Notes on Topics in Probability Theory and Statistics for Economics

Matthew Lee Chen*

University of Cambridge
mlc82 [at] cam [dot] ac [dot] uk

I made these notes in preparation for my Mathematics and Statistics exam in the second-year of my undergraduate degree in economics at Cambridge University.

These notes use the lecture notes, problem sets, textbooks (namely *Probability: An Introduction* by G. Grimmett & D. Welsh, *Statistical Inference* by G. Casella & R. Berger, and *Probability, Statistics and Econometrics* by O. Linton) and numerous online sources.

I take full responsibility for all errors and if I become aware, I will be more than happy to promptly make corrections and adjustments.

Contents

*These notes are based on lecture notes and problem sets by Oliver Linton. All errors are solely my own.

1 Probability Theory

1.1 Definitions in Set-Theoretic Probability

The basis of probability theory is an **experiment** which has a measurable outcome that is unknown ahead of time.

A **probability space** $(\mathcal{S}, \mathcal{A}, \mathcal{P})$ is a model of an uncertain process that consists of three distinct components that we consider below.

Definition 1. *The **sample space** $\mathcal{S}$ is the set of all possible outcomes.*

For instance, flipping a coin may yield $\mathcal{S} = \{H, T\}$, in which case the sample space is finite. Often, we have cases of **subjective probability** such as with the outcome of an election. The sample space is still finite but unlike the coin toss, we cannot repeat the experiment and we are very unlikely to agree on the outcome.

We may also have a sample space that is **countably infinite**. Take an experiment where we toss a coin until the first head appears. In theory, this could be an infinite number of tosses but we can represent every attempt as a natural number.

Alternatively, we can have a sample space that is **uncountably infinite** such as an experiment where we take the average height of a person drawn randomly from a group. The sample space is thus any real number over a reasonable interval, of which there are uncountably infinitely many.

Definition 2. *An **event** A is any subset of $\mathcal{S}$.*

Definition 3. *Let $\mathcal{A}$ be a non-empty class of subsets of $\mathcal{S}$. We say that $\mathcal{A}$ is an **algebra** on $\mathcal{S}$ if it satisfies*

- ***Complementation Axiom**: $A \in \mathcal{A} \implies A^c \in \mathcal{A}$*
- ***Additivity Axiom**: $A_1, A_2 \in \mathcal{A} \implies A_1 \cup A_2 \in \mathcal{A}$*

Using Definition 3, we may argue that $\mathcal{S} \in \mathcal{A}$ and $\varnothing \in \mathcal{A}$ i.e. the empty set is in $\mathcal{A}$. The argument proceeds by invoking the Complementation Axiom which suggests that if $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$ and by the Addivitity Axiom, $A \cup A^c \in \mathcal{A}$. Since $\mathcal{S} = A \cup A^c$, we have $\mathcal{S} \in \mathcal{A}$ which again by complementation implies $\mathcal{S}^c = \varnothing \in \mathcal{A}$.

Definition 4. *$\mathcal{A}$ is a σ-**algebra** on $\mathcal{S}$ if it is an algebra (i.e. complementation and additivity hold)* *and further the axiom of* ***countable additivity*** *holds i.e.* $A_n \in \mathcal{A}, \ \ \forall n \in \mathbb{N} \implies \bigcup\limits_{n=1}^{\infty} A_n \in \mathcal{A}$.

Definition 3 implies that a σ-algebra is an extension of an algebra which abides the stronger condition of countable additivity that extends the additivity axiom to infinite unions. Note that if countable additivity holds, then additivity must hold. Thus, every σ-algebra is also an algebra.

If we have finite sets, repeating the additivity axiom allows us to define an algebra on $\mathcal{S}$. However, we require the stronger axiom of countable additivity and defining σ-algebras when $\mathcal{S}$ is uncountable e.g. a real-valued interval.

We may have algebras which are not σ-algebras. To have this property, we require that additivity holds but not countable additivity.

To see an example, consider the set $\mathcal{S} = \{1, 2, 3, ...\}$ i.e. the set of the positive integers. Suppose we have $\mathcal{A}$ consisting of all sets which are either finite or have finite complement. Then, $\varnothing, \mathcal{S} \in \mathcal{A}$.

Let $A_i = \{2, 4, 6, ..., 2i\}$ such that A_i is the set of all even numbers up to the *ith* even number; for instance, $A_1 = \{2\}$, $A_2 = \{2, 4\}$ and $A_3 = \{2, 4, 6\}$. Clearly, complementation is satisfied since any A_i^c is the set of odd numbers up to the *ith*, which is finite so in $\mathcal{A}$. Additivity is satisfied since $\bigcup\limits_{i=1}^{n} A_i \in \mathcal{A}$ as the finite union operation generates a set of a finite number of even numbers. However, countable additivity is not satisfied since $\bigcup\limits_{i=1}^{\infty} A_i = \mathbb{Z}_{>0} \notin \mathcal{A}$.

In a set, there may be many σ-algebras, all of which obey standard results in set-theoretic mathematics summarised below.

Proposition 1. *Sets A, B and C are such that*

- ***Commutativity****.* $A \cup B = B \cup A$ *and* $A \cap B = B \cap A$
- ***Associativity****.* $A \cup (B \cup C) = (A \cup B) \cup C$ *and* $A \cap (B \cap C) = (A \cap B) \cap C$
- ***Distributive Law****.* $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ *and* $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- ***De Morgan's Laws****.* $(A \cup B)^c = A^c \cap B^c$ *and* $(A \cap B)^c = A^c \cup B^c$

Definition 5. *The* ***power set*** $\mathcal{P}(\mathcal{S})$ *is the set of all subsets of* $\mathcal{S}$ *and it is the largest* σ*-algebra on* $\mathcal{S}$*. The power set has cardinality* 2^n *when* $\mathcal{S}$ *contains* n *elements.*

Think of a set $\mathcal{S} = \{x, y\}$ consisting of two elements x and y. The power set $\mathcal{P}(\mathcal{S}) = \{\varnothing, \{x\}, \{y\}, \{x, y\}\}$ has cardinality $2^2 = 4$. Note that we call a set of only one element a

singleton, and that every set contains the empty set $\varnothing$.

Proposition 2. *The smallest σ-algebra on $\mathcal{S}$ is $\{\varnothing, \mathcal{S}\} \subset \mathcal{P}(\mathcal{S})$, sometimes called the* ***trivial σ-algebra****.*

It is possible to generate a σ-algebra from any collection of subsets by adding to the set the complements and unions of its elements.

For example, if we have $\mathcal{S} = \{1,2,3,4,5,6\}$ and $\mathcal{B} = \{\varnothing, \mathcal{S}, \{1,2,3\}\}$, then the σ-algebra $\sigma(\mathcal{B}) = \{\varnothing, \mathcal{S}, \{1,2,3\}, \{4,5,6\}\}$. The logic here is to consider the complements of each element of $\mathcal{B}$. Clearly, the complements of both $\varnothing$ and $\mathcal{S}$ are respectively one another. Their union is $\mathcal{S}$. Also, $\{1,2,3\} \cup \varnothing = \{1,2,3\}$ and $\{1,2,3\} \cup \mathcal{S} = \mathcal{S}$. The only other case is $\{1,2,3\}^c = \{4,5,6\}$ and so by exhaustion we have our $\sigma(\mathcal{B})$.

For a more complex example, consider A_1, A_2 as disjoint subsets of $\mathcal{S}$. The σ-algebra generated by A_1, A_2 is the union of A_1 and A_2 along with all their complements and unions.

This is often extremely tedious to find. For example, in general, we might have $\sigma(A_1, A_2) = \{A_1, A_2, A_1^c, A_2^c, A_1 \cup A_2, A_1 \cup A_2^c, A_1^c \cup A_2, A_1^c \cup A_2^c, (A_1 \cup A_2)^c, A_1 \cup (A_1 \cup A_2)^c, A_2 \cup (A_1 \cup A_2)^c, ...\}$.

However, if we have $A_3 = (A_1 \cup A_2)^c$, then $A_1^c = A_2 \cup A_3$, for instance. Hence, $\sigma(A_1, A_2) = \{A_1, A_2, A_3, A_2 \cup A_3, A_1 \cup A_3, A_1 \cup A_2, A_1 \cup A_2 \cup A_3\}$ which happens to equal $\{S, \varnothing\}$.

For countable sets, we typically take the σ-algebra to be the power set.

For uncountable sets, the power set is too big (namely the cardinality of the power set is greater than the cardinality of the set itself). Hence, for many applications such as with $\mathcal{S} = \mathbb{R}$, we look at the following concept.

Proposition 3. ***Closure of σ-algebras under countable intersections****. Suppose $A_i \in \mathcal{A}$. Then, $\bigcap\limits_{i=1}^{\infty} A_i \in \mathcal{A}$.*

Proof. $A_i \in \mathcal{A} \implies A_i^c \in \mathcal{A}$ by complementation. In turn, this implies $\bigcup\limits_{i=1}^{\infty} A_i^c \in \mathcal{A}$ by countable additivity and $(\bigcup\limits_{i=1}^{\infty} A_i^c)^c \in \mathcal{A}$ by complementation.

By De Morgan's Laws, $\bigcup\limits_{i=1}^{\infty} A_i^c = (\bigcap\limits_{i=1}^{\infty} A_i)^c$ and also $\left((\bigcup\limits_{i=1}^{\infty} A_i)^c\right)^c = \bigcap\limits_{i=1}^{\infty} A_i$.

Combining gives $\bigcap\limits_{i=1}^{\infty} A_i \in \mathcal{A}$. $\square$

Definition 6. *The **Borel σ-algebra** of open intervals $\mathcal{B} = \{(a,b) : a,b \in \mathbb{R}\}$, defined $\sigma(\mathcal{B})$, consists of all intervals and countable unions (and countable intersections) of open intervals and complements. We can equivalently define the Borel σ-algebra over the set of closed intervals.*

Observe that the Borel σ-algebra over open intervals is the same as the Borel σ-algebra over closed intervals. This follows from the fact that any open interval can be written as a countable union of closed sets i.e. $(a,b) = \bigcup\limits_{i=1}^{\infty} [a + \dfrac{1}{n}, b - \dfrac{1}{n}]$. Similarly any closed interval can be written as a countable union of open sets i.e. $[a,b] = \bigcup\limits_{i=1}^{\infty} (a + \dfrac{1}{n}, b - \dfrac{1}{n})$. We can do a very similar exercise but writing arbitrary intervals as countable intersections of open, closed and "half-open" intervals.

Finding a set that is not in the Borel σ-algebra is difficult, however the cardinality of the Borel σ-algebra is the same as that of $\mathbb{R}$ (since we simply take countable operations on unions of the real line to find it) so it is a strict subset of $\mathcal{P}(\mathbb{R})$, which has cardinality greater than $\mathbb{R}$.

Proposition 4. *If $\mathcal{A}_1$ and $\mathcal{A}_2$ are σ-algebras, then $\mathcal{A}_1 \cap \mathcal{A}_2$ is a σ-algebra.*

Proof. First, check complementation: $X \in \mathcal{A}_1 \cap \mathcal{A}_2 \implies X \in \mathcal{A}_1, X \in \mathcal{A}_2 \implies X^c \in \mathcal{A}_1, X^c \in \mathcal{A}_2 \implies X^c \in \mathcal{A}_1 \cap \mathcal{A}_2$.

Next, additivity: $X_1, X_2 \in \mathcal{A}_1 \cap \mathcal{A}_2 \implies X_1, X_2 \in \mathcal{A}_1, \mathcal{A}_2 \implies X_1 \cup X_2 \in \mathcal{A}_1, \mathcal{A}_2 \implies X_1 \cup X_2 \in \mathcal{A}_1 \cap \mathcal{A}_2$. □

Note that we may recursively apply Proposition 4 to give the result that any number of countably infinite intersections of σ-algebras also yields a σ-algebra.

However, this is not generally true for unions. We can see this from the analogue of the proof above. Observe that $X \in \mathcal{A}_1 \cup \mathcal{A}_2 \not\Rightarrow X \in \mathcal{A}_1, X \in \mathcal{A}_2$ and $X_1, X_2 \in \mathcal{A}_1 \cup \mathcal{A}_2 \not\Rightarrow X_1, X_2 \in \mathcal{A}_1; X_1, X_2 \in \mathcal{A}_2$.

Definition 7. *Given sample space $\mathcal{S}$ and σ-algebra $\mathcal{A}$, we define a **probability measure** as $\mathcal{P} : \mathcal{A} \to \mathbb{R}$ such that*

- $0 \leq \mathcal{P}(A) \leq 1$ *for all* $A \in \mathcal{A}$
- $\mathcal{P}(\mathcal{S}) = 1$
- *If* $A_i \cap A_j = \varnothing$ *for all* $i \neq j$, *then* $\mathcal{P}(\bigcup\limits_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

We may thus say that $\mathcal{P}(A)$ is the probability that event A occurred. Note that A is $\mathcal{A}$-**measurable**, meaning that it only makes sense to define $\mathcal{P}$ on $\mathcal{A}$.

Proposition 5. *For any sets $A, A_i, B \in \mathcal{A}$, the following properties hold*

1. $\mathcal{P}(\varnothing) = 0$
2. $\mathcal{P}(A) \leq 1$
3. $\mathcal{P}(A^c) = 1 - \mathcal{P}(A)$
4. ***Law of Total Probability***. $\mathcal{P}(A) = \mathcal{P}(A \cap B) + \mathcal{P}(A \cap B^c)$
5. $A \subset B \implies \mathcal{P}(A) \leq \mathcal{P}(B)$
6. $\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B)$
7. $\mathcal{P}(\bigcup\limits_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mathcal{P}(A_i)$

The proofs of (1)-(7) consist of manipulating sets to obtain disjoint sets and then applying the rules we have seen so far. Most of these can be verified with a simple Venn Diagram.

For instance, take (3): observe that $1 = \mathcal{P}(S) = \mathcal{P}(A \cup A^c) = \mathcal{P}(A) + \mathcal{P}(A^c)$ as A, A^c are mutually exclusive and exhaustive.

Slightly more involved is the proof of (6): it is easier to use a Venn Diagram in this case. This allows us to verify that $A \cup B = (A \cap B^c) \cup (A^c \cap B) \cup (A \cap B)$ and that $A = (A \cap B) \cup (A \cap B^c)$ (similarly $B = (A \cap B) \cup (A^c \cap B)$).

Since A and B are mutually exclusive, $\mathcal{P}(A \cup B) = \mathcal{P}(A \cap B^c) + \mathcal{P}(A^c \cap B) + \mathcal{P}(A \cap B)$, $\mathcal{P}(A) = \mathcal{P}(A \cap B) + \mathcal{P}(A \cap B^c)$ and $\mathcal{P}(B) = \mathcal{P}(A \cap B) + \mathcal{P}(A^c \cap B)$. Combining $\mathcal{P}(A) + \mathcal{P}(B)$ reveals the result.

1.2 Conditional Probability

The **conditional probability** of an event A given an event B is denoted $\mathcal{P}(A|B) = \dfrac{\mathcal{P}(A \cap B)}{\mathcal{P}(B)}$ and defined when $\mathcal{P}(B) > 0$.

It should follow easily that if A and B are mutually exclusive, $\mathcal{P}(A|B) = 0$. Also if $A \subset B$, then $\mathcal{P}(A|B) = \dfrac{\mathcal{P}(A)}{\mathcal{P}(B)} \geq \mathcal{P}(A)$ with strict inequality unless $\mathcal{P}(B) = 1$. Finally, if $B \subset A$, then $\mathcal{P}(A|B) = 1$.

Proposition 6. $\mathcal{P}(.|B) : \mathcal{A} \to \mathbb{R}$ *is a probability measure.*

Proof. We proceed by checking the three conditions for being a probability measure.

First, we check that $\mathcal{P}(.|B) \geq 0$ which is true since $\mathcal{P}(. \cap B) \geq 0$ and $\mathcal{P}(B) \geq 0$ by definition. By the law of total probability, $\mathcal{P}(.|B) = \dfrac{\mathcal{P}(. \cap B)}{\mathcal{P}(. \cap B) + \mathcal{P}(. \cap B^c)} \leq 1$ since $\mathcal{P}(. \cap B^c) \geq 0$ by definition.

The second property is trivial. Observe that $\mathcal{P}(\mathcal{S}|B) = 1$ because $\mathcal{P}(\mathcal{S} \cap B) = \mathcal{P}(B)$.

For the third property, let $A_1, ..., A_\infty \in \mathcal{A}$ be disjoint such that $A_i \cap A_j = \varnothing$ for all $i \neq j$. Thus, $\mathcal{P}(\bigcup\limits_{i=1}^{\infty} A_i|B) = \dfrac{\mathcal{P}((\bigcup\limits_{i=1}^{\infty} A_i) \cap B)}{\mathcal{P}(B)} = \dfrac{\mathcal{P}(\bigcup\limits_{i=1}^{\infty} A_i \cap B)}{\mathcal{P}(B)} = \sum\limits_{i=1}^{\infty} \dfrac{\mathcal{P}(A_i \cap B)}{\mathcal{P}(B)}$, completing the proof. □

Definition 8. *Suppose* $\mathcal{P}(A), \mathcal{P}(B) > 0$*. Then,* A *and* B *are independent events if* $\mathcal{P}(A \cap B) = \mathcal{P}(A)\mathcal{P}(B)$*. By definition of conditional probability, this is equivalent to* $\mathcal{P}(A|B) = \mathcal{P}(A)$ *and* $\mathcal{P}(B|A) = \mathcal{P}(B)$*).*

Note that independence is a symmetric relationship and so A being independent of B is the same as saying that B is independent of A. Since knowing A and B means that we know A^c and B^c, then if A is independent of B, then A is independent of B^c and A^c is independent of B, and A^c is independent of B^c.

However, independence is not transitive. That is, A being independent of B and B being independent of C does not imply that A is independent of C. Consider the following example of rolling a fair die giving $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ and events $A = \{1, 2, 3\}$, $B = \{1, 2, 3, 4\}$ and $C = \{1, 3, 5\}$. We have $\mathcal{P}(A \cap B) = \mathcal{P}(\{2, 4\}) = \dfrac{1}{3} = \mathcal{P}(A)\mathcal{P}(B) = \dfrac{1}{2} \times \dfrac{2}{3}$ and $\mathcal{P}(B \cap C) = \mathcal{P}(\{1, 3\}) = \dfrac{1}{3} = \mathcal{P}(B)\mathcal{P}(C) = \dfrac{2}{3} \times \dfrac{1}{2} = \dfrac{1}{3}$. However, we also have $\mathcal{P}(A \cap C) = \varnothing$ but $\mathcal{P}(A)\mathcal{P}(C) = \dfrac{1}{4}$.

Independence is really a special case. In general, we expect dependence i.e. observing event B has an effect on the conditional probability of A. We say that we have **positive dependence** if $\mathcal{P}(A|B) > \mathcal{P}(A)$ and **negative dependence** if $\mathcal{P}(A|B) < \mathcal{P}(A)$.

Proposition 7. ***Bayes' Rule****. For sets* A *and* B *with* $\mathcal{P}(A) > 0$*:*

$$\mathcal{P}(B|A) = \frac{\mathcal{P}(A|B)\mathcal{P}(B)}{\mathcal{P}(A)} = \frac{\mathcal{P}(A|B)\mathcal{P}(B)}{\mathcal{P}(A|B)\mathcal{P}(B) + \mathcal{P}(A|B^c)\mathcal{P}(B^c)}$$

Bayes' Rule is a simple application of the definition of conditional probability and the law of total probability. We call $\mathcal{P}(B)$ the **prior probability** and $\mathcal{P}(A|B)$ the **likelihood**, while $\mathcal{P}(B|A)$ is the **posterior probability** and $\mathcal{P}(A)$ is the **marginal probability**. These ideas capture the notion that before information is known, a prior probability is made from existing belief. Then, upon the observation of new information, the prior is adjusted to give a posterior that captures the post-information state.

Often, a different version of Bayes' Rule is stated where $\mathcal{P}(A)$ is treated as some normalising constant so that

$$\mathcal{P}(B|A) \propto \mathcal{P}(A|B)\mathcal{P}(B)$$

1.3 Functions over Random Variables

Definition 9. *$X : \mathcal{S} \to \mathbb{R}$ is a real-valued **random variable** on $(\mathcal{S}, \mathcal{A}, \mathcal{P})$ if and only if X is $\mathcal{A}$-measurable. A random variable is essentially an experiment with induced sample space $\mathcal{S}_X \subset \mathbb{R}$ and induced σ-algebra $\mathcal{A}_X$ with probability measure $\mathcal{P}_X : \mathcal{A}_X \to [0, 1]$.*

Multiple random variables may be defined on the same $\mathcal{S}$. Consider for instance $\mathcal{S} = \{(1,1), (1,2), (1,3), ..., (6,6)\}$ consisting of 36 outcomes of rolling two dice. We may define an induced sample space $\mathcal{S}_X = \{2, 3, ..., 11, 12\}$ being the sum of outcomes, or a different induced sample space $\mathcal{S}_Y = \{1, 2, ..., 36\}$ being the product of outcomes. Then, X and Y are two random variables defined on the same $\mathcal{S}$.

With our example of rolling two dice and X corresponding to the sum of outcomes, consider $A = \{(1,1)\}$ which occurs with probability $\mathcal{P}_X(\{2\}) = \dfrac{1}{36}$. We can generate σ-algebra $\mathcal{A}$ by taking complements and countable unions along with $\varnothing$ and $\mathcal{S}_X$, giving $\mathcal{A}_X = \{\{2\}, \{3, ..., 12\}, \varnothing, \mathcal{S}\}$.

Intuitively, we can consider **discrete** random variables which takes values in a set of countable cardinality. Alternatively, they can be **continuous** where the values taken are in a set of uncountable cardinality. In general, we may have a combination of discrete and continuous random variables.

Definition 10. *A function $F_X(x) = \mathcal{P}(X \leq x)$ is a **cumulative distribution function (cdf)** of a random variable X defined on $x \in \mathbb{R}$ if and only if*

1. $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$.

2. *F is non-decreasing i.e. $x' \geq x \iff F(x') \geq F(x)$.*

3. F is everywhere right-continuous i.e. $\forall x_0$, $\lim_{x \to x_0^+} F(x) = F(x_0)$.

The cdf effectively replaces probability measure $\mathcal{P}_X$ since we can construct an unknown $\mathcal{P}_X$ using just the information from a known cdf. For example, $P_X(a,b] = P(a < X \leq b) = F_X(b) - F_X(a)$.

For a continuous random variable, we have the cdf being continuous everywhere whereas for a discrete random variable, the cdf is a step function.

We say that random variables X and Y are **identically distributed** if $\mathcal{P}(X \in A) = \mathcal{P}_X(A) = \mathcal{P}(Y \in A) = \mathcal{P}_Y(A)$ for all $A \in \mathcal{A}$.

This implies that it is also true to say that $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$.

Definition 11. *Let X and Y be random variables. We say X **first-order stochastically dominates** Y ie. $X \succsim_{FSD} Y$ if and only if $F_X(x) \leq F_Y(x)$ for all $x \in \mathbb{R}$ with strict inequality for some x.*

In the case of discrete random variables, we can define a **probability mass function (pmf)** $f_X(x) = \mathcal{P}(X = x)$.

In the case of continuous random variables, the analogue is a **probability density function (pdf)** $f_X(x) = F'_X(x)$ meaning that $F_X(x) = \int\limits_{-\infty}^{x} f_X(t)dt$.

However, it is important to note that if X is continuous, $\mathcal{P}(X = x) = \mathcal{P}(x - \epsilon < X \leq x) = F_X(x) - F_X(x - \epsilon) \to 0$ as $\epsilon \to 0$ by the continuity of F_X at x.

Since the singleton set $\{x\}$ is such that the probability of realising x exactly is zero, we don't consider the difference between weak and strict inequalities.

Both pmfs and pdfs are such that $f_X(x) \geq 0$ for all x.

In the discrete case, we have $\sum\limits_{x} f_X(x) = 1$ and in the continuous case, $\int\limits_{-\infty}^{\infty} f_X(x)dx = 1$.

In the discrete case, $f_X(x) \leq 1$. However, this is not true for the continuous case. Pdfs can take on values greater than 1 unlike probability measures.

For instance, the uniform distribution on $[0, 0.5]$ has $f_X(x) = 2$ for $x \in [0, 0.5]$ and $f_X(x) = 0$ for x otherwise. We can verify this with cdf $F_X(x) = 2x$ for $x \in [0, 0.5]$ which differentiates to $f_X(x) = 2$ over the same interval.

Definition 12. *A* ***quantile function*** $Q_X : [0,1] \to \mathbb{R}$ *is an inverse of the cdf and* $Q_X(\alpha)$ *solves* $F_X(Q_X(\alpha)) = \alpha$.

In general, there may be no solution, a unique solution or many solutions to $F_X(Q_X(\alpha)) = \alpha$.

When F_X is continuous and strictly increasing at $Q_X(\alpha)$, then the quantile exists and is uniquely defined. This is because a unique F_X^{-1} exists and we can write the median, for instance, as $M = Q_X(1/2) = F_X^{-1}(1/2)$. Or alternatively, we can compute the interquartile range as $Q_X(3/4) - Q_X(1/4)$.

However, if we do not have continuous and strictly increasing F_X, we may have no unique quantile.

For instance, suppose a random variable X takes values 0, 1 and 2 with equal probability. Then, suppose we are trying to find the median M corresponding to $F_X(M) = 0.5$. No such M exists because obviously $\mathcal{P}(X \leq 0) = 1/3$ and $\mathcal{P}(X \leq 1) = 2/3$. Hence, we need to specify a quantile function that ensures the uniqueness of the value given.

The conventional quantile function used is $Q_X(\alpha) = \inf\{x : F_X(x) \geq \alpha\}$ for all $\alpha \in [0,1]$.

This would give a median of 1 as when $\alpha = 1/2$, we are looking for the greatest lower bound of the set of all x such that $\mathcal{P}(X \leq x) \geq 1/2$.

Another possible quantile function is $Q_X(\alpha) = \sup\{x : F_X(x) \leq \alpha\}$ for all $\alpha \in [0,1]$, which gives median 0.

Another is $Q_X(\alpha) = \dfrac{\inf\{x : F_X(x) \geq \alpha\} + \sup\{x : F_X(x) \leq \alpha\}}{2}$, which gives median 0.5.

1.4 Transformations of Random Variables

If X is a random variable with cdf $F_X(x)$, then any measurable real-valued function of X i.e. $Y = g(X)$ is also a random variable.

We are able to define $\mathcal{P}(g(X) \in B) = \mathcal{P}(Y \in B) = \mathcal{P}_Y(B)$ for any set B in the σ-algebra on Y.

We want to be able to express $F_Y(y) = \mathcal{P}(Y \leq y)$ in terms of F_X and g and likewise to find a neat expression for the density f_Y when X and Y are continuous.

As an example, consider $Y = a + bX$ with random variables X and Y and scalar constants a

and b. Then, we have $F_Y(y) = \mathcal{P}(a + bX \leq y)$ which can be rearranged into $\mathcal{P}(X \leq \dfrac{y-a}{b})$ or $F_X(\dfrac{y-a}{b})$ providing $b > 0$.

If $b < 0$, then we have $1 - F_X(\dfrac{y-a}{b})$. Differentiating by the chain rule with respect to y gives the pdf $f_Y(y) = \dfrac{1}{|b|} f_X(\dfrac{y-a}{b})$ for $b \neq 0$.

As a different example, suppose $Y = X^k$ for $X > 0$ with X and Y being random variables and k a non-zero constant. Then $F_Y(y) = \mathcal{P}(Y \leq y) = \mathcal{P}(X^k \leq y) = \mathcal{P}(X \leq y^{1/k}) = F_X(y^{1/k})$.

By differentiation, we can obtain the density $f_Y(y) = \dfrac{1}{k} y^{\frac{1-k}{k}} f_X(y^{\frac{1}{k}})$ if $y > 0$ and 0 otherwise.

Proposition 8. *Suppose density f_X and inverse mapping $g^{-1}(.)$ are well-defined and continuously differentiable. Then, we have*

$$F_Y(y) = \begin{cases} F_X(g^{-1}(y)), & \text{if } g \text{ is increasing} \\ 1 - F_X(g^{-1}(y)), & \text{if } g \text{ is decreasing} \end{cases}$$

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}$$

defined for all $y \in Y \backslash 0$.

Proposition 9. ***Probability Integral Transform***. *Let X have continuous cdf $Y = F_X(x)$. Then $Y \sim U[0,1]$. That is, $\mathcal{P}(Y \leq y) = y$ for any $y \in [0,1]$.*

Proof. $F_Y(y) = \mathcal{P}(Y \leq y) = \mathcal{P}(F_X(x) \leq y) = \mathcal{P}(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y$ □

Proposition 10. *If X has cdf F_X and U is uniform on support $[0,1]$, then $Y = F_X^{-1}(U)$ has the same distribution as X.*

As an example, consider the logistic distribution which has cdf $F_X(x) = \dfrac{1}{1+e^{-x}}$. Then, we can find the inverse CDF which is $Q_X(\alpha) = \dfrac{\alpha}{1-\alpha}$ and evaluating at U which is uniform on $[0,1]$ gives $Y = ln(\dfrac{U}{1-U})$ which has the logistic distribution i.e. the same distribution as X despite $X \neq Y$.

We can verify that Y has the logistic distribution. We have $F_Y(y) = \mathcal{P}(Y \leq y) = \mathcal{P}(ln(\frac{U}{1-U}) \leq y) = \mathcal{P}(U \leq \frac{1}{1+e^{-y}}) = \frac{1}{1+e^{-y}}$.

1.5 Moments and Cumulants

Definition 13. *For any (measurable) function g such that $\sum\limits_x |g(x)| f_X(x)dx < \infty$ or $\int\limits_{-\infty}^{\infty} |g(x)| f_X(x)dx < \infty$ in the discrete and continuous cases respectively, the* ***expectation*** *is defined as*

$$E(g(X)) = \sum_x g(x) f_X(x)$$

$$E(g(X)) = \int\limits_{-\infty}^{\infty} g(x) f_X(x)dx$$

For computing the expectation of a random variable X instead of a function, we have the special cases as follows (again assuming $\sum\limits_x |x| f_X(x)dx < \infty$ or $\int\limits_{-\infty}^{\infty} |x| f_X(x)dx < \infty$ in the discrete and continuous cases respectively)

$$E(X) = \sum_x x f_X(x)$$

$$E(X) = \int\limits_{-\infty}^{\infty} x f_X(x)dx$$

Sometimes, we have functions that fail to satisfy Definition 13. For instance, consider the density for the Cauchy distribution $f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}$. It satisfies the definition of a probability density since it integrates to 1 and is everywhere non-negative.

Since the density is symmetric around 0, it is tempting to argue that $E(X) = 0$. However, it is true that $\int\limits_{0}^{\infty} x f_X(x)dx = \infty$ and $\int\limits_{-\infty}^{0} x f_X(x)dx = -\infty$ so $E(X)$ is not well-defined as $E|X|= \infty$.

Proposition 11. *For measurable g_1 and g_2 and constants $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$, we have*

1. ***Linearity of expectation.*** $E(\alpha_1 g_1(X) + \alpha_2 g_2(X) + \alpha_3) = \alpha_1 E(g_1(X)) + \alpha_2 E(g_2(X)) + \alpha_3$.

2. ***Monotonicity.*** *$g_1(x) \geq g_2(x)$ for all $x \iff E(g_1(X)) \geq E(g_2(X))$.*

3. ***Existence of expectation with higher moments****. For all $j \leq k$, $E(|X|^k) < \infty \implies E(|X|^j) < \infty$.*

Proposition 12. ***Jensen's Inequality****. Suppose $g(x)$ is convex i.e. $g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$ for all $x, y \in \mathbb{R}$ and all $\lambda \in [0,1]$.*

Then, $E(g(X)) \geq g(E(X))$ with strict inequality if we have strict convexity and X takes more than 1 possible value.

We reverse the inequality in the event of concavity and have an analogous definition for strict concavity.

Proof. Let $L(x) = a + bx$ be the tangent line to convex $g(x)$ at $E(X)$. Then, $g(x) \geq a + bx$ for all x. This implies $E(g(x)) \geq a + bE(X) = g(E(X))$. The reverse case applies for concavity of $g(x)$. Inequalities become strict in the event of strict convexity/concavity. □

Proposition 13. *Suppose $E(X^2)$ is finite and exists. Then $E(X)$ is the unique minimiser of the mean-squared error $E(X-\theta)^2$ with respect to θ.*

Proof. $E(X-\theta)^2 = E(X^2) - 2\theta E(X) + \theta^2$ which is strictly convex. Thus, $\theta = E(X)$ is the optimal θ that minimises mean-squared error. □

We may have other moments such as the **variance** σ_X^2 defined as $Var(X) = E(X - E(X))^2 = E(X^2) - (E(X))^2 \geq 0$. The square root σ_X is called the **standard deviation**. The variance satisfies $Var(aX+b) = a^2 Var(X)$ and the standard deviation is such that $SD(aX+b) = |a|SD(X)$. An alternative measure of dispersion is the interquartile range given by the difference in quantile functions $Q_X(3/4) - Q_X(1/4)$. For a standard normal distribution, $IQR = 1.3$ so we standardise it by using $IQR/1.3$ to compare it to the standard deviation.

Often, we are interested in measuring dispersion but we face the problem of scale variance. That is, if we multiply a random variable X by a scalar, typical measures of dispersion such as the standard deviation will change even though the underlying distribution is the same.

If we are interested in a scale invariant measure of dispersion, we could use the **coefficient of variation** given by $CV(X) = \dfrac{SD(X)}{E(X)}$ for random variable $X > 0$. Note that $CV(\alpha X) = CV(X)$

since the standard deviation and expectation change by the same proportion under any $\alpha \in \mathbb{R}\backslash 0$, so the changes cancel one another.

Some higher moments are also of interest. For instance, **skewness** $\kappa_3(X) = \dfrac{E(X - E(X))^3}{(Var(X))^{3/2}}$ measures the asymmetry of a distribution of X.

If X is symmetric about $E(X)$ then it has $\kappa_3(X) = 0$, but the reverse argument does not hold.

It is often said that if mean exceeds median, the distribution is "right-skewed" or has positive skewness (i.e. the right tail of the distribution is longer) and if the median exceeds the mean, the distribution is "left-skewed" or has negative skewness (i.e. left tail of the distribution is longer).

Another higher moment is the **excess kurtosis** $\kappa_4(X) = \dfrac{E(X - E(X))^4}{(Var(X))^2} - 3$. The **kurtosis** which is the first part of the equation measures the thickness of the tails and the peakedness of the middle of the distribution relative to the normal distribution which has kurtosis 3. We often normalise kurtosis to excess kurtosis to standardise the normal excess kurtosis to 0.

Heavy tails correspond to $\kappa_4(X) > 0$ which is called the **leptokurtic** case. Thin tails (i.e. $\kappa_4(X) < 0$) correspond to the **platykurtic** case. The **mesokurtic** case is where $\kappa_4(X) = 0$.

1.6 Univariate Distributions

There exist some common univariate distributions with some key properties.

The most basic discrete distribution is a **Bernoulli distribution** which depends solely on parameter $p \in [0, 1]$.

$$X = \begin{cases} 1, & \text{with probability p} \\ 0, & \text{with probability 1-p} \end{cases}$$

We can deduce that $E(X) = 1(p) + 0(1 - p) = p$ and $Var(X) = E(X^2) - p^2 = p - p^2 = p(1 - p)$.

If we repeat Bernoulli trials n times with sample space $\{0, 1\}^n$, then the binomial random variable Y denotes the number of 1's in n trials which takes values $\{0, 1, ..., n\}$. We have a **Binomial distribution** which depends on parameters n and p.

$$\mathcal{P}(Y = x|n, p) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x \in \{0, 1, ..., n\}$$

We can derive $E(X) = E(X_1 + ... + X_n) = np$ and $Var(X) = E(X_1^2 + ... + X_n^2) - (np)^2 = np(1-p)$.

An important distribution which considers the number of trials until success is the **geometric distribution**. The pmf below gives the probability that trial x is the first success

$$\mathcal{P}(X = x) = (1-p)^{x-1}p, \quad x = 1, 2, ...$$

which has $E(X) = \sum\limits_{k=1}^{\infty} kp(1-p)^{k-1} = \frac{1}{p}$ and $Var(X) = \frac{1-p}{p^2}$.

If we are interested in modelling the occurrence of rare events, the **Poisson distribution** is a good approximation of the binomial distribution when n is large and p is small, allowing for $p(n) = \frac{\lambda}{n}$.

$$\mathcal{P}(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x}, \quad x = 0, 1, 2, ...$$

The Poisson distribution has $E(X) = Var(X) = \lambda$.

Consider a distribution that is **uniform** on $[a, b]$. Then, the pdf and cdf respectively are

$$f(x|a,b) = \begin{cases} \frac{1}{b-a}, & if x \in [a,b] \\ 0, & \text{otherwise} \end{cases}$$

$$F(x|a,b) = \int\limits_a^x f(z|a,b)dz = \frac{x-a}{b-a}$$

We can deduce that $E(X) = \frac{a+b}{2}$ and $Var(X) = \frac{(b-a)^2}{12}$. Note also that if $X \sim U[a,b]$, then $X - a \sim U[0, b-a]$ and further $\frac{X-a}{b-a} \sim U[0,1]$.

The most common continuous distribution is the **Normal distribution**. We say that $X \sim N(\mu, \sigma^2)$ whenever

$$f(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\frac{-(x-\mu)^2}{2\sigma^2}, \quad x \in (-\infty, \infty)$$

where $E(X) = \mu$ and $Var(X) = \sigma^2$. The normal distribution is symmetric about μ and is

unimodal meaning that its mean is equal to its median, and in turn equal to its mode.

We may standardise the normal distribution by observing that $\frac{X-\mu}{\sigma} = Z \sim N(0,1)$ where Z is called the **standard normal distribution** with density $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp -\frac{z^2}{2}, \quad z \in (-\infty, \infty)$ and cdf Φ which can be computed by integrating ϕ with respect to z.

The normal density has very thin tails, meaning that $E(|X|^n) < \infty$ for all n so higher moments are all defined. In fact, $E(\exp(tX))$ is called the **moment generating function** of X. For a normal X, we can compute the moment generating function by computing $E(\exp(tX)) = \int\limits_{-\infty}^{\infty} \exp(tx) f_X(x) dx$ where f is the normal pdf. This integrates to give $\exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right) < \infty$ for all t.

Importantly, the moment generating function uniquely generates a particular probability distribution and so we can use it as part of proofs involving a particular probability distribution.

For the standard normal distribution, most of the mass is found within 3 standard deviations of the mean so extreme values are unlikely.

1.7 Multivariate Random Variables

Definition 14. *A* ***multivariate random variable*** $X : \mathcal{S} \to \mathbb{R}^k$. *The cdf of* X *is defined for* $\boldsymbol{x} = (x_1, ..., x_k) \in \mathbb{R}^k$ *where* $F_X(x) = \mathcal{P}(X_1 \leq x_1, ..., X_k \leq x_k)$.

For continuous random variables, the density $f_X \geq 0$ satisfies

$$F_X(x) = \int\limits_{-\infty}^{x_1} ... \int\limits_{-\infty}^{x_k} f_X(z_1, ..., z_k) dz_1 ... dz_k$$

$$f_X(x) = \frac{\partial^k F_X(x_1, ..., x_k)}{\partial x_1, ..., \partial x_k}$$

where $\int f_X(x)dx = 1$. For discrete random variables, we have $\sum f_X(x) = 1$ and the pmf $f_X(x_1, ..., x_k) = \mathcal{P}(X_1 = x_1, ..., X_k = x_k)$ and cdf $F_X(x) = \sum_x f_X(x)$.

The multivariate pdf follows from iterating the fundamental theorem of calculus.

We call $f_X(x)$ a **joint probability distribution** where X denotes a **random vector**. From the joint pdf, we are able to compute the **marginal distributions** of some subset X_1 of $X = (X_1, X_2)$ where $X_1 \in \mathbb{R}^{k_1}$ and $X_2 \in \mathbb{R}^{k_2}$ such that $k_1 + k_2 = k$.

$$f_{X_1}(x_1) = \sum_{x_2} f_X(x_1, x_2)$$

$$f_{X_1}(x_1) = \int\limits_{-\infty}^{\infty} f_X(x_1, x_2) dx_2$$

It is important to note that knowing the marginals, we cannot compute the joint distribution as knowing f_{X_1} and f_{X_2} does not uniquely determine f_X unless X_1 and X_2 are independent or perfectly dependent i.e. $X_1 = cX_2$.[1]

Definition 15. *The conditional mass function (or density in the continuous case) is* $f_{Y|X}(y|z) = \dfrac{f_{Y,X}(y,x)}{f_X(x)}$ *where* $f_{Y|X}(y|x) \geq 0$ *and* $\int_{-\infty}^{\infty} f_{Y|X}(y|x)dy = 1$ *or* $\sum_y f_{Y|X}(y|x) = 1$.

We may define the conditional cdf through the conditional density or vice versa

$$F_{Y|X}(y|x) = \int\limits_{-\infty}^{y} f_{Y|X}(y'|x)dy'$$

$$f_{Y|X}(y|x) = \frac{\partial F_{Y|X}(y|x)}{\partial y}$$

Within the context of conditional distributions, we can define expectations, higher moments and conditional quantiles respectively as follows

$$E(Y|X = x) = \int\limits_{-\infty}^{\infty} y f_{Y|X}(y|x)dy$$

$$Var(Y|X = x) = E(Y^2|X = x) - (E(Y|X = x))^2$$

[1] However, we are able to establish a general relation between the joint cdf and marginal cdfs in a more general setting by examining the so-called **Frechet Bounds on the Joint Distribution** which state that $\max\{F_X(x) + F_Y(y) - 1, 0\} \leq F_{XY}(x,y) \leq \min\{F_X(x), F_Y(y)\}$. We can prove this result as follows.

Let $A = \{X \leq x\}$ and $B = \{Y \leq y\}$. The joint cdf is the probability of realising $A \cap B$ while the marginal cdf of X is the probability of A. We can see that $\mathcal{P}(A \cap B) \leq \mathcal{P}(A)$ and that $\mathcal{P}(A \cap B) \leq \mathcal{P}(B)$ which implies $F_{XY}(x,y) \leq \min\{F_X(x), F_Y(y)\}$. We also know that $\mathcal{P}(A \cap B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cup B) \geq \mathcal{P}(A) + \mathcal{P}(B) - 1$.

$$Q_{Y|X}(\alpha) = \inf\{y : F_{Y|X}(y|x) \geq \alpha\}, \quad \alpha \in [0,1]$$

Definition 16. *Y and X are **independent** i.e. $Y \perp\!\!\!\perp X$ if $\mathcal{P}(Y \in A, X \in B) = \mathcal{P}(Y \in A)\mathcal{P}(X \in B)$ for all events A and B in respective σ-algebras on $\mathcal{S}_Y$ and $\mathcal{S}_X$ i.e. if for all $x, y \in \mathbb{R}$, $F_{Y,X}(y,x) = F_Y(y)F_X(x)$.*

Alternatively, we sometimes say that Y and X are independent if any of the following holds:

$$f_{Y,X}(y,x) = f_Y(y)f_X(x)$$

$$f_{Y|X}(y|x) = f(y)$$

$$f_{X|Y}(x|y) = f(x)$$

for all $x, y \in \mathbb{R}$. Note that if $Y \perp\!\!\!\perp X$, then $g(X) \perp\!\!\!\perp h(Y)$ for any measurable functions g and h.

Definition 17. *Let X, Y be random variables such that $Var(X), Var(Y) < \infty$. Then, define the **covariance** $\sigma_{XY}(X,Y) = E((X - E(X))(Y - E(Y)) = E(XY) - E(X)E(Y)$.*

Note that covariance remains unaffected by the units of measurement as $\sigma_{XY}(aX + c, bY + d) = E((aX - aE(X))(bY - bE(Y)) = ab\sigma_{XY}(X,Y)$. The covariance essentially measures the degree of co-movement between X and Y. Positive covariance corresponds to positive co-movement and negative covariance to negative co-movement. Zero covariance corresponds to no co-movement.

Covariance is intimately related to the formula for additive variances.

Observe that $Var(X + Y) = E[(X - E(X) + Y - E(Y)]^2 = E(X - E(X))^2 + E(Y - E(Y))^2 + 2E(X - E(X))(Y - E(Y)) = Var(X) + Var(Y) + 2Cov(X,Y)$.

If $Cov(X,Y) = 0$, then $Var(X + Y) = Var(X) + Var(Y)$.

If $Cov(X,Y) < 0$, then $Var(X + Y) < Var(X) + Var(Y)$ and we can flip the inequality to get the opposite result.

We can extend the result to say that $Var(X + Y + Z) = Var(X) + Var(Y) + Var(Z) + 2Cov(X,Y) + 2Cov(X,Z) + 2Cov(Y,Z)$ and so on for the variance of the sum of even more random variables.

Definition 18. *The **correlation** is the typical measure of association described by the coefficient*

$\rho_{XY} = \dfrac{\sigma_{XY}(X,Y)}{\sigma_X(X)\sigma_X(Y)}$ *which satisfies* $\rho_{XY} \in [-1,1]$.

Just as with covariance, correlation is invariant to the units of measurement. Hence, $\rho_{a+bX,c+dY}$ is independent of constants a and c and will be unchanged as long as b and d have the same sign. If they have different signs, then the correlation coefficient will change sign but magnitude will be unchanged.

Proposition 14. *X and Y are independent random variables* $\implies \sigma_{XY}(X,Y) = 0$.

Proof. $E(XY) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} xyf(x,y)dydx = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} xyf(y)f(x)dydx = \int\limits_{-\infty}^{\infty} xf_X(x)dx \int\limits_{-\infty}^{\infty} yf_Y(y)dy =$ $E(X)E(Y)$ under independence. Thus, $\sigma_{XY}(X,Y) = 0$. □

Proposition 14 is a sufficiency argument. It is not a necessity argument in general. For instance, we could have $E(XY) - E(X)E(Y) = 0$ when $E(XY) \neq E(X)E(Y)$. For instance, we need $E(XY)$ and one of $E(X)$ or $E(Y)$ to be zero for this to hold.

An example where this holds is the case of $X = \cos\theta$ and $Y = \sin\theta$ where $\theta \sim U[0, 2\pi]$. Clearly, $Y^2 = 1 - X^2$ by the identity that $\cos^2\theta + \sin^2\theta \equiv 1$ so X and Y are functionally related and not independent.

However, $\sigma_{XY}(X,Y) = \int\limits_0^{2\pi} \cos\theta\sin\theta.d\theta - \int\limits_0^{2\pi}\cos\theta.d\theta \int\limits_0^{2\pi}\sin\theta.d\theta = 0.$

Definition 19. *X and Y are* ***bivariate normal*** *if*

$$f_{X,Y}(x,y|\mu_x,\mu_y,\sigma_X^2,\sigma_Y^2,\rho_{XY}) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}}\exp\left(-\frac{u(x,y)}{2(1-\rho_{XY}^2)}\right)$$

$$u(x,y) = \left(\frac{x-\mu_x}{\sigma_X}\right)^2 + \left(\frac{y-\mu_y}{\sigma_Y}\right)^2 - \frac{2\rho_{XY}(x-\mu_x)(y-\mu_y)}{\sigma_X\sigma_Y}$$

where $\mu_x = E(X)$, $\mu_y = E(Y)$, $\sigma_X^2 = Var(X)$, $\sigma_X^2 = Var(Y)$, $\sigma_{XY} = \sigma_X\sigma_Y\rho_{XY}$ *where* $\rho_{XY} = Corr(X,Y)$.

Using Definition 19, we can extend Proposition 14 to a necessary and sufficient condition in one particular case.

Proposition 15. (X, Y) *are jointly normal independent random variables* $\iff \rho_{XY}(X, Y) = 0$.

Proof. Sufficiency is already established by Proposition 14.

To prove necessity, suppose $\rho_{XY}(X, Y) = 0$. The bivariate normal pdf reduces to $f(x, y) = \dfrac{1}{2\pi\sigma_X\sigma_Y}\exp\left(-\dfrac{u(x,y)}{2(1-\rho_{XY}^2)}\right)$ where $u(x, y) = \left(\dfrac{x-\mu_x}{\sigma_X}\right)^2 + \left(\dfrac{y-\mu_y}{\sigma_Y}\right)^2$.

From there, it can be shown (tediously) that $E(XY) = E(X)E(Y)$. □

We can generalise the notions of expectation, variance and covariance to the multivariate case with the following concepts.

Definition 20. *The* ***mean vector*** *is* $\mu = E(X) = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} \in \mathbb{R}^k$ *and the* ***covariance matrix***

$$\Sigma = E[(X-E(X))(X-E(X))^T] = (Cov(X_i, X_j))_{i,j} = E(XX^T)-\mu\mu^T = \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \dots & \sigma_{kk} \end{pmatrix} \in \mathbb{R}^{k\times k}$$

has variances on the main diagonal and covariances elsewhere.

The covariance matrix is symmetric, which follows from the fact that $\sigma_{ij} = \sigma_{ji}$ which holds since all covariances commute.

Due to symmetry, for a $k \times k$ covariance matrix, the number of unique entries is $\frac{1}{2}k(k+1)$ i.e. the main diagonal and either above or below the main diagonal.

Proposition 16. *The covariance matrix* Σ *is always positive semi-definite.*

Proof. By definition of positive definiteness, consider $a^T\Sigma a = a^TE[(X - E(X))(X - E(X))^T]a = E[(aX - E(aX))(aX - E(aX))^T = Var(a^TX) \geq 0$. □

It is positive definite unless there exists a variable that is an exact linear function of other variables. This is in line with the notion of linear dependence of the rows/columns of Σ which implies singularity. Hence, $\det(\Sigma) \neq 0$ only exists when Σ is positive definite i.e. no variable can be expressed as an exact linear function of the other variables.

Using the notation from Definition 20, we can write the normal pdf as

$$f_X(x|\mu, \Sigma) = \frac{1}{\sqrt{2\pi \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

In the bivariate case, we can write this compactly as

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}\right) = N(\mu, \Sigma)$$

and we can write that (assuming positive definite Σ)

$$\det(\Sigma) = \sigma_X^2 \sigma_Y^2 - \sigma_{XY}^2 = \sigma_X^2 \sigma_Y^2 (1 - \rho_{XY}^2)$$

$$\Sigma^{-1} = \frac{1}{\det \Sigma} \begin{pmatrix} \sigma_Y^2 & -\sigma_{XY} \\ -\sigma_{XY} & \sigma_X^2 \end{pmatrix}$$

Proposition 17. *If X and Y are bivariate normal, then*

1. *$X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$.*
2. *Conditional distributions are such that $f_{Y|X}(y|x) \sim N(\mu_{Y|X}, \sigma_{Y|X}^2)$ where $\mu_{Y|X} = E(Y|X) = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(X - \mu_X)$ and $\sigma_{Y|X}^2 = \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}$.*
3. *If the marginals and conditionals are normal, the joint is also normal.*

For any pair of bivariate normals X and Y, we can say that $Y = \alpha + \beta X + \epsilon$ with $\epsilon \sim N(0, \sigma_{Y|X}^2)$ and also $\alpha = \mu_Y - \mu_X \frac{\sigma_{XY}}{\sigma_X^2}$ and $\beta = \frac{\sigma_{XY}}{\sigma_X^2}$.

Proposition 18. *Suppose X and Y are bivariate standard normal with correlation ρ. Then, we can write that $Y = \rho X + \sqrt{1-\rho^2} Z$ with Z being standard normal.*

Using Proposition 18, we can write $Y^2 = \rho^2 X^2 + (1-\rho^2)Z^2 + 2\rho\sqrt{1-\rho^2} XZ$ which computing covariances gives us that $Cov(X^2, Y^2) = \rho^2$.

Definition 21. ***Bayes' Theorem for Densities**.*

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} = \frac{f_{Y|X}f_X(x)}{\displaystyle\int f_{YX}(y,x)dx} \propto f_{Y|X}(y|x)f_X(x)$$

Proposition 19. *Suppose that $X \sim N(\mu_X, \sigma_X^2)$ and $Y|X = x \sim N(x, \sigma_u^2)$.*

$$\textit{Then } X|Y = y \sim N\left(\frac{\sigma_u^2}{\sigma_u^2+\sigma_X^2}\mu_X + \frac{\sigma_X^2}{\sigma_u^2+\sigma_X^2}y, \quad \frac{\sigma_X^2\sigma_u^2}{\sigma_u^2+\sigma_X^2}\right).$$

1.8 Conditional Expectations

The **conditional expectation** $E(Y|X = x)$ is sometimes called (in statistical inference) the **regression function**. It measures the response of Y to X. Note that $E(X|Y) \neq E(Y|X)$.

However, conditional expectations obey the monotonicity and linearity properties of expectation such that $E(\alpha X + \beta Y|Z = z) = \alpha E(X|Z = z) + \beta E(Y|Z = z)$ for random variables X, Y and Z and $\alpha, \beta \in \mathbb{R}$.

We can construct $Y = E(Y|X)+Y-E(Y|X) = m(X)+\epsilon$ with $E(Y|X) = m(X)$ and $E(\epsilon|X) = 0$. We describe $m(X)$ as systematic and ϵ as random.

On average, we have that the random part is zero, however $Var(\epsilon|X) = \sigma^2(X)$ need not be independent of X and so we could have the variance changing with X which is called **heteroskedasticity**.

In statistical inference, it is often assumed that the regression function is linear in parameters i.e. $Y = \alpha + \beta X + \epsilon$ where ϵ is independent of X with mean zero and variance $\sigma^2 < \infty$.

This follows from the joint normality of Y and X but normality is unnecessary for the linearity of $E(Y|X) = E(\alpha + \beta X + \epsilon|X) = \alpha + \beta X$.

Proposition 20. ***Law of Iterated Expectations**. $E(Y) = E(E(Y|X))$*

Proof. We have $E(Y) = \int y f_Y(y)dy = \int y \int f_{YX}(y,x)dxdy = \int\int y f_{YX}(y,x)dydx = \int\int y f_{Y|X}(y|x)dy)f_X(x)dx = \int E(Y|X)f_X(x)dx = E(E(Y|X))$. □

A key implication of the Law of Iterated Expectations is that if $E(\epsilon|X) = 0$ (a special case of **mean independence** i.e. where $E(\epsilon|X) = E(\epsilon)$), then $E(h(X)\epsilon) = 0$ for all measurable h. That is, we must have $E(X\epsilon) = 0$, $E(X^2\epsilon) = 0$ etc.

This follows because, for instance, $E(X\epsilon) = Cov(X, \epsilon) + E(X)E(\epsilon) = 0$ since $E(\epsilon) = 0$ by the case of mean independence above.

Another key implication can be summarised as follows. Suppose there are professional forecasters and the general public both trying to forecast an event Y. Let X denote the random variable representing public information and Z represents information available only to professional forecasters.

The Law of Iterated Expectations implies $E(Y|X) = E(E((Y|X, Z)|X)$ where $E(Y|X, Z)$ is the professional forecast. That is, the forecast given only public information is equal to the average forecast made by the professional forecasters.

Suppose $\hat{E}$ is the unbiased forecast of Y based on information $\{X, Z\}$, then the forecast error $Y - \hat{E}$ is such that $E(Y - \hat{E}|X) = 0$ i.e. on average, the forecast error will be zero.

Proposition 21. ***Law of Iterated Variances*** *or the* ***Analysis of Variance Formula****.* $Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$.

Proof. $Var(Y) = E((Y - E(Y)^2) = E((Y - E(Y|X) + E(Y|X) - E(Y))^2) = E((Y - E(Y|X))^2) + E((E(Y|X) - E(Y))^2) + 2E((Y - E(Y|X))(E(Y|X) - E(Y))$.

By the Law of Iterated Expectations, rewriting first and second terms gives $E((Y - E(Y|X))^2) = E[E(Y - E(Y|X))^2|X] = E(Var(Y|X))$ and $E((E(Y|X) - E(Y))^2) = Var(E(Y|X))$.

Since $\epsilon = Y - E(Y|X)$ is such that $E(\epsilon|X) = E(\epsilon) = 0$, and $E(Y|X) - E(Y)$ is measurable on X, the third term becomes (denoting $h(X) = E(Y|X) - E(Y)$) $E(\epsilon h(X)) = E(E(\epsilon h(X)|X)) = E[h(X)E(\epsilon|X)] = 0$. Combine to complete the proof. □

Clearly, $Var(Y) \geq Var(E(Y|X))$ with $Var(E(Y|X)$ being the systematic part and $E(Var(Y|X))$ being the random part.

Indeed we can use the population **coefficient of determination** which is the population R^2 to measure the proportion of variation explained by the systematic part, expressed as $R^2 = \dfrac{Var(E(Y|X))}{Var(Y)} = 1 - \dfrac{E(Var(Y|X))}{Var(Y)} \in [0, 1]$ by the Law of Iterated Variances.

A useful application of the Law of Iterated Variances allows us to write (in the regression function setting) that $Var(Y) = E(Var(Y|X)) + Var(E(Y|X)) = E(\sigma^2) + Var(\alpha + \beta X) = \sigma^2 + \beta^2 Var(X)$.

Proposition 22. ***Law of Iterated Covariances***. *Define the* ***conditional covariance*** *as* $Cov(Y, X|Z) = E[(Y - E(Y|Z))(X - E(X|Z))|Z]$.

The Law of Iterated Covariances states $Cov(Y, X) = E(Cov(Y, X|Z)] + Cov(E(Y|Z), E(X|Z))$.

Proof. From invoking the Law of Iterated Expectations, we can write that

$$Cov(Y, X) = E(XY) - E(X)E(Y) = E(E(XY|Z)) - E(E(X|Z))E(E(Y|Z))$$

$$= E[Cov(X, Y|Z) + E(X|Z)E(Y|Z)] - E(E(X|Z))E(E(Y|Z))$$

$$= E(Cov(X, Y|Z)) + E[E(X|Z)E(Y|Z)] - E(E(X|Z))E(E(Y|Z))$$

$$= E(Cov(X, Y|Z)) + Cov(E(X|Z), E(Y|Z)).$$ □

If we have $Cov(Y, X) = 0$, it is possible to have $Cov(Y, X) \neq 0$ for some Z. For instance, consider $X \sim N(0, 1)$ and let Y represent a coin toss. $Y = X$ if heads, and $Y = -X$ when tails.

Clearly, $\mathcal{P}(Y \leq y) = \mathcal{P}(Y \leq y|H)\mathcal{P}(H) + \mathcal{P}(Y \leq y|T)\mathcal{T} = 0.5\Phi(y) + 0.5\Phi(y) = \Phi(y)$ where Φ is the normal cdf. Hence, $Y \sim N(0, 1)$ as the cdf encodes all the necessary information to find the expectation and variance.

Note that $Cov(X, Y) = E(XY) - E(X)E(Y) = E(XY) = 0.5E(X^2) - 0.5E(X^2) = 0$, whereas $Cov(X, Y|H) = Cov(X, X) = Var(X) = 1$ while $Cov(X, Y|T) = Cov(X, -X) = -Var(X) = -1$.

It is also possible to have $Cov(Y, X|Z) = 0$ for all Z but $Cov(X, Y) \neq 0$.

For instance, consider $Y = Z + \epsilon$ and $X = Z + \eta$ where ϵ and η are such that $\epsilon \perp\perp \eta$ and $\epsilon, \eta \perp\perp Z$. Then, $Cov(X, Y|Z = z) = Cov(\epsilon, \eta) = 0$.

However, $Cov(Y, X) = Var(Z) + Cov(Z, \eta) + Cov(Z, \epsilon) + Cov(\eta, \epsilon) = Var(Z) > 0$ as Z is a random variable.

In general, we may have that conditioning on variable Z reverses the sign of the covariance.

1.9 Mean-Squared Minimisation and Marginal Effects

Proposition 23. *For random variables* X *and* Y *with* $E(Y^2) < \infty$, *we have* $E(Y|Z)$ *that minimises* $E[(Y - g(X))^2]$ *over all measurable functions* g.

Hence, we call $m(X) = E(Y|X) = \int y f_{Y|X}(y|x)dy$ *the* ***best predictor*** *of* Y *by* X *with respect to the mean-squared error.*

If $m(X)$ does not vary with X, then Y is mean independent of X and we must have $m(X) = E(Y)$.

Proof. By the Law of Iterated Expectations, we write $E[(Y - g(X))^2] = E[E(Y - g(X))^2|X]$.

With some algebra, $E[E(Y - g(X))^2|X] = E[E(Y - E(Y|X) + E(Y|X) - g(X))^2|X]$

$= E[E[(Y - E(Y|X))^2|X] + E(E(g(X) - E(Y|X))^2|X) + E[E(Y - E(Y|X))(E(Y|X) - g(X))|X]$

$= Var(Y|X) + E(g(X) - E(Y|X))^2 \geq Var(Y|X)$ for all functions g and realisations X with equality if and only if $g(X) = E(Y|X)$ with probability 1. □

Proposition 24. *For random variables X and Y with $E(Y^2) < \infty$, we define $E_L(Y|X)$ as the unique minimiser of $E[(Y - g(X))^2]$ and $E[(m(X) - g(X))^2]$ over all linear functions $g(x) = \alpha + \beta x$ where*

$$E_L(Y|X) = \alpha_L + \beta_L X = m_L(X)$$

with $\alpha_L = E(Y) - \beta_L E(X)$ and $\beta_L = \dfrac{Cov(Y,X)}{Var(X)}$ provided $Var(X) > 0$.

We say $E_L(Y|X)$ is the ***best linear predictor****. Note that generally $m_L(x) \neq m(x)$.*

If $Cov(Y,X) = 0$, then $E_L(Y|X) = \alpha_L = E(Y)$.

Proof. Using calculus, let $Q(\alpha, \beta) = E[(Y - \alpha - \beta X)^2]$.

Then, $\dfrac{\partial Q}{\partial \alpha} = -2E[(Y - \alpha - \beta X)]$ and $\dfrac{\partial Q}{\partial \beta} = -2E[X(Y - \alpha - \beta X)]$.

Setting derivatives equal to zero gives two equations in two unknowns, whose solution is written in matrix form as $\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \dfrac{1}{E(X^2) - E(X)^2} \begin{pmatrix} E(X^2) & -E(X) \\ -E(X) & 1 \end{pmatrix} \begin{pmatrix} E(Y) \\ E(XY) \end{pmatrix}$ which has a unique solution (α_L, β_L) providing $Var(X) > 0$.

The second order condition is satisfied since the Hessian is $2 \begin{pmatrix} 1 & E(X) \\ E(X) & E(X^2) \end{pmatrix}$ which is positive definite whenever $Var(X) > 0$ (which by definition holds) and so the objective function is strictly convex and so the first-order conditions will guarantee a minimum.

The same argument applies for $Q(\alpha, \beta) = E[(m(X) - \alpha - \beta X)^2]$ except we notice that $E(XE(Y|X)) = E(E(XY|X)) = E(XY)$ by the Law of Iterated Expectations and the fact that in

$E(E(XY|X))$, X behaves like a constant. Hence, $Cov(X,Y) = Cov(m(X),X)$. □

Also, observe that $E[(Y-E(Y|X))^2] \leq E[(Y-E_L(Y|X))^2] \leq E[(Y-E(Y))^2]$. That is, the regression function (and best predictor) $E(Y|X)$ is more accurate according to the mean-squared error than the best linear predictor, which is in turn more accurate than the unconditional mean.

If $E(Y|X)$ is linear, then $E(Y|X) = E_L(Y|X)$ and so $E[(Y-E(Y|X)^2] = E[(Y-E_L(Y|X))^2]$.

By using the law of iterated expectations i.e. that $E(Y) = E(E(Y|X))$, we are able to write that $Y = m(X)+\epsilon$ where $E(\epsilon|X) = 0$ and also $Y = \alpha_L + \beta_L X + \eta = m_L(X) + \eta$ where $E(\eta) = 0$ and so $E(\eta X) = 0$ (also $Cov(X,\eta) = 0$) but not necessarily $E(\eta|X) = 0$.

Note that $E(E_L(Y|X)) = \alpha_L + \beta_L E(X) = E(Y)$.

We can extend this to the multivariate case with $X \in \mathbb{R}^k$ where we have $E_L(Y|X) = \beta_L^T X = m_L(X)$ with $\beta_L = E(XX^T)^{-1}E(XY)$ provided the matrix $E(XX^T)$ is nonsingular to ensure β_L is uniquely defined.

Nevertheless, even if we have $E(XX^T)$ being singular, then we can uniquely define a best linear predictor based on projection arguments where many β_L may yield the same $\beta_L^T X$.

Often, we are not interested in the whole regression function $E(Y|X)$ but just in the **average marginal effect** $\delta = E(m'(X)) = \int m'(X)f_X(x)dx$. In the special case of a linear regression function $m(x) = \alpha + \beta x$ for constants α and β, then $\delta = \beta_L = \beta$ but otherwise not.

When we are interested in just the average marginal effect, it may make sense to ask whether δ can be approximated by β_L. Heuristically, one possible approximation to a nonlinear function is the Taylor approximation at its mean. Hence, consider $m(X) \approx m(\mu) + (X-\mu)m'(\mu)$. If we let $\beta_T = m'(\mu)$ and $\alpha_T = m(\mu) - \mu m'(\mu)$ and define $m_T(x) = \alpha_T + \beta_T x$, we can linearly approximate m. In general, it will turn out that $m_T(x) \neq m_L(x)$.

To see when the linearisation by Taylor polynomials is a good approximation, we can show that $E[(m(X) - m_T(x))] \leq 0.5Var(X)\sup_x|m''(x)|$ for all x.

That is, the expected difference between the best predictor and the Taylor approximation is bounded by half of the variance multiplied by the greatest possible absolute value of the second derivative of the best predictor.

This holds for any x and so if $Var(X)$ or $\sup|m''(x)|$ are small, then the approximation error may be very small.

Clearly, if m is linear then $m'' = 0$ and so the approximation is perfect.

In general, it is possible to have the linear approximation being very close but generally not equivalent.

For the same x, we usually have $E[(Y - m(X))^2] \leq E[(Y - m_L(X))^2] \leq E[(Y - m_T(X))^2]$ meaning that the best linear predictor is better than the Taylor approximation to the best predictor.

When the best predictor is linear, all of the expectations above are equivalent.

Proposition 25. *Suppose that $X \sim N(\mu_X, \sigma_X^2)$ and that $Y = m(X) + \epsilon$ where $E(\epsilon|X) = 0$ and m is nonlinear. Then, $\delta = \beta_L$.*

That is, for normally distributed random covariates, the slope of the best linear predictor measures perfectly the average marginal effect irrespective of m.

The proof of Proposition 25 is involved but relies on another result in probability theory that is stated below.

Lemma 1. ***Stein's Lemma***. *Suppose $X \sim N(\mu, \sigma^2)$ and m is a function on which $E(m(X)(X-\mu))$ and $E(m'(X))$ exist. Then, $E(m(X)(X - \mu)) = \sigma^2 E(m'(X))$.*

Proof. The left hand side is $E(m(X)(X - \mu)) = \dfrac{1}{\sqrt{2\pi}\sigma} \displaystyle\int m(x)(x - \mu) \exp\left(-\dfrac{(x - \mu)^2}{2\sigma^2}\right) dx$.

This integrates by parts with $u = m(x)$ and $dv = (x - \mu) \exp\left(-\dfrac{(x - \mu)^2}{2\sigma^2}\right)$ to give the right hand side. □

1.10 Asymptotic Theory

Asymptotic theory is all about examining the properties of estimators as the sample size becomes very large. There are some useful inequalities that we can use to place upper or lower bounds on specific unknown values.

For instance, we can take expectations of the triangle inequality to give $E|X+Y| \leq E|X| + E|Y|$. This works because if the triangle inequality holds for all random variables X and Y, then taking expectations should allow it to still hold.

Proposition 26. ***Cauchy-Schwarz Inequality for Expectations***. *For random variables X and Y, $(E(XY))^2 \leq E(X^2)E(Y^2)$.*

The triangle inequality and Cauchy-Schwarz inequality for expectations can be crude inequalities. That is, suppose X and Y are standard normal with correlation ρ, then $(E(XY))^2 = \rho^2$.

If $\rho = 0$, then the Cauchy-Schwarz inequality for expectations is very strong as the right-hand side is 1 and the left-hand side zero. If $\rho = -1$, then $X + Y = 0$ and so the triangle inequality for expectations will be very strong.

Proposition 27. ***Markov's Inequality***. *For $\eta > 0$ and any random variable X, we can write the upper bound $\mathcal{P}(|X| \geq \eta) \leq \dfrac{E|X|}{\eta}$.*

Proof. $\mathcal{P}(|X| \geq \eta) = E(\mathbf{1}(|X| \geq \eta))$ where $\mathbf{1}$ is an indicator function which takes value 1 if $|X| \geq \eta$ and 0 otherwise.

Thus in every case $\mathbf{1}(|X| \geq \eta) \leq \dfrac{|X|}{\eta}$ so $E(\mathbf{1}(|X| \geq \eta)) \leq \dfrac{E|X|}{\eta}$. □

Proposition 28. ***Chebyshev's Inequality***. *For random variable X with finite variance, we can write an upper bound $\mathcal{P}(|X - E(X)| \geq k\sigma) \leq \dfrac{1}{k^2}$.*

Proof. Apply Markov's Inequality to random variable $(X - E(X))^2$ and $\eta = (k\sigma)^2$. □

The toolset of asymptotic theory involves generalising the usual notion of convergence to random variables.

Definition 22. *A sequence of random variables $\{X_n\}_{n=1}^{\infty}$* ***converges in probability*** *to a random variable (or constant) X i.e. $X_n \xrightarrow{P} X$, if for all $\epsilon > 0$, $\lim_{n\to\infty} \mathcal{P}(|X_n - X| > \epsilon) = 0$ (sometimes written shorthand as $X = \operatorname{plim}_{n\to\infty} X_n$).*

Definition 23. *Presuming $E(X_n^2), E(X^2) < \infty$, a sequence of random variables $\{X_n\}_{n=1}^{\infty}$* ***converges in mean squared*** *to a random variable (or constant) X i.e. $X_n \xrightarrow{MS} X$, if $\lim_{n\to\infty} E[(X_n) - X)^2] = 0$ i.e. the limit of the mean-squared error is zero.*

Note that when X is a constant, we have $E((X_n - X)^2) = Var(X_n) + (E(X_n - X))^2$ and so a necessary and sufficient condition for convergence in mean squared is for $E(X_n) \to X$ and $Var(X_n) \to 0$.

Proposition 29. $X_n \xrightarrow{MS} X \implies X_n \xrightarrow{P} X$

Proof. Applying Markov's Inequality to $|X_n - X|^2$ gives $\mathcal{P}(|X_n - X|^2 \geq \eta) = \mathcal{P}(|X_n - X| \geq \sqrt{\eta}) \leq \dfrac{E[(X_n - X)^2]}{\sqrt{\eta}}$.

Taking limits gives $\lim_{n\to\infty} \mathcal{P}(|X_n - X| \geq \sqrt{\eta}) \leq \lim_{n\to\infty} \dfrac{E[(X_n - X)^2]}{\sqrt{\eta}} = 0$ by definition of convergence in mean-squared.

By definition of probabilities as non-negative, $\lim_{n\to\infty} \mathcal{P}(|X_n - X| \geq \sqrt{\eta}) = 0$.

Letting $\epsilon = \sqrt{\eta}$ without loss of generality completes the proof. □

However, convergence in probability does not imply convergence in mean-squared.

For instance, suppose X_n took value n with probability $1/n$ and 0 with probability $1 - 1/n$. For any $\epsilon > 0$, $\mathcal{P}(X_n \geq \epsilon) = 1/n \to 0$ and so this random variable exhibits convergence in probability.

However, $E(X_n) = 1$ which does not converge to zero and since X is a constant (i.e. 0), we do not have convergence in mean-squared.

Definition 24. *A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ **converges in distribution** to a random variable X i.e. $X_n \xrightarrow{D} X$, if for all x at which $\mathcal{P}(X \leq x)$ is continuous, we have* $\lim_{n\to\infty} \mathcal{P}(X_n \leq x) = \mathcal{P}(X \leq x)$.

Proposition 30. $X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$

Generally, convergence in distribution does not imply convergence in probability.

For example, consider $X_n = (-1)^n X$ where X is standard normal. Clearly, X and $-X$ are both standard normal so trivially we have convergence in distribution. However, the sequence never converges in probability to X as it oscillates between X and $-X$ evermore.

However, there is a special case in which convergence in distribution does imply convergence in probability.

Proposition 31. *When X is a constant, $X_n \xrightarrow{P} X \iff X_n \xrightarrow{D} X$.*

Summarising, we have that convergence in mean-squared necessarily implies convergence in probability which in turn implies convergence in distribution.

Proposition 32. ***Weak Law of Large Numbers.*** *If $X_1, ..., X_n$ are independent and identically distributed random variables, if $E|X_i|< \infty$ for all i, then $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu$.*

Proof. We assume $E|X_i^2|< \infty$[2] i.e. where variance is finite.

Note that $Var(T_n) = \frac{\sigma^2}{n}$ and $E(T_n) = \mu$.

Using Chebyshev's inequality, $\mathcal{P}(|X_n - \mu|\geq \epsilon) \leq \frac{Var(T_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$. □

Proposition 33. ***Lindeberg-Levy Central Limit Theorem.*** *Let $X_1, ..., X_n$ be independent and identically distributed with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$.*

Then, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) \xrightarrow{D} N(0, \sigma^2)$.

There are many generalisations of the Central Limit Theorem to random variables that are not independent and identically distributed.

Proposition 34. ***Lindeberg's Condition.*** *Let $X_1, ..., X_n$ be independent random variables with $E(X_i) = \mu_i$ and $Var(X_i) = \sigma_i^2$. Let $s_n^2 = \sum_{i=1}^{n} \sigma_i^2$.*

If $\frac{1}{s_n^2}\sum_{i=1}^{n} E[(X_i - \mu_i)^2 \boldsymbol{1}((X_i - \mu_i)^2 > \epsilon s_n^2)] \to 0$ *for all* $\epsilon > 0$, *then* $\frac{1}{s_n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{D} N(0, 1)$ *where **1** is an indicator function.*

If X_i are independent and identically distributed, Lindeberg's condition is automatically satisfied but the condition extends the Central Limit Theorem to cases where we have heterogeneously distributed random variables.

[2] This is a sufficient condition but not necessary for the Weak Law of Large Numbers. A necessary condition is that X is symmetrically distributed around zero and $\lim_{n\to\infty} n\mathcal{P}(|X|> n) = 0$.

Proposition 35. ***Lyapunov's Condition***. *Under the same conditions as in Proposition 34, if* $\sum_{i=1}^{n} \frac{E|X_i - \mu_i|^3}{s_n^3} \to 0$, *then* $\frac{1}{s_n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{D} N(0,1)$.

Lyapunov's Condition holding necessarily implies that Lindeberg's Condition holds, but not vice versa. Therefore Lindeberg's Condition is a weaker sufficient condition for the Central Limit Theorem than Lyapunov's, making it a more powerful result.

Note that both Lyapunov's Condition and Lindeberg's Condition are sufficient conditions for the Lindeberg-Levy Central Limit Theorem to apply. More generally, by writing the proposition below, we establish a necessary condition.

Proposition 36. ***Necessity of Lindeberg's Condition***. *If the sequence of random variables satisfies* $\frac{\max_{1\leq i\leq n} \sigma_i^2}{s_n^2} \to 0$, *then Lindeberg's Condition is necessary and sufficient for the Lindeberg-Levy Central Limit Theorem to hold.*

Intuitively, Proposition 36 requires that the sum is not dominated by a single random variable.

To see this more clearly, suppose Z_i are independent and identically distributed with mean zero and variance 1.

Consider the weighted random variables $X_i = w_i Z_i$ where w_i are deterministic weights i.e. constants.

Defining $s_n^2 = \sum_{i=1}^{n} w_i^2$ and $E|Z_i|^3 < \infty$, the condition $\frac{\max_{1\leq i\leq n} w_i^2}{\sum_{i=1}^{n} w_i^2} = \left(\frac{\max_{1\leq i\leq n}|w_i|}{\sqrt{\sum_{i=1}^{n} w_i^2}}\right)^2 \to 0$ implies Lyapunov's Condition because for each i, $E|X_i|^3 = |w_i^3|E|Z_i|^3 \leq w_i^2(\max_{1\leq i\leq n}|w_i|E|Z_i|^3)$.

Note however that Proposition 36 does not generally imply Lyapunov's Condition, but only when the random variables have a weighted independent and identically distributed structure.

As an example of a typical distribution where the Central Limit Theorem does not hold, consider

$$X_i = \begin{cases} \frac{1}{i}, & p = \frac{1}{2} \\ \frac{-1}{i}, & p = \frac{1}{2} \end{cases}$$

Or we can write this as $X_i = \dfrac{Z_i}{i}$ with $Z_i = 1$ with probability half and -1 with probability half. Clearly the weak law of large numbers holds because $E(X_i) = 0$ and $E(X_i^2) = i^{-2}$ so $\sum_{i=1}^{\infty} i^{-2} = \dfrac{\pi^2}{6} < \infty$.

However, to check if the Central Limit Theorem holds, we must consider either Lindeberg's Condition or Lyapunov's Condition.

Testing Lindeberg's Condition, we have $\epsilon > 0$, $\sum_{i=1}^{n} \dfrac{1}{i^2} E[Z_i^2 \mathbf{1}(Z_i^2 > \epsilon i^2)] = \sum_{i=1}^{n} \dfrac{1}{i^2} \mathbf{1}(1 > \epsilon i^2) = \sum_{i=1}^{1/\sqrt{\epsilon}} \dfrac{1}{i^2} \mathbf{1}(1 > \epsilon i^2) \nrightarrow 0$.

Intuitively, the reason for the Lindeberg Condition not being satisfied is that the distribution of X_i shrinks with i and is dominated by X_1.

The problem has the structure of a weighted sums of independent and identically distributed random variables. Hence, we can apply the sufficient condition for Lyapunov's Condition which is $\dfrac{\max_{1 \le i \le n} w_{ni}^2}{\sum_{i=1}^{n} w_{ni}^2} \to 0$.

Intuitively, this says that each term should be individually negligible. In the above case, when $w_{ni} = i^{-1}$, this is not true since $\max_{1 \le i \le n} w_{ni}^2 = 1$ and $\sum_{i=1}^{n} w_{ni}^2 = \sum_{i=1}^{n} \dfrac{1}{i^2}$ is finite.

Proposition 37. ***Continuous Mapping Theorem.*** *Suppose g is continuous. Then, if $X_n \xrightarrow{D} X$ then $g(X_n) \xrightarrow{D} g(X)$. Also, if $X_n \xrightarrow{P} X$ then $g(X_n) \xrightarrow{P} g(X)$.*

A key example is that if $X_n \xrightarrow{D} X$ where $X \sim N(0,1)$ then $X_n^2 \xrightarrow{D} X^2 \sim \chi_1^2$.

Continuity of g in the Continuous Mapping Theorem is a sufficient but not necessary condition. If we don't have a continuous g, then results may hold or may not.

For example, consider $T_n = \dfrac{1}{n} \to 0$ with probability 1. Let $g(t) = \mathbf{1}(t > 0)$. Then, $g(T_n) = 1$ for all n but $g(0) = 0$.

As another example, suppose $T_n = \dfrac{\sum X_i}{\sqrt{n}}$ wth X_i being independent and identically distributed with $E(X_i) = 0$ and $Var(X_i) = 1$. Let $g(t) = \mathbf{1}(t > 0)$. Then, observe that $T_n \xrightarrow{D} T \sim N(0,1)$ by the Central Limit Theorem.

Also, observe that there are two discontinuities in the cdf of $g(T_n)$ at 0 and 1. Where the cdf is continuous i.e. where $t \in (0,1)$, we have $g(T_n) \xrightarrow{D} g(T)$ because for $t \in (0,1)$, $\mathcal{P}(g(T_n) \leq t) = \mathcal{P}(T_n \leq 0) \rightarrow \mathcal{P}(T \leq 0)$.

Proposition 38. ***Slutsky's Theorem*** *If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} \alpha$, then*

1. $X_n + Y_n \xrightarrow{D} X + \alpha$
2. $X_n Y_n \xrightarrow{D} \alpha X$
3. $\dfrac{X_n}{Y_n} \xrightarrow{D} \dfrac{X}{\alpha}$ *providing $\alpha \neq 0$.*

Both Slutsky's Theorem and the Continuous Mapping Theorem are nothing but generalisations of obvious results for deterministic sequences.

Note that it is very important that α is a constant in Slutsky's Theorem, otherwise it does not hold.

Consider the following example where we suppose that $X_n = \dfrac{\sum A_i}{\sqrt{n}}$ and $Y_n = \dfrac{\sum B_i}{n}$ where A_i, B_i are independent and identically distributed with $E(A_i) = 0$, $Var(A_i) = \sigma_A^2$ and $E(B_i) = \mu_B$.

Then, we have that $X_n + Y_n = \dfrac{\sum A_i}{\sqrt{n}} + \dfrac{\sum B_i}{n} \xrightarrow{D} X + \mu_B \sim N(\mu_B, \sigma_A^2)$.

Furthermore, $X_n Y_n = \left(\dfrac{\sum A_i}{\sqrt{n}}\right)\left(\dfrac{\sum B_i}{n}\right) \xrightarrow{D} X\mu_B \sim N(0, \mu_B^2 \sigma_A^2)$.

Finally, we have $\dfrac{X_n}{Y_n} \xrightarrow{D} \dfrac{X}{\mu_B} \sim N\left(0, \dfrac{\sigma_A^2}{\mu_B^2}\right)$.

We can extend our notion of convergence in probability to vectors with the following definition.

Definition 25. *A random vector $X_n = (X_{n1}, ..., X_{nk})^T$ is such that $X_n \xrightarrow{P} X$ if for all $\epsilon > 0$, we have $\mathcal{P}\left(\sum_{j=1}^{k} (X_{nj} - X_j)^2 > \epsilon\right) \rightarrow 0$, which holds if and only if for all $j = 1, ..., k$, $|X_{nj} - X_j| \xrightarrow{P} 0$.*

Definition 25 tells us that we must check every element of a vector. If every element converges in probability, then the whole vector necessarily converges in probability. Next, we extend the notion of convergence in distribution to random vectors below.

Definition 26. *A sequence of random vectors* $\{X_n\}_{n=1}^{\infty}$ *converges in distribution to random vector* $X \in \mathbb{R}^k$ *if for all* $x \in \mathbb{R}^k$ *at which* $\mathcal{P}(X \leq x)$ *is continuous,* $\lim_{n\to\infty} \mathcal{P}(X_n \leq x) = \mathcal{P}(X \leq x)$.

In this case, it is not sufficient that $X_{nj} \xrightarrow{D} X_j$ for each j. We also care about the covariance between the elements. That is, if $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} Y$, it does not automatically follow that $X_n + Y_n \xrightarrow{D} X + Y$. Hence, we require a further result.

Proposition 39. ***Cramer-Wold Theorem***. *A random vector* X_n *converges in distribution to* X *if and only if* $\sum_{j=1}^{k} c_j X_{nj} \xrightarrow{D} \sum_{j=1}^{k} c_j X_j$ *for every* $c_1, ..., c_k$.

As an example, suppose that $X_n = \dfrac{\sum_{i=1}^{n} A_i}{\sqrt{n}}$ and $Y_n = \dfrac{\sum_{i=1}^{n} B_i}{\sqrt{n}}$ where A_i and B_i are independent and identically distributed with $E(A_i) = E(B_i) = 0$ and $Var(A_i) = \sigma_A^2$ and $Var(B_i) = \sigma_B^2$.

Then, $c_1 X_n + c_2 Y_n = \dfrac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i$ where $Z_i = c_1 A_i + c_2 B_i$. Note that $E(Z_i) = 0$ and so $c_1 X_n + c_2 Y_n \xrightarrow{D} N(0, c_1^2 \sigma_A^2 + c_2^2 \sigma_B^2 + 2c_1 c_2 Cov(A_i, B_i))$ for all c_1, c_2 by the Central Limit Theorem.

We may extend our analysis of asymptotic theory to general nonlinear functions of random variables that obey the Central Limit Theorem.

Proposition 40. ***Delta Method***. *Univariate Case. Suppose* $X_n = \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} X \in \mathbb{R}$ *for some random variable* X *and that* $f : \mathbb{R} \to \mathbb{R}$ *is continuously differentiable with* $f'(\theta_0) \neq 0$. *Then,* $\sqrt{n}(f(\hat{\theta}) - f(\theta_0)) \xrightarrow{D} N(0, (f'(\theta_0)^2 \sigma^2)$.

Multivariate Case. Suppose that $X_n = \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} \boldsymbol{X} \in \mathbb{R}^k$ *for some random vector* $\boldsymbol{X}$ *and that* $f : \mathbb{R}^k \to \mathbb{R}$ *is continuously differentiable with* $f'(\theta_0) \neq 0$. *Then,* $\sqrt{n}(f(\hat{\theta}) - f(\theta_0)) \xrightarrow{D} N(0, \nabla f(\theta_0)^T \Sigma \nabla f(\theta_0))$.

Proof. We prove the Delta Method for the scalar case.

By the mean value theorem, $f(\hat{\theta}) = f(\theta_0) + (\hat{\theta} - \theta_0) f'(\theta^*)$. Scaling up gives $\sqrt{n}(f(\hat{\theta}) - f(\theta_0)) = f'(\theta^*)\sqrt{n}(\hat{\theta} - \theta_0)$ where $\theta^* \in [\theta_0, \hat{\theta}]$ (that is, $|\theta^* - \theta_0| \leq |\hat{\theta} - \theta_0|$).

Since $\hat{\theta} \xrightarrow{P} \theta_0 \implies \theta^* \xrightarrow{P} \theta_0 \implies f'(\theta^*) \xrightarrow{P} f'(\theta_0) \neq 0$ by the Continuous Mapping Theorem. By Slutsky's Theorem, $f'(\theta^*)\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} f'(\theta_0) N(0, \sigma^2) = N(0, \sigma^2 (f'(\theta_0))^2)$. □

As an example of the Delta Method, suppose we have X_i for $i = 1, ..., n$ being independent and identically distributed with mean μ_0 and variance σ_0^2 and let $\hat{\mu} = n^{-1}\sum_{i=1}^{n} X_i$.

By the Central Limit Theorem, we observe that $\sqrt{n}(\hat{\mu} - \mu_0) \xrightarrow{D} N(0, \sigma_0^2)$. If we seek to find the limiting distribution of $\hat{\mu}^3$, we can apply the Delta method directly to obtain $\sqrt{n}(\hat{\mu}^3 - \mu^3) \xrightarrow{D} N(0, 9\sigma^2\mu^4)$. Note that this only works with $\mu \neq 0$.

If we have $\mu = 0$, then we could simply cube both sides of $\sqrt{n}(\hat{\mu}) \xrightarrow{D} N(0, \sigma_0^2)$ to obtain that $n\sqrt{n}\hat{\mu}^3 \xrightarrow{D} N(0, \sigma_0^2)^3$.

In this case, the Delta Method serves as a quick result that simplifies the asymptotic analysis. However, sometimes it is easier to work directly.

Consider the following example. Suppose $X_n = \dfrac{\sum_{i=1}^{n} A_i}{n}$ and $Y_n = \dfrac{\sum_{i=1}^{n} B_i}{n}$ where A_i and B_i are independent and identically distributed with $E(A_i) = \mu_A \neq 0$ and $E(B_i) = \mu_B \neq 0$.

Also, $Var(A_i) = \sigma_A^2 < \infty$ and $Var(B_i) = \sigma_B^2 < \infty$. Define $\tilde{X}_n = \dfrac{1}{\sqrt{n}}\sum_{i=1}^{n}(A_i - \mu_A)$ and $\tilde{Y}_n = \dfrac{1}{\sqrt{n}}\sum_{i=1}^{n}(B_i - \mu_B)$.

Then, we are able to write that $X_n = \mu_A + \dfrac{\tilde{X}_n}{\sqrt{n}}$ and $Y_n = \mu_B + \dfrac{\tilde{Y}_n}{\sqrt{n}}$.

Thus, if we look at $X_nY_n = \mu_A\mu_B + \mu_A\dfrac{1}{\sqrt{n}}\tilde{Y}_n + \dfrac{1}{\sqrt{n}}\tilde{X}_n\mu_B + \dfrac{1}{n}\tilde{X}_n\tilde{Y}_n$, then rearranging and scaling gives $\sqrt{n}(X_nY_n - \mu_A\mu_B) = \mu_A\tilde{Y}_n + \mu_B\tilde{X}_n + \dfrac{\sqrt{n}}{n}\tilde{X}_n\tilde{Y}_n = \dfrac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i + \dfrac{\sqrt{n}}{n}\tilde{X}_n\tilde{Y}_n$ where $W_i = \mu_A(B_i - \mu_B) + \mu_B(A_i - \mu_A)$ and $E(W_i) = 0$ and $Var(W_i) = \sigma_W^2 < \infty$.

These conditions enable the Central Limit Theorem to apply to the first term involving W_i.

We could also look at the ratio $\dfrac{X_n}{Y_n} = \dfrac{\mu_A + \dfrac{1}{\sqrt{n}}\tilde{X}_n}{\mu_B + \dfrac{1}{\sqrt{n}}\tilde{Y}_n} = \dfrac{\mu_A}{\mu_B} + \dfrac{1}{\sqrt{n}}\dfrac{\tilde{X}_n}{\mu_B} - \dfrac{\mu_A}{\mu_B^2}\dfrac{1}{\sqrt{n}}\tilde{Y}_n + ...$ from the Taylor approximation to the denominator term. Rearranging and scaling gives $\sqrt{n}\left(\dfrac{X_n}{Y_n} - \dfrac{\mu_A}{\mu_B}\right) = \dfrac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i + ...$ where $W_i = \dfrac{\mu_A(B_i - \mu_B)}{\mu_B^2} - \dfrac{A_i - \mu_A}{\mu_B}$ has $E(W_i) = 0$ and $Var(W_i) < \infty$. Once again, the Central Limit Theorem applies to the first term on the right hand side.

Our asymptotic tools may also help with the properties of some key sample statistics.

Proposition 41. *Let $X_1, ..., X_n$ be a random sample from a population of mean μ, variance σ^2, skewness κ_3 and kurtosis κ_4.*

Then, $E(\hat{\mu}) = \mu$ and $Var(\hat{\mu}) = \frac{\sigma^2}{n}$.

Furthermore, $\hat{\mu} \xrightarrow{P} \mu$ and by the Central Limit Theorem, $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{D} N(0, \sigma^2)$.

Moreover, if $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$ is an estimator of σ^2, $E(s^2) = \sigma^2\frac{n-1}{n}$ with $s^2 \xrightarrow{P} \sigma^2$ and by the Central Limit Theorem, $\sqrt{n}(s^2 - \sigma^2) \xrightarrow{D} N(0, (\kappa_4 + 2)\sigma^4)$.

Proof. The asymptotic properties of the sample mean are trivial applications of the Weak Law of Large Numbers and Central Limit Theorem. The sample variance case is more involved.

With some algebra, we can take expectations of s^2 and obtain $E(s^2) = \frac{n-1}{n}\sigma^2$. Observe that $s^2 \xrightarrow{P} \sigma^2$.

Furthermore, $\sqrt{n}(s^2 - \sigma^2) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 - \sigma^2\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}((X_i - \mu)^2 - \mu^2) + R_n$ where $R_n = -\sqrt{n}(\mu - \bar{X})^2 \xrightarrow{P} 0$.

It follows that denoting $Z_i = (X_i - \mu)^2 - \sigma^2$ allows us to argue that $\sqrt{n}(s^2 - \sigma^2) \xrightarrow{D} N(0, Var(Z_i))$.

We can compute that $Var(Z_i) = E[(X_i - \mu)^2 - \sigma^2]^2 = E(X_i - \mu)^4 - \sigma^4$.

Hence, $\sigma^4\left(\frac{E(X_i - \mu)^4}{\sigma^4} - 1\right) = \sigma^4(\kappa_4 + 2)$. □

We can find the joint asymptotic distribution of the vector $T_n = (\sqrt{n}(\hat{\mu} - mu), \sqrt{n}(s^2 - \sigma^2))^T$ using the Cramer Wald Theorem since we have both elements of the vector converging to normal distributions, and so any linear combination of the vector elements will converge.

As such, we can say that $T_n \xrightarrow{D} N(0, \Sigma)$ where $\Sigma = \begin{pmatrix} \sigma^2 & \sigma^2\kappa_3 \\ \sigma^2\kappa_3 & \sigma^4(\kappa_4 + 2) \end{pmatrix}$ and $\sigma^2\kappa_3 = Cov(\sqrt{n}(\hat{\mu} - \mu), \sqrt{n}(s^2 - \sigma^2) = E[(X_i - \mu)^3]$.

2 Statistical Inference

2.1 Frequentist and Bayesian Inference

A **statistic** is just a measurable function of the sample $X_1, ..., X_n$, that is, some $T(X_1, ..., X_n)$. The **frequentist** approach to statistical inference involves some model $\mathbf{P}$ for the data. $\mathbf{P}$ is a collection of probability measures which constitute a unique true measure (or true distribution) $\mathcal{P}$ and other false ones $\mathcal{Q} \neq \mathcal{P}$.

We are interested in **parameters** $\theta(\mathcal{P})$ and obtain a sample $\mathcal{X}^n = \{X_1, ..., X_n\}$ from the true $\mathcal{P}$ and estimate $\theta(\mathcal{P})$ by some function $\hat{\theta}$ of the data. We treat $\hat{\theta}$ as a random variable because it arose as a realisation of the sampling process delivered by $\mathcal{P}$. We are thus interested in the variability of $\hat{\theta}$ across different potential samples.

Sampling is most commonly conducted randomly. That is, we draw every sample "with replacement" from an infinite population. This allows for us to say that every sample is independent and identically distributed.

However, we might want to weaken the assumption that the population is the same for each observation.

We may be interested in heterogeneous individuals or allow for dependence between samples.[3]

In the **classical parametric** approach, we define $\mathbf{P} = \{\mathcal{P}_\theta, \theta \in \Theta\}$ for $\Theta \subset \mathbb{R}^k$ as the **parameter space** for each θ. $\mathcal{P}_\theta$ is a probability measure (or density function in the continuous case).

In this approach, the entire distribution of the data is specified except from the unknown parameters. For example, we may write any distribution in terms of unknown parameters. So we

[3] Consider the following problem which illustrates this dependence. If we have a small population of n individuals, and p of them are of interest for a study, while the remaining q are not (such that $q + p = n$), we might be interested in what happens if we draw m people randomly without replacement. Denoting X as the random variable of the number of people of interest, we can write that

$$\mathcal{P}(X = k) = \frac{\binom{p}{k}\binom{n-p}{n-k}}{\binom{n}{m}}, \quad k = 1, ..., m$$

This is called a **hypergeometric** distribution and has the properties of $E(X) = \frac{pm}{n}$ and $Var(X) = \frac{pm}{n} \times \frac{n-p}{n} \times \frac{n-m}{n-1}$.

could write a normal random variable $X \sim N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_+$ as having density function

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

Definition 27. *The* ***likelihood function*** *is $L(\theta|\mathcal{X}^n) = f(X_1, ..., X_n|\theta)$. It gives the density of the data given the parameter but is treated as a function of θ for given data.*

Under independence and identical distribution of $X_1, ..., X_n$, we may write the likelihood as $L(\theta|\mathcal{X}^n) = \prod_{i=1}^{n} f(X_i|\theta)$ and the **log-likelihood** as $\mathcal{L}(\theta|\mathcal{X}^n) = \sum_{i=1}^{n} logf(X_i|\theta)$.

As an example, consider the Bernoulli population where X takes 1 with probability p and 0 with probability $1-p$, where $p \in [0,1]$ is unknown.

All variables are independent and identically distributed and so we may write $L(p|\mathcal{X}^n) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = \begin{pmatrix} n \\ k \end{pmatrix} p^k(1-p)^{n-k}$ with k ones and $n-k$ zeros.

Similarly, we may write $\mathcal{L}(p|\mathcal{X}^n) = log \begin{pmatrix} n \\ k \end{pmatrix} + klog(p) + (n-k)log(1-p)$.

If we have a continuous distribution such as the normal distribution, then we may write that $L(\mu, \sigma^2|\mathcal{X}^n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}(\frac{X_i-\mu}{\sigma})^2 \right)$ and $\mathcal{L}(\mu, \sigma^2|\mathcal{X}^n) = -\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$.

Next, we discuss **identification** which is about the theoretical question of whether, after observing an infinite number of observations, the true parameters can be ascertained.

This is true if and only if different values of the parameters generate different probability distributions.

Definition 28. *Two parameter points θ and θ' in a parameter set Θ are* ***observationally equivalent*** *if for all $\mathcal{X}^n$, $L(\theta|\mathcal{X}^n) = L(\theta'|\mathcal{X}^n)$ with probability 1.*

A parameter point θ in Θ is ***identifiable*** *if there is no other $\theta' \in \Theta$ with $\theta' \neq \theta$ that is observationally equivalent to it.*

If there is no $\mathcal{X}^n$ for which the likelihoods are different for some $\theta' \neq \theta$, then the parameter is unidentifiable.

For example, suppose X_i are independent and identically distributed normal random variables with mean μ and variance one.

Then, we have $\mathcal{L}(\mu|\mathcal{X}^n) = -\frac{n}{2}log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2$. For any $\mu' \neq \mu$, we have that

$$\mathcal{L}(\mu|\mathcal{X}^n) - \mathcal{L}(\mu'|\mathcal{X}^n) = -\frac{1}{2}\left(\sum_{i=1}^{n}(X_i - \mu)^2 - \sum_{i=1}^{n}(X_i - \mu')^2\right)$$

$$= -\frac{1}{2}\left(\sum_{i=1}^{n}(X_i - \mu' + \mu' - \mu)^2 - \sum_{i=1}^{n}(X_i - \mu')^2\right) = -\frac{1}{2}\left(\sum_{i=1}^{n}(\mu' - \mu)^2 + 2(\mu' - \mu)\sum_{i=1}^{n}(X_i - \mu')\right)$$

Clearly if (without loss of generality) $\mu' > \mu$, then for some $\mathcal{X}^n$, we have that $(\mu' - \mu)\sum_{i=1}^{n}(X_i - \mu') > 0$ and so this model is identifiable as the likelihoods for different parameters are not observationally equivalent.

However, we may often encounter unidentifiable models if they are observationally equivalent for different parameters.

For instance, if we have independent and identically distributed X_i normal random variables with mean $\mu_1 + \mu_2$ and variance 1.

Then, for any μ_1', μ_2' with $\mu_1' + \mu_2' = \mu_1 + \mu_2$, $\mathcal{L}(\mu_1, \mu_2|\mathcal{X}^n) = -\frac{n}{2}log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu_1 - \mu_2)^2 = \mathcal{L}(\mu_1', \mu_2'|\mathcal{X}^n)$.

Similarly, for independent and identically distributed normal X_i with mean $\sin(\theta)$ and variance 1. Then, for any $\theta' = \theta + 2\pi$, $\mathcal{L}(\theta|\mathcal{X}^n) = -\frac{n}{2}log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i - \sin(\theta))^2 = \mathcal{L}(\theta'|\mathcal{X}^n)$.

We may conclude that in these unidentifiable cases, the model is badly specified or there are too many unknown parameters.

Often, we have to specify a likelihood function that depends on the realisation of an observation. For instance, we may have a model $U[\theta_0, \theta_0 + 1]$.

Suppose that $U[0, 1]$ represents the true density. The likelihood function of the true density given one observation X is $L(0, 1|X) = 1$ if $X \in [0, 1]$ and 0 otherwise.

The likelihood function of the model given X is $L(\theta_0, \theta_0 + 1|X) = 1$ if $X \in [\theta_0, \theta_0 + 1]$ and 0 otherwise.

Recall that so long as the likelihoods are not always the same for all observations, parameter points are identifiable.

If an X observed is in $[\theta_0, \theta_0 + 1]$ but outside $[0, 1]$, then the likelihoods will not be the same. Since this occurs for some X, we have that 0 and 1 are identifiable parameters.

Definition 29. ***In-Sample Identification***. *A parameter point θ in Θ is* ***unidentifiable*** *in sample if for any sample $\mathcal{X}^n$, there exists $\theta' \in \Theta$ such that $L(\theta|\mathcal{X}^n) = L(\theta'|\mathcal{X}^n)$.*

That is, if θ is unidentifiable, we are unable to distinguish between θ and $\theta' \neq \theta$ given the data available.

In-sample identification is about whether one sample observation $\mathcal{X}^n$ allows us to correctly identify a parameter point θ.

It differs from the population definition of identification since in-sample unidentifiability typically results from a lack of observations.

For instance, if we have one observation X in the uniform case of true density $U[0, 1]$ and model density $U[\theta_0, \theta_0 + 1]$, then if $X \in [0, 1]$ and $X \in [\theta_0, \theta_0 + 1]$, then the likelihood functions are equivalent and parameter points 0 and 1 are unidentifiable.

However, if we select a different sample and obtain an X such that the likelihoods differ, then the parameter points become identifiable in the new sample.

For small-sample observations, the **Bayesian** approach to statistical inference is often used. Here, parameters θ are also random variables and have prior density function $\pi(\theta)$ which reflects knowledge about θ without seeing the sample.

The objective in Bayesian statistics is to update the prior using the sample data as represented by the likelihood and obtain the posterior by Bayes' Theorem where $\pi(\theta|\mathcal{X}^n) \propto f(\mathcal{X}^n|\theta)\pi(\theta)$. Then, we can report various features of the posterior as an estimator of θ.

As an example, consider the case where we observe $X = 1$. Then, the likelihood of the sample (distribution of the data) is $f(X = 1|p) = p^X(1 - p)^{1-X} = p$.

Suppose the prior distribution $\pi(p)$ is uniform on $[0, 1]$. That is, it places equal probability on different p within this interval.

The posterior density after observing $X = 1$ is $\pi(p|X) = \dfrac{f(X|p)\pi(p)}{\displaystyle\int f(X|p)\pi(p)dp} = \dfrac{p}{\displaystyle\int_0^1 pdp} = 2p.$

Next suppose we observe $X = 2$ as a second observation.

Then, the prior is $f(X_1 = 1, X_2 = 0|p) = p^1(1-x)^0p^0(1-x)^1 = p(1-p)$.

The posterior density of p after observing X_1, X_2 is $\dfrac{p(1-p)}{\displaystyle\int_0^1 p(1-p)} = 6p(1-p)$.

When we have a Poisson distribution where $X \sim Po(\lambda)$, we have an unknown parameter $\lambda \geq 0$.

Typically we allow for $\lambda \in \mathbb{R}_+$ and so we can use an ignorance prior that takes value 1 for every non-negative real number.

Obviously this doesn't integrate to one but the posterior density is still well-defined. For instance, suppose we observe $X = 1$.

Then, we have posterior density $\dfrac{\lambda\exp(-\lambda)}{\displaystyle\int_0^\infty \lambda\exp(-\lambda)} = \lambda\exp(-\lambda)$ for all $\lambda \in \mathbb{R}_+$.

In the binomial case, if we observe k values of $X = 1$ and $n-k$ values of $X = 0$, then the likelihood of the sample (distribution of the data) is $f(\mathcal{X}^n|p) = \begin{pmatrix} n \\ k \end{pmatrix} p^k(1-p)^{n-k}$.

Again suppose that the prior distribution $\pi(p)$ is uniform on $[0,1]$.

The posterior density of p is given by $\pi(p|\mathcal{X}^n) = \dfrac{f(\mathcal{X}^n|p)\pi(p)}{\displaystyle\int f(\mathcal{X}^n|p)\pi(p)dp} = \dfrac{p^k(1-p)^{n-k}}{\displaystyle\int_0^1 p^k(1-p)^{n-k}dp} =$ $\dfrac{(n+1)!}{k!\,(n-k)!}p^k(1-p)^{n-k}$.

Note the integration here where $\left(\displaystyle\int_0^1 p^k(1-p)^{n-k}dp\right)^{-1} = \dfrac{(n+1)!}{k!\,(n-k)!}$ is very tedious.

As another example, suppose we observe a single data point $X \sim N(\mu_X, \sigma_X^2)$ where σ_X^2 is a known quantity.

Suppose the prior of μ_X is $N(\mu, \sigma^2)$ where μ, σ^2 are known. Then, the posterior density is given by an application of Proposition 19 below

$$\mu_X|X \sim N\left(\frac{\sigma_X^2}{\sigma^2+\sigma_X^2}\mu + \frac{\sigma^2}{\sigma^2+\sigma_X^2}X,\ \left(\frac{1}{\sigma^2}+\frac{1}{\sigma_X^2}\right)^{-1}\right)$$

We can interpret the equation above as follows. The posterior mean is a weighted average of the prior mean and the data mean, whereas the posterior variance is the harmonic average i.e. $\left((\sigma^2)^{-1}+(\sigma_X^2)^{-1})\right)^{-1}$ of the variance of the prior and the variance of the data.

As an extension, suppose we now have n observations $\mathcal{X}^n$, and so $\bar{X} \sim N(\mu_X, \sigma_X^2/n)$. It can be shown by using Proposition 19 that

$$\mu_X|\mathcal{X}^n \sim N\left(\frac{\sigma_X^2/n}{\sigma^2+\sigma_X^2/n}\mu + \frac{\sigma^2}{\sigma^2+\sigma_X^2/n}\bar{X},\ \left(\frac{1}{\sigma^2}+\frac{n}{\sigma_X^2}\right)^{-1}\right)$$

Asymptotically, we have that $\dfrac{\sigma_X^2/n}{\sigma^2+\sigma_X^2/n}\mu + \dfrac{\sigma^2}{\sigma^2+\sigma_X^2/n}\bar{X} \xrightarrow{P} \mu_X$.

That is, as our observations become large, our sample mean converges in probability to the population mean.

2.2 Estimation

Proposition 42. ***Method of Moments***. *Suppose $X \in \mathbb{R}^k$ is a random variable with $\theta \in \mathbb{R}^p$ being a vector of unknown parameters.*

Suppose also that m and f are $p-$vectors of functions with f being invertible and corresponding inverse g such that

$$E(m(X)) = f(\theta) \iff \theta = g(E(m(X)))$$

Let $\{X_1, ..., X_n\}$ be a sample of data. Then, the ***method of moments estimator (MoM)*** *is*

$$\hat{\theta}_{MoM} = g\left(\frac{1}{n}\sum_{i=1}^{n} m(X_i)\right)$$

For instance, suppose $\mu = E(X)$ and $\sigma^2 = Var(X)$. Then, we have that the skewness is

$\kappa_3 = \dfrac{E[(X - E(X)^3]}{Var(X)^{3/2}}$ as a function $g(E(X), E(X^2), E(X^3))$.

By the method of moments, we can write estimators such that $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = s^2$ and also $\hat{\kappa}_3 = \dfrac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^3}{\left(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right)^{3/2}} = g\left(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n X_i^2, \frac{1}{n}\sum_{i=1}^n X_i^3\right)$.

The standard t-distribution is given by $f(x|v) \propto (1 + x^2/v)^{-(v+1)/n}$ where v is the number of degrees of freedom.

In this case, we have that $Var(X) = E(X^2) = \dfrac{v}{v-2} \implies v = \dfrac{2E(X^2)}{E(X^2) - 1}$ whenever $v > 2$. Hence, by the method of moments, we can estimate v by $\hat{v} = \dfrac{2s^2}{s^2 - 1}$.

Proposition 43. *The* ***Maximum Likelihood Estimator (MLE)*** $\hat{\theta}$ *is the function of* $\mathcal{X}^n$ *that maximises* $L(\theta|\mathcal{X}^n)$ *or equivalently maximises* $\mathcal{L}(\theta|\mathcal{X}^n)$ *with respect to* $\theta \in \Theta \subset \mathbb{R}^k$*. The true parameter is denoted* $\theta_0 \in \Theta$*.*

With any estimation method, there are two fundamental issues we must consider. First, the question of whether there exists a maximum. In general, this is a difficult question but we have a sufficiency condition for this. If $\theta \mapsto \mathcal{L}(\theta|\mathcal{X}^n)$ is continuous and the parameter space Θ is compact with respect to $\mathbb{R}^k$, then Weierstrass' Theorem tells us that there exists a maximum.

Secondly, the question of uniqueness. That is, whether the set of maximisers is a singleton, a non-single finite set or an uncountable set. Lack of uniqueness might point to a more fundamental problem i.e. whether or not the model is identifiable. We can say that if $\theta \mapsto \mathcal{L}(\theta|\mathcal{X}^n)$ is globally strictly concave, then $\hat{\theta}$ is unique. Otherwise, we could select some subset of the set of maximisers so this may not be a problem.

Suppose there does exist a maximum. It could either be on the boundary of Θ or in the interior. If the maximum is in the interior, we can use calculus-based methods to find it.

Under the relevant smoothness conditions for which calculus works, we say that the optimal $\hat{\theta}$ satisfies $\dfrac{\partial \mathcal{L}}{\partial \theta}(\hat{\theta}|\mathcal{X}^n) = 0$ where the partial derivative is called the **score function**.

Note that this is a necessary condition for an interior maximum. A sufficient condition for an interior local maximum is that the second derivative satisfies $\frac{\partial^2 \mathcal{L}}{\partial \theta^2}(\hat{\theta}|\mathcal{X}^n) < 0$.

To ensure a global maximum, we must check all local maxima. It suffices to be a global maximum if we have the second derivative negative for all θ i.e. we have global concavity.

For example, suppose we have $X \sim N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2) \in \mathbb{R}, \mathbb{R}_+$ where the parameter space is not compact but $\mathcal{L}$ is globally concave. Then, we can write the log-likelihood

$$\mathcal{L}(\theta|\mathcal{X}^n) = -\frac{n}{2} log(2\pi) - \frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

The derivatives with respect to the parameters are

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)$$

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2$$

Setting these equations to zero allows us to solve both equations uniquely to give

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

Next consider the Binomial case. The parameter space is $[0, 1]$ and $X_i \in [0, 1]$. We can write the log-likelihood

$$\mathcal{L}(p|\mathcal{X}^n) = \sum_{i=1}^{n} X_i log(p) + (1 - X_i) log(1 - p)$$

Taking derivatives with respect to p and setting equal to zero gives

$$\frac{\partial \mathcal{L}}{\partial p} = \sum_{i=1}^{n} \frac{X_i}{p} - \frac{1 - X_i}{1 - p} = \sum_{i=1}^{n} \frac{X_i - p}{p(1 - p)}, \quad p \neq 0, 1 \iff \hat{p}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

If $X_i = 1$ for all i, then $\mathcal{L}(p|\mathcal{X}^n) = nlog(p)$ and so $\hat{p}_{MLE} = 1$ but since this is on the boundary of the parameter space, the first-order condition is not satisfied.

In the Poisson case, we have parameter space $\mathbb{R}_+$ with $X_i \in \{0, 1, 2, ...\}$ with log-likelihood

$$\mathcal{L}(\lambda|\mathcal{X}^n) = \sum_{i=1}^{n} X_i log(\lambda) - n\lambda - \sum_{i=1}^{n} log(X_i)$$

The derivatives are

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{i=1}^{n} \frac{X_i}{\lambda} - n = 0$$

$$\frac{\partial^2 \mathcal{L}}{\partial \lambda^2} = -\sum_{i=1}^{n} \frac{X_i}{\lambda^2} \leq 0$$

Hence, the first and second-order conditions are satisfied and so rearranging the first-order condition gives us that $\hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

We cannot always use calculus to find estimators.

For instance, consider the uniform distribution $U[0, \theta]$ where $\theta > 0$. The likelihood is equal to

$$L = \prod_{i=1}^{n} \frac{1}{\theta}\mathbf{1}(0 \leq X_i \leq \theta) = \begin{cases} \frac{1}{\theta^n}, & \text{if } \max X_i \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

The maximum is at $\hat{\theta} = \max_{1 \leq i \leq n} X_i$ and does not satisfy a first-order condition.

The method of moments estimator uses $E(X) = \theta/2$ and hence $\hat{\theta}_{MoM} = 2\bar{X}$.

So far, all of our expressions for estimators have been written in closed form i.e. explicit formulae of some known g.

More generally, the MLE may not be explicitly defined.

For example, consider the Cauchy distribution $f(x|\theta) = \frac{1}{\pi}\frac{1}{1 + (x - \theta)^2}$ for which the log likelihood function and score function respectively are

$$\mathcal{L}(\theta|\mathcal{X}^n) = -log(n\pi) - \sum_{i=1}^{n} log(1 + (X_i - \theta)^2)$$

$$\frac{\partial \mathcal{L}}{\partial \theta}(\theta|\mathcal{X}^n) = 2\sum_{i=1}^{n} \frac{X_i - \theta}{1 + (X_i - \theta)^2}$$

This gives a nonlinear equation in θ, namely $g(\hat{\theta}, \mathcal{X}^n) = 0$ which must be solved by numerical methods.

The estimator $\hat{\theta}$ is a random variable with a distribution (say P_{n,θ_0}) that depends on the truth θ_0 and the sample size n.

A widely agreed upon criterion for measuring estimator performance is the mean-squared error. In general, the MSE of an estimator $\hat{\theta}$ depends on θ_0 and n.

Recall that we may define the mean-squared error as $MSE_{\theta_0}(\hat{\theta}) = E[(\hat{\theta} - \theta_0)^2] = \mathcal{M}(\theta_0)$ where expectation is taken using the distribution of the data P_{n,θ_0}.

Definition 30. ***Multivariate Mean-Squared Error.*** *Suppose $\theta \in \mathbb{R}^p$ and the parameter estimator $\hat{\theta}$ has finite variance. Then, mean-squared error is $MSE_{\theta_0}(\hat{\theta}) = E[(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^T]$.*

The mean-squared error is a symmetric, positive semi-definite matrix.

In the multivariate case, we cannot necessarily compare two mean-squared error matrices and so we may need to work with other scalar functions of the mean-squared error such as the trace or determinant of the matrices. However, this becomes tricky since ranking by one measure may differ from a ranking constructed through another measure.

An ideal estimator would have zero MSE regardless of the truth. Unfortunately this is impossible unless the problem is trivial.

We can always construct an estimator with zero MSE at a single parameter point. For example, suppose $X \sim N(\theta_0, 1)$. The sample mean has mean squared error equal to $1/n$ for any θ_0. Consider the estimator $\hat{\theta} = 0$. This estimator has MSE equal to θ_0^2 which is equal to zero if and only if $\theta_0 = 0$. Hence, if $\theta_0 = 0$, this estimator is perfect, but with any other truth, the estimator is very bad.

We are looking for an estimator that performs well over a range of θ_0 since we don't know the truth. If we did, the problem would be trivial.

Note that we can write the mean-squared error as nothing but the sum of the variance and the squared bias. That is, $E(\hat{\theta} - \theta_0)^2 = E(\hat{\theta} - E(\hat{\theta})^2 + (E(\hat{\theta} - \theta_0))^2$.

This implies a trade-off between bias and variance. A large bias could be offset by a small variance and vice versa.

In many cases, calculating the exact MSE is difficult and we may work with an asymptotic approximation where $n \to \infty$ called the **asymptotic mean-squared error (AMSE)**.

Proposition 44. *An estimator is* ***consistent*** *if for all $\theta_0 \in \Theta$, $\hat{\theta} \xrightarrow{P} \theta_0$.*

*A sufficient condition for consistency is that $MSE_{\theta_0}(\hat{\theta}) \to 0$. Equivalently, if $E(\hat{\theta}) \to \theta_0$ (****asymptotic unbiasedness****) and $Var(\hat{\theta}) \to 0$.*

Definition 31. *An estimator is* ***asymptotically normal*** *if for all θ_0, there is positive finite $Var(\theta_0)$ such that $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, Var(\theta_0))$ and $AMSE_{\theta_0}(\hat{\theta}) = \dfrac{Var(\theta_0)}{n}$.*

Hence, asymptotic normality implies that we are able to rank asymptotically normal estimators according to their variance only.

Take an example where $X \sim N(\mu, \sigma^2)$ and let $\hat{\mu} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ and $\hat{\mu}_E = \dfrac{2}{n}\sum_{i=1}^{n/2} X_{2i}$.

Then, $E(\hat{\mu}) = \mu$ and $Var(\hat{\mu}) = \sigma^2/n$ so that the bias is zero and the $MSE(\hat{\mu}) = \sigma^2/n$ and $MSE(\hat{\mu}_E) = 2\sigma^2/n$ for all μ.

In addition, $\hat{\mu} \xrightarrow{P} \mu$ and $\hat{\mu}_E \xrightarrow{P} \mu$ but $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{D} N(0, \sigma^2)$ and $\sqrt{n}(\hat{\mu}_E - \mu) \xrightarrow{D} N(0, 2\sigma^2)$.

Hence, the estimator $\hat{\mu}_E$ has greater variance so is a worse estimator than $\hat{\mu}$. We can say this due to both being asymptotically normal.

Consider another comparison of estimators. Suppose $X \sim N(\mu, \sigma^2)$ and consider the following estimators of the variance σ^2.

First, we have $s^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ and second is $s_*^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

We have $E(s_*^2) = \sigma^2$ and $E(s^2) = \dfrac{n-1}{n}\sigma^2$ so s_*^2 is unbiased whereas s^2 is not for small n.

However, $Var(s_*^2) = \dfrac{2\sigma^4}{n-1}$ but $Var(s^2) = \dfrac{2\sigma^4(n-1)}{n^2}$.

Therefore, $MSE(s^2) = \dfrac{\sigma^4}{n^2} + \dfrac{2\sigma^4(n-1)}{n^2} = \dfrac{(2n-1)\sigma^4}{n^2} \leq \dfrac{2\sigma^4}{n-1} = MSE(s_*^2)$ and so the estimator s^2 is weakly better according to the MSE irrespective of σ^2.

Thus, despite being biased, its lower variance sufficiently well offsets this.

However, as $n \to \infty$, the difference disappears. We have that $\sqrt{n}(s^2 - \sigma^2) \xrightarrow{D} N(0, 2\sigma^4)$ and $\sqrt{n}(s_*^2 - \sigma^2) \xrightarrow{D} N(0, 2\sigma^4)$.

Thus, the two estimators are asymptotically equivalent where $\sqrt{n}(s_*^2 - s^2) \xrightarrow{P} 0$.

Thinking about the multivariate case, we can extend our discussions to the multivariate case with some matrix algebra.

For instance, suppose $X \in \mathbb{R}^d$ and $X \sim N(\mu, \Sigma)$ where $\mu \in \mathbb{R}^d$. Covariance matrix Σ is such that every entry $\sigma_{ij} = Cov(X_i, X_j)$. We can write out the log-likelihood

$$\mathcal{L}(\mu, \Sigma | \mathcal{X}^n) = c - \frac{n}{2} logdet(\Sigma) - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^T \Sigma^{-1}(X_i - \mu)$$

Then, we can obtain a score function for μ as

$$\frac{\partial \mathcal{L}(\mu, \Sigma | \mathcal{X}^n)}{\partial \mu} = -\sum_{i=1}^{n} \Sigma^{-1}(X_i - \mu)$$

MLEs can be calculated from the score functions for μ and Σ respectively to give

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})^T$$

It turns out that the mean vector μ is unbiased but the covariance matrix Σ is not.

The unbiased covariance matrix replaces n by $n - 1$, as with the univariate case.

2.3 Efficiency

Measuring the quality of estimators by lowest mean-squared error runs into some issues. For instance, a best MSE estimator uniformly over θ is impossible. As such, we may resolve this in a number of ways.

One basic way is to try to do the best against the worst that nature can do. This is the so-called

minimax approach which involves choosing $\hat{\theta}$ to solve $\min\sup_{\theta\in\Theta} MSE_\theta(\hat{\theta})$.

For example, suppose $X_i \sim N(\mu, 1)$ and consider sample mean $\bar{X}$. We have the mean-squared error of $\bar{X}$ equalling $1/n$ for all μ so that $\sup_{\mu\in\mathbb{R}} MSE_\mu(\bar{X}) = \dfrac{1}{n}$ and so minimising the highest possible MSE is trivial since MSE doesn't change with the parameter.

However, if we had an estimator $\hat{\theta} = 0$, then $\sup_{\mu\in\mathbb{R}} MSE_\mu(0) = \infty$ which is a very bad estimator. In general, the minimax estimator can be difficult to compute, but in this case, $\bar{X}$ turns out to be the minimax estimator.

Another strategy involves restricting the class of estimators to linear, unbiased estimators. Sometimes we look for **Uniformly Minimum Variance Unbiased Estimators (UMVUE)** or equivalently the **Best Unbiased Estimator (BUE)** but these are difficult to find. As such, by restricting ourselves to linear unbiased estimators, we have a desirable result.

Proposition 45. ***Gauss-Markov Theorem***. *Suppose $X_i \in \mathbb{R}^k$ are independent and identically distributed with mean vector μ and covariance matrix Σ.*

*The sample mean estimator $\hat{\mu} = \sum_{i=1}^n X_i/n$ is **Best Linear Unbiased Estimator (BLUE)**.*

That is, $Var(\hat{\mu}) \leq Var(\tilde{\mu})$ for any other $\tilde{\mu} = \sum_{i=1}^n c_{ni}X_i$ where c_{ni} are deterministic weights.

Proof. For $\tilde{\mu}$ to be unbiased, we need $E(\tilde{\mu}) = \mu\displaystyle\sum_{i=1}^n c_{ni}$ so that $\displaystyle\sum_{i=1}^n c_{ni} = 1$.

$Var(\hat{\mu}) = \dfrac{\Sigma}{n}$ and $Var(\tilde{\mu}) = \displaystyle\sum_{i=1}^n c_{ni}^2\Sigma$.

Therefore, $Var(\tilde{\mu}) - Var(\hat{\mu}) = \Sigma\left(\displaystyle\sum_{i=1}^n c_{ni}^2 - \frac{1}{n}\right) = \Sigma\left(\sum_{i=1}^n c_{ni}^2 - \frac{1}{n}\left(\sum_{i=1}^n c_{ni}\right)^2\right) \geq 0.$ □

We can think of efficiency within the class of maximum likelihood estimators.

For example, for any parameter values $\theta \in \Theta$, we can define the likelihood or log-likelihood, as well as the score function. The derivative of the score function in the multivariate case is the Hessian matrix.

Proposition 46. *If the support is independent of θ, $E\left(\dfrac{\partial\mathcal{L}}{\partial\theta}(\theta_0|\mathcal{X}^n\right) = 0$.*

Proof. $\int \frac{\partial \mathcal{L}}{\partial \theta}(\theta_0|\mathcal{X}^n) f(\mathcal{X}^n|\theta_0) d\mathcal{X}^n = \frac{\partial \mathcal{L}}{\partial \theta}(\theta_0|\mathcal{X}^n) \int f(\mathcal{X}^n|\theta_0) d\mathcal{X}^n = \frac{\partial}{\partial \theta} 1 = 0.$ □

Definition 32. *The* ***information*** *is the variance of the score function at the truth.*

That is, $\mathcal{I}_n(\theta_0) = E\left(\frac{\partial \mathcal{L}}{\partial \theta}(\theta_0|\mathcal{X}^n)^2\right)$.

Proposition 47. *If the support is independent of* θ, $E\left(\frac{\partial \mathcal{L}}{\partial \theta}(\theta_0|\mathcal{X}^n)^2\right) = -E\left(\frac{\partial^2 \mathcal{L}}{\partial \theta^2}(\theta_0|\mathcal{X}^n)\right)$.

For instance, consider $X \sim B(n,p)$. Then, we have the score function

$$\frac{\partial \mathcal{L}}{\partial p}(p|\mathcal{X}^n) = \frac{1}{p(1-p)} \sum_{i=1}^{n} (X_i - p)$$

The information is the variance of the score function at the truth or equally -1 times the expectation of the Hessian at the truth. That is, $\mathcal{I}_n(p) = \frac{n}{p(1-p)}$.

As another example, suppose $X \sim N(\mu, \sigma^2)$. Then, we have the score function

$$\frac{\partial \mathcal{L}}{\partial \mu}(\mu|\mathcal{X}^n) = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)$$

The information is $\mathcal{I}_n(\mu) = \frac{n}{\sigma^2}$.

One example of an exception to Proposition 47 is the case of uniform distribution $U[0,\theta]$ where $\theta > 0$.

The likelihood and score function respectively are

$$L(\theta|\mathcal{X}^n) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbf{1}(0 \leq X_i \leq \theta) = \begin{cases} \frac{1}{\theta^n}, & \text{if } \max X_i \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

$$\frac{\partial \mathcal{L}}{\partial \theta}(\theta|\mathcal{X}^n) = \begin{cases} -\frac{n}{\theta}, & \text{if } \max X_i \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

The support of X depends on θ so this example does not obey either Proposition 46 nor 47.

We can combine Propositions 46 and 47 into the multivariate case, which gives us the following.

Definition 33. *The* ***information matrix*** *is defined as* $\mathcal{I}_n(\theta) = E\left(\frac{\partial \mathcal{L}}{\partial \theta}(\theta|\mathcal{X}^n)\frac{\partial \mathcal{L}}{\partial \theta^T}(\theta|\mathcal{X}^n)\right) = -E\left(\frac{\partial^2 \mathcal{L}}{\partial \theta \partial \theta^T}(\theta|\mathcal{X}^n)\right)$.

This is nothing but an extension of the notion of information in the scalar or univariate case.

Proposition 48. ***Cramer-Rao Theorem****. Let $\tilde{\theta}$ be an unbiased estimator of θ. Assume the support of X is independent of θ and some other regularity conditions which we will not specify.*

Then, $Var(\tilde{\theta}) \geq \mathcal{I}_n^{-1}(\theta)$.

If the Cramer-Rao bound is binding, then the estimator $\tilde{\theta}$ is called ***efficient****.*

Proof. Let $\tilde{\theta}(\mathcal{X}^n)$ be an unbiased estimator and suppose X are continuously distributed. Then, for all $\theta \in \Theta$, $E(\tilde{\theta}) = \int \tilde{\theta}(\mathcal{X}^n) L_n(\theta|\mathcal{X}^n) d\mathcal{X}^n = \theta$.

The limits of the integral are independent of θ. Hence, differentiating with respect to θ gives $1 = \int \tilde{\theta}(\mathcal{X}^n) \frac{\partial \mathcal{L}_n(\theta|\mathcal{X}^n)}{\partial \theta} L_n(\theta|\mathcal{X}^n) d\mathcal{X}^n$.

Since $\int L_n(\theta|\mathcal{X}^n) d\mathcal{X}^n = 1$ (as $L_n(\theta|\mathcal{X}^n)$ is a density for each θ), $0 = \int \frac{\partial L_n(\theta|\mathcal{X}^n)}{\partial \theta} d\mathcal{X}^n = \int \frac{\partial \mathcal{L}_n(\theta|\mathcal{X}^n)}{\partial \theta} L_n(\theta|\mathcal{X}^n) d\mathcal{X}^n$.

Combining the two results allows us to say that $\int (\tilde{\theta}(\mathcal{X}^n) - \theta) \frac{\partial \mathcal{L}_n(\theta|\mathcal{X}^n)}{\partial \theta} L_n(\theta|\mathcal{X}^n) d\mathcal{X}^n = 1$.

We use the fact that $(\int fg)^2 \leq \int f^2 \int g^2$ (**Cauchy-Schwarz Inequality for Integrals**) taking $f = (\tilde{\theta}(\mathcal{X}^n) - \theta) L_n(\theta|\mathcal{X}^n)^{1/2}$ and $g = \frac{\partial \mathcal{L}_n(\theta|\mathcal{X}^n)}{\partial \theta} L_n(\theta|\mathcal{X}^n)^{1/2}$.

This gives $\int (\tilde{\theta}(\mathcal{X}^n) - \theta)^2 L_n(\theta|\mathcal{X}^n) d\mathcal{X}^n \geq \frac{1}{\int (\frac{\partial \mathcal{L}_n(\theta|\mathcal{X}^n)}{\partial \theta})^2 L_n(\theta|\mathcal{X}^n) d\mathcal{X}^n}$ as required. □

The Cramer-Rao Theorem also holds in the multivariate case but where $\theta, \tilde{\theta}$ are matrices, and so we cannot directly compute $Var(\tilde{\theta})$. Instead, we must compare in some more technical sense, which can again be challenging.

Proposition 49. ***Asymptotic Cramer-Rao Bound***. *Under some general conditions which we will not specify, it can be shown that*

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow{D} N(0, \mathcal{I}^{-1}(\theta))$$

where $\mathcal{I}(\theta) = \lim_{n\to\infty} \frac{1}{n}\mathcal{I}_n(\theta)$.

Let $\tilde{\theta}$ be any other estimator such that $\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{D} N(0, Var(\theta))$.

Then, for all θ, $Var(\theta) \geq \mathcal{I}^{-1}(\theta)$. The result implies that asymptotically, the MLE is efficient.

Just as with Cramer-Rao holding in multiple dimensions, the Asymptotic Cramer Rao Bound also holds in the multivariate case with a more technical argument.

Consider an example in the multivariate case where $X \sim N(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)^T$.

The information matrix for vector θ is the diagonal matrix $\mathcal{I}_n(\theta) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$.

Note that when we have a normal distribution, the mean and variance are independent and so covariance terms are zero. However, this is not a general result.

The MLE of μ is best unbiased according to the mean-squared error but this is not the case with σ^2. However, as with the univariate case, the MLE is asymptotically efficient.

Similarly, suppose $X \sim N(\mu, \Sigma)$ where covariance matrix Σ is known.

Then, $E(\hat{\mu}) = \mu_0$ and $Var(\hat{\mu}) = E[(\hat{\mu} - E(\hat{\mu}))(\hat{\mu} - E(\hat{\mu}))^T] = \frac{\Sigma}{n}$.

Furthermore, the information matrix for μ is $\mathcal{I}_{n\mu\mu}(\mu, \Sigma) = n\Sigma^{-1}$.

The MLE of μ is best unbiased according to matrix mean-squared error.

2.4 Hypothesis Testing

Consider a parametric model $\{P_\theta, \theta \in \Theta\}$ where Θ is the parameter space.

We can evaluate whether the data we have is compatible with certain preconceived **hypotheses**.

Definition 34. *One reduction of Θ, namely $\Theta_0 \subseteq \Theta$, is the* ***null hypothesis*** *H_0.*

The ***alternative hypothesis*** *H_1 is the complement $\Theta_1 \subseteq \Theta$ such that $\Theta_0 \cap \Theta_1 = \varnothing$ and $\Theta_0 \cup \Theta_1 = \Theta$.*

In hypothesis testing, we distinguish between a **simple hypothesis** in which the data distribution under the hypothesis is completely specified, and a **composite hypothesis** in which the hypothesis does not completely define the distribution due to some **nuisance parameters** not being specified.

We may also distinguish between a single hypothesis which has one restriction and multiple hypotheses where we have multiple restrictions on parameters.

Regarding the alternative, we distinguish between a one-sided and two-sided alternative.

For instance, if we suppose $X \sim N(\mu, 1)$ and $H_0 : \mu = \mu_0$ is a simple single null hypothesis, then $H_1 : \mu \neq \mu_0$ is a composite two-sided alternative.

Also, if $X \sim N(\mu, \sigma^2)$, then $H_0 : \mu = \mu_0, \sigma^2 = \sigma_0^2$ is a simple multiple null hypothesis.

As an example of a nuisance parameter, we might have $X \sim N(\mu, \sigma^2)$ with $H_0 : \mu = \mu_0$ being a composite null because σ^2 is unrestricted.

$H_1 : \mu \neq \mu_0$ is an example of a composite two-sided alternative while $H_1 : \mu > \mu_0$ is an example of a composite one-sided alternative.

A **hypothesis test** is a decision rule $\phi : \mathcal{X}^n \to \{Accept, Reject\}$ which requires us to partition the data into two regions.

The intermediate step involves computing a **test statistic** $T(\mathcal{X}^n, \theta) \in \mathbb{R}$ that can measure the consistency with H_0.

The **logic** of the test should be such that large values of T are incompatible with the null whereas small values are compatible.

To determine the size of T before rejection, we make two considerations. First, the **significance level** $\alpha \in (0, 1)$ and second, the **critical region** R_α which is the complement of the interval $[\underline{c_\alpha}, \bar{c_\alpha}]$ with the limits being called **critical values**.

The test is such that if $T \in R_\alpha$, then reject the null hypothesis. If $T \notin R_\alpha$, then do not reject the null hypothesis.

We could choose the critical region in two ways.

First, we could have an **exact** critical region where $\mathcal{P}(T \in R_\alpha | H_0 \ true) = \alpha$.

Second, we could have an **approximate** critical region where $\mathcal{P}(T \in R_\alpha | H_0 \ true) \to \alpha$ as $n \to \infty$.

In the latter case, considering asymptotic definitions makes sense.

Definition 35. *A test based on statistic T is consistent if both $\mathcal{P}(T \in R_\alpha | H_0 \ true) \to \alpha$ and $\mathcal{P}(T \in R_\alpha | H_0 \ false) \to 1$.*

When the null hypothesis is true, the rule yields a rejection decision α proportion of times (at least approximately) and so if the test were applied to many datasets generated by H_0, α proportion of the time the null would be rejected.

We choose α such that we are trying to measure the standard of evidence against the null that leads us to reject it.

The **p-value** is the marginal significance level at which we are indifferent between acceptance and rejection. We want to choose a T that satisfies the logic of the test and allows us to compute $\mathcal{P}(T \in R_\alpha | H_0)$. A formal definition of the p-value is below.

Definition 36. *Suppose T is a given test and T_{obs} is an observation. Then, the one-sided p-value is given by $\mathcal{P}(T \geq T_{obs} : H_0 \ true)$.*

The two-sided p-value is given by computing the probability for $|T| \geq |T_{obs}|$. A low p-value indicates evidence against the null hypothesis.

Proposition 50. *If a test statistic T has a continuous distribution $F(t) \equiv \mathcal{P}(T < t)$ for all t under the null where F^{-1} exists, the p-value has a uniform distribution on $[0, 1]$ under the null. Under the alternative, the p-value converges in probability to 0.*

Proof. Let α_{obs} be the p-value. We have $\mathcal{P}(\alpha_{obs} < p) = \mathcal{P}(F(T) < p)$ since α_{obs} is a function of the test statistic.

Hence, taking inverses gives $\mathcal{P}(F(T) < p) = \mathcal{P}(T < F^{-1}(p)) = F(F^{-1}(p)) = p$ which gives $\alpha_{obs} \sim U[0,1]$. □

Suppose that we have J test statistics each with some known distribution under the null. We expect that αJ tests will reject on average if the null is true.

This is called the **multiple testing problem**. One way to combine evidence across tests is to let α_{obs} be an arbitrary p-value and $\alpha_{obs}(j)$ be the p-value associated with the jth test. Under the null, this is uniform on $[0,1]$.

If the J tests are independent, we have that $\hat{T} = -2\sum_{j=1}^{J} log(\alpha_{obs}(j)) \sim \chi^2(2J)$, where large values of $\hat{T}$ are evidence against the null hypothesis. This allows us to combine evidence from different studies.

As an example of a single restriction test, suppose $\mathcal{X}^n$ is independent and identically distributed normal with mean μ and variance σ^2 with known σ^2.

Suppose $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$. Then, $T = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ under H_0. Let $R_\alpha = \{x : |x| \geq z_{\alpha/2}\}$.

The rule is to reject $T \in R_\alpha$ or equivalently $|T| \geq z_{\alpha/2}$ where $\Phi(z_\alpha) = 1 - \alpha$.

Critical values are $\pm z_{\alpha/2}$. If the alternative is one-sided, then reject if $T > z_\alpha$ i.e. critical value is z_α.

Suppose now that σ^2 is unknown. Then, we use the **t-test** $T = \dfrac{\bar{X} - \mu_0}{s_*/\sqrt{n}} \sim t(n-1)$ where we reject if $|T| > t_{\alpha/2}(n-1)$ or $T > t_\alpha(n-1)$.

The test is asymptotically valid since $T = \dfrac{\bar{X} - \mu_0}{s_*/\sqrt{n}} \xrightarrow{D} N(0,1)$.

As an example of a multiple restriction test, suppose we are interested in whether Zodiac sign affects income at age 40. We have a sample of n_j individuals for all $j = 1, ..., 12$ and $X_j \sim N(\mu_j, \sigma_0^2)$ where σ_0^2 is known but μ_j are not.

In this case, we might be interested in testing the null $H_0 : \mu_1 = ... = \mu_{12} = \mu_0$ where μ_0 is known versus the alternative that the μ_j are not all the same.

If we have sample sizes n_j, then a simple joint test statistic is $\sum_{j=1}^{12}\left(\frac{\bar{X}_j-\mu_0}{\sigma_0/\sqrt{n_j}}\right)^2 = \sum_{j=1}^{12} n_j\left(\frac{\bar{X}_j-\mu_0}{\sigma_0}\right)^2 \sim \chi^2_{12}$.

If we don't know σ_0^2, we need to estimate it and our test of multiple restrictions becomes an **F-test**.

Asymptotically, this F-statistic would converge in distribution to a $\chi^2(12)$ assuming each of the 12 restrictions is independent.

Under the alternative, the F-statistic would converge in probability to ∞ and so this test would be consistent.

Hypothesis testing makes two types of error. First, a **Type I error** occurs where we reject H_0 when H_0 is true.

A **Type II error** occurs where we accept H_0 when it is false.

Note that $\alpha = \mathcal{P}(Type\ \ I\ \ error) = \mathcal{P}(Reject\ \ H_0|H_0\ \ true)$ and $\beta = \mathcal{P}(Type\ \ II\ \ error) = \mathcal{P}(Accept\ \ H_0|H_1\ \ true)$. There is generally a trade-off between α and β.

α is the **size** or significance level of the test whereas $\pi = 1-\beta$ is called the **power** of the test. Ideally, we want high power in any hypothesis test.

Suppose that $X \sim N(\mu, 1)$ and $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$.

We have for all μ that $T = \sqrt{n}(\bar{X}-\mu_0) = \sqrt{n}(\bar{X}-\mu) + \sqrt{n}(\mu-\mu_0) \sim N(\sqrt{n}(\mu-\mu_0), 1)$.

Note that under the null, the distribution is standard normal, and under the alternative, there is convergence in probability to ∞.

Therefore, we reject if $T > z_\alpha$ and $\pi = \mathcal{P}(T \geq z_\alpha|\mu) = \mathcal{P}(N(0,1) + \sqrt{n}(\mu-\mu_0) \geq z_\alpha) = 1 - \Phi(z_\alpha - \sqrt{n}(\mu-\mu_0))$.

This is clearly an asymptotically consistent test since as under the alternative, as $n \to \infty$, $\pi \to 1$.

When $\mu > \mu_0$, then $\sqrt{n}(\mu-\mu_0) > 0$ and $\pi > \alpha$. The larger $\sqrt{n}(\mu-\mu_0)$, the larger the power of the test.

Definition 37. *Suppose we have null $\theta \in \Theta_0$ and alternative $\theta \in \Theta_1 = \Theta_0^c$. The **likelihood ratio***

statistic *is*

$$\lambda(\mathcal{X}^n) = \frac{\max_{\theta\in\Theta} L(\theta|\mathcal{X}^n)}{\max_{\theta\in\Theta_0} L(\theta|\mathcal{X}^n)} = \frac{L(\hat{\theta}_{MLE}|\mathcal{X}^n)}{L(\hat{\theta}_{RMLE}|\mathcal{X}^n)}$$

where $\hat{\theta}_{MLE}$ is the MLE over Θ and $\hat{\theta}_{RMLE}$ is the restricted MLE over Θ_0.

Small values of λ point to the veracity of the null while large values go against it.

In this definition, $\lambda \geq 1$. However, we can also define it upside down or with $\max_{\theta\in\Theta}$ replaced by $\max_{\theta\in\Theta_1}$. This is defined for any number of parameter restrictions.

For instance, let $X \sim N(\mu, 1)$ and $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Thus, the denominator is

$$L(\mu_0|\mathcal{X}^n) = \frac{1}{(2\pi)^{n/2}} \exp\left(-0.5\sum_{i=1}^{n}(X_i - \mu_0)^2\right)$$

The MLE $\hat{\mu}_{MLE} = \bar{X}$, and so we have

$$\lambda(\mathcal{X}^n) = \frac{(2\pi)^{-n/2}\exp\left(-0.5\sum_{i=1}^{n}(X_i - \bar{X})^2\right)}{(2\pi)^{-n/2}\exp\left(-0.5\sum_{i=1}^{n}(X_i - \mu_0)^2\right)} = \exp\left(0.5\sum_{i=1}^{n}(X_i - \mu_0)^2 - 0.5\sum_{i=1}^{n}(X_i - \bar{X})^2\right)$$

with $2\log\lambda(\mathcal{X}^n) = n(\bar{X} - \mu_0)^2$.

A Likelihood Ratio test is any test that rejects when $\lambda \geq c$ for some $c \geq 1$ i.e. when we have a rejection region $R^{\lambda}_{\alpha} = \{\lambda : \lambda \geq c\}$.

The key issue is to find the c to control the significance level.

Monotonic transformations yield the same test. For instance, $\lambda \geq c \iff 2log\lambda \geq 2log(c) \iff \sqrt{n}|\bar{X} - \mu_0| \geq \sqrt{2log(c)}$ for any c.

We know that for $T = \sqrt{n}|\bar{X} - \mu_0|$, $R^T_\alpha = T : T \geq z_{\alpha/2}$. Therefore, for λ, we can take $c = \exp(0.5z^2_{\alpha/2})$ which gives exactly the same region.

Optimally, we would like to simultaneously minimise α and maximise π but they are generally conflicting. Hence, we often first specify α and find the test with the highest π given α.

Consider the class of all tests with size α. We seek to find the test for which $\pi(\theta)$ is uniformly maximised over Θ_1, called the **Uniformly Most Powerful** test. Such a test need not exist in general.

Proposition 51. ***Neyman-Pearson Lemma***. *Assume special case $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, i.e. $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$.*

Suppose there exists a likelihood ratio region $R^*_\alpha = \{x : \lambda(x) \geq c\}$ such that $\mathcal{P}(X \in R^*_\alpha|\theta_0) = \alpha$.

Let R_α be any critical region with $\mathcal{P}(X \in R_\alpha|\theta_0) = \alpha$.

Then, $\mathcal{P}(X \in R^*_\alpha|\theta_1) \geq \mathcal{P}(X \in R_\alpha|\theta_1)$.

That is, a test based on the likelihood ratio has better power.

Definition 38. ***General Linear Hypothesis***. *Suppose $\theta \in \mathbb{R}^p$ and the null is $R\theta = r$ where R is a $q \times p$ matrix and r is a $q \times 1$ vector with $q \leq p$.*

If $q < p$, we have a composite null i.e. not all of θ is restricted.

The alternative is two-sided where $R\theta \neq r$.

As an example, suppose $X_i \sim N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)^T$. The null $\mu = \mu_0$ can be written $R\theta = r$ where $r = \mu_0$ and $R = (1, 0)$. The variance here is clearly unrestricted.

Alternatively, we can return to the Zodiac example where the null is now $H_0 : \mu_1 = ... = \mu_{12} = \mu$ where μ is unknown.

This can be written as $\mu_1 - \mu_2 = 0$, $\mu_2 - \mu_3 = 0$,..., $\mu_{11} - \mu_{12} = 0$ i.e. 11 restrictions since we are not imposing any further restriction on μ (if we were, we would have 12 restrictions).

In matrix form, taking account of the fact that $\sigma^2_1, ..., \sigma^2_{12}$ are unrestricted, we have $R\theta = r$ with R being an 11×24 matrix (although the right half of the matrix is an 11×12 zero matrix because the variances are unrestricted), θ being a 24×1 vector and r being a zero-vector in $\mathbb{R}^{11}$.

Definition 39. ***Wald Test***. *In the case of a single parameter restriction, the Wald statistic $W = \dfrac{(\hat{\theta} - \theta_0)^2}{Var(\hat{\theta})}$ where $\hat{\theta}$ is the sample MLE.*

In the case of multiple parameter restrictions, let $\hat{\theta}_n$ be the sample MLE (a vector) which follows $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, I^{-1}(\theta))$. Hence, $W = (R\hat{\theta} - r)^T (R I_n^{-1}(\hat{\theta}) R^T)^{-1} (R\hat{\theta} - r)$.

Notice that $W \geq 0$. The logic here is that if the null is true then $R\hat{\theta} - r$ should be small (or in the univariate case $\hat{\theta} - \theta_0$ should be small) and so a larger Wald statistic provides more evidence of the null being false.

Definition 40. ***Lagrange Multiplier Test****. In the single restriction case, the Lagrange multiplier statistic* $LM = \left(\dfrac{\partial \mathcal{L}}{\partial \theta}\right)^2 I_n^{-1}(\tilde{\theta})$ *where* $\tilde{\theta}$ *is the MLE under the restriction* $R\theta = r$.

In the multiple restriction case, we have $LM = \dfrac{\partial \mathcal{L}(\tilde{\theta})}{\partial \theta^T} I_n^{-1}(\tilde{\theta}) \dfrac{\partial \mathcal{L}(\tilde{\theta})}{\partial \theta}$ *where* $\tilde{\theta}$ *is the MLE vector under the restriction* $R\theta = r$.

Recall that at the unrestricted MLE, we have that the score function is exactly zero.

The logic of the Lagrange multiplier test is that if the null is true, then the score function at the restricted MLE should be close to zero.

We have in general that at the true parameter value θ, $\sqrt{n}\dfrac{\partial \mathcal{L}(\theta)}{\partial \theta} \xrightarrow{D} N(0, I(\theta))$.

Definition 41. ***Likelihood Ratio Test****. The likelihood ratio statistic* $LR = 2(\mathcal{L}(\hat{\theta}) - \mathcal{L}(\tilde{\theta}))$.

The logic is that if the null hypothesis is true, then the log likelihood evaluated at the unrestricted MLE is close to the log likelihood evaluated at the MLE restricted by the general linear hypothesis.

Consider the following example where $X_i \sim B(p)$ with $\theta = p$.

The null is $p = p_0$ and in this case, $I_n(\theta) = \dfrac{n}{p(1-p)}$.

We can deduce that $W = \dfrac{n(\bar{X} - p_0)^2}{\bar{X}(1 - \bar{X})}$ where $\bar{X}$ is the sample mean which is the MLE.

We can also show that $LM = \dfrac{n(\bar{X} - p_0)^2}{p_0(1 - p_0)}$ and $LR = 2n\left(\dfrac{\bar{X}log(\bar{X})}{p_0} + \dfrac{(1-\bar{X})log(1-\bar{X})}{1-p_0}\right)$.

For a multiple restrictions example, we have $X_i \sim N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)^T$ and null $\mu = \mu_0$. In this case, recall that we have $\dfrac{\partial \mathcal{L}}{\partial \mu}(\mu|\mathcal{X}^n) = -\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)$ and $I_n(\theta) = \begin{pmatrix} \dfrac{n}{\sigma^2} & 0 \\ 0 & \dfrac{n}{2\sigma^4} \end{pmatrix}$.

Therefore, we have that $W = \dfrac{n(\bar{X} - \mu_0)^2}{s^2}$ where $s^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Also, $LM = \dfrac{n(\bar{X}-\mu_0)^2}{\tilde{s}^2}$ where $\tilde{s}^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)^2$, and $LR = n(log(\tilde{s}^2) - log(\tilde{s}^2))$.

In fact, the three tests are monotonic transformations of one another.

Proposition 52. *Under some regularity conditions, under H_0, $LR, W, LM \xrightarrow{D} \chi^2(q)$ with q restrictions. Under H_1, $LR, W, LM \xrightarrow{P} \infty$.*

The tests are all consistent and asymptotically equivalent.

However, observe that the Wald test only requires estimation under the alternative whereas the Lagrange Multiplier test requires only estimation under the null.

The Likelihood Ratio test requires both but recall that by the Neyman-Pearson Lemma, when we have well-defined exact hypotheses, the likelihood ratio test has the highest power.

2.5 Confidence Intervals

One way to summarise information about the precision of an estimator $\hat{\theta}$ of θ is a **confidence set** $\mathcal{C}_\alpha(\mathcal{X}^n) \subseteq \Theta \subseteq \mathbb{R}^p$.

This is chosen either such that it is exact in the sense that $\mathcal{P}(\theta \in \mathcal{C}_\alpha(\mathcal{X}^n)) = 1-\alpha$ or approximate such that $\mathcal{P}(\theta \in \mathcal{C}_\alpha(\mathcal{X}^n)) \to 1$.

We describe the coverage of the set as $1-\alpha$. That is, if the size of the test is $\alpha \approx 0.05$, then the set has 95% coverage.

For a univariate parameter $\mathcal{C}_\alpha(\mathcal{X}^n)$ is called a **confidence interval** $[L(\mathcal{X}^n), U(\mathcal{X}^n)] \subset \Theta \subset \mathbb{R}$

The smaller the interval, the more confidence we have in the estimate.

We often take equal tail symmetric intervals since they usually generate the smallest confidence interval. This is not a general property, but does hold for most practical examples.

For example, suppose $X \sim N(\mu, 1)$ and the coverage $1-\alpha$ interval $\mathcal{C}_\alpha(\mathcal{X}^n) = \{\bar{X} - \frac{z_{\alpha_1}}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha_1}}{\sqrt{n}}\}$ where $\alpha_1 + \alpha_2 = \alpha$.

Let $\alpha_2 = \theta$ and $\alpha_1 = \alpha - \theta$ so we consider minimising the length of the interval which is equivalent to saying $\min_{\theta \in [0,\alpha]} Q(\theta) = \Phi^{-1}(1-\theta) - \Phi^{-1}(\alpha - \theta)$.

This has first-order condition $\frac{1}{\phi(\Phi^{-1}(1-\theta))} = \frac{1}{\phi(\Phi^{-1}(\alpha-\theta))} = 0$ that is solved when $\theta = 0.5\alpha$. That is, when $\Phi^{-1}(1-\alpha/2) = \Phi^{-1}(\alpha/2)$.

A general method for constructing confidence intervals is based on some pivotal statistic $T(\mathcal{X}^n, \theta)$ whose distribution is known exactly or approximately and does not depend on θ.

We let $\mathcal{C}_\alpha(\mathcal{X}^n) = \{\theta : I(\mathcal{X}^n) \leq T(\mathcal{X}^n, \theta) \leq u(\mathcal{X}^n)\}$ where I, u are such that either (exactly) $\mathcal{P}(I(\mathcal{X}^n) \leq T(\mathcal{X}^n, \theta) \leq u(\mathcal{X}^n)) = 1-\alpha$ or (approximately) $\mathcal{P}(I(\mathcal{X}^n) \leq T(\mathcal{X}^n, \theta) \leq u(\mathcal{X}^n)) \to 1-\alpha$ with the probability being calculated under θ.

For example, suppose $X \sim N(\mu, \sigma^2)$ with known σ^2. Then, $T = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

It follows that a two-sided coverage $1-\alpha$ confidence interval for μ can be expressed as $\mathcal{C}_\alpha(\mathcal{X}^n) = \{\mu : -z_{\alpha/2} \leq T \leq z_{\alpha/2}\} = \{\mu : -z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\}$.

We can rewrite this interval as $\mathcal{C}_\alpha(\mathcal{X}^n) = \{\bar{X} \pm \frac{\sigma z_{\alpha/2}}{\sqrt{n}}\}$.

There is an intimate relationship between the confidence interval and the acceptance region of a hypothesis test.

The confidence interval is the region of the parameter space determined by the data that contains the true parameter with given probability.

The acceptance region is the region of the data determined under the null that contains the test statistic with given probability.

The acceptance region of a hypothesis test using null hypothesis $\mu = \mu_0$ against two-sided alternative using $T = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ is $\{T : -z_{\alpha/2} \leq T \leq z_{\alpha/2}\}$ which can be rewritten for $\bar{X}$ as $\mathcal{C}_\alpha(\mu_0) = R^c_\alpha(\mu_0) = \left\{x : x \in \left(\mu_0 - \frac{\sigma z_{\alpha/2}}{\sqrt{n}}, \mu_0 + \frac{\sigma z_{\alpha/2}}{\sqrt{n}}\right)\right\} = \mu_0 \pm \frac{\sigma z_{\alpha/2}}{\sqrt{n}}$.

We accept when $\bar{X} \in \mathcal{C}_\alpha(\mu_0)$ and we have for any $\mu_0 \in \mathcal{C}_\alpha(\mathcal{X}^n)$ that $\bar{X} \in \mathcal{C}_\alpha(\mu_0)$.

Suppose that $X \sim N(\mu, \sigma^2)$ with unknown σ^2.

Then, a two-sided coverage $1-\alpha$ confidence interval for μ is given by $\mathcal{C}_\alpha(\mathcal{X}^n) = \bar{X} \pm \frac{s t_{\alpha/2}(n-1)}{\sqrt{n}}$.

The acceptance region of the test of $\mu = \mu_0$ is the set where $\bar{X}$ lies i.e. $\mathcal{C}_\alpha(\mu_0) = \mu_0 \pm \dfrac{st_{\alpha/2}(n-1)}{\sqrt{n}}$.

A two-sided asymptotic coverage $1-\alpha$ confidence interval for μ is given by $\mathcal{C}_\alpha(\mathcal{X}^n) = \bar{X} \pm \dfrac{sz_{\alpha/2}}{\sqrt{n}}$.

We can consider asymptotics of confidence intervals.

In general, $T_n = \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, V(\theta, \phi))$ where $V(\theta, \phi)$ is a known function of the unknown θ and nuisance parameters ϕ.

Provided V is continuous in θ and ϕ, and we have a consistent estimator $\hat{\phi} \xrightarrow{P} \phi$, then $\dfrac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{V(\hat{\theta}, \hat{\phi})}} \xrightarrow{D} N(0,1)$ and so $\mathcal{C}_\alpha(\mathcal{X}^n) = \hat{\theta} \pm \dfrac{z_{\alpha/2}\sqrt{V(\hat{\theta}, \hat{\phi})}}{\sqrt{n}}$ shrinks with sample size.

We can generate confidence intervals for likelihood ratios for the simple null of $\theta = \theta_0$.

The likelihood ratio statistic is $\lambda(\mathcal{X}^n, \theta_0) = \dfrac{\max_{\theta \in \Theta} L(\theta|\mathcal{X}^n)}{L(\theta_0|\mathcal{X}^n)}$.

A likelihood ratio test is any test that accepts when $\lambda(\mathcal{X}^n) \leq c_\alpha$ for some determined c_α such that $\mathcal{C}_\alpha(\lambda) = \{\lambda : \lambda \leq c_\alpha\}$.

The confidence interval is $\mathcal{C}_\alpha(\mathcal{X}^n) = \{\theta_0 : \lambda(\mathcal{X}^n, \theta_0) \leq c_\alpha\}$.

Suppose $X \sim N(\mu, 1)$ where $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. In this case, we have that $2\log\lambda \sim \chi^2_1$. Therefore, the confidence interval is $\{\mu_0 : n(\bar{X} - \mu_0)^2 \leq \chi^2_\alpha(1)\}$.

Rearranging for μ_0 gives $\left\{\mu_0 : \bar{X} - \dfrac{z_{\alpha/2}}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + \dfrac{z_{\alpha/2}}{\sqrt{n}}\right\}$.

As another example, suppose $X \sim B(p)$. We are interested in finding an interval for binary proportion p.

We can try using the Wald statistic for $H_0 : p = p_0$ and $H_1 : p \neq p_0$. We have $T = \dfrac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{\hat{p}(1 - \hat{p}}}$ where $\hat{p} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$. We know that under H_0, $T \xrightarrow{D} N(0,1)$.

It follows that a two-sided coverage $1-\alpha$ confidence interval for p can be expressed as $\mathcal{C}_\alpha(\mathcal{X}^n) = \{p : -z_{\alpha/2} \leq T \leq z_{\alpha/2}\}$ which can be rewritten as $\hat{p} \pm \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} z_{\alpha/2}$.

This confidence interval works except when $\hat{p} = 0$ or $\hat{p} = 1$, which can happen when the sample size is small and p_0 is very close to 0 or 1.

As an alternative, we could consider a Lagrange multiplier test statistic $T = \dfrac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1 - p_0)}}$ which is also asymptotically standard normal under the null hypothesis.

The corresponding confidence interval is $\mathcal{C}_\alpha(\mathcal{X}^n) = \{p_0 : -z_{\alpha/2} \leq T \leq z_{\alpha/2}\}$. Such a confidence interval works also when $\hat{p} = 0$ or $\hat{p} = 1$.

Bayesian statistics has an analogue of confidence intervals known as a **Bayesian interval**.

If we observe k values of $X = 1$ and $n - k$ values of $X = 0$, then this gives us a likelihood of the sample (distribution of the data) to be $f(\mathcal{X}^n|p) = \begin{pmatrix} n \\ k \end{pmatrix} p^k(1 - p)^{n-k}$.

Suppose the prior distribution $\pi(p) \sim U[0, 1]$.

In this case, the posterior density of p is $\pi(p|\mathcal{X}^n) = \dfrac{f(\mathcal{X}^n|p)\pi(p)}{\displaystyle\int f(\mathcal{X}^n|p)\pi(p)dp} = \dfrac{(n + 1)!}{k!\,(n - k)!}p^k(1-p)^{n-k}$.

A Bayesian interval $[L, U]$ satisfies $\displaystyle\int_L^U \pi(p|\mathcal{X}^n)dp = 1 - \alpha$.

Simulation methods might be needed to calculate such an integral.

Consider the special case that $k = 0$ so that $\hat{p} = 0$ in which case we have $(n+1)\displaystyle\int_0^U (1-p)^n dp = 1 - \alpha = (1 - (1 - U)^{n+1})$ and therefore $U = 1 - \alpha^{1/n+1}$.

When we take significance level $\alpha = 0.05$ and $n = 10$, we obtain $U = 0.238$ and so with 95% confidence, we can assert the random set $[0, 0.238]$ contains p.

2.6 Simulation Methods

Suppose X_i is independent and identically distributed according to some distribution $F(.|\theta)$ and we want to calculate the distribution of some statistic $T(\mathcal{X}^n)$ which is complicated. This means that we are looking for $H_n(x, \theta) = \mathcal{P}(T(\mathcal{X}^n \leq x)$. Also, suppose n is small so we cannot rely on asymptotic approximations. This means we must resort to computational methods.

Now, in the case that θ is known, we can do the following.

First, generate a sample from the distribution using a random number generator, and compute

$T(\mathcal{X}_*^n)$ from the sample $\mathcal{X}_*^n$.

Repeat this S times and calculate the empirical distribution of $\{T_1^*, ..., T_S^*\}$ and use this distribution in place of $H_n(x, \theta)$ i.e. the true distribution.

Hence, compute $\hat{H}_S(x, \theta) = \dfrac{1}{S}\displaystyle\sum_{s=1}^{S} \mathbf{1}(T_S^* \leq x)$.

By the Law of Large Numbers, we have $\hat{H}_S(x, \theta) \xrightarrow{P} H_n(x, \theta)$ for every n and every x, giving us an exact asymptotic distribution.

Now, suppose we have no idea of the parameter of the distribution θ. Then, we might approach the problem by estimating θ from the data by $\hat{\theta}$.

Then, we generate a sample $\mathcal{X}_*^n$ from $F(.|\hat{\theta})$ using a random number generator.

Once again, we compute $T(\mathcal{X}_*^n)$ and repeat S times and calculate the empirical distribution of $\{T_1^*, ..., T_S^*\}$.

Theoretically, this approach works but requires a large sample size.

As an example, suppose $X \sim N(\mu, \sigma^2)$ independently and with identical distribution and let $T = \bar{X} \sim N(\mu, \sigma^2/n)$.

Then, we may estimate μ by $\bar{X}$ and σ^2 by s^2, generating random variables X_i^* from $N(\bar{X}, s^2)$. Conditional on the data, the distribution of $T^* = \bar{X}^*$ is $N(\bar{X}, s^2/n)$ which is approximately $N(\mu, \sigma^2/n)$ when n is large.

Next, consider a more general setting where $X_1, ..., X_n$ are independent and identically distributed with distribution function F that is completely unknown.

The **Bootstrap** is a method of conducting inference from such a distribution. Computationally, this method is preferred to that which was discussed above. The bootstrap works in a large variety of situations - we will just examine the very simplest ones.

We have a statistic $T(\mathcal{X}^n, \theta)$ which is a function of the data $X_1, ..., X_n$ and parameter value θ that may be estimated by $\hat{\theta}$.

We are looking to find $H_n(x, \theta) = \mathcal{P}(T(\mathcal{X}^n, \theta) \leq x)$.

But now since we have no clue of θ, the Bootstrap principle says that we should treat the empirical distribution $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x)$ as the population and then to sample from this population which is now known.

Definition 42. ***Bootstrap Algorithm***. *This method works in theory if the sample size is large.*

1. *Generate sample $\mathcal{X}_*^n$ from empirical distribution F_n (drawn with replacement from $\{X_1, ..., X_n\}$).*
2. *Compute $T(\mathcal{X}_*^n, \hat{\theta})$ where $\hat{\theta}$ is the sample estimate.*
3. *Repeat S times.*
4. *Calculate the empirical distribution of $\{T_1^*, ..., T_S^*\}$ and use this in place of $H_n(x, \theta)$. That is, $\hat{H}_B(x, \hat{\theta}) = \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}(T_s^* \leq x)$.*
5. *Critical values can be calculated as quantiles e.g. $\hat{H}_B^{-1}(\alpha, \hat{\theta})$.*

Suppose X_i are independent and identically distributed and $T = \sqrt{n}(\bar{X} - \mu)$ is the statistic of interest where $T = \sqrt{n}(\bar{X} - \mu) \xrightarrow{D} N(0, \sigma^2)$.

We can calculate the first two conditional moments of the bootstrap statistic $\bar{X}^*$. First, observe that X_i^* are drawn independently from the following **multinomial distribution**

$$X_i^* = \begin{cases} X_1 & prob\ 1/n \\ & \vdots \\ X_n & prob\ 1/n \end{cases}$$

In particular, $E(X_i^*|\mathcal{X}^n) = \sum_{i=1}^{n} \frac{1}{n} X_i = \bar{X}$ and $Var(X_i^*|\mathcal{X}^n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

It follows that $E(\bar{X}^*|\mathcal{X}^n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^*|\mathcal{X}^n) = E(X_i^*|\mathcal{X}^n) = \bar{X}$ and that $Var(\bar{X}^*|\mathcal{X}^n) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i^*|\mathcal{X}^n) = \frac{1}{n}s^2$.

This argument says that the random variable $T^* = \sqrt{n}(\bar{X}^* - \bar{X})$ has conditional mean zero and conditional variance given by the sample variance of the original data.

When n is large, $s^2 \xrightarrow{P} \sigma^2$ and so conditional on the data, the asymptotic property is that $T^* = \sqrt{n}(\bar{X}^* - \bar{X}) \xrightarrow{D} N(0, \sigma^2)$.

This says that the distribution of T^* is close to that of T, and we can approximate the distribution of T^* by bootstrapping.

Proposition 53. *If bootstrapping is applied to a statistic that is asymptotically normal, it gives a good approximation to its distribution when the sample size is large.*

Suppose that $T(\mathcal{X}^n, \theta) \xrightarrow{D} N(0, V)$ where V might be very complicated.

That is, $H_n(x, \theta) \to H_A(x, V)$ where $H_A(x, V)$ is the limiting cdf (in this case $N(0, V)$).

The asymptotic approximation is based on $H_A(x, \hat{V})$ where $H_A(x, \hat{V}) \to H_A(x, V)$.

However, the bootstrap approximation is based on $\hat{H}_B(x, \hat{\theta})$ where $\hat{H}_B(x, \hat{\theta}) \to H_A(x, V)$.

We can use the Bootstrap procedure to obtain confidence intervals or critical values for tests. For a one-sided interval of coverage $1 - \alpha$, we have $\mathcal{C}_\alpha(\mathcal{X}^n) = \{\theta : T(\mathcal{X}^n, \theta) \leq \hat{H}_A^{-1}(1 - \alpha)\}$, which then ensures that $\mathcal{P}(\theta \in \mathcal{C}_\alpha(\mathcal{X}^n)) \to 1 - \alpha$.

When we use the Bootstrap, we simply replace the asymptotic critical value $\hat{H}_A^{-1}(\alpha)$ by the bootstrap critical value $\hat{H}_B^{-1}(\alpha)$ obtained from bootstrap samples of T^* such that $\mathcal{C}_\alpha^*(\mathcal{X}^n = \{\theta : T(\mathcal{X}^n, \theta) \leq \hat{H}_B^{-1}(1 - \alpha)\}$. Then, also we have that $\mathcal{P}(\theta \in \mathcal{C}_\alpha^*(\mathcal{X}^n) | \mathcal{X}^n) \to 1 - \alpha$.

For two-sided intervals of coverage $1-\alpha$, $\mathcal{C}_\alpha(\mathcal{X}^n) = \{\theta : \hat{H}_A^{-1}(\alpha/2) \leq T(\mathcal{X}^n, \theta) \leq \hat{H}_A^{-1}(1-\alpha/2)\}$.

Usually $\hat{H}_A(.)$ is symmetric about zero and $\hat{H}_A^{-1}(\alpha/2) = -\hat{H}_A^{-1}(1 - \alpha/2)$. For the bootstrap method, we have $\mathcal{C}_\alpha^*(\mathcal{X}^n) = \{\theta : \hat{H}_B^{-1}(\alpha/2) \leq T(\mathcal{X}^n, \theta) \leq \hat{H}_B^{-1}(1 - \alpha/2)\}$.

In this case, there is no reason to expect $\hat{H}_B(.)$ to be symmetric about zero so we typically calculate two separate critical values.

For example, suppose X is independent and identically distributed with mean μ and variance σ^2 where $T = \sqrt{n}(\bar{X} - \mu) \xrightarrow{D} N(0, \sigma^2)$. The bootstrap confidence interval based on $T = \sqrt{n}(\bar{X}^* - \bar{X})$ is $\left\{ \bar{X} - \dfrac{\hat{H}_B^{-1}(\alpha/2)}{\sqrt{n}}, \bar{X} + \dfrac{\hat{H}_B^{-1}(1 - \alpha/2)}{\sqrt{n}} \right\}$.

Sometimes, $H(x, F) = \lim_{n\to\infty} H_n(x, F)$ does not depend on F in which case we call T an **asymptotic pivot**.

In the case of a pivotal limit, the bootstrap achieves a **refinement**.

That is, the approximation error of the bootstrap is strictly smaller than the asymptotic approach, such that for any x, we have $\lim_{n\to\infty} \dfrac{|H_n(x) - H_B(x)|}{|H_n(x) - H_A(x)|} = 0$.

As an example, suppose X is independent and identically distributed with mean μ and variance σ^2 with $T = \dfrac{\sqrt{n}(\bar{X} - \mu)}{s} \xrightarrow{D} N(0,1)$.

In this case, the large sample confidence interval for μ is $\bar{X} \pm \dfrac{z_{\alpha/2}s}{\sqrt{n}}$. The bootstrap confidence interval is based on $\hat{H}_B$ of $T^* = \dfrac{\sqrt{n}(\bar{X}^* - \bar{X})}{s^*}$ and $(s^2)^* = \dfrac{1}{n}\sum_{i=1}^{n}(X_i^* - \bar{X}^*)^2$.

It is given by $[\bar{X} - \dfrac{\hat{H}_B^{-1}(\alpha/2)s}{\sqrt{n}}, \bar{X} + \dfrac{\hat{H}_B^{-1}(1-\alpha/2)s}{\sqrt{n}}]$.

In practice, we can calculate the quantile by ordering the bootstrap numbers and taking the corresponding order statistic i.e. choose from $T^*_{(1)} \leq T^*_{(2)} \leq ... \leq T^*_{(S)}$.

Generally, we need not choose $S \to \infty$.

Instead, we can take $\alpha(S+1)$ to be an integer so the relevant quantiles of the bootstrap distribution are uniquely defined.

For instance, $S = 199$ works for $\alpha = 0.01$ and $\alpha = 0.05$.