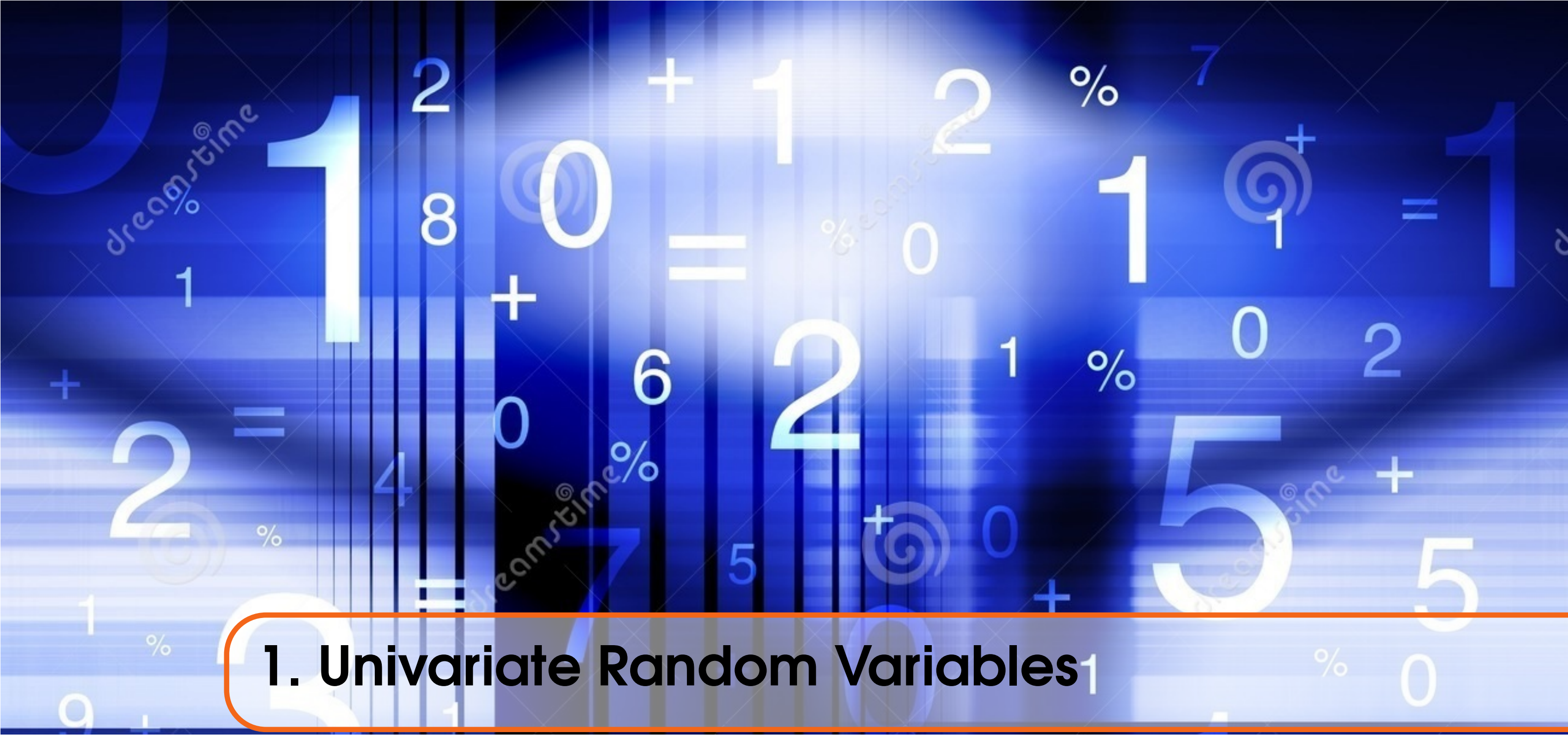
STAT330 Review Notes

University of Waterloo

The One And Only
Waterloo 76er
Bill Zhuo

Índice general



1. Univariate Random Variables

1.1 Probability Space

Definition 1.1.1 Sample Space
S of a random experiment is the set of all distinct possible outcomes of the random experiment.

Definition 1.1.2 σ-Algebra
A collection of subsets of S is called σ-algebra denoted as $\mathscr{F}$, if it satisfies

1. $\emptyset \in \mathscr{F}$
2. $A \in \mathscr{F}$ implies that $A^c \in \mathscr{F}$
3. If $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathscr{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathscr{F}$.

Definition 1.1.3 Event
An element $A \in \mathscr{F}$ is called an event if the result ω of the random experiment belongs to A, $w \in A$, we say event A happens.

Definition 1.1.4 Probability Measure
A probability measure $\mathbf{P}$ is a set function

$$\mathbf{P} : \{A_1, A_2, \dots\} \to \mathbb{R}$$

defined on $\mathscr{F}$ satisfying

1. $\mathbf{P}(A) \geq 0, \forall A \in \mathscr{F}$ (Positivity)
2. $\mathbf{P}(S) = 1$ (Unity)

3. If $\{A_i\}_{i\in\mathbb{N}}$ with $A_i \cap A_j = \emptyset, \forall i \neq j$, then

$$\mathbf{P}\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mathbf{P}(A_i)$$

1.2 Properties of Probabilities

Definition 1.2.1 Probability Space
We call the triplet $(S, \mathscr{F}, \mathbf{P})$ to be a probability space

Proposition 1.2.1 Basic Properties of Probabilities

1. $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$
2. $\mathbf{P}(A) \leq 1, \forall A \in \mathscr{F}$
3. $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$ (subadditivity) and

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B) \text{ (ex-inc principle)}$$

4. $A, B \in \mathscr{F}, A \subseteq B$ implies that $\mathbf{P}(A) \leq \mathbf{P}(B)$
5. **Boole's Inequality:**

$$\mathbf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbf{P}(A_i)$$

1.3 Conditional Probability and Independence

Definition 1.3.1 Conditional Probability
The conditional probability of an event A given event B is defined by

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}, \mathbf{P}(B) > 0$$

R Given $B \in \mathscr{F}$ such that $\mathbf{P}(B) > 0$. $\mathbf{P}(\cdot|B)$ is also a probability set function based on the same σ-algebra.

Theorem 1.3.1 Total Probability Formula
If $B_1, \ldots, B_k$ are disjoint events and

$$\bigcup_{i=1}^{k} B_k = S$$

then

$$\mathbf{P}(A) = \sum_{i=1}^{k} \mathbf{P}(A|B_i)\,\mathbf{P}(B_i), \forall A \in \mathscr{F}$$

Theorem 1.3.2 Bayes' Rule
If $B_1, \ldots, B_k$ are disjoint events and

$$\bigcup_{i=1}^{k} B_k = S$$

then $\forall A \in \mathscr{F}$ such that $\mathbf{P}(A) > 0$, we have

$$\mathbf{P}(B_j|A) = \frac{\mathbf{P}(A|B_j)\mathbf{P}(B_j)}{\sum_{i=1}^{k}\mathbf{P}(A|B_i)\mathbf{P}(B_i)}$$

R It should be clear that

$$\sum_{i=1}^{k}\mathbf{P}(A|B_i)\mathbf{P}(B_i) = \mathbf{P}(A) \textbf{ as the total probability}$$

Definition 1.3.2 Event Independence
Two events are called independent, $A, B \in \mathscr{F}$, if $\mathbf{P}(A\cap B) = \mathbf{P}(A)\mathbf{P}B$, we denote this by $A \perp\!\!\!\perp B$. If $\mathbf{P}(B) > 0$m then A and B are independent if and only if $\mathbf{P}(A|B) = \mathbf{P}(A)$.

R Note that $A \perp\!\!\!\perp B$ is different from $A \cap B = \emptyset$!

1.4 Random Variables

Definition 1.4.1 Random Variable
A random variable X is a function from the sample space S to $\mathbb{R}$ such that $\{X \leq x\} := \{\omega : X(\omega) \leq x\} \in \mathscr{F}, \forall x \in \mathbb{R}$. In simpler words, the probability measure $\mathbf{P}(X \leq x) = \mathbf{P}(\{\omega : X(\omega) \leq x\})$ is well-defined.

Definition 1.4.2 CDF
The cumulative distribution function (CDF) of random variable X is defined by

$$\mathbf{F}(x) = \mathbf{P}(X \leq x), \forall x \in \mathbb{R}$$

Proposition 1.4.1 Properties of CDF

1. $\mathbf{F}$ is non-decreasing
$$\mathbf{F}(x_1) \leq \mathbf{F}(x_2), \forall x_1 < x_2$$
2. $\operatorname{lím}_{x\to-\infty}\mathbf{F}(x) = 0$ and $\operatorname{lím}_{x\to\infty}\mathbf{F}(x) = 1$
3. **Right-continuous:**
$$\operatorname*{lím}_{x\to a^+}\mathbf{F}(x) = \mathbf{F}(a)$$

Theorem 1.4.2 Getting Probabilities From CDF

1. $\mathbf{P}(a < x \leq b) = \mathbf{P}(X \leq b) - \mathbf{P}(X \leq a) = \mathbf{F}(b) - \mathbf{F}(a), \forall a < b$
2. $\mathbf{P}(X = a) = \mathbf{F}(a) - \operatorname{lím}_{x\to a^-}\mathbf{F}(a)$ (Vertical Jump)

1.5 Discrete and Continuous Random Variables

1.5.1 Discrete Random Variables

Definition 1.5.1 Discrete Random Variable
A random variable X is called discrete, if there exists an at most countable set $\{x_1, x_2, \dots\}$ such that

$$\sum_{i=1}^{\infty} \mathbf{P}(X = x_i) = 1$$

R X only takes values from an at most countable set. The distribution of a discrete random variable is characterized by its **probability mass function (pmf)**

$$\mathbf{f}(x) := \mathbf{P}(X = x), x \in \mathscr{A}$$

where $\mathscr{A}$ is the support set.

Proposition 1.5.1 Properties of pmf

1. $\mathbf{f}(x) \geq 0, \forall x \in \mathscr{A}$
2. $\sum_{i=1}^{\infty} \mathbf{f}(x_i) = 1$

R On the other hand, any function satisfying these two properties is the pmf os some discrete random variable.

R The CDF of a discrete random variable is a step function

$$\begin{aligned}\mathbf{F}(x) &= \mathbf{P}(X \leq x) \\ &= \sum_{x_i \leq x} \mathbf{f}(x_i) \\ &= \sum_{x_i} \mathbf{f}(x_i) \mathbf{1}_{\{x_i \leq x\}}\end{aligned}$$

Thus, for a discrete random variable from pmf to cdf we have

$$\mathbf{F}(x) = \sum_{x_i \leq x} \mathbf{f}(x_i) \Longrightarrow \mathbf{f}(x) = \mathbf{F}(x) - \lim_{y \to x^-} F(y)$$

1.5.2 Continuous Random Variable

Definition 1.5.2 Continuous Random Variable (CDF & pdf)
A random variable X is called continuous if there exists a function $\mathbf{f}(x)$ such that the CDF $\mathbf{F}(x)$ of X can be written as

$$\mathbf{F}(x) = \int_{-\infty}^{x} \mathbf{f}(t) dt, \forall x \in \mathbb{R}$$

the function $\mathbf{f}(x)$ is called the **probability density function (pdf)** of X.

Theorem 1.5.2 From CDF to pdf:
If X is continuous, $\mathbf{F}(x)$ is continous everywhere and differentiable almost everywhere we can take

$$\mathbf{f}(x) = \frac{\partial}{\partial x} \int_{-\infty}^{x} \mathbf{f}(t) dt = \frac{\partial}{\partial x} \mathbf{F}(x)$$

when the derivative exists, and $\mathbf{f}(x) = 0$ otherwise and then this $\mathbf{f}(x)$ is the pdf of X.

R

1. Note that pdf is not a probability measure, i.e
$$\mathbf{f}(a) = \frac{1}{2} \not\Leftrightarrow \mathbf{P}(X = a) = \frac{1}{2}$$
we can even have $\mathbf{f}(x) > 1$. It only gives probability only when integrated.
2. For random variable X, it always has CDF, but not always a pdf or pmf.

Proposition 1.5.3 Properties of pdf

1. $\mathbf{f}(x) \geq 0$
2. $\int_{-\infty}^{\infty} \mathbf{f}(x)dx = 1$

R On the other hand, any function satisfying these two conditions is the pdf of some continuous random variable.

Theorem 1.5.4 From pdf to Probability

$$\begin{aligned}\mathbf{P}(x_1 < X \leq x_2) &= \mathbf{P}(X \in (x_1, x_2]) \\ &= \mathbf{P}(X \leq x_2) - \mathbf{P}(X \leq x_1) \\ &= \mathbf{F}(x_2) - \mathbf{F}(x_1) \\ &= \int_{x_1}^{x_2} \mathbf{f}(t)dt\end{aligned}$$

R For continuous random variable we have
$$\mathbf{P}(X \in [x_1, x_2]) = \mathbf{P}(X \in (x_1, x_2]) = \mathbf{P}(X \in [x_1, x_2)) = \mathbf{P}(X \in (x_1, x_2))$$

1.6 Expectation

Definition 1.6.1 Expectation

The expectation (mean) of a discrete random variable X with pmf $\mathbf{f}(x)$ is defined by
$$\mathbb{E}(X) = \sum_{x \in \mathscr{A}} x\mathbf{f}(x) = \sum_{x \in \mathscr{A}} x\mathbf{P}(X = x)$$
for continuous random varaible with pdf, the expectation is defined by
$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\mathbf{f}(x)dx$$

R Let h be a nice function, then $h(X)$ is also a random variable.

Theorem 1.6.1 Expectation of $h(X)$
If X is a random variable with pdf/pmf $\mathbf{f}(x)$ and h is a function, then

$$\mathbb{E}\left(h(X)\right) = \begin{cases} \sum_{x\in\mathscr{A}} h(x)\mathbf{f}(x) & X \textbf{ discrete} \\ \int_{\infty}^{\infty} h(x)\mathbf{f}(x)dx & x \textbf{ continuous} \end{cases}$$

R The expectation of a random variable does not always exist!

Theorem 1.6.2 Linearity of Expectation
Let X be a random variable $a,b \in \mathbb{R}$, g,h be two nice functions, then

$$\mathbb{E}\left(ag(X)+bh(X)\right) = a\mathbb{E}\left(g(X)\right)+b\mathbb{E}\left(h(X)\right)$$

Definition 1.6.2 Variance
The variance of a random variable X is defined by

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

Proposition 1.6.3 Properties of Variance

1. $\mathbf{Var}\left(X\right) = \mathbb{E}\left(X^2\right) - \left(\mathbb{E}\left(X\right)\right)^2$
2. $\mathbf{Var}\left(aX+b\right) = a^2\mathbf{Var}\left(X\right)$

Definition 1.6.3 k-th Moments
The k-th moment of random variable X is the $\mathbb{E}\left(X^k\right)$.
The k-th moment about mean of X is $\mathbb{E}\left(\left(X-\mathbb{E}\left(X\right)\right)^k\right)$.

1.7 Moment Generating Functions

Definition 1.7.1 MGF
Let X be a random variable, then the function

$$\mathbf{M}\left(t\right) := \mathbb{E}\left(e^{tX}\right)$$

is called the moment generating function (MGF) of X, if the expectation exists for all $t \in (-h,h)$ for some $h > 0$.

R It is important to check the existence of the expectation. The MGF is only well-defined if the expectation exists.

Theorem 1.7.1 Properties of MGF

1. $\mathbf{M}\left(0\right) = 1$

2. $\mathbf{M}^{(k)}0 = \mathbb{E}\left(X^k\right), k = 1, 2, \ldots$
$$\frac{\partial^k}{\partial t^k}\mathbf{M}(t)\bigg|_{t=0}$$

Theorem 1.7.2 Uniqueness of MGF
The MGF completely determines the distribution of a random variable

$$\mathbf{M}_X(t) = \mathbf{M}_Y(t), \forall t \in (-h, h) \textbf{ for some } h > 0$$

then $\mathbf{F}_X(s) = \mathbf{F}_Y(s), \forall s \in \mathbb{R}$.

1.8 Special Distributions

1.8.1 Discrete Random Variable Examples

Bernoulli (Toss a Coin Once) $X \sim Bin(1, p)$

Definition 1.8.1 A Bernoulli distribution is characterized as $X \sim Bin(1, p)$ with the

1. **pmf**
$$\mathbf{f}(x) = \begin{cases} 1-p & x = 0 \\ p & x = 1 \end{cases}$$
2. **Expectation:** $\mathbb{E}(X) = p$
3. **Variance:** $\mathbf{Var}(X) = p(1-p)$
4. **MGF:** $\mathbf{M}(t) = \mathbb{E}\left(e^{tX}\right) = (1-p) + pe^t$

Geometric (# of Tails Until a Success) $X \sim Geo(p)$

1. **pmf**
$$\mathbf{f}(x) = p(1-p)^{x-1}, x = 1, 2, \ldots; p \in (0, 1]$$
2. **CDF**
$$\mathbf{F}(x) = 1 - (1-p)^x$$
We also have something called **survival function**: $1 - \mathbf{F}(x) = (1-p)^n$
3. **Expectation:** $\mathbb{E}(X) = \frac{1}{p}$
4. **Variance:** $\mathbf{Var}(X) = \frac{1-p}{p^2}$
5. **MGF:** $\mathbf{M}(t) = \frac{pe^t}{1-(1-p)e^t}$ for $t < -\log(1-p)$

Theorem 1.8.1 Memoryless Property
Let $X \sim Geo(p)$, then

$$\mathbf{P}(X > m + n | X > m) = \mathbf{P}(X > n), \forall m, n \in \mathbb{N}$$

Knowing the first m trials have passed without success will not affect how many trials are still needed to get the next success.

Poisson Distribution (# of rare events happen in a fixed amount of time) $X \sim Poi(\lambda)$

1. **pmf**
$$\mathbf{f}(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \lambda > 0, x = 0, 1, 2, \ldots$$
2. **Expectation:** $\mathbb{E}(X) = \lambda$

3. **Variance:** $\mathbf{Var}\left(X\right) = \lambda$
4. **MGF:** $\mathbf{M}\left(t\right) = e^{\lambda(e^t-1)}, t \in \mathbb{R}$

1.8.2 Continuous Distribution Examples

Exponential Distribution $X \sim Exp(\lambda)$

1. **pdf**

$$\mathbf{f}(x) = \lambda^{-\lambda x}, \lambda > 0, x > 0$$

2. **CDF**

$$\mathbf{F}(x) = 1 - e^{-\lambda x}, x > 0$$

3. **Expectation:** $\mathbb{E}\left(X\right) = \frac{1}{\lambda}$
4. **Variance:** $\mathbf{Var}\left(X\right) = \frac{1}{\lambda^2}$
5. **MGF:** $\mathbf{M}\left(t\right) = \frac{\lambda}{\lambda - t}, t < \lambda$

Theorem 1.8.2 Memoryless Property
Let $X \sim Exp(\lambda)$, then

$$\mathbf{P}(X > t + s | X > s) = \mathbf{P}(X > t), \forall t, s \in \mathbb{R}$$

R The future lifetime does not depend on the current age.

R **Memoryless distribution Uniqueness**

1. Discrete $\iff$ Geometric Distribution
2. Continuous $\iff$ Exponential Distribution

Normal Distribution $X \sim N(\mu, \sigma^2)$

1. **pdf**

$$\mathbf{f}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \mu \in \mathbb{R}, \sigma > 0$$

2. **CDF**: Φ
3. **Expectation:** $\mathbb{E}\left(X\right) = \mu$
4. **Variance:** $\mathbf{Var}\left(X\right) = \sigma^2$
5. **MGF:** $\mathbf{M}\left(t\right) = e^{\mu + \frac{1}{2}\sigma^2 t^2}$

1.9 Location and Scale Parameters

Definition 1.9.1 Location Parameter
Let $f(x;\theta)$ or $F(x;\theta)$ be a group of (pmf/pdf) or CDF with a parameter θ, θ is called a location parameter if one of the following is true:

1. $\mathbf{f}(x;\theta) = \mathbf{f}_0(x-\theta)$ where $\mathbf{f}_0(x) = \mathbf{f}(x;0)$ or
2. $\mathbf{F}(x;\theta) = \mathbf{F}_0(x-\theta)$ where $\mathbf{F}_0(x) = \mathbf{F}(x;0)$

Definition 1.9.2 Scaled Parameter
Let $f(x;\theta)$ or $F(x;\theta)$ be a group of (pmf/pdf) or CDF with a parameter θ, θ is called a scale parameter if one of the following is true:

1. $\mathbf{f}(x;\theta) = \frac{1}{\theta}\mathbf{f}_1\left(\frac{x}{\theta}\right)$ where $\mathbf{f}_1(x) = \mathbf{f}(x;1)$

2. $\mathbf{F}(x;\theta) = \mathbf{F}_1\left(\frac{x}{\theta}\right)$ where $\mathbf{F}_1(x) = \mathbf{F}(x;1)$

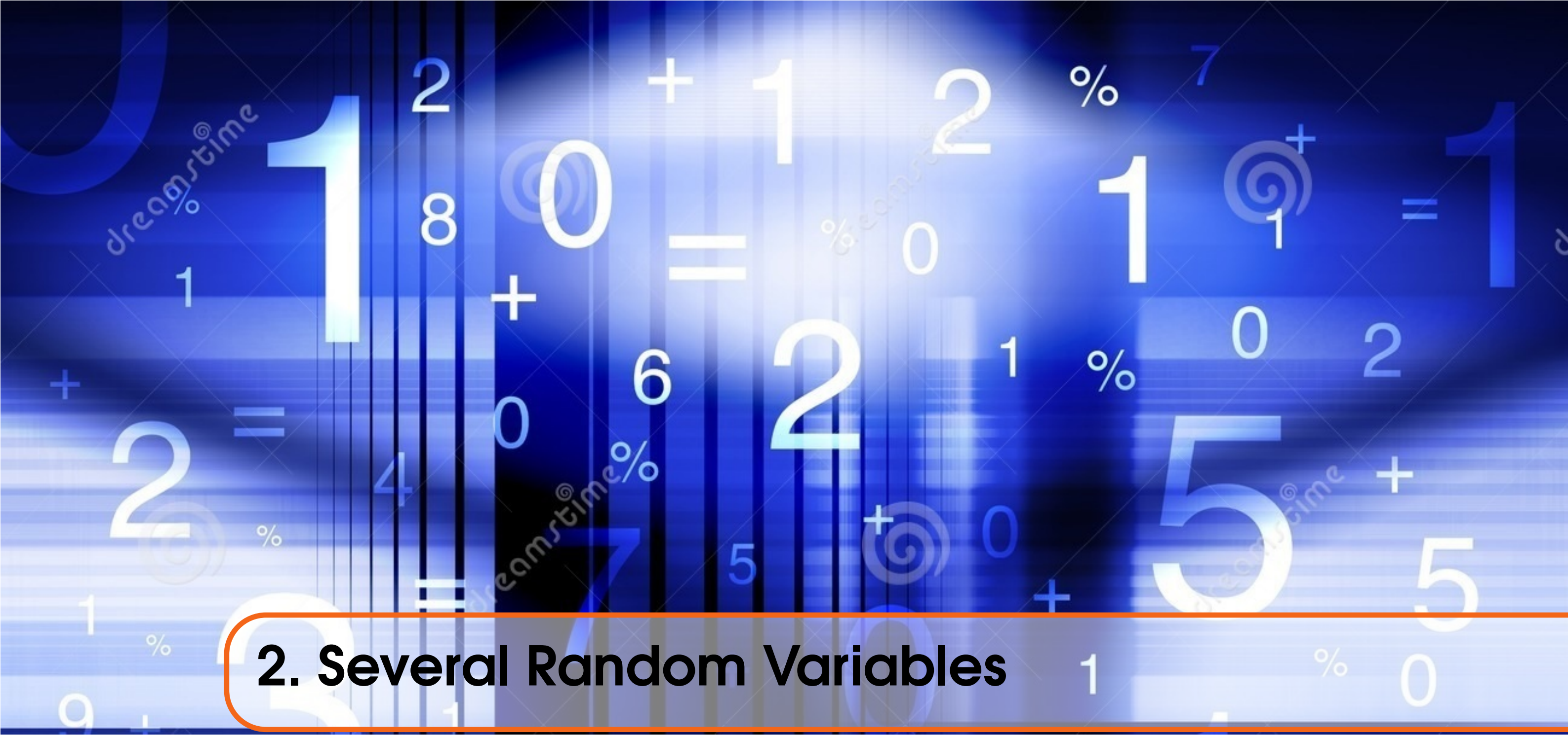
1.10 Probability Inequalities

Theorem 1.10.1 Markov's Inequality
Let X be a **non-negative** random variable, then

$$\mathbf{P}(X \geq c) = \frac{\mathbb{E}(X)}{c}, c > 0$$

Theorem 1.10.2 Chebyshev's Inequality
Let X be a random variable with mean μ and variance σ^2, then

$$\mathbf{P}(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$



2. Several Random Variables

2.1 Joint CDF

Definition 2.1.1 Joint CDF
Let X and Y be two random variables defined on a probability space, then the joint CDF of X and Y is defined as

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

R We can talk about the joint behaviour of several random variabes only if they are defined on the same probability space.

Proposition 2.1.1 Properties of Joint CDF

1. $\mathbf{F}$ is non-decreasing in x and y
2. $\operatorname{lím}_{x\to-\infty} \mathbf{F}(x,y) = \operatorname{lím}_{y\to-\infty} \mathbf{F}(x,y) = 0$
3. $\operatorname{lím}_{x\to\infty, y\to\infty} \mathbf{F}(x,y) = 1$
4. The CDF of the single variable X or Y is called the marginal CDF of X or Y

$$\operatorname*{lím}_{y\to\infty} \mathbf{F}(x,y) = \mathbf{F}(x,\infty) = \mathbf{F}_X(x)$$
$$\operatorname*{lím}_{x\to\infty} \mathbf{F}(x,y) = \mathbf{F}(\infty,y) = \mathbf{F}_Y(x)$$

2.2 Joint Discrete Distribution

Definition 2.2.1 Joint pmf
Let X and Y be discrete random varaibles defined on a probability space. Then, the joint probability mass function of X and Y is defined as

$$\mathbf{f}(x,y) = \mathbf{P}(X = x; Y = y)$$

the set $\mathscr{A} := \{(x,y) : f(x,y) > 0\}$ is called the **support** of (X,Y).

Proposition 2.2.1 Properties of Joint pmf

1. $\mathbf{f}(x,y) \geq 0, \forall x,y \in \mathbb{R}$
2. $$\sum_{(x,y)\in\mathbb{R}^2} \mathbf{f}(x,y) = \sum_{(x,y)\in\mathscr{A}} \mathbf{f}(x,y) = 1$$
3. $$\begin{aligned} \mathbf{P}((X,Y)\in B) &= \sum_{(x,y)\in B} \mathbf{f}(x,y) \\ &= \sum_{(x,y)\in\mathscr{A}\cap B} \mathbf{f}(x,y) \end{aligned}$$
4. **Marginal pmf**
$$\begin{aligned} \mathbf{f}_X(x) &= \sum_y \mathbf{f}(x,y) \\ \mathbf{f}_Y(x) &= \sum_x \mathbf{f}(x,y) \end{aligned}$$

2.3 Joint Continuous Distribution

Definition 2.3.1 Joint pdf
A pair of random variable (X,Y) is called continuous if there exits a function $\mathbf{f}(x,y)$ such that thee joint CDF can be written as

$$\mathbf{F}(x,y) = \int_{\infty}^{y}\int_{-\infty}^{x} \mathbf{f}(s,t)dsdt, \forall x,y \in \mathbb{R}$$

then $\mathbf{f}(x,y)$ is called the joint pdf of (X,Y).

R This can be extended to higher dimensions, for the purpose of review, we simply don't care...

Theorem 2.3.1 From Joint CDF to Joint pdf
If (X,Y) is continuous, then

$$\mathbf{f}(x,y) = \begin{cases} \frac{\partial^2}{\partial x \partial y}\mathbf{F}(x,y) & \textbf{when it exits} \\ 0 & \textbf{otherwise} \end{cases}$$

Proposition 2.3.2 Properties of Joint pdf

1. $\mathbf{f}(x,y) \geq 0, \forall x,y \in \mathbb{R}$
2. $\mathbf{P}((X,Y)\in B) = \int\int_B \mathbf{f}(x,y)dxdy$
3. **Joint pdf to Marginal pdf**:
$$\begin{aligned} \mathbf{f}_X(x) &= \int_{-\infty}^{\infty} \mathbf{f}(x,y)dy \\ \mathbf{f}_Y(x) &= \int_{-\infty}^{\infty} \mathbf{f}(x,y)dx \end{aligned}$$

2.4 Independence of Random Variables

Definition 2.4.1 Independence of Random Variables
Two random variables X,Y are called independent if

$$\mathbf{P}(X \in A, Y \in B) = \mathbf{P}(X \in A)\mathbf{P}(Y \in B), \forall A, B \subseteq \mathbb{R}$$

In other words, $\{X \in A\}, \{Y \in B\}$ are independent events for all sets $A, B \subseteq \mathbb{R}$. We note this by $X \perp\!\!\!\perp Y$.

Theorem 2.4.1 Independent Joint CDF and pmf/pdf Breakdown
Two random variables X,Y are independent if and only if

1. $\mathbf{F}(x,y) = \mathbf{F}_X(x)\mathbf{F}_Y(y), \forall x,y \in \mathbb{R}$ or equivalently
2. $\mathbf{f}(x,y) = \mathbf{f}_X(x)\mathbf{f}_Y(y), \forall x,y \in \mathbb{R}$ where $\mathbf{f}$ is a joint pdf/pmf of X and Y.

Theorem 2.4.2 Functions of Independent Random Variables Are Independent
If X and Y are independent, then $h(X)$ and $g(Y)$ are also independent where h,g are real-valued functions.

R It is usually quite hard to check independence of random variables without using the following theorem.

Theorem 2.4.3 Factorization Theorem
Let X,Y be two random variables with joint pdf/pmf $\mathbf{f}(x,y)$. Let $\mathscr{A}$ be the support of X and Y. Then, X and Y are independent random variables if and only if

1. $\mathbf{f}(x,y) = g(x)h(y)$ for some g,h **non-negative** functions, and
2. $\mathscr{A} = \mathscr{A}_1 \times \mathscr{A}_2, \forall \mathscr{A}_1, \mathscr{A}_2 \subseteq \mathbb{R}$.

R We need to support to breakdown into a cartesian product on $\mathbb{R}^2$. If you see that the support region is not a rectangle, you should know that these two random variables are not independent immediately.

2.5 Conditional Distribution

Definition 2.5.1 Conditional pdf/pmf
Let X,Y be two random variables with joint pdf/pmf $\mathbf{f}(x,y)$ and marginal pdfs/pmfs $\mathbf{f}_X(x)$ and $\mathbf{f}_Y(y)$ respectively. The conditional pdf/pmf of X given $Y = y$ is given by

$$\mathbf{f}_{X|Y}(x|y) = \frac{\mathbf{f}(x,y)}{\mathbf{f}_Y(y)} \textbf{ for } y \textbf{ such that } \mathbf{f}_Y(y) > 0$$

and define $\mathbf{f}_{X|Y}(x|y) = 0$ otherwise.

R The conditional distributions are legitimate distributions. They have all the properties of distribution that we have seen before.

Theorem 2.5.1 Product Rule

$$\mathbf{f}(x,y) = \mathbf{f}_{X|Y}(x|y)\mathbf{f}_Y(y) = \mathbf{f}_{Y|X}(y|x)\mathbf{f}_X(x)$$

Theorem 2.5.2 Independence on Conditional pmf/pdf
Let X,Y be two random variables with marginal pdfs/pmfs $\mathbf{f}_X(x)$ and $\mathbf{f}_Y(y)$ be conditional pdfs/pmfs, $\mathbf{f}_{X|Y}(x|y)$ and $\mathbf{f}_{Y|X}(y|x)$, then

$$X \perp\!\!\!\perp Y \iff \mathbf{f}_{X|Y}(x|y) = \mathbf{f}_X(x), \forall x,y \in \mathbb{R}, y \in \mathscr{A}_2 \textbf{(support of } Y\textbf{)}$$

2.6 Joint Expectation

Definition 2.6.1 Joint Expectation
Suppose $h(x,y)$ is a real-value function. For X,Y discrete, with a joint pmf $\mathbf{f}(x,y)$, we have

$$\mathbb{E}(h(X,Y)) = \sum_{(x,y)\in\mathscr{A}} h(x,y)\mathbf{f}(x,y)$$

For X,Y continuous, with a joint pdf $\mathbf{f}(x,y)$, we have

$$\mathbb{E}(h(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{f}(x,y)dxdy$$

R As in the single random variable case, the expectation does not always exist.

Proposition 2.6.1 Properties of Joint Expectation

1. **Linearity:** $\mathbb{E}(aX+bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$
2. $\mathbb{E}(ag(X,Y)+bh(X,Y)) = a\mathbb{E}(g(X,Y)) + b\mathbb{E}(h(X,Y))$

Theorem 2.6.2 Independent Expectation Breakdown
If $X \perp\!\!\!\perp Y$, then

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$$

Corollary 2.6.3 1. If $X \perp\!\!\!\perp Y$, then $\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$

2. Let $X_1,\dots,X_n$ be independent random variables and $h_1,\dots,h_n$ be real-valued functions, then

$$\mathbb{E}\left(\prod_{i=1}^{n} h_i(X_i)\right) = \prod_{i=1}^{n}\mathbb{E}(h_i(X_i))$$

Definition 2.6.2 Covariance
The covariance between X,Y is defined as

$$\mathbf{Cov}(X,Y) = \mathbb{E}((X-\mathbb{E}(x))(Y-\mathbb{E}(Y)))$$

Proposition 2.6.4 Properties of Covariance

1. $\textbf{Cov}(X,Y) = \textbf{Cov}(Y,X)$
2. $\textbf{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
3. If $X \perp\!\!\!\perp Y$, then $\textbf{Cov}(X,Y) = 0$ **(the converse is not generally true)**
4. $\textbf{Var}(aX + bY + c) = a^2\textbf{Var}(X) + b^2\textbf{Var}(Y) + 2ab\textbf{Cov}(X,Y)$
5. $\textbf{Cov}(a_1X_1 + b_1, a_2Y + b_2) = a_1a_2\textbf{Cov}(X,Y)$

Corollary 2.6.5 Let $X_1, \dots, X_n$ be independent, then

$$\textbf{Var}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2\textbf{Var}(X_i)$$

Definition 2.6.3 Correlation Coefficient
The correlation coefficient of random variables X,Y is defined as

$$\rho(X,Y) = \frac{\textbf{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

Theorem 2.6.6 Linear Relation Implication by ρ

1. $-1 \le \rho(X,Y) \le 1$
2. $\rho(X,Y) = 1$ if and only if $Y = aX + b$ for some $a > 0$
3. $\rho(X,Y) = -1$ if and only if $Y = aX + b$ for some $a < 0$

2.7 Conditional Expectation

Definition 2.7.1 Conditional Expectation
The conditional expectation of $g(Y)$ given $X = x$ is defined as

$$\mathbb{E}(g(Y)|X = x) = \begin{cases} \sum_y g(y)\mathbf{f}_{Y|X}(y|x) & Y \textbf{ discrete} \\ \int_{-\infty}^{\infty} g(y)\mathbf{f}_{Y|X}(y|x)dy & Y \textbf{ continuous} \end{cases}$$

R A special case that requires some attention

$$\textbf{Var}(Y|X = x) = \mathbb{E}\left((Y - \mathbb{E}(Y|X = x))^2|X = x\right) = \mathbb{E}\left(Y^2|X = x\right) - (\mathbb{E}(Y|X = x))^2$$

Theorem 2.7.1 Independent Expectation

$$X \perp\!\!\!\perp Y \implies \mathbb{E}(Y|X = x) = \mathbb{E}(Y)$$

Corollary 2.7.2

$$X \perp\!\!\!\perp Y \implies \mathbb{E}(g(Y)|X = x) = \mathbb{E}(g(Y))$$

Theorem 2.7.3 Iterated Expectation

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$$

Corollary 2.7.4

$$\mathbb{E}\left(\mathbb{E}\left(g(Y)|X\right)\right) = \mathbb{E}\left(g(Y)\right)$$

Theorem 2.7.5 Decomposition of Variance (EVVE's Law)

$$\mathbf{Var}\left(Y\right) = \mathbb{E}\left(\mathbf{Var}\left(Y|X\right)\right) + \mathbf{Var}\left(\mathbb{E}\left(Y|X\right)\right)$$

2.8 Joint Moment Generating Function

Definition 2.8.1 Joint MGF
Let X, Y be random variables, then

$$\mathbf{M}\left(t_1, t_2\right) = \mathbb{E}\left(e^{t_1 X + t_2 Y}\right)$$

is called the joint MGF of X and Y if expectation exists for all $t_1 \in (-h_1, h_1), t_2 \in (-h_2, h_2)$ for some $h_1, h_2 > 0$.

R We can extend this idea to higher dimension, but for the purpose or review, we simply don't care.

Proposition 2.8.1 Properties of Joit MGF

1. $\mathbf{M}_X\left(t\right) = \mathbb{E}\left(e^{tX}\right) = \mathbf{M}\left(t, 0\right)$
2. If $X \perp\!\!\!\perp Y$, then $\mathbf{M}\left(t_1, t_2\right) = \mathbf{M}_X\left(t_1\right)\mathbf{M}_Y\left(t_2\right)$
3. The joint MGF completely determines the joint distribution
4. $$\frac{\partial^{m+n}}{\partial t_1^m \partial t_2^n}\mathbf{M}\left(t_1, t_2\right)\bigg|_{(0,0)} = \mathbb{E}\left(X^m Y^n\right)$$
5. If $X \perp\!\!\!\perp Y$ and $Z = X + Y$, then $\mathbf{M}_Z\left(t\right) = \mathbf{M}_X\left(t\right)\mathbf{M}_Y\left(t\right)$

2.9 Binomial and Multinomial Distribution

2.9.1 Binomial Distribution

Definition 2.9.1 Binomial Distribution $X \sim Bin(n, p)$
A binomial distribution models the number of successes in n iid trials. Let $X_1, \dots, X_n$ be Bernoulli, then $S_n = \sum_{i=1}^n X_i \sim Bin(n, p)$.

1. **pmf**
$$\mathbf{b}(x; n, p) = \mathbf{P}(S_n = x) = \binom{n}{x} p^x (1-p)^{n-x}$$
2. **Expectation:** $\mathbb{E}\left(S_n\right) = \mathbb{E}\left(X_1 + \cdots + X_n\right) = np$
3. **Variance:** $\mathbf{Var}\left(S_n\right) = np(1-p)$
4. **MGF:** $\mathbf{M}_{S_n}\left(t\right) = \left((1-p) + pe^t\right)^n$

R As we see, using pmf/pdf to derive other quantity expectation/MGF is not always the best way. Taking advantage of specific structure leads to much easier solutions.

2.9.2 Multinomial Distribution

Definition 2.9.2 Multinomial Distribution

$(X_1,\dots,X_k)$ is said to follow multinomial distribution with parameter n and $p_1,\dots,p_k$ if it has the joint pmf to be

$$\mathbf{f}(x_1,\dots,x_k) = \frac{n!}{x_1!\dots x_k!} p_1^{x_1}\dots p_k^{x_k}$$

where $x_i = 0,1,\dots,n$ for $i = 1,\dots,k$ and

$$\sum_{i=1}^{k} x_i = n$$

and $p_i > 0, i = 1,\dots,k$ with

$$\sum_{i=1}^{k} p_i = 1$$

Denote as $(X_1,\dots,X_k) \sim Multi(n, p_1,\dots,p_k)$.

Proposition 2.9.1 Properties of Multinomial Distribution

1. Any subset of $X_1,\dots,X_k$ follows a multinomial distribution. In particular, the marginal distributions follows

$$X_i \sim Bin(n, p_i)$$
$$X_i + X_j \sim Bin(n, p_i + p_j)$$

2. **Cov** $(X_i, X_j) = -np_ip_j$
3. Conditional distribution of any subset $X_1,\dots,X_k$ given some other constraints is multinomial. In particular,

$$X_i | X_j = x_j \sim Bin\left(n - x_j, \frac{p_i}{1-p_j}\right)$$

4. $X_i | X_i + X_j = t \sim Bin\left(t, \frac{p_i}{p_i+p_j}\right)$
5. **MGF:**

$$\mathbf{M}(t_1,\dots,t_k) = \left(\sum_{i=1}^{k} p_i e^{t_i}\right)^n$$

2.10 Bivariate Normal Distribution

Definition 2.10.1 Bivariate Normal Distribution

$X = (X_1, X_2)^T$ follows a bivariate normal distribution with mean vector $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and a covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

we say that $X \sim BVN(\mu, \Sigma)$.

1. **Joint pdf**

$$\mathbf{f}(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right), x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

2. **Joint MGF**

$$\mathbf{M}(t_1,t_2) = \exp\left(\mu^T t + \frac{1}{2}t^T \Sigma t\right), t = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

Proposition 2.10.1 Properties of BVN

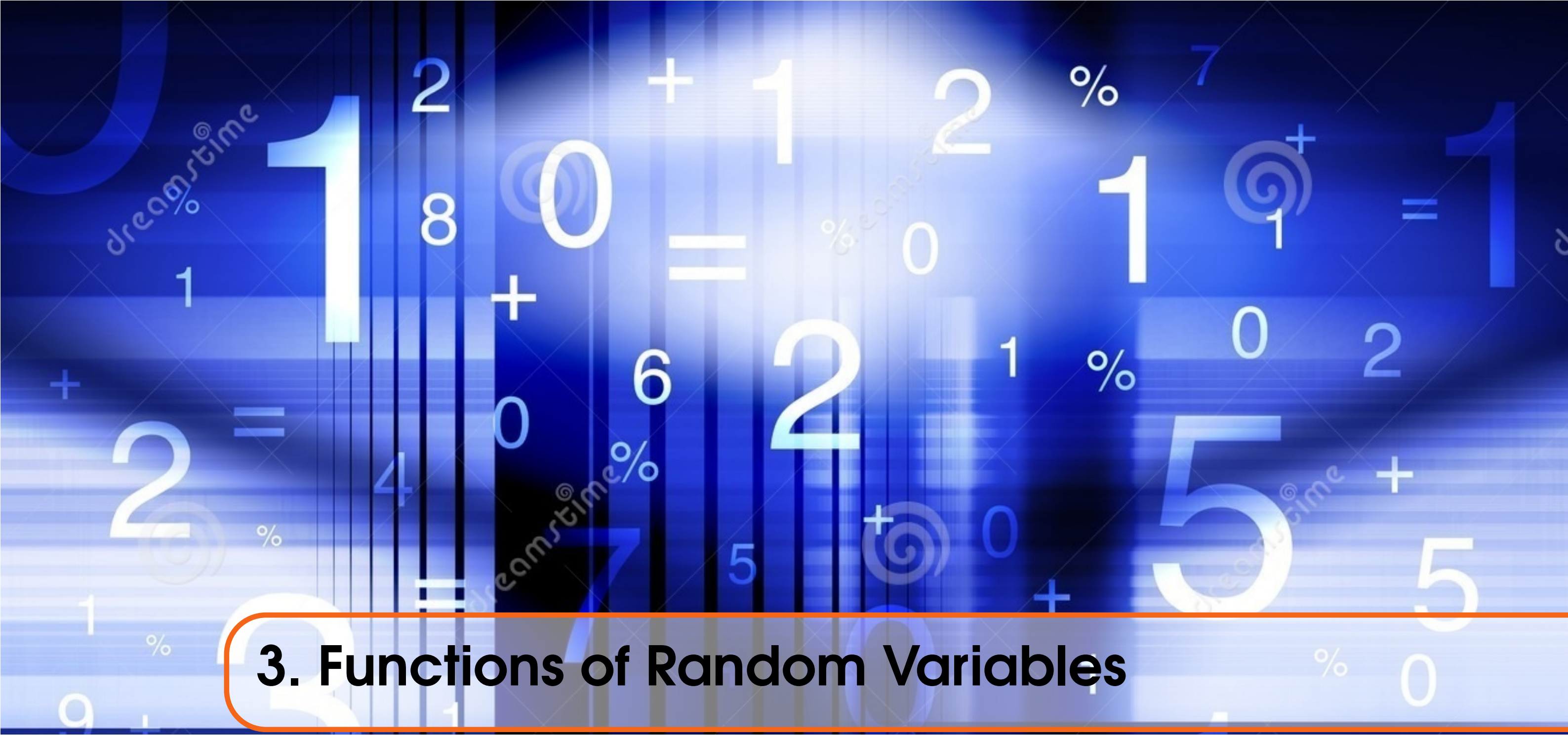
1. Any linear combination of X_1 and X_2 follows normal distribution. Say $C = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$, then

$$C^T X = c_1 X_1 + c_2 X_2 \sim N(C^T \mu, C^T \Sigma C)$$

2. BVN has normal marginal distributions

$$X_1 \sim N(\mu_1, \sigma_1^2) \qquad\qquad X_2 \sim N(\mu_2, \sigma_2^2)$$

3. $\mathbf{Cov}(X_1, X_2) = \rho \sigma_1 \sigma_2$
4. $X_1 \perp\!\!\!\perp X_2 \iff \rho = 0$
5. $Y = AX + B \sim BVN(A\mu + B, A\Sigma A^T)$ where A is a 2×2 non-singular matrix.
6. $X_2 | X_1 = x_1 \sim N\left(\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1-\rho^2)\right)$
7. $(X-\mu)^T \Sigma^{-1} (X-\mu) \sim \mathscr{X}^2(2)$



3. Functions of Random Variables

R Given some random variables, how to find the distribution of the functions of them?

3.1 Single Random Variable $h(X)$

For the following techniques, we assume h to be an injective function.

3.1.1 Method 1: CDF Technique

Theorem 3.1.1 CDF Technique

$$\begin{aligned}\mathbf{F}_Y(y) = \mathbf{P}(Y \le y) = \mathbf{P}(h(X) \le y) \\ \implies \mathbf{P}(X \le h^{-1}(y)) \\ \implies \int_{x \in h^{-1}((-\infty,y])} f_X(x)dx\end{aligned}$$

and $\mathbf{f}_Y(y) = \mathbf{F}'_Y(y)$.

3.1.2 Method 2: Change of Variable Formula

Theorem 3.1.2 Change of Variable Formula
Let X be a continuous random variable with pdf f, $Y = h(X)$ where h is injective and differentiable. Then, the pdf of Y is given by

$$g(y) = \mathbf{f}(h^{-1}(y)) \cdot \left| \frac{\partial}{\partial y} h^{-1}(y) \right|$$

3.1.3 Method 3: MGF Method

Theorem 3.1.3 MGF Method
Calculate the MGF of Y

$$\mathbf{M}_Y(t) = \mathbb{E}\left(e^{tY}\right) = \mathbb{E}\left(e^{th(X)}\right)$$

then compare $\mathbf{M}_y$ with the MGFs of the known distribution.

R **What if h is not injective?**
We can add the contributions from all the preimages.

$$g(y) = \sum_i \mathbf{f}(x_i)\frac{\partial}{\partial y}h^{-1}(y)$$

where x_is are the preimages of y and h^{-1} is the local inverse of h of x_i.

3.2 Two Random Variables $U = h_1(X,Y), V = h_2(X,Y)$

3.2.1 Method 1: CDF Method

Theorem 3.2.1 CDF Method
For any $u, v \in \mathbb{R}$, we can find

$$\mathbf{F}(u,v) = \mathbf{P}(U \le u, V \le v) = \mathbf{P}((X,Y) \in \vec{h}^{-1}((-\infty,u],(-\infty,v]))$$

$$= \int\int_{(x,y)\in\vec{h}^{-1}((-\infty,u],(-\infty,v])} \mathbf{f}(x,y)dxdy$$

then take the cross partial derivatives to get joint pdf.

R The CDF method is generally much more difficult compared to the single random variable because the region of integration is usually irregular.

3.2.2 Method 2: Change of Variable Formula

Theorem 3.2.2 Change of Variable Formula
Let X,Y be two continuous random variables with joint pdf $\mathbf{f}(x,y)$ and let $U = h_1(X,Y), V = h_2(X,Y)$ be two injective transformation with inverses. $X = w_1(U,V), Y = w_2(U,V)$. If the Jacobian

$$J = \begin{vmatrix} \frac{\partial w_1}{\partial u} & \frac{\partial w_2}{\partial v} \\ \frac{\partial w_2}{\partial u} & \frac{\partial w_2}{\partial v} \end{vmatrix}$$

always exists over the range of the transformation, then the joint pdf of (U,V) is

$$g(u,v) = \mathbf{f}(w_1(u,v), w_2(u,v))|J|$$

3.2.3 Method 3: Joint MGF Method

Theorem 3.2.3 Joint MGF Method
Calculate the joint MGF of U,V

$$\mathbf{M}_{U,V}(t_1,t_2) = \mathbb{E}\left(e^{t_1U+t_2V}\right) = \mathbb{E}\left(e^{t_1h_1(X,Y)+t_2h_2(X,Y)}\right)$$

then compare $\mathbf{M}_{U,V}$ to the joint MGF of known distribution.

4. Limiting Distribution

4.1 Convergence in Distribution

Definition 4.1.1 Convergence in Distribution
Let $X_1,\dots,X_n$ be a sequence of random variables with CDF $\mathbf{F}_1,\dots\mathbf{F}_n$ and X be a random variable with CDF $\mathbf{F}$. We say $\{X_n\}_{n=1}^{\infty}$ converges in distribution to X denoted as

$$X_n \longrightarrow_d X$$

if $\text{lím}_{n\to\infty}\mathbf{F}_n(x) = \mathbf{F}(x)$ at all the points x at which $\mathbf{F}$ is continuous. $\mathbf{F}$ is called the limiting distribution/asymptotic distribution of $\{X_n\}$.

R Although we talk about $X_n \to_d X$, the convergence in distribution is really the distribution. It does not tell anything aobut the convergence of $X_n(\omega)$ to fixed ω. It is also called weak convergence.

Theorem 4.1.1 Convergence in Distribution of Discrete Case
Let $\{X_n\}$ and X be (non-negative) integer-valued random variable. Then,

$$X_n \to_d X \Longleftrightarrow \mathbf{P}(X_n = x) \to \mathbf{P}(X = x), x = 0,1,2,\dots$$

4.1.1 MGF and Convergence Theorem

Theorem 4.1.2 Continuity Theorem
Let $\{X_n\}$ be a sequence of random variables with MGF $\{\mathbf{M}_n(t)\}$ and X be a random variable with MGF $\mathbf{M}(t)$. If there exists $h > 0$ such that

$$\mathbf{M}_n(t) \longrightarrow \mathbf{M}(t), \forall t \in (-h,h)$$

then $X_n \to_d X$.

4.2 Convergence in Probability

Definition 4.2.1 Convergence in Probability
A sequence of random variables $\{X_n\}$ converges in probability to a random variable X if for any $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}(|X_n - X| \geq \varepsilon) = 0$$

denote it as $X_n \to_p X$.

Theorem 4.2.1 Convergence in Probability$\Longrightarrow$ Convergence in Distribution

$$X_n \to_p X \Longrightarrow X_n \to_d X$$

R In general, the converse is not true. But when the limit is a constant, the converse is indeed true.

Theorem 4.2.2 Converse of Previous Theorem With Constant Limit
If $X_n \to_d c$, where $c \in \mathbb{R}$, then $X_n \to_p c$.

Proposition 4.2.3 Properties of Convergences

1. If $X_n \to_p a$ and g is a function which is continuous at a, then $g(X_n) \to_p g(a)$.
2. If $X_n \to_p a, Y_n \to_p b$ and $g(x,y)$ is a function continuous at $(x,y) = (a,b)$, then

$$g(X_n, Y_n) \to_p g(a,b)$$

.

Theorem 4.2.4 Slutsky Theorem
If $X_n \to_d X$, $Y_n \to_p b$, $g(x,y)$ is continuous at (x,b) for all $x \in \mathscr{A}_X$, then

$$g(X_n, Y_n) \to_d g(X, b)$$

Corollary 4.2.5 In particulars, given $X_n \to_d X$, $Y_n \to_p b$, we have

1. $X_n + Y_n \to_d X + b$
2. $X_n Y_n \to_d bX$
3. $\frac{X_n}{Y_n} \to_d \frac{X}{b}$, $b \neq 0$.

4.3 Limiting Theorems

4.3.1 Weak Law of Large Numbers (WLLN)

Theorem 4.3.1 Weak Law of Large Numbers (WLLN)
Let $\{X_n\}$ be a sequence of independent random variables with $\mathbb{E}(X_i) = \mu$, $\mathbf{Var}(X_i) = \sigma^2 < \infty$, then

$$\frac{X_1 + \cdots + X_n}{n} = \overline{X}_n \to_p \mu$$

4.3.2 Central Limit Theorem (CLT)

Theorem 4.3.2 Central Limit Theorem (CLT)
Suppose $\{X_n\}$ a sequence of iid random variables with $\mathbb{E}(X_i) = \mu$, $\mathbf{Var}(X_i) = \sigma^2 < \infty, i = 1, 2, \ldots,$ then

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \rightarrow_d Z \sim N(0,1)$$

1. The CLT gives the normal distribution, a special case as the (rescaled) limit of the sum of iid random variables with any distribution, as long as the mean and variance exist.
2. The result of CLT can be generalize to since cases where $X_1, \ldots, X_n$ be are not identically distributed, but only have the same mean and variance.
3. WLLN tells us for iid samples the sample mean will converge to the expectation. CLT tells us how fast this convergence happens

   $$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$$

   Intuitively, the difference between the sample mean and the true mean decreases at the speed of $\sqrt{n}$ at the spread as $\frac{1}{\sqrt{n}}$.

4.3.3 Delta Method

Theorem 4.3.3 Delta Method
Suppose $n^b(X_n - a) \rightarrow_d X$ for $b > 0, a \in \mathbb{R}$. g is differentiable at a and $g'(a) \neq 0$, then

$$n^b\left(g(X_n) - g(a)\right) \longrightarrow_d g'(a)X$$

4.4 Sampling Distribution

Derive the distributions of certain functions of random samples (very important in statistics)

4.4.1 Sum of Normals

Theorem 4.4.1 Sum of Normal
If $X_1, \ldots, X_n$ are independent random samples follow $N(\mu_i, \sigma_i^2)$, then

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

4.4.2 $\mathscr{X}^2$ Distribution

Definition 4.4.1 Y is said to follow $\mathscr{X}^2$ distribution with degree of freedom of k. Then,

$$Y \sim Gam\left(2, \frac{k}{2}\right)$$

or equivalently, Y has pdf

$$\mathbf{f}(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, x > 0$$

we say $Y \sim \mathscr{X}^2(k)$. Recall that

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt, x > 0 \textbf{ and } \Gamma(n) = (n-1)!$$

1. **Expectation:** $\mathbb{E}(Y) = k$
2. **Variance:** $\textbf{Var}(Y) = 2k$
3. **MGF:**
$$\mathbf{M}(t) = (1-2t)^{-\frac{k}{2}}, t < \frac{1}{2}$$

Theorem 4.4.2 Sum of $\mathscr{X}^2$ Random Variables
Let $Y_1, \dots, Y_n$ be independent random variables follows $\mathscr{X}^2(k_i)$ for $1 \le i \le n$, then

$$V = \sum_{i=1}^n Y_i \sim \mathscr{X}^2\left(\sum_{i=1}^n k_i\right)$$

Theorem 4.4.3 $\mathscr{X}^2$ and Standard Normal
If $Z \sim N(0,1)$, then $Z^2 \sim \mathscr{X}^2(1)$.

Corollary 4.4.4 Let $X_1, \dots, X_n$ be iid $N(\mu, \sigma^2)$, then

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \mathscr{X}^2(n) \iff \frac{n(\overline{X}_n - \mu)^2}{\sigma^2} \sim \mathscr{X}^2(1) \iff \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim N(0,1)$$

Theorem 4.4.5 Construction of S^2 (Unbiased Estimator of σ^2)
Let $X_1, \dots, X_n$ be iid $N(\mu, \sigma^2)$, then

1. $\overline{X}$ and $\{X_i - \overline{X}\}_{i=1,\dots,n}$ are independent
2. $\overline{X}$ and
$$S^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$$
are independent.
3. $$\frac{(n-1)S^2}{\sigma^2} \sim \mathscr{X}^2(n-1)$$

4.4.3 Gamma Distribution

Definition 4.4.2 Gamma Distribution

1. **First Parametrization** $Gam(\theta, k)$ with pdf
$$\frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-\frac{x}{\theta}}, x > 0$$
2. **Second Parametrization** $Gam(\alpha, \beta)$ with pdf
$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, x > 0$$

R We can see that $k = \alpha$ is a shape parameter, $\theta = \frac{1}{\beta}$ is a scale parameter.

Proposition 4.4.6 MGF of Gamma Distribution
If $Y \sim Gam(\theta, k)$, then

$$\mathbf{M}(t) = (1-\theta t)^{-k}, t < \frac{1}{\theta}$$

R Gamma distribution is a large family of distributions, containing many often used distributions and special cases.

$$Gam\left(\frac{1}{\theta}, 1\right) = Exp(\theta)$$
$$Gam\left(2, \frac{k}{2}\right) = \mathscr{X}^2(k)$$

these two give us $\mathscr{X}^2(2) = Exp\left(\frac{1}{2}\right)$.

Theorem 4.4.7 Sum of Gamma
Let $X_1, \ldots, X_n$ be independent random variables following $Gam(\theta, k_i), i = 1, \ldots, n$ (same θ), then

$$\sum_{i=1}^{n} X_i \sim Gam(\theta, \sum_{i=1}^{n} k_i)$$

Theorem 4.4.8 Sum of Poisson
Let $X_1, \ldots, X_n$ be independent random variables following $Poi(\lambda_i), i = 1, \ldots, n$ then

$$\sum_{i=1}^{n} X_i \sim Poi(\sum_{i=1}^{n} \lambda_i)$$

Theorem 4.4.9 Sum of Binomial
Let $X_1, \ldots, X_n$ be independent random variables following $Bin(k_i, p), i = 1, \ldots, n$ then

$$\sum_{i=1}^{n} X_i \sim Bin(\sum_{i=1}^{n} k_i, p)$$

4.4.4 Student's t Distribution

Definition 4.4.3 Student' t Distribution
A random variable X is said to follow t-distribution with degree of freedom n, if it has the pdf

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

denote it as $X \sim t(n)$.

R Note that as $n \to \infty$, we have

$$\mathbf{f}(x) \longrightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \sim N(0,1)$$

Thus, the student t distribution will converge to $N(0,1)$ in distribution.

Theorem 4.4.10 **μ Estimator With Unknown Variance**
Let $X_1, \dots, X_n$ be iid random variables and $X_i \sim N(\mu, \sigma^2)$, then

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t(n-1)$$

where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

4.4.5 F Distribution

Definition 4.4.4 **F Distribution**
A random variable X follows a F distribution if it has the pdf

$$\mathbf{f}(x; d_1, d_2) = \frac{1}{\beta\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}\right)^{-\frac{d_1+d_2}{2}}, x > 0$$

where $\beta(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$. Denote it as $X \sim F(d_1, d_2)$.

Theorem 4.4.11 **Relation Between $\mathscr{X}^2$ and F**
Let $U_1 \sim \mathscr{X}^2(d_1), U_2 \sim \mathscr{X}^2(d_2)$ and $U_1 \perp\!\!\!\perp U_2$, then

$$\frac{U_1/d_1}{U_2/d_2} \sim F(d_1, d_2)$$

Theorem 4.4.12 **Variance Comparison**
Let $X_1, \dots, X_n$ be iid normal samples with $X_i \sim N(\mu_1, \sigma_1^2)$. And let $Y_1, \dots, Y_m$ be iid normal samples with $Y_j \sim N(\mu_2, \sigma_2^2)$. Then, for

$$S_1^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \qquad S_2^2 = \frac{\sum_{i=1}^{m}(Y_i - \overline{Y})^2}{m-1}$$

then

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1, m-1)$$

4.5 Order Statistics

Definition 4.5.1 **Order Statistics**
Let $X_1, \dots, X_n$ be random variables. Rearrange them by increasing order we have

$$\{X_{(1)}, \dots, X_{(n)}\}$$

to be the order statistics of $X_1, \dots, X_n$.

R Assuming $X_1, \dots, X_n$ to be iid with pdf $\mathbf{f}$ and CDF $\mathbf{F}$, then we can have the general form

$$\mathbf{f}_{X_{(k)}}(x) = \frac{n!}{(n-k)!(k-1)!}\mathbf{F}^{k-1}(x)\left(1 - \mathbf{F}(x)\right)^{n-k}\mathbf{f}(x)$$

In particular,

$$\mathbf{F}_{X_{(1)}}(x) = 1 - (1 - \mathbf{F}(x))^n \Longrightarrow \mathbf{f}_{X_{(1)}}(x) = n\mathbf{f}(x)(1 - \mathbf{F}(x))^{n-1}$$

$$\mathbf{F}_{X_{(n)}}(x) = \mathbf{F}^n(x) \Longrightarrow \mathbf{f}_{X_{(n)}}(x) = n\mathbf{f}(x)\mathbf{F}^{n-1}(x)$$

Theorem 4.5.1 Joint pdf of Order Statistics
Let $X_1, \ldots, X_n$ be a random sample with pdf $\mathbf{f}(x)$, then the joint pdf of $X_{(1)}, \ldots, X_{(n)}$ is

$$g(y_1, \ldots, y_n) = n!\mathbf{f}(y_1)\mathbf{f}(y_2)\ldots\mathbf{f}(y_n)\mathbf{1}_{y_1<y_2<\ldots<y_n}$$

5. Estimation

5.1 Basic Settings

R Let $X := X_1, \dots, X_n$ be iid random variables representing a random sample from the distribution with pmf/pdf $\mathbf{f}(x;\theta)$, where θ is/are **unkown parameter(s)**, we have $\theta \in \Omega$ in the **parameter space**. Note that θ can be a vector. For the joint distribution pmf/pdf

$$\prod_{i=1}^{n} \mathbf{f}(x_i;\theta)$$

we also have observed data denoted as $x_1, \dots, x_n$. Our goal is to use the observed data to find a way to estimate the true value of θ.

Definition 5.1.1 Statistic
A statistic is denoted as $T = T(X_1, \dots, X_n)$ is a function of random variables $X_1, \dots, X_n$ which does not depend on any unknown parameters.

Definition 5.1.2 Estimator of $\tau(\theta)$
A statistic $T = T(X_1, \dots, X_n)$ is used to estimate a function of the parameter θ, say $\tau(\theta)$ is called the estimator of $\tau(\theta)$. The observed value $t = T(x_1, \dots, x_n)$ is called an estimate of $\tau(\theta)$.

5.2 Maximum Likelihood Estimation (One Parameter)

Definition 5.2.1 Likelihood Function
Let $X_1, \dots, X_n$ be a random sample from a distribution with pmf/pdf $\mathbf{f}(x;\theta)$ and $x_1, \dots, x_n$ are observations. Then, the likelihood function is defined to be

$$L(\theta) = L(\theta; x_1, \dots, x_n) = \prod_{i=1}^{n} \mathbf{f}(x_i;\theta)$$

Definition 5.2.2 Log-likelihood Function
With the same setting as the previous definition, the log-likelihood function is defined to be

$$l(\theta) = l(\theta; x_1, \dots, x_n) = \log L(\theta) = \sum_{i=1}^{n} \log \mathbf{f}(x_i; \theta)$$

R Later, we will need to take derivative of the likelihood/log-likelihood function. It turns out it is much easier to take derivative of the log-likelihood function compared to likelihood function. Due to this reason, log-likelihood function is often used.

Definition 5.2.3 Maximum Likelihood Estimate
The value θ that maximizes the likelihood function $L(\theta)$ or equivalently maximizes the log-likelihood function $l(\theta)$ is called the maximum likelihood estimate (MLE). Note $\hat{\theta} = \hat{x}$ to be the maximum likelihood estimate and $\hat{\theta} = \hat{\theta}(X)$ to be the maximum likelihood estimator.

R As we see in class that the first derivative $l'(\theta)$ plays an important role in finding $\hat{\theta}$.

Definition 5.2.4 Score Fucntion
The function $S(\theta) = S(\theta; X) = \frac{\partial}{\partial \theta} \mathrm{l}(\theta)$ is called the score function.

Definition 5.2.5 Information Function
The information function is defined to be

$$I(\theta) = I(\theta; X) = -l''(\theta) = -\frac{\partial^2}{\partial \theta^2} \log L(\theta)$$

Definition 5.2.6 Fisher Information
The Fisher information is defined to be

$$J(\theta) = \mathbb{E}\left(I(\theta; X)\right) = \mathbb{E}\left(-\frac{\partial^2}{\partial \theta^2} l(\theta; X)\right)$$

R Note that if $X_1, \dots, X_n$ is an iid random sample from some $f(x; \theta)$. Then,

$$\begin{aligned} J(\theta) &= \mathbb{E}\left(-\frac{\partial^2}{\partial \theta^2} \log L(\theta)\right) \\ &= \mathbb{E}\left(-\frac{\partial^2}{\partial \theta^2} \sum_{i=1}^{n} \log f(X_i; \theta)\right) \\ &= \sum_{i=1}^{n} \mathbb{E}\left(-\frac{\partial^2}{\partial \theta^2} \log f(X_i; \theta)\right) \\ &= n\mathbb{E}\left(-\frac{\partial^2}{\partial \theta^2} \log f(X_1; \theta)\right) \\ &= nI(\theta) \end{aligned}$$

This is just the Fisher information for a single observation, true only for iid. Fisher information is additive and linear. It increases linearly with the sample size n.

Theorem 5.2.1 Simpler Way to Compute Fisher Information
Under certain regularity condition,

$$J(\theta) = n\mathbb{E}\left(\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right)$$

Demostración. We shall show the single random variable case.

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\log f(X;\theta) &= \frac{\partial}{\partial\theta}\frac{\frac{\partial}{\partial\theta}f(X;\theta)}{f(X;\theta)}\\ &= \frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)} - \frac{\left(\frac{\partial}{\partial\theta}f(X;\theta)\right)^2}{f^2(X;\theta)}\end{aligned}$$

then, note that

$$\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right) = \mathbb{E}\left(\frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)}\right) + \mathbb{E}\left(-\frac{\left(\frac{\partial}{\partial\theta}f(X;\theta)\right)^2}{f^2(X;\theta)}\right) \tag{5.2.1}$$

and we found that

$$\begin{aligned}\mathbb{E}\left(\frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)}\right) &= \int\frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)}f(X;\theta)dx\\ &= \int\frac{\partial^2}{\partial\theta^2}f(X;\theta)dx\\ &= \frac{\partial^2}{\partial\theta^2}\int f(X;\theta)dx\\ &= \frac{\partial^2}{\partial\theta^2}1 = 0\end{aligned}$$

Thus, we have that

$$\mathbb{E}\left(-\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right) = \mathbb{E}\left(\frac{\left(\frac{\partial}{\partial\theta}f(X;\theta)\right)^2}{f^2(X;\theta)}\right) = \mathbb{E}\left(\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right)$$

■

5.2.1 Invariance Principle of MLE

Theorem 5.2.2 Invariance Principle of MLE
Suppose $\tau = h(\theta)$ is a one-to-one function of θ and suppose $\hat{\theta}$ is the MLE of θ, then

$$\hat{\tau} = h(\hat{\theta}) \text{ is the MLE of } \tau$$

R $l(\theta;x) = \sum_{i=1}^{n}\log f(x_i;\theta) = \sum_{i=1}^{n}\log f(x_i;\tau = h(\theta)) = l(\tau = h(\theta);x)$

This implies that θ maximizes $l(\theta;x)$ if and only if $\tau = h(\theta)$ maximizes $l(\tau = h(\theta);x)$

5.3 Criteria for Evaluating Estimators

Definition 5.3.1 Unbiasness
The bias of an estimator T of the function $\tau(\theta)$ is defined as the difference between the $\mathbb{E}(()T)$ and the true value of $\tau(\theta)$.

$$\mathbb{E}_{(\theta)}(T) - \tau(\theta)$$

when the bias is 0 or say $\mathbb{E}_{(\theta)}(T) = \tau(\theta), \forall\theta \in \Omega$, we say that T **is an unbiased estimator of** θ.

Definition 5.3.2 Consistency
Let $\{T_n\}$ be a sequence of estimators of $\tau(\theta)$. They are said to be **consistent**, if $T_n \to_p \tau(\theta)$ as $n \to \infty$. Very often, T_n is the estimator corresponding to the sample size. For example,

$$\overline{X_n} = \frac{1}{n}(X_1 + \cdots + X_n)$$

for each $n \geq 1$, we have a different estimator, but the sequence converges to the mean in probability by **WLLN**.

■ **Example 5.1** Let $X_1, \ldots, X_n$ be a random sample iid with $X_i \sim N(\mu, \sigma^2)$ with unknown variance and we would like to estimate μ. We given a set of observations $x_1, \ldots, x_n$.

1. Find the MLE of μ

$$\begin{aligned} L(\mu) &= \prod_{i=1}^{n} f(x_i;\mu) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ l(\mu) &= \sum_{i=1}^{n}\left(-\log(\sqrt{2\pi}\sigma)) - \frac{(x_i-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

Set $\frac{\partial}{\partial\mu} l(\mu) = 0$. A lot of useless constants, garbage...we can get

$$\sum_{i=1}^{n} \frac{2(x_i-\mu)}{2\sigma^2} = 0 \Longrightarrow n\mu = \sum_{i=1}^{n} x_i \Longrightarrow \mu = \frac{1}{n}(x_1 + \cdots + x_n) = \overline{x}$$

this is just an estimate, we can easily get the estimator to be

$$\hat{\mu} = \overline{X}$$

2. Check the performance of this estimator.
   a) **Unbiasness:**

$$\begin{aligned} \mathbb{E}(\hat{\mu}) &= \mathbb{E}\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) \\ &= \frac{1}{n}(\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)) = \mu \textbf{ since iid} \end{aligned}$$

   Thus, $\hat{\mu}$ is unbiased.

   b) **Consistency:** By WLLN, we know that

$$\hat{\mu} = \overline{X} \to_p \mu$$

   actually $\hat{\mu}$ is a sequence of estimators for different sample sizes. Thus, $\hat{\mu}$ is also consistent.

Both unbiased and consistent!!! Good estimator!!!

■

5.3.1 More Concepts

Definition 5.3.3 Relative Likelihood Function
The relative likelihood function $R(\theta)$ is defined as

$$R(\theta) = R(\theta; x) = \frac{L(\theta)}{L(\hat{\theta})}, \theta \in \Omega$$

where $\hat{\theta}$ is the MLE. And $R(\theta) \in [0,1]$ always. It tells us how likely the data is to appear with this parameter θ compared to the case with $\hat{\theta}$.

Definition 5.3.4 Log Relative Likelihood Function
The log relative likelihood function is

$$r(\theta) = r(\theta; x) = \log R(\theta) = \log L(\theta) - \log L(\hat{\theta}) = l(\theta) - l(\hat{\theta}) < 0$$

5.4 Confidence Interval

Let $X_1, \ldots, X_n$ be a random iid sample and suppose L and U are two statistics.

$$L = l(X_1, \ldots, X_n), \qquad U = u(X_1, \ldots, X_n)$$

After obtaining observations $x_1, \ldots, x_n$. Thus, we have

$$l(x_1, \ldots, x_n), u(x_1, \ldots, x_n)$$

Definition 5.4.1 Confidence Interval
An interval $(l(x_1, \ldots, x_n), u(x_1, \ldots, x_n))$ is called a $100p\,\%$ **confidence interval** (CI) for θ if

$$P(l(X_1, \ldots, X_n) < \theta < u(X_1, \ldots, X_n)) = p$$

R

Interpretation of Confidence Interval It is very important to have a clear understanding of the confidence interval. A $95\,\%$ CI **should not be** interpreted as “the random true value of θ falls in the given interval with probability of $95\,\%$"; it **should be** interpreted as the “random interval contains the (fixed) true value of θ with probability $95\,\%$".

$$P(l(X_1, \ldots, X_n) < \theta < u(X_1, \ldots, X_n)) = p, \ \boldsymbol{\theta} \textbf{ is fixed}$$

We can say 95 out of 100 intervals contain θ. The CI is not unique: any random interval (l,u) with

$$P(l(X_1, \ldots, X_n) < \theta < u(X_1, \ldots, X_n)) = p$$

is a $100p\,\%$ confidence interval.
Typically, people can choose the interval to be **symmetric**:

$$(\hat{\theta} - \varepsilon, \hat{\theta} + \varepsilon)$$

Or to be one-sided:

$$[0,u), (l, \infty), \ldots$$

5.4.1 CI Construction

1. How to find a CI?

   Use **pivotal quantities!!!**

Definition 5.4.2 Pivotal Quantites
Let $Q = q(X_1, \dots, X_n; \theta)$ be a function of $X_1, \dots, X_n$ and θ. Then, Q is called a **pivotal quantity**, if its distribution does not depend on θ or any other unknown parameters.

Corollary 5.4.1 Once we have a pivotal quantity Q, we can take q_1, q_2 such that

$$P(q_1 < Q < q_2) = p$$

where $Q = q(X_1, \dots, X_n; \theta)$. If Q is monotone in θ (increasing or decreasing), then this can be rearranged as

$$P(l(X_1, \dots, X_n) < \theta < u(X_1, \dots, X_n)) = p$$

There we have it! $100p\%$ CI is (l, u).

■ **Example 5.2** Let $X_1, \dots, X_n$ be a random sample follows $N(\mu, \sigma^2)$ where σ^2 is known and we would like to estimate μ by constructing a $100(1-\alpha)\%$ confidence interval. Recall that

$$\hat{\mu} = \overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

This is the MLE of μ. We proceed to construct the confidence interval for μ. We use the following pivotal quantity.

$$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(\mu, \sigma^2) \textbf{ this is an exact pivotal quantity}$$

Demostración.

$$\begin{aligned}
\overline{X} &\sim N\left(\mu, \frac{\sigma^2}{n}\right) \\
\overline{X} - \mu &\sim N\left(0, \frac{\sigma^2}{n}\right) \\
\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} &\sim N(0, 1)
\end{aligned}$$

Thus, Z is a pivotal quantity when σ is known. ■

Let z_1, z_2 be such that

$$\Phi(z_1) = \frac{\alpha}{2}, \Phi(z_2) = 1 - \frac{\alpha}{2}$$

where Φ is the cdf of $N(0,1)$. Then,

$$P(z_1 < Z < z_2) = \Phi(z_2) - \Phi(z_1) = 1 - \alpha$$

Now,

$$\begin{aligned}
1-\alpha &= P(z_1 < Z < z_2) \\
&= P\left(z_1 < \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} < z_2\right) \\
&= P\left(\sigma z_1 < \sqrt{n}(\overline{X}-\mu) < \sigma z_2\right) \\
&= P\left(\frac{\sigma z_1}{\sqrt{n}} < \overline{X}-\mu < \frac{\sigma z_2}{\sqrt{n}}\right) \\
&= P\left(\overline{X}-\frac{\sigma z_2}{\sqrt{n}} < \mu < \overline{X}-\frac{\sigma z_1}{\sqrt{n}}\right) \\
&= P\left(\overline{X}-\frac{\sigma z_2}{\sqrt{n}} < \mu < \overline{X}+\frac{\sigma z_2}{\sqrt{n}}\right)
\end{aligned}$$

Note that $-z_1 = z_2 > 0$. Thus,

$$\left(\overline{X}-\frac{\sigma z_2}{\sqrt{n}}, \overline{X}+\frac{\sigma z_2}{\sqrt{n}}\right)$$

is a $100(1-\alpha)\,\%$ CI for μ, where $z_2 = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$.

1. For a $95\,\%$ CI, $\alpha = 0{,}05$,

$$z_2 = \Phi^{-1}\left(1-\frac{0{,}05}{2}\right) = \Phi^{-1}(0{,}975) = 1{,}96$$

Thus, a $95\,\%$ CI for μ is

$$\left(\overline{X}-1{,}96\frac{\sigma}{\sqrt{n}}, \overline{X}+1{,}96\frac{\sigma}{\sqrt{n}}\right)$$

■

5.5 Asymptotic Behaviour

Asymptotic distribution of MLE

Theorem 5.5.1 Let $\hat{\theta}_n = \hat{\theta}_n(X_1,\dots,X_n)$ be the MLE of θ. Under certain regularity conditions, we have the following results

1. $\hat{\theta}_n \to_p \theta_0$, where θ_0 is the true value of θ. (This is describing consistency)
2. $J(\theta_0)^{\frac{1}{2}}\left(\hat{\theta}_n-\theta_0\right) \to_d Z \sim N(0,1)$
3. $-2\log R(\theta_0;X) = 2\left(l(\hat{\theta}_n,X)-l(\theta_0,X)\right) \to_d W \sim \mathscr{X}^2(1)$

Corollary 5.5.2 The above theorem tells us:

1. The MLE $\hat{\theta}_n$ is consistent
2. Approximately

$$\hat{\theta}_n-\theta_0 \sim N\left(0,\frac{1}{J(\theta_0)}\right)$$

thus, **Var** $\left(\hat{\theta}_n\right) \cong J(\theta_0)^{-1}$ and $J(\theta_0)^{-1}$ is called the **asymptotic variance** of $\hat{\theta}_n$.

3. Will be used to build the approximate CIs and construct the **likelihood ratio test**.

In practice, we do not known θ_0, therefore $J(\theta_0)$ is also unknown. However, since $\hat{\theta}_n \to_p \theta_0$, if we assume J is continuous in θ, then we have

$$J(\hat{\theta}_n) \to_p J(\theta_0)$$

Originally, we have $J(\theta_0)^{\frac{1}{2}}\left(\hat{\theta}_n-\theta_0\right)\rightarrow_d Z\sim N(0,1)$, but we can do better as

$$J(\hat{\theta}_n)^{\frac{1}{2}}\left(\hat{\theta}_n-\theta_0\right)\rightarrow_d Z\sim N(0,1)$$

. This can be understood from the observation that

$$(J(\hat{\theta}_n)^{\frac{1}{2}}/J(\theta_0)^{\frac{1}{2}})J(\theta_0)^{\frac{1}{2}}\left(\hat{\theta}_n-\theta_0\right)=J(\hat{\theta}_n)^{\frac{1}{2}}\left(\hat{\theta}_n-\theta_0\right)$$

and apply the **Slutsky's Theorem**. Hence,

$$\mathbf{Var}\left(\hat{\theta}_n\right)\cong J\left(\hat{\theta}_n\right)^{-1}$$

as well. Moreover, we can also look at the information instead of the Fisher information,

$$\begin{aligned}
\frac{1}{n}I(\theta;x) &= \frac{1}{n}\left(-\frac{\partial^2}{\partial\theta^2}\left(\sum_{i=1}^{n}l(\theta;X_i)\right)\right) \tag{5.5.1}\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(-\frac{\partial^2}{\partial\theta^2}\left(l(\theta;X_i)\right)\right) \tag{5.5.2}\\
&\rightarrow_p \mathbb{E}\left(-\frac{\partial^2}{\partial\theta^2}\left(l(\theta;X_i)\right)\right)\text{ since they are iid and we use WLLN} \tag{5.5.3}\\
&= \frac{1}{n}J(\theta) \tag{5.5.4}
\end{aligned}$$

As a result, we also have

$$I(\hat{\theta}_n;X)^{\frac{1}{2}}\left(\hat{\theta}_n-\theta_0\right)\rightarrow_d Z\sim N(0,1) \tag{5.5.5}$$

In most cases, an exact pivotal quantity cannot be constructed for a finite sample size. As a result, we are not able to give an exact CI. However, when n is large, we can build approximate CIs using the **asymptotic pivotal quantities** that we discussed above.

■ **Example 5.3** $J(\hat{\theta}_n)^{\frac{1}{2}}\left(\hat{\theta}_n-\theta_0\right)\rightarrow_d Z\sim N(0,1)$
In order to find a $100p\%$ approximate CI, let $a>0$ such that

$$P(-a<Z<a)=p, Z\sim N(0,1)$$

what we are trying to do here is to construct a **symmetric confidence interval** using the symmetry of the standard normal. In general, we discuss both sides, but for symmetric confidence interval, it is quite convenient to just consider

$$P(Z<a)=P(Z\leq a)=\frac{1+p}{2}, Z\sim N(0,1)\Longleftrightarrow a=\Phi^{-1}\left(\frac{1+p}{2}\right)$$

This formula is good for us to check the table.
Then,

$$\begin{aligned}
p &= P(-a<Z<a) \tag{5.5.6}\\
&\cong P(-a<J(\hat{\theta}_n)^{\frac{1}{2}}\left(\hat{\theta}_n-\theta_0\right)<a) \tag{5.5.7}\\
&= P(-aJ(\hat{\theta}_n)^{\frac{1}{2}}<\hat{\theta}_n-\theta_0<aJ(\hat{\theta}_n)^{\frac{1}{2}}) \tag{5.5.8}\\
&= P(\hat{\theta}_n-aJ(\hat{\theta}_n)^{\frac{1}{2}}<\theta_0<\hat{\theta}_n+aJ(\hat{\theta}_n)^{\frac{1}{2}}) \tag{5.5.9}
\end{aligned}$$

Now, we know the **approximate** $100p\%$ confidence interval for θ is

$$(\hat{\theta}_n-aJ(\hat{\theta}_n)^{\frac{1}{2}},\hat{\theta}_n+aJ(\hat{\theta}_n)^{\frac{1}{2}})$$

■

Note that the precision increases as the sample size increases, in fact, it is of $\mathscr{O}(\sqrt{n})$, i.e, we need n^2 sample size to shrink the length by $\frac{1}{n}$. This is a consistent observation with our previous results on exact pivotal quantities.

■ **Example 5.4** Given a random sample $X_1, X_2, \ldots, X_n$, where $X_i \sim Bin(1, \theta)$ iid and θ is unknown. Recall from a previous result that we know the MLE of θ is $\hat{\theta} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. From this, we can find the approximate confidence interval since there is no simple pivotal quantity for θ, so we need to build an approximate confidence interval.

$$\begin{aligned}
I(\theta) &= -\frac{\partial^2}{\partial \theta^2} l(\theta) \tag{5.5.10}\\
&= -\frac{\partial^2}{\partial \theta^2}\left(\sum_{i=1}^{n} X_i \log\theta + \left(n - \sum_{i=1}^{n} X_i\right)\log(1-\theta)\right) \tag{5.5.11}\\
&\textbf{skipping a lot of steps here, but this is the log likelihood function} \tag{5.5.12}\\
&\frac{\partial^1}{\partial \theta^1}\left(-\frac{1}{\theta}\sum_{i=1}^{n} X_i \log\theta + \frac{1}{1-\theta}\left(n - \sum_{i=1}^{n} X_i\right)\right) \tag{5.5.13}\\
&= \frac{1}{\theta^2}\left(\sum_{i=1}^{n} X_i\right) + \frac{1}{(1-\theta)^2}\left(n - \sum_{i=1}^{n} X_i\right) \tag{5.5.14}
\end{aligned}$$

We can get the Fisher Information as follow: using the iid property and the linearity of expectation, we can easily get,

$$\begin{aligned}
J(\theta) &= \mathbb{E}(I(\theta)) \tag{5.5.15}\\
&= \frac{1}{\theta^2} n\theta + \frac{1}{(1-\theta)^2}(n - n\theta) \tag{5.5.16}\\
&= \frac{n}{\theta(1-\theta)} \tag{5.5.17}
\end{aligned}$$

Now, to get $J(\hat{\theta}_n)$, we plug in the MLE $\hat{\theta}$, we get a $100p\%$ confidence interval for θ:

$$(\hat{\theta}_n - aJ(\hat{\theta}_n)^{\frac{1}{2}}, \hat{\theta}_n + aJ(\hat{\theta}_n)^{\frac{1}{2}}) = \left(\hat{\theta}_n - a\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}), \hat{\theta}_n + a\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}})\right)$$

where $a = \Phi^{-1}\left(\frac{1+p}{2}\right)$.

■

BZ: note that this CI expression is the same one derived from CLT approximation result. Interesting! When are these results identical for all approximation method?

5.5.1 Likelihood Interval

Definition 5.5.1 The set of θ for which

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} \geq q$$

is called a $100q\%$ likelihood region for θ.

If it is actually an interval, then it is called the **100q % likelihood interval (LI)** for θ.

Since

$$-2\log R(\theta_0;X) \to_d W \sim \mathscr{X}^2(1)$$

■ **Example 5.5** we can find CI from LI. For example, we know that

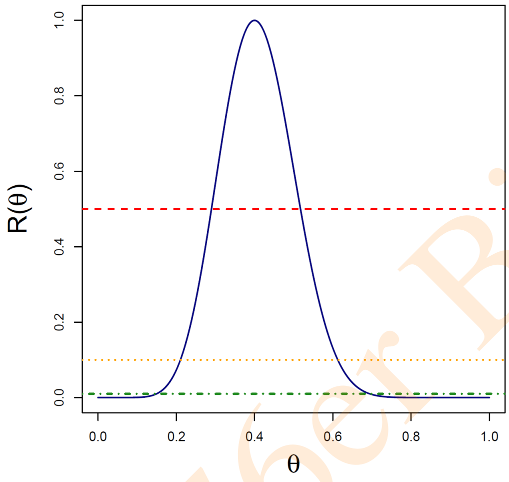
$$P(-1{,}96 < Z < 1{,}96) = 0{,}95, Z \sim N(0,1)$$

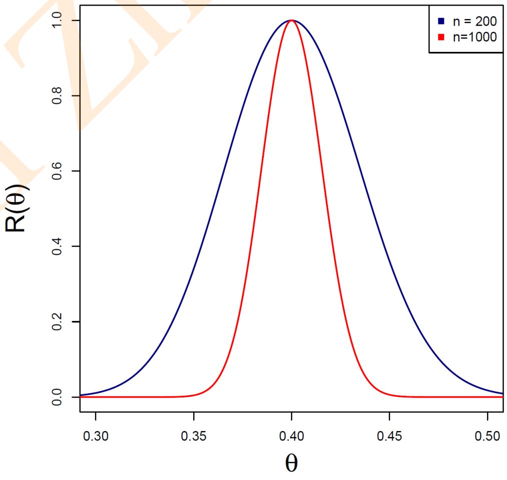
this corresponds to $Z^2 < 3{,}84$ and $Z^2 \sim \mathscr{X}^2(1)$. Then, we have that

$$\begin{aligned} P(W < 3{,}84) = 0{,}95 \qquad & (5.5.18)\\ \cong P(-2\log R(\theta_0;X) < 3{,}84) \qquad & (5.5.19)\\ \cong P(R(\theta_0;X) > 0{,}15) \qquad & (5.5.20)\\ = P(\theta_0 \in \{\theta : R(\theta : X) > 0{,}15\}) \qquad & (5.5.21) \end{aligned}$$

This proves that the 95 % Confidence Interval is approximately a 15 % Likelihood Interval. ■

Why do we usually get an interval? The reason is that, so far, we work with **unimodal** distributions, shown in the graphs below. Thus, the likelihood intervals displayed are connected intervals.





5.6 MLE in Multi-Parameter Case

Let's now consider the parameter vector as follow

$$\theta = (\theta_1,\ldots,\theta_k)^T$$

and we consider log-likelihood function

$$l(\theta_1,\ldots,\theta_k) = \log L(\theta_1,\ldots,\theta_k)$$

similarly, we have k-dimensional estimator

$$\hat{\theta} = (\hat{\theta}_1,\ldots,\hat{\theta}_k)$$

it is pretty much the same thing except for solving a larger system of equations.

$$\frac{\partial}{\partial\theta_j} l = 0, j = 1,\ldots,k$$

a system of equations.

Definition 5.6.1 Score Vector

$$S(\theta) = S(\theta;X) = \left(\frac{\partial}{\partial\theta_1} l, \dots, \frac{\partial}{\partial\theta_k} l\right)^T, \theta \in \Omega$$

Definition 5.6.2 Information Matrix

$$I(\theta) = I(\theta;X)$$

where

$$I_{ij}(\theta) = -\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\theta)$$

Definition 5.6.3 Fisher Information Matrix

$$J(\theta) = \mathbb{E}(I(\theta;X)) = \left\{\mathbb{E}\left(-\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\theta;X)\right)\right\}_{i,j=1,\dots,k}$$

Theorem 5.6.1 Invariance Property of MLE
The invariance property of MLE is still true

$$\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k) \text{ is the MLE of } \theta$$

$$h(\hat{\theta}_1, \dots, \hat{\theta}_k) \text{ is the MLE of } h(\theta_1, \dots, \theta_n)$$

5.6.1 Random Normal Sample

■ **Example 5.6** Let $X_1, \dots, X_n$ is a random sample from $N(\mu, \sigma^2)$. We want to estimate both μ and σ^2.

1. **Likelihood Function:**

$$L(\mu,\sigma^2) = \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \tag{5.6.1}$$

$$= (2\pi)^{-\frac{n}{2}}\left((\sigma^2)^{-\frac{n}{2}}\right) e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)} \tag{5.6.2}$$

2. **Loglikelihood Function:**

$$l(\mu,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu)^2\right) \tag{5.6.3}$$

$$\frac{\partial}{\partial\mu} l = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(-2x_i + 2\mu) \tag{5.6.4}$$

$$\frac{\partial}{\partial\sigma^2} l = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\left(\sum_{i=1}^{n}(x_i-\mu)^2\right) \tag{5.6.5}$$

3. Set $\frac{\partial}{\partial\mu} l$ and $\frac{\partial}{\partial\sigma^2} l$ to 0, we have

$$\frac{\partial}{\partial\mu} l = 0 \Longrightarrow \sum_{i=1}^{n}(x_i-\mu) = 0 \Longrightarrow \mu = \frac{1}{n}(x_1 + \cdots + x_n) = \bar{x} \tag{5.6.6}$$

$$\hat{\mu} = \overline{X} \tag{5.6.7}$$

$$\frac{\partial}{\partial\sigma^2} l = 0 \Longrightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{5.6.8}$$

Since we are solving this with $\frac{\partial}{\partial \mu} l = 0$, we know that $\mu = \bar{x}$ and we have

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Thus, $\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$.

4. **Let's check performance!**
   a) **Unbiasness:**

$$\mathbb{E}(\hat{\mu}) = \mathbb{E}\left(\overline{X}\right) = \mu \Longrightarrow \textbf{Unbiased}$$

For $\hat{\sigma^2}$, recall that

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

and we know that

$$\frac{(n-1)S^2}{\sigma^2} \sim \mathscr{X}^2(n-1)$$

and $\mathbb{E}\left(\frac{(n-1)S^2}{\sigma^2}\right) = n-1$ and $\textbf{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1)$. Thus, it is quite clear that

$$\mathbb{E}\left(S^2\right) = \sigma^2$$

and S^2 is an unbiased estimator of σ^2. As a result, the MLE $\hat{\sigma^2} = \frac{n-1}{n} S^2$ is biased. This actually shows that MLE is not necessarily unbiased. There is a trade off.

5. **Information Matrix**

$$-\frac{\partial^2 l}{\partial \mu^2} = \frac{n}{\sigma^2} \qquad -\frac{\partial^2 l}{\partial \mu \partial \sigma^2} = \frac{n(\bar{x} - \mu)}{\sigma^4} \qquad -\frac{\partial^2 l}{(\partial \sigma^2)^2} = -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\left(\sum_{i=1}^{n} (x_i - \mu)^2\right)$$

$$I(\mu, \sigma^2) = \begin{bmatrix} \frac{n}{\sigma^2} & \frac{n(\bar{x}-\mu)}{\sigma^4} \\ \frac{n(\bar{x}-\mu)}{\sigma^4} & -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\left(\sum_{i=1}^{n}(x_i - \mu)^2\right) \end{bmatrix}$$

Note that in practice, we cannot see $I(\mu, \sigma^2)$, instead, we have the **observed information** by replacing μ with $\hat{\mu}$ and σ^2 with $\hat{\sigma^2}$, then

$$I(\hat{\mu}, \hat{\sigma}^2) = \begin{bmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{bmatrix}$$

6. **Fisher Information**

$$J(\mu, \sigma^2) = \mathbb{E}\left(I(\mu, \sigma^2)\right)$$

$$J(\mu, \sigma^2) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix} \Longrightarrow J(\mu, \sigma^2)^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^2}{n} \end{bmatrix}$$

Note that

$$\textbf{Var}(\hat{\mu}) = \textbf{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$$

$$\textbf{Var}\left(\hat{\sigma}^2\right) = \textbf{Var}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{2(n-1)\sigma^4}{n^2} \cong_{n \to \infty} \frac{2\sigma^4}{n}$$

Since we have the observation that

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 \sim \frac{\sigma^2}{n} \mathscr{X}^2(n-1)$$

since $\frac{(n-1)S^2}{\sigma^2} \sim \mathscr{X}^2(n-1)$. From previous discussion, we know that

$$\mathrm{Cov}(\hat{\mu}, \hat{\sigma}^2) = 0, n \to \infty$$

since $\mathrm{Cov}(\overline{X}, S^2) = 0$ as $n \to \infty$. Hence, we verify that $J(\mu, \sigma^2)^{-1}$ the “asymptotic coveriance matrix".

7. **Confidence Region:** this can be derived similarly as in the single parameter case using the following **pivotal quantities:**

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1) \qquad \frac{(n-1)S^2}{\sigma^2} \sim \mathscr{X}^2(n-1)$$

■

R In general, the asymptotic behaviour in the parameter case is similar...

Theorem 5.6.2 Let $X_1, \ldots, X_n$ be a random sample from

$$f(x;\theta), \theta = (\theta_1, \ldots, \theta_n)^T, \hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$$

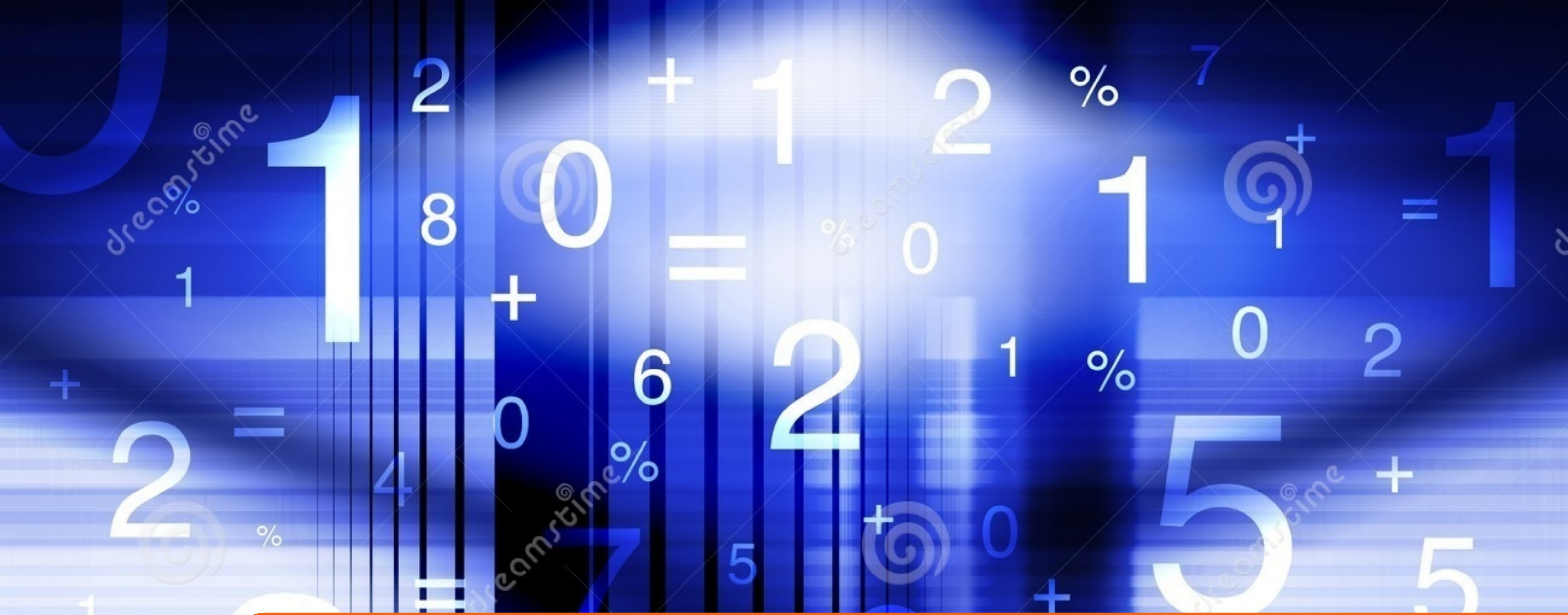
is the MLE. Then, under certain regularity conditions, we have the following:

1. $\hat{\theta}_n \to_p \theta_0$ convergence in components
2. $J(\theta_0)^{\frac{1}{2}} \left(\hat{\theta}_n - \theta_0\right) \to_d Z \sim MVN(\mu, I)$ where we have Z has the same distribution to $(Y_1, \ldots, Y_k)$ iid and $Y_i \sim N(0,1)$.
3. $-2\log(R(\theta_0;X)) = 2(l(\hat{\theta}_n;X) - l(\theta_0;X)) \to_d W \sim \mathscr{X}^2(k)$

Corollary 5.6.3 Under the same conditions as the theorem above, we have the following approximation result:

$$J(\theta_0)^{-1} \cong J(\hat{\theta}_n)^{-1} \cong I(\hat{\theta}_n)^{-1}$$

are asymptotic covariance matrices.



6. Hypothesis Testing

Definition 6.0.1 Statistical Test
A statistical test is a procedure to check the strength of the evidence (from the data) against an hypothesis, usually known as the **null hypothesis**, H_0. The opposite the null hypothesis is called the **alternative hypothesis**, H_1.
Very often, $H_0 : \theta \in \Omega_0$ for some $\Omega_0 \in \Omega$.

Definition 6.0.2 p-value
How (un)likely the observed data can happen if the null hypothesis is true.

We ue a “statistic". Then, the p-value is the probability of observing a value of the test statistic at least as extreme (e.g. as large) as the observed value if the null hypothesis is true.

1. **Small p-value:** the observed data is highly unlikely to happen under H_0, we reject the null hypothesis (probably false)
2. **Large p-value:** we fail to reject the null hypothesis (H_0 can still be false, but we cannot tell from the data using this particular statistic)

6.1 Likelihood Ratio Test For Simple Hypothesis

Definition 6.1.1 Simple Hypothesis vs. Composite Hypothesis
Let $X_1, \dots, X_n$ be a random sample from $f(x;\theta)$ where $\theta = (\theta_1, \dots, \theta_k)^T$.

1. **Simple Hypothesis:** is in the form of $H_0 : \theta = \theta_0$ as a specific value.
2. **Composite Hypothesis:** is te form of $H_0 : \theta \in \Omega_0 \subseteq \Omega$, which is an open set in $\mathbb{R}^q$ with $0 < q < k$

Definition 6.1.2 The Likelihood Ratio (L.R.) Statistic
The likelihood ratio statistic is called as

$$\Lambda = -2\log R(\theta_0;X) = 2(l(\hat{\theta};X) - l(\theta_0;X))$$

R Λ is the test statistic we are going to use for hypothesis testing. We have seen that if θ_0 is the true value of θ, then

$$\Lambda = -2\log R(\theta_0;X) \longrightarrow_d W \sim \mathscr{X}^2(k)$$

thus, the p-value is approximately (k is number of unknown parameters)

$$p \cong \mathbf{P}(W \geq -2\log R(\theta_0;x))$$

■ **Example 6.1 Normal Random Sample**
Let $X_1,\ldots,X_n$ be the random sample from $N(\mu,\sigma^2)$ with known σ^2. The null hypothesis is $H_0 : \mu = \mu_0$. We have seen that the MLE of μ is $\hat{\mu} = \overline{X}$.

$$\begin{aligned}
l(\theta;x) &= \log\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_i-\mu)^2}{s\sigma^2}\right)\right) \tag{6.1.1}\\
&= \sum_{i=1}^{n}\left(-\frac{1}{2}\log(2\pi\sigma^2) - (x_i-\mu)^2/2\sigma^2\right) \tag{6.1.2}\\
&= -\frac{n}{2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{6.1.3}\\
&\Longrightarrow \tag{6.1.4}\\
-2\log R(\theta_0;X) &= 2(l(\hat{\mu};X) - l(\mu;X)) \tag{6.1.5}\\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{\mu}-\mu_0)(2X_i-\mu_0-\hat{\mu}) \tag{6.1.6}\\
&= \frac{1}{\sigma^2}(\hat{\mu}-\mu_0)\sum_{i=1}^{n}(2X_i-\mu_0-\hat{\mu}) \tag{6.1.7}\\
&= \frac{\hat{\mu}-\mu_0}{\sigma^2}n(\hat{\mu}-\mu_0) = \frac{n}{\sigma^2}(\overline{X}-\mu_0)^2 \tag{6.1.8}
\end{aligned}$$

The p-value is

$$\mathbf{P}\left(W \geq \frac{n}{\sigma^2}(\overline{X}-\mu_0)^2\right), W \sim \mathscr{X}^2(1)$$

■

R From the above example, we have $\Lambda \sim \mathscr{X}^2(1)$ exactly since our sample is actual normal. However, in general (not-necessarily normal) case, what we get is the approximate p-value.

6.2 L.R. Test For Composite Hypothesis

Definition 6.2.1 Composite Hypothesis
Composite hypothesis is $H_0 : \theta \in \omega_0$, where Ω_0 is a q dimensional open set in Ω where Ω has k dimensions ($q < k$).

■ **Example 6.2** 1. Let $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$. Let $\Omega = \{(\mu,\sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$, the upper half-plane

$$H_0 : \mu = \mu_0 \ (\sigma^2 \text{ is unknown})$$

we have $k = 2$ and $q = 1$.

2. $(X_1,Y_1),\ldots,(X_n,Y_n)$ a random sample, $X_i \sim Exp(\theta_1)$ and $Y_1 \sim Exp(\theta_2)$, with $X_i \perp\!\!\!\perp Y_i, i = 1,2,\ldots,n$. Our null hypothesis is
$$H_0 : \theta_1 = \theta_2$$
the parameter space $\Omega = \{(\theta_1,\theta_2) : \theta_1 > 0, \theta_2 > 0\}$; and
$$\Omega_0 = \{(\theta_1,\theta_2) : \theta_1 = \theta_2 > 0\}$$
then, $k = 2$ and $q = 1$.

■

R The likelihood ratio statistic becomes

$$\Lambda(X) = -2\log\left(\frac{\text{máx}_{\theta\in\Omega_0} L(\theta;X)}{\text{máx}_{\theta\in\Omega} L(\theta;X)}\right) = 2\left(l(\hat{\theta};X) - \underset{\theta_0\in\Omega}{\text{máx}}\, l(\theta;X)\right)$$

In the simple hypothesis case, Ω_0 is the singleton $\{\theta_0\}$.

Theorem 6.2.1 Asymptotic Behaviour of $\Lambda(X)$
The aymptotic distribution of L.R. statistic is, if H_0 is true,

$$\Lambda(X) \longrightarrow_d W \sim \mathscr{X}^2(k-q)$$

where k is the total degree of freedom and q is the degree of freedom contained in H_0. Then,

$$k-q : \text{ is the remaining degree of freedom}$$

In other words, it is the number of free parameters which exist in Ω but not in Ω_0.

R The approximate p-value is

$$p \cong \mathbf{P}(W \geq \Lambda(x)), W \sim \mathscr{X}^2(k-q)$$

R In practice, we preset a “threshold", called “level", which is oftern $5\,\%$. We reject the null hypothesis H_0 if the p-value is smaller than this threshold (level).

END OF THE COURSE NOTE