left(X^2\right) = \mathsf{Var}(X) + (\mathsf{E}(X_i))^2.$$

Definition 11.4 (Convergence in r^{th} mean)
If $X_1, X_2, \ldots$ is a sequence of random variables, then $X_n \to X$ in r^{th} mean, if

$$\mathsf{E}|(X_n - X)|^r \to 0 \text{ as } n \to \infty$$

[10] This also holds for finite measure.

We illustrate this concept with an artificial example

Example 11.3
If $X_n \in \{0, n\}$, then

$$\mathsf{P}(X_n = 0) = 1 - \frac{1}{2^n}, \quad \mathsf{P}(X_n = n) = \frac{1}{2^n}$$

and

$$\mathsf{E}\,(X_n - 0)^2 = \frac{n^2}{2^n} \to 0 \text{ as } n \to \infty$$

and even $X_n \to 0$ in r$^{\text{th}}$ mean for $r \geq 1$.

Now, choosing instead $\mathsf{P}(X_n = 0) = 1 - \frac{1}{n^2}$ and $\mathsf{P}(X_n = n) = \frac{1}{n^2}$, then this time $\mathsf{E}\,(X_n - 0)^2 = \frac{n^2}{n^2} \nrightarrow 0$, but $\mathsf{E}|X_n - 0| = \frac{n}{n^2} \to 0$ and so we conclude that $X_n \to 0$ in 1^{st} mean, but not 2^{nd} mean. This motivates the following

Theorem 11.5
Let $1 \leq r \leq s$. Then $X_n \to X$ in s^{th} mean implies that $X_n \to X$ in r^{th} mean.

Proof Using Hölder inequality, or more cleverly Lyapunov inequality, we have

$$0 \leq (\mathsf{E}|X_n - X|^r)^{\frac{1}{r}} \leq (\mathsf{E}|X_n - X|^s)^{\frac{1}{s}}$$

and as the right hand side converges to zero, then so does the left hand side. ■

Theorem 11.6
Let X and $X_1, X_2, \ldots$ be random variables; for $r \geq 1$, $X_n \to X$are random vectors in r^{th} mean, then $X_n \xrightarrow{\mathsf{p}} X$. The former concept of convergence, although stronger, if often easier to establish.

Proof Using Markov inequality, we have

$$\begin{aligned} 0 \leq \mathsf{P}(|X_n - X| > \varepsilon) &= \mathsf{P}(|X_n - X|^r > \varepsilon^r) \\ &\leq \frac{\mathsf{E}|X_n - X|^r}{\varepsilon^r} \to 0 \text{ as } n \to \infty \end{aligned}$$

■

Definition 11.7 (Convergence in distribution)
Let $X_1, X_2, \ldots$ be a sequence of random variables. Then X_n converges in distribution (in

law, or weakly), to a random variable $X \rightsquigarrow X$, if

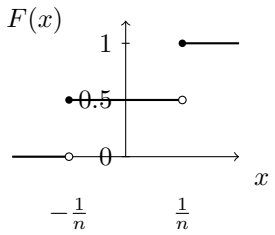
$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for all $x \in \mathbb{R}$ **such that** F **is continuous in** x.

Example 11.4

Let $X_n \in \left\{-\frac{1}{n}, \frac{1}{n}\right\}$ and $\mathsf{P}(X_n = -1/n) = \mathsf{P}(X_n = 1/n) = 1/2$.

The CDF of X_n is of the form



and

$$F_{X_n}(x) \to \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

and so we define

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0 \end{cases}$$

and $F_{X_n}(x) \to F(x)$ $allx \in (-\infty, 0) \cup (0, \infty)$ and so $X_n \rightsquigarrow 0$.

Remark

1. If $X_n \xrightarrow{\mathsf{p}} X$ and $X_n \to X$, then X is almost surely unique.
2. If $X_n \rightsquigarrow X$, then the distribution of X is unique.

Example 11.5

If $U_1, \ldots U_n$ are IID $\mathcal{U}(0,1)$, and suppose we look at $\max(U_1, \ldots, U_n) = U_{(n)},$, then the density of $U_{(1)}$ is

$$f_n(y) = u^{n-1} n \mathbf{1}_{(x\in(0,1))}$$

and the corresponding CDF of the maximum is

$$F_n(u) = \begin{cases} 0, & u < 0 \\ u^n, & u \in [0,1) \\ 1, & u \geq 1 \end{cases}$$

and

$$F_n(u) \xrightarrow[n\to\infty]{} = \begin{cases} 0, & u < 1 \\ 1, & u \geq 1 \end{cases}$$

the CDF of a constant 1, so $U_{(n)} \rightsquigarrow 1$. Looking at a rescaled version of the variable, $X_n \coloneqq n(1 - U_{(n)})$, then

$$\mathsf{P}(X_n \leq x) = \begin{cases} 0, & x < 0 \\ \mathsf{P}(n(1-U_{(n)}) \leq x) = \mathsf{P}(1 - \frac{x}{n} \geq U_{(n)}) = 1 - \left(1 - \frac{x}{n}\right)^n, & x \in [0,n)n, \\ 1, & x \geq n \end{cases}$$

and the limiting CDF as $n \to \infty$ is

$$\mathsf{P}(X \leq x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-x} & x > 0 \end{cases}$$

and $X_n \rightsquigarrow \mathcal{E}(1)$.

Theorem 11.8

We have

1. $X_n \xrightarrow{\mathsf{p}} X \Rightarrow X_n \rightsquigarrow X$
2. $X_n \rightsquigarrow c \Rightarrow X_n \xrightarrow{\mathsf{p}} c$, for c constant

Proof

1. Let x be a continuity point of F_X, then to have convergence in distribution means

$$F_{X_n}(x) \xrightarrow[n\to\infty]{} F_X(x).$$

Let $\varepsilon > 0$ be arbitrary, then

$$\begin{aligned}\mathsf{P}(X_n \leq x) &= \mathsf{P}(X_n \leq x, X \leq x+\varepsilon) + \mathsf{P}(X_n \leq x, X > x+\varepsilon) \\ &\leq F_X(x+\varepsilon) + \mathsf{P}(|X_n - X| > \varepsilon)\end{aligned}$$

and as $n \to \infty$, we get the second term on the RHS tends to zero. Taking $\varepsilon \to 0$, we get convergence. We however need a lower bound, found by taking

$$\begin{aligned}\mathsf{P}(X \leq x-\varepsilon) &= \mathsf{P}(X \leq x-\varepsilon, X_n \leq x) + \mathsf{P}(X \leq x-\varepsilon, X_n > x) \\ &\leq \mathsf{P}(X_n \leq x) + \mathsf{P}(|X_n - X| > \varepsilon).\end{aligned}$$

Together,

$$F_X(x-\varepsilon) + \mathsf{P}(|X_n - X| > \varepsilon) \leq F_{X_n}(x) \leq F_X(x+\varepsilon) + \mathsf{P}(|X_n - X| > \varepsilon)$$

and as $n \to \infty$, taking the limit $\mathsf{P}(|X_n - X| > \varepsilon) \to 0$, and $\forall\, \varepsilon > 0$,

$$F_X(x-\varepsilon) \leq \lim_{n\to\infty} F_{X_n}(x) \leq F_X(x+\varepsilon)$$

and both converge to $F_X(x)$ using continuity.

2. We know that $F_{X_n}(x) \to 0$ if $x < c$ and $F_{X_n}(x) \to 1$ if $x > c$. If we look at

$$\begin{aligned}\mathsf{P}(|X_n - c| > \varepsilon) &= \mathsf{P}(X_n - c > \varepsilon) + \mathsf{P}(c - X_n > \varepsilon) \\ &= \mathsf{P}(X_n > c+\varepsilon) + \mathsf{P}(X_n < c-\varepsilon) \\ &\leq 1 - \underbrace{\mathsf{P}(X_n \leq c+\varepsilon)}_{\to 1} + \underbrace{\mathsf{P}(X_n \leq c-\varepsilon)}_{\to 0}\end{aligned}$$

■

Theorem 11.9

Convergence in distribution is equivalent to pointwise convergence ($\forall\, t$ for which M_X exists) of the characteristic functions or the moment generating functions. In other words

Suppose that X_n, X have moment generating functions. Then, if $X_n \rightsquigarrow X \Leftrightarrow M_{X_n}(t) \xrightarrow{n\to\infty} M_X(t)$

1. $X_n \rightsquigarrow X \Leftrightarrow C_{X_n}(t) \Rightarrow C_X(t)\ \forall\, t \in \mathbb{R}$.

Theorem 11.10 (Central limit theorem)
Let $X_1, X_2, \ldots$ be an IID sequence of random variables such that $\mathsf{E}(X_i) = \mu, \mathsf{Var}(X_i) = \sigma^2 < \infty$, both moments assumed to exist. Then

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \rightsquigarrow \mathcal{N}(0,1)$$

or equivalently,

$$\sqrt{n}(\overline{X}_n - \mu) \rightsquigarrow \mathcal{N}(0,\sigma^2).$$

Proof (Sketch) Assume that X_i has a MGF. We could write the above as

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)\right)$$

and $\frac{X_i - \mu}{\sigma}$ are IID with mean zero and variance is one. Hence, assume wlog $\mu = 0, \sigma^2 = 1$. Thus, it suffices to show that $\sqrt{n}\overline{X}_n \overset{?}{\rightsquigarrow} \mathcal{N}(0,1)$. The MGF of $\overline{X}_n$ is

$$M_{\overline{X}_n}(t) = \left(M_X\left(\frac{t}{n}\right)\right)^n \Rightarrow M_{\sqrt{n}\overline{X}_n}(t) \qquad = \left(M_X\left(\frac{t}{\sqrt{n}}\right)\right)^n \xrightarrow[?]{n\to\infty} e^{t^2/2}$$

and since $M_X(0) = 1$ and M_X is differentiable at an arbitrary degree. Thus, taking the Taylor's expansion,

$$M_X\left(\frac{t}{\sqrt{n}}\right) = 1 + M_X'(0)\frac{t}{\sqrt{n}} + M_X''(0)\frac{t^2}{2n} + o\left(\frac{1}{n}\right)$$

hence

$$M_X\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)$$

and so

$$\left(1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \xrightarrow[n\to\infty]{} e^{t^2/2}$$

■

Example 11.6
Consider the simple case $X_n \sim \Gamma(n,1)$. We know that $X_n \overset{d}{=} Y_1 + \cdots + Y_n$ where $Y_i's$ are

IID $\mathcal{E}(1)$. By the CLT,

$$\sqrt{n}(\overline{Y}_n - 1)\ n \overset{\rightsquigarrow}{\to} \infty\ \mathcal{N}(0,1)$$

If n is large enough,

$$\sqrt{n}(\overline{Y}_n - 1) \approx \mathcal{N}(0,1)$$

and we approximate

$$\begin{aligned}
\mathsf{P}(X_n \leq x) &= \mathsf{P}(Y_1 + \cdots + Y_n \leq x) \\
&= \mathsf{P}\left(\sqrt{n}(\overline{Y}_n - 1) \leq \sqrt{n}\left(\frac{x}{n} - 1\right)\right) \\
&\approx \Phi\left(\sqrt{n}\left(\frac{x}{n} - 1\right)\right)
\end{aligned}$$

and similar results can be derived for $\mathcal{P}(\lambda), \chi^2(\nu)$ if $\lambda, \nu = n$ and n is large enough. What about the case where $X_n \sim \Gamma(\alpha, 1)$, then $X_n \overset{d}{n} Y_1 + \cdots + Y_n \sim F\left(\frac{\alpha}{n}, 1\right)$

Remark

Some results are available for non-IID variables (still independent), then CLT can be extended. The extensions are known for as Lindeberg CLT and Lyapunov Central limit theorem.

Theorem 11.11 (Continuous mapping theorem)

Let $X_1, X_2, \ldots$ be a sequence of random vectors in $\mathbb{R}^d, X$ a random variable $\in \mathbb{R}^d$. Let also $g : \mathbb{R}^d \to \mathbb{R}^n$ such that g is continuous in A and $\{(X \in A) = 1$.

1. If $X_n \xrightarrow{\text{a.s.}} X$, then $g(X_n) \xrightarrow{\text{a.s.}} g(X)$
2. If $X_n \xrightarrow{\mathsf{p}} X$, then $g(X_n) \xrightarrow{\mathsf{p}} g(X)$
3. If $X_n \rightsquigarrow X$, then $g(X_n) \rightsquigarrow g(X)$.

The proof will be omitted due to lack of time. Here is a proof in the univariate case (from MATH 356).

Proof

1. Given that $\mathsf{P}(\{\omega : X_n(\omega) \to X(\omega)\}) = 1$, once you fix ω, you get a sequence of numbers and continuity of the mapping
   $\mathsf{P}(\{\omega : g(X_n(\omega)) \to g(X(\omega))\}) = 1$ if continuous on $\mathbb{R}$.

2. $\forall\, k > 0, g$ is uniformly continuous on $[-k, k]$ by definition if

$$\big(\ \forall\ \varepsilon\ >\ 0\big)\big(\exists\, \delta_{(\varepsilon,k)}\ >\ 0\big)\big(\ \forall\ x\ \in\ [-k,k]\big)\big(|x-y|\ <\ \delta_{(\varepsilon,k)}\ \Rightarrow\ |g(x)-g(y)|\ <\ \varepsilon\big)$$

equivalent to saying that $|g(x) - g(y)| \geq \varepsilon \Rightarrow |x - y| \geq \delta_{(\varepsilon,k)}$. For all ϵ positive, we have by the Law of total probability

$$\begin{aligned}
&\mathsf{P}\big(|g(X_n) - g(X)| \geq \varepsilon\big) \\
&\quad = \mathsf{P}\big(|g(X_n) - g(X)| \geq \varepsilon, |X| \leq k\big) + \mathsf{P}\big(|g(X_n) - g(X)| \geq \varepsilon, |X| > k\big) \\
&\quad \leq \mathsf{P}\big(|X_n - X| \geq \delta_{(\varepsilon,k)}, |X| \leq k\big) + \mathsf{P}\big(|X| > k\big) \\
&\quad \leq \mathsf{P}\big(|X_n - X| \geq \delta_{(\varepsilon,k)}\big) + \mathsf{P}\big(|X| > k\big)
\end{aligned}$$

by first enlarging the probability and secondly using the fact that $\mathsf{P}(A \cap B) \leq \mathsf{P}(B)$. Since both probabilities are non-negative, we obtain

$$0 \leq \lim_{n\to\infty} \mathsf{P}\big(|g(X_n) - g(X)| \geq \varepsilon \leq 0 + P\big(|X| > k\big) \xrightarrow{k\to\infty} 0$$

as $n \to \infty$ so the limit must indeed be zero.

■

Lemma 11.12 (Slutsky)
Let $\{X_n\}$ and $\{Y_n\}$ be sequences of random variables.

1. If $X_n \rightsquigarrow X$ and $Y + n \xrightarrow{\mathsf{p}} c$, then

$$X_n + Y_n \rightsquigarrow X + c$$

2. If $X_n \rightsquigarrow X$ and $Y_n \xrightarrow{\mathsf{p}} c$, then

$$X_n Y_n \rightsquigarrow cX$$

and to zero if $c = 0$.

This follows from the continuous mapping theorem (CMT) because one can prove $(X_n, Y_n) \rightsquigarrow (X, c)$.

Example 11.7
By the Central limit theorem (CLT), $\sqrt{n}(\overline{X}_n - \mu) \rightsquigarrow \mathcal{N}(0, \sigma^2)$ this actually implies that

$\overline{X}_n \xrightarrow{\mathsf{p}} \mu$ by Slutsky theorem, then

$$\underbrace{\sqrt{n}(\overline{X}_n - \mu)}_{\rightsquigarrow \mathcal{N}(0,\sigma^2)} \underbrace{\left(\frac{1}{\sqrt{n}}\right)}_{\to 0} \rightsquigarrow 0$$

but here we need some existence requirements for variance.

Coming back to previous example, if $U_1, \dots U_n \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$, then $n(1 - U_{(n)}) \rightsquigarrow \mathcal{E}(1)$ and

$$\underbrace{n(1 - U_{(n)})}_{\rightsquigarrow \mathcal{E}(1)} \underbrace{\frac{1}{n}}_{\to 0} \rightsquigarrow 0$$

Observation

Suppose that $r > 0, c \in \mathbb{R}$ $\{X_n\}$ a sequence of random variables such that $n^r(X_n - a) \rightsquigarrow X$, then $X_n \xrightarrow{\mathsf{p}} a$.

Example 11.8 (Odds ratio)

If $X_1, X_2, \dots$ are IID Bernoulli, then we may be interested in the odds ratio $\theta = p/(1-p)$, where p is the probability of 1. An estimator using the sample proportion is

$$\theta_n = \frac{\overline{X}_n}{1 - \overline{X}_n}$$

By the CLT,

$$\sqrt{n}\frac{(\overline{X}_n - p)}{\sqrt{p(1-p)}} \rightsquigarrow \mathcal{N}(0,1)$$

and $\theta_n = g(\overline{X}_n), g(x) = x/(1-x)$ the continuous mapping theorem (CMT) would give us

$$g\left(\sqrt{n}(\overline{X}_n - p)\right) \rightsquigarrow g(X)$$

where $X \sim \mathcal{N}(0, p(1-p)$ but this is of no use. Rather, we need $n^r(\theta_n - \theta) = n^r(g(\overline{X}_n) - g(p))$; we know that $\theta_n \xrightarrow{\mathsf{p}} \theta$. If we take $g(\overline{X}_n)$, where g is assumed to be differentiable, we could take a Taylor series expansion to get the approximation

$$g(\overline{X}_n) \approx g(p) + g'(p)(\overline{X}_n - p)$$

and rearranging these terms,

$$\sqrt{n}\left(g(\overline{X}_n) - g(p)\right) \approx g'(p) \underbrace{\sqrt{n}(\overline{X}_n - p)}_{\rightsquigarrow \mathcal{N}(0,p(1-p))} \rightsquigarrow \mathcal{N}\left(0, \left\{g'(p)^2\right\} p(1-p)\right)$$

and this motivates the following approximation, under some smoothness conditions and assuming that the scaling is such that $g'(p) \neq 0$

Theorem 11.13 (Delta method)
Let $Z_1, Z_2, \ldots$ a sequence of random variables such that $\sqrt{n}(Z_n - \theta) \rightsquigarrow \mathcal{N}(0, \sigma^2)$. Suppose that g is twice differentiable and $g'(\theta) \neq 0$.

Example 11.9
Continuing the odds ratio example, we have $\theta = p/(1-p) = g(p)$ and $g'(p) = \frac{1-p+p}{(1-p)^2} = (1-p)^{-2}$ non-zero if $p \in (0,1)$ Then

$$\sqrt{n}\left(\frac{\overline{X}_n}{1-\overline{X}_n} - \frac{p}{1-p}\right) \rightsquigarrow \mathcal{N}\left(0, \frac{p(1-p)}{(1-p)^4}\right)$$

and the variance is $p/(1-p)^3$.

Example 11.10
If $X_1, X_2, \ldots \overset{\text{iid}}{\sim} \mathcal{E}(\lambda)$ for $\lambda > 0$, then $\lambda_n = 1/\overline{X}_n$ and so taking $g(x) = 1/x$, we get $g'(x) = -1/x^2$ and $g'(\lambda) > 0$.

$$\sqrt{n}(\lambda_n - \lambda) = \sqrt{n}\left(\frac{1}{\overline{X}_n} - \lambda\right) \rightsquigarrow \mathcal{N}\left(0, \frac{1}{\lambda^2}\frac{1}{\lambda^4}\right)$$

We omit the proof of the Delta method, which can be found in the book and deals with details in a careful fashion.

Example 11.11
If $X_1, \ldots, X_n$ are random variables $X_i \sim \mathcal{B}(p)$. Then $\mathsf{Var}(X) = p(1-p) = \sigma^2$, and $\sigma_n^2 = \overline{X}_n(1-\overline{X}_n)$. Now, if we take $g(x) = x(1-x), g'(x) = 1-2x$ and we are in big trouble if $p = \frac{1}{2}$. However if $p \neq \frac{1}{2}$, then

$$\sqrt{n}\left(\sigma_n^2 - \sigma^2\right) \rightsquigarrow \mathcal{N}\left(0, p(1-p)(1-2p)^2\right)$$

and if $p = \frac{1}{2}$, we will pursue with a quick and dirty calculation, and we need to get the appropriate scaling parameter. However, if we go further in the Taylor's series expansion,

$$g(\overline{X}_n) = g(\theta) + g'(t)(Z_n - \theta) + g''(\theta)/2(Z_n - \theta)^2$$

and $g(Z_n) - g(\theta) \approx g''(\theta)/2(Z_n - \theta)^2$ and we need to multiply by n. Indeed, we know that $\sqrt{n}(Z_n - \theta) \rightsquigarrow \mathcal{N}(0, \sigma^2)$ as

$$n(g(Z_n) - g(\theta)) \approx \frac{g''(\theta)}{2}\left(\sqrt{n}(Z_n - \theta)\right)^2$$

Theorem 11.14 (Delta Method II)

If Z_n and θ as before, that is $\sqrt{n}(Z_n - \theta) \rightsquigarrow \mathcal{N}(0, \sigma^2)$. Suppose that g is three times differentiable, $g'(\theta) = 0, g''(\theta) \neq 0$ then

$$n(g(Z_n) - g(\theta)) \rightsquigarrow \frac{g''(\theta)}{2}\sigma^2\chi^2(1)$$

Example 11.12

We have $n\left(\overline{X}_n(1 - \overline{X}_n) - p(1-p)\right) \rightsquigarrow -\frac{1}{4}\chi^2(1)$.

License