# Higher Theory of Statistics
# Math2901 UNSW

Hussain Nawaz
hussain.nwz000@gmail.com

2022T2

## Contents

# 1 Introduction

## 1.1 Experiments, Sample Space and Events

**Experiments** An experiment is any process that leads to a recorded observation.

**Outcome and Sample Space** An outcome is possible result of the experiment. The set of all possible outcomes is called the sample space. The sample space is often denoted by $\Omega$.

Observe that not all sample spaces are countable. An uncountable example would be the set of all real number between 0 and 1.

**Events** An event is a set of outcomes that is, a subset of the sample space $\Omega$.

**Mutual Exclusion** Events $A, B$ are mutually exclusive (disjoint) if they have no outcomes in common. That is, $A \cap B = \emptyset$.

**Set Operation Revision** If you have trouble recalling the following laws, for associativity and distributivity, you may replace $\cap$ with $\times$ and $\cup$ with $+$.

TODO: Associative and Distributive Law

## 1.2 Sigma Algebra

The $\sigma$ algebra must be defined for rigorously working with probability. The formalization of this, is beyond the scope of this course.

The $\sigma$-algebra can be thought of as the family of all possible subsets or events in a sample space. Analogously, this may be conceptualised as the power-set of the sample space.

**Probability** The probability is a set function, often denoted by $\mathcal{P}$ that maps events from the $\sigma$-algebra to $[0, 1]$ and satisfies certain properties.

**Probability Space** The triplet $\Omega, \mathcal{A}, \mathbb{P}$ is the probability space where

- $\Omega$ is the sample space,
- $\mathcal{A}$ is the $\sigma$-algebra,
- $\mathbb{P}$ is the probability function.

**Properties of Probability** Given the probability/sample space $\Omega, \mathcal{A}, \mathbb{P}$, the probability function $\mathbb{P}$ must satisfy

- For all set $A \in \mathcal{A}$, $\mathbb{P}(A) \geq 0$
- $\mathbb{P}(\Omega) = 1$
- Countable additive. Suppose that the family of set $A_i$

**Theorem: Continuity from below** Given an increasing sequence of events $A_1 \subset A_2 \subset \cdots \subset A_n$ then,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} = \mathbb{P}(A_n)$$

**Theorem: Continuity from above** Given a decreasing sequence of events $A_1 \supset A_2 \supset \ldots A_n$ then,

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} = \mathbb{P}(A_n)$$

**More Probability Lemmas**

- $\mathbb{P}(\emptyset) = 0$,
- For any $A \in \mathcal{A}$, $\mathbb{P}(A) \leq 1$ and $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$,
- Suppose $A, B \in \mathcal{A}$ and $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

## 1.3 Conditional Probability and Independence

**Conditional Probability** The conditional probability that an event $A$ occurs given that the event $B$ has already occurred is denoted by

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

**Independence** The events $A$ and $B$ are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

**A lemma on independence** Given two events $A, B$, then

$$\mathbb{P}(A|B) = A \quad \text{if and only if} \quad \mathbb{P}(B|A) = B$$

**Pairwise Independence of Sequences** A countable sequence of events $A_{i\in\mathbb{N}}$ is pairwise independent if

$$\mathbb{P}(A_i 0 \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j) \quad \forall i \neq j.$$

**Independence of Sequences** A countable sequence of events $A_{i\in\mathbb{N}}$ is independent is independent if for any sub-collection $A_{i_1}, \ldots, A_{i_n}$ we have

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}) = \prod_{j=1}^{n} \mathbb{P}(A_{i_j}).$$

**Multiplicative Law** Given $A, B$ are events, then,

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

This is equivalent to the multiplication down a decision tree.

**Additive Law** Let $A, B$ be events. Then,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

This is analogous to the inclusion-exclusion principle from set theory.

**Law of Total Probability** Suppose that $(A_i)_{i=1,\dots,k}$ are mutually exclusive and exhaustive of $\Omega$. That is,

$$\bigcup_{i=1}^{k} A_i = \Omega.$$

Then for any event $B$, we have

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

## 1.4 Descriptive Statistics and R

**Sample Variance and Mean** Suppose that we are given observations $x$ such that $x = (x_1, x_2, \dots, x_n)$.

Then, the **sample mean** is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The **sample variance** is given by

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

# 2 Random Variables

## 2.1 Random Variables

**Definition: Random Variables** Suppose that we work on a probability space $(\Omega, \mathcal{A}, \mathbb{P}$ And the outcomes in $\Omega$ are denoted by $\omega$.

Then, a random variable (r.v) $X$ is a function from $\Omega$ to $\mathbb{R}$ such that $\forall x \in \mathbb{R}$, the set $A_x = \{\omega \in \Omega, X(\omega) \leq x\}$. That is, a random variable is a function that maps *Omega* to some space.

**Convention on Random Variables** Random variables are often denoted by capital letters while, the outcomes are denoted by the lower-case equivalent of the random variable.

**Cumulative Distributive** The cumulative distribution of a r.v $X$ is defined by

$$F_X(x) = \mathbb{P}(\{\omega : X(\omega \leq x)\}) = \mathbb{P}(X \leq x).$$

**Cumulative Distribution Theorems** Suppose that $F_X$ is cumulative distribution function of $X$. Then,

- It is bounded between zero and one and

$$\lim_{x\downarrow-\infty} = 0 \quad \text{and} \lim_{x\uparrow\infty} = 1.$$

- It is non-decreasing. That is, if $x \leq y$ then, $F_X(x) \leq F_X(y)$.
- For any $x \leq y$,

$$\mathbb{P}(x < X < y) = \mathbb{P}(X \leq y) - \mathbb{P}(X \leq x) = F_X(y) - F_X(x).$$

- It is right continuous. That is,

$$\lim x \uparrow \infty F_X\left(x + \frac{1}{n}\right) = F_X(x).$$

- it has finite left-hand limit and

$$\mathbb{P}(X < x) = \lim_{n\to\infty} F_x\left(x - \frac{1}{n}\right),$$

  denoted by $F_X(x-)$. It is useful to observe that,

$$\mathbb{P}(X = x) = F_X(X) - F_X(x-) \coloneqq F_X(x).$$

**Discrete Random Variables** A r.v. is said to be discrete if the image of $X$ consists of countable many values $x$ where $\mathbb{P}(X = x) > 0$. The probability function is $\Delta F_X(x) = \mathbb{P}(X = x)$ and satisfies

$$\sum_{\text{all}x} \mathbb{P}(X = x) = 1.$$

**Continuous Random Variables and Probability Density Functions** A r.v is continuous if the image of $X$ takes a continuum of values.

The probability density function of a r.v is a real-valued function $f_x$ on $\mathbb{R}with$ the property that

$$\mathbb{P}(X \in A) = \int_A f_x(y)dy,$$

for any *Borel* subset of $\mathbb{R}$

**Required Properties of a Density Function** Valid density functions $f : \mathbb{R} \to \mathbb{R}$ must satisfy the following properties:

- $f(x) \geq 0, \forall x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f(x)dx = 1.$

**Useful Properties of a Continuous Random Variable** For all continuous random variables $X$, with density $f_x$,

1. If $A = (-\infty, x]$ and creating a cumulative distribution function $F_x$ such that $F_X(x) = \mathbb{P}(X \in A) = \mathbb{P}(X \leq x)$ then,
$$F_X(x) = \int_{-infty}^{x} f_x(y)dy.$$
2. For all $a < b$,
$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) = \int_{(} a)^b f_X(x)dx.$$
3. By the fundamental theorem of calculus and property 1,
$$F'_X(x) = \frac{d}{dx}\int_{-\infty}^{x} f_x(y)dy = f_X(x).$$

## 2.2 Expectation and Variance

**Expectation** The expectation of a r.v $X$, denoted by $\mathbb{E}(X)$ may be computed depending on when $X$ is discrete or continuous.

**Expectation of Discrete Random Variables** If $X$ is a discrete random variable then,
$$\mathbb{E}(X) \coloneqq \sum_{\text{all } x} x\mathbb{P}(X = x) = \sum_{\text{all } x} x\Delta F_x(x).$$

**Expectation of continuous Random Variables** If $X$ is a continuous random variable then,
$$\mathbb{E}(X) \coloneqq \int_{-infty}^{infty} x f_X(x)dx$$

**Interpreting the Expectation** Often $\mathbb{E}(x)$ is called the *mean* of $X$. Observe that mean and average are not neccesarily the same. $\mathbb{E}(X)$ may be thought as the long-run average of the outcomes of $X$. That is, the average observation of $X$ converges to $\mathbb{E}(X)$.

Where our density function represents a physical model, $\mathbb{E}(X)$ is equivalent to the center of mass.

**Linearity of the Expectation** We note that the expectation is linear. That is, for all constants $a, b \in \mathbb{R}$,
$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$

**Variance** Let $X$ be a r.v and set $\mu = \mathbb{E}(x)$. Then,
$$\text{Var}(X) \coloneqq \mathbb{E}\left((X - \mu)^2\right).$$

The standard deviation is the square root of variance.

**Properties of Variance** Given a random variance $X$ then, for any constants $a, b \in \mathbb{R}$,

1. $\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.,
2. $\text{Var}(aX) = a^2\text{Var}(X)$,
3. $\text{Var}(X + b) = \text{Var}(X)$,
4. $\text{Var}(b) = 0$.

**Covariance** Recall that variance is $\mathbb{E}((X-\mu)^2)$. Suppose that $X, Y$ are random variables such that $\mathbb{E}(X) = \mu_X, \mathbb{E}(Y) = \mu_Y$. Then, the covariance is defined as

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)).$$

Observe that $\text{Cov}(X, X) = \text{Var}(X)$.

**Variance and Linearity** The variance is not linear. That is,

$$\text{Var}(X + Y) \neq \text{Var}(X) + \text{Var}(Y).$$

However,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y).$$

## 2.3 Moment Generating Functions

**Moments** A moment of the random variable is denoted by

$$\mathbb{E}[X^r], \quad r = 1, 2, \ldots$$

Moments measure mean, variance, skewness, and kurtosis, all ways of looking at the shape of the distribition.

Suppose that $f(x)$ is a probability density function. Then,

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

**Kurtosis** The kurtosis is the standards 4th moment. It measures how *fat* the tail is. A positive kurtosis implies a thinner tail than negative kurtosis.

**Moment Generating Function** A moment generating function (MGF) is denoted as

$$M_X(u) = \mathbb{E}(e^{uX}) = \int_{\text{all } x} e^{uX} f_X(x) dx.$$

We say that the MGF of $X$ exists if $M_X(u)$ is finite in some interval containing zero.

**Using Moment Generating Function to Find Moments** Suppose that the moment genrating funciton exists. Then,

$$\mathbb{E}(X^r) = \lim_{u\to 0} M_X^(r)(u) =: \lim_{x\to 0} \frac{d^r}{du} M_x(u).$$

**Equivalence of Moment Generation Functions** Let $X, Y$ be two random variables and suppose that $M_X(u) = M_Y(u)$ for all $u$ in some interval containing 0. Then,

$$F_X(x) = F_Y(x), \forall x \in \mathbb{R}.$$

That is, a moment generating function (when it exists), uniquely characterises a cumulative distribution function of a random variable.

**Existence of Moments and Moment Generating Functions** If the moment generating function exists then all moments can be computed. However, the converse is not neccesarily true. That is, if all the moments exist and are finite, this does not imply the moment generating function exists.

**Useful Inequalities**

**Markov Inequality - Chebychev's First Inequality** For all non-negative r.v $X$, for $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Equivalently,

$$\int_a^\infty f(x)dx \leq \int_{-\infty}^\infty xf(x)dx.$$

**Chebychev's Second Inequality** Suppose that $X$ is any r.v with $\mathbb{E}(X) = \mu, \text{Var}(x) = \sigma^2$ and $k > 0$. Then

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

**Convex and Concave Functions** In probability, we may want to know if a function is concave but, cannot use the usual method of the second derivative as the function is not neccesarily twice differentiable.

A function $h$ is convex if for any $\lambda \in [0, 1]$ and $x_1, x_2$ in the domain of $h$,

$$h(\lambda x_1 + (1 - \lambda)x_2) \leq (\geq)\lambda h(x_1) + (1 - \lambda)h(x_2).$$

**Jensen's Inequality** Let $X$ be a random variable. Suppose that $h$ is a convex function. Then

$$h(\mathbb{E}(X)) \leq \mathbb{E}(h(X)).$$

If $h$ is concave then,

$$h(\mathbb{E}(X)) \leq \mathbb{E}(h(X)).$$

**Applications of Jensen's Inequality** Using Jesnen's inequality, it can be shown that

$$\text{Arithmetic Mean} \geq \text{Geometric Mean} \geq \text{Harmonic Mean}.$$

# 3 Common Distributions

## 3.1 Common Discrete Distributions

**Bernoulli Distributions** A Bernoulli trial is an experiment with two outcomes; success and failure. A random variable $X$ is defined with

$$X = \begin{cases} 1 & \text{if success,} \\ 0 & \text{if failure.} \end{cases}$$

Let $p \in [0,1]$ be probability of success. Then, we denote $X \sim \text{Bernoulli}(p)$

1. $\mathbb{P}(X = 1) = p$,
2. $\mathbb{P}(X = 0) = 1 - p$,
3. $\mathbb{E}(X) = p$,
4. $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 1 \times p - p^2 = p(1-p)$.

**Binomial Distribution** When there are $n$ independent bernoulli trials with a success rate of $p$, and $X :=$ total number of succeses. Then, $X$ is a Binomial r.v with parameter $n$ and $p$ such taht we write $X \sim \text{Bin}(n,p)$.

Let $(Y_i)_{i=1,\ldots,n}$ be a sequence of independent bernoulli trials with success rate $p$.. Then

$$X := \sum_{i=1}^{n} Y_i \text{ is } \text{Bin}(n,p).$$

Expectation exists as

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} \mathbb{E}(Y_i) = np.$$

Alternatively, using combinatorics,

$$\mathbb{P}(X = k) = C_k^n p^k (1-p)^{n-k}.$$

**Poisson Distribution** A random variable $X$ follows the poisson distribution with parameter $\lambda$ if its probability function is

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

Observe that

$$\mathbb{E}(X) = \lambda.$$

**Use** The poisson distribution is used to model count data. That is, counting the number of times an event occurs within a time period. The parameter $\lambda$ epresents the average number of times the event occurs in the time period of interest.

**Hypergeometric Distribution** A random variable has hypergeometric distribution with parameter $N, m, n$ and is written as $X \sim \text{Hyp}(n, m, N)$ if

$$\mathbb{P}(X = k) = \frac{C_x^m \times C_{n-x}^{N-m}}{C_n^N}, \quad \text{where } x = 1, \ldots, n.$$

**Example of Hypergeometric** Given a box of $N$ balls, $m$ are red and $N - m$ are black. Draw $n$ balls at random and let $X$ be the number of red balls drawn. Then, $X \sim \text{Hyp}(n, m, N)$.

**Remark** The "I feel like skipping this... discrete problems are not very interesting" was stated while covering this. Interpretation of this is left as an exercise to the reader.

## 3.2 Continous Distributions

**Gaussian - Normal Random Variable** Given parameters $\mu, \sigma^2$ has a probability density function as

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

The expecation and variance are $\mu, \sigma^2$ respectively.

**Linear Transforms** Let $X$ be a r.v with a probability density function $f_X$. Let $y = a + bX$ then for $b > 0$ and $a \in \mathbb{R}$, then

$$f_Y(x) = \frac{1}{b} f_X\left(\frac{x-a}{b}\right).$$

**Linear Tranformation of Normally Distributed Random Variable** Suppose that $X \sum \mathcal{N}(\mu, \sigma^2)$ and $a \in \mathbb{R}$ and $b > 0$. Then, the random variable $Y := a + bX$ is normally distributed as

$$Y \sim \mathcal{N}(a + b\mu, b^2\sigma^2).$$

That is, normally distributed random variable are closed under linear transformation.

**Indicator Function** An indicator function of a set $A$ is defined by

$$I_A(X) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}.$$

Commonly, we may see indicator functions over intervals that follow the notation $I_{[a,b]}$.

The indicator unifies expectation and probability since, the probability is the expectation of the indicator function. Therefore, it may be written that

$$\mathbb{P}(X \in A) = \int_A f_X(x)dx = \int_{-\infty}^{\infty} I_A(x) f_X(x)dx = \mathbb{E}(I_A(X)).$$

# 4 Bivariate Distributions

**Definition: Bivariate Distribution** The joint density function of two continuous random variables $X, Y$ is given by a bivariate functiosn $f_{x,y}$ with the following properties

1. $f(x, y) \geq 0, \forall (x, y) \in \mathbb{R}$.
2. $$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 0.$$
3. For any (measurable) sets $A, B$,
$$\mathbb{P}(X \in A, Y \in B) = \int_A \int_B f_{X,Y}(x, y)dxdy.$$

**Min and Max Notation** We may write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

**Revision: Double Integral Theorems**

1. **Tonelli's Theorem:** Suppose $f : \mathbb{R}^2 \to \mathbb{R}_+$. Then,
$$\int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y)dxdy) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y)dydx).$$
2. **Fubini's Theorem:** Suppose that $f : \mathbb{R}^2 \to \mathbb{R}$. If
$$\int_{\mathbb{R}} \int_{\mathbb{R}} |f(x, y))| \, dxdy < \infty, \text{ or } \int_{\mathbb{R}} \int_{\mathbb{R}} |f(x, y))| \, dydx < \infty,$$
Then,
$$\int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y))dxdy < \infty, = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y))dydx < \infty, .$$

**Lemma: Expected Value of Bounded Borel Function** For any bounded Borel function $g : \mathbb{R} \to \mathbb{R}$ and random variables $X, Y$, if the following sums / integrals are finite then,

$$\mathbb{E}(g(X, Y)) = \begin{cases} \sum_{\forall x} \sum_{\forall y} g(x, y)\mathbb{P}(X = x, Y = y) & \text{discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)dxdy. \end{cases},$$

**Marginal Probability Density** This is the name of the individual density functions of a joint density function. That is, $f_X, f_Y$ are the marginal densities of $f_{X,Y}$.

The marginal densities are given as

$$f_X(x) = \int_{\mathbb{R}} f_{x,y} dy..$$

Similarly, for the discrete case,

$$\mathbb{P}(X = x) = \sum_{\forall y} \mathbb{P}(X = x, Y = y)..$$

**Independence** Two random discrete variables $X, Y$ are independent if

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y), \forall x, y.$$

Similarly, for the continuous case, with a join probability of $f_{X,Y}$,

$$f_{X,Y} = f_x(x) f_y(y), \forall x, y.$$

**Independent Product of Expectation** If $X, Y$ are independent for bounded functions $f, g : \mathbb{R} \to \mathbb{R}$ then

$$\mathbb{E}(g(X)f(Y)) = \mathbb{E}(g(X))\mathbb{E}(f(Y)).$$

**Conditional Probability** Conditional probability for functions extends the standard case of sets. That is, in the discrete case,

$$\mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}..$$

Similarly for the continuous case,

$$f_{x|y} \coloneqq \frac{f_{x,y}(x, y)}{f_y(y)}.$$

**Multivariate Gaussian** A random vector $X = (X_1, X_2)$ is said to be Gaussian with $\mu_X = (\mu x_1, \mu x_2)$ and covariance matrix $V$ if

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^d |V|}} \exp\left(-\frac{1}{2}(X - \mu_X)^T V^{-1}(X - \mu_X)\right).$$

That is

$$f_X(x).$$

$d = 2$ (dimensions), $V^{-1}$ is the matrix inverse of $V$ and $|V|$ is the determinant of $V$.

**Variance Matrix** The variance matrix is a symmetric matrix with entries

$$V_{ij} = \text{Cov}(X_i, X_j), \text{ where } i, j \in 1, \dots d.$$

If $X = (X_1, X_2)$ is multivariate Gaussian then $(X_i)$ for $i = 1, 2$ must be one-dimensional Gaussian but, the converse is not true.

**Conditional Variance and Expecation** Given any bound Borel function $g$, the conditional expecation of $g(x)$ given the set $\{Y = y\}$ is

$$\mathbb{E}(g(x)|Y = y) = \begin{cases} \sum_x g(x)\mathbb{P}(X = x|Y = y) & \text{discrete,} \\ \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)dx & \text{continuous.} \end{cases}.$$

Conditional variance follows such that when given $Y = y$,

$$\text{Var}(X|Y = y) = \mathbb{E}(X^2|Y = y) - (\mathbb{E}(X|Y = y))^2.$$

**Independent Conditional Variance and Expecation** Let $X, Y$ be independent random variables. Then,

$$\mathbb{E}(X|Y = y) = \mathbb{E}(X) \text{ and } \text{Var}(X|Y = y) = \text{Var}(X).$$

**Covariance** Let $X, Y$ be random variables. Then,

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X)))(Y - \mathbb{E}(Y)).$$

**Properties of Covariance** For random variables $X, Y$,

1. $\text{Cov}(X, X) = \text{Var}(X)$.
2. $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.
3. $\text{Cov}(X, Y) = 0$ if $X, Y$ are independent. However a zero covariance does not imply independence.
4. The covariance is bilinear such that $\forall a, bc \in \mathbb{R}$, and random variables $X, Y, Z$,

   $$\text{Cov}(aX + bY, Z) = a\text{Cov}(X, Y) + b\text{Cov}(Y, Z).$$

   The same principle holds for $\text{Cov}(X, aY + bZ)$.

**Correlation** The correlation may be thought of as a measure of linear independence. Note that though independence implies variables are uncorrelated, uncorrelated variables are not neccesarily independent.

The correlation can be thought of as a normalisation of covariance. That is,

$$\text{Corr}(X, Y) :== \frac{\text{Cov}(X, Y)}{\text{Var}(X)\text{Var}(Y)}.$$

**Properties of Correlation** Let $X, Y$ be two random variables.

1. $|\text{Corr}(X, Y)| \leq 1$
2. $\text{Corr}(X, Y) = 1$ if and only if there exists an $a, b \in \mathbb{R}$ where $b < 0$ such that $\mathbb{P}(Y = a + bX) = 1$.
3. $\text{Corr}(X, Y) = 1$ if and only if there exists an $a, b \in \mathbb{R}$ where $b > 0$ such that $\mathbb{P}(Y = a + bX) = 1$.

**Transformations - Continuous Random Variables** A real valued function $h$ over $A \subset \mathbb{R}$ is said to be monotone on $A$ if $h$ is strictly increasing or decreasing over $A$.

Note that $h$ is invertible.

**Calculating PDF of Monotone Transformations** Suppose that $X$ is a random variable with density $f_x$. If $h$ is monotone over $\{x : f_X(x) > 0\}$ then, the probability density of $Y \coloneqq h(X)$ is given by

$$f_Y(y) = f_X(x) \left| \frac{dy}{dx} \right| = f_X \circ h^{-1}(y) \left| \frac{dh^{-1}(y)}{dy} \right|.$$

## 4.1 Bivariate Transformations

Suppose that $X, Y$ are random variables over $\mathbb{R}$ with tranforms $U, V$ respectively. Then, the joint probability density is given by

$$f_{U,V} = f_{X,Y}(x, y) |\det(J)|.$$

Observe that $X = U^{-1}, V = Y^{-1}$. Recall from multivariate calculus that $J$ is the Jacobian Matrix such that

$$J = \begin{pmatrix} \frac{dx}{du} & \frac{dx}{dv} \\ \frac{dy}{du} & \frac{dy}{dv} \end{pmatrix}.$$

# 5 Sums of Variables

**Convolution Formula** Suppose $X, Y$ are independent, continuous r.v. with density $f_X$, $f_Y$. Let $Z = X + Y$. Then,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(Z - Y) f_Y(y) dy.$$

In the discrete case,

$$\mathbb{P}(X + Y = z) = \sum_y \mathbb{P}(X = z - y)\mathbb{P}(Y = y).$$

**Convolutions and Exponential / Gamma** The method for convolutions can be use to show that if $X \sim \Gamma(\alpha, 1)$ and, $Y \sim \Gamma(\beta, 1)$, where $X, Y$ are independent, then,

$$Z \coloneqq X + Y \sim \Gamma(\alpha + \beta, 1)..$$

Observe $0 < y < z < \infty$.

**Convolutions and Moment Generating Functions** Suppose that $X, Y$ are random variables with moment generating functions. Then,

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

Generally, if $\{X_i\}_{i=1}^n$ is an independent sequence of random variables then

$$M_{\sum_{i=1}^n x_i}(T) = \prod_{i=1}^n M_{X_i}(t).$$

**Useful Results about Sums** Let $(X_i)_{i=1,\dots,n}$ be an independent sequence of random variables and set $Y \coloneqq \sum_{i=1}^n X_i$. Then,

$$\begin{aligned}
X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) &\implies Y \sim \mathcal{N}\left(\sum \mu_i, \sum \sigma_i^2\right) \\
X_i \sim \exp(\lambda) &\implies Y \sim \Gamma(n, \lambda) \\
X_i \sim \Gamma(1, \lambda) &\implies Y \sim \Gamma(n, \lambda) \\
X_i \sim \Gamma(\alpha_i, \beta) &\implies Y \sim \Gamma\left(\sum \alpha_i, \beta\right) \\
X_i \sim \text{Poisson}(\lambda_i) &\implies Y \sim \text{Poisson}\left(\sum \lambda_i\right) \\
X_i \sim \text{Bernoulli}(p_i) &\implies Y \sim \text{Bin}(n, p) \\
X_i \sim \text{Bin}(n_i, p) &\implies Y \sim \text{Bin}\left(\sum n_i, p\right)..
\end{aligned}$$

# 6 Central Limit Theorem and Convergence of Random Variables

## 6.1 Central Limit Theorem

**Definition: Central Limit Theorem** Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of random variables with common mean $\mu = \mathbb{E}(X_1)$ and variance $\sigma^2 = \text{Var}(X_1) < \infty$. Let $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ then,

$$\frac{\overline{x}_n - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0, 1).$$

**Normal Approximation to the Binomial Distribution** This is really just an extension of the central limit theorem but for binomial distributions. Suppose $X \sim \text{Bin}(n, p)$. Then,

$$\frac{X - np}{\sqrt{np(1-p)}} \xrightarrow{d} Z \sim \mathcal{N}(0, 1).$$

## 6.2 Convergence of Random Variables

**Convergence of Random Variables** Let $(X_i)_{i\in\mathbb{N}}$ be a random sequence of random variables. $X_n$ converges to $X$ in terms of distribution if, for all $x$ where $F_X(x)$ is continuous,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

This is denoted as $X_N \xrightarrow{d} X$.

**Convergence of Moment Generating Functions and Existence of CDF** Suppose $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables with moment generating function $M_{X_N}(t)$. Suppose

$$M(t) = \lim_{n\to\infty} M_{X_n}(t)$$

exists. Then, there exists a unique valid cumulative distribution function $F$ and random variable $X$ such that $F_X = F$.

The moments are uniquely determined by $M$ and for all points of continuity,

$$\lim_{n\to\infty} F_{X_n}(X)S = F(X) + F_X(x).$$

**Convergence: Epsilon Definition for Probability** Consider a sequence of variable $(X_n)_{n=1...,}$ converges to a random variable $X$ if for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

**Law of Large Numbers** (Strong Version) Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables with mean $\mu$, and finite variable $\sigma^A$. Set $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\overline{X}_n \xrightarrow{a.s} \mu.$$

(Weak Version): Same Thing but using *almost surely* probability for convergence.

**Almost Surely Convergence** Two random variables converge *almost surely* if for random variables $X, Y$ $\mathbb{P}(Y = X) = 1$. The same can be said for sequences, written as $X \xrightarrow{a.s} X$.

**Slutsky's Theorem** Suppose that $(X_n)_{n\in\mathbb{N}}$ is a sequence of independent random variables that converge to $X$ in distribution and $(Y_i)_{i\in\mathbb{N}}$ is a sequence of independent random variables that converge in probability to a constant $C$.

**Delta Method** Recall that the central limit theorem informs us us that sequences of random variables converge to a normal distribution. The delta-method is concerned with the square of this random variable.

Suppose

$$\frac{X_n - \theta}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} Z \sim \mathcal{N}(0,1)$$

and $g$ is a differentiable function in the neighbourhood of $\theta$ and $g'(\theta) \neq 0$. Then,

$$\sqrt{n}\left(g(X_n) - g(\theta) \xrightarrow{d} \mathcal{N}(0, \theta^2[g'(\theta)]^2)\right).$$

Equivalently (and better),

$$\frac{g(X_n) - g(\theta)}{\sigma g'(\theta)/\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

You should be thinking of this as

$$\frac{X_n - \mathbb{E}(X_n)}{\sqrt{\text{Var}(X_n)}}.$$

# 7 Estimators

## 7.1 Data and Models

**Samples and Data** A random sample is a collection of (random) observations, namely $(X_1, \ldots, X_n)$. The sample data is $x_1, \ldots, x_n$. For the sake of laziness, this may be represented as $X[1,n]$ and $x[1,n]$ respectively. This is not standard notation.

**Parametric Model** A parametric model for a random sample $X[1,n]$ is a family of probability density functions $f(x:\theta)$ for $\theta in \Theta$ where $\Theta \in \mathbb{R}^d$ is called the parameter space.

Data being modelled by a parametric family $f(x;\theta) : \theta \in Theta$ can be written as

$$X[1,n] \sim \{f(x;\theta) : \theta \in \Theta\}, \text{ for } x \in \mathbb{R}.$$

## 7.2 Estimators

**Estimators** Suppose that $X[1,n] \sim \{f(x;\theta) : \theta \in \Theta\}$. As estimator of $\theta$ is denoted by $\left(\hat{\theta}\right)_n$. That is,

$$\hat{\theta}_n = \hat{\theta}_n(X[1,n]) = g(X[1,n]),$$

where $g : \mathbb{R}^n \to mathbbR$.

**Estimators of Mean of Normal Distribution** Suppose that $X[1,n]$ are independent samples from a normal distribution $\mathcal{N}(\mu, \sigma^2)$.

Then, there are a few methods of calculating the mean:

- Average: $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$
- Median: $\hat{\mu}_n = X_{n/2}$
- $\hat{\mu}_n = \frac{X_1 + X_n}{2}$.

By the law of large numbers, the first estimator approaches $\mu$ and is the best estimator.

**Bias** The bias is a way to assess how good an estimator is. Is is calculated as

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\theta) - \hat{\theta}.$$

If $\text{Bias}(\hat{\theta}) = 0$ then, $\theta$ is an unbiased estimator.

**Variance Estimators** From the definition of variance, we may infer that if $\hat{\mu}_n = \overline{X}_n$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$

As such $\text{Bias}(\hat{\sigma}^2 = \mathbb{E}(\hat{sigma}^2) - \sigma^2)$. Equivalently,

$$\text{Bias}(\hat{\sigma}^2) = \sigma^2\left(1 - \frac{1}{n}\right) - \sigma^2 = sigma^2\left(-\frac{1}{n}\right).$$

On average, $\hat{sigma}^2$ will underestimate the variance but, the becomes unbiased for large $n$.

**Student $t$-Distribution** A random variable $T$ has a $t$-distribution with a degree of freedom $v$ if it has a probability density function of

$$f_T(x) = \frac{\Gamma(\nu/2)}{\Gamma(\nu/2)\Gamma(1/2)}v^{-\frac{1}{2}}\left(1 + \frac{x^2}{\nu}\right)^{(1-v)/2}.$$

This looks like a normal distribution. In fact, as $\nu \to \infty$, $T_\nu \to Z \sim \mathcal{N}(0,1)$. The $t$-distribution has a fatter tail.

**Normal and $t$-Distribution** Let $Y, Z$ be independent random variables with $Y \sim \mathcal{N}(0,1)$ and $Z \sim \chi^2_\nu$.

Recall that $\chi^2_n$ is equivalent to the distribution $\Gamma(n, \frac{1}{2})$.

## 7.3 Errors

**Standard Error** The standard error of a estimator $\hat{\sigma}$ is

$$\text{Se}(\hat{\theta}) = \frac{1}{\sqrt{n}}\sqrt{\text{Var}(\hat{\theta})}.$$

Similarly, the estimated standard error is

$$\hat{\text{Se}}(\hat{\theta}) = \frac{1}{\sqrt{n}}\left[\sqrt{\text{Var}(\hat{\theta})}\right]_{\theta=\hat{\theta}}.$$

**Mean Square Error** Define Mean Square Error (MSE) of an estimator to be

$$\text{MSE}(\hat{\theta}) \coloneqq \mathbb{E}\left(\left(\hat{\theta} - \hat{\theta}\right)^2\right).$$

Equivalently,

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2.$$

This is estimated as

$$\hat{\text{MSE}}(\hat{\theta}) = \text{Bias}(\hat{\theta})^2 + \text{Se}(\hat{\theta}).$$

**Variance Bias Tradeoff** For a fixed MSE, there is a trade-off between variance and bias. In statistics, we care for reducing bias most of the time. However, in other applications such as machine learning, there is a need for low variance.

**Comparing Estimators with MSE** Suppose that $\hat{\theta_1}, \hat{\theta_2}$ are two estimators of $\theta$. Then, $\hat{\theta_1}$ is better than $\hat{\theta_2}$ at $\theta_0$ if

$$\text{MSE}_{\theta_0}(\hat{\theta_1}) < \text{MSE}(\hat{\theta_2}).$$

A estimator $\hat{\theta_1}$ is uniformly better than $\hat{\theta_2}$ if, $\hat{\theta_1}$ is better for all $\theta \in \Theta$.

**Asymptotic Properties of the Estimator** An estimator $\hat{\theta}$ is a consistent estimator of $\theta$ if as $n \to \infty$,

$$\hat{\theta} \xrightarrow{\mathbb{P}} \theta.$$

By the weak law of large numbers $\overline{X}$ is a consistent estimator of $\mu$. We say that the estimators are *asymptoticlly* unbiased.

**TODO:** MOVE DISTS FOR NORMAL TO OWN SECTION AND FINISH

**Distribution Related to Normal Distribution** Given a random sample $X[1, n]$ that follows $\mathcal{N}(\mu, \sigma^2)$ where all $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2$ with independent $X_i$.

**Consistency** Consistency of an estimator is its performance as the amount of data increases. For reasonable estimators, we expect that $\theta = \hat{\theta}$ improves for large $n$. We define an estimator $\hat{\theta}_n$ to be consistent for $\theta$ if

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} ..$$

In cases where checking convergence of probability is difficult, we may equivalently test that

$$\lim_{n \to \infty} \text{MSE}(\hat{\theta}_n) = 0.$$

**Asymptomatic Normality** The estimator is asymptotically normal if

$$\frac{\hat{\theta} - \theta}{\text{Se}(\hat{\theta})} \xrightarrow{d} \mathcal{N}(0, 1)..$$

We know from the central limit theorem that $\hat{\mu} = \overline{X}$ and, a sample proportion $\hat{p}$ are asymptotically normal.

## 7.4 Notation and Common Practice

**Good Notation For Standard Error and Observed Values** It is common to add the standard error in parenthesis after reporting an observed value. That is, if $x_i = 100$ with a standard error 10, then we may state

$$x_i = 100(10).$$

**Tau Transformation for Gamma Models** For a $\Gamma(\alpha, \beta)$ distribution, we may often want to use the transformation $\tau = \alpha\beta$. This corresponds to the mean of the distribution.

## 7.5 Confidence Intervals

The value of an estimator is not sufficient to let us know of the inherent variability of that estimator. Confidence intervals ameliorate this by proving a *range* of values.

Suppose that $X[1, n]$ is a random sample from a model with a know parameter $\theta$. Let

$$L = L(X[1, n]) \quad \text{and} \quad U = U(X[1, n])$$

be statistics (functions) for $X_i$'s such that

$$\mathbb{P}(L < \theta < U) \geq 1 - \alpha, \quad \text{for all } \theta > \Theta.$$

Then, $(L, U)$ is a $1 - \alpha$ or, $100(1 - \alpha)\%$ confidence interval.

**Confidence Interval For a Normal Sample** Recall that for $X[1, n] \sim \mathcal{N}(\mu, \sigma)$, we derive that

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

allow for exact confidence intervals for $\mu, \sigma$.

A $1 - \alpha$ confidence interval for $\mu$ is

$$\left(\overline{X} - T_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, X + t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right).$$

**Confidence For Comparing Random Normal Samples** Suppose that

$$X[1, n] \sim \mathcal{N}(\mu_X, \sigma_X^2), Y[1, n] \sim \mathcal{N}(\mu_Y, \sigma_Y^2).$$

Then, to compare their ostensible means, we take a confidence interval for $\mu_X - \mu_Y$ is

$$\hat{X} - \hat{Y} \pm t_{n_X = N_y - 2, 1-\frac{a}{2}} S_P \sqrt{\frac{1}{n_x} + \frac{1}{n_Y}}.$$

Here, we may call the following the *pooled sample variance*

$$S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1S_Y^2}{n_X + n_Y - 1}.$$

**Confidence Interval for Paired Normal Random Sample** Suppose that $(X, Y)$ is a paired normal random sample. We can construct the confidence interval for the mean difference $D = X - Y$.

# 8 Methods for Parameter Estimation and Inference

**Estimates vs Estimators** It is important to differentiate the following. An estimate of a parameter $\theta$ is a function $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$.

An estimator is the same function $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ ove observable random variables. That is, the estimator is itself a random variable with examinable properties while the estimate is an actual number. That is, the realised value of an estimator

**Notation** Denote $f_x$ as $f$ because writing that subscript is alot of effort.

## 8.1 Method of Moments

**Method of Moments Estimation** Let $x_1, \ldots, x_n$ be observations from the model $f$ where,

$$f(x; \theta_1, \ldots, \theta_k).$$

As there are $k$ parameters a systemof $k$ equations forms, that equates the moments of $f_x$ with their sample counterparts.

$$\begin{aligned}
\mathbb{E}(X^1) &= \frac{1}{n}\sum_i x_i \\
\mathbb{E}(X^2) &= \frac{1}{n}\sum_i x_i^2 \\
&\vdots \\
\mathbb{E}(X^k) &= \frac{1}{n}\sum_i x_i^k.
\end{aligned}$$

The *method of moments* estimates are the solutions of these equations in terms of $\theta_1, \ldots, \theta_k$.

**Consistency of Method of Moments Estimators** By the weak law of large numbers, we can deduce that $\hat{\theta}_j \xrightarrow{\mathbb{P}} \theta_j$. That is, the method of moments leads to consistent estimations. However, this is not optimal as we can do better in terms of standard error and, mean squared error.

## 8.2 Maximum Likelihood Estimator

**Likelihood Function** Let $x[1, n]$ be observations from a random variable with the pdf $f$ where

$$f(x) = f(x; \theta), \theta \in \Theta.$$

The *likelihood function* $\mathcal{L}$ of $\theta$ is

$$\mathcal{L}(\theta) = f(x_1; \theta) \cdots f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta), \quad \theta \in \Theta.$$

Similarly, we have the log-likelihood function of $\theta$,

$$\ell(\theta) = \ln \mathcal{L}(\theta) = \sum_i \ln f(x_i; \theta).$$

**Maximum Likelihood Estimate** Suppose there is $x[1, n]$ observations from the function $f$ where

$$f(x) = f(x; \theta), \quad \text{for } \theta \in \Theta.$$

The maximum likelihood estimate of $\theta$ is the choice

$$\hat{\theta} = \theta, \text{ maximising } \mathcal{L}(\theta) \text{ over } \theta \in \Theta.$$

**Equivalence of Log Likelihood Maximisation** The place where $\theta$ achieves its maximum over $\theta \in Theta$ is also where

$$\ell(\theta) = \ln\{\mathcal{L}(\theta)\} = \sum_i \ln\{f(x_i; \theta)\}$$

attains its maximum. Therefore, the maximum likelihood estimate of $\theta$ is equivalently,

$$\hat{\theta} = \theta \text{ that maximises } \ell(\theta) \text{ over } \Theta.$$

**Non-Smooth Functions and Indicators** Not all functions $f(x; \theta)$ are smooth. As such, we may use the indicator function for a logical condition $\mathcal{P}$ given by

$$\mathbb{I}(\mathcal{P}) = \begin{cases} 1 & = \text{ if } \mathcal{P} \text{ is true} 0 \\ = \text{ if } \mathcal{P} \text{ is false} & \end{cases}.$$

Then, we may try to maximise $\mathcal{L}(\theta) \cdot \mathbb{I}(\mathcal{P})$.

**Indicators and Intersection** For any logical conditions $\mathcal{P}, \mathcal{Q}$,

$$\mathbb{I}(\mathcal{P} \cap \mathcal{Q}) = \mathbb{I}(\mathcal{P})\mathbb{I}(\mathcal{Q}).$$

**Consistency of Maximul Likelihood Estimators** The maximum likelihood estimator $\hat{\theta}_n$ of $\theta$ is consisted, that is $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$, if the following conditions hold.

- The domain of $x$ does not depend on $\theta$. That is, all $f(x; \theta)$ are non-zero over the same set, irrespective of $\theta$.
- If $\theta \neq \vartheta$ then $f(x; \theta) \neq f(x; \vartheta)$. Equivalently, if $\theta = \vartheta$ then $f(x; \theta) = f(x; \vartheta)$.
- The MLE $\hat{\theta}$ is unique and lies in the interior of $\Theta$.

**Equivariance under Function Transformation** If $\hat{\theta}$ is a MLE of $\theta$ then for all $g$,

$$g(\hat{\theta}) \text{ is the MLE of } g(\theta).$$