STAT3913 - Statistical Inference (Advanced)

A Students' Notes for Undergraduate Statistics

Jerry Xu

October 2018

https://github.com/jerry-ye-xu
jerryxu2500@gmail.com

Contents

1 Probability Theory

1.1 Introduction

In our world, there are actions we can take where the outcome cannot be predicted with certainty. Suppose we roll a normal dice. You cannot say for sure how it will land every time you throw it.

In statistics, we can attempt to quantity the amount of uncertainty in an action that is taken. If an action can be repeated under the same conditions and terminated with an outcome, we call this a statistical investigation. When such an investigation has an uncertainty in the outcome, it is called a random experiment.

In a **random experiment**, the collection of all possible outcomes is called the **sample space**, denoted by Ω.

Example:

Suppose you have 2 dice, one blue and one yellow. Rolling the dice results in an ordered pair (B_i, Y_i) where they are the i^{th} outcome for each dice respectively. In this experiment there are 36 possible outcomes.

Say we are interested in certain outcomes that lie within the sample space C. If we perform experiments and the outcome is in C, then we say **that the event C has occurred**.

If we perform the experiments n times, then we have the **frequency** f times that a particular event occurred. The ratio $\frac{f}{n}$ is called the **relative frequency**.

Relative frequencies for a small n can sometimes truly depend on luck, but for a purely random experiment the relative frequency will stabilise as n increases. We denote p as being approximately equal (or in some cases equal) to the relative frequency in the long run. In probability, we cannot predict each individual event with certainty, but over repeated experiments we can begin to see patterns of an event occurring. The symbol p is called the **probability of the event C**.

Thus our motivation is clear - we want to create mathematical models to determine the stabilising p for a random experiment. If we are able to do this, then we can make inferences about the investigation. We discuss inference much later.
Now dive very shallowly into set theory, which contains some cool results related to probability that you should know as well.

1.2 Set Theory

In **Naive Set Theory**, a set is considered a collection of distinct objects. These objects can be anything from letters to people, and the order in which they appear is not relevant. We usually write members of a set as

$$\mathcal{A} = \{\text{Red}, \text{Blue}, \text{Green}\}$$

When a mathematician by the name of Bertrand Russell discovered what is known as the **Russell's Paradox**, which you can read about here, people have gone on to, after more paradoxes were found, propose axioms and develop the field even further with more formal definitions of sets. However, as you can imagine, if we continued talking about set theory you'd be here for at least another 50 pages. So let's just keep it simple for now.

If you are really interested in this very important aspect of mathematics, I would recommend taking a math course. It would definitely help in your statistics as well!

1.2.1 Basic Results

In set theory, there are some common notation and basic results that you should (hopefully) know:

(i) If a set is empty, we call it the **null set** and denote it with ϕ

(ii) A set consisting all elements of $\mathit{\Omega}$ that are not elements belonging in another subset A is denoted by A^c. A^c is called the complement of A.

(iii) $A \cap B$ is called the **intersection** of A and B, where both events occur.

(iv) $A \cup B$ is called the **union** of A and B, where one of the 2 events occurs.

(v) 2 events A and B are **mutually exclusive** is denoted by $A \cap B = \phi$

(vi) 2 events A and B are **mutually exclusive and exhaustive** if $A \cap B = \phi$ and $A \cup B = \Omega$

(vii) The **Power set** of a sample space Ω is the set of all subsets including the null set and Ω itself.

(viii) Both union and intersection results satisfy **distributive rules** i.e. $A \cup (B \cap B) = (A \cup B) \cap (A \cup B)$ and vice versa.

1.2.2 De Morgan's Law

We look at an interesting result, the **De Morgan's** law, which states that:

For a set of events from the same sample space
(i) The complement of the union of n sets is equal to the intersection of their complements.

$$(\cup A_i)^c = \cap (A_i^c) \tag{1}$$

(ii) and vice versa; the complement of the intersection of n sets is equal

$$(\cap A_i)^c = \cup (A_i^c) \tag{2}$$

Let's prove this.

For part (i); suppose that we have some $x \in (\cup A_i)^c$. Then by definition, $x \notin (\cup A_i)$.

In words, this is saying that x is not in any of the events specified by A_i. Thus this would mean that x exists in every complement of A_i that doesn't include A_j where $i \neq j$.

Hence $x \in \cap (A_i^c)$, as required.

The result in part (ii) can also be proved similarly.

1.2.3 Probability Set Function

σ-Algebra

Here we introduce the notation of the **σ-Algebra** (or σ-Field), which ...
The σ-Algebra has certain properties, as follows;

(i) $\mathcal{F}$ is not a null set $\longrightarrow \phi \in \mathcal{F}$

(ii) $\mathcal{F}$ is closed under complements $\longrightarrow A \in \mathcal{F}$ then $A^c \in \mathcal{F}$

(iii) $\mathcal{F}$ is closed under **countable unions** $\longrightarrow A_1, A_2, ...\mathcal{F}$ gives $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$

Now we introduce the **Probability Set Function**.

$\mathit{\Omega}$ be a sample space and A be the set of events, with $\mathcal{P}$ a real-valued function defined on A. $\mathcal{P}$ is considered probability set function if $\mathcal{P}$ satisfies the following 3 conditions:

(i) $\mathcal{P}(A) \geq 0$ for all $A \in \mathit{\Omega}$

(ii) $\mathcal{P}(\mathit{\Omega}) = 1$

(iii) If $A_1, A_2, \ldots$ are mutually exclusive events such that $A_i \cap A_j =$ null for all $i \neq j$, then $\mathcal{P}\left(\cup_{i=1}^{\infty}\right) = \sum_{i=1}^{\infty} \mathcal{P}\left(A_i\right)$.

The triplet $(\mathit{\Omega}, \mathcal{F}, \mathcal{P})$ is called the **probability space** associated with $\mathit{\Omega}$.

Remember that the $\mathit{\Omega}$ is defined as a collection of all possible outcomes.

Important Properties of Probability

Let's do some intuitive ones first:

(i) Show that $P =$ null-zero where null-zero is an impossible or null event.

Let $A \in \mathit{\Omega}$, making $A \cap \phi$ mutually exclusive.

Hence $A \cup \phi = A \longrightarrow P(A) + P(\phi) = P(A)$, $P(\phi) = 0$.

(ii) For any $A \subset \mathit{\Omega}$, $0 \leq P(A) \leq P(\mathit{\Omega})$ show that $P(A^c) = 1 - P(A)$.

We know that A and A^c are mutually exclusive, and thus $P(A) + P(A^c) = 1$, which gives $P(A^c) = 1 - P(A)$.

(iii) For $B \subset A$ show that $P(B) \leq P(A)$.

From the above we know that $B \cup A = A$ and $B \cap A = B$.

Hence we know that $A = B \cup \{B^c \cap A\}$, giving

$$\begin{aligned} P(A) &= P(B) + P(B^c \cap A) \\ &\geq P(B) \ \ \textbf{since} \ P(B^c \cap A) \geq 0 \end{aligned} \tag{3}$$

(iv) **Addition Law**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$\begin{aligned} P(A \cup B) &= P(A \cap B) + P(A \cap B^c) + P(A^c \cap B) \\ &= P(A \cap B) + P(A) - P(A \cap B) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned} \tag{4}$$

The next 2 properties can prove using induction.

(v) **Boole's Inequality**: $P(\cup_{i=1}^{k} A_i \leq \sum_{i=1}^{k} P(A_i)$

for $k = 2$;

$$\begin{aligned} P\left(A_1 \cup A_2\right) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &\leq P(A_1) + P(A_2) \end{aligned} \tag{5}$$

assume $k = n$ is true;

$$P\left(\cup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i) \tag{6}$$

prove that $k = n + 1$ is true;

We have, assuming $k = n$ is true

$$\begin{aligned} P\left(\cup_{i=1}^{n} A_i\right) \cup P\left(A_{n+1}\right) &\leq \left[\sum_{i=1}^{n} P(A_i)\right] \cup P\left(A_{n+1}\right) \\ &\leq \left[\sum_{i=1}^{n+1} P(A_i)\right] - P\left[\sum_{i=1}^{n} P(A_j) \cap P(A_{n+1})\right] \\ &\leq \left[\sum_{i=1}^{n+1} P(A_i)\right] \end{aligned} \tag{7}$$

as required. Note that we are using $P(\sum(A_k) \cup A_{k+1}) = P(\sum A_k) + P(A_{k+1}) - P(\sum(A_k) \cap A_{k+1})$.

Boole's inequality provides an upper-bound for the union of events.

(vi) **Bonferroni's Inequality** $P\left(\cap_{i=1}^{k} A_i\right) \geq \sum_{i=1}^{k}(P(A_i)) - (k-1)$

for $k = 2$;

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1) + P(A_2) - P(A_1 \cup A_2^c) - P(A_1^c \cup A_2) \\ &\leq P(A_1) + P(A_2) \\ &\geq P(A_1) + P(A_2) - 1 \end{aligned} \tag{8}$$

If we use the inequality $1 \geq P(A_1) + P(A_2) - P(A_1 \cap P(A_2))$.

assume $k = n$ is true;

$$P\left(\cap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n}(P(A_i)) - (n-1) \tag{9}$$

prove that $k = n + 1$ is true;

$$\begin{aligned} P\left(\cap_{i=1}^{n}(A_i) \cap A_{n+1}\right) &\geq P\left(\cap_{i=1}^{n} A_i\right) + P(A_{n+1}) - 1 \ \text{ using } k = 2 \\ &\geq \sum_{i=1}^{n}(P(A_i)) - (n-1) + P(A_{n+1}) - 1 \\ &\geq \sum_{i=1}^{n+1}(P(A_i)) - n \end{aligned} \tag{10}$$

as required. In contrast, the Bonferroni inequality provides the lower-bound for the intersection of multiple events.

The most common application, which you might have already seen in multiple testing, is using the Bonferroni correction. This is conservative, but applicable when you cannot assume independence between the tests.

1.3 Conditional Probability

If we are interested in the probability of it raining tomorrow if it already rained today, then we are looking at **conditional probability**. Conditional probability is a measure of, say event A occurring given that another event B has already occurred.

If the events are independent, then whether B has occurred does not really matter and thus conditional probability does not apply.

We define the probability of an event A given B as:

$$\begin{aligned} P(A_1|A_2) &= \frac{P(A_1, A_2)}{P(A_2)} \\ &= \frac{P(A_2|A_1)P(A_1)}{P(A_2)} \end{aligned} \tag{11}$$

Rewriting the intersection of events, we have

$$P(A_1 \cap A_2 \cap A_3) = P(A_3|A_1 \cap A_2)P(A_2|A_1)P(A_1)$$

We can extend this case to n events:

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_n|A_1 \cap A_2 \cap .... \cap A_{n-1})P(A_{n-1}|A_2 \cap A_2 \cap ... \cap A_{n-2})...P(A_2|A_1)P(A_1) \tag{12}$$

This can be proved using induction.
However, in many cases conditional probability is incorrectly interpreted and can be quite tricky.

Very quickly, we say 2 events are independent if and only if

$$P(A_1 \cap A_2) = P(A_1)P(A_2)$$

Extending this to n events gives

$$P\left(\cap_{i=\mathcal{I}} A_i\right) = \prod_{i=\mathcal{I}} P(A_i) \tag{13}$$

for a set of events $\{A_i \mid i \in \mathcal{I}\}$

1.3.1 The Law of Total Probability

The **Law of Total Probability** draws a relationship between marginal and conditional probabilities. Suppose you are interested in an event A that can be broken down into mutually exclusive conditional events $B_1, B_2, ..., B_n$.

Summing these conditional events of gives us the total probability of A. In other words, we have

$$\begin{aligned} P(A \cap B_i) &= P(A \cap B_1) \cup P(A \cap B_2) \cup ... \cup P(A \cap B_n) \\ &= P(A|B_1) + P(A|B_2) + ... + P(A|B_n) \\ &= \sum_{i=1}^{n} P(A|B_i)P(B_i) \end{aligned} \tag{14}$$

Example:

(i) In a certain factory there are four machines L, Q, R and S producing springs of the same length. On each day, these machines produce 25, 30, 25 and 20 percent of the total production containing 2, 1, 2 and 3 percent of defective springs. If one spring is selected at random from the total production (on Monday), what is the probability that it is defective?

Let $P(A)$ be the probability that a spring is defective. $P(A|B)$ is then the probability that a spring is defective given a particular machine produced it. $P(B)$ is the probability that the spring is produced by a particular machine.

By using the Law of Total Probability

$$\begin{aligned} P(A) &= \sum_{i=1}^{n} P(A|B_i)P(B_i) \\ &= \frac{25}{100} \cdot \frac{2}{100} + \frac{30}{100} \cdot \frac{1}{100} + \frac{25}{100} \cdot \frac{2}{100} + \frac{20}{100} \cdot \frac{3}{100} \\ &= 1.9\% \end{aligned}$$

(ii) If the selected spring is defective, what is probability that it was produced by machine R?

$$\begin{aligned} P(B_R|A) &= \frac{P(A|B_R)P(B_R)}{P(A)} \\ &= \frac{0.02 \times 0.25}{0.019} \\ &= 0.263 \end{aligned}$$

1.4 Random Variables

You're going to be dealing with random variables for the rest of the course, and I assume you know enough things. I'm just going to define random variables here and more onto the next section.

By definition, a random variable X is a mapping of the sample space $\varOmega$ to a measurable space E; $X : \varOmega \to E$.

If S is a subset of E where $S \subseteq E$, then the probability that X takes on a value in set S is

$$P(X \in S) = P\left(\{w \in \varOmega \,|\, X(w) \in S\}\right) \tag{15}$$

Often, the measurable space is in the real-values $E = \mathbb{R}$. Let's close this part off with a quick example.

Example:

Suppose that X is a degenerate (discrete) random variable with PMF $P_X(x) = 1$, $x = x_0$ and zero otherwise. Find the CDF, mean, median and variance of X.

X is degenerate, which means that it has one possible value with $P(X = x_0) = 1$

$$P_X(x) = \begin{cases} 1 & \text{if } x = x_0 \\ 0 & \text{otherwise} \end{cases}$$

Hence we have (i) The CDF is given by

$$\sum_{x=0}^{\infty} P_X(x) = \begin{cases} 1 & \text{if } x \geq x_0 \\ 0 & \text{otherwise} \end{cases}$$

(ii) $E[X] = x_0$
(iii) $Var[X] = 0$
(iv) The median is x_0 given that we have $0.5 \leq \sum_{x=0}^{\infty} P_X(x) = P_X(x_0)$

1.5 Important Inequalities

1.5.1 Existence of k Moments

Let X be a random variable and $m \in \mathbb{Z}^+$, a positive integer. Suppose $E\left[X^m\right]$ exists. If $k \in \mathbb{Z}^+$ and $k \leq m$, then show that $E\left[X^k\right]$ exists.

We prove for the continuous case, and the discrete case is similar.

$$\begin{aligned}
\int_{-\infty}^{\infty} |x|^k f(x)\ \delta x &= \int_{|x|\leq 1} |x|^k f(x)\ \delta x + \int_{|x|>1} |x|^k f(x)\ \delta x \\
&\leq \int_{|x|\leq 1} f(x)\ \delta x + \int_{|x|>1} |x|^m f(x)\ \delta x \\
&\leq \int_{-\infty}^{\infty} f(x)\ \delta x + \int_{-\infty} \infty |x|^m f(x)\ \delta x \\
&\leq 1 + E\left[|X|^m\right] < \infty
\end{aligned}$$

as required.

Note: There is a slight difference between a moment existing and being finite.

1.5.2 Markov's Inequality

Markov's Inequality provides and probabilistic upper-bound for a **non-negative function** $g(X)$. This upper-bound is proportional to the expectation of the function, determined by a positive constant c.

$$P\left[g(X) \geq c\right] = \frac{E\left[g(X)\right]}{c} \tag{16}$$

Intuitively, this means that for any probability of a given $g(x)$, the greater the difference between the expected value of $g(X)$ and c, the less likely it will happen. This intuition is generalised by Chebyshev's inequality (which we will see later on), but first we give the proof for this inequality.

We use the definition of the expectation and the fact that we are looking for the probability of $g(X) \leq c$ to yield

$$\begin{aligned}
E\left[g(X)\right] &= \int_{-\infty}^{\infty} g(x)f(x)\ \delta x \\
&= \int_{0}^{c} g(x)f(x)\ \delta x + \int_{c}^{\infty} g(x)f(x)\ \delta x \\
&\geq \int_{c}^{\infty} g(x)f(x)\ \delta x \ \text{ and since we have defined } g(X) \leq c \text{ in this integral,} \\
&\geq c\int_{c}^{\infty} f(x)\ \delta x \\
&\geq cP\left[g(X) \geq c\right]
\end{aligned}$$

which you can rearrange to reach Markov's Inequality.

1.5.3 Chebyshev's Inequality

The **Chebyshev's Inequality** bounds the distribution of a random variable, as opposed to only the probability

Suppose we now have a random variable X with a distribution that contains a finite variance (and thus $E[X]$ exists). Then for a positive constant k

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{17}$$

Essentially this suggests that for k standard deviations, the probability of the random variable being within that range is bounded by 1 over k^2.

This is more conservative than the standard deviation bounds you might have seen for the normal distribution $(68 - 95 - 99.7\%)$. For $k = 2$, the probability of the confidence interval $(\mu - 2\sigma, \mu + 2\sigma)$ is $1 - \frac{1}{2^2} = 0.25$. For $k = 3$, this becomes $1 - \frac{1}{3^2} = 0.88\dot{8}$.

You can easily prove the Chebyshev's inequality by substituting $g(X) = (X-\mu)^2$ and using $c = k^2\sigma^2$, but you can also show this directly as well

$$
\begin{aligned}
E\left[|X-\mu| \geq k\sigma\right] &= E\left[\mathbb{1}_{|X-\mu|\geq k\sigma}\right] \\
&= E\left[\mathbb{1}_{\left(\frac{(X-\mu)}{k\sigma}\right)^2 \geq 1}\right] \\
&\leq E\left(\frac{(X-\mu)}{k\sigma}\right)^2 \\
&\leq \frac{1}{k^2}
\end{aligned}
$$

This inequality can be applied to any distribution that has a defined mean and variance.

1.5.4 Jensen's Inequality

For a convex function $\varphi(\cdot)$ on an open interval (essentially (a, b)), and if X is a random variable that has support within $mathcalI$ and $E[X] < \infty$ you can apply **Jensen's Inequality** to get:

$$\varphi\left(E[X]\right) \leq E\left[\varphi(X)\right] \tag{18}$$

and I'm just as confused as you are. How does having a convex function on an open interval with some random variable give you the aforementioned result?

First, let's define what a convex function is - and yes I like to think of it as the smiley face - so you're looking for $\varphi'(x) < \varphi'(y)$ (or $\varphi'(x) \leq \varphi'(y)$ for strictly convex), provided that x and y are both within a differentiable open interval (e.g. (a, b) as we saw earlier). We can also look into the second derivative as well but I'm sure you know what to do there.

You might gain some intuition for Jensen's inequality after looking figure below.
Here we can that a secant to a convex function is always greater than the function itself. We are actually able to represent the secant as a weighted mean of the convex function and thus show that

$$\varphi(wx_1 + (1-w)x_2) \leq w\varphi(x_1) + (1-w)\varphi(x_2)$$

for $w \in [0, 1]$. The generalisation of this result leads to Jensen's inequality. You can prove this finite form by using induction, but for our purposes we can show the statistical result by using the Taylor Expansion about the expectation.

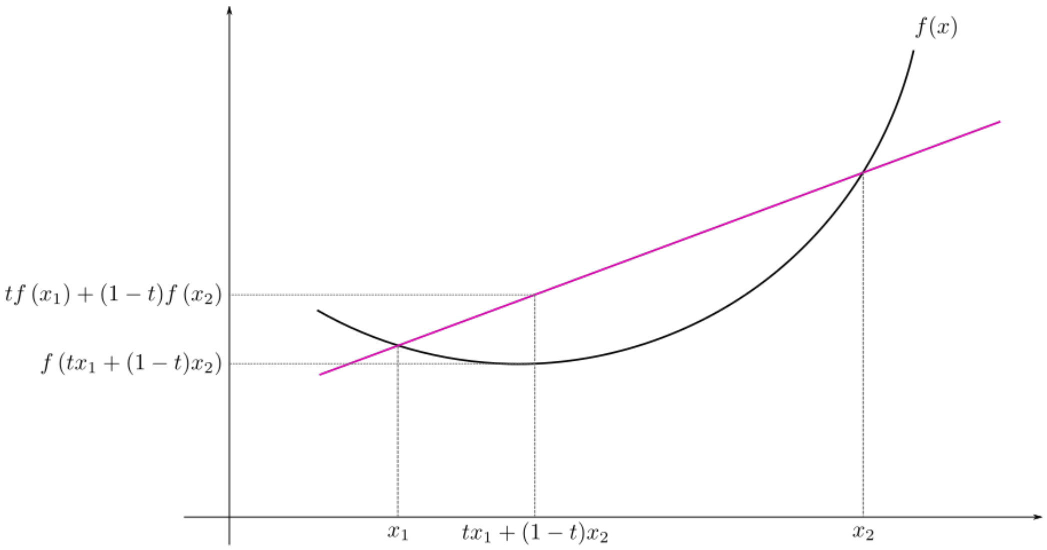

Figure 1: Secant of a Convex Function

$$
\begin{aligned}
\varphi(x) &= \varphi(\mu) + \varphi'(\mu)(x-\mu) + \frac{\varphi''(\zeta)(x-\mu)^2}{2!} \\
&\geq \varphi(\mu) + \varphi'(\mu)(x-\mu)
\end{aligned}
$$

and since the 2nd and 3rd term are both non-negative, you can take the expectation to achieve the desired results.

There are much more complicated measure-theoretic proves, but I don't think it's too important to touch on for now.

2 Moment Generating Functions

2.1 Introduction to Moments

When we were first introduced to **Moment Generating Functions** there was no clear motivation given for why it is such a useful concept in statistics. Let's begin the discussion with the concept of moments.

In statistics, we define the n^{th} **moment** as the $E[X^n]$. The 1st and 2nd moments of the a random variable X should be familiar, where $E[X]$ is the mean and can be combined with $E[X^2]$ to find the variance as $Var[X] = E[X^2] + E[X]^2$.

In addition, we define the n^{th} **central moment** as $E[(X - \mu)^n]$ and the n^{th} **standardised moment** as $E[\frac{(X-\mu)}{\sigma}^n]$. This is immediately useful, as the **skew** of a distribution is defined using the $3rd$ standardised moment $E[\frac{(X-\mu)}{\sigma}^3]$.

The skew of a distribution conveys information about the asymmetry of the distribution. Standardising here is important, as it removes the magnitude of the values being measured. Moreover, given that the first central moment is zero, then we can use the 3rd moment. Any odd central moment can give us information about asymmetry, but the 3rd moment is the easiest to calculate and has relatively less noise. However, just like how the median is another representation of the centre of a distribution, there are other ways to look at skewness.

Sometimes we are also interested in how heavy (or long) the tails of the distribution are. We can use the 4th standardised moment, the **Kurtosis** as a measure of the tails. A large Kurtosis suggests a sharp peak in the central, with relatively lower shoulders and heavy tails.

We define Kurtosis as

$$Kurt(X) = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^n\Big] - 3 \tag{19}$$

Note: without the 3, the function is called the **excess Kurtosis**

Now that we have defined moments, we return to the moment generating function. The moment generating function (MGF) can be used to evaluate a distribution's moments. We will get to that shortly.

It must be noted that knowledge of up to the n^{th} moment is insufficient in identifying a unique distribution. 2 random variables can have equal n^{th} moments but different distributions. There are some conditions required for this to be overturned, but we won't discuss it here.

Notation

In this class, we use the following notation:

$$\mu'_r = E[X^r] \tag{20}$$

for moments and

$$\mu_r = E[(X - \mu)^r] \tag{21}$$

Later on when we talk about bivariate moments, this will come in handy. For now, just keep this in mind.

2.2 Defining the Moment Generating Function

We define the moment generating function $\Psi_x(t)$ as

$$\Psi_X(t) = E\left[e^{\boldsymbol{t}^T\boldsymbol{X}}\right] \tag{22}$$

Recall that earlier we stated that the MGF can be used to derive the moments. Recall that

$$e^{tX} = \sum_{i=1}^{n} \frac{X^n t^n}{n!}$$

which gives

$$\begin{aligned} E[e^{tX}] &= \mathit{\Psi}_X(t) \\ &= \sum_{i=1}^{n} \frac{E[X^n]t^n}{n!} \\ &= 1 + \frac{E[X]t}{1!} + \frac{E[X^2]t^2}{2!} + ... + \frac{E[X^n]t^n}{n!} \end{aligned} \tag{23}$$

and thus we have

$$\begin{aligned} E[X^n] &= \mathit{\Psi}_X^{(n)}(t=0) \\ &= \frac{\delta^{(n)}}{\delta t^{(n)}}\big[\mathit{\Psi}_X(t=0)\big] \end{aligned} \tag{24}$$

This shows that the n^{th} derivative of the MGF for any distribution gives the n^{th} moment when the MGF derivative is evaluated at $t = 0$. This is an important alternative to specifying the probability distribution of a random variable.

Another function of interest is the **characteristic function**, which unlike the MGF, always exists for any distribution. We discuss this sibling later.

We also introduce the cumulant generating function (CGF) as

$$\mathcal{K}_X(t) = log\big[\mathit{\Psi}_X(t)\big] \tag{25}$$

This is another relation to the MGF itself. Whilst the n^{th} moment can be derived using the MGF, the central moments of a random variable can be similarly represented with the CGF, evaluating $\mathcal{K}_X^(n)(t=0)$ to reach the n^{th} central moment.

We won't go into much more detail about this. Another function of interest is the **characteristic function**, which we discuss later.

We can show that the MGF of the transformation of a random variable $Y = aX + b$ can be written as

$$\begin{aligned} E\big[e^{tY}\big] &= E\big[e^{t(aX+b)}\big] \\ &= e^{bt}E\big[e^{(at)X)}\big] \\ &= e^{bt}\mathit{\Psi}_X(at) \end{aligned} \tag{26}$$

and independent random variables can be split apart, as follows

$$\begin{aligned}\Psi_{X+Y}(t) &= E\big[e^{t(X+Y)}\big] \\ &= E\big[e^{(t)X)}\big]E\big[e^{(t)Y)}\big] \\ &= \Psi_X(t)\Psi_Y(t)\end{aligned} \tag{27}$$

Unlike moments, 2 random variables X and Y are equal if their MGF is identical.

2.3 MGFs for Common Distributions

Now we introduce various MGF for common distributions.

Binomial Distrbution

The $f_X(x)$ for a binomial is $\binom{n}{x}(p)^x(1-p)^{n-x}$ Thus

$$\begin{aligned}\Psi_X(t) &= E\big[e^{tX}\big] \\ &= \sum_{x=1}^{n} e^{tx}\cdot\binom{n}{x}(p)^x(1-p)^{n-x} \\ &= \sum_{x=1}^{n}\binom{n}{x}(p\cdot e^t)^x(1-p)^{n-x} \\ &= (e^tp+(1-p))^n\end{aligned}$$

as $(a+b)^n = \sum_{n=1}^{m}\binom{m}{n}(a)^n(b)^{m-n}$.

Geometric Distribution

The $f_X(x)$ for a geometric is $(1-p)^x(p)$ Thus

$$\begin{aligned}\Psi_X(t) &= E\big[e^{tX}\big] \\ &= \sum_{x=0}^{\infty} e^{tx}\cdot(1-p)^x(p) \\ &= p\sum_{x=0}^{\infty}[(1-p)\cdot e^t]^x \\ &= \frac{p}{1-((1-p)e^t)}\end{aligned}$$

as this the summation is a geometric series where $\sum_{x=0}^{\infty} ar^x = \frac{a}{1-r}$ if $|r| \leq 1$, and thus $(1-p)e^t \leq 1$.

Negative Binomial

The $f_X x$ for the negative binomial is $\binom{x+k-1}{k-1}(1-p)^x(p)^k$

$$\begin{aligned}
\Psi_X(t) &= E\left[e^{tX}\right] \\
&= \sum_{x=1}^{\infty} e^{tx} \cdot \binom{x+k-1}{k-1}(1-p)^x(p)^k \\
&= \sum_{x=1}^{\infty} \binom{x+k-1}{k-1}[(1-p)e^t]^x(p)^k \\
&= \frac{p^k}{(1-[(1-p)e^t]^k})
\end{aligned}$$

because $\sum_{x=1}^{\infty} \binom{x+k-1}{k-1}(y)^x = \frac{1}{(1-y)^k}$.

Poisson Distribution

The $f_X(x)$ for the poisson is $e^{-\lambda}\left[\frac{\lambda^x}{x!}\right]$

$$\begin{aligned}
\Psi_X(t) &= E\left[e^{tX}\right] \\
&= \sum_{x=0}^{\infty} e^{tx} \cdot e^{-\lambda}\left[\frac{\lambda^x}{x!}\right] \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \left[\frac{[e^{-t}\lambda]^x}{x!}\right] \\
&= e^{-\lambda} e^{e^t \cdot \lambda} \\
&= e^{-\lambda(1-e^t)}
\end{aligned}$$

as $\sum_{n=0}^{\infty} \frac{(x)^n}{n!} = e^x$

Normal Distribution

The $f_X(x)$ for the Normal is $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$

$$
\begin{aligned}
\Psi_X(t) &= E\big[e^{tX}\big] \\
&= \int_{-\infty}^{\infty} e^{tx} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big]\delta x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\big[-\frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2 - 2t\sigma^2)\big]\delta x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\big[-\frac{1}{2\sigma^2}(x^2 - 2x(\mu - t\sigma^2) + \mu^2 + 2\mu t\sigma^2 + t^2\sigma^4 - 2\mu t\sigma^2 - t^2\sigma^4)\big]\delta x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\big[-\frac{1}{2\sigma^2}(x - (\mu + t\sigma^2)^2 - 2\mu t\sigma^2 - t^2\sigma^4)\big]\delta x \\
&= \exp\big[\mu t + \frac{1}{2}t^2\sigma^2\big]\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\big[-\frac{1}{2\sigma^2}(x - (\mu + t\sigma^2))^2\big]\delta x \\
&= \exp\big[\mu t + \frac{1}{2}t^2\sigma^2\big]
\end{aligned}
$$

after some careful manipulation, where you complete the square and extract out the extra components.

Gamma Distribution

The $f_X(x)$ for the Gamma is $\frac{x^{\alpha} e^{\frac{x}{\beta}}}{\beta^{\alpha} \varGamma(\alpha)}$

$$
\begin{aligned}
\mathit{\Psi}_X(t) &= E\big[e^{tX}\big] \\
&= \int_{-\infty}^{\infty} e^{tx} \cdot \frac{x^{\alpha-1}e^{\frac{-x}{\beta}}}{\beta^\alpha \mathit{\Gamma}(\alpha)}\delta x \\
&= \int_{-\infty}^{\infty} \frac{x^{\alpha-1}e^{-x(\frac{1}{\beta}-t)}}{\beta^\alpha \mathit{\Gamma}(\alpha)}\delta x
\end{aligned}
$$

substitute $y = \big(\frac{1}{\beta} - t\big)x$ and $\delta y = \big(\frac{1}{\beta} - t\big)\delta x$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \frac{x^{\alpha-1}e^{-x(\frac{1}{\beta}-t)}}{\beta^\alpha \mathit{\Gamma}(\alpha)} \cdot \Big[\frac{1-\beta t}{\beta}\Big] \cdot \Big[\frac{\beta}{1-\beta t}\Big]\delta x \\
&= \int_{-\infty}^{\infty} \Big[\frac{\beta}{1-\beta t}\Big]^{\alpha-1} \cdot \frac{y^{\alpha-1}e^{-y}}{\beta^\alpha \mathit{\Gamma}(\alpha)} \cdot \Big[\frac{\beta}{1-\beta t}\Big]\delta y \\
&= \Big[\frac{\beta}{1-\beta t}\Big]^{\alpha} \cdot \Big[\frac{\mathit{\Gamma}(\alpha)}{\beta^\alpha \mathit{\Gamma}(\alpha)}\Big] \\
&= \frac{1}{(1-\beta t)^\alpha}
\end{aligned}
$$

2.4 Characteristic Function

The **characteristic function** is defined as

$$\mathit{\Psi}_X(it) = E\big[e^{\boldsymbol{it}^T\boldsymbol{X}}\big] \tag{28}$$

Moment generating functions are without a doubt useful in their applications. Now, we move to consider transformations of random variables.

3 Transformation of Random Variables

Now we move to dealing with multivariate random variables and their transformations. We start with the bivariate case and then generalise to higher dimensions. Multivariate distributions are extremely important in statistics, especially for conceptual understanding. I've tried to include simple examples to set the stage, and move into more difficult ones as the intuition is slowly built.

3.1 Multivariate Distributions

Let's start with the bivariate case. Suppose we have 2 random variables $\boldsymbol{X} = [X_1, X_2]$, where $\boldsymbol{X}$ is a random vector, then the **space** of $\boldsymbol{X}$ is the set of ordered pairs $\mathcal{D} = \{(x_1, x_2) | X_1 = x_1, X_2 = x_2\}$.

Suppose we are interested in an event $\mathcal{A}$ where the probability of this event $P_{X_1,X_2}(\mathcal{A})$, then we can write the cumulative distribution function as

$$\begin{aligned} F_{X_1,X_2}(x_1, x_2) &= P[(X_1 < x_1) \cap (X_2 < x_2)] \\ &= P[X_1 < x_1, X_2 < x_2] \end{aligned} \tag{29}$$

for all possible $(x_1, x_2) \in \mathbb{R}^2$.

As always, we must concern ourselves with both the discrete and continuous cases.

In the **discrete** case, we have the **joint probability mass function**

$$P_{X_1,X_2}(x_1, x_2) = P[X_1 = x_1, X_2 = x_2] \tag{30}$$

for all $(x_1, x_2) \in \mathcal{D}$.

In the **continuous** case, we need to first define the cumulative function as

$$F_{X_1,X_2}(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{X_1,X_2}(w_1, w_2)\, \delta w_1\, \delta w_2 \tag{31}$$

Notice that we integrate with respect to an arbitrary variable up until $x_{1,2}$. With this, we can state the **joint probability density function** as

$$\frac{\delta^2}{\delta x_1 \delta x_2} F_{X_1,X_2}(x_1, x_2) = f_{X_1,X_2}(x_1, x_2) \tag{32}$$

Example:

Suppose we have a joint distribution with the pdf

$$f_{X_1,X_2}(x_1, x_2) = \begin{cases} 6x_1^2 x_2 & 0 < x_1 < 1, 0 < x_2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

Evaluate $P\left(0 < X_1 < \frac{3/4}{,}\frac{1}{3} < X_2 < 2\right)$.

$$\begin{aligned}
P\left(0 < X_1 < \frac{3}{4}, \frac{1}{3} < X_2 < 2\right) &= \int_{\frac{1}{3}}^{2} \int_{0}^{\frac{3}{4}} 6x_1^2 x_2 \ \delta x_1 \delta x_2 \\
&= \int_{\frac{1}{3}}^{1} \int_{0}^{\frac{3}{4}} 6x_1^2 x_2 \ \delta x_1 \delta x_2 + \int_{1}^{2} \int_{0}^{\frac{3}{4}} 6x_1^2 x_2 \ \delta x_1 \delta x_2 \\
&= \int_{\frac{1}{3}}^{1} 2x_1^3 x_2 \Big|_{0}^{\frac{3}{4}} \delta x_2 + 0 \\
&= \int_{\frac{1}{3}}^{1} 2x_2 \left(\frac{27}{64}\right) \delta x_2 \\
&= \left(\frac{27}{64}\right) x_2^2 \Big|_{\frac{1}{3}}^{1} \\
&= \frac{3}{8}
\end{aligned}$$

which is a relatively straightforward example of calculating joint probabilities.

We can easily extend this concept to n-dimensions, for both discrete and continuous distributions. Consider a $\boldsymbol{X} = [X_1, X_2, ..., X_k]$ k-dimensional random vector, where X is the realisation of a sample space mapped to real values, $\boldsymbol{X} : \Omega \to \mathbb{R}^k$. We can hence define the joint CDF as:

$$F(x_1, x_2, ..., x_k) = P(X_1 \leq x_1, X_2 \leq x_2, ..., X_k \leq x_k) \tag{33}$$

For a **discrete distribution**, we have

$$F(x_1, x_2, ..., x_k) = \sum_{x_1 \leq m_1} \sum_{x_2 \leq m_2} ... \sum_{x_1 \leq m_k} P(X_1 = m_1, X_2 = m_2, ..., X_k = m_k) \tag{34}$$

Similarly, for a **continuous distribution**

$$F(x_1, x_2, ..., x_k) = \int_{m_1 \leq x_1} \int_{m_2 \leq x_2} ... \int_{m_k \leq x_k} m_1 m_2 ... m_k \ \delta m_1 \delta m_2 ... \delta m_k \tag{35}$$

For the **marginal** of $\mathbf{X}_A = [X_i, X_j ... X_z]$ we can calculate the marginal by summing/integrating over the $X_{B \in \Omega}$ random variables, where $A \cup B = \Omega$. For example, the marginal of (X_1, X_3) is

$$F(x_1, \infty, x_3, ..., \infty) = \int_{x_2 \in \mathbb{R}} \int_{x_4 \in \mathbb{R}} ... \int_{x_k \in \mathbb{R}} m_2 m_4 ... m_k \ \delta m_2 \delta m_4 ... \delta m_k \tag{36}$$

Example:
Consider the joint PDF $f_{X,Y}(x, y) = 6(x - y)$ where $0 < x < y < 1$.

(i) Find the marginal densities

(ii) Find $E[X|Y]$

(iii) Verify $E[E[X|Y]] = E[X]$

(i) The marginals are given by

$$\begin{aligned} f_X(x) &= 6 \int_x^1 (y - x) \delta y \\ &= 6 \left(\frac{1}{2} y^2 - xy \right) \Big|_x^1 \\ &= 6 (\frac{1}{2} - x + \frac{x^2}{2}) \ \text{ for } 0 < x < 1 \end{aligned}$$

and

$$\begin{aligned} f_Y(y) &= 6 \int_0^y (y - x) \delta x \\ &= 6 \left(xy - \frac{1}{2} x^2 \right) \Big|_0^y \\ &= 3y^2 \ \text{ for } 0 < y < 1 \end{aligned}$$

(ii) The conditional expectation, using the Bayes theorem and definition of expectation

$$\begin{aligned}
f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\
&= \frac{6(y-x)}{3y^2} \\
&= \frac{2(y-x)}{y^2}
\end{aligned}$$

$$\begin{aligned}
E[X|Y=y] &= \int_0^y x f_X(x)\ \delta x \\
&= \frac{2}{y^2}\int_0^y (yx - x^2)\ \delta x \\
&= \frac{2}{y^2}\left(\frac{1}{2}yx^2 - \frac{1}{3}x^3\right)\Big|_0^y \\
&= 2(\frac{1}{2}y - \frac{1}{3}y) \\
&= \frac{1}{3}y
\end{aligned}$$

(iii) First we have $E[E[X|Y]] = E\left[\frac{Y}{3}\right] = \frac{E[Y]}{3}$. Now we just need to check the previous expression is equivalent to $E[X]$.

Now we have

$$\begin{aligned}
E[Y] &= \int_0^1 x \cdot 3y^2\ \delta y \\
&= \frac{3}{4}
\end{aligned}$$

$$\begin{aligned}
E[X] &= \int_0^1 6x(\frac{1}{2} - x + \frac{x^2}{2})\ \delta x \\
&= 6\left(\frac{1}{4}x^2 - \frac{1}{3}x^3 + \frac{1}{8}x^4\right) \\
&= \frac{3}{2}x^2 - 2x^3 + \frac{3}{4}x^4|_0^1 \\
&= -\frac{1}{2} + \frac{3}{4} = \frac{1}{4}
\end{aligned}$$

Therefore $E[E[X|Y]] = \frac{E[Y]}{3} = \frac{1}{4} = E[X]$

3.2 Conditional Distributions and Expectations

From the previous example, we have caught a glimpse of the conditional distribution and expectation. The conditional probability of $X|Y$ is given by

$$P(X = x_i|Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} \quad -\infty < i < \infty$$

and we have to make sure that $P(Y = y_j) > 0$. For a fixed value of y_j the $P(X = x_i, Y = y_j)$ is a proper probability distribution. If the X and Y are **functionally related**, for example $X = h(y)$, then the conditional distribution is degenerate, i.e.

$$f_Y(y) = \begin{cases} 1 & x_i = h(y_i) \\ 0 & \text{otherwise} \end{cases}$$

In the continuous case with joint density, we have

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{37}$$

Also note that

$$E\left[h(X,Y)|Y = y\right] = E\left[h(X,y)|Y = y\right] \tag{38}$$

3.3 Transformation of Variables

3.3.1 Univariate

Suppose X is a random variable and $g(\cdot) : \mathbb{R} \to \mathbb{R}$. Assume that $g(\cdot)$ is strictly increasing. Suppose we are interested in a transformation of the unknown random variable X to another random variable Y, where $Y = g(X)$. In order to derive the distribution $F_Y(y)$ we have

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P(g(X) \leq y) \\
&= P(X \leq g^{-1}(y)) \\
&= F_X(g^{-1}(y)) \\
&= F_X(x)
\end{aligned} \tag{39}$$

hence by the chain rule,

$$f_Y(y) = f_X(x)\frac{\delta x}{\delta y} \tag{40}$$

If $g(\cdot)$ is strictly increasing $f_Y(y) = -f_X(x)\frac{\delta x}{\delta y}$ with $\frac{\delta x}{\delta y} < 0$. Thus using the absolute value covers both cases.

$$f_Y(y) = f_X(x)\left|\frac{\delta x}{\delta y}\right| \tag{41}$$

Let's look at 2 examples, one where you can apply the formula and the other where you cannot apply the **change of variables formula**.

Example:

Suppose we have $X \sim N(0,1)$, and let $Y = e^X$. This is the **log-normal distribution**, and let us denote this by $\psi(x)$. Hence swapping the x and y and solving for y gives

$$g^{-1}(x) = log(x)$$

$$\begin{aligned}
f_Y(y) &= f_X(x)\left|\frac{\delta x}{\delta y}\right| \\
&= f_X(g^{-1}(y))\left|\frac{\delta}{\delta y}(g^{-1}(y))\right| \\
&= f_X(g^{-1}(y))\left|\frac{\delta}{\delta y}(log(y))\right| \\
&= f(log(y))\frac{1}{y}
\end{aligned}$$

Example:

Suppose we have $X \sim N(0,1)$, and let $Y = X^2$. We cannot apply the change of variables formula because $g(x) = x^2$ is not one-to-one. Thus we have to take the CDF approach.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(g(X) \leq y) \\ &= P(-\sqrt{y} \leq x \leq \sqrt{y}) \\ &= F_X(y) - F_X(-y) \end{aligned}$$

and thus we have

$$\begin{aligned} f_Y(y) &= \frac{1}{2\sqrt{y}} \left(f_x(\sqrt{y}) + f_x(-\sqrt{y}) \right) \\ &= \frac{1}{\sqrt{y}} f_x(\sqrt{y}) \text{ because } f_X(x) \text{ is symmetrical} \end{aligned}$$

One has to be careful with these transformations, as sometimes the change of variables rule applies and other times you have to evaluate it from the perspective of the CDF

3.3.2 Multivariate

For multivariate transformations, we have to deal with the **Jacobian matrix** or more accurately, the determinant of the Jacobian (sometimes just referred to as the Jacobian).

Now we have a continuous random vector $\boldsymbol{X} = [X_1, X_2, ..., X_k]$ with joint PDF $f_{\boldsymbol{X}}(\boldsymbol{x})$ and we are again, interested in a $\boldsymbol{Y} = g(\boldsymbol{X})$ where $g(\cdot) : \mathbb{D}^k \to \mathbb{D}^k$ where $\mathbb{D} \subseteq \mathbb{R}$ such that $\mathbb{D}$ contains the support of $\boldsymbol{X}$.

Assume that for the transformation, all partial-derivatives $\frac{\delta x_i}{\delta y_j}$ exist and are continuous. Hence we can derive the **Jacobian matrix**

$$\frac{\delta \boldsymbol{x}}{\delta \boldsymbol{y}} = \begin{pmatrix} \frac{\delta x_1}{\delta y_1} & \frac{\delta x_1}{\delta y_1} & \cdots & \frac{\delta x_1}{\delta y_k} \\ \frac{\delta x_2}{\delta y_1} & \frac{\delta x_2}{\delta y_2} & \cdots & \frac{\delta x_2}{\delta y_k} \\ \vdots & & \ddots & \vdots \\ \frac{\delta x_k}{\delta y_1} & \frac{\delta x_k}{\delta y_2} & \cdots & \frac{\delta x_k}{\delta y_k} \end{pmatrix}$$

Assume that this matrix is non-singular (and hence the inverse exists, the determinant is non-zero), we can say that

$$\begin{aligned} f_{\boldsymbol{Y}}(y) &= f_{\boldsymbol{X}}(x)\left|\frac{\delta \boldsymbol{x}}{\delta \boldsymbol{y}}\right| \\ &= f_{\boldsymbol{X}}(g^{-1}(y))\left|\frac{\delta}{\delta \boldsymbol{y}} g^{-1}(y)\right| \end{aligned} \tag{42}$$

where $\left|\frac{\delta \boldsymbol{x}}{\delta \boldsymbol{y}}\right|$ is the absolute value of the determinant of the Jacobian. Another important thing to note is that $\left|\frac{\delta \boldsymbol{x}}{\delta \boldsymbol{y}}\right| = \left|\frac{\delta \boldsymbol{y}}{\delta \boldsymbol{x}}\right|^{-1}$

Important: Transformations of a discrete random variables does NOT require the Jacobian. For example, if $Y = X^3$ has $X = 1, 2, ..., n$ then

$$P(Y = y) = P(X = y^{\frac{1}{3}})$$

Let's have a look at an example of variable transformation in 2 dimensions.

Example:

Let $U \sim U(0, 2\pi])$ and $T \sim \exp(1)$ be independent of U. Define a transformation $X = \sqrt{2T}\cos(U)$ and $Y = \sqrt{2T}\sin(U)$. Find the joint PDF $[X, Y]$.

The joint PDF of $[U, T]$ is

$$f_{U,T}(u, t) = \frac{1}{2\pi} e^{-t}$$

Now, looking at X and Y together gives $X^2 + Y^2 = 2T(sin^2(U) + cos^2(U)) = 2T$. We first look at the inverse transformations. Note that normally we're after $g^{-1}(x, y)$ but we already have what we need with the result we just mentioned.

Now we just need the Jacobian

$$\frac{\delta(x, y)}{\delta(u, t)} = \begin{pmatrix} \frac{\delta x}{\delta u} & \frac{\delta x}{\delta t} \\ \frac{\delta y}{\delta u} & \frac{\delta y}{\delta t} \end{pmatrix}$$

which gives you

$$\frac{\delta(x, y)}{\delta(u, t)} = \begin{pmatrix} -\sqrt{2t} sin(u) & \frac{1}{\sqrt{2t}} cos(u) \\ \sqrt{2t} sin(u) & \frac{1}{\sqrt{2t}} sin(u) \end{pmatrix}$$

with the $|\det(\mathcal{J})| = 1$. Now we have enough information to do the transformation.

$$\begin{aligned} f_{X,Y}(x,y) &= f_{U,T}(u,t)\left|\frac{\delta(u,t)}{\delta(x,y)}\right| \\ &= \frac{1}{2\pi}e^{-t}\cdot 1 \\ &= \frac{1}{2\pi}e^{-\frac{1}{2}x^2+y^2} \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2} \end{aligned} \tag{43}$$

This looks like the product of 2 standard normals! This is known as the **Box-Muller** method. The Box-Muller method is a pseudo-random sampling method for generating pairs of independent, standard normal random variables giving uniformly distributed random numbers.

3.4 Bivariate Moments

Now we look at bivariate variables from the perspective of moments. For any bivariate distribution $f(x,y)$, you have

$$E[g(X,Y)] = \begin{cases} \sum_i \sum_j P(X = x_i, Y = y_j) \\ \int\int_{\boldsymbol{R}} g(x,y)f(x,y)\ \delta x\delta y \end{cases}$$

When you have the expectation of the joint density, then the moments of $[X,Y]$ are

$$\mu'_{rs} = E[X^rY^s] \tag{44}$$

and the central moments are

$$\mu_{rs} = E[(X-\mu_{10})^r(Y-\mu_{01})^s] \tag{45}$$

Furthermore, we can represent the covariance as μ_{11} and extending that the correlation as

$$\rho = \frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}} \tag{46}$$

And of course, we have the bivariate moments as

$$\Psi_{X,Y}(s,t) = E\left[e^{sX+tY}\right] \tag{47}$$

with partial moments found by differentiation and setting $s = t = 0$.

$$\mu'_{ij} = \frac{\delta^{i+j}\Psi(s,t)}{\delta^i s\ \delta^j t}_{|s=t=0} \tag{48}$$

Let's look at an example, and we solve

Example:

Suppose we have a bivariate distribution for 2 random variables X and Y as shown below.

X, Y	0	1	$P(Y)$
0	0.2	0.3	0.5
1	0.4	0.1	0.5
$P(X)$	0.6	0.4	1

(i) Find $\Psi(s,t)$

(ii) Hence or otherwise, find μ_X, μ_Y, μ'_{11} and μ_{11}

(iii) Lastly, find ρ

(i) Since we have the entire discrete distribution mapped out, we have

$$\begin{aligned}
\Psi(s,t) &= E\left[e^{sX+tY}\right] \\
&= \sum_y \sum_x e^{sX+tY} P(X=x, Y=y) \\
&= e^{s(0)+t(0)}P(X=0,Y=0) + e^{sX+t(0)}P(X=x,Y=0) \\
&+ e^{s(0)+tY}P(X=0,Y=y) + e^{s(1)+t(1)}P(X=1,Y=1) \\
&= 0.2 + 0.4e^s + 0.3e^t + 0.1e^{s+t}
\end{aligned}$$

(ii) Using the MGF from above, we can proceed to evaluate

$$\begin{aligned}
\mu_X &= \left[\frac{\delta}{\delta s}\Psi(s,t)\right]_{|s=t=0} \\
&= 0.4e^0 + 0.1e^{0+0} \\
&= 0.5
\end{aligned}$$

$$\begin{aligned}\mu_Y &= \left[\frac{\delta}{\delta t}\Psi(s,t)\right]_{|s=t=0} \\ &= 0.3e^0 + 0.1e^{0+0} \\ &= 0.4\end{aligned}$$

$$\begin{aligned}\mu_{11} &= \left[\frac{\delta^2}{\delta s\ \delta t}\Psi(s,t)\right]_{|s=t=0} \\ &= \frac{\delta}{\delta t}\left[0.4e^s + 0.1e^{s+t}\right] \\ &= 0.1e^{0+0} \\ &= 0.1\end{aligned}$$

$$\begin{aligned}\mu'_{11} &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E[XY] - \mu_X\mu_Y \\ &= 0.1 - (0.5 \times 0.4) \\ &= -0.1\end{aligned}$$

(iii) We know the correlation is given by $\frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}}$.
$\mu_{20} = \frac{\delta}{\delta s}\left[0.4e^s + 0.1e^{s+t}\right] = 0.5$ and $\mu_{20} = \frac{\delta}{\delta t}\left[0.3e^t + 0.1e^{s+t}\right] = 0.4$

Hence we have

$$\begin{aligned}\rho &= \frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}} \\ &= \frac{-0.1}{\sqrt{(0.5-0.5^2)(0.4-0.4^2)}} \\ &= -0.408\end{aligned}$$

3.5 Convolution Formulae

Suppose we have independent variables from the same distribution and we are interested in the distribution of the sample sum. Using a concept called **convolution** we evaluate the distribution of the sum of 2 independent variables. Repeated application gives the sum of n independent variables.

If we have $Z = X + Y$ with PDFs $f_X(x)$ and $f_Y(y)$. $f_Z(z)$ is the convolution $f_X(x)$ and $f_Y(y)$. We can show that

$$\begin{aligned} f_Z(z) &= P\left(X + Y \leq z\right) \\ &= P\left(X \leq z - Y\right) \text{ and with the Law of Iterated Expectation} \\ &= E\left[P(X \leq z - Y) | Y = y\right] \\ &= E\left[F_X(z - Y)\right] \end{aligned}$$

In the discrete case, we have

$$f_Z(z) = \sum_y f_X(z - y) f_Y(y)$$

and in the continuous case

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y)\ \delta y$$

Let's look at 2 examples, one in the discrete case and the other in the continuous case.

Example: Discrete

Suppose you have 2 discrete random variables X and Y with

$$P(X = x) = \begin{cases} \frac{1}{2} & x \in \{0, 1\} \\ 0 & \text{otherwise} \end{cases}$$

$$P(Y = y) = \begin{cases} \frac{1}{2} & y \in \{0, 1\} \\ 0 & \text{otherwise} \end{cases}$$

and you are interested in the PMF of $Z = X + Y$ where $Z \in \{0, 1, 2\}$. The PMF of Z can be evaluated separately as follows:

$$\begin{aligned} P(Z = 0) &= \sum_{y=0}^{1} f_X(0 - y) f_Y(y) \\ &= P(X = 0)P(Y = 0) + P(X = -1)P(Y = 1) \\ &= \frac{1}{2} \cdot \frac{1}{2} + 0 \\ &= \frac{1}{4} \end{aligned}$$

$$
\begin{aligned}
P(Z=1) &= \sum_{y=0}^{1} f_X(1-y) f_Y(y) \\
&= P(X=1)P(Y=0) + P(X=0)P(Y=1) \\
&= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \\
&= \frac{1}{2}
\end{aligned}
$$

$$
\begin{aligned}
P(Z=0) &= \sum_{y=0}^{1} f_X(2-y) f_Y(y) \\
&= P(X=2)P(Y=0) + P(X=1)P(Y=1) \\
&= 0 + \frac{1}{2} \cdot \frac{1}{2} \\
&= \frac{1}{4}
\end{aligned}
$$

Hence the distribution of $f_Z(z)$ is

$$
P(Z=z) = \begin{cases} \frac{1}{4} & z=0 \\ \frac{1}{2} & z=1 \\ \frac{1}{4} & z=2 \\ 0 & \text{otherwise} \end{cases}
$$

Example: Continuous

Suppose you have 2 discrete continuous variables X and Y with

$$
P(X=x) = \begin{cases} 1 & x \in \{0,1\} \\ 0 & \text{otherwise} \end{cases}
$$

$$
P(Y=y) = \begin{cases} e^{-y} & y \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}
$$

and you are interested in the PDF of $Z = X + Y$ where $Z \in [0,\infty)$. The PDF of Z can be evaluated separately as follows:
$\nVdash$

$$\begin{aligned}
f_Z(z) &= \int_0^\infty f_X(z-Y)f_Y(y)\ \delta y \\
&= \int_0^\infty \mathbb{1}\left\{0 \leq z-y \leq 1\right\} e^{-y}\ \delta y \\
&= \int_0^\infty \mathbb{1}\left\{z-1 \leq y \leq z\right\} e^{-y}\ \delta y \\
&= \int_{\max\{z-1,0\}}^{z} e^{-y}\ \delta y \\
&= -e^{-z} + e^{-\max\{z-1,0\}}
\end{aligned}$$

Hence we have

$$f_Z(z) = \begin{cases} e^{1-z} - e^{-z} & \text{if } z \geq 1 \\ 1 - e^{-z} & \text{if } 0 \leq z < 1 \end{cases}$$

4 Multivariate Normal Variables

Multivariate normal variables are actually quite straightforward if you are familiar with linear algebra. We first present some basic properties of the multivariate standard normal, followed by their linear transformation and then finish off with the bivariate normal. The following properties should not be too complicated.

4.1 Properties of Multivariate Normals

The multivariate normal vector $\boldsymbol{X} = [X_1, X_2, ..., X_k]$ with $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has probability distribution

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} (\det[\boldsymbol{\Sigma}])^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{49}$$

where $x \in \mathbb{R}^k$.

Suppose we have a multivariate standard normal $\boldsymbol{Z} = [Z_1, Z_2, ..., Z_k]$, where $\boldsymbol{Z} \sim N(0, I_k)$. If we are interested in a linear transformation $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$. The $\boldsymbol{X}$ are unrelated (Although you could chose a transformation to reach $\boldsymbol{X}$)! Then we have expectation

$$\begin{aligned} E[\boldsymbol{X}] &= E[\boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}] \\ &= E[\boldsymbol{A} \cdot 0] + E[\boldsymbol{\mu}] \\ &= \mu \end{aligned} \tag{50}$$

and variance-covariance matrix

$$\begin{aligned} Cov[\boldsymbol{X}] &= E[\boldsymbol{X} - E[\boldsymbol{X}]][\boldsymbol{X} - E[\boldsymbol{X}]]^T \\ &= E[(\boldsymbol{A}\boldsymbol{Z})(\boldsymbol{A}\boldsymbol{Z})^T] \\ &= \boldsymbol{A}E[\boldsymbol{Z}\boldsymbol{Z}^T]\boldsymbol{A}^T \\ &= \boldsymbol{A}\boldsymbol{A}^T \end{aligned} \tag{51}$$

and lastly the MGF

$$\begin{aligned}\psi_{\boldsymbol{X}}(\boldsymbol{t}) &= E[e^{\boldsymbol{t}^T\boldsymbol{X}}] \\ &= E[e^{\boldsymbol{t}^T\boldsymbol{AZ}+\boldsymbol{\mu}}] \\ &= e^{\boldsymbol{t}^T\boldsymbol{\mu}}\cdot\psi_{\boldsymbol{X}}(\boldsymbol{t}^T\boldsymbol{A}) \\ &= e^{\boldsymbol{t}^T\boldsymbol{\mu}}\cdot e^{\frac{1}{2}(\boldsymbol{t}^T\boldsymbol{A})(\boldsymbol{t}^T\boldsymbol{A})^T} \\ &= e^{\left(\boldsymbol{t}^T\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{t\Sigma t}\right)}\end{aligned} \tag{52}$$

Let's have a look at how to utilise the multivariate standard normal to reach the Cauchy distribution.

Example:

Suppose that Z_1 and Z_2 are two *iid* random variables with $Z_{1,2} \sim N(0,1)$. Find the PDF of $Y = \frac{Z_1}{Z_2}$, where $Z_2 \neq 0$.

The trick is to create a $Y_2 = Z_2$ and relabel $Y = Y_1$. Using this, we have $Z_1 = Y_1Y_2$ and $Z_2 = Y_2$. The Jacobian can be obtained as

$$\mathcal{J} = \begin{pmatrix} \frac{\delta z_1}{\delta y_1} & \frac{\delta z_1}{\delta y_2} \\ \frac{\delta z_2}{\delta y_1} & \frac{\delta z_2}{\delta y_2} \end{pmatrix} = \begin{pmatrix} Y_2 & Y_1 \\ 0 & 1 \end{pmatrix}$$

This gives $\det\mathcal{J} = |y_2|$. Using the change of variable formula, we have

$$\begin{aligned} g_{\boldsymbol{Y}}(y_1, y_2) &= \frac{1}{2\pi}\exp\left\{-\frac{1}{2}(z_1^2 + z_2^2)\right\}\cdot|\det\mathcal{J}| \\ &= \frac{1}{2\pi}\exp\left\{-\frac{1}{2}(y_1^2y_2^2 + y_2^2)\right\}\cdot|y_2| \end{aligned}$$

In order to find the distribution of Y_1 we take the marginals, and thus we want to evaluate

$$g_{Y_1}(y_1) = \frac{1}{2\pi}\int_{-\infty}^{0} -y_2 \exp\left\{\frac{y_1^2y_2^2+y_2^2}{-2}\right\}\;\delta y + \frac{1}{2\pi}\int_{0}^{\infty} +y_2 \exp\left\{\frac{y_1^2y_2^2+y_2^2}{-2}\right\}\;\delta y$$

substituting $u = -\frac{1}{2}(y_1^2+1)y_2^2$ and differentiating

$du = -(y_1^2+1)y_2\;dy$ and preparing for substition of u gives

$$\begin{aligned}
&= \frac{1}{2\pi}\cdot\frac{1}{y^2+1}\left(\int_{-\infty}^{0} -y_2(y_2^2+1)\exp\{\dots\}\;\delta y - \int_{0}^{-\infty} -y_2(y_2^2+1)\exp\{\dots\}\;\delta y\right)\\
&= \frac{1}{2\pi}\cdot\frac{1}{y^2+1}\left(\int_{-\infty}^{0} e^u\;\delta u - \int_{0}^{-\infty} e^u\;\delta u\right)\\
&= \frac{1}{2\pi}\cdot\frac{1}{y^2+1}\left[(1-0)-(0-1)\right]\\
&= \frac{1}{\pi(y^2+1)}
\end{aligned}$$

which is the **Cauchy Distribution!**

4.2 Bivariate Normal Variables

Bivariate normals are what you might be expect - a random vector containing 2 normal variables $[X_1, X_2]^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = [\mu_1, \mu_2]^T$ and $\boldsymbol{\Sigma}$

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

where $\rho \neq \pm 1$ because $\rho = \text{Corr}(X_1, X_2)$. In general, a correlation equal to zero does not imply independent but if $\rho = 0$, then X_1 and X_2 are independent random variables - a special case for normal random variables.

Suppose that $\boldsymbol{C} = \Sigma_{12}\Sigma_{22}^{-1}$, then using the above we can show that $E\left[X_1 - CX_2\right]X_2^T = 0$.

$$\begin{aligned}
E\left[X_1X_2^T - \boldsymbol{C}X_2X_2^T\right] &= E\left[X_1X_2^T\right] - E\left[\boldsymbol{C}X_2X_2^T\right]\\
&= \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{-1}\\
&= 0
\end{aligned}$$

Given that $\boldsymbol{C} = \Sigma_{12}\Sigma_{22}^{-1} = \rho\frac{\sigma_1}{\sigma_2}$.

So if $\boldsymbol{C}$ is the above special result then we have $[X_1 - \boldsymbol{C}X_2]$ being independent of X_2 and we can use this to derive the conditional variance of X_1.

We already know that the mean is zero, and thus for the variance

$$\begin{aligned}
Var\left[X_1 - \mu_1 - \boldsymbol{C}(X_2 - \mu_2)\right] &= E\left[\left(X_1 - \mu_1 - \rho\frac{\sigma_1}{\sigma_2}(X_2 - \mu_2\right)^2\right] - 0 \\
&= E\left[X_1^2\right] - 2E\left[X_1\right]\mu_1 + \mu_1^2 - 2[E\left[X_1\right] - \mu_1] \\
&\cdot [E\left[X_2\right] - \mu_2]\rho\frac{\sigma_1}{\sigma_2} + \rho^2\frac{\sigma_1^2}{\sigma_2^2}Var(X_2) \\
&= Var(X_1) - 2\text{Cov}(X_1, X_2)\rho\frac{\sigma_1}{\sigma_2} + \rho^2\frac{\sigma_1^2}{\sigma_2^2}Var(X_2) \\
&= \sigma_1^2 - 2[\rho\sigma_1\sigma_2]\rho\frac{\sigma_1}{\sigma_2} + \rho^2\frac{\sigma_1^2}{\sigma_2^2}\cdot\sigma_2^2 \\
&= \sigma_1^2(1 - \rho^2)
\end{aligned}$$

thus giving $[X_1|X_2 = x_2] \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1 - \rho^2)\right)$.
Suppose we want to generate our bivariate distribution by applying a transformation onto a bivariate vector of standard normals. In order to do that, we first define $\boldsymbol{Z} = [Z_1, Z_2]$ which is a bivariate standard normal where the 2 random variables are *iid*, and then apply

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$$

where $\boldsymbol{\mu} = [\mu_1, \mu_2]^T$ and

$$\boldsymbol{A} = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1 - \rho^2} \end{pmatrix}$$

Example:
For a bivariate standard normal $\boldsymbol{Z}$ and $\boldsymbol{\mu} = [2, 3]^T$ with $\sigma_1 = 5, \sigma_2 = 6$ and $\rho = 0.5$, write down the distribution of $\boldsymbol{X}$ where $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$.

$$\begin{aligned}
E[\boldsymbol{X}] &= E[\boldsymbol{A}\boldsymbol{Z}] + E[\boldsymbol{\mu}] \\
&= \boldsymbol{A}\cdot 0 + E[\boldsymbol{\mu}] \\
&= \boldsymbol{\mu}
\end{aligned}$$

Now the variance is

$$\begin{aligned}\text{Var}[\boldsymbol{X}] &= \text{Var}[\boldsymbol{AZ}] \\ &= \boldsymbol{A\mathcal{I}A}^T\end{aligned}$$

Writing in matrix form gives

$$\boldsymbol{AA}^T = \begin{pmatrix} 25 & 15 \\ 15 & 36 \end{pmatrix}$$

Alright, one last example

Example:

Suppose that the joint PDF of X and Y is given by

$$f(x,y) = \frac{1}{\pi\sqrt{3}} \exp\left\{-\frac{2}{3}\left(x^2 - xy + y^2\right)\right\}$$

for $-\infty < x, y < \infty$. Suppose that the marginal distribution of X and Y are standard normal and $\rho = 0.5$.

(i) Find the conditional density of $Y|X = x$ and show that it is $N(x/2; 3/4)$

(ii) Hence or otherwise, evaluate $P(Y \leq 2|X = 1.2)$

We have previously shown that $[X_1|X_2 = x_2] \sim N\left(\mu_1 - \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1 - \rho^2)\right)$.

(i) Since we know that the marginals are standard normal, we have $[X, Y] \sim N(0, 0, 1, 1, 0.5)$ which means that

$$\begin{aligned}(Y|X = x) &\sim N\left(\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X), \sigma_Y^2(1 - \rho^2)\right) \\ &\sim N\left(0 + 0.5 \cdot 1(x - 0), 1(1 - 0.5^2)\right) \\ &\sim N\left(\frac{x}{2}, \frac{3}{4}\right)\end{aligned}$$

(ii) Since we know that the conditional density is $P(Y \leq 2|X = 1.2)$ is normal, we can transform this into a standard normal and evaluate the expression using the standard z-test.

$$\begin{aligned}P(Y \leq 2|X = 1.2) &= P(Z \leq \frac{y - \mu_Y}{\sigma_Y}) \\ &= P(Z \leq \frac{2 - 0.5 \cdot 1.2}{\sqrt{0.75}}) \\ &= P(Z \leq 1.616) \\ &\approx 0.947\end{aligned}$$

4.3 Partition of a Random Vector

Suppose we have a multivariate normal distribution $\boldsymbol{X} \in R^n$, sometimes we may want to represent this single multivariate random variable by partitioning it into 2 or more multivariate random variables. For example, we can let $\boldsymbol{X}_1$ be a subvector of $\boldsymbol{X}$ where $\boldsymbol{X}_1 \in \mathbb{R}^m$ and $n = m + k$ and thus we also have $\boldsymbol{X}_2 \in \mathbb{R}^k$. Partitioning using this logic can also be done to the mean and variance. Thus we have

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}$$

with mean $E[\boldsymbol{X}]$ and variance $\mathrm{Var}[\boldsymbol{X}]$

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

In this case, the covariance of $\boldsymbol{X}_1$ is $\boldsymbol{\Sigma}_{11}$ etc.

Being able to partition a multivariate normal vector like this is extremely useful, because the mean and variance of the marginal distribution of $\boldsymbol{X}_i$ can be derived directly from the partial vectors.

5 Limiting Distributions

In simple words, a **limiting (or asymptotic) distribution** is what happens to a probability distribution when the sample size of X becomes extremely large, or approaches infinity if you like.

In the real world, not all data you work with will have a normal distribution. Sometimes, there's not much you can do about that and certain limitations acknowledged and you move onto a different strategy that doesn't require the normality assumption.

A practical benefit of limiting distributions, especially ones that reach normality, is the fact that we can say "alright how about we collect more data?"

All jokes aside though, a solid understanding of convergence will go a long way for you.

5.1 Convergence in Probability

We introduce 2 important notions of convergence, the first being convergence in probability.

Given a sequence of n random variables X_n, we say that X_n **converges in probability** to a random variable X if, for all $\epsilon > 0$

$$\lim_{n \to \infty} P\left[X_n - X \geq \epsilon\right] = 0 \tag{53}$$

and thus

$$X_n \xrightarrow{P} X \tag{54}$$

In statistics, the limiting X is often a constant (i.e. a degenerate random variable with all its mass at some constant c).

An important result relevant to convergence in probability is the **Weak Law of Large Numbers**.
The Weak Law of Large Numbers (WLLN) states that for a sequence random variable X_n with common mean μ and finite variance $\sigma^2 < \infty$. Then

$$\overline{X_n} \longrightarrow \mu \tag{55}$$

where $\overline{X_n}$ is the mean of X_n. We can show this using Chebyshev's inequality:

$$\begin{aligned}\lim_{n\to\infty} P\left[X_n - X \geq \epsilon\right] &= \lim_{n\to\infty} P\left[X_n - X \geq \epsilon\left[\frac{\sqrt{n}}{\sigma}\right]\left[\frac{\sigma}{\sqrt{n}}\right]\right] \\ &\leq \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0\end{aligned}$$

where $\mathrm{Var}\left[\overline{X_n}\right] = \frac{\sigma^2}{n}$. This result states that the mean of a sequence of random variables with the same distribution will eventually reach the true mean μ.

There are some useful results in probability convergence that you should remember

(i) $X_n + Y_n \xrightarrow{P} X + Y$

(ii) $X_n Y_n \xrightarrow{P} XY$

Another useful concept comes from **Slutsky's Theorem**, which states that if $Y_n \xrightarrow{P} c$ for some constant c, then they are jointly convergent.

(i) $X_n + Y_n \xrightarrow{P} X + c$

(ii) $X_n Y_n \xrightarrow{P} X \cdot c$

(iii) If $X_n \xrightarrow{P} X$ then $aX_n \xrightarrow{P} aX$

(iv) If $X_n \xrightarrow{P} a$ then $g(X_n) \xrightarrow{P} g(a)$

Now we discuss convergence in distribution.

5.2 Convergence in Distribution

A sequence of random variables $\boldsymbol{X}_n$ with CDF $F_{\boldsymbol{X}_n}(x)$ will converge to $\boldsymbol{X}$ if

$$F_{X_n}(x) \xrightarrow{D} F_{X_n}(x) \tag{56}$$

at all points x, where $F_X(x)$ is continuous. We can also write this as

$$P(X_n \leq x) \xrightarrow{D} P(X \leq x)$$

If the corresponding MGF converges, $\psi_{X_n}(t) \longrightarrow \psi_X(t)$, then this implies a convergence in distribution. This can be proved using characteristic functions.

This concept can be used to show the **Central Limit Theorem**. We won't write the proof here, but basically this theorem states that a sequence of $\boldsymbol{X}_n$ normal variables with finite variance can be transformed to $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$, which importantly has a limiting normal distribution.

$$\lim_{n \longrightarrow \infty} P\left(\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n}\sigma} \leq x\right) = \Phi(x) \tag{57}$$

which you can rearrange to reach

$$P\left(\frac{\overline{X}_n - \mu)}{\sigma/\sqrt{n}} \leq x\right) = \Phi(\frac{x - \mu}{\sigma/\sqrt{n}}) \tag{58}$$

5.3 Limiting Distributions of Variable Transformations

Suppose that we know the limiting distribution of X_n and are interested in the limiting distribution of $g(X_n)$. Typically we are interested in the mean and variance of a distribution.

To find the variance and mean, we use Taylor's Expansion about the mean, μ. Approximate the $g(x)$

$$g(x) \approx g(\mu) + (x - \mu)g'(\mu) \tag{59}$$

and using this, we have

$$\begin{aligned} E[g(x)] &= E[g(\mu) + (x - \mu)g'(\mu)] \\ &= E[g(\mu)] + E[(x - \mu)g'(\mu)] \\ &= E[g(\mu)] \end{aligned} \tag{60}$$

$$\begin{aligned} \text{Var}[g(x)] &= \text{Var}[g(\mu) + (x - \mu)g'(\mu)] \\ &= \text{Var}[xg'(\mu)] \\ &= [g'(\mu)]^2\sigma^2 \end{aligned} \tag{61}$$

Hence we have $(g(X) - g(\mu)) \sim N(0, [g'(\mu)]^2\sigma^2)$.

This is known as the **delta method**, which we talk about in a bit more detail later.
Example:

Let X_i for $i = 1, 2, ..., n$ are *iid* $\mathrm{B}(1, p)$. Find the limiting distribution of $g(\overline{X}_n) = \overline{X}_n(1 - \overline{X}_n)$.

We know that $\overline{X}_n \sim N(p, \frac{p(1-p)}{n})$. Hence we have $g(\overline{X}_n) = \overline{X}_n(1 - \overline{X}_n)$.

Hence we have $E[g(\overline{X}_n)] = E[\overline{X}_n(1 - \overline{X}_n)] = p(1 - p)$. Since we have $g'(\overline{X}_n) = 1 - 2\overline{X}_n$. Hence we have $\mathrm{Var}[g(\overline{X}_n)] = \frac{1}{n}p(1-p)(1-2p)^2$.

Hence we have $g(\overline{X}_n) \xrightarrow{\infty} N(p(1-p), \frac{1}{n}p(1-p)(1-2p)^2)$.

5.4 Variance-stabilising Transformations

Sometimes during the transformation of a random variable we end up with a variance that depends on the mean - $\mathrm{Var}(X) = \Omega(\mu)$.

This could be problematic when considering limiting distributions as it should not depend on an unknown parameter. As you have already seen, you have $\sqrt{n}(\overline{X}_n - \mu \xrightarrow{D} N(0, \Omega(\mu))$, which gives you a limiting distribution

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{D} N(0, [g'(\mu)]^2\, \Omega(\mu))$$

Now we want to re-express $\mathrm{Var} g(\overline{X}_n)$ as a constant, and we can do this by writing

$$[g'(\mu)]^2\, \Omega(\mu) = C^2 \tag{62}$$

Rearranging this, we can obtain an expression

$$g(x) = \int_{-\infty}^{x} \frac{C}{[\Omega(\mu)]^{1/2}}\, \delta\mu \tag{63}$$

Let's look at an example to see how it all ties together.

Example:

Suppose that $X_1, X_2, ..., X_n$ are *iid* Pois(λ).

(i) Find the limiting distribution of $\overline{X}_n$.

(ii) Find a suitable transformation of $g(\overline{X}_n)$ to stabilise the variance in (i)

(iii) Write down the corresponding distribution using the result in (ii)

(i) We know that for Poisson distribution, the limiting distribution of the mean is $\overline{X}_n \sim N(\lambda, \frac{\lambda}{n})$.

(ii) Knowing that the variance is a function in λ, hence we have

$$\begin{aligned} g(x) &= \int_{-\infty}^{x} \frac{C}{[\mathit{\Omega}(\mu)]^{1/2}} \, \delta\mu \\ &= \int_{-\infty}^{x} \frac{C}{[\lambda]^{1/2}} \, \delta\lambda \\ &= 2C\sqrt{y} \\ &= C'\sqrt{y} \end{aligned}$$

(iii) Hence $\sqrt{n}(g(\overline{X}_n) - g(\lambda)) \sim N\left(0, [g'(\lambda)]^2 \, \mathit{\Omega}(\lambda)\right)$

$$\begin{aligned} N\left(0, [g'(\lambda)]^2 \, \mathit{\Omega}(\lambda)\right) &= N\left(0, \left[C' \frac{1}{2\sqrt{\lambda}}\right]^2 \mathit{\Omega}(\lambda)\right) \\ &= N\left(0, \frac{(C')^2}{4\lambda} \cdot \frac{\lambda}{n}\right) \\ &= N\left(0, \frac{(C')^2}{4n}\right) \end{aligned}$$

6 An Introduction to Statistical Inference

6.1 Definition

The field of statistical inference involves collecting and interpreting informative data and drawing conclusions. Often, we are interested in the underlying properties of a population given sample data. In practice, this may involve estimating parameters that model the population of interest, and investigating whether the estimates are plausible.

As a side note, descriptive statistics refer to understanding the properties of the observed data and does not concern itself with the idea of this sample data being drawn from a population.

6.2 Estimators

As mentioned in the definition, statistical inference concerns itself with making accurate estimations of some statistic. This could be the mean or the variance, for example.

First let us define a set of n independent and identically distributed (*iid*) random variables drawn from some population of interest:

$$\mathbf{X} = X_1, X_2, X_3, ...X_n$$

each with mean μ and variance σ^2. We then define any function of that vector $\mathbf{X}$ as:

$$T_n(\mathbf{X}) = (X_1, X_2, X_3, ...X_n)' \tag{64}$$

where $T_n(\mathbf{X})$ is called a *statistic*, which estimates some parameter that decribes the properties of the underlying population.

For example, we can define a *statistic* for the mean of some population as

$$T_n^*(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

or we can construct a *statistic* that describes the variance

$$T_n^{**}(\mathbf{X}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

You may have seen the sample mean and variance being denoted as $\bar{X}$ and S^2, and that is the widely accepted notation. However, the $T_n(\mathbf{X})$ is a general notation for any *statistic.*

6.3 Basic Properties of Estimators

We now review some properties of any estimator.

6.3.1 Unbiased Estimators

Let's define what a bias is first. In plain English, bias is how "far off" your estimate is from the real parameter. If an estimator is unbiased, its expectation is equal to the true parameter:

$$E[T_n(\mathbf{X})] - \theta = 0 \tag{65}$$

It is important to note here that the expectation of the *statistic* means all possible distributions of this estimator, and not just the one instance you obtain from the sample data. One way to interpret an unbiased estimator is understanding that the distribution of such an estimator is concentrated around the true value of θ.

In real life it is almost impossible to obtain a sufficiently representative sample of your data from which to produce this unbiased estimate, even if your estimator is theoretically unbiased. If possible, we would prefer an unbiased estimator but in many practical scenarios this may not be feasible or it is favourable to trade some bias for other benefits (e.g. lower variance in MSE).

Example:

Suppose that $X_1, X_2, X_3, ...X_n$ is a random sample from a population with mean μ and variance σ^2.

Show that S_n^2 is an unbiased estimator of σ^2

$$\begin{aligned}
S_n^2 &= \frac{1}{n-1}E\Big[\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\Big] \\
&= \frac{1}{n-1}E\Big[\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2)\Big] \\
&= \frac{1}{n-1}E\Big[\sum_{i=1}^{n}(X_i^2 - \bar{X}_n^2)\Big] \\
&= \frac{1}{n-1}E\Big[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} - \mu^2)\Big] \\
&= \frac{1}{n-1}E\Big[n(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} - \mu^2)\Big] \\
&= \sigma^2
\end{aligned}$$

Sometimes we may want to introduce a **bias-corrected version** of some estimator.

Example:

Suppose that $\hat{\theta}$ is a biased estimator for a parameter θ. If $bias(\hat{\theta}) = c\theta$, find an unbiased estimator θ^* such that $\theta^* = k\hat{\theta}$.

Given

$$\begin{aligned}
bias(\hat{\theta}) &= c\theta \\
&= E[\hat{\theta}] - \theta
\end{aligned}$$

and we know that

$$\theta^* = k\hat{\theta}$$

therefore

$$\begin{aligned}
E[\theta^*] &= kE[\hat{\theta}] \\
&= k(1+c)\theta
\end{aligned}$$

Hence if $\theta^* = \dfrac{1}{1+c}\hat{\theta}$ gives an unbiased estimator for θ^*

It is important also know that if $\hat{\theta}$ is an unbiased estimator of θ, $g(\hat{\theta})$ is **NOT** necessarily an unbiased estimator of $g(\theta)$.

One more thing - there are also mean-unbiased estimators and median-unbiased estimators. Just keep that in mind, and being mean-unbiased does not necessarily mean you are median-unbiased. I would hazard a guess that if the population distribution is symmetric, then they are equivalent.

6.3.2 Consistent Estimators

A consistent estimator $\hat{\theta}_n$ will converge in probability to a true parameter θ as $n- > \infty$. We formally define this as

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0 \tag{66}$$

Sometimes this is referred to as *weakly* consistent.

Methods that are used to show convergence in probability can sometimes also be used to show the consistency of the estimator. It is possible to use the inequality

$$P[h(T_n(X) - \theta) \geq \epsilon] \leq \frac{E[h(T_n(X) - \theta))]}{h(\epsilon)} \tag{67}$$

For example, using the Chebyshev's inequality

$$P[|X - \mu| \geq \epsilon] \leq \frac{E[(\hat{\theta} - \theta)^2]}{\epsilon^2} \tag{68}$$

For an estimator to be *strongly* consistent, the convergence probability has to be *almost surely* convergent. We won't delve into the details here.

As you might have guessed, inconsistent estimators are probably not so good in practice because we cannot 'improve' our estimators by collecting more data.

6.3.3 Bias vs. Consistency

Just because an estimator is biased doesn't mean it is consistent and vice-versa. Let's look at it with some examples.

Suppose you let $T(X) = x_1$ be the estimator of the population mean E[X], where x_i comes from an *iid* sample $x_1, x_2, ..., x_n$. In this case we are simply using the first observed value as our statistic, and sure, $E[T(X)] = X$ because as you average across all possible distributions of $T(X)$ it will be an unbiased estimate, but this test statistic does not converge to anything and thus cannot be called consistent.

What about an unbiased but consistent estimator? Well, consistency is just converging to the true parameter so if $T(X) = \frac{1}{n}\sum_{i=1} nX_i + \frac{1}{n}$, then as $n \to \infty$ the estimate converges.
Okay, let's talk about some common estimators. In particular, we are going to delve into the mean squared error estimator (MSE) and the minimum variance unbiased estimator (MVUE).

6.3.4 The Mean Squared Error (MSE) of an Estimator

You may have since this estimator when studying Linear Regression, but let's break it down anyway. In particular, we look at how the decomposition changes slightly when you're considering a distribution of MSE estimators as opposed to a single estimate from one sample.

$$\begin{aligned}
MSE[\hat{\theta}] &= E[(\hat{\theta} - \theta)^2] \\
&= E[\hat{\theta}^2] - 2E[\hat{\theta}]E[\theta] + E[\theta^2] \\
&= Var[\hat{\theta}] + E[\hat{\theta}]^2 - 2E[\hat{\theta}]E[\theta] + Var[\theta] + E[\theta]^2 \\
&= Var[\theta] + Var[\hat{\theta}] + (E[\hat{\theta} - \theta])^2 \\
&= \sigma^2 + Var[\hat{\theta}] + (E[\hat{\theta} - \theta])^2
\end{aligned} \tag{69}$$

where

σ^2 = irreducible error of the true function describing the population.
$Var[\hat{\theta}]$ = error variance of the estimate.
$(E[\hat{\theta} - \theta])^2$ = bias of the estimate.

Now, which 2 of the 3 are what we commonly refer to as the "bias-variance" trade-off?

6.3.5 Minimum Variance Unbiased Estimator (MVUE)

There's actually a lot to talk about here, we save the bulk of the discussion for Chapter 4 Properties of Estimators. First let's look at the **Cramer-Rao (lower) bound (CRLB)**, which states that the variance of any unbiased estimator is at the very minimum the inverse of the Fisher Information I_n.

Cramer-Rao Bound

Suppose we have some statistic $T(X)$ that is an estimator of θ based on n *iid* observations with $Var(T[X]) \leq \infty$.

In its simplest form, we define this bound as

$$var(\hat{\theta}) \geq \frac{1}{I_n(\theta)}$$

which states that the variance of the estimator must be at the very least the CRLB, which is in this case the inverse Fisher Information which you know is the asymptotic variance of the MLE estimate. So this makes a lot of sense. . A related term, efficiency, is defined as:

$$e(\theta) = \frac{[I_n(\theta)]^{-1}}{var(\hat{\theta})}$$

and we know that $e(\theta)$ is less than one. This is a measure of how good your estimate is with respect to the CRLB.

Suppose your $T(X)$ is an **unbiased estimator** that achieves a variance equal to the CRLB, then it must have the minimum variance amongst all unbiased estimators. Such an estimator is called an **efficient** estimator. Tying it back to the MSE, this unbiased estimator will also have the lowest MSE amongst all unbiased methods.

It is important to note that achieving the CRLB is sufficient for an unbiased estimator to be a MVUE, but it is not a necessary condition.

If we are instead dealing with a function $g(X)$ of the MLE, we define the CRLB as

$$var(\hat{\theta}) \geq \frac{[\psi'(\theta)]^2}{I_n(\theta)} \tag{70}$$

where $\psi(\theta) = E[T(X)])$ is the expectation of the statistic $T(X)$ .... We use the formula for correlation to find the proof.

First of all let's just quickly acknowledge that

$$E(T(X)) = \int T(X)f(x)\,\delta x$$

and that

$$E(\frac{\delta}{\delta x} ln(f(X;\theta))) = 0$$

and that the integral and derivative signs are interchangeable.

Since we are looking for $\psi'(\theta)$

$$\begin{aligned}
\psi'(\theta) &= \frac{\delta}{\delta\theta}\int T(X)f(X;\theta)\,\delta x \\
&= \int T(X)\frac{\delta}{\delta\theta}f(X;\theta)\,\delta x \\
&= \int T(X)\frac{\delta}{\delta\theta}[ln(f(X;\theta))]f(X;\theta)\,\delta x \\
&= E\Big[(T(X)\ \frac{\delta}{\delta\theta}[ln(f(X;\theta))]\Big]
\end{aligned}$$

Knowing that the expectation of the score function is zero given the interchangeability condition, we can then write the above as

$$\psi'(\theta) = Cov\Big[T(X)\ \frac{\delta}{\delta\theta}[ln(f(X;\theta))]\Big]$$

and using the correlation function we get

$$
\begin{aligned}
Corr\Big[T(X)\ \frac{\delta}{\delta\theta}[\ln(f(X;\theta))]\Big] \leq 1 \\
&= \frac{Cov\Big[T(X)\ \frac{\delta}{\delta\theta}[ln(f(X;\theta))]\Big]}{\sqrt{Var[T(X)]\ Var(\frac{\delta}{\delta\theta}[ln(f(X;\theta))])}} \\
&= \frac{\psi'(\theta)}{\sqrt{Var[T(X)]\ I_n(\theta)}}
\end{aligned}
$$

Rearranging the result gives

$$
Var[T(X)] \leq \frac{[\psi'(\theta)]^2}{I_n(\theta)}
$$

Now let's go back and formally define the concept of a MVUE.

Suppose that we are interested in estimating a parameter function $g(\theta)$ from some sample data $X_1, X_2, ...X_n$ which are *iid*, and the samples belong to a family of densities $p_\theta, \theta \in \varTheta$, where $\varTheta$ is the parameter space.

An unbiased estimator $\delta(\mathbf{X})$ of $g(\theta)$ is $MVUE$ if $\forall\ \theta \in \varTheta$,

$$
var(\delta(\mathbf{X})) \leq var(\widetilde{\delta}(\mathbf{X})) \tag{71}
$$

This is quite an important topic that will be discussed further.

6.3.6 Asymptotic Normality of an Estimator

An estimator $\hat{\theta}_n$ is said to be asymptotically normal if

$$
\frac{\hat{\theta}_n - \theta}{SE[\theta_n]} \rightsquigarrow N(0,1) \text{ as } n \to \infty \tag{72}
$$

Now we move to talk about MLE. Originally I thought about discussing further topics of estimators first, but as they involve likelihood functions it might be better to cover MLE completely first before coming back. Rest assured, everything ties in together in statistics!

7 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is one of the classic frequentist methods for parameter estimation given sample data. From a Bayesian perspective however, it is just a special case of Maximum a Posteriori Estimation (MAP) under the assumption of a uniform prior. We'll talk about this later.

Intuitively speaking, the MLE attempts to uncover a set parameters that maximises the probability of the observed sample data. This is mathematically represented by the **likelihood function**, and more practically the **log likelihood function**. From the log likelihood function we can also derive the **Score function**, which is the sensitivity of the likelihood with respect to the parameter of estimation. We also explore the **Fischer Information**, which tells us how influential a single observed point of data is in bringing us closer to the true value of the parameter, as so to speak.

It is really important to understand how all 3 are related, and their uses in parameter estimation and subsequent inference.

So what are we actually trying to estimate?

$$\hat{\theta} \in \{\underset{\theta \in \Theta}{argmax}\ \mathcal{L}(\theta; X)\} \tag{73}$$

7.1 The Likelihood Function

Alright, let's look at the likelihood, or its practical-friendly brother, the log-likelihood function.

We define the likelihood function as

$$\mathcal{L}(\theta; X) = \prod_{i=1}^{n} f(X; \theta) \tag{74}$$

Look at this closely for a moment, notice that we're multiplying all the probabilities together. If you think back to basic probability theory, the product of various probabilities can only be the joint probability right? Also! If you think about why we multiple probabilities, that occurs when the probabilities are independent right? So what does this tell us about the assumptions of using MLE?

As mentioned already, we like to use the log likelihood

$$\begin{aligned} log\big[\mathcal{L}(\theta;X)\big] &= log\big(\prod_{i=1}^{n} f(X;\theta)\big) \\ &= \ell(\theta;X) \end{aligned} \tag{75}$$

From which we differentiate and set the derivative to zero. This will obtain the estimates for the parameters for which the joint probability of the observed data is maximised.

7.1.1 Estimating the MLE

Like any other mathematical concept, it might be easier to digest this with a simple example. After this, we're going to look at deriving the MLE for a normal distribution, which involves 2 parameters and thus is a little bit more tricky.

Example:

Suppose we have $\mathbf{X} = X_1, X_2, ...X_n$ *iid* random variables, each following a Bernoulli distribution with probability θ.

$$X_i \sim B(\theta) \text{ where i = 1, 2, ..., n}$$

$$\begin{aligned} \ell(\theta;X) &= log\big(\prod_{i=1}^{n} f(X;\theta)\big) \\ &= log\big(\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{n-x_i}\big) \\ &= \sum_{i=1}^{n} \big(x_i log(\theta) + (1-x_i) log(1-\theta)\big) \\ &= (\sum_{i=1}^{n} x_i) log(\theta) + (n - \sum_{i=1}^{n} x_i) log(1-\theta) \end{aligned}$$

Now we differentiate and set the derivative to zero

$$\frac{\delta}{\delta\theta}\ell(\theta;X) = \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{1}{1-\theta}(n - \sum_{i=1}^{n} x_i)$$

and by setting the derivative to 0 we get

$$\begin{aligned}\frac{\delta}{\delta\theta}\ell(\theta;X) &= 0 \\ &= \frac{1}{\hat{\theta}}\sum_{i=1}^{n} x_i - \frac{1}{1-\hat{\theta}}(n - \sum_{i=1}^{n} x_i)\end{aligned}$$

and thus

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Note: Have a look at when the $\hat{\theta}$ appeared

You have to remember that closed-form solutions for this does not always exist. That's when numerical optimisation comes in. We also discuss this later. The common probability distributions almost always a nice answer though. And yes, most of the time it's just an average.

7.1.2 Properties of a Maximum Likelihood Estimator

This estimator has some convenient properties, and also many assumptions to go with it. Let's discuss the properties first.

(i) The MLE is **consistent**: $\hat{\theta}_n \xrightarrow{P} \theta$ where θ is the true parameter.

(ii) The MLE is **equivariant**: if $\hat{\theta}_n$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

(iii) The MLE is **asymptotically normal**: $\frac{\hat{\theta}_n - \theta}{\sqrt{I_n^{-1}}} \rightsquigarrow N(0,1)$

(iv) The MLE is **asymptotically optimal** or **efficient**: That means that the MLE has the lowest asymptotic variance among all consistent estimators of θ (link this back to the MVUE and the Rao-Cramer lower bound.)

We will discuss some of these properties shortly.

7.1.3 Assumptions of a Maximum Likelihood Estimator

When the MLE first appeared over 100 years ago by Ronald Fisher, there were many assumptions that were made in order to accommodate for many possibilities. We briefly discuss below.

(i) The model is **identifiable** - without a model to use we can't make any estimations!

(ii) The **support** of the $f_X(x;\theta)$ does NOT depend on θ for all $\theta \in \Theta$.

(iii) The true parameter θ is **not at the boundary** of Θ - remember we are taking derivatives and sometimes boundary estimates may cause issues.

(iv) $I(\theta)$ is **invertible** in a neighbourhood of θ which means that the determinant will be greater than zero. $I(\theta) = 0$ would cause issues because the curve would be totally flat. That means your estimate could be unpredictably bad.

Alright, we now we have an estimate of the parameter which also happens to be a pretty good one. Luckily for us keen statistics students, the fun doesn't end there.

I promise.

7.2 The Score Function

Let's dive into the Score function. This guy is incredibly useful, and ties directly with the Fischer Information which we cover next.

One thing you have to remember is that people like to write this in a few thousand different ways, so it might be easier to just remember this in words. The score function is the derivative of the log-likelihood function. That's it.

We define the Score function as

$$
\begin{aligned}
S(\theta; X) &= \frac{\delta}{\delta\theta} ln[\mathcal{L}(\theta; X)] \\
&= \frac{\delta}{\delta\theta} \ell(\theta; X) \\
&= \sum_{i=1}^{n} \frac{\delta}{\delta\theta} ln[f(\theta; X)]
\end{aligned}
\tag{76}
$$

Our lecturer, not surprisingly, also likes to write things in 50 different ways, so I've just condensed the story down to 3 equations. They are all the same thing. It would be best to convince yourself that is the case, before moving on.

The score function sometimes is explained as the derivative of an the log-likelihood for an individual observation. Other times it is explained as the sum of all observed scores obtained from the sample data. If the score value is evaluated at a specific value of θ in a score test, for example, then the score can also be considered as a statistic.

Either way, just remember it is the derivative of the log-likelihood function. In the multi-parameter case, it is the vector of partial derivatives with respect to the parameter of interest θ.

7.2.1 Some Assumptions about the Score Function

There are some very important **regularity** assumptions we're making about the score function here. Namely,

(i) The **support** of the pdf $\{x : f(x; \theta)\}$ is independent of θ.

(ii) $f(x; \theta)$ is twice differentiable.

(iii) The derivative can be evaluated by **interchanging the order of differentiation and integration** (or summation in the discrete case).

The first assumption is intuitive - you can't keep changing the space of which x belongs with different θ values as you can't consistently estimate them. If the pdf is not twice differentiable, you can't get the Fischer information, which also means it's hard to verify that your estimate is a maximum. If you can't interchange the derivative and integral signs, the relationship between the score and Fischer Information becomes messy. You'll see why in a moment.

With the above 3 assumptions, we can show that $E[S(\theta;X)] = 0$

Example:

Show that the expectation of the score function is zero.

$$\begin{aligned}
E[S(\theta;X)] &= \int S(\theta;X)f(X;\theta)\delta x \\
&= \int \frac{\delta}{\delta\theta} ln[f(X;\theta)]f(X;\theta)\ \delta x \\
&= \int \frac{f^{'}(X;\theta)}{f(X;\theta)} f(X;\theta)\ \delta x \\
&= \int f^{'}(X;\theta)\ \delta x \ \ \text{(and now assuming 3) is true)} \\
&= \frac{\delta}{\delta x}\int f(X;\theta)\ \delta x \\
&= \frac{\delta}{\delta x} 1 \\
&= 0
\end{aligned} \tag{77}$$

Example:

Let's go back to the Bernoulli example again. Remember that we've already derived the derivative of the log likelihood because of our need to estimate the parameter.

$$\begin{aligned}
S(\theta;X) &= \frac{\delta}{\delta\theta}\ell(\theta;X) \\
&= \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{1}{1-\theta}(n - \sum_{i=1}^{n} x_i) \\
\text{and thus} & \\
E\big[S(\theta;X)\big] &= \frac{1}{\theta} nE[X] - \frac{1}{1-\theta} n - nE[X]) \\
&= n - n = 0
\end{aligned}$$

This is, of course, consistent with the the expectation of the score function as we already had shown above.

7.3 Fischer Information

First of all. Let me just clarify something. There is a difference between **Fisher** Information and **Information**.

The exact specifics of Information is left for another resource that is able to cover Information theory in more detail. We will only discuss what is relevant to MLE.

Let's bring out attention back to the Score function for a moment.

I didn't actually explain what the score function is earlier because I felt like the 2 concepts are so interwined that saving it for when I introduce the Fisher Information might help with the understanding.

The score function is essentially a derivative right? And calculus tells us the derivative means the rate of change of the function, which is the likelihood. What this means then, is that each point of data induces a degree of "movement" within the likelihood itself, and the score allows us to quantify that sensitivity.

The Fisher Information is, under some regularity conditions, the variance of the Score function. So they are very closely related in the sense that a bigger variance indicates a sharper curve of $\ell(\theta; x)$ and we can be confident that our estimates provide us a lot of information as it moves closer and close towards that true parameter.

$$\begin{aligned} I_1(\theta) &= Var\Big[\frac{\delta}{\delta\theta}\ell(\theta;X)\Big] \\ &= E\Big[\Big(\frac{\delta}{\delta\theta}\ell(\theta;X)\Big)^2\Big] \text{ as} \\ E\Big[\frac{\delta}{\delta\theta}\ell(\theta;X)\Big] &= E[S(\theta;X)] \\ &= 0 \end{aligned}$$

Sometimes (read as "in your exams") it is easier to compute the Fisher Information in another form:

$$I_1(\theta) = -E\Big[\frac{\delta^2}{\delta^2\theta}\ell(\theta; X)\Big]$$

We'll show a quick proof of their equivalence.

Example:

$$\begin{aligned}
I_1(\theta) &= -E\Big[\frac{\delta^2}{\delta^2\theta}\ell(\theta; X)\Big] \\
&= -\int\Big[\frac{\delta^2}{\delta^2\theta}ln(f(X;\theta))\Big]\,f(X;\theta)\delta x \\
&= -\int\Big[\frac{\delta}{\delta\theta}\frac{f'(X;\theta)}{f(X;\theta)}\Big]\,f(X;\theta)\delta x \\
&= \int\Big[\frac{f''(X;\theta)f(X;\theta) - (f'(X;\theta))^2}{f(X;\theta)^2}\Big]\,f(X;\theta)\delta x \\
&= \int\Big[\frac{(f'(X;\theta))^2}{f(X;\theta)}\Big]\delta x - \int f''(X;\theta)\delta x \\
&= \int\Big[\frac{(f'(X;\theta))^2}{f(X;\theta)}\Big]\delta x + 0
\end{aligned}$$

assuming interchangibility of integral and derivative signs.

Now we calculate it starting from the other form

$$\begin{aligned}
I_1(\theta) &= E\Big[\Big(\frac{\delta}{\delta\theta}\ell(\theta; X)\Big)^2\Big] \\
&= \int\Big[\frac{f'(X;\theta)}{f(X;\theta)}\Big]^2\,f(X;\theta)\delta x \\
&= \int\Big[\frac{(f'(X;\theta))^2}{f(X;\theta)}\Big]\,\delta x
\end{aligned}$$

7.4 The Asymptotic Distribution

Recall earlier that we briefly mentioned the asymptotic distribution of the MLE is **approximately normal**. We'll show a brief explanation of why that is the case, using a combination of the CLT and Taylor's Polynomial Expansion.

Consider a Taylor series expansion of $\ell(\theta; X)$ about the true parameter θ_*.

$$\begin{aligned}\ell(\theta; X) &= \sum_{n=0}^{\infty} \frac{\ell^n(\theta_*)}{n!}(\theta - \theta_*)^n \\ &= \ell(\theta_*) + \frac{\delta\ell(\theta_*; X)}{\delta\theta}(\theta - \theta_*) + \frac{1}{2!}\frac{\delta^2\ell(\theta_*; X)}{\delta\theta^2}(\theta - \theta_*)^2 + ...\end{aligned}$$

Given the regularity conditions, we know that this polynomial will be finite up to the 3rd order. We take that and set the derivative to zero to give

$$\begin{aligned}\frac{\delta\ell(\hat{\theta}; X)}{\delta\theta} &= 0 \\ &= \frac{\delta\ell(\theta_*; X)}{\delta\theta} + \frac{\delta^2\ell(\theta_*; X)}{\delta\theta^2}(\theta - \theta_*)\end{aligned}$$

Rearranging the equation gives and manipulating n gives

$$\sqrt{n}(\hat{\theta} - \theta_*) = \Big[-\frac{1}{n}\frac{\delta^2\ell(\theta_*; X)}{\delta\theta^2}\Big]^{-1}\Big[\frac{1}{\sqrt{n}}\frac{\delta\ell(\theta_*; X)}{\delta\theta}\Big]$$

We can break the above equation down to its 2 parts and solve for its converging result.

$$-\frac{1}{n}\frac{\delta^2\ell(\theta_*; X)}{\delta\theta^2} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\delta^2\ell(\theta_*; X_i)}{\delta\theta^2}$$

and due to the Law of Large Numbers, this empiriral quantity must converge

$$\longrightarrow -E[\ell''(\theta_*; \mathbf{X})]$$

whilst the other part, under the CLT, converges to the a normal distribution with $N(0, I_n(\theta))$ because that guy is the Score function, as we know that it's expectation is zero.

$$\frac{1}{\sqrt{n}}\frac{\delta\ell(\theta_*; X)}{\delta\theta} = \sqrt{n}\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{\delta\ell(\theta_*; X_i)}{\delta\theta}\Big]$$

Now we can put everything together as say that

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta_*) &= I_n(\theta)^{-1} N(0, I_n(\theta)) \\ &= N(0, I_n(\theta)^{-1})\end{aligned}$$

Knowing that placing the inverse Fisher Information into the distribution gives the square.

Phew! When I was staring at this for the first time the lecturer slides wrote it in a slightly different form, and it took me a few minutes to actually recognise all these things. That's why it is important to pick a style and stick to it.

Knowing the limiting distribution of the MLE allows us to find the confidence intervals for our estimates, which we briefly cover.

7.4.1 Confidence Intervals

We'll talk about finding confidence intervals with an example.

Example:
Let $X_1, X_2, ..., X_n$ be a random sample from $Poi(\lambda)$. Find an approximate (asymptotic) 95% CI for λ.
Using the MLE, we can show that $\hat{\theta} = \overline{X}$ and $I_1(\theta) = \frac{1}{\theta}$, which gives

$$\sqrt{n}(\hat{\theta} - \theta) \rightsquigarrow N(0, \theta)$$

Since this is an asymptotic normal and we want a 5% alpha

$$P\left[\left|\frac{\hat{\theta} - \theta}{\sqrt{\theta/n}}\right| \leq \Phi^{-1}(0.95)\right] = 0.95$$

where

$$\begin{aligned}\left[\frac{\hat{\theta} - \theta}{\sqrt{\theta/n}}\right]^2 &\leq \Phi^{-1}(0.95)^2 \\ &= 1.96^2\end{aligned}$$

which gives

$$\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2 \leq 1.96^2 \cdot \frac{\theta}{n}$$
$$n\theta^2 - (2n\hat{\theta} + 1.96^2)\theta + n\hat{\theta}^2 \leq 0$$

a quadratic in θ. Thus, when we solve this using the quadratic equation, we get our desired confidence interval bounds which is

$$\theta_\ell = \overline{X} + \frac{1.96^2}{2n} - \frac{1.96}{\sqrt{2n}}\sqrt{2\overline{X} + \frac{1.96^2}{2n}}$$
$$\theta_u = \overline{X} + \frac{1.96^2}{2n} + \frac{1.96}{\sqrt{2n}}\sqrt{2\overline{X} + \frac{1.96^2}{2n}}$$

7.5 Putting It All Together

Alright. We all this new knowledge, we're ready to try computing MLE and its paraphernalia for the Gaussian distribution

Example:

Consider a vector $\mathbf{X} = X_1, X_2, ...X_n$ of *iid* random variables from a Gaussian (normal) distribution.
We know that it's probability density function is:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{X-\mu}{\sigma})^2}$$

7.6 Functions of the MLE - Delta Method

The Delta method is extremely useful as it allows us to find the asymptotic distribution of a continuous and differentiable function of an MLE (this can be generalised to any consistent and asymptotically normal estimator).

Suppose you have a maximum likelihood estimator with an asymptotic distribution

$$\sqrt{n}(\hat{\theta} - \theta) \sim AN(0, I_1(\theta)^{-1})$$

Then the distribution of $g(\hat{\theta})$

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \sim AN(0, [g'(\hat{\theta})]^2 I_1(\theta)^{-1}) \tag{78}$$

We show a quick proof, using the Mean Value Theorem (see in the Appendix).

$$g'(\theta^*) = \frac{g(X_n) - g(\theta)}{X_n - \theta}$$

and rearranging gives

$$g(X_n) = g(\theta) + g'(\theta^*) g(X_n) - g(\theta)$$

where $X_n \leq \theta^* \leq \theta$ and $X_n \xrightarrow{\rho} \theta$, giving $\theta^* \xrightarrow{\rho} \theta$. If we assume that $g'(\theta)$ is continuous, and we can apply the **continuous mapping theorem** to yield

$$g'(\theta^*) \xrightarrow{\rho} g'(\theta)$$

The continuous mapping theorem states that continuous functions are **limit-preserving** even if their arguments are sequences of *random variables.*

And because

$$\sqrt{n}(X_n - \theta) \xrightarrow{\infty} N(0, \sigma^2)$$

we can use the **Slutsky's theorem** and argue that because the first derivative of the function $g(\theta)$ is a constant, the asymptotic distribution of the MLE function becomes

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \sim AN(0, [g'(\hat{\theta})]^2 I_1(\theta)^{-1})$$

Utilising this in practice is not difficult.

Example:

Suppose that $X_n \sim B(n, \theta)$ where X_i's are *iid* with probability θ.
Let $g(\theta) = \hat{\theta}(1 - \hat{\theta})$. Find the asymptotic distribution of $g(\theta)$

$$g'(\theta) = 1 - 2\theta$$

and the Fisher information is

$$I_n(\theta) = n\theta(1 - \theta)$$

$$\begin{aligned}\sqrt{n}(g(\hat{\theta}) - g(\theta)) &\sim AN(0, [g'(\hat{\theta})]^2 I_1(\theta)^{-1}) \\ &\sim AN(0, [1-2\theta]^2 \theta(1-\theta))\end{aligned}$$

The multivariate case is similar.

7.7 Newton-Raphson Method

The Newton-Raphson method allows us to iteratively improve our estimations towards the root of a function using the derivative of the function. This method stems from something you have seem from high school:

$$y - y_0 = m(x - x_0)$$

which is the equation for the tangent at point (x_0, y_0).
This diagram might help:
We take the approximation for x_{n+1} at y_0 and thus the equation becomes

$$\begin{aligned}0 - y_0 &= m(x - x_0) \text{ and} \\ x &= x_0 - \frac{y_0}{m}\end{aligned}$$

Now, just to make the connection clear, we know that m is the gradient, which is often written as $g'(x_0)$.

Thus we write the Newton-Raphson method as:

$$x = x_0 - \frac{g(x_0)}{g'(x_0)}$$

The MLE case is exactly the same, replacing $g(x)$ by $\ell'(\theta; X)$, and it looks like:

$$\hat{\theta}_{i+1} = \hat{\theta}_i - \frac{\ell'(\theta; X)}{\ell''(\theta; X)} \tag{79}$$

remember that you're looking for the root of the **derivative of the log-likelihood!**

As with any numerical method, we have to set the number of iterations or some condition for the Newton-Raphson to stop. The rate of convergence for a zero of multiplicity one is at least quadratic.

Example:

Write down the Newton-Rapson algorithm (based on n observations) to find the MLE from the Cauchy distribution, with

$$f_X(x) = \frac{1}{\pi} \cdot \frac{1}{1+(x-\theta)^2}, \quad -\infty < x < \infty$$

Now we need $\ell'(\theta; X)$

$$\begin{aligned}
\ell(\theta; X) &= log\Big[\prod_{i=1}^{n} \frac{1}{\pi} \cdot \frac{1}{1+(x-\theta)^2}\Big] \\
&= -nlog(\pi) - \sum_{i=1}^{n} log(1+(x-\theta)^2) \text{ and the derivative is} \\
\ell'(\theta; X) &= \Big[\sum_{i=1}^{n} \frac{2(x-\theta)}{1+(x-\theta)^2}\Big] \text{with Hessian} \\
\ell''(\theta; X) &=
\end{aligned}$$

7.8 Method of Moments

We close this section by discussing the "Method of Moments (MOM)", which is an alternative method to estimate parameters.

Recall that:

$$m_r = \mu_r^{'} = E(X^r) \tag{80}$$

is the r^{th} non-central moment.

and we estimate each theoretical moment by

$$\hat{m}_r = \hat{\mu}_r^{'} = \frac{1}{n}\sum_{i=1}^{n} X_i^r \text{ r} = 1,\, 2,\, ... \tag{81}$$

Thus, if we have with m unknown parameters we need at least m equations. Often, if we are trying to estimate higher k theoretical moments, our solution will not be unique due to the $m > k$ and thus we take a **generalised method of moments** (GMM) approach which involves using a cost function. We won't talk about GMM here.

So technically with MOM we are equating the empirical with the theoretical. So if the observed data is indeed from this known distribution, then it is expected that

$$m_r \approx \hat{m}_r$$

Let's look at the $X \sim$ **Pois**(λ) distribution and see how MOM estimates work.

$$
\begin{aligned}
m_1 &= E[X] \\
&= \sum_{i=1}^{\infty} i\big(\frac{e^{-\lambda}\lambda^i}{i!}\big) \text{ i = 0, 1, ...} \\
&= \lambda e^{-\lambda}\big(\frac{\lambda^{i-1}}{(i-1)!}\big) \\
&= \lambda e^{-\lambda}(e^{\lambda}) \\
&= \lambda \\
\text{and therefore we have} \\
\hat{m} &= \hat{\lambda} = \overline{X}_n \\
\text{What about the 2nd moment?} \\
m_2 &= E[X^2] \\
&= \sum_{i=1}^{\infty} i^2\big(\frac{e^{-\lambda}\lambda^i}{i!}\big) \text{ i = 0, 1, ...} \\
&= \sum_{i=1}^{\infty} (i(i-1)+i)\big(\frac{e^{-\lambda}\lambda^i}{i!}\big) \\
&= \sum_{i=1}^{\infty} i(i-1)\big(\frac{e^{-\lambda}\lambda^i}{i!}\big) + \sum_{i=1}^{\infty} i\big(\frac{e^{-\lambda}\lambda^i}{i!}\big) \\
\text{and expanding the sum and simplifying leads to} \\
&= e^{-\lambda}\cdot\lambda^2(1+\lambda+\lambda^2+...)+\lambda \\
&= e^{-\lambda}\cdot\lambda^2\cdot e^{\lambda}+\lambda \\
&= \lambda^2+\lambda \text{which means we need to solve} \\
\hat{m}_2 &= \overline{X_n^2} = \lambda^2+\lambda
\end{aligned}
$$

Let's end our discussion with a multi-parameter case.
Suppose that $\mathbf{X} = X_1, X_2, ...X_n$ is from a $N(\mu, \sigma^2)$

$$
\begin{aligned}
m_1 &= E[X] = \mu \quad \text{and} \\
\hat{m}_1 &= \overline{X}_n \\
\text{therefore} \\
\hat{\mu} &= \overline{X}_n \\
\text{alright, now we want another equation for } \sigma^2 \\
m_2 &= E[X^2] \\
&= \sigma^2 + \mu^2 \\
\text{so we simply solve for} \\
\hat{\sigma}^2 &= \overline{X}_n - \hat{\mu}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2
\end{aligned}
$$

8 Exponential Family

We briefly discuss the **exponential family** which is an extremely useful group of probability distributions. There are a lot of familiar faces like the Gaussian, Bernoulli, Poisson, Exponential, Gamma, Beta, Dirichlet etc. These guys are extremely useful, both in their specific forms and properties.

Have a look at the form of this family and then break down its individual components, and we will then reveal why this family is so useful for many applications.

We'll start simple with the one-parameter form, and then move to the scalar, vector parameter form. They are of course, identical conceptually.

8.1 Scalar-parameter Form

We'll start with the simplest one-parameter case. Even here, there are numerous ways to represent the form of this family.

In the lecture notes, it is written as

$$\begin{aligned} f(x;\theta) &= \exp\left[c(\theta)\cdot T(X) + d(\theta) + s(x)\right]\mathbb{1}_A(x) \text{ or} \\ &= \exp\left[c(\theta)\cdot T(X) + d(\theta)\right]\mathbb{1}_A(x)\cdot h(x) \text{ or} \\ &= \exp\left[c(\theta)\cdot T(X)\right]\mathbb{1}_A(x)\cdot h(x)\cdot d(\theta) \end{aligned} \tag{82}$$

where $c(\theta)$, $T(X)$, $d(\theta)$ and $s(x)$ are suitable (non-unique) functions.

Some people also like to use this notation.

$$f(x;\theta) = \exp\left[\eta(\theta)\cdot T(X) - A(\eta) + S(x)\right]\mathbb{1}_A(x)\text{or} \tag{83}$$

where $d(\theta)$ has been re-parametrised to contain the canonical parameter $\eta(\theta)$, which when $\eta(\theta) = \theta$ is said to be in its **canonical form**. In the canonical form, its corresponding $T(X)$ becomes the link function for GLMs when determining the distribution of the response variable.

More on this later (well, GLMs probably belong in a different course).

Just aware that people like to explain one of the 2 forms whilst completely disregarding the poor student who tries to read material elsewhere and ends up confused (or at least I did.)

Let's showcase 2 examples of this.

Example:

Re-write the following probability functions of the distributions in its exponential form.

1) Bernoulli

$$\begin{aligned} f(x;\theta) &= \exp\Big[c(\theta)\cdot T(X) + d(\theta) + s(x)\Big]\mathbb{1}_A(x) \\ &= \exp\Big[\log\big((\theta)^x(1-\theta)^{n-x}\big)\Big] \\ &= \exp\Big[xlog(\theta) + (n-x)log(\theta)\Big] \\ &= \exp\Big[xlog\Big(\frac{\theta}{1-\theta}\Big) + nlog(1-\theta)\Big] \end{aligned}$$

where

$$\begin{aligned} c(\theta) &= log\Big(\frac{\theta}{1-\theta}\Big) & d(\theta) &= nlog(1-\theta) \\ T(X) &= x & s(x) &= 1 \end{aligned}$$

2) Normal Distribution with known variance.

$$\begin{aligned} f(x;\theta) &= \exp\Big[c(\theta)\cdot T(X) + d(\theta) + s(x)\Big]\mathbb{1}_A(x) \\ &= \exp\Big[-\frac{1}{2}\Big(\frac{X-\mu}{\sigma}\Big)^2 - log(\sqrt{2\pi\sigma^2})\Big] \\ &= \exp\Big[-\frac{X^2}{2\sigma^2} + \frac{X\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - log(\sqrt{2\pi\sigma^2})\Big] \end{aligned}$$

where

$$c(\theta) = \frac{\mu}{\sigma^2} \qquad d(\theta) = \frac{-\mu}{2\sigma^2}$$
$$T(X) = x \qquad s(x) = \frac{-X^2}{2\sigma^2} - log(\sqrt{2\pi\sigma^2})$$

We will showcase the situation with both unknown mean and variance later.

8.2 Properties

The exponential family has an important property, where the **support of x does not depend on the parameter(s) θ of the distribution**. Hence the uniform distribution is immediately discounted from being a member of this family.

The exponential family is **minimal** if $c(\theta)$ and $T(X)$ are linearly independent. This can be achieved with re-parametrisation.

Amongst all the different forms you see, the η form is the most important.

8.2.1 The η Form

The form with η (83) is actually called the **natural or canonical parameterisation** of the exponential family. This holds important properties that gives us many of results you have seem in other areas of statistics.

The first one is of course to do with $T(X)$, which is the **natural, sufficient statistic** of the distribution (more on sufficient statistics in the next chapter).

As mentioned before the $\eta(\theta)$ is called the natural (canonical) parameter, whilst $A(\eta)$ is the **log-partition function** because of its role in normalising the probability density such that it can be integrated to 1.

$$\begin{aligned}
\int f(x;\theta)\delta x &= \int \exp\big[\eta(\theta)\cdot T(X) - A(\eta) + s(x)\big]\mathbb{1}_A(x)\delta x \\
&= \exp(-A(\eta))\int \exp\big[\eta(\theta)\cdot T(X) + s(x)\big]\mathbb{1}_A(x) \\
&= 1 \text{ and hence} \\
\exp(A(\eta)) &= \int \exp\big[\eta(\theta)\cdot T(X) + s(x)\big]\mathbb{1}_A(x) \\
A(\eta) &= log\Big[\int \exp\big[\eta(\theta)\cdot T(X) + s(x)\big]\mathbb{1}_A(x)\Big]
\end{aligned}$$

where $A(\eta)$ can be differentiated to find the first and second moments of the test statistic $T(X)$.

To show this, we must be able to assume that the gradient and integral signs can be interchanged (by the **dominated convergence theorem**). Let's have a look.

$$\begin{aligned}
A(\eta) &= log\Big[\int \exp\big[\eta(\theta)\cdot T(X) + s(x)\big]\mathbb{1}_A(x)\Big] \\
\frac{\delta A}{\delta\eta} &= \frac{\delta}{\delta\eta}\left[log\int \exp\big[\eta(\theta)\cdot T(X) + s(x)\big]\,\delta x\right] \\
&= \frac{\int T(X)\exp\{\eta(\theta)\cdot T(X) + s(x)\}\,\delta x}{\int \exp\{\eta(\theta)\cdot T(X) + s(x)\}\,\delta x} \\
&= \frac{\int T(X)\exp\{\eta(\theta)\cdot T(X) + s(x)\}\,\delta x}{\exp\{A(\eta)\}} \\
&= \int T(X)\exp\{\eta(\theta)\cdot T(X) - A(\eta) + s(x)\}\,\delta x \\
&= E[T(X)]
\end{aligned}$$

and the $Var[X]$) is

$$
\begin{aligned}
\frac{\delta^2 A}{\delta \eta^2} &= \frac{\delta A}{\delta \eta} \int T(x) \exp\{\eta(\theta) \cdot T(X) - A(\eta) + s(x)\} \, \delta x \\
&= \int T(X)(T(X) - \frac{\delta}{\delta \eta} A(\eta)) \exp\{\eta(\theta) \cdot T(X) - A(\eta) + s(x)\} \, \delta x \\
&= \int T(X)(T(X) - E[T(X)]) \exp\{\eta(\theta) \cdot T(X) - A(\eta) + s(x)\} \, \delta x \\
&= E[T(X)^2] - E[T(X)E[T(X)]] \\
&= E[T(X)^2] - E[T(X)]^2 \\
&= Var[X]
\end{aligned}
$$

8.3 Vector-parameter Form

The vector parameter form looks similar, but it I include it for the sake of completeness. Take note of the inner product between η and θ.

$$
f(x; \boldsymbol{\theta}) = \exp \Big[\mathbf{c}(\boldsymbol{\theta}) \cdot \mathbf{T}(X) + \mathbf{d}(\boldsymbol{\theta}) + \mathbf{s}(x) \Big] \mathbb{1}_A(x)
$$

Example:

Let's revisit the Gaussian distribution again, but this time with both parameters unknown.

$$
\begin{aligned}
f(x; \theta) &= \exp \Big[\mathbf{c}(\boldsymbol{\theta}) \cdot \mathbf{T}(X) + \mathbf{d}(\boldsymbol{\theta}) + \mathbf{s}(x) \Big] \mathbb{1}_A(x) \\
&= \exp \Big[-\frac{1}{2} \Big(\frac{X - \mu}{\sigma} \Big)^2 - log(\sqrt{2\pi\sigma^2}) \Big] \\
&= \exp \Big[-\frac{X^2}{2\sigma^2} + \frac{X\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - log(\sqrt{2\pi\sigma^2}) \Big]
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{c}(\boldsymbol{\theta}) &= \Big[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \Big] & \mathbf{d}(\boldsymbol{\theta}) &= \frac{-\mu}{2\sigma^2} + log(|\sigma|) \\
\boldsymbol{T}(X) &= [x, x^2] & s(x) &= log(\sqrt{2\pi})
\end{aligned}
$$

We have to be careful to distinguish from the vector parameters and vector of variables as well.

9 Properties of Estimators

This is such a huge topic. So where do I begin? Let's start by thinking about the whole purpose of statistics. We want to make sense of some data, and by that I mean trying to estimate parameters that give us information about the distribution of a population. We do that by utilising that sample data we've collected, but the sample data is never perfect, and consequently neither will our estimates of the parameters.

So we need to evaluate the quality of the estimates we make. Previously we touched upon the mean squared error, the bias and consistency. Now, we dive deeper into slightly more complicated notions of a quality estimate.
Before we do that though, I want to state our end goal - finding a UMVUE for any statistic $T(X)$. In order to do that, we'll need to confirm that several properties for our statistic do hold, and apply Rao-Blackwell*isation*, a really awesome property which, when combined with the Lehmann-Scheffe's Theorem, gives us what we want.
First let's explore what other properties a statistic might have and return to the Rao-Blackwell Theorem once we are have the tools we need.
Let's start with **sufficiency**.

9.1 Sufficiency

It's easy to say that we're looking for a sufficient statistic $T(\mathbf{X})$ that gives us information about the parameter of interest θ and sufficiency occurs when we can obtain sufficient information from this statistic, and we can't get anything more out of $T(X)$.

That's basically going in circles. Try again.

Let's go back to basics. Given some data, what do we use to estimate the parameters? Take the expectation? Think more generalised. We just spent a whole section on this. Exactly, we tend to use the maximum likelihood!

9.1.1 An intuitive approach with MLE

One way to think about sufficiency is whether our statistic $T(X)$ can be used to compute an estimate using some method. If we can use this $T(X)$ guy to calculate our likelihood, then we can say that it's sufficient. Let's start with an example.

Example:

Suppose you have an *iid* sample of data $X_1, X_2, ...X_n$ and this distribution follows a Bernoulli.

We define a test statistic $T(X) := \sum_{i=1}^{n} x_i$

The log-likelihood function for a Bernoulli is

$$\begin{aligned}\ell(\theta;\mathbf{X}) &= \prod_{i=1}^{n} f(\mathbf{X};\theta)\\ &= p^{\sum x_i}(1-p)^{n-\sum x_i}\\ &= p^{T(X)}(1-p)^{T(X)}\end{aligned}$$

and thus we can see that $T(X)$ enables us to calculate the likelihood and thus we can say it's sufficient.

9.1.2 Sufficiency with Conditional Distribution

That might be good enough, but some may be wondering, where is the math in this? Well there is a much more nicer way to define sufficiency, and that method has to do with conditional distributions.

Consider a statistic $T(X)$ for some *iid* $\mathbf{X} = X_1, X_2, ..., X_n$, we say that this statistic is sufficient if the conditional distribution of $\mathbf{X}$ given $T(X)$ is independent of the parameter θ.

In math, this means

$$P_\theta(X = x|T = t) = P(X = x|T = t) \tag{84}$$

If we go back to the Bernoulli example:

$$
\begin{aligned}
P_\theta(X = x|T = t) &= \frac{P_\theta\big((X = x) \cap P_\theta(T = t)\big)}{P_\theta(T = t)} \\
&= \frac{P_\theta(X = x)}{P_\theta(T = t)} \\
&= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&= \frac{1}{\binom{n}{t}} \\
&= P(X = x|T = t)
\end{aligned}
$$

Checking for sufficiency this way may sometimes be infeasible if the conditional distribution is difficult to compute. There exists an even easier way if we utilise the **Factorisation Theorem**.

9.1.3 Fisher-Neyman Factorisation Theorem

The Factorisation theorem allows us to (easily) check whether a statistic $T(X)$ is sufficient by rewriting the pdf as

$$f(X;\theta) = g(T(X), \theta)h(x) \tag{85}$$

which decomposes the pdf down into a product where one factor does not depend on X and the other interacts with θ through X.

I want to look at an example of this with the Gaussian.

Example:

Let $\mathbf{X} = X_1, X_2, ..., X_n$ be *iid* $N(\mu, \sigma^2)$ with unknown μ and unknown σ^2, and let $\theta := (\mu, \sigma^2)$. The joint density of $\mathbf{X}$ is given by

$$
\begin{aligned}
f(X;\theta) &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\Big[-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-u)^2\Big]\\
&= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\Big[-\frac{n\mu^2}{2\sigma^2}\Big]\exp\Big[-\frac{1}{2\sigma^2}\Big(\sum_{k=1}^{n}(x_k)^2-2\mu\sum_{k=1}^{n}x_k\Big)\Big]\\
&\text{where}\\
g(T(X),\theta) &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\Big[-\frac{n\mu^2}{2\sigma^2}\Big]\exp\Big[-\frac{1}{2\sigma^2}\Big(\sum_{k=1}^{n}(x_k)^2-2\mu\sum_{k=1}^{n}x_k\Big)\Big]\\
&\text{and}\\
h(x) &= 1
\end{aligned}
$$

This would mean that

$$
T(\mathbf{X}) := \Big(\sum_{k=1}^{n}X_k, \sum_{k=1}^{n}X_k^2\Big)
$$

is a sufficient statistic. However, the statistic for the variance isn't something that is commonly used, and in practice it might be better to use

$$
T(\mathbf{X}) := \Big(\bar{X}_n, \frac{1}{n}\sum_{k=1}^{n}(X_k-\bar{X}_n)^2\Big)
$$

if the variance is known, the result is similar, we treat σ^2 as a constant.

Knowing this theory, we can revisit the exponential family and show that $T(X)$ is indeed the sufficient statistic, since the form can be rewritten as:

$$
\begin{aligned}
f(X;\theta) &= \exp\big[c(\theta)\cdot T(X)+d(\theta)\big]\mathbb{1}_A(x)\cdot h(x)\\
&= g(T(X),\theta)\cdot h(x)
\end{aligned}
$$

where

$$
g(T(X),\theta) = \exp\big[c(\theta)\cdot T(X)+d(\theta)\big]
$$

and

$$
h(x) = \mathbb{1}_A(x)\cdot h(x)
$$

Hence the result.

9.2 Minimal Statistic

So we can find a sufficient statistic $T(X)$, this would mean this provides enough information, we can afford to discard our observed data and still be able to compute an estimate. The existance of a sufficient statistic is unique though, and many statistics may be sufficient. Rather than deciding which one to choose, we elect to use the one with the smallest dimension which should hopefully make our lives easier.
When I was taught this concept, we were simply told to use this formula and I didn't quite grasp the motivation. Hopefully the paragraph above helps make the method ot demonstrate sufficiency more intuitive.
Suppose you had 2 test statistics $T(X)$ and $T(Y)$. The statistic $T(X)$ is minimal if and only if the ratio of the 2 likelihoods are independent of θ when $T(X) = T(Y)$.
That is:

$$\frac{\ell(\theta; X)}{\ell(\theta; Y)} \tag{86}$$

is independent of θ IFF $T(X) = T(Y)$ for some predefined $T(X)$

9.3 Complete Statistic

Alright, so let's complete our exploration.
We say a statistic is complete for θ if $T(X)$ expressed as a function $g(T(X)) \longrightarrow \mathbb{R}$ defined for the range $T(X)$ that satisfies

$$E[g(T(X))] = 0 \text{ for } \theta \in \Theta \tag{87}$$

which also happens to imply that $P_\theta(g(T(X)) = 0) = 1$

9.4 Rao-Blackwell Theorem

This is where things start to get interesting. The **Rao-Blackwell theorem** demonstrates a means of improving our estimators when evaluating them with respective to the MSE.

Suppose you are interested in a parameter θ and $S(X)$ is an estimate of that parameter.

Rao-Blackwell Theorem states that the conditional expectation of $S(X)$ given $T(X)$, where $T(X)$ is a sufficient statistic, is at least better than the original estimator $S(X)$ if $E[|S(X)|] < +\infty$. To show this, we recall that we have

$$S^*(X) := E\big[S(X)\,|\,T(X)\big] \tag{88}$$

and the result

$$E\Big[(S^*(X)-\theta)^2\Big] \leq E\Big[(S(X)-\theta)^2\Big] \tag{89}$$

with equality if $S*(X) = S(X)$ if $Var[S(X)] < +\infty$

In order to prove this, we require a few results that we've learnt previously. Recall that the MSE can be decomposed into the Variance and bias components. We first show that the bias of both $S(X)$ and $(S^*(X)$ are equal, and then the variance of $S^*(X)$ offering a smaller variance, completing the proof.

$$\begin{aligned} E[S(X)-\theta]^2 &= E\big[E[S(X)|T(X)]-\theta\big]^2 \\ &= E\big[E[S^*(X)]-\theta\big]^2 \end{aligned}$$

as $E[X] = E\big[E[X|Y]\big]$. Hence the bias are equal. Now for the inequality of variances;

$$\begin{aligned} Var\big(S(X)\big) &= Var\big[E[S(X)|T(X)]\big] + E\big[Var(S(X)|T(X))\big] \\ &\geq Var\big[E[S(X)|T(X)]\big] \\ &\geq Var\big(S^*(X)\big) \end{aligned}$$

by the definition of $S^*(X)$ and using $Var(X) = Var(E[X|Y]) + E[Var(X|Y)]$.

Alright. So now we have an estimator $S^*(X)$ that has a lower MSE compared to the original $S(X)$ after Rao-Black*isation*. However, this is still not as optimal as we would like. Instead of finding a better estimator, we are more interested in the "best" estimator. In this case, we are interested in achieving the minimum variance. However, it turns out that if we were given an unbiased estimator $S(X)$ and $T(X)$ satisfies a few more conditions, we can find the Uniform Minimum Variance Unbiased Estimator (UMVUE).

To get there, we'll hop into the Lehmann-Scheffe's Theorem.

9.5 Lehmann-Scheffe's Theorem

This time, instead of just any $S(X)$, suppose that $S(X)$ is an unbiased estimator. If there exists a $T(X)$ that is **minimal**, **complete** and **sufficient**, $S^*(X) := E\big[S(X)|T(X)\big]$ is the UMVUE.
Now we try to show this.

10 Optimal Tests

All the back way in first year we were taught hypothesis testing and what types of tests to use for different types of data. However, we never really touched upon how these tests came about. How do we even that these tests are effective?

10.1 Revision of Assumed Knowledge

Simple and Composite Hypothesis

A simple hypothesis completely specifies the parameter of interest whereas the composite hypothesis does not. Typically in hypothesis testing, a range is specified. E.g.

Simple Hypothesis:

$$H_0 : \theta = \theta_0$$

Composite Hypothesis:

$$H_0 : \theta > \theta_0$$

Note that if you have 2 parameters in a distribution and one of them is not specified, then the hypothesis is a composite one as the distribution was not completely specified!

In hypothesis testing, a performance indicator may be the probability of the outcome of the test, the p-value, is correct whether that is the accepting or rejecting the null hypothesis.

‘

To this effect, there are 2 types of errors which we summarise in the table below:

	Accept H_0	Reject H_0
H_0 is true		Type I Error (α)
H_1 is true	Type II Error (β)	Power ($1-\beta$)

The **Power** of a test defines the probability for a test to correctly reject.

We care about this more because, as you will see below, many of the 'good' hypothesis tests we make involve holding α fixed and trying to maximise the power.

This is different to the **power function**, which is defined as

$$\beta(\theta) := P_\theta(\text{rejection}), \quad \theta \in \Theta \tag{90}$$

and the power function can represent different things depending on the sample space θ is part of. If $\theta \in \Theta_0$, then $\beta(\theta)$ is the probability of a Type I Error. If $\theta \in \Theta_1$ then it is the "Power" of a hypothesis test as defined before. A subtle difference to take note of.

Of course, we are always making a trade-off between α and β and they are not linearly inversely proportional to each other.

Example:
Suppose we have a new drug which is believed to help with recovery from a disease. From previous studies, the probability of recovery for a group of patients is 20% without any treatment. Let R_n be the number of patients that have recovered after using the drug. Treating the outcome for each patient as binary, we can model the number of successful treatments as a binomial distribution.

Thus, we have

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta > \theta_0 \end{aligned}$$

Now, if we are interested in the Type I Error then we assume $\theta_0 = 0.2$ is true.

$$P(R_n \geq k) = \sum_{i=k}^{n} \binom{n}{i} (\theta_0)^i (1-\theta_0)^{n-i}$$

Comparatively, the Type II Error is

$$P(R_n < k) = \sum_{i=0}^{k-1} \binom{n}{i} (\theta_0)^i (1-\theta_0)^{n-i}$$

Suppose that $\theta_1 = 0.5$, $k = 9$, $n = 20$, then we have

$$\begin{aligned} P(R_{20} \geq 9) &= \sum_{i=9}^{20} \binom{20}{i} (0.2)^i (1-0.2)^{20-i} \\ &= 0.01 \end{aligned}$$

$$\begin{aligned} P(R_{20} < 9) &= \sum_{i=0}^{9-1} \binom{20}{i} (0.5)^i (1-0.5)^{20-i} \\ &= 0.252 \end{aligned}$$

Given a set of hypothesises and a known distribution to model the problem, we can calculate the related statistics such as power and Type I Error. However, we need a more systematic way of finding a good test given any situation. This is where the Neyman-Pearson Lemma comes in.
Before we jump in, we need to define the Best Critical Region.

Best Critical Region

Consider a hypothesis test with a simple alternative $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$. If there are several rejection regions with significance $P(x \in A; \theta_0) = P(x \in B; \theta_0)) = ... = \alpha$, then $P(x \in A; \theta_0)$ is the **Best Critical Region** if for all other rejection regions $R = \{B, ...\}$

$$P(x \in A; \theta_1) \geq P(x \in B; \theta_1) \tag{91}$$

which maximises the Power.

10.2 Neyman-Pearson Lemma

You might remember that I said we are often interested in fixing the α and then maximising the Power of a test. This is exactly what the **Neyman-Pearson Lemma** attempts to formalise. Under a simple hypothesis, there exists a 'best' test that maximises the power such that Type I Errors are held at a constant.

Suppose that we are interested in testing the unknown parameter θ, from a probability distribution $f_X(x;\theta)$, by constructing a simple null and alternative hypothesis as follows

$$H_0 : \theta = \theta_0, \qquad \text{vs.} H_1 : \theta = \theta_1$$

First, we decide on the significance level α and impose this constraint on our critical region. This is

$$P(X \in A) = \int_S \mathbb{1}_A(x) f_X(x;\theta_0)\ \delta x \leq \alpha$$

where A is the rejection region for X and $A \subseteq S$ where $S \subseteq \mathbb{R}$ is the support of $f_X(x;\theta_0)$.

With this in-place we now maximise Power of the test

$$P(X \in A) = \int_S \mathbb{1}_A(x) f_X(x;\theta_1)\ \delta x$$

These two concepts allow us to then define the **Best Critical Region (BCR)** as

$$A_c^* := \{x \in S | f_X(x;\theta_1) - c f_X(x;\theta_0)\} \tag{92}$$

where c is chosen to obey the constraint

$$\int_S \mathbb{1}_{A_c^*}(x) f_X(x;\theta_0)\ \delta x = \alpha$$

This satisfies the constraint we placed earlier. We still need to show that A_c^* is the optimal solution to this problem though.

Since we have chosen the *BCR* to be equal to α, in theory this should yield a larger rejection region compared to any value less than the chosen α. Hence we have

$$\begin{aligned}\int_S \mathbb{1}_A f(x;\theta_1)\ \delta x - c\alpha &\leq \int_S \mathbb{1}_A f(x;\theta_1)\ \delta x - c\int_S \mathbb{1} f(x;\theta_0)\ \delta x\\ &= \int_S \mathbb{1}_A (f(x;\theta_1 - cf(x;\theta_0))\ \delta x\\ &\leq \int_S \mathbb{1}_{A_c^*} (f(x;\theta_1 - cf(x;\theta_0))\ \delta x\\ &= \int_S \mathbb{1}_{A_c^*} f(x;\theta_1)\ \delta x - c\alpha\end{aligned}$$

which shows that the BRC is greater than the rejection region A.

Example:

You have a sequence of random variables $X_1, X_2, ..., X_n$ from a normal distribution with mean μ and $\sigma^2 = 16$. Find the most powerful test for

$$\begin{aligned}H_0 :&\mu_0 = 5\\ H_1 :&\mu_1 = 10\end{aligned}$$

where $n = 16$, and $\alpha = 0.05$.

Since the distribution is specified completely, we can apply the Neyman-Pearson Lemma here

$$A_c^* = \left\{x \in [0, 1, ..., n]| \frac{f(x;\mu_1)}{f(x;\mu_0)} \geq k\right\}$$

$$\begin{aligned}k &\leq \frac{\left[\frac{1}{\sqrt{2(16)\pi}}\}\right]^{16} \exp\left[\frac{-1}{2(16)}\sum_{i=0}^{16}(x_i - 10)^2\right]}{\left[\frac{1}{\sqrt{2(16)\pi}}\}\right]^{16} \exp\left[\frac{-1}{2(16)}\sum_{i=0}^{16}(x_i - 5)^2\right]}\\ &\leq \exp\left[\frac{-1}{2(16)}\left(\sum_{i=0}^{16}(x_i - 10)^2 - \sum_{i=0}^{16}(x_i - 5)^2\right)\right]\\ leq &\exp\left[\frac{-1}{2(16)}\left(-10\sum_{i=0}^{16} x_i + 1200\right)\right]\end{aligned}$$

Rearranging to make $\sum x_i$ the subject, we have

$$\begin{aligned}\sum x_i &\geq \frac{32\ln(k)+1200}{10}\\ \bar{x} &\geq \frac{32\ln(k)+1200}{160}\\ &= k^*\end{aligned}$$

Hence the rejection region for the most powerful test is given by $\bar{x} \geq k^*$. Now we can obtain a k^* where α is constrained to be at most 0.05. Hence we are looking for

$$\begin{aligned}0.05 &= \int_{k^*}^{\infty} f(x;\mu_0)\ \delta x\\ &= \int_{k^*}^{\infty} \frac{1}{(\sqrt{2(16)\pi})^{16}} \exp\left[-\frac{(x-5)^2}{2(16)}\right]\ \delta x\end{aligned}$$

which you can find in R with $\text{pnorm}(0.95, 5, 1)$.

The Neyman-Pearson Lemma finds the BRC for only simple hypothesis tests, and ideally we would like to find the best test for a composite hypothesis as well.

10.3 Uniformly Most Powerful Tests

Uniformly Most Powerful (UMP) tests do not exist in all cases, but in the event that it does, a *UMP* test is one that maximises the Power for every parameter value in a composite alternative hypothesis. From the previous example, in the case of a normal distribution, there exists a UMP test where the value of k^* is independent of μ_1 once we know whether $\mu_1 > 0$ or $\mu_1 < 0$.

Let's look at the Binomial distribution and see if we can get a UMP test.

Example:

Consider the binomial distribution with $f(x;\theta) = \binom{n}{x}(\theta)^x(1-\theta)^{n-x}$ and that we are interested in testing for

$$\begin{aligned} H_0 :&\theta = \theta_0 \\ H_1 :&\theta \neq \theta_0 \end{aligned}$$

$$A_c^* = \left\{ x \in [0, 1, ..., n] | \frac{f(x; \mu_1)}{f(x; \mu_0)} \geq k \right\}$$

hence we have after some simplifying

$$k \leq \left[\frac{1 - \theta_1}{1 - \theta_0} \right]^n \cdot \left[\frac{\theta_1(1 - \theta_0)}{\theta_0(1 - \theta_1)} \right]^x$$

We know that $\left[\frac{1-\theta_1}{1-\theta_0} \right]^n > 0$, and after some checking we see that

$$\frac{\theta_1(1 - \theta_0)}{\theta_0(1 - \theta_1)} \in \begin{cases} (1, +\infty) & \text{if } \theta_1 > \theta_0 \\ 1 & \text{if } \theta_1 = \theta_0 \\ (0, 1) & \text{if } \theta_1 < \theta_0 \end{cases}$$

together, we cannot find a UMP because the critical regions for these possibilities are all different. However, if we have a one-sided hypothesis test then there exists a UMP, e.g. $H_1 : \theta > \theta_0$, where

$$x \geq \left(\ln \left[\frac{\theta_1(1 - \theta_0)}{\theta_0(1 - \theta_1)} \right] \right)^{-1} \cdot \ln(k')$$

where $k' = k \cdot \left[\frac{1-\theta_1}{1-\theta_0} \right]^{-n}$. In we are looking for $H_1 : \theta < \theta_0$ instead, then

$$\begin{aligned} x &\geq \left(\ln \left[\frac{\theta_1(1 - \theta_0)}{\theta_0(1 - \theta_1)} \right] \right)^{-1} \cdot \ln(k') \\ &\geq \left(-\ln \left[\frac{\theta_0(1 - \theta_1)}{\theta_1(1 - \theta_0)} \right] \right)^{-1} \cdot \ln(k') \\ &\leq \left(\ln \left[\frac{\theta_0(1 - \theta_1)}{\theta_1(1 - \theta_0)} \right] \right)^{-1} \cdot \ln(k') \end{aligned}$$

since we know that $\ln \left[\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)} \right] < 0$ for $\theta_1 < \theta_0$.

Now let's look an example where a UMP does not exist.

Example:

Suppose we are interested in finding an UMP test for the Cauchy distribution, that has $f(x;\theta)$
Our null and alternative are

$$\begin{aligned} H_0 :&\theta = \theta_0 \\ H_1 :&\theta > \theta_0 \end{aligned}$$

By the Neyman-Pearson Lemma, we have

$$\begin{aligned} k &\leq \frac{f(x;\theta_1)}{f(x;\theta_0)} \\ &\leq \frac{\pi(1+(x-\theta_0)^2}{\pi(1+(x-\theta_1)^2)} \end{aligned}$$

which when you try to simplify, you get a quadratic that you don't really want:

$$\begin{aligned} k &\leq \frac{\pi(1+(x-\theta_0)^2}{\pi(1+(x-\theta_1)^2)} \\ k-1 &\leq (x-\theta_0)^2 - k(x-\theta_1)^2 \\ k-1-\theta_0^2+\theta_1^2 &\leq x^2(1-k) - 2x(\theta_0+\theta_1) \end{aligned}$$

And it's a little bit hard to actually quantify what we are looking for, unlike the previous two examples. If you draw out the likelihood ratios, then you will be able to see that the rejection region changes depending on the alternate hypothesis. For example, $\theta_1 = 5$ vs. $\theta_1 = 10$ would yield different BCRs. Thus there does not exist a UMP for the Cauchy distribution.

Actually, there is an important property we did not talk about that is important for determining whether a UMP exists.

10.3.1 Monotone Likelihood Ratio

The **monotone likelihood ratio** refers to the ratio of 2 PDFs. Suppose we have $f(x;\theta_A)$ and $f(x;\theta_B)$, then for every $x_1 > x_0$ in the support of X,

$$\frac{f(x_1;\theta_A)}{f(x_1;\theta_B)} \geq \frac{f(x_0;\theta_A)}{f(x_0;\theta_B)} \tag{93}$$

and we say that this distribution has the MLR property in $T(X)$ if for a test statistic $T(X)$ in the likelihood ratio the property holds. Think back to the example with the normal distribution. Can you see what the $T(X)$ was?

For a $T(X) = \sum x_i$, the UMP exists for the exponential, binomial, poisson and normal (if σ is known).
It turns out, that whilst a one-sided alternative in a

Example:

10.3.2 Uniformly Most Powerful Unbiased Test

An **unbiased hypothesis test** is defined as the probability of a correct rejection is always greater than the probability of rejecting incorrectly.

In other words, we are looking to constrain the Power to be at equal to or larger than the α.

$$P_\theta(X \in A) \leq \alpha, \;\; \theta \in \Theta_0 \tag{94}$$

$$P_\theta(X \in A) \leq \alpha, \;\; \theta \in \Theta_0 \tag{95}$$

for the rejection region A. An unbiased test with uniformly higher power is called the uniformly most powerful unbiased (UMPU) test.

10.4 Likelihood Ratio Tests

We discussed most powerful tests for simple hypothesise and extended this to composite hypothesises. Now, we want to generalise this concept to any test.

Consider a hypothesis test as follows:

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta_1 \end{aligned}$$

We define the **Likelihood Ratio** as

$$\Lambda = \frac{\sup\{\mathcal{L}(\theta; x) : \theta \in \Theta_1\}}{\sup\{\mathcal{L}(\theta; x) : \theta \in \Theta_0\}}$$

When certain asymptotic properties are met, the logarithmic test statistic will converge to a chi-squared distribution

$$2 \ln [\Lambda] \overset{n \to \infty}{\longrightarrow} \chi^2_p \tag{96}$$

under the assumption that the null hypothesis is true, where p is the difference in dimensionality between Θ_1 and Θ_0.

The p-value is determined by $P(\chi^2_p < 2 \ln [\Lambda])$

11 The Bayesian

Alright, we're wrapping to a different dimension now. Buckle your seatbelts because we're going Bayesian.

In this section we give an introduction to Bayesian methods and attempt to uncover the why Bayesian methods are becoming more popular nowadays and of course their advantages compared to the classical frequentist approach.

11.1 Introduction - Frequentist vs. Bayesian Thinking

What makes Bayesian different?

Let's talk about the philosophies of the **frequentist** approach vs. the **bayesian** ideology.

In classical statistics, probability is defined as a fixed parameter. What is the probability of a coin toss? Well, it's 50% for an unbiased coin, but if it is biased, that probability is p where $\{p \in (0,1)\}$ and the correct answer is just a number. As we increase our sample size, the parameter estimate should converge to the "true" probability.

The Bayesian philosophy, as I read somewhere, comes from an "evidence-based" approach. Sure, you know that coin tosses usually end up in a 50/50. What if you toss a coin 1000 times and get 300 heads? Does that mean the coin is biased? How would you check that? Toss it another 1000 times?

Regardless of how you deal with biased coins, Bayesian thinking says you can describe that probability p with a distribution. And what's more interesting is that you can construct this distribution from both the observed data AND your prior knowledge about this event.

In practice, Bayesian inference allows you to incorporate information from different sources and multiple levels of randomness. If you're wondering how this works, keep wondering. I was wondering for a while too.

Alright, let's try and make sense of the concept of Bayesian using our prior knowledge.

11.2 The Mechanics of Bayesian

We mentioned before that Bayesian methods concern themselves with probability distributions more so than point-estimates of probability. However, a gentle introduction to the idea of prior and likelihoods using point-estimates will give you a taste of why we set up the Bayes theorem

Not surprisingly, Bayesian methods resemble the Bayes theorem. Let's break it down.

$$\begin{aligned} p(\theta|y_{1:n}) &= \frac{p(y_{1:n}|\theta)p(\theta)}{\int p(y_{1:n}|\theta)p(\theta)\ \delta\theta} \\ &= \frac{p(y_{1:n}|\theta)p(\theta)}{c_n} \\ &\propto p(y_{1:n}|\theta)p(\theta) \end{aligned} \tag{97}$$

where

$$c = p(y_{1:n}) = \int p(y_{1:n}|\theta)p(\theta)\ \delta\theta$$

is called the **normalising constant** or **marginal likelihood**. The proportion expressed in the last line is often called the **un-normalised posterior density**.

Let's break this down more methodically. In Bayesian $p(\theta|y_{1:n})$ is what we're trying to figure out. We call this the **posterior probability distribution**, the probability distribution obtained after observing the data.

$p(y_{1:n}|\theta)$ is the "sampling density" which is proportional to the Likelihood. This is the information you can obtain from the data itself. $p(\theta)$ is then the prior distribution for the parameter.

We use the proportion here because the normalising constant acts to make the posterior density proper. So in a sense, we have the information we want simply from the prior and likelihood.

11.2.1 Single-Parameter Models

The first example we're going to look at is the Bernoulli distribution.

Example: Bernoulli Distribution
Let X be a random variable from the Bernoulli distribution with probability mass function $f(x) = \theta^x(1-\theta)^x$. Suppose the prior density is uniform distribution with $U \sim (0,1)$. Assume that we are interested in a sample of X's of size n.

(i) Find the posterior distribution.

(ii) Using (i) or otherwise, find the Bayes estimator of θ.

(i) Recall that the posterior distribution is given by:

$$\begin{aligned} f(\theta|x) &= \frac{f(x|\theta)\cdot f_\theta(\theta)}{f_X(x)} \\ &= \frac{f(x|\theta)\cdot f_\theta(\theta)}{\int f_{X|\theta}(x|\theta)f_\theta(\theta)\ \delta\theta} \end{aligned}$$

The likelihood function is

$$f(x|\theta) = \binom{n}{x}(\theta)^x(1-\theta)^{n-x}$$

and the prior is

$$f_\theta(\theta) = 1$$

And lastly the marginal is

$$\int f_{X|\theta}(x|\theta)f_\theta(\theta)\ \delta\theta = \int \binom{n}{x}(\theta)^x(1-\theta)^{n-x}\ \delta\theta$$

The integral of the marginal is a beta distribution where

$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 (\theta)^{\alpha-1}(1-\theta)^{\beta-1}\ \delta\theta$$

Putting all of this together gives you

$$
\begin{aligned}
f(\theta|x) &= \frac{\binom{n}{x}(\theta)^x(1-\theta)^{n-x}}{\binom{n}{x}\frac{\Gamma(x+1)\Gamma(n-y+1)}{\Gamma(n+2)}} \\
&= \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-y+1)}(\theta)^x(1-\theta)^{n-x}
\end{aligned}
$$

The posterior and prior have the same distribution, which is convenient for reasons we shall discuss later. Just note that it is important for now.
(ii) Since we are looking for the Bayes estimator of θ. Let the Bayes estimator be $\hat{\theta} = E[\theta|x]$ and $y = \sum_{i=1}^{n} X_i$ gives

$$
\begin{aligned}
E[\theta|x] &= \frac{\Gamma(n+2)}{\Gamma(\sum X+1)\Gamma(n-y+1)}\int_0^1 \theta\ (\theta)^{\sum x}(1-\theta)^{n-\sum x} \text{ let } y = \sum X \text{ for convenience} \\
&= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\int_0^1 \theta\ (\theta)^{y}(1-\theta)^{n-y} \\
&= \frac{\Gamma(n+2)\Gamma(y+2)\Gamma(n-y+1)}{\Gamma(x+1)\Gamma(n-y+1)\Gamma(n+3)} \\
&= \frac{y+1}{n+2} \\
&= \frac{\sum X_i + 1}{n+2}
\end{aligned}
$$

11.3 Non-informative Priors

Disciplines belonging to the school of **subjectivism** believe that the prior distribution should reflect our subjective opinion and be consistent with available prior information.

However, it may not always be the case we have previous information to work with, or it may not be feasible to use. Hence if we want to utilise Bayesian techniques we will need to minimise the influence of the prior distribution on the posterior. In this case we adopt what is commonly called the **non-informative prior**.

An obvious choice is to simply use a constant c, where $f_\theta(\theta) \propto c$ but take care to note that

$$\int f_\theta(\theta)\ \delta\theta = \infty$$

which means that this prior distribution is not a probability density. Such a prior is called an **improper prior**. Improper priors can still be used with the likelihood to obtain the unnormalised posterior, but we must be careful to check that this posterior

Another thing to note is that flat priors are not necessarily **transformation invariant**. Applying a $g(X)$ transformation does not result in another flat prior.

11.3.1 Jeffrey's Prior

Instead of using a straight uniform distribution or priors obtained by evaluating the limit of conjugate distributions, we can use a prior that is proportional to the Fisher Information.

We define **Jeffrey's Prior** as:

$$f_\theta(\theta) \propto [I_n(\theta)]^{\frac{1}{2}}$$

The Jeffrey's Prior is dependent on the set of parameter variables chosen to describe the parameter space. Use of the Jeffrey's Prior violates the strong version of the likelihood principle.

We'll quickly discuss the **likelihood principle**.

The likelihood principle states that, for a given statistical model, all the evidence relevant for parameter estimation is held within the likelihood function. What this means in human language is that, because 2 likelihood functions are equivalent is one is a scalar multiple of the other, the conclusions we draw about the parameter of interest should not change depending on the observable X that we used.

The **strong version** of the **likelihood principle** asserts that this should hold when considering sequential case studies or observed sets of sample data.

Interestingly enough, undertaking hypothesis tests and drawing conclusions using p-values often violates this principle and we say that this is due to experimental design.

Example:

The prior for the Bernoulli distribution would be

$$f_\theta(\theta) \propto [\theta(1-\theta)]^{-\frac{1}{2}}$$

which is a beta density $\beta(\frac{1}{2}, \frac{1}{2})$, which is very close to a uniform density.

If a non-informative prior is undesirable, we can use a **weakly informative prior** that is adjusted to either constrain the posterior to a reasonable distribution, use a highly informative prior and broaden it to account for uncertainties in our estimates.

11.4 Bayesian Linear Regression

One of the immediate applications that we see in Bayesians statistics is towards linear regression. Instead of considering the response y_i to be a single point, we instead are interested in realising the response as a distribution, and subsequently estimate the distribution of β_i coefficients.

I'm going to assume that you've done a substantial amount of work on linear regression and understand roughly how it works. So just to freshen your mind, I'll just quickly walk through it again to save you some Googling.

In linear regression, we have a response y_i to be tagged to a series of predictors represented by x_{ij} for the i^{th} observation and j^{th} feature variable. Suppose we have n samples and p predictors, then we have $y \in \mathbb{R}^{n\times 1}$ with $x \in \mathbb{R}^{p\times n}$ and $\beta \in \mathbb{R}^{p\times 1}$ for $i = 1, 2, ..., n$ and $j = 0, 1, 2, ..., p$. Note that we've included the intercept turn by letting $x_{ij} = 1$ if $j = 0$.

$$y_i = x_i^T \beta + \epsilon_i$$

where ϵ_i are independent, identically distributed normal variables with $\epsilon \sim N(0, \sigma^2)$. This is considered to be the 'true model scenario' where the true values of the response can be modelled with a residual error that is supposed to account for all the latent factors that cannot be captured.

Suppose we want to estimate y_i's with our model, we first need to find the β coefficients with our sample data by either minimising the residual sum of squares (RSS),

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{n}(y_i - \hat{y}) \\ &= \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta}) \end{aligned}$$

or via maximum log likelihood (MLE),

$$\hat{\beta} = (X^TX)^TX^TY$$

and both methods are equivalent provided that the response is assumed to be a Gaussian distribution.

Provided we satisfy some assumptions, we can take the estimate

$$\hat{y}_{i'} = x_{i'}^T\hat{\beta}$$

for a new sample $x_{i'}$.

Alright, now let's talk about Bayesian linear regression. Since we consider the response variables y_i to be follow a normal distribution, it is then natural to assume that in the Bayesian approach the distribution of the likelihood is the same.

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp\Big(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}^T\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}^T\boldsymbol{\beta})\Big)$$

So 'all' we need to do now is pick a prior. Not all priors give you an analytical solution, but we can overcome this issue with **Monte Carlo methods** that we will discuss shortly.

Before that, let's quickly talk about preferably a conjugate one.

11.4.1 Using a Conjugate Prior

Since we have 2 unknown parameters, we need a conjugate prior that has the same functional form as the likelihood with respect to β and σ^2. In order to find something like this, let's rewrite the exponential section of the likelihood as follows:

$$(\mathbf{y}-\mathbf{X}^T\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}^T\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}^T\hat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}^T\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$$

Rewriting the likelihood function in this form yields

$$p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2) \propto (\sigma^2)^{-\frac{v}{2}} \exp\Big(-\frac{vs}{2\sigma^2}\Big) \cdot (\sigma^2)^{-\frac{n-v}{2}} \exp\Big(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\Big)$$

where $vs = (\mathbf{y}-\mathbf{X}^T\hat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}^T\hat{\boldsymbol{\beta}})$ and $v = n-k$, where k is the number of regression coefficients.
Looking at this form, we can use a neat trick to break the prior density as follows:

$$p(\beta,\sigma^2) = p(\beta|\sigma^2)p(\sigma^2)$$

where

$$p(\boldsymbol{\beta}|\sigma^2) \propto (\sigma^2)^{-\frac{k}{2}} \exp\Big(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu_0})^T\boldsymbol{\Lambda}_0(\boldsymbol{\beta}-\boldsymbol{\mu_0})\Big)$$

with $p(\boldsymbol{\beta}|\sigma^2)$ following a normal distribution having parameters $N(\boldsymbol{\mu_0},\sigma^2\Lambda_0^{-1})$. $\boldsymbol{\Lambda}_0$ is sometimes called the **precision matrix** which determines the strength of the prior.

$$p(\sigma^2) \propto (\sigma^2)^{-\frac{v_0}{2}-1} \exp\Big(-\frac{v_0 s_0}{2\sigma^2}\Big)$$

which follows a inverse-Gamma distribution $\gamma(\alpha_0,\beta_0)$ with $\alpha_0 = \frac{v_0}{2}$ and $\beta_0 = v_0 s_0^2$.

I don't think we've discussed the **inverse-Gamma distribution** before, but this distribution has probability density function

$$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\gamma(\alpha)}(\frac{1}{x})^{\alpha+1}\exp\big(-\frac{\beta}{x}\big)$$

and mean $E[X] = \frac{\beta}{\alpha-1}$ and variance $Var[X] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$
Of course, you are free to use your own priors as well.

11.5 Monte Carlo Integration

So what if you decide to use a non-conjugate prior? Or what if the chosen prior gives an non-closed form of the posterior distribution? In this case, we will need to use numerical integration to evaluate the posterior and its predictive distribution.

This were **Monte Carlo methods** come in. Monte Carlo methods are a family of computational algorithms for obtaining numerical results using random sampling. By using randomness, this method aims to estimate a deterministic outcome that, in the case of Bayesian statistics, allows us to map the posterior distribution given a likelihood and prior without necessarily having to find a closed-form solution.

In this section we're going to set up the Monte Carlo method with some motivation, and in the next few sections we will discuss some basic algorithms that are used to do Bayesian computation.

Let's start with an easy example to however draw you into the world of Monte Carlo.

Example:

What is the probability of obtaining 3, 6 or 9 heads if a coin is flipped 10 times?

Using Monte Carlo methods, we would

(i) Define $X_i \sim \mathcal{U}(0,1)$ with the decision boundary $x_i < 0.5$ as tails, otherwise heads.

(ii) Define $S =$ the number of desired outcomes.

(iii) Simulate 10 draws from $X_i \sim \mathcal{U}(0,1)$, and increment S if there are 3, 6 or 9 heads using the decision rule in (i).

(iv) Repeat (iii) a desired number of N times (a lot!).

(v) Calculate the simulated probability S/N.

Of course, we won't need to do numerical simulation a lot of times as we can easily evaluate the probability by summing the appropriate binomial distributions. We are assuming here that the asymptotic value of $P(X_n) \to P(X)$, which holds due to the law of large numbers

One of the most common applications of the Monte Carlo method is Monte Carlo integration which, as you might have guessed, means applying Monte Carlo methods to evaluating integrals. Consider an definite integral with limits a and b. In the situation that direct evaluation is impossible or too cumbersome, we choose to use Monte Carlo integration.

11.5.1 Our goal

We can use Monte Carlo integration to evaluate properties of distributions, probabilities, confidence intervals etc. and for Bayesian computation we have a specific goal.

For our purposes, we are looking to evaluate

$$\mathcal{I} = E[h(X)] = \int h(x)f(x)\ \delta x \tag{98}$$

where X could be high-dimensional, and more often than not $\mathcal{I}$ is analytically intractable. Technically $h(X)$ could be anything;

$$\begin{aligned} E[\theta; \mathbf{y}] &= \int \theta f(x)\ \delta x \\ Var[\theta; \mathbf{y}] &= \int (\theta - E[\theta; \mathbf{y}])^2 f(x)\ \delta x \\ P(\theta; \mathbf{y}) &= \int \mathbb{1}_A(x) f(x)\ \delta x \end{aligned}$$

where $\mathbb{1}_A(x) = 1$ if $\theta \in A$

.

Suppose that we are able to generate N *iid* samples $X_i \sim f(x)$ for $i = 1, 2, ...N$. We can use this random sample to evaluate the integral

$$\begin{aligned} \hat{\mathcal{I}}_N = E[\theta; \mathbf{y}] &= \int \theta f(x)\ \delta x \\ &\approx \frac{1}{N}\sum_{i=1}^{N} h(X_i) \end{aligned}$$

As we mentioned earlier, this works due to the law of large numbers where $\hat{\mathcal{I}}_N \xrightarrow{p} \mathcal{I}$. Also, $Var[\hat{\mathcal{I}}_N] = Var[h(X_N)]$ regardless of the dimension of the integral. As you can imagine, this poses problem with high-dimensional problems. Choosing random points on the 'grid' will help, but it has its limitations.

So now we know what we're looking for, the next step is to generate random samples from the prior distribution. Complete randomness by a computer is difficult. So can we generate truly random samples from a deterministic machine?

Let's discuss this.

11.5.2 Random Number Generation

In order to ensure randomness, we seek the aid of **Random Number Generators** - algorithms that generate a series of N numbers starting from s_0 to s_N using the transformation rule $G(x)$ where $s_i = G(s_{i-1})$ for $i = 1, 2, ..., S$ such that $s \in \mathbb{R}$.

The most basic random number generator is the **Linear Congruential Generator**. This algorithm is quite simple and has a lot of tweaks to it, but the basic formula is

$$N_{j+1} = (aN_j + c) \bmod (m) \tag{99}$$

where $0 < a < m$ is the multiplier, $0 \leq c < m$ is the increment, $0 \leq X_0 < m$ is the seed and m the modulus. As a minimal, you can set $a = 7^5 = 16807$, $c = 0$, and $m = 2^{31} - 1$. We can make improvements into the algorithm by setting multiple seeds so that we don't get stuck in a local area or if you need a higher period you can use multiple generators and find their lowest common multiple period.

In practice though, try to use algorithms from the GNU Scientific Library.

Whatever algorithm we chose to implement, it is also good to test that the sequence of numbers follows an *iid* uniform distribution. A conservative test we can use is the **Kolmogorov-Smirnov test** which comes from the family of **empirical distribution function (EDF)** tests. Other less conservative tests include the **Cramer-von Mises test**, and the **Anderson-Darling**

test.

Note: You can use the **Shapiro-Wilk test** for testing normality (only).

11.5.3 Inverse Method

If an inverse exists for the CDF $F(x)$ of a given PDF $f(x)$, then we can directly sample from that distribution. We know, of course, that

$$F(w) = \int_0^w f(x)\, \delta x$$

and for any given probability distribution by definition. Hence if we draw y where $y \sim \mathcal{U}(0,1)$ then

$$F^{-1}(Y) = X$$

where X is the random variable used to calculate the PDF/CDF.

Example:

Suppose we have the CDF of an exponential distribution

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

then we sample a draw from $X \sim exp(\lambda)$ as

$$F^{-1}(y) = \frac{-log(1-y)}{\lambda}$$

11.5.4 Rejection Sampling

If the inverse distribution cannot be computed or the CDF is difficult to estimate, we can use **rejection method**.

Suppose you have a probability density distribution $f_X(x)$. We represent this distribution as an improper density with a normalisation constant C. In other words, we have

$$f_X(x) = \frac{k(x)}{C} \tag{100}$$

The rejection method intuitively relies on finding an alternative probability density $g_X(x)$ where $Dg_X(x) \geq k(x)$ for x in the support of $f_X(x)$, the probability function of interest.

Note: I like to think of it as $Dg_X(x) \geq Cf_X(x)$

The rejection sampling method is as follows:

(i) Generate $X_i \sim g_X(x)$ and $U_i \sim U(0,1)$ for $i = 1, 2, ..., n$

(ii) If $U_i \cdot Dg_X(x_i) \leq k(x_i)$, accept X_i and add it to the observed sample data. Otherwise, reject X_i.

(iii) Repeat (ii) until the desired sample size has been obtained.

The probability of accepting X_i is

$$\begin{aligned} P(U_i \cdot Dg_X(x_i) \leq k(x_i)) &= P\Big(U_i \leq \frac{k(x_i)}{Dg_X(x_i)}\Big) \\ &= P\Big(U_i \leq \frac{Cf_X(x)}{Dg_X(x_i)}\Big) \\ &= \frac{C}{D} \end{aligned} \tag{101}$$

This method is extremely inefficient if $f_X(x)$ is very different compared to $g_X(x)$

The rejection sampling method provides a generic method to simulate form any density $f_X(x)$ that is known up to a normalising constant. It is also useful for sampling from posterior distributions. However, it might be difficult to select $g(X)$ and D.

Let's run through a quick example before we move to importance sampling.

Example:

Suppose we have a truncated normal density where $f_X(x) \propto k(x) = \exp(-\frac{x^2}{2})\mathbb{1}_{x\geq 0}$. We use the exponential distribution $exp(\lambda)$ where $g_X(x) = \lambda\exp(-\lambda x)$.

This gives us

$$\begin{aligned}\frac{k(x)}{g(x)} &= \frac{1}{\lambda}\exp\Big(-\frac{x}{2}+\lambda x\Big)\\ &\leq \frac{1}{\lambda}\exp(\frac{\lambda^2}{2})\end{aligned}$$

This might seem a little arbitary, but consider a normal distribution with mean λ and variance $\sigma^2 = 1$. Without the normalisation constant, it might look like:

$$\frac{1}{\lambda}exp\Big[-\frac{1}{2}(\lambda - x)^2)\Big] \geq \frac{1}{\lambda}$$

Expanding the brackets and rearranging gives

$$\begin{aligned}\frac{1}{\lambda}exp\Big[-\frac{1}{2}(\lambda^2 - 2\lambda x + x^2)\Big] \geq \frac{1}{\lambda}\\ \frac{1}{\lambda}exp\Big[-\frac{1}{2}(\lambda^2)\Big] &\geq \frac{1}{\lambda}exp\Big[\frac{1}{2}(-2\lambda x + x^2)\Big]\\ \frac{1}{\lambda}exp\Big[\frac{1}{2}(\lambda^2)\Big] &\geq exp\Big[-\frac{1}{2}(-2\lambda x + x^2)\Big]\end{aligned}$$

since we have an upper bound, we look for the minimum to maximise the efficiency of the algorithm.

$$\begin{aligned}\frac{\delta}{\delta\lambda}\Big[\frac{1}{\lambda}\exp(\frac{\lambda^2}{2}\Big] &= \exp\Big(\frac{\lambda^2}{2}\Big)(1-\lambda^2) \text{ and finding the roots gives}\\ 0 &= (1-\lambda)(1+\lambda)\end{aligned}$$

Hence the upper bound is minimised at $\lambda = 1$, so the best $g(x)$ is $exp(1)$.

$$\frac{k(x)}{g(x)} \leq D = \sqrt{e}$$

11.5.5 Importance Sampling

Importance sampling involves generating samples using a probability density that differs from the true probability density related to the distribution of interest. Unlike rejection sampling, where you focus on having a distribution "above" the distribution of interest, importance sampling introduces a density that may differ quite dramatically from the density of the distribution itself.

Normally, we would draw samples from the density of interest in order to approximate the distribution of interest. Sometimes we may not be able to do that so we need an alternative. However, even if we can, it turns out that if we choose a good distribution, you can create an even better estimator when sampling from a distribution from which the .

Suppose you have a density function $f_X(x)$ and you are interested in a $E[h(X)]$ where $h(X)$ comes from $f_X(x)$. We choose a distribution $q_X(x)$ from which we will sample from instead of $f_X(x)$

Now, we re-write the expectation of $E[h(X)]$ as

$$\begin{aligned} E[h(X)] &\approx \frac{1}{n}\sum_{i=1}^{n} h(X_i) \\ &= \int h(X) f_X(x)\ \delta x \\ &= \int h(X)\frac{f_X(x)}{q_X(x)} q_X(x)\ \delta x \end{aligned}$$

where $q_X(x) = 0$ implies $f_X(x) = 0$. This is known as **absolute continuity**.

We can approximate the above as

$$\begin{aligned} E[h(X)] &\approx \frac{1}{n}\sum_{i=1}^{n} h(X)\frac{f_X(x)}{q_X(x)} \\ &= \frac{1}{n}\sum_{i=1}^{n} h(X_i) w(X_i) \end{aligned}$$

where $w(X_i)$ is called the **importance weights**.They are equivalent because we are drawing from $q_X(x)$ instead of $f_X(x)$.

This is still a Monte Carlo estimator and since we know that the Monte Carlo estimator is unbiased, then this estimator is unbiased as well!

If the variance of the weights are finite, then we can calculate the effective sample size as

$$S_{\text{eff}} = \frac{1}{\sum_{s=1}^{S} [w_{norm}(X_s)]^2} \tag{102}$$

where $w_{norm(X_s)}$ is the normalised weight defined as

$$w_{norm(X_s)} = \frac{w(X_s)}{\sum_{s=1}^{S} w(X_s)} \tag{103}$$

Importance sampling is not effective if the weights differ substantially across the support. It is also important to check that the estimate is not dominated by outliers, especially as the tails. Lastly, we can use the jackknife to estimate the variance of the importance sampling estimator.
Alright, we now we've discussed some approaches to generating random samples, whether that's using a uniform distribution first to reach the function, or via other methods like rejection/importance sampling.
Now we look at methods that are available to use when it is not possible to sample your parameter of interest θ directly from the posterior distribution.

11.6 Markov Chain Monte Carlo

In this section we deal with using Monte Carlo methods based on **Markov Chains**, which is a type of stohastic model describing a sequence of events where the probability of each event is dependent only on the state attained in the previous event. This means that for some event at x_{t+1}

$$P(x_{t+1}|x_0, x_1, ..., x_t) = P(x_{t+1}|x_t)$$

where X is a random variable in $X_t : t \in T$ with $X \in \chi$, called the **state space** and T being the **index set** (in this case time).

The motivation for using **Markov Chain Monte Carlo (MCMC)** is such that it is extremely difficult to propose distributions in high dimensions that cover the true distribution adequately and accurately. In one dimension, we may be able to propose an appropriate $q_X(x)$ for importance sampling, but this may not be the case in high dimensions.

The key point here is to construct a Markov chain with an stationary distribution identical to the distribution you want to obtain a sample from. As we know that this stationary distribution is reached as a limiting distribution.

Recall that the empirical expectation is obtained if we sample from a large enough *iid* sample, but in Markov Chains the X_t's are correlated. In order to make this work, we'll need the **Ergodic Theorem**. The Ergodic theorem is defined as:

If $X_1, X_2, X_3, ...X_t$ is irreducible (all states communicate), , discrete AND time-homogeneous (transition operator is constant through time, i.e. independent of n. This is also called stationary Markov Chain.) with stationary distribution π then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{n\to\infty}{\longrightarrow} E[h(X)] \text{ where } X \sim \pi$$

Note: Transition operator is the probability of jumping from one class to another.

Furthermore if the Markov Chain is aperiodic (periodicity = 1), then

$$P(X_n = x | X_0 = x_0) \overset{n\to\infty}{\longrightarrow} \pi(x)$$

What this means is that the distribution of X_n is converging to the distribution of $\pi(n)$, and thus if we run the simulation for a long time regardless of its initial state, we will be able to reach a value of X_n that we can use to be our sample, as it is close to the distribution of interest $\pi(x)$ that we've set to be our stationary distribution. This is only an asymptotic guarantee, so sometimes it helps to pick a good initial distribution or know the rate of convergence.

If all these properties are satisfied, we call the Markov Chain an **Ergodic Markov Chain**. Alright, so now we have what we need.

Let's discuss 2 examples of such algorithms, the Metropolis-Hastings and Gibbs-Sampling algorithms which are both based on the random-walk Markov Chains.

11.6.1 Metropolis-Hastings Algorithm

Let's start our discussion with the **Metropolis Algorithm**, (which is different to the Metropolis-Hastings algorithm!) the original MCMC algorithm. After this, we extend our discussion to the Metropolis-Hastings algorithm to make the motivation clear.

First we define a **proposal matrix** or distribution, in either the discrete or continuous cases. A proposal matrix is also sometimes known as a stohastic matrix, which is non-negative and the rows sum to 1.

We look at the continuous case. Suppose we have a $f_X(x)$ that is proportional to the desired probability distribution.

(i) Select a proposal distribution $p_X(x)$ and an initial state x_0. The proposal distribution must be symmetric.

(ii) Sample from the distribution using a random walk model, where $x^* = x_t + \epsilon$ with $\epsilon \sim p_X(x)$.

(iii) Sample from a uniform $U_t \sim U(0,1)$.

(iv) For x^* we can reject or accept the value. We accept the new value if $u_t \leq \frac{f_X(x^*)}{f_X(x_t)}$ and set $x_{t+1} = x^*$. Else we retain the old value x_t and have $x_{t+1} = x_t$.

(v) Continue (ii)-(iv) until you have enough values.

(vi) Discard the early values (known as **burning phrase**) which are influenced by the choice of initial state x_0

For $p_X(x)$, some common choices are Gaussian, uniform, Student's t-distribution, Levy's-flight update (do not ask me what this is) etc.

If a Gaussian is chosen for ϵ then the variance σ^2 becomes a tuning parameter.

Sometimes we define $\alpha(x^*, x_t) = \min\left(\frac{f_X(x^*)}{f_X(x_t)}, 1\right)$ as the acceptance function and compare this to u_t.

Intuitively, why this algorithm works is because we always accept a shift towards to a point that has higher probability than the correct state. In the other case, there is a random chance that we will reject the move. This will lead to the returned samples approximately following the distribution.

Now we introduce the Metropolis-Hastings algorithm, where our proposed distribution is no longer symmetric.

Now, instead of using a straight sample using ϵ and plugging it into $f_X(x)$, we now use another $Q_X(x)$ which acts as a **Hastings Ratio** to correct the bias.
Instead of using

$$\alpha(x^*, x_t) = \min\Big(\frac{f_X(x^*)}{f_X(x_t)}, 1\Big)$$

and have instead

$$\alpha(x^*, x_t) = \min\Big(\frac{f_X(x^*)Q_X(x^*, x_t)}{f_X(x_t)Q_X(x_t, x^*)}, 1\Big)$$

If $Q_X(x_t, x^*) = Q_X(x^*, x_t)$, then you have the Metropolis Algorithm.

An alternative is to use the independent Metropolis-Hastings Algorithm, which draws from a fixed distribution $p_X(x)$ instead of using a random walk with $x_t + \epsilon$.

Note: This is different as x_{t+1} is independent from the previous state x_t.

11.6.2 Gibbs-Sampling

We have seen cases where the joint distribution is sampled from directly because the desired parameter of interest is not easily computable, or simply cannot be derived in closed-form. However, there may also be cases where the joint distribution itself is difficult to sample from.

In this case, we turn to **Gibbs sampling** which samples the conditional density instead of the joint density.

Suppose we have a distribution with p parameters of interest and a time index t. We sample from the conditional distribution of $f_{\mathbf{X}}(\mathbf{x})$:

$$
\begin{aligned}
x_{t+1,1} &\sim f_{\mathbf{X}}(x_1|X_2 = x_{t,2}, X_3 = x_{t,3}, ..., X_p = x_{t,p}) \\
x_{t+1,2} &\sim f_{\mathbf{X}}(x_2|X_1 = x_{t+1,1}, X_3 = x_{t,3}, ..., X_p = x_{t,p}) \\
x_{t+1,3} &\sim f_{\mathbf{X}}(x_3|X_1 = x_{t+1,1}, X_2 = x_{t+2,2}, ..., X_p = x_{t,p}) \\
&... \\
&... \\
x_{t+1,p} &\sim f_{\mathbf{X}}(x_p|X_1 = x_{t+1,1}, X_2 = x_{t+,2}, ..., X_{p-1} = x_{t+1,p-1})
\end{aligned}
$$

Take careful note - there conditional sampling here relies on X_{t+1} index up to the $p-1^{th}$ sampled variable and from the $p+1^{th]}$ variables onwards the X_t index is used. What this means is that the computation continues, the latest sampled values of the random variables are being used.

To obtain your posterior sample, simply repeat the above a desired number of times. Gibbs sampling results in nice properties that should be noted.

First is that this sample will approximate the joint distribution of all the variables, and the marginal distribution of any subset can be approximately by considering the sample of the subset (that's actually insane!). The expectation $E[X]$ of any variable can therefore be approximated by averaging over the samples.

Example:

Gibbs sampling is reasonably efficient for independent distributions, but highly dependent distributions will reduce its effectiveness.

11.7 Assessing Convergence and Accuracy

In this section we make a brief outline of points to consider when implementing these algorithms.

11.7.1 Assessing Convergence

It is important to remember that Markov Chains by definition induce autocorrelation amongst X_i. This would mean that inference from the generated sample would be less effective given the same number of independent draws. However, at convergence the draws are considered independent and identically distributed. In practice though, we have to be careful.

We discuss one way of handling this problem.

(i) Simulate multiple sequences with initial values dispersed throughout the parameter space

(ii) Determine the convergence of parameters of interest and compare variation between the sequences. Low/equal variation across the sequences compared to the variation within each individual sequence suggest reasonable approximation of the target distribution.

(iii) Alter the algorithm if the simulation efficiency is low. This can include transformation, reparameterisation or data augmentation.

Another concept is the **burn-in**, or **warm-up** period, where we discard the initial sample values because the Markov Chain has not yet stabilised. As a conservative measure, we can discard the first half, or if convergence has not been reached, sample more and discard the initial run.

Sometimes, it may be appropriate to retain every k^{th} sample, also known as *thinning*. This may be appropriate when large numbers of parameters are involved and storage concerns are prioritised.

There are also challenges associated with monitoring convergence. It is common to diagnose convergence by checking for **mixing** and **stationarity**.
For us to believe with a reasonable degree of confidence that the simulation has converged, we need the multiple chains we have simulated to have 'mixed' (i.e. map out a common distribution) AND have reached stationarity.

We discuss a fairly simple way to assess this criteria. This involves splitting all the chains, after discarding the burn-in period, into half. We then proceed to this whether the half-sequences have mixed. This simple test is a crude way of assessing both mixing and stationarity.

Let m be the number of chains and n be the total length of each chain. $m \geq 4$ as you must have at least 2 chains. Our criteria is **between** and

within sequence varianaces. Let $i = 1, 2, ..., n$ and $j = 1, 2, ..., m$ with the variance being θ_{ij}

We compute the between sequence variance as

$$\mathcal{B} = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\theta}_{.j} - \bar{\theta}_{..})^2 \tag{104}$$

where $\bar{\theta}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij}$ and $\bar{\theta}_{..} = \frac{1}{m}\sum_{i=1}^{m}\theta_{.j}$. Note that $\mathbb{B}$ is scaled by a factor of n because of the individual sequences $\bar{\theta}_{.j}$ being calculated as a mean.

The within sequence variance is easily computed as

$$\mathcal{W} = \frac{1}{m}\sum_{j=1}^{m}s_j^2 \tag{105}$$

where $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_{.j})^2$

The marginal posterior variance can then be estimated with

$$v\hat{a}r^{+}(\theta|y) = \frac{n-1}{n}\mathcal{W} + \frac{1}{n}\mathcal{B} \tag{106}$$

This an overestimation assuming that the starting distributions are appropriately dispersed, but is unbiased if the distributions are stationary. The within variance $\mathcal{W}$ should be an underestimate because the individual chains have yet to traverse across a majority of the distribution.

We can calculate the **potential scale reduction factor (PSRF)**, which estimates the potential decrease in within-chain variability compared to between-chain variability, using

$$\hat{\mathcal{R}} = \sqrt{\frac{v\hat{a}r^{+}(\theta|y)}{\mathcal{W}}} \tag{107}$$

If the PSRF is high, then there is evidence to suggest that proceeding with further simulations may improve inference of the target distribution.

11.7.2 Simulation Draws

If we were to treat our simulations as independent, then we would have $m \times n$ number of samples to work with. However this is not the case as we know that the simulations are autocorrelated, and the between-chain variance $\mathcal{B}$ will be larger compared to an independent sample.

However, there are methods to determine the *effective number of independent simulation draws* given a simulated sample size. One such method involves considering the **statistical efficiency** of the simulations $\bar{\theta}_{..}$ as an estimate of the posterior mean $E[\theta|y]$.

We introduce an equivalent statement without proof:

$$\lim_{n \to \infty} mn(var(\bar{\theta}_{..})) = \Big(1 + 2\sum_{t=1}^{\infty} \rho_t\Big) var(\theta|y) \tag{108}$$

using the RHS we can compute the effective sample size. ρ_t is the autocorrelatino of the sequence ψ at lag t. If the n simulation draws from each of the m chains were independent, then the asymptotic variance of the total sum of squares variance would be $\frac{1}{mn} var(\theta|y)$
The effective sample size is then defined as

$$n_{eff} = \frac{mn}{1 + 2\sum_{t=1}^{\infty} \rho_t} \tag{109}$$

It is not ideal to sum from i to ∞ because the correlation becomes much too noisy as i increases. Instead, we sum until autocorrelation estimates for the chain is negative for 2 successive lags $\hat{\rho}_t$ and $\hat{\rho}_{t+1}$ is negative.

So instead we have

$$n_{eff} = \frac{mn}{1 + 2\sum_{t=1}^{\mathcal{T}} \rho_t} \tag{110}$$

where $\mathcal{T}$ is the first odd positive integer where $\hat{\rho}_{\mathcal{T}}$ is negative.

In order to estimate the correlation, we'll need to compute ρ.
We use $v\hat{a}r^{+}(\theta|y)$ from before and the **variogram** V_t at index t.

$$\mathcal{V}_t = \frac{1}{m(n-t)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_{ij} - \theta_{i-t,j})^2 \tag{111}$$

and given that

$$\mathcal{V}_t = 2(1-\rho_t) v\hat{a}r^{+}(\theta|y)$$

we have

$$\rho_t = 1 - \frac{V_t}{2v\hat{a}r^{+}(\theta|y)} \tag{112}$$

All the calculations above are performed after discarding the burn-in period. This methodology rely on the mean and variance, which could pose problems if the distribution of the posterior is far from a Normal distribution. Asymptotically the posterior will approach normality, but in practice this is not the case especially since we often work with small amounts of data. If possible, we can use transformations to make the data better behaved.

In general, we would like $\hat{\mathcal{R}}$ to be close to as 1 as possible. How strict you have to be really depends on the problem at hand, but a good threshold to begin with may be 1.1?

For $\hat{n}_{eff}$, it is recommended to be at least $5m$, which is 10 independent draws per chain. Unfortunately, all of these checks are not rigorous in the sense that they are hypothesis tests. They are meant to be a guide, for practical significance.

11.8 Hamiltonian Monte Carlo