MSO-201 : Probability and Statistics

Lecture 1: Random Experiments and Probability

1. INTRO: There are lot of things around us about which we cannot make any precise prediction. For example the weather, the outcome of a coin toss, the roll of a die, etc. Such processes are often referred to as "random experiments". So outcomes of random experiments are uncertain. However in most of these situations we know what are the possible outcomes. Each such outcome is called an elementary event. The set of all elementary events is called the sample space often denoted as Ω.

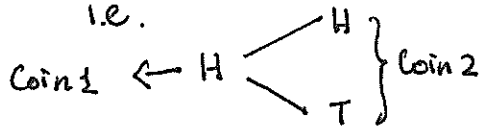
Example 1: If we toss a coin it is not sure in which way it will turn up when it hits the ground. One may say apart from head or tail one may argue that on a field for example a coin may get stuck and remain vertical. Thus in order to build a theory of probability we need make more idealistic assumption So we talk of a fair coin, which can either be head or tail, when it falls to the ground. Thus

$$\Omega = \{H, T\}, \quad : H \equiv \text{Head}, \ T \equiv \text{Tail}.$$

Example 2: If we roll a fair die, then the sample space is given as,

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

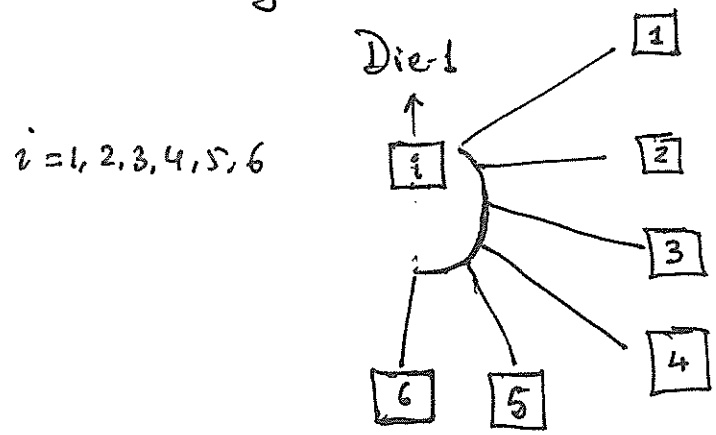
Example 3: Now let us toss two coins simultaneously. What is Ω? Note that for each outcome of the first coin, there are two outcomes of the second coin, i.e.



Thus

$$\Omega = \{HH, HT, TH, TT\}.$$

Example 4: If we role a fair ~~dy~~ die, then role another fair die, then there are 36 possible outcomes, since for each of the six outcomes of the first die, there corresponds six outcomes of the second die. This is given as



$$\Omega = \left\{ \begin{matrix} (1,1), (1,2), (1,3), (1,4), (1,5), (1,6) \\ (2,1), (2,2), (2,3), (2,4), (2,5), (2,6) \\ (3,1), (3,2), (3,3), (3,4), (3,5), (3,6) \\ (4,1), (4,2), (4,3), (4,4), (4,5), (4,6) \\ (5,1), (5,2), (5,3), (5,4), (5,5), (5,6) \\ (6,1), (6,2), (6,3), (6,4), (6,5), (6,6) \end{matrix} \right\}$$ □

All our examples above may make us feel that for any random experiments the number of all possible outcomes is finite. But these need not be the case. For example consider the number of car accidents that may happen in Kanpur in the next 6 ~~not~~ months. In fact theoretically any non-negative integer would do, i.e.

$$\Omega = \{0, 1, 2, 3, 4 \ldots\ldots\}$$

As we will see later that there can be a sample space with uncountably infinite ~~number~~ of sample points. We shall however in this chapter consider only sample spaces with finite number of sample points.

2. Events and Probability.

To be precise, for a finite sample space Ω, we can define any subset of Ω to be an <u>event.</u> Thus the power set of Ω, ie 2^{Ω} is called the set of all events associated with random experiment with sample space Ω. Remember the outcome of a ~~ra~~ trial of a random experiment is some $\omega \in \Omega$. If $\omega \in A \subset \Omega$, then we say that the event A has occured

For example let us consider the rolling of a fair die and let us say A be the event that an odd number appears. Then A is said to happen if the face of the die shows 1, 3 or 5. Thus $A = \{1, 3, 5\}$, which is a subset of $\Omega = \{1, 2, 3, 4, 5, 6\}$.

We say that two events A and B are mutually exclusive if $A \cap B = \phi$, We say events $A_1, A_2, \ldots A_K$ are mutually exclusive if $A_i \cap A_j = \phi$, $i, j = 1 \ldots K$ & $i \neq j$. ϕ or the empty-set denotes the null-event, sometimes often referred to as the impossible event.

Let us consider a random experiment which has N outcomes, ~~Gen~~ i.e. $\Omega = \{\omega_1, \omega_2, \ldots \omega_N\}$, i.e. $\text{Card}(\Omega) = N$. Note $\{\omega_1\}, \{\omega_2\}, \ldots \{\omega_N\}$ forms a set of mutually exclusive events. For example if we toss a "fair" coin, then either "head" or "tail" appears. Both cannot appear at the same time. However we also "feel" that "head" or "tail" must have an equal chance of ~~appearin~~ materializing. This is what one says as "Head & Tail are equally likely events". In fact each $\omega_i \in \Omega$, is called an elementary event. So we are now ready to give the "Classical" ~~De~~ definition of probability.

Classical ~~events~~ Definition of Probability

Let a random experiment result in N, mutually exclusive, equally likely and exhaustive, elementary events. Let A be an event such that card (A) = n_A. Then probability of the occurence of the event A, is the number P(A) given as

$$\boxed{P(A) = \frac{n_A}{N}}$$

* Exhaustive means that apart ~~from~~ these N outcomes there are no other outcomes

Certain facts are immediately clear.

i) $0 \leq P(A) \leq 1$

ii) $P(\Omega) = \frac{N}{N} = 1,$

iii) $P(\phi) = \frac{0}{N} = 0$

iv) If card (A) = n_A and card (B) = n_B, then

$$P(A \cup B) = \frac{\text{Card}(A \cup B)}{N} = \frac{n_A + n_B - n_{A \cap B}}{N}$$

$$\therefore P(A \cup B) = \frac{n_A}{N} + \frac{n_B}{N} - \frac{n_{A \cap B}}{N}$$

$$= P(A) + P(B) - P(A \cap B)$$

v) If $A \cap B = \phi \Rightarrow P(A \cup B) = P(A) + P(B)$.

vi) A^c ~~is called~~ denotes the complement of the event A; (ie. NOT A)

In fact card $(A^c) = N - n_A$. Thus

$$P(A^c) = \frac{N - n_A}{N} = 1 - \frac{n_A}{N}$$

$$\therefore \boxed{P(A^c) = 1 - \frac{n_A}{N}}$$

$$\therefore \boxed{P(A^c) = 1 - P(A)}$$

Another way to look at the last inequality is the following

$$A \cup A^c = \Omega,$$

Since $A \cap A^c = \phi, \Rightarrow P(\Omega) = P(A) + P(A^c)$

$$\Rightarrow \quad P(A^c) = 1 - P(A).$$

Drawback: One might note that the word "equally likely" must refer to the fact that each elementary event has an equal "probability" of occurence, or that they are equiprobable. Thus the definition might appear circular. In the next chapter we shall see how Kolmogorov, the great Russian mathematician provided a solid mathematical foundation to probability.

The term equiprobable means: $P(\{\omega_i\}) = \frac{1}{N}$, for $i = 1, \ldots, N$.

Thus when you toss a coin : $P(\text{Head}) = 0.5$, $P(\text{Tail}) = 0.5$

Thus when you roll a die : $P(i) = \frac{1}{6}$, $i = 1, 2, \ldots, 6$.

In fact if you repeatedly toss a coin you will see after a large number of tosses, the number of heads and tails are almost equal. Thus using these kinds of experimentation one may have a feel that the classical definition is not bad at all. We will soon see its power, but let us just begin with a story. Chevalier de Mere, a French gambler who was a pose posed this problem to his mathematician friend Blaise Pascal

Which is more likelier!

Rolling at least one six in four throws of a die

OR

Getting a double six in 24 throws of a pair of die

The gambler reasoned as follows:

He felt that the average number of succesful roll is same in both cases:

Case-I: Probability of getting six in a single throw is $\frac{1}{6}$. So the expected value in 4 throws is $4 \times \frac{1}{6} = \frac{2}{3}$

Case-II: Probability of getting a double six when a double die is rolled is $\frac{1}{36}$, so the average in 24 throws is $24 \times \frac{1}{36} = \frac{2}{3}$.

But 'Chevalier de Mere' observed that he lost frequently when he used the second gamble. This problem excited Blaise Pascal, and he wrote about this to his friend Pierre de Fermat about this issue, and between them they formed the mathematical edifice of probability theory. We shall solve 'Chevalier de Mere's problem in Lecture 3.

Let us now use the classical definition to solve an interesting problem, Let there be N people in ~~this~~ a party. What is probability that at least two of them have the same birthday. Let us ignore leap years. This problem is solved by first computing the ~~complement~~ probability of the complementary event, i.e. no two people have their birthdays on the same day of the year. Let A^c denote that event

Now $\text{card}(\Omega) = 365^N$ in this case.

$$P(A^c) = \frac{365 \times 364 \times 363 \times \cdots \times (365-(N-1))}{365^N}$$

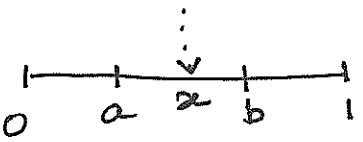
The first person can have his birthday on any of the 365 days and second one on the remaining 364 days, the third on the remaining 363 days and so on. Thus our event A has the probability

$$P(A) = 1 - \left(\frac{365 \times 364 \times 363 \times \cdots \times (365-N+1)}{365^N} \right)$$

If $N = 25$, then one can show that $P(A^c) \approx 0.40$, showing that $P(A) \approx 0.60$, thus the probability of the birthdays of two people matching ~~be~~ increases beyond 0.5 if $N \geq 25$. So if N is large such an event is equally like

3. Infinite Sample Spaces & Bertrand's Paradox

Consider the following problem. Let us be given the line $[0,1]$

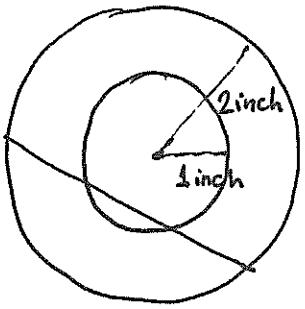


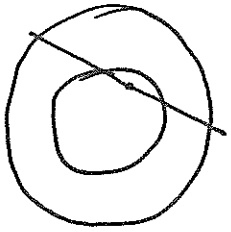
We ~~be~~ throw a stone or a pin on the line. We would like to know if our pin falls in the interval (a, b), $a < b$, & $0 < a < b < 1$, ~~So our sample space~~ and what is the probability that ~~he~~ it will ~~do so be so~~ fall there. It is intuitively clear that $\Omega = [0,1]$. One way of calculating the probability of the event $A = \{\omega : a < \omega < b\}$, is to consider ~~as~~ the lengths as a kind of "cardinality" of these intervals. Thus we may write

$$P(A) = \frac{\text{length}(a,b)}{\text{length}[0,1]} = b - a.$$

However observe that if this is the way we compute the probability in the above case then if $\tilde{A} = \{\omega : \omega = a\}$, then $P(\tilde{A}) = 0$. Thus it is impossible to through a pin at an exact point on $[0,1]$. However such approaches have its own perils while we are handling the inf case of infenite sample spaces. This is aptly demonstrated by the Bertrand's paradox, where in the same event is shown to have to different probabilities.

Bertrand's Paradox: Consider a ~~circle~~ circle of radius 2 inch and a concentric circle of radius 1 inch. A chord is drawn at random in the circle of radius 2 inch. What is the probability of the chord cutting through the circle of radius one inch?

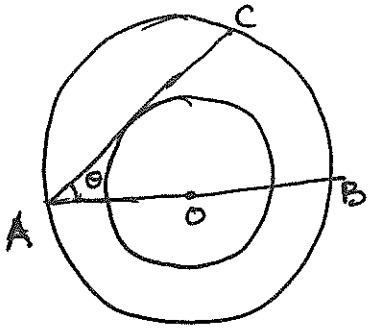




Solution: Approach 1: The chord of course is unique if it is not a diameter, in the sense that it can be identified through its mid-point. The chord to cut through the smaller concentric circle the mid point of the chord must lie in the circular disc formed by the smaller circle. Let us denote by A as the event that the chord cuts through the smaller circle. Here the sample space is the circular disc of radius 2 inch. We shall get the probability of A as a ratio of areas

i.e $$P(A) = \frac{\text{Area of smaller circle}}{\text{Area of bigger circle}} = \frac{\pi}{\pi 4.} = \frac{1}{4}$$

Solution: Approach 2



Fix a diameter ~~and draw~~ AB of the bigger circle and observe that any chord AC starting at A will vary from 0 to π while θ varies from 0 to $\pm\frac{\pi}{6}$ to cut through the smaller circle. So, we have

$$P(A) = \frac{\frac{2\pi}{6}}{\pi} = \frac{1}{3}.$$

Note that for any chord AC we actually draw a horizontal diameter AB and argue as above.

You can see the contradiction!!

Lecture 2: Kolmogorov's Axioms for Probability

We have seen that though for finite sample spaces the idea of defining probability through counting works pretty well, we run into rough weather the moment we try to extend such an idea to an infinite sample space setting.

It was the great mathematicians, Emile Borel, Henri Lebesgue, Felix Hausdorff and Cantelli who realized that at a much deeper level the notion of probability is intimately linked with notion of measure of a set and thus notion of probability needed a fresh mathematization, which finally was completed by Kolmogorov. Borel and his co-workers gained a tremendous insight into the nature of probability by considering questions about infinite coin tosses and thus each instance can be represented as a point in the interval $[0,1]$. This finally led to the strong law of large numbers which we will mention later in the course. Roughly what happens is the following. When we make an arbitrarily large number of trials the proportion of ~~heads~~ is heads is nearly $1/2$ and remains there forever with probability one.

Borel and his co-workers understood the fact that every subset of an infinite sample space may not be an event! It was believed that events must have the following property

i) Ω, & ϕ must be events (sure event & null event must be there)

ii) If A is an event then so is A^c.

iii) If $A_1, A_2 \ldots A_n \ldots$ be a countable sequence of events then so is $\bigcup_{i=1}^{\infty} A_i$.

The class of such subsets of Ω is often denoted by $\mathcal{F}$ and $\mathcal{F}$ is often referred to as the σ-algebra or σ-fields of events.

One might be worried that we spoke of a countably infinite sequence of events but did not say anything about finite number events. But it is simple to show that if $A_1 \dots A_k$ are a finite number of events then $\bigcap_{i=1}^{k} A_i$ is also an event, i.e. a member of $\mathcal{F}$. Set $A_{k+1} = \phi$, $A_{k+2} = \phi$ .....

Since by i) ϕ is an event, we have

$$\bigcup_{i=1}^{k} A_i = A_1 \cup A_2 \cup \dots \cup A_k \cup \phi \cup \phi \cup \phi \cup \dots \cup \phi \dots$$

is also an event.

Further if $A_1 \dots A_n \dots$ is a countable sequence of events then $\bigcap_{i=1}^{\infty} A_i$ is also an event (Try this out as an assignment).

Kolmogorov assumed or rather viewed that given the pair $(\Omega, \mathcal{F})$, we can define a function $P : \mathcal{F} \to \mathbb{R}$ (A set function actually) which satisfies the following rules

i) For any $A \in \mathcal{F}$, $0 \le P(A) \le 1$ (This is clear from defn of P)

ii) $P(\Omega) = 1$

iii) If $\{A_n\}_{n=1}^{\infty}$ is a sequence of mutually exclusive events (i.e. $A_i \cap A_j = \phi$, if $i \neq j$), then

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

~~The third rule or axi~~ The above rules are often called Kolmogorov's axioms of probability. The third rule is a central one and provides us with some important facts about the function P which is called the probability function or probability measure

Our first result which is a consequence of the axiom is as follows,

$$\boxed{P(\phi) = 0}$$

One might wonder why this was not mentioned in the axiom. In fact $P(\phi) = 0$ is a consequence of the third axiom, since if we set $A_i = \phi$, $\forall\, i \in \mathbb{N}$, (Here $\mathbb{N}$ denotes the set of natural numbers), then $\{A_n\}_{n=1}^{\infty}$ is a mutually disjoint sequence of sets, and iii) gives us

$$P(\phi) = \sum_{n=1}^{\infty} P(\phi)$$

$$P(\phi) = P(\phi) + P(\phi) + \cdots + P(\phi) + \cdots$$

This equation is only possible if $P(\phi) = 0$ which is what we already know from the classical approach, but deriving this fact using Kolmogorov's axioms does not put any restriction on the sample space Ω.

Our next result is the following, which is again a consequence of the third rule and the fact that $P(\phi) = 0$. Let $A_1, A_2 \ldots A_k$ be k, mutually disjoint events, then

$$P\left(\bigcup_{n=1}^{k} A_n\right) = \sum_{n=1}^{k} P(A_k).$$

The strategy is the same; set $A_{k+1} = \phi$, $A_{k+2} = \phi \ldots$

Thus

$$\bigcup_{n=1}^{k} A_n = A_1 \cup A_2 \cup A_3 \ldots \cup A_k \cup \phi \cup \phi \ldots \cup \phi$$

Thus using (iii) we have

$$P\left(\bigcup_{n=1}^{k} A_n\right) = \sum_{n=1}^{k} P(A_n) + \sum_{n=k+1}^{\infty} P(\phi)$$

$$= \sum_{n=1}^{k} P(A_n) \quad (\because P(\phi) = 0).$$

Thus we have that if $A_1 \cdots A_k$ are mutually disjoint events, then

$$\boxed{P\left(\bigcup_{n=1}^{k} A_n\right) = \sum_{n=1}^{k} P(A_n).} \longrightarrow (\#)$$

Now as $A \cup A^c = \Omega$, using (#) we have

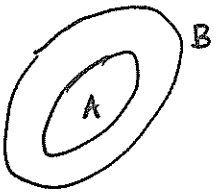
$$P(\Omega) = P(A \cup A^c)$$
$$= P(A) + P(A^c) \text{ by } (\#)$$
$$\therefore P(A) + P(A^c) = 1 \quad (\text{since } P(\Omega) = 1, \text{ from second axiom})$$

Hence irrespective of the nature of Ω we have

$$\boxed{P(A^c) = 1 - P(A)}$$

We will now derive some more results about the probability function based on the Kolmogorov's axioms.

- Let $A \subseteq B$, then $P(A) \leq P(B)$.



In this case we have

$$B = A \cup (B \setminus A) \quad [\text{Show that } B \setminus A \in \mathcal{F}]$$
$$\therefore P(B) = P(A) + P(B \setminus A), \quad (\text{Using } (\#))$$
$$\therefore P(B) - P(A) = P(B \setminus A)$$

But by i) $P(B \setminus A) \geq 0 \Rightarrow P(B) \geq P(A)$.

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, for any $A, B \in \mathcal{F}$.

Now $A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B)$

We observe that above three sets on the left are mutually exclusive.

Hence using (*) we have

$$P(A \cup B) = P(A \setminus B) + P(B \setminus A) + P(A \cap B) \longrightarrow (*)$$

Again

$$A = (A \setminus B) \cup (A \cap B).$$

But as $(A \setminus B)$ and $(A \cap B)$ are disjoint, i.e. mutually exclusive we have using (*)

$$P(A) = P(A \setminus B) + P(A \cap B)$$

$$\therefore \quad P(A \setminus B) = P(A) - P(A \cap B)$$

Further

$$B = (B \setminus A) \cup (A \cap B)$$

and by similar arguments as before

$$P(B) = P(B \setminus A) + P(A \cap B)$$

$$\therefore \quad P(B \setminus A) = P(B) - P(A \cap B).$$

Putting these expressions back in (*) we have

$$P(A \cup B) = P(A) + P(B) - 2P(A \cap B) + P(A \cap B)$$

$$\therefore \boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)} \longrightarrow (**)$$

We can immediately conclude that

$$\boxed{P(A \cup B) \leq P(A) + P(B)}$$ since $P(A \cap B) \geq 0$

We can extend this using mathematical induction to the following result. If $A_1, \ldots, A_k$ are elements of $\mathcal{F}$ then

$$\boxed{P\left(\bigcup_{n=1}^{k} A_n\right) \leq \sum_{n=1}^{k} P(A_n)}$$

This is called Boole's Inequality.

The equality (4.4) can be easily generalized for more than two events and is given as:-

$$P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P(A_k) - \sum_{k_1<k_2}^{n} P\left(A_{k_1} \cap A_{k_2}\right) + \sum_{k_1<k_2<k_3}^{n} P\left(A_{k_1} \cap A_{k_2} \cap A_{k_3}\right) - \sum_{k_1<k_2<k_3<k_4}^{n} P\left(A_{k_1} \cap A_{k_2} \cap A_{k_3} \cap A_{k_4}\right) + \cdots + (-1)^{n+1} P\left(\bigcap_{k=1}^{n} A_n\right)$$

This is sometimes referred to as the "Principal of Inclusion and Exclusion".

Since the Kolmogorov's axiom involves, infinite sequences of events let us look into some properties of such sequences.

- Consider the set of non-decreasing events, ie $A_n \subset A_{n+1}$. The set $A = \bigcup_{n=1}^{\infty} A_n$ is then called the limit of $\{A_n\}$
- Similarly consider the sequence of non-increasing events, i.e. $A_{n+1} \subset A_n$. Then $A = \bigcap_{n=1}^{\infty} A_n$ is call the limit of $\{A_n\}$

Let us look at the following result, which is important enough to be mentioned as a theorem

Theorem 2.1: Consider a sequence $\{A_n\}$ of non-decreasing event. Then

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

Proof: Let us set $A = \bigcup_{n=1}^{\infty} A_n$. We can write A as follows

$$A = A_n + \bigcup_{j=n}^{\infty} (A_{j+1} \setminus A_j)$$

Again using the third axiom of Kolmogorov, (Think why??) we have

$$P(A) = P(A_n) + \sum_{j=n}^{\infty} P(A_{j+1} \setminus A_j)$$

Note that

$$0 \leq \sum_{j=1}^{\infty} P(A_{j+1} \setminus A_j) \leq 1$$

Thus $\sum_{j=1}^{\infty} P(A_{j+1} \setminus A_j)$ is a convergent series of non-negative numbers. ~~that~~ Thus

$$\lim_{n\to\infty} \sum_{j=n}^{\infty} P(A_{j+1} \setminus A_j) = 0 \quad \# \text{ (see details)}$$

Hence

$$P(A) = \lim_{n\to\infty} P(A_n) + \lim_{n\to\infty} \sum_{j=n}^{\infty} P(A_{j+1} \setminus A_j)$$

$$\therefore \boxed{\lim_{n\to\infty} P(A_n) = P(A)}$$

— x —

\# Let $S = \sum_{i=1}^{\infty} a_i$, $S_n = \sum_{i=1}^{n} a_i$, Then for any $\varepsilon > 0$, we have

~~$n_0 \in \mathbb{N}$~~, such that for all $n \geq n_0$

$$|S_n - S| < \varepsilon$$

$$\Rightarrow |a_{n+1} + a_{n+2} + \dots| < \varepsilon$$

$$\Rightarrow \left| \sum_{i=n+1}^{\infty} a_i \right| < \varepsilon$$

$$\Rightarrow \lim_{n\to\infty} \sum_{i=n+1}^{\infty} a_i = 0.$$

Lecture 3: Conditional Probability, Independence & Bayes Theorem

Sec 1: Conditional Probability

In this lecture we shall introduce the reader to a very important notion of probability theory called "Conditional Probability". Mathematically stated it means the following: Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and events $A, B \in \mathcal{F}$, we ask the question as to what is the ~~probability~~ probability of the event A given the fact or conditioned on the fact that the event B has already occured"?

We ask such questions pretty often asked in real-life scenarios. For example given that a train has departed from the originating station with a ten minute delay, and we may be interested to know the probability that it will reach its' destination on time.

The probability of an event A, given that the event B has occured is called the conditional probability of A given that B, and is denoted by $P(A|B)$.

Definition 3.1 : Conditional Probability : Given $(\Omega, \mathcal{F}, \mathbb{P})$ to be a probability space, and $\mathbb{P}(B) > 0$; where $B \in \mathcal{F}$. Then the conditional probability $P(A|B)$, for any $A \in \mathcal{F}$ is given as

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

- Let us view this definition from the classical angle. Let Ω be a sample space, with N possible out comes. The event B has occured me means ~~n_B number~~ that some $\omega \in B$ has occured. Now what is the probability that A will occur. So now our total number of possibilities have shrunk to only those which represent the event B. So among those ~~events~~ possibilities we need to what is proportion that is common to ~~A~~ the event A, i.e. the sample points which represent the occurence $A \cap B$

$$\therefore P(A|B) = \frac{n_{A \cap B}}{n_B} = \frac{\frac{n_{A \cap B}}{N}}{\frac{n_B}{N}} = \frac{P(A \cap B)}{P(B)}, \quad \text{where } P(B) > 0.$$

Example 3.1 Let us consider an urn having 5 red & 5 black balls. Find the probability that a red ball appears on the 2nd draw given that the first draw resulted in a black ball. The first ball is not replaced back in the urn.

Let B be the event that the first ball drawn was black. Let A be the event that second ball drawn is red.

$$\therefore \quad P(A|B) = \frac{5}{9}$$

How did we come to this value. We have already drawn a black ball. So before making the second draw we have only 9 balls left so that is exactly our total possible outcomes and we have 5 balls are red, which gives the above value

Observe that from the definition of ~~$P(B|A)$~~ $P(A|B)$ we have

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

So in Example 3.1 we have

$$P(A \cap B) = P(A|B)\,P(B)$$

Here $P(B) = \frac{5}{10} = \frac{1}{2}$ $\quad \therefore \boxed{P(A \cap B) = \frac{5}{9} \times \frac{1}{2} = \frac{5}{18}}$

Of course $A \cap B$ is the event that a black ball appears in the first draw and a red in the second

Given any $B \in \mathcal{F}$ with $P(B) > 0$, we define the conditional probability function given B as $P(\cdot|B) : \mathcal{F} \longrightarrow [0,1]$, i.e.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \forall A \in \mathcal{F}.$$

Now let us see if the axioms of Kolmogorov are satisfied. For any $A \in \mathcal{F}$, it is obvious that $P(A|B) \geq 0$. Further as $A \cap B \subset B$, $\Rightarrow P(A \cap B) \leq P(B)$. Thus $P(A|B) \leq 1$.

Now if $A = \Omega$ we have

$$P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$$

Let $\{A_n\}$ be a sequence of mutually ~~exclusive~~ disjoint events, ~~i.e~~ then

$$P\left(\bigcup_{n=1}^{\infty} A_n \mid B\right) = \frac{P\left(\left(\bigcup_{n=1}^{\infty} A_n\right) \cap B\right)}{P(B)}$$

$$= \frac{P\left(\bigcup_{n=1}^{\infty} (A_n \cap B)\right)}{P(B)}$$

As $\uplus \{A_n \cap B\}$, forms a ~~dis~~ mutually disjoint sequence of events. ~~such that~~ Hence

$$P\left(\bigcup_{n=1}^{\infty} (A_n \cap B)\right) = \sum_{n=1}^{\infty} P(A_n \cap B)$$

$$\therefore P\left(\bigcup_{n=1}^{\infty} A_n \mid B\right) = \sum_{n=1}^{\infty} \frac{P(A_n \cap B)}{P(B)}$$

$$= \sum_{n=1}^{\infty} P(A_n \mid B).$$

Thus all the three laws of Kolmogorov holds. Thus we have shown that for any given event B, with $P(B)>0$, $P(\cdot \mid B)$ is a probability measure or probability function.

We shall now state an important result called the "Theorem of Total Probabilities" which will lead to do computations with the famous Bayes Theorem, which we will discuss in the next section.

Theorem of Total Probability

Theorem 3.1.1 : Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\{B_1, \ldots B_n\}$ be a collection of n mutually disjoint collection of events such that $\Omega = \bigcup_{i=1}^{n} B_i$. Further assume that for each $i \in \{1, 2, \ldots n\}$, $P(B_i)>0$. Then for any $A \in \mathcal{F}$

$$\boxed{P(A) = \sum_{i=1}^{n} P(B_i) P(A \mid B_i)}$$

Proof: Note that since $\{B_1 \ldots B_n\}$ forms a mutually disjoint partition of Ω, then A either occurs joints with B_1 or B_2 or .... or B_n, i.e.

$$A = \bigcup_{i=1}^{n} (A \cap B_i)$$

while we know that $(A \cap B_1) \ldots (A \cap B_n)$ are mutually disjoint. Then

$$P(A) = P\left(\bigcup_{i=1}^{n} (A \cap B_i)\right)$$

$$= \sum_{i=1}^{n} P(A \cap B_i) \quad \left[\because A \cap B_i,\ i = 1 \ldots n \text{ are mutually exclusive}\right]$$

→ By using exclusion-inclusion principle.

$$\therefore \quad \boxed{P(A) = \sum_{i=1}^{n} P(B_i)\, P(A|B_i)}$$

This theorem remains true if $n = \infty$ Think how.

Section: Bayes Theorem

Thomas Bayes was a priest, an English priest. The theorem that now bears his name, was published after his notes were edited by Robert Price. Thomas Bayes had not published his result. The idea of Bayes later led to the field of Bayesian statistics which has huge applications in modern medicine and machine learning, for example.

Bayes Theorem Let us consider the probability space $(\Omega, \mathcal{F}, P)$. Let $\{B_1 \ldots B_n\}$ be a finite sequence of events such that $P(B_i) > 0$, $\forall\, i = 1 \ldots n$ & $\Omega = \bigcup_{i=1}^{n} B_i$. Let $A \in \mathcal{F}$ such that $P(A) > 0$, then

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

$$\boxed{P(B_i|A) = \frac{P(B_i)\, P(A|B_i)}{\sum_{i=1}^{n} P(B_i)\, P(A|B_i)}}$$

This theorem also holds for $n = \infty$. Now we shall provide the simple proof.

$$P(B_i \mid A) = \frac{P(B_i \cap A)}{P(A)}$$

$$P(B_i \cap A) = P(A)\, P(B_i)\, P(A \mid B_i)$$

while by the theorem of total probability

$$P(A) = \sum_{i=1}^{n} P(B_i) P(A \mid B_i).$$

A more simple version of the Bayes Theorem can be provided with two events $A \& B \in \mathcal{F}$ with $P(A) > 0$, $P(B) > 0$.

Note that if $B \in \mathcal{F} \Rightarrow B^c \in \mathcal{F}$ and we know that

$$\Omega = B \cup B^c$$

$$\therefore \quad P(A) = P(A \cap B) + P(A \cap B^c)$$

$$= P(B) P(A \mid B) + P(B^c) P(A \mid B^c)$$

$$\therefore \quad \boxed{P(B \mid A) = \frac{P(B) P(B \mid A) P(A \mid B)}{P(B) P(A \mid B) + P(B^c) P(A \mid B^c)}}$$

What the Bayes theorem does is as follows. It can help us answering the following questions. For example in an urn there n_W white balls and n_B black balls. Suppose balls are drawn at random without replacement. Suppose we draw the balls twice and suppose that the second ball is white: Then we can ask the question as to what is the probability that the first ball drawn was white or the first ball drawn is black. Bayes theorem helps us in answering such questions.

If B be the event that the first ball drawn is black then B^c is naturally the event that the first ball drawn is white. Let W be the event that the second ball drawn is white. So

$$P(B \mid W) = \frac{P(B) P(W \mid B)}{P(W)} \quad \text{(By Bayes Theorem)}.$$

Now $P(B) = \frac{n_B}{n_B + n_W}$

$P(W|B) = \frac{n_W}{n_B + n_W - 1}$ $\quad \frac{n_W}{n_B + n_W - 1}$

While

$P(W) = P(W \cap B) + P(W \cap B^c)$

$= P(B)\,P(W|B) + P(B^c)\,P(W|B^c)$

$= \frac{n_B}{(n_B + n_W)} \cdot \frac{n_W}{}$

$= \frac{n_B}{n_B + n_W} \cdot \frac{n_W}{n_B + n_W - 1} + \frac{n_W}{n_B + n_W} \cdot \frac{n_W - 1}{n_B + n_W - 1}$

$= \frac{n_B n_W + n_W (n_W - 1)}{(n_B + n_W)(n_B + n_W - 1)}$

$= \frac{n_B n_W + n_W^2 - n_W}{(n_B + n_W)(n_B + n_W - 1)}$

$\therefore P(B|W) = \frac{\frac{n_B}{n_B + n_W} \cdot \frac{n_W}{n_B + n_W - 1}}{\frac{n_B n_W + n_W^2 - n_W}{(n_B + n_W)(n_B + n_W - 1)}}$

$\therefore P(B|W) = \frac{n_B n_W}{n_B n_W + n_W^2 - n_W}$

Since $n_W \in \mathbb{N}$, i.e. $n_W \geq 1 \Rightarrow n_W^2 \geq n_W \Rightarrow n_W^2 - n_W \geq 0$

$\Rightarrow n_B n_W + n_W^2 - n_W \geq n_B n_W$

$\Rightarrow P(B|W) \leq 1$. and thus fulfilling the basic requirement of the probability function.

$$\boxed{P(B|W) = \frac{n_B n_W}{n_B n_W + n_W^2 - n_W}}$$

• we urge the reader to compute $P(B^c|A) = ??$

Sec 3

Independence of Events

I do not even remember which year but I remember reading a book called "The Enigmas of Chance", by the famous mathematician Mark Kác. where he discussed an important idea called "Independence of Events" i.e. two events A and B, are called independent if and only if

$$P(A \mid B) = P(A), \quad \text{if } P(B) > 0$$

$$P(B \mid A) = P(B), \quad \text{if } P(A) > 0$$

i.e. $\boxed{P(A \cap B) = P(A) \cdot P(B)}$ ⟶ (the key fact)

This idea as I read in the book was due to Hugo Steinhaus, who developed this notion, when he was in hiding during the second world war. S. Steinhaus was a famous Polish mathematician, who made major contributions to an area of mathematics.

Let us see how this can be applied. Consider the simple situation where you toss a coin twice. What is the probability that head (H) appears in the second toss for the first time

Let A be the event that head appears in the second toss for the first time; and B be the event that tail appears Thus we are trying to find the probability the configuration HT We know that $P(HT) = \frac{1}{4}$ from direct computation. But

$$HT = B \cap A$$

$$\therefore P(HT) = P(B \cap A) = \frac{1}{4}$$

$$\therefore P(B \cap A) = \frac{1}{2} \cdot \frac{1}{2} = P(B) \cdot P(A)$$

Thus B & A are independent events and which appears to be intuitively true.

Let us now introduce the notion of independence of three events $A, B, C \in \mathcal{F}$. The definition of independence now means the following: Three events A, B & C are independent events if and only if

$$P(A \cap B) = P(A) \cdot P(B)$$
$$P(B \cap C) = P(B) \cdot P(C)$$
$$P(C \cap A) = P(C) \cdot P(A)$$
$$P(A \cap B \cap C) = P(A) P(B) P(C).$$

It is important that pairwise independence of events do not mean they are independent. The following example from the book [1] by Moods, Graybill and Boes, ~~shows~~ establishes what we have just mentioned above.

Example: Pairwise independence does not mean independence.

Let

A_1: Event that odd face appears in the first die

A_2: Event that odd face appears in the second die.

A_3: The sum of the two numbers in a random throw of two die is odd.

$$P(A_1) = \frac{3}{6} = \frac{1}{2} \quad \& \quad P(A_2) = \frac{3}{6} = \frac{1}{2}$$

$$P(A_1 \cap A_2) = \frac{9}{36} = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A_1)\,P(A_2).$$

We urge the reader to show that

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

$$P(A_3) = \frac{18}{36} = \frac{1}{2} \quad \left(A_3 = \{(1,2),(2,1),(1,4),(4,1),(1,6)(6,1),(2,3),(3,2) (5,6),(6,5),(3,6),(6,3),(4,3),(3,4) (5,4),(4,5),(2,5),(5,2)\} \right)$$

$$P(A_1 \cap A_3) = \frac{9}{36} = \frac{1}{4} \quad \text{(Finding the cases in } A_3 \text{ with odd first term)}$$
$$= \frac{1}{2} \cdot \frac{1}{2} = P(A_1) \cdot P(A_3)$$

We leave $P(A_2 \cap A_3) = P(A_2)\,P(A_3)$ verification to the reader. It is symmetric.

$$P(A_1 \cap A_2 \cap A_3) = 0.$$

But $P(A_1) \cap P(A_2) \cap P(A_3) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$

$$\therefore \boxed{P(A_1 \cap A_2 \cap A_3) \neq P(A_1)\,P(A_2)\,P(A_3)}$$

—x—

At the end of this section let us solve the "Chevalier de Mere" problem and also revisit the birthday problem.

For the first case in the "Chevalier de Mere" problem one is looking at the appearance of atleast one six in the throw of four dice or four throws of a die.

Let A_i be the event that six does not appear on the i-th throw then $P(A_i) = \frac{5}{6}$, $i = 1, \ldots, 4$

Thus if A is the event the six does not appear in any of the four throws, then $A = \bigcap_{i=1}^{4} A_i$

$$\therefore P(A) = P\left(\bigcap_{i=1}^{4} A_i\right) = P(A_1)\,P(A_2)\,P(A_3)\,P(A_4)$$

[as A_1, A_2, A_3, A_4 are independent events]

$$= \left(\frac{5}{6}\right)^4$$

~~∴ If B is the probability~~

Hence A^c is the event at least one six appears we have

$$P(A^c) = 1 - \left(\frac{5}{6}\right)^4$$

His second case was a double six appearing on 24 throws of a pair of die.

Let B_i, be the event that a double six does not occur in the i-th throw, then $P(B_i) = \left(\frac{35}{36}\right)$, $i = 1, \ldots, 24$.

Let B be the event that a double six never occurs, then

$$B = \bigcap_{i=1}^{24} B_i$$

Hence the independence of B_i's showed that

$$P(B) = \left(\frac{35}{36}\right)^4$$

Our required event is B^c, and hence

$$P(B^c) = 1 - \left(\frac{35}{36}\right)^4.$$

which is bigger $P(B^c)$ or $P(A^c)$? They don't appear to be equal in the face of it.

From the internet for example from mathword.wolfram.com. we gather that

$$\mathbb{P}(A^c) = 1 - \left(\frac{5}{6}\right)^4 \approx 0.5177$$

$$\mathbb{P}(B^c) = 1 - \left(\frac{35}{36}\right)^4 \approx 0.4914$$

and thus $P(A^c) > P(B^c)$. and that is why Chevalier de Mere was mostly losing when we chose the second bet.

Let us revisit the "Birthday problem" in Terms of conditional probability. Instead of asking the question, as what is the probability of at least two people having the same birthday we ask the question, what in a group of people what is the probability that none of the members have the same birthday. We provide the following approach from [S1].

Suppose we choose two people from the group. What is the probability that they have different birthdays? Call that event B_2, then

$$\mathbb{P}(B_2) = 1 - \frac{1}{365.}$$ [Not considered leap year.

What happens if we have three people. What is $\mathbb{P}(B_3)$? In fact we can write

$$B_3 = (A_3 \cap B_2)$$

if A_3 is the event that the third person has a birthdate which does not match with any one of the other two.

$$P(B_3) = P(A_3 \cap B_2) = P(A_3 \mid B_2) P(B_2)$$

Now $\mathbb{P}(A_3 \mid B_2) = \frac{363}{365} = 1 - \frac{2}{365}$

$$\therefore \mathbb{P}(B_3) = \left(1 - \frac{2}{365}\right)\left(1 - \frac{1}{365}\right) \approx 0.9918$$

Thus in general we can write when we are looking at a group of n-people we can write

$$B_n = A_n \cap B_{n-1}$$

$$\therefore \mathbb{P}(B_n) = \mathbb{P}(A_n \mid B_{n-1})\, \mathbb{P}(B_{n-1})$$

$$= \left(1 - \frac{(n-1)}{365}\right) \mathbb{P}(B_{n-1})$$

$$= \left(1 - \frac{n-1}{365}\right)\left(1 - \frac{n-2}{365}\right) \mathbb{P}(B_{n-2})$$

$$\vdots \qquad \vdots$$

$$= \left(1 - \frac{n-1}{365}\right) \cdots \left(1 - \frac{2}{365}\right) \mathbb{P}(B_2)$$

$$= \left(1 - \frac{n-1}{365}\right) \cdots \left(1 - \frac{2}{365}\right)\left(1 - \frac{1}{365}\right)$$

It was shown for example in [S1], that $\mathbb{P}(B_{22}) \approx 0.5243$, while $\mathbb{P}(B_{23}) \approx 0.4927$. So if we have 23 or more people in the group their probability of having different birthdays goes down below 0.5 and has less chance of occurence

[S1]: F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, & L. E. Meester, A Modern Introduction to Probability and Statistics: Understanding why and how. Springer 2010

Lecture 4: Random Variables, Their Distribution and their Expectation

Sec 1: RANDOM VARIABLES

We are all accustomed to handling real numbers. Thus it might be a good idea to shift the arena for computing probability from the zone of abstract events to the real lines. This is done through the vehicle of random variables. But before we embark into the definition of a random variable let us get a fact straight: Random Variable is a function.

> A random variable $X : \Omega \longrightarrow \mathbb{R}$, a function from the sample space Ω, to the real line $\mathbb{R}$, such that for any real number $\alpha \in \mathbb{R}$, such that the set
>
> $$S_\alpha = \{\omega \in \Omega : X(\omega) \leq \alpha\} \subset \mathcal{F}$$
>
> i.e. the set S_α is an event.

Consider the random experiment of tossing a coin, i.e. $\Omega = \{H, T\}$
Define random variable

$$X(H) = 1, \quad X(T) = 0.$$

Suppose we consider a σ-algebra made out of open intervals in $\mathbb{R}$, i.e. taking their unions, complements, etc. We call such a σ-algebra as the Borel σ-algebra. In fact for a random variable the set

$$S_I = \{\omega : X(\omega) \in I\},$$

where I is an open interval is also an event. Define for each I

$$P_X(I) = P\{\omega : X(\omega) \in I\},$$

$\mathfrak{F}$ We leave it to the reader should be able to show that $P_X(I)$ indeed satisfies the Kolmogorov Axioms. Thus if $\mathcal{B}$ is the notation for Borel σ-algebra, then $(\mathbb{R}, \mathcal{B}, P_X)$ is a probability space induced by the random variable X.

Suppose the range of X is finite, i.e. Range $X = \{x_1, \ldots, x_n\}$

Then consider the sets

$$\Omega_i = \{\omega : X(\omega) = x_i\}, \quad i = 1 \ldots n$$

Then $\Omega_i \cap \Omega_j = \phi, \ \forall\ i \neq j$ & $\bigcup_{i=1}^{n} \Omega_i = \Omega$, i.e. a random variable X with a finite range induce a partition of the sample space Ω. We urgue the reader to extend this idea to a random variable (r.v. for short) with a countable range.

Any random variable X whose range is either finite or countable is called discrete random variable. When studying discrete random variable we often ask the question, what is the probability that X takes the value say x_k or just say x.

We denote this by $P(X = x)$, which means

$$P(X = x) = P\{\omega \in \Omega : X(\omega) = x\}$$

Now suppose the range of the r.v. is countable, i.e

$$\text{range } X = \{x_1 \ldots x_n \ldots\} \subseteq \mathbb{R}$$

Any ~~x which~~ $x \in \mathbb{R}$, which is not in the range has zero probability. Thus it is meaningful to ask the question

What is $P(X = x_k) = P\{(\{\omega \in \Omega : X(\omega) = x_k\})$

$$= P(\Omega_k)$$

As $\Omega = \bigcup_{k=1}^{\infty} \Omega_k \Rightarrow P(\Omega) = \sum_{k=1}^{\infty} P(\Omega_k)$

$$\therefore \boxed{\sum_{k=1}^{\infty} P(X = x_k) = 1}$$

i.e. the total probability is one.

Now for any $x \in \mathbb{R}$, even for a discrete random variable X, it is meaningful to ask the question, what is the probability that X takes values less than x. Thus

$$P(X \leq x) = \sum_{x_k \leq x} P(X = x_k)$$

So more formally we write

$$\boxed{f_X(x) = P(X = x)}$$ (p.m.f for short)

which is called the probability mass function of a discrete random variable, and

$$\boxed{F_X(x) = P(X \leq x) = \sum_{x_k \leq x} f_X(x_k)}$$

is called the cumulative distribution function (cdf), or distribution function of the random variable.

Consider the following simple example of a throwing a fair die. Define a random variable $X : \Omega \to \mathbb{R}$ as follows;

$$X(\omega) = \begin{cases} 1, & \text{if } \omega \text{ is odd} \\ 0, & \text{if } \omega \text{ is even} \end{cases}$$

The range of X is given as $\text{range } X = \{0, 1\}$. The probability mass function thus given as

$$f_X(x) = \begin{cases} P(X = 1) = 1/2 \\ P(X = 0) = 1/2 \end{cases}$$

Note: $P(X=1) = P\left(\{\omega : X(\omega) = 1\}\right) = P\left(\{1,2,3\}\right) = 1/2$, Same approach for calculating $P(x=0)$.

The distribution function is given as

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/2 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Observe the way we do the segmenting of the variable, x. The custom

is to have strict inequality on the right-hand side which allows us to graphically represented as a step function, the cdf of a discrete r.v. X. In our example of the die, the graph of ~~F(x) is given~~ F_X is given as

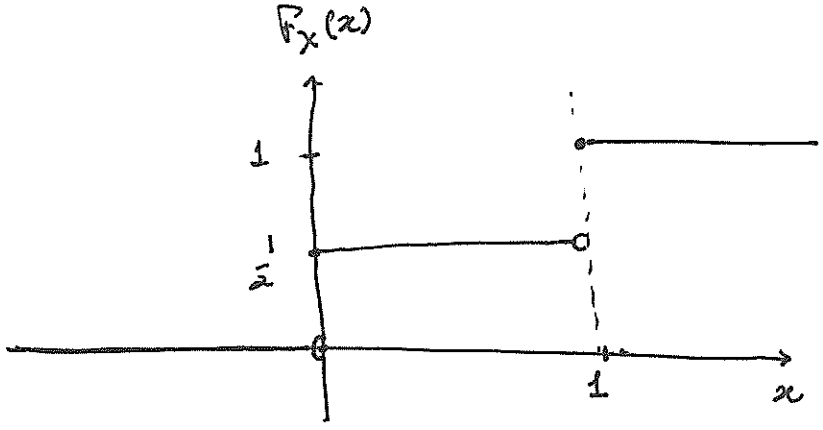


Fig 4.1 c.d.f of X in the previous example

The c.d.f. or distribution function has the following properties.

i) F_X is a non-negative function, and naturally $F_X(x) \leq 1$

ii) If $x < y$, we have $F_X(x) \leq F_X(y)$.

This is what we get just by observing the ~~figures~~ above figure of the cdf. It is not a difficult task to prove the above two from the very definition of a c.d.f.

Another important result shows us how to compute a p.m.f. value if the c.d.f. of a discrete r.v. is given.

Let us arrange the values of a r.v. X with finite range in the the ascending order

$$x_1 < x_2 < x_3 < \ldots\ldots < x_n$$

and $f(x_1) = F(x_1)$, and

$$f(x_i) = F_X(x_i) - F_X(x_{i-1}), \quad i = 2 \ldots n$$

This can be proved from the definition of F_X as

$$F_X(x_i) = F_X(x_{i-1}) + f(x_i),$$

Further observe that F_X is right continuous, i.e.

$$\lim_{x \to a+} F(x) = F(a)$$

It is however not left continuous. Note that we are writing all these properties by looking at the graph of the c.d.f in Fig 4.1. Also observe that

$$\lim_{x \to -\infty} F_X(x) = 0$$

while

$$\lim_{x \to +\infty} F_X(x) = 0$$

You might wonder why the above two properties uses limits. It is done so to consider the ~~countably infin~~ random variables which has a countably infinite range. Let us now summarize the properties of the distribution function F_X of a discrete random variable X, that we have learnt.

i) $F_X \geq 0$ & $F_X \leq 1$

ii) $\lim_{x \to a+} F_X(x) = F_X(a)$

iii) If $x < y \Rightarrow F_X(x) \leq F_X(y)$

iv) $\lim_{x \to -\infty} F_X(x) = 0$ & $\lim_{x \to +\infty} F_X(x) = 1$

Every random variable need not have a finite or countable range. A random variable may have a range which is uncountably ~~finite~~ infinite. A r.v. X is called continuous if its range is uncountably infinite

Suppose we want to measure height of all students here at IIT Kanpur. Then we cannot say that there are only a finite set of ~~point~~ height measures possible. In fact we can reasonably say that height of students are in the interval $[4,7]$ i.e. between 4ft and 7ft. So the height of a student is a random variable

$$\Omega \equiv \text{Students} \xrightarrow{\text{Height} = X} \mathbb{R}$$

So if X is the random variable denoting the height of a student, then

$$4 \leq X(\omega) \leq 7$$

and thus ~~X~~ X has an ~~co~~ uncountably infinite range. Now which question is more meaningful,

$P(X(\omega) = 5\text{ft})$ or $P(4.8 \leq X(\omega) \leq 5.2)$.

So lets first consider the issue of fact $P(4.8 \leq X(\omega) \leq 5.2)$, Since whenever we make a measurement, random errors and observational errors creep in and it is impossible to say that an interval is of exactly ~~this~~ a particular length. In fact if we measure the same line segment ~~for~~ five times, each time we are likely to get very slightly different results. Keeping our intuition of the classical approach we may write

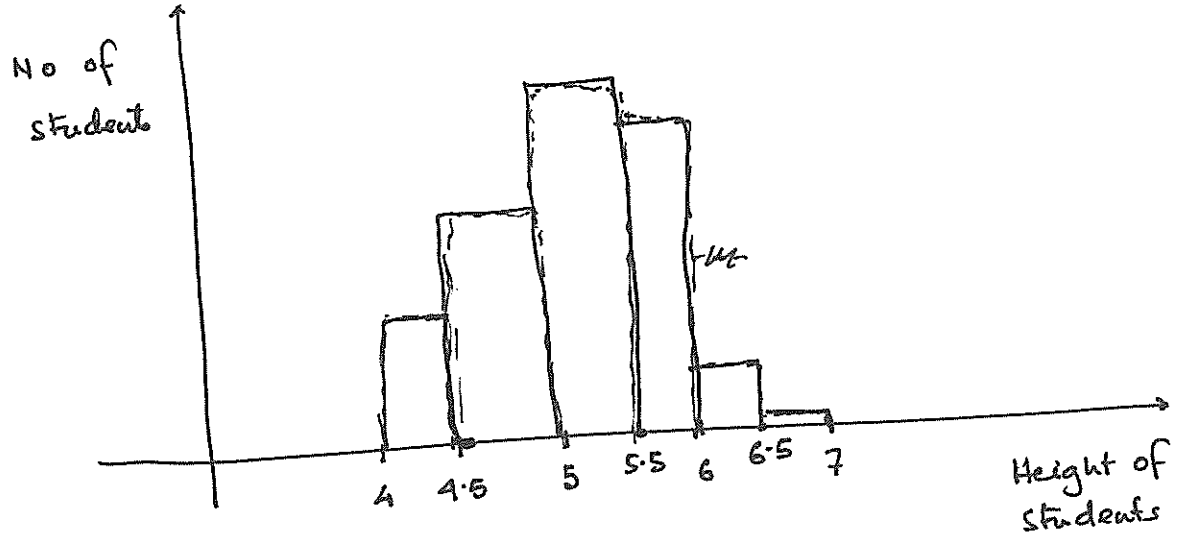
$$P(4.8 \leq x \leq 5.8) = \frac{\text{length of } [4.8, 5.8]}{\text{length of } [4, 7]} = \frac{1}{3}.$$

However $P(\cancel{4.8} x = 5) = \dfrac{\text{length of } \{5\}}{\text{length of } [4,7]} = 0$

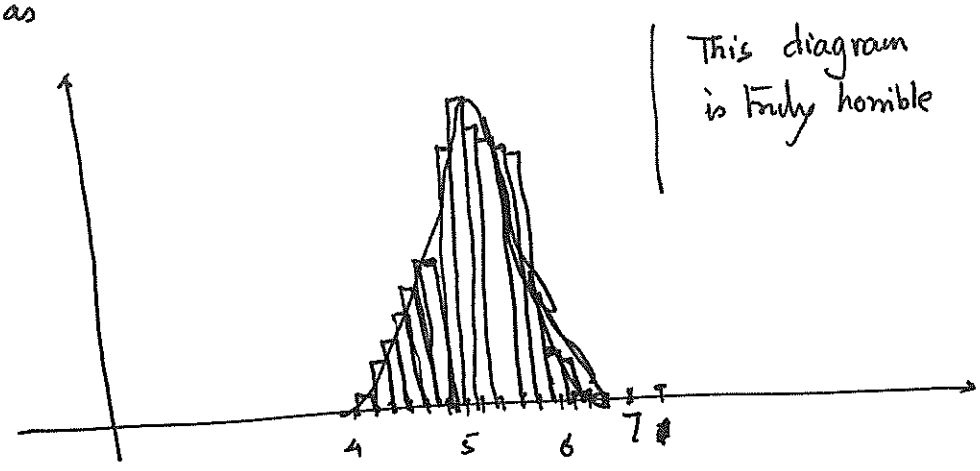
* We are aware of the perils of such an approach Recall Bertrand's paradox.

But is that the only way we deal with the case when range of X is uncountably infinite. There could be many situations where our intuition ~~to~~ from classical probability may fail. A better path is ~~to~~ through pictorical representation of data.

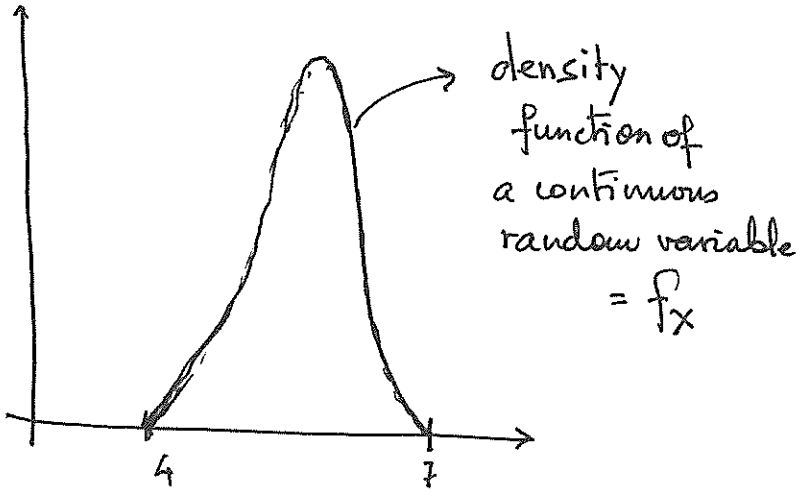
Consider ~~a.g.~~ again the heights of students at IIT Kanpur. Let us divide the height range $[4, 7]$ into intervals of length 0.5 along the horizontal axis, and in the vertical axis the number of ~~peo~~ students with height in ~~that interval~~ those intervals. The resulting diagram is called a histogram which we depict below



If we make more finer partitions of the interval, then we have a diagram as



So as we make the partitioning intervals smaller and smaller we can represent the histogram by a smooth curve which we will call the density function, since each rectangle in some sense represents the density of the people with heights in that interval.



Note that the values of the density function of a continuous random variable denoted as f_X does not represent the probabilites, as at individual points the probability is zero while as we see the density function is not.

From a formal perspective, the probability density function of a continuous random variable X, is a function f_X which has the following properties

i) $f_X(x) \geq 0, \quad \forall x$

ii) $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

iii) $P(a \leq x \leq b) = \int_a^b f_X(x)\,dx.$

Another approach to view a continuous random variable X is to assume that it has a distribution function F_X which is continuous. Indeed we have

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(x)\,dx.$$

using (iii). Remember that all other properties of the distribution function remains same, as we have seen for the discrete variable.

Thus for a continuous random variable, X, we have

$$P(a \le X \le b) = \int_a^b f_X(x)\,dx$$

$$= \int_{-\infty}^{b} f_X(x)\,dx - \int_{-\infty}^{a} f_X(x)\,dx$$

$$= F_X(b) - F_X(a).$$

Thus

$$\boxed{P(a \le X \le b) = F_X(b) - F_X(a)}$$

One can also show that $\boxed{\frac{dF_X}{dx} = f_X.}$

One might have a sense that somehow, all these we have discussed for the continuous case is largely intuitive. However there is a result which we will not prove, but will every thing clear.

We have already discussed that given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and X be a random variable associated with it. Then using X we can shift our working space to $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra in $\mathbb{R}$. The following Theorem shows how a unique probability measure $\hat{\mathbb{P}}$ can be defined for the space $(\mathbb{R}, \mathcal{B})$. We present the result as given in Robert Ash*

Theorem 4.1 : Let f_X be a non-negative function on $\mathbb{R}$, with $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$. Then there exists a unique probability measure $\hat{\mathbb{P}}$ such that for any Borel subset of $B \in \mathcal{B}$ we have

$$\hat{\mathbb{P}}(B) = \int_B f_X(x)\,dx$$

is a probability measure on $\mathbb{R}$.

The reader can show that $\hat{\mathbb{P}}(B)$ satisfies the Kolmogorov Axioms.

* Robert B. Ash : Basic Probability Theory, Dover 2008.

As an example consider, the following c.d.f or distribution function.

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

It is a continuous function. It is ~~increasing~~ non-decreasing also and $\lim_{x \to -\infty} F_X(x) = 0$ & $\lim_{x \to +\infty} F_X(x) = 1$. Hence F_X represents a continuous distribution function. Hence X is a continuous ~~fund~~ random variable.

The probability density function associated with X is

$$f_X(x) = \begin{cases} e^{-x}, & \text{when } x \geq 0 \\ 0 & \text{when } < 0 \end{cases}$$

(We just differentiated F_X, ie we set $f_X(x) = \frac{dF_X}{dx}$).

Let us now check that f_X is truly a ~~f~~ density function. Of course $f_X \geq 0$, $\forall x \in \mathbb{R}$ &

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \int_0^{\infty} e^{-x}\,dx$$

$$= \left[-e^{-x} \right]_0^{\infty} = 1.$$

There are ~~several~~ other random variables, which are of the mixed type, i.e. it has both discrete part and continuous.

For our purposes, we shall stick to discrete and continuous variables and in the next chapter we shall look at some special (means largely applicable) distribution of discrete random variables, followed by a chapter on distributions for continuous random variables. Among the continuous variable distributions the normal distribution will play a crucial role in statistics.

[* Mathematicians by the way term the normal distribution as the Gaussian distribution, after the Gauss, the originator of the idea. More on that later.]

Section 2: Expectation of a Random Variable

If we take a simplistic approach, then expectation means average. We are also aware of the idea of weighted average.

Let us consider n observations (could be height or weight... blah blah)

$$x_1, x_2, \ldots, x_n.$$

To each of these observation assign weights $w_1, w_2 \ldots w_n$ respectively & $w_i \geq 0$, for all $i = 1 \ldots n$. Then weighted mean means is defined as

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}$$

If we set $\bar{w}_i = \dfrac{w_i}{\sum_{i=1}^{n} w_i}$, where $\sum_i w_i = w_1 + \cdots + w_n$,

then

$$\bar{x} = \bar{w}_1 x_1 + \bar{w}_2 x_2 + \cdots + \bar{w}_n x_n$$

Now for each i, $0 \leq \bar{w}_i \leq 1$ & $\sum_{i=1}^{n} \bar{w}_i = 1$, and thus the weights $\bar{w}_i$ can be viewed as probability ~~associated with occurrence of the i-th obsrv~~ of the i-th observation turning out to be x_i. Thus ~~for~~ taking a clue from the weighted mean the expectation of a discrete random variable, X is given as

$$E(X) = \sum_{i=1}^{\infty} x_i P(X = x_i) = \sum_{i=1}^{\infty} x_i f_X(x_i)$$ [provided the series is convergent]

The definition of expectation for a continuous random variable is quite analogous and is given as

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx,$$

provided the integral is finite.

Given a random variable $X : \Omega \longrightarrow \mathbb{R}$, & g: Range $X \longrightarrow \mathbb{R}$, the the function $Y = g \circ X$, is a function from $\Omega \longrightarrow \mathbb{R}$. Is this also a random variable? (The value of Y is computed as $Y(\omega) = g(X(\omega))$

Now consider the set $\{\omega : Y(\omega) \in I\}$, for any interval I, we need to show that $\{\omega : Y(\omega) \in I\} \in \mathcal{F}$, if $(\Omega, \mathcal{F}, \mathbb{P})$ is the probability space we are working on.

Observe that $\{\omega : Y(\omega) \in I\} = Y^{-1}(I)$, i.e. to prove that Y is a random variable we have to show that $Y^{-1}(I) \in \mathcal{F}$ for any interval I in $\mathbb{R}$, Observe that

$$Y^{-1}(I) = (g \circ X)^{-1}(I)$$

$$= X^{-1} \circ g^{-1}(I)$$

Now $g^{-1}(I) \subset$ Range X. ~~Thus for~~ For ~~any~~ Any subset of Range X is a Borel set and hence $X^{-1}\left(g^{-1}(I)\right) \in \mathcal{F}$ as X is a random variable.

So for $Y = g(X)$, we have, when X is discrete

$$E(Y) = E[g(X)] = \sum_{i=1}^{\infty} g(x_i) f_X(x_i)$$ (provided the series converges)

and when X is continuous

$$E(Y) = E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx,$$ (provided the integral is finite).

E.g. If $g(X) = X + c$, where X is constant then

$$E[X + c] = \int_{-\infty}^{\infty} (x + c) f_X(x)\, dx = \int_{-\infty}^{\infty} x f_X(x) dx + c \int_{-\infty}^{\infty} f_X(x) dx$$

$$= E(X) + c$$

$\because \int_{-\infty}^{\infty} f_X(x) dx.$

You can check out some properties in the assignments. Consider again the random variable X with the distribution

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

Then $E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$

$$= \int_0^{\infty} x e^{-x}\, dx$$

$$= -x e^{-x}\Big|_0^{\infty} - \int_0^{\infty} \frac{dx}{dx}(-e^{-x})\, dx$$

$$= 0 + \int_0^{\infty} e^{-x}\, dx = 1$$

i.e. $E(X) = 1$.

However for every distribution one need not have a finite expectation. Consider the discrete distribution, where the random variable has countable values, ie X has values $\{0 \& 1, 2, 3, 4 \ldots\}$.

Let us have

$$f_X(x) = \frac{1}{2^x}$$

Consider $g(X) = 2^X$. Then

$$E(g(X)) = \sum_{x=0}^{\infty} 2^x \frac{1}{2^x} = 1 + 1 + \ldots + 1 + \ldots$$

Thus the series $1 + 1 + \ldots 1 + \ldots$ is not convergent and hence $E(g(X))$ does not exist. This is often called the St. Petersburg Paradox.

In statistics, $E(X)$ is traditionally given the symbol μ. Another important measure in statistics is that of variance. Variance is a measure of dispersion. It measures the average-square deviation of the random variable values from the mean. Thus for a random variable X

$$\sigma_X^2 = \text{Variance of } X = \text{Var}(X) = E(X-\mu)^2$$

Thus
$$\begin{aligned} E(X-\mu)^2 &= E(X^2 - 2X\mu + \mu^2) \\ &= E(X^2 - 2\mu X) + \mu^2 \\ &= E(X^2) - 2\mu E(X) + \mu^2 \quad \text{(check this)} \\ &= E(X^2) - 2E(X)^2 + E(X)^2 \\ &= E(X^2) - E(X)^2 \end{aligned}$$

$$\therefore \boxed{\text{Variance of } X = \text{Var}\, X = E(X^2) - E(X)^2}$$

Let us finish this chapter by computing the variance of the random variable X, whose pdf is

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

Thus
$$\begin{aligned} \text{Var}(X) &= E(X^2) - E(X)^2 \\ &= E(X^2) - 1 \\ &= \int_0^\infty x^2 e^{-x}\, dx - 1 \\ &= \left[-x^2 e^{-x}\right]_0^\infty - 2\int_0^\infty x\,(-e^{-x})\, dx - 1 \\ &= 0 + 2E(X) - 1 \\ &= 2 - 1 = 1. \end{aligned}$$

Section 3: Chebyshev's Theorem

We begin with a general form of Chebyshev's Theorem:

Theorem 4.2 Let g be a non-negative function of a random variable X, defined on a probability space $(\Omega, \mathcal{F}, P)$. Then for any $k > 0$ we have

$$\boxed{P\left(g(X) \geq k\right) \leq \frac{E\left[g(X)\right]}{k}}$$

Proof: Observe that if X is a continuous random variable,

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

$$= \int_{\{x:\, g(x) \geq k\}} g(x) f_X(x)\, dx + \int_{\{x:\, g(x) < k\}} g(x) f_X(x)\, dx$$

$$\geq \int_{\{x:\, g(x) \geq k\}} g(x) f_X(x)\, dx \quad \left[\text{Since } \int_{\{x:\, g(x) < k\}} g(x) f_X(x)\, dx \geq 0\right]$$

$$\geq k \int_{\{x:\, g(x) \geq k\}} f_X(x)\, dx = k\, P\left[g(X) \geq k\right]$$

$$\therefore \quad P\left[g(X) \geq k\right] \leq \frac{E\left[g(X)\right]}{k}. \qquad \square$$

Further if $g(X) = (X - \mu)^2$ we have $P\left[(X-\mu)^2 \geq k\right] \leq \frac{Var(X)}{k}$

Chebyshev's inequality says that

$$P\left(|x-\mu| \geq r\sigma_x\right) \leq \frac{1}{r^2}, \quad \text{where } \sigma_x^2 = \text{Var}(x) \neq 0$$

In fact Observe that

$$P\left(|x-\mu| \geq r\sigma_x\right) = P\left((x-\mu)^2 \geq r^2\sigma_x^2\right)$$

$$\leq \frac{\text{Var}(x)}{r^2\sigma_x^2} \quad \text{(By Theorem 4.2)}$$

$$\Rightarrow \quad P\left(|x-\mu| \geq r\sigma_x\right) \leq \frac{\sigma_x^2}{r^2\sigma_x^2}$$

$$\Rightarrow \quad \boxed{P\left(|x-\mu| \geq r\sigma_x\right) \leq \frac{1}{r^2}}$$

This is Chebyshev's inequality. What it says is the following

$$-P\left(|x-\mu| \geq r\sigma_x\right) \geq -\frac{1}{r^2}$$

$$\Rightarrow \quad 1 - P\left(|x-\mu| \geq r\sigma_x\right) \geq 1 - \frac{1}{r^2}$$

i.e $$\boxed{P\left(|x-\mu| < r\sigma_x\right) \geq 1 - \frac{1}{r^2}} \longrightarrow \odot$$

In fact this says that

$$\boxed{P\left(\mu - r\sigma_x < x < \mu + r\sigma_x\right) \geq 1 - \frac{1}{r^2}}$$

This says that the probability, that the value of the random variable X falls in the open interval $(\mu - r\sigma_x, \mu + r\sigma_x)$ increases as r increases. We will make use of this idea later in our study.

Section 4: Moments and Moment Generating function

The concept of moments play an important role in the practical applications. The r-th moment is simply the expectation of of the ~~random variable~~ function $g(X) = X^r$ of the random variable X, i.e.

$$\mu'_r = E(X^r)$$

If $r=1$, then $\mu'_1 = E(X) = \mu_X$, the mean or expectation.

There is also a notion of central moment, ie moment centered at ~~a~~ the value μ_X, i.e

$$\mu_r = E[(X-\mu_X)^r]$$

In this case $\mu_1 = 0$, while $\mu_2 = \sigma_X^2$.

But how do we compute moments. This is done through the device of moment generating function, which we now discuss.

The moment generating function of a r.v. X, or mgf for short is given as

$$m_X(t) = E[e^{tX}]$$

For the continuous case we have

$$m_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tX} f_X(x)\,dx$$

of course for $m_X(t)$ to be finite this integral must exist for ~~every~~ the given t. In fact it is much better that $m_X(t)$ is finite at least over an open interval, ~~for~~ i.e. $\exists\, h>0$, s.t $\forall\, t \in (-h,h)$, the function $m_X(t)$ is finite.

For the discrete case of course we have

$$m_X(t) = E(e^{tX}) = \sum_x e^{tx} f_X(x).$$

Now if g_1 & g_2 are functions of the random variable X, then one can easily show that

$$E[g_1(x) + g_2(x)] = E[g_1(x)] + E[g_2(x)]$$

Can this be generalized to the countable case. Let

$$Y = \sum_{i=1}^{\infty} g_i(x)$$

and assume that $Y(w)$ is finite for all w and $E(Y)$ exists, and if $E[g_i(x)]$ exist for each i, then

$$E[Y] = \sum_{i=1}^{\infty} E(g_i(x)) \quad \left[\begin{array}{c}\text{of course under}\\ \text{some conditions}\end{array}\right]$$

provided $\sum_{n=1}^{\infty} E[g_i(x)]$ is also convergent.

$$E[e^{tX}] = E\left[1 + \frac{tx}{1!} + \frac{t^2x^2}{2!} + \frac{t^3x^3}{3!} + \cdots + \frac{t^n x^n}{n!} + \cdots\right]$$

$$= 1 + \mu_1' t + \mu_2' \frac{t^2}{2!} + \cdots + \mu_n' \frac{t^n}{n!} + \cdots.$$

In fact the coefficient of t^n, is μ_n', for $n = 1, 2, \ldots$.

If the expectation $E[e^{tX}]$ is finite then one can show that the above series also converges. Suppose we can integrate under the differential sign. Suppose we can differentiate under the integral sign. Suppose we differentiate the mgf, r times, then we get

$$\frac{d^r}{dt^r} m(t) = \int_{-\infty}^{\infty} x^r e^{tx} f_X(x)\, dx$$

$$\Rightarrow \boxed{\frac{d^r}{dt^r} m(0) = \int_{-\infty}^{\infty} x^r f_X(x)\, dx = \mu_r'}$$

Lecture -5

Chapter 5: Binomial & Poission Distributions

In this chapter we seek to study some specialized probability distributions, which has a lot of applicability. In this chapter in particular we shall focus on Binomial and Poission distributions which are important distributions for discrete random variables. For a discrete random variable X, its probability distribution is simply an algebraic expression of its probability mass function. Of course special distributions are for particular type of random variables, describing special situations.

Section 1: Binomial Distributions

Suppose we make repeated coin tossing. Of course a fair coin is used. Suppose we toss a coin 100 times and ask the question that what is the probability that head appears 25 times? The Binomial Distribution answers such questions.

Binomial Distribution deals with repeated trials and each trial has only two possible outcomes, which mark as success or failure. Each such trial is often called a Bernoulli trial. For example when we toss a coin for example we can consider the appearance of head (H) as success and appearance of tail as failure. If we denote success as S and failure as F, then any repeated random trial with n trials may look like the following cases

SSFF...FSSS.....S

So there could be k successes out of n trials. We ask ourselves what is the probability of such an event if each success has a fixed probability p. Note when n trials are carried out k successes can appear in $\binom{n}{k}$ ways.

Thus if X denotes the random variable denoting the number of success, then

$$\boxed{P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}}$$

~~In fact S denot.~~

So the Binomial p.m.f is usually written as

$$\boxed{f_X(x) = \binom{n}{x} p^x q^{n-x}, \quad \text{where } q = 1-p}$$

Let us first check that f_X is indeed a probability mass function.

i) $f_X(x) \geq 0$, is of course clear as $p \geq 0$

ii) $$\sum_{x=1}^{n} f_X(x) = \sum_{x=1}^{n} \binom{n}{x} p^x q^{n-x}$$

$$= (p+q)^n = 1. \quad \left(\begin{array}{l}\text{Last step used the Binomial Theorem}\end{array}\right)$$

So why it is called a Binomial distribution is clear. Before providing any example of a numerical problem let us compute two important measures associated with a random variable, namely expectation and variance. Thus if X is a binomial random variable, then we will compute $E(X)$ & $\text{Var}(X)$. There are two approaches to it. We shall choose the approach using the moment generating function. Now the mgf of the binomial random variable is

$$m_X(t) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x}$$

$$= (pe^t + q)^n.$$

Thus

$$\boxed{m_X(t) = (pe^t + q)^n}$$

$$\therefore \quad m'_X(t) = n(pe^t+q)^{n-1} pe^t$$

$$\therefore \quad m'_X(0) = n(p+q)^{n-1} p$$

$$m'_X(0) = np \quad (\text{as } p+q=1)$$

Hence $\mu_X = E(X) = m'_X(0) = np.$

Now to compute the variance we compute $m''_X(0)$. So

$$m''_X(t) = n(n-1)(pe^t+q)^{n-2} pe^t pe^t + n(pe^t+q)^{n-1} pe^t$$

$$\therefore \quad m''_X(0) = n(n-1)(p+q)^{n-2} \cancel{(pe^0)} p^2 + n(p+q)^{n-1} p.$$

$$\therefore \quad m''_X(0) = n(n-1)p^2 + np$$

$$= (n^2 - n)p^2 + np$$

$$= n^2p^2 - np^2 + np$$

$$= n^2p^2 + np - np^2$$

$$= n^2p^2 + np(1-p)$$

$$= n^2p^2 + npq.$$

$$\therefore E(X^2) = n^2p^2 + npq$$

$$\therefore \text{Var}(X) = E(X^2) - [E(X)]^2 = n^2p^2 + npq - n^2p^2$$

$$\therefore \boxed{\text{Var}(X) = npq}$$

The whole idea of using m.g.f might appear counter-intuitive as if something is forced on to us, without motivation and does not seem to have immediate ~~motivation~~ connect with the definition. Of course we can definitely compute $E[X]$ and $Var(X)$ of a binomial random variable, using brute force, which we demonstrate below.

Brute force computation :

$$\mu_X = E(X) = \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x \frac{n!}{x!\,(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} \frac{n!}{(x-1)!\,(n-x)!} p^x q^{n-x}$$

$$\therefore \quad \mu_X = np \sum_{x=0}^{n} \frac{(n-1)!}{(x-1)!\,(n-x)!} p^{x-1} q^{n-x}$$

$$= np \sum_{x=0}^{n} \binom{n-1}{x-1} p^{x-1} q^{n-x}$$

$$= np\,(p+q)^{n-1}$$

$$= np.$$

However computing $E(X^2)$ using brute force technique is not so simple. There is a small trick. This is as follows.

$$E(X^2) = E(X^2 - X + X)$$

$$= E(X^2 - X) + E(X)$$

$$= E(X(X-1)) + E(X).$$

$$E\left(X(X-1)\right) = \sum_{x=0}^{n} x(x-1)\binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1)\frac{n!}{x!\,(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1)\frac{n!}{x(x-1)(x-2)!\,(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} \frac{n!}{(x-2)!\,(n-x)!} p^x q^{n-x}$$

$$= n(n-1)p^2 \sum_{x=0}^{n} \frac{(n-2)!}{(x-2)!\,(n-2)!} p^{x-2} q^{n-x}$$

$$= n(n-1)p^2 (p+q)^{n-2}$$

$$= n(n-1)p^2$$

$$\therefore \quad E(x^2) = n(n-1)p^2 + np .$$

Hence Var(X) is given as

$$Var(x) = E(x^2) - [E(x)]^2$$

$$= n(n-1)p^2 + np - n^2p^2$$

$$= n^2p^2 - np^2 + np - n^2p^2$$

$$= np(1-p) = npq.$$

The square root of the Vart variance of X is called the standard deviation of X. $\sigma_x = \sqrt{Var(x)}$. Let us now look into a numerical example, where we can see the application of binomial distribution.

Example 5.1: A jumbo-jet, say Boeing-747-400 has four engines, the operate independently. Let each engine has a probability p q of failure and for a succesful flight at least two engines should work. What is the probability that we with the flight will be successful.

Solution: The probability of engine not failing is p

Let X be the number of engines working during the flight

We have to find the probability that $X \geq 2$, ie

$$P(X \geq 2) = 1 - P(X < 2)$$

Of course X follows a binomial distribution with $n = 4$ & probability of succes p. Thus

$$P(X \geq 2) = 1 - (P(X=0) + P(X=1))$$

$$= 1 - \left(\binom{4}{0} p^0 q^{4-0} + \binom{4}{1} p^1 q^{4-1} \right)$$

$$= 1 - \left[q^4 + 4pq^3 \right]$$

$$\therefore \boxed{P(X \leq 2) = 1 - q^4 - 4pq^3} \quad \square$$

Observe that if X is a binomial random variable, then its distribution depends on two parameters, the number of trials n and the probability of succes p. So often one write

$$\boxed{X \sim B(n, p)},$$

which means that X is a binomial random variable with parameters n & p.

Section 2: Poisson Distribution

Suppose you are editing an essay with 5,000 words and you want to know ~~if there~~ what is the probability, that there are committing five spelling mistakes, given that the probability of ~~finding~~ the mistake is 0.005. If we consider that ~~find~~ occurence of a spelling mistake is a succes, then modelling through the Binomial distribution shows that

$$f_X(5) = \binom{5000}{5} (0.005)^5 (0.995)^{4995}$$

Of course this is a cumbersome computation. The question is can we make it simpler. We are in a situation where n is large and p small while the product np is moderate. Our question is such cases, what form would the Binomial pmf take. Trying to figure out this leads us to the Poisson distribution.

Let us assume the $np = \lambda > 0$, then setting $p = \frac{\lambda}{n}$, we can rewrite the Binomial distribution as

$$f_X(x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{n!}{x!\,(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}$$

$$= \frac{n(n-1)(n-2)\cdots(n-x+1)(n-x)!}{x!\,(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}$$

Thus

$$f_X(x) = \frac{1\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{(x-1)}{n}\right)}{x!} \lambda^x \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}$$

Now we need to see what happens to the right hand side when $n \to \infty$

We of course know the celebrated formula

$$\lim_{n\to\infty} \left(1 + \frac{z}{n}\right)^n = e^z$$

Thus

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$$

Further $\left(1-\frac{\lambda}{n}\right)^{-x} \to 1$ as $n \to \infty$. Hence the rhs becomes (as $n\to\infty$)

$$\boxed{\frac{\lambda^x e^{-\lambda}}{x!}}$$

This gives us a new kind of probability distribution, which we call a Poisson distribution with parameter λ.

$$\boxed{f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}}, \quad x = 0, 1, 2, 3, \ldots$$

Note that though x can take any non-negative value, in practice it is small, but even if it is large we can always use Stirlings approximation to $x!$. Let us now compute the mean and variance of the Poission ~~rand~~ random variable. In fact we write $X \sim$ Poisson(λ), i.e X follows a Poisson distribution with parameter λ. Let us use the m.g.f technique and leave the brute force technique to the reader

$$m_X(t) = \sum_{x=0}^{\infty} \frac{e^{tx}\lambda^x e^{-\lambda}}{x!}$$

$$= e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^t)^x}{x!}$$

$$= e^{-\lambda}e^{\lambda e^t}$$

Now

$$m'_X(t) = e^{-\lambda}\lambda e^{\lambda e^t}e^t = \lambda e^{-\lambda}e^t e^{\lambda e^t}$$

$$\boxed{m'_X(0) = \mu_X = \lambda}$$

$$m''_x(t) = \lambda e^{-\lambda}\left[e^t e^{\lambda e^t} + \lambda e^t e^{\lambda e^t} e^t\right]$$

$$m''_x(0) = \lambda e^{-\lambda}\left[e^{\lambda} + \lambda e^{\lambda}\right]$$

$$\boxed{m''_x(0) = \lambda + \lambda^2}$$

Thus

$$\begin{aligned} \text{Var}(x) &= E(x^2) - (\mu_x)^2 \\ &= m''_x(0) - (\mu_x)^2 \\ &= \lambda + \lambda^2 - \lambda^2 \\ &= \lambda \end{aligned}$$

$$\boxed{\text{Var}(x) = \lambda}$$

I find this amazing, both mean and variance have the same value, λ the parameter of the Poisson distribution.

In his small book called "Facts from Figures", M.J. Moroney almost magically brings up the Poisson Distribution, for a discrete random variable X, with countable range, i.e.

$$\text{Range } X = \{0, 1, 2, 3, \ldots\ldots\}$$
$$= \mathbb{N} \cup \{0\}.$$

Observe that,

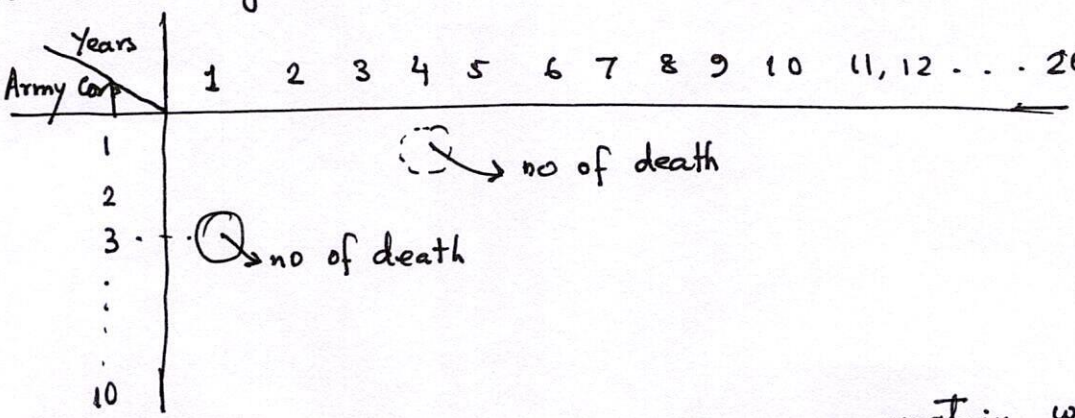
~~$+ = e^{-z} e^{z} = e^{-z}\left(1 + \frac{z}{1!} + \frac{z^2}{2!} + \right.$~~

$$1 = e^{-\lambda} e^{\lambda} = e^{-\lambda}\left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \ldots\ldots + \frac{\lambda^n}{n!} + \ldots\ldots\right)$$

$$1 = e^{-\lambda} + \frac{\lambda e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} + \ldots\ldots \frac{\lambda^n e^{-\lambda}}{n!} + \ldots\ldots$$

Observe that each term on the rhs is a term of the Poisson

distribution for different values of x. In fact to show that how Poission distribution works effectively Moroney tested it on a data collected by the great statistician R.A. Fisher. This a data on the death of a cavalryman getting killed by a horsekick in the course of a year. The data has been collected over twenty years from ten army corps (of Britain I guess) thus has 200 readings



So the data is represent is a 10 × 20 matrix, which has 200 entries. It can be summarized as follows

No of deaths	Frequency of occurence of such deaths [No cells in the above matrix with the given no]
0	109
1	65
2	22
3	3
4	1
5	0
6	0

Total death = $(0\times109)+(1\times65)+(2\times22)+(3\times3)+(4\times1)+(5\times0)+(6\times0)$

$=122$ (in twenty years) among 200 observations

So average death per year per corps $= \frac{122}{200} = 0.61$

So $\lambda = 0.61$ in this case. Once we fix it we have $e^{-\lambda} \approx 0.543$. Let us assume that number of deaths is a random variable following Poission with the assumption that $\lambda = 0.61$ is fixed.

No of deaths	Poisson Prob.	Poisson freq = 200 × Poisson Prob	Actual
0	0.543	109	109
1	0.331	66.3	65
2	0.101	20.2	22
3	0.021	4.1	3
4	0.003	0.6	1

[Table is taken from page 98 of "Facts and Figures" by M.J. ~~Morner~~ Moroney, Pelican, 1951]

So you see this data can be indeed ~~very~~ modelled very well by Poisson distributions.

"Thus statistics works"

—x—

Lecture 6 : Distributions of Continuous Random Variables

In this chapter we will be more focused with collecting and collating information, rather than delving into conceptual issues. Let us note that we shall list in this chapter some important distributions associated with continuous random variables. We are not going to make detailed computation of mean and variances or mgf of each and every distribution, that we list here. But we will of course do so for all some of the most well known ones, like the exponential distribution, Gamma distribution and above all the jewel of statistics, the normal distribution.

Lets start with a simple one first, the uniform distribution. This is a distribution whose density function remains constant over an interval

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

It is clear that f_X is a density function. Observe that $f_X(x) \ge 0 \ \forall x$ and

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \int_a^b \frac{1}{b-a}\,dx = 1.$$

Further $E(X) = \int_a^b x \frac{1}{b-a}\,dx = \frac{1}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2}\right) = \frac{b+a}{2}$

while $\text{Var}(X) = \frac{(b-a)^2}{12}$ (Compute it yourself)

The mgf for the uniform distribution is given as

$$m_X(t) = E\left[e^{tx}\right] = \frac{e^{bt} - e^{at}}{(b-a)\,t}$$

Before we discuss the normal distribution, let us ~~also~~ mention few important distributions which are used in statistics.

Section 1: The Gamma Distribution and Exponential ~~distri~~ distribution

The Gamma distribution depends on the Gamma function

The Gamma Function is given as

$$\Gamma(z) = \int_0^\infty e^{-x}\, x^{z-1}\, dx, \quad z>0$$

If $z \in \mathbb{N}$ then

$$\Gamma(z) = (z-1)!$$

Further if $z = 1/2$ then it is well-known that $\Gamma(1/2) = \sqrt{\pi}$

The Gamma function, is an improper integral. For more details see any good book on real analysis or advanced calculus. In fact

$$\boxed{\Gamma(z+1) = z\Gamma(z)}$$

Just for fun let us see why $\Gamma(1/2) = \sqrt{\pi}$. Note that

$$\Gamma(1/2) = \int_0^\infty \frac{e^{-x}}{\sqrt{x}}\, dx \quad \text{Set } x = t^2, \text{ i.e. } dx = 2t\,dt$$

$$\therefore \Gamma(1/2) = \int_0^\infty \frac{e^{-t^2}}{t}\, 2t\,dt = 2\int_0^\infty e^{-t^2}\, dt$$

There is a famous formula in calculus, i.e

$$\int_0^\infty e^{-t^2}\, dt = \frac{\sqrt{\pi}}{2}.$$

Thus $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

A random variable X is said to follow the Gamma distribution if it has the density

$$\boxed{f_X(x) = \frac{\lambda}{\Gamma(r)} (\lambda x)^{r-1} e^{-\lambda x}}, \quad x > 0 \quad (\lambda > 0, r > 0)$$

Here r and λ is the parameter. We say that $X \sim \text{Gamma}(r, \lambda)$.

[$f_X(0)$ cannot be defined in general]

What happens if $r = 1$. Then

$$f_X(x) = \frac{\lambda}{\Gamma(1)} (\lambda x)^0 e^{-\lambda x}, \quad x \geq 0, \lambda > 0$$

$$\therefore \boxed{f_X(x) = \lambda e^{-\lambda x}} \longrightarrow (*), \quad x \geq 0$$

Note that $\Gamma(1) = 1$, since $0! = 1$ (By convention)

The expression in $(*)$ is also a p.d.f which is called exponential distribution. Note for the exponential pdf $f_X(0)$ can be defined as λ, since $\lim_{x \to 0} f_X(x) = \lambda$.

The exponential distribution has the following amazing property called "Memoryless-ness". Let us see why it is called so.

Let $X \sim \exp(\lambda)$, i.e. X is a random variable with exponential distribution with parameter λ. Then

$$P[X > a+b \mid X > a] = P[X > b]. \quad a > 0, b > 0$$

It does not keep the information $X > a$, used in the conditioning.

$$P[X > a+b \mid X > a] = \frac{P[X > a+b \text{ & } X > a]}{P[X > a]}$$

$$= \frac{P[X > a+b]}{P[X > a]}$$

$$P[X > a+b] = 1 - P[X \leq a+b]$$

$$= 1 - \int_0^{a+b} \lambda e^{-\lambda x}\, dx$$

$$= 1 - \left[\lambda \int_0^{a+b} e^{-\lambda x}\, dx\right]$$

$$= 1 - \left[\lambda \left[\frac{e^{-\lambda x}}{-\lambda}\right]_0^{a+b}\right]$$

$$= 1 - \left[-e^{-\lambda x}\right]_0^{a+b}$$

$$= 1 - \left[-e^{\lambda(a+b)} + 1\right]$$

$$= 1 + e^{-\lambda(a+b)} - 1$$

$$= e^{-\lambda(a+b)}$$

Similarly $P[X > a] = e^{-\lambda a}$. Thus

$$P[X > a+b \mid X > a] = \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda b} = P[X > b].$$

The idea of memoryless-ness is used in reliability theory. For example let X denote the random variable which represents the the lifetime of a component in a device counted for example in hours. Then

$$P[X > a+b \mid X > a]$$

seeks to find the probability, that given that the component has functioned for more than a hours, what is the probability that it will work for more than additional b hours, i.e. it will work for $a+b$ hours. Memoryless-ness tells us that this probability is same as the probability that the component has worked more than b hours.

For seeing the application of exponential distribution see for example the book "Probability and Statistics, with reliability, queuing and computer Science applications" by Kishore Trivedi (Wiley 2002).

We ask the reader to compute the mean and variance of an exponential distribution. The answers are

$$E[X] = \lambda \quad \text{and} \quad \text{Var}(X) = \frac{1}{\lambda^2}$$

For the Gamma distribution, we shall compute the mean and variance using mgf technique. Of course we have

$$m_X(t) = E[e^{tX}]$$

$$= \int_0^\infty e^{tx} \frac{\lambda^r}{\Gamma(r)} e^{-\lambda x} x^{r-1} dx$$

$$= \int_0^\infty \frac{\lambda^r}{\Gamma(r)} e^{-(\lambda - t)x} x^{r-1} dx$$

$$= \frac{\lambda^r}{(\lambda-t)^r} \int_0^\infty \frac{1}{\Gamma(r)} e^{-(\lambda-t)x} x^{r-1} dx$$

$$= \lambda^r \int_0^\infty \frac{1}{\Gamma(r)} x^{r-1} e^{-(\lambda-t)x} dx$$

Set $(\lambda - t)x = y \Rightarrow dy = (\lambda - t)dx$

$$\therefore \quad m_X(t) = \frac{\lambda^r}{(\lambda-t)^r} \int_0^\infty \frac{(\lambda-t)^r}{\Gamma(r)} x^{r-1} e^{-(\lambda-t)x} dx \quad (\text{Assume } \lambda > t)$$

$$= \frac{\lambda^r}{(\lambda-t)^r} \frac{1}{\Gamma(r)} \int_0^\infty y^{r-1} e^{-y} dy$$

$$= \frac{\lambda^r}{(\lambda-t)^r} \frac{1}{\Gamma(r)} \Gamma(r) = \frac{\lambda^r}{(\lambda-t)^r}$$

$$\therefore m'_X(t) = r\lambda^r (\lambda - t)^{-r-1} \Rightarrow \boxed{E(X) = m'_X(0) = \frac{r}{\lambda}}$$

To compute $E[X^2]$, we need to compute $m_X''(0)$. Observe that

$$m_X''(t) = r(r+1)\lambda^r(\lambda - t)^{-r-2}$$

Hence

$$m_X''(0) = \frac{r(r+1)}{\lambda^2}$$

$$\therefore \quad \boxed{\text{Var}(X) = \frac{r(r+1)}{\lambda^2} - \frac{r^2}{\lambda^2} = \frac{r}{\lambda^2}}$$

The Gamma distribution plays a key role in ~~queu~~ queuing theory.

Section 2: The Beta Distribution

The ~~notion~~ notion of Beta distribution depends on the idea of Beta-functions an important class of functions in mathematical analysis, which is given as

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx, \qquad a>0,\ b>0.$$

To begin ~~wi~~ with we will first explore the relation between Beta and Gamma function. The ~~reati~~ relation is

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

(See any book on advanced calculus for a proof. You need double integrals for proving this). The Beta distribution has the pdf

$$f_X(x) = \frac{1}{B(a,b)}\, x^{a-1}(1-x)^{b-1}; \qquad 0<x<1$$

with $a>0$ & $b>0$.

It is of course simple to show that f_X is a p.d.f

We say $X \sim \text{Beta}(a,b)$. If $a=1$ & $b=1$, then the Beta distribution is just the uniform distribution. The distribution function or cdf can be in fact very compactly represented

$$F_X(x) = \begin{cases} 0, & x \le 0 \\ \int_0^x \frac{1}{B(a,b)} z^{a-1}(1-z)^{b-1}\,dz, & x \in (0,1) \\ 1, & x \ge 1. \end{cases}$$

Let us compute mean and variance of the beta distribution. We present here the approach given in the book "Introduction to the Theory of Statistics" by Mood, Graybill and Boes, Mc Graw Hill, 1974 Third Edn. We first compute

$$E[X^k] = \frac{1}{B(a,b)} \int_0^1 x^k\, x^{a-1}(1-x)^{b-1}\,dx$$

$$= \frac{1}{B(a,b)} \int_0^1 x^{k+a-1}(1-x)^{b-1}\,dx$$

$$= \frac{B(k+a,b)}{B(a,b)} = \frac{\Gamma(k+a+b)}{\cancel{\Gamma(k+a)\Gamma(b)}} \cdot \frac{\cancel{\Gamma(a+b)}}{\Gamma(a)\Gamma(b)}$$

$$= \frac{\Gamma(k+a)\Gamma(b)}{\Gamma(k+a+b)} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

$$\therefore E[X^k] = \frac{\Gamma(k+a)\Gamma(a+b)}{\Gamma(k+a+b)\Gamma(a)}$$

$$\therefore \text{For } k=1 \quad E[X] = \frac{\Gamma(a+1)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+1)}$$

$$= \frac{a\Gamma(a)\Gamma(a+b)}{\Gamma(a)\cdot(a+b)\Gamma(a+b)}$$

$$= \frac{a}{a+b}$$

Now we will set $k=2$

$$\therefore \quad E[X^2] = \frac{\Gamma(a+2)\,\Gamma(a+b)}{\Gamma(2+a+b)\,\Gamma(a)}$$

$$= \frac{(a+1)\,\Gamma(a+1)\,\Gamma(a+b)}{(a+b+1)\,\Gamma(a+b+1)\,\Gamma(a)}$$

$$= \frac{a\,(a+1)\,\Gamma(a)\,\Gamma(a+b)}{(a+b)\,(a+b+1)\,\Gamma(a+b)\,\Gamma(a)}$$

$$= \frac{a\,(a+1)}{(a+b)(a+b+1)}$$

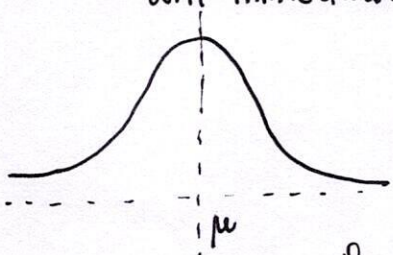
$$\text{Var}(X) = \frac{a\,(a+1)}{(a+b)(a+b+1)} - \left(\frac{a}{a+b}\right)^2$$

$$= \frac{ab}{(a+b)^2\,(a+b+1)}$$

So we now see the advantage of computing $E[X^n]$

Section 3: Normal distribution

Normal distribution is the poster-boy of statistics. Even many members of the general public have heard about it and will immediately recognize the bell-shaped curve, which you see on the left. On a more formal level we say that a continuous random variable X has a normal distribution if it has a p.d.f



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

Here μ and σ are the parameters, and we say ~~or rather~~ that X follows a normal distribution with ~~mean μ and~~ parameters μ & σ^2 and symbolically $X \sim N(\mu, \sigma^2)$.

It will turn out that μ and σ^2 are the ~~var~~ mean and variance of a normal random variable. We shall call a continuous random variable, normal random variable, if it follows the normal distribution. Mind you, ~~mathematics~~ mathematicians do not like the term normal distribution. They want to use the term Gaussian distribution, after Karl, Federich Gauss, the great mathematician who introduced the idea ~~whi~~ of the normal curve ~~white~~ while tabulating ashronomical data. Gauss was trying to measure the radius of the moon. As he repeated his measurement he got a different value (even if the difference was very less). When he plotted the values of his measurement he found that they seem to fall on a bell-shaped curve which the statisticians called as normal curve. The normal curve is also called an ~~error~~ curve as errors in measurements lie on the bell-shaped curve. We will say more on this later.

First let us show that f_x is a p.d.f. The fact that $f_x(x) \geq 0$ is obvious. Now observe that

$$\int_{-\infty}^{\infty} f_x(x)\,dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx$$

Let $\frac{x-\mu}{\sqrt{2}\,\sigma} = t \Rightarrow dx = \sqrt{2}\sigma\,dt$. Hence

$$\int_{-\infty}^{\infty} f_x(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2} \frac{1}{\sqrt{\pi}}\,dt \int_{-\infty}^{\infty} e^{-t^2}\,dt$$

$$= \frac{\sqrt{\pi}}{\sqrt{\pi}} = 1. \quad \left(\because \int_{-\infty}^{\infty} e^{-t^2}\,dt = 1, \text{ famous result in analysis}\right)$$

Ofcourse that shows that f_x is indeed a p.d.f

Suppose you do not know ~~to~~ the fact that $\int_{-\infty}^{\infty} e^{-t^2}\,dt = \sqrt{\pi}$. Then in that case we can show that $\int_{-\infty}^{\infty} f_x(x)\,dx$; using the techniques double integral. This attempt is made in Mood, Graybill Boes, [Though I believe they tacitly use the above formula].

Theorem 6.1: If $X \sim N(\mu, \sigma^2)$, Then

$$E[X] = \mu, \quad Var[X] = \sigma^2$$

and

$$m_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

Proof: You can of course directly compute the mean and the variance directly. However we will compute the $m_X(t)$ and through that compute μ & σ^2 through that.

$$m_X(t) = E[e^{tX}]$$

$$= E[e^{tX - t\mu + t\mu}]$$

$$= e^{t\mu} E[e^{t(x-\mu)}]$$

$$= e^{t\mu} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{t(x-\mu)} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}} dx$$

$$= e^{t\mu} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left[(x-\mu)^2 - 2\sigma^2 t^2 (x-\mu)\right]} dx$$

$$= e^{t\mu} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}}$$

Now

$$(x-\mu)^2 - 2\sigma^2 t^2 (x-\mu) = (x-\mu)^2 - 2\sigma^2 t (x-\mu) + \sigma^4 t^2 - \sigma^4 t^2$$

$$= (x - \mu - \sigma^2 t)^2 - \sigma^4 t^2$$

$$\therefore \quad m_X(t) = e^{t\mu} e^{\frac{\sigma^2 t^2}{2}} \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{[(x-\mu-\sigma^2 t)^2]}{2\sigma^2}} dx$$

$$= e^{t\mu + \frac{\sigma^2 t^2}{2}}$$

$$\because \quad \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-(\mu+\sigma^2 t))^2}{2\sigma^2}} dx = 1 \quad \text{(Think why!!)}.$$

$$\therefore \quad m_X(t) = e^{t\mu + \frac{1}{2}\sigma^2 t^2}$$

$$m_X'(t) = (\mu + \sigma^2 t)\, e^{t\mu + \frac{1}{2}\sigma^2 t^2}$$

$$\boxed{E(X) = m_X'(0) = \mu}$$

$$m_X''(t) = \sigma^2 e^{t\mu + \frac{1}{2}\sigma^2 t^2} + (\mu + \sigma t)(\mu + \sigma t)\, e^{t\mu + \frac{1}{2}\sigma^2 t^2}$$

$$m_X''(0) = \sigma^2 + \mu^2 = E(X^2)$$

$$\therefore \quad Var(X) = E(X^2) - [E(X)]^2$$

$$= \sigma^2 + \mu^2 - \mu^2$$

$$\boxed{Var(X) = \sigma^2}$$

If $\mu = 0$ and $\sigma^2 = 1$, we say X follows a standard normal distribution, and its pdf is

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

In fact if we consider $X \sim N(\mu, \sigma^2)$, then the random variable

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1).$$

The distribution function of a standard normal variable has a symbol $\Phi(x)$, ie

$$\Phi(x) = P(X \le x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}x^2}\, dx.$$

Now consider any $X \sim N(\mu, \sigma^2)$

then $$P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

Now observe

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

Now $\frac{X-\mu}{\sigma} \sim N(0,1)$. Thus if $X \sim N(\mu, \sigma^2)$

Then

~~$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$~~

$$\boxed{F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)}$$

The values of the function Φ is tabulated and these tables are available.

$$\therefore \quad P(a \leq X \leq b) = F_X(b) - F_X(a).$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$\therefore \quad \boxed{P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

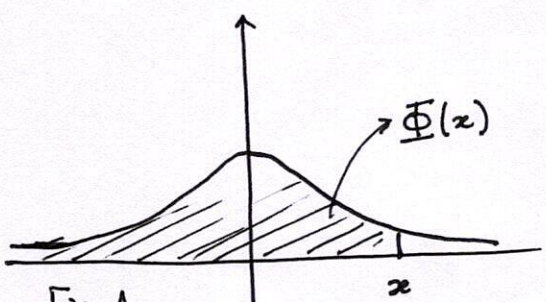


Fig A:

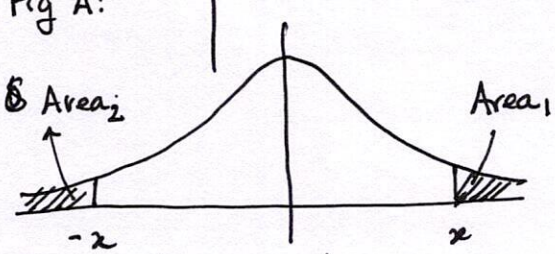


Fig B: Area 1 = Area 2 by symmetry.

For a more historical aside, let us note the De Moivre also had independently discovered the normal curve as a distribution of errors in measurement. Now observe that

$$\Phi(x) = P(X \leq x) = 1 - P(X > x)$$

$$= 1 - \int_{x}^{+\infty} f_X(x)\, dx$$

Now using Fig B we have

$$P(X > x) = P(X \leq -x)$$

Thus

$$\bar{\Phi}(x) = 1 - P(x \leq -x) \quad \left[\text{Always keep in mind that } P(x=x)=0\right]$$

$$= 1 - \bar{\Phi}(-x)$$

$$\boxed{\bar{\Phi}(x) = 1 - \bar{\Phi}(-x)}$$

The binomial distribution which we know as a discrete distribution actually links up ~~beab~~ beautifully with normal distribution when the number of trials is very large. ~~and thus links up beautifully with the normal distrib~~

Thus the binomial distribution is a link between the discrete and the continuous. The following result asserts this

Theorem 8.1: Let $X \sim B(n,p)$, then

$$Z = \frac{X - np}{\sqrt{npq}}$$

approaches the standard normal distribution with mean zero and variance 1 as $n \to \infty$.

Proof: (A sketch only)

How do we detect whether a given r.v. tends to some other distribution. Here we use the following approach. We show that as $n \to \infty$, the mgf of the binomial random variable will tend to the mgf of standard normal. We know from an assigment problem, that

- $m_{\frac{X+a}{b}}(t) = e^{\frac{a}{b}t} m_X\left(\frac{t}{b}\right)$ [See practice problem set No 2. Problem]

Since $X \sim B(n,p)$, we have

$$m_X(t) = \left(pe^t + (1-p)\right)^n, \qquad 1-p = q \text{ (in the notes)}$$

$\therefore$ For Z we have

$$m_Z(t) = e^{-\frac{np}{\sigma}t}\left[pe^{t/\sigma} + (1-p)\right]^n$$

where $\sigma = \sqrt{npq}$.

Let us set $np = \mu$. Then

$$m_Z(t) = e^{-\frac{\mu t}{\sigma}} \left[1 + p\left(e^{\frac{t}{\sigma}} - 1\right) \right]^n$$

In order to proceed to the limit we need to delink n & p since we will consider $n \to \infty$.

$$\therefore \quad \ln\left(m_Z(t)\right) = -\frac{\mu t}{\sigma} + n \ln\left[1 + p\left(e^{t/\sigma} - 1\right)\right]$$

$$= -\frac{\mu t}{\sigma} + n \ln\left[1 + p\left(\frac{t}{\sigma} + \frac{t^2}{2!\sigma^2} + \frac{1}{3!}\frac{t^3}{\sigma^3} + \cdots\right)\right]$$

Remember the series $\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \cdots$ converges $|x| < 1$. Assuming that such conditions hold we have (Think why)

$$\ln\left(m_Z(t)\right) = -\frac{\mu t}{\sigma} + np\left[\frac{t}{\sigma} + \frac{1}{2!}\frac{t^2}{\sigma^2} + \frac{1}{3!}\frac{t^3}{\sigma^3} + \cdots\right]$$

$$- \frac{np^2}{2}\left[\frac{t}{\sigma} + \frac{1}{2!}\frac{t^2}{\sigma^2} + \frac{1}{3!}\frac{t^3}{\sigma^3} + \cdots\right]^2$$

$$+ \frac{np^3}{3}\left[\frac{t}{\sigma} + \frac{1}{2!}\frac{t^2}{\sigma^2} + \frac{1}{3!}\frac{t^3}{\sigma^3} + \cdots\right] - \cdots$$

Now by collecting the powers of t we obtain

$$\ln\left(m_Z(t)\right) = \left(\frac{-\mu}{\sigma} + \frac{np}{\sigma}\right)t + \left(\frac{np}{2\sigma^2} - \frac{np^2}{2\sigma^2}\right)t^2$$

$$+ \left(\frac{np}{6\sigma^3} - \frac{np^2}{2\sigma^3} + \frac{np^3}{3\sigma^3}\right)t^3 + \cdots$$

$$= \frac{1}{\sigma^2}\left(\frac{np - np^2}{2}\right)t^2 + \left(\frac{np}{6\sigma^3} - \frac{np^2}{2\sigma^3} + \frac{np^3}{3\sigma^3}\right)t^3 + \cdots$$

$$= \frac{1}{\sigma^2} \cdot \frac{\sigma^2}{2} t^2 + \frac{n}{\sigma^3}\left[\frac{p - 3p^2 + 2p^3}{6}\right] t^3$$

$\therefore \ln$

$$\therefore \ln\left(m_Z(t)\right) = \frac{1}{2}t^2 + \frac{n}{\sigma^3}\left[\frac{p - 3p^2 + 2p^3}{6}\right]t^3 + \cdots\cdots$$

$$= \frac{1}{2}t^2 + \frac{n}{\left(\sqrt{npq}\right)^3}\left[\frac{p - 3p^2 + 2p^3}{6}\right]t^3 + \cdots\cdots$$

$$= \frac{1}{2}t^2 + \frac{n}{n^{3/2}\,p^{3/2}(1-p)^{3/2}}\left[\frac{p - 3p^2 + 2p^3}{6}\right]t^3 + \cdots$$

So $$\ln m_Z(t) = \frac{1}{2}t^2 + \frac{1}{\sqrt{n}\left(p^{3/2}(1-p)^{3/2}\right)}\left[\frac{p - 3p^2 + 2p^3}{6}\right]t^3 + \cdots$$

We can now safely argue, that for $r > 2$ the coefficient of t^r will go to zero, as $n \to \infty$. Hence

$$\ln m_Z(t) \longrightarrow \frac{1}{2}t^2 \quad \text{as } n \to \infty$$

as $\frac{1}{2}t^2$ is the mgf of the standard normal random variable.

What we Theorem 8.1, is a special case of the central limit theorem we will learn later on in the course.

[The sketch of the proof given here is from the book:

John. E. Freund's Mathematical Statistics by Miller & Miller Pearson : 2014.]

—x—

Lecture

Chapter 7: Joint and Conditional Distribution & Stochastic Independence

Suppose X and Y are two random variables on the same probability space. Can we give any meaning to the following question?

> Do these two variables taken jointly can have distribution?

Are these two variables independent of each other or are they related? Means we ask the question

> When is X and Y stochastically independent?

and

> If X and Y are related then we how do we measure it?

We seek to answer these questions here.

Section 1: Joint distribution of a random vector

In this chapter our aim is to study a random vector and its distribution. A random vector X is the following mapping

$$X : \Omega \longrightarrow \mathbb{R}^k$$

i.e. if $\omega \in \Omega$, then $X(\omega) = (x_1, x_2, \ldots, x_k)$

Note that each $x_1 \ldots x_k$ depend on $\omega \in \Omega$, thus can be viewed as a realization of a random variable. So we can write

$$X(\omega) = (X_1(\omega) \ldots X_k(\omega))$$

where $X_i(\omega) = x_i$, $i = 1, \ldots, k$. In short

$$X = (X_1, \ldots, X_k)$$

The expression

$$X \leq x \iff X_1 \leq x_1, \ldots, X_k \leq x_k$$

In fact

$$\{X \le x\} = \{X_1 \le x, \ldots X_k \le x\} = \bigcap_{i=1}^{k} \{ \bullet X_i \le x_i \}$$

The distribution function of this random vector X, is given as

$$F_X(x) = F_{X_1 \ldots X_k}(x_1 \ldots x_k) = P(X_1 \le x_1 \ldots \ldots X_k \le x_k)$$

Instead of focusing on k-random variables we just focus on two random variables X and Y in order to make the discussion simpler at least for the time being. For the sake of curiosity largely we state the properties of the distribution functions of two random variables, X and Y

- Properties of $F_{X,Y}(x,y)$ (never mind if you forget these immediately after reading)

i) $\lim_{x \to -\infty} F(x,y) = 0$, for all y & $\lim_{y \to -\infty} F(x,y) = 0$, for all x.

and

$$\boxed{\lim_{\substack{x \to \infty \\ y \to \infty}} F(x,y) = 1}$$

ii) $F(x,y)$ is right continuous in each variable

$$\lim_{h \to 0+} F(x+h, y) = \lim_{h \to 0+} F(x, y+h) = F(x,y).$$

Further one can relate the computing the probability using distribution functions. ~~in~~ Let $x_1 < x_2$ & $y_1 < y_2$, such that

$$P[x_1 < X \le x_2, \; y_1 < Y \le y_2]$$
$$= F(x_2, y_2) - F(x_2, y_1) - F(x_1 y_2) + F(x_1, y_1)$$

- **Discrete bi-variate r.v.**

Let X and Y be discrete random variable. Then its the joint probability mass function or joint pmf of X and Y, is given by

$$f_{X,Y}(x,y) = P[X=x, \; Y=y]$$

Of course one must have

- $f_{X,Y}(x,y) \geq 0, \quad \forall \; (x,y) \in Ran(X) \times Ran(Y)$
- ~~$\sum_x \sum_y f(x,y) = \int 1 = \sum_y \sum_x f(t$~~
- $\sum_x \sum_y f_{X,Y}(x,y) = 1 = \sum_y \sum_x f_{X,Y}(x,y)$.

Example: 6.1: Consider an urn having 3 red, 4 black and 1 green ball. Two balls are drawn at random. Let X be the random variable denoting the number of ~~green~~ red balls among the two drawn balls and Y denotes the number of black balls. Can we write down the pmf joint probability distribution

X	Y
0	0
1	1
2	2

X \ Y	0	1	2	$\sum_y f_{X,Y}(x,y)$
0	0	$\frac{4}{28}$	$\frac{6}{28}$	$\frac{10}{28}$
1	$\frac{3}{28}$	$\frac{12}{28}$	0	$\frac{15}{28}$
2	$\frac{3}{28}$	0	0	$\frac{3}{28}$
$\sum_x f_{X,Y}(x,y)$	$\frac{6}{28}$	$\frac{16}{28}$	$\frac{6}{28}$	1

Probability table

$\sum_x \sum_y f_{X,Y}(x,y)$

- The details of how we fill the table.

Some parts can be filled by the reader

There are 8 balls and we are choosing just two, ie in $\binom{8}{2}$ ways.

$$\binom{8}{2} = \frac{8!}{2!\,6!} = \frac{6!\times 7\times 8}{2!\,6!}$$

$$= \frac{56}{2} = 28$$

So there are 28 all possible out comes. Let us first compute

$$f_{X,Y}(0,0) = 0 \quad \text{(Think why!)}$$

$$f(0,1) = \frac{\binom{3}{0}\binom{4}{1}\binom{1}{1}}{28} = \frac{4}{28} \qquad [\,0! = 1 \text{ remember}\,]$$

$$f(0,2) = \frac{\binom{3}{0}\binom{4}{2}\binom{1}{0}}{28} = \frac{6}{28}$$

$$f(1,0) = \frac{\binom{3}{1}\binom{4}{0}\binom{1}{1}}{28} = \frac{3}{28}$$

$$f(1,1) = \frac{\binom{3}{1}\binom{4}{1}\binom{1}{0}}{28} = \frac{12}{28}$$

$$f(2,0) = \frac{\binom{3}{2}\binom{4}{0}\binom{1}{0}}{28} = \frac{3}{28}$$

$$f(2,1) = 0 \quad \& \quad f(2,2) = 0$$

I have skipped writing X, Y every time, so $f(0,1) = f_{X,Y}(0,1)$.

Looking at the table one can see that all the properties of the joint pmf is satisfied. □

Of course we can now write down the distribution function as follows

$$F_{X,Y}(x,y) = \sum_{s\le x}\sum_{t\le y} f(s,t), \qquad x\in\mathbb{R} \,\&\, y\in\mathbb{R}.$$

Exercise. Using the ~~above~~ table in the previous page compute $F_{XY}(2,2)$.

Now we can of course think of ~~independent~~ continuous random variables X, Y. Thus $f_{X,Y}(x,y)$ is a joint probability density function if

$$f_{X,Y}(x,y) \geq 0 \qquad \forall\ (x,y) \in \text{Ran}(X) \times \text{Ran}(Y)$$

and

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\, dy\, dx = 1$$

which is same as saying

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx\, dy = 1.$$

The distribution function can now be written as

$$F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(x,y)\, dx\, dy, \qquad x \in \mathbb{R},\ y \in \mathbb{R}$$

$$= P(X \leq x,\ Y \leq y).$$

For your happiness only.

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x\, \partial y}$$

~~I urge~~ We urge the reader to extend this definition for k random variables.

Example: Continuous case: Consider

$$f(x,y) = K(x+y), \qquad 0 < x < 1,\ 0 < y < 1$$

If $K \geq 0$, then $f(x,y) \geq 0$, $\forall\ (x,y) \in [0,1] \times [0,1]$

Now to be a joint ~~pmf~~ density we must have

$$\int_0^1\int_0^1 K(x+y)\, dx\, dy = 1$$

$$\therefore \quad K\int_0^1\left[\int_0^1 (y+x)\, dx\right] dy = 1 \Rightarrow K\int_0^1\left[y + \tfrac{1}{2}\right] dy = 1$$

$$\Rightarrow K\int_0^1 y\, dy + K\int_0^1 \tfrac{1}{2}\, dy = 1$$

$$\Rightarrow \frac{K}{2} + \frac{K}{2} = 1$$

$$\Rightarrow K = 1.$$

Knowing the joint distribution of (X, Y) can we get the individual distributions of X and Y.

The marginal distribution of X : $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$

The marginal distribution of Y: $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$

We have to show that $f_X(x)$ and $f_Y(y)$ are actually densities.

Of course we have ~~to show that~~ $f_X(x) \geq 0$ & $f_Y(y) \geq 0$, $\forall (x,y)$ $\in Ran(X) \times Ran(Y)$.

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\right]dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx = 1$$

$$\int_{-\infty}^{\infty} f_Y(y)\,dy = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\right]dy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1$$

We can also define all the above ideas for the discrete case.

Following on the ideas of conditional ~~distribution~~ probability, we define, what is called conditional pmf and conditional p.d.f

Let X and Y be two continuous random variables. Then the conditional density of Y given $X = x$ is given as

$$\boxed{f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}}$$

[This is defined if $f_X(x) > 0$ and not defined $f_X(x) = 0$]

Let X and Y be discrete, then, with $P[X = x_i] > 0$, $\forall x_i \in Ran(X)$, then

$$f_{Y|X}(y_j|x_i) = P[Y = y_i \,|\, X = x_i] = \frac{P[X = x_i,\ Y = y_j]}{P[X = x_i]}$$

Our question again remains the same. If $f_{Y|X}(y|x)$ a density of (X,Y) are jointly continuous random variable. Let us check this fact. If $f_{Y|X}(\cdot|x)$ is a density function of y. The fact that $f_{Y|X}(y|x) \geq 0$, $\forall y$ is clear. Then.

$$\int_{-\infty}^{\infty} f_{Y|X}(y|x)\,dy = \int_{-\infty}^{\infty} \frac{f_{Y|X}(y|x)\, f_{X,Y}(x,y)}{f_X(x)}\,dy$$

$$= \frac{1}{f_X(x)} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$

$$= \frac{1}{f_X(x)} \cdot f_X(x) = 1$$

Hence $f_{Y|X}(\cdot|x)$ is truly a density function. How can one write down the cumulative distribution function in this case. If $f_X(x) > 0$ then then the cdf of Y given $X = x$ is given as

$$F_{Y|X}(y|x) = \int_{-\infty}^{y} f_{Y|X}(z|x)\,dz$$

Let us now look into the important concept of independence of the random variables X and Y.

We say X and Y are two given random variables.

If X and Y are discrete

Independence $\Leftrightarrow$ Joint pmf $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ (marginal pmf)

If X and Y are continuous

Independence $\Leftrightarrow$ Joint density = Product of marginals.

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

$$\Rightarrow f_{Y|X}(y|x) = f_Y(y)$$

$$\& \; f_{X|Y}(x|y) = f_X(x)$$

A collection of k random variables $X_1, X_2, \ldots, X_k$ iff

$$f_{X_1 \ldots X_n}(x_1, x_2 \ldots x_n) = f_{X_1}(x_1) \cdot \ldots \cdot f_{X_n}(x_n).$$

Depends on wether you consider pmf a pdf.

E.g. $f_{X,Y}(x,y) = e^{-(x+y)}$; $x \geq 0, y \geq 0$

Is X and Y independent. Let us first compute the marginals

$$f_X(x) = \int_0^\infty e^{-(x+y)}\, dy = e^{-x}$$

$$f_Y(y) = \int_0^\infty e^{-(x+y)}\, dx = e^{-y}$$

$$f_{X,Y}(x+y) = e^{-(x+y)} = e^{-x}e^{-y} = f_X(x) f_Y(x).$$

———x———

tation, Covariance and Conditional Expectation

look at finding the expectation of a random function of a random vector $X: \Omega \to \mathbb{R}^k$ joint distribution is known.

$Y = g(X)$ be a function of random vector X and it is real valued, i.e. $g(X(\omega)) \in \mathbb{R}$, and $g : \text{Ran}(X) \to \mathbb{R}$.

We shall also study the way to measure the relation between two random variables X and Y, using an idea called covariance and correlation coefficient.

~~Sec 1 Ex~~ Section 1: Expectation of $g(x)$

Let $X : \Omega \longrightarrow \mathbb{R}^k$ be random-vector given as

$$X = (X_1, X_2, \ldots, X_k)$$

If each $X_1, X_2 \ldots X_k$ be discrete random variables and let $g : \mathbb{R}^k \longrightarrow \mathbb{R}$, then expectation of $g(X)$ is given as.

$$E[g(X_1 \ldots X_k)] = \sum g(x_1, x_2, \ldots x_k) f_X(x_1 \ldots x_k)$$

If $X_1, \ldots, X_k$ be are continuous random variables, then

$$E[g(X_1, \ldots, X_k)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1 \ldots x_k) f_X(x_1 \ldots x_k) dx_1 \ldots dx_k.$$

Theorem 8.1: Let X and Y be two continuous random variable with $f_{XY}(\cdot,\cdot)$, representing their joint density function. Then if $E(X) < \infty$ & $E(Y) < \infty$ and $E(X+Y) < \infty$ then

$$E(X+Y) = E(X) + E(Y).$$

[• Here $g(X,Y) = X+Y$. For two variables the density function is give written as $f_{X,Y}(x,y)$.]

Proof: $E(X+Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y) f_{X,Y}(x,y)\,dx\,dy$

$$\therefore E(X+Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f_{X,Y}(x,y)\,dx\,dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f_{X,Y}(x,y)\,dx\,dy$$

$$= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\right] dx + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} \left[f_{X,Y}(x,y)\,dx\right] dy$$

$$= \int_{-\infty}^{\infty} x\, f_X(x)\,dx + \int_{-\infty}^{\infty} y\, f_Y(y)\,dy \quad \left[\begin{array}{l}\text{Using the defn}\\ \text{of marginal}\\ \text{distribution}\end{array}\right]$$

$$\therefore E(X+Y) = E(X) + E(Y) \qquad \square$$

Think how will you extend it to more than 2 variables. In fact one can show that

$$E\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} E[X_i]. \longrightarrow (\#)$$

We urge the reader to prove the above result in Thm 8.1, for the discrete case.

What about $Var(X+Y)$, where X and Y are two random variables, with finite expectations.

$$Var(X+Y) = E\left[(X+Y - E(X+Y))^2\right]$$

$$= E\left[(X+Y - E(X) - E(Y))^2\right] \quad \text{Using Thm 8.1}$$

$$= E\left[(X - E(X) + (Y - E(Y))^2\right]$$

$$= E\left[(X-E(X))^2 + (Y-E(Y))^2 + 2(X-E(X))(Y-E(Y))\right]$$

$$= E\left[(X-E(X))^2\right] + E\left[(Y-E(Y))^2\right] + 2\,E\left[(X-E(X))(Y-E(Y))\right]$$

$\longrightarrow$ Using (#)

$$\therefore \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\,E\left[(X-E(X))(Y-E(Y))\right]$$

We shall now focus on the term

$$E\left[(X-E(X))(Y-E(Y))\right]$$

$$= E\left[XY - XE(Y) - YE(Y) + E(X)E(Y)\right]$$

$$= E[XY] - E[Y]E[X] - E[Y]E[X] + E[X]E[Y]$$

$$= E[XY] - E[X]E[Y]$$

$$\therefore \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\left[E[XY] - E[X]E[Y]\right]$$

What happens if X and Y are independent. What happens to the term $E[XY] - E[X]E[Y]$ in such a case.

Let just check out this for the continuous case. We will first compute $E[XY]$, Thus

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{X,Y}(x,y)\,dx\,dy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x) f_Y(y)\,dx\,dy \qquad \left[\begin{array}{l} f_{XY}(x,y) \\ \quad = f_X(x) f_Y(y) \\ \text{as } X \text{ and } Y \\ \text{are independent} \end{array}\right]$$

$$= \left(\int_{-\infty}^{\infty} x\, f_X(x)\,dx\right)\left(\int_{-\infty}^{\infty} y\, f_Y(y)\,dx\right)$$

$$= E[X]E[Y]$$

$$\therefore \boxed{E[XY] = E[X]E[Y]}$$

Theorem 8.2 Let X and Y are random variables, with finite expectations. Then If X and Y are independent then

$$E[XY] = E[X]E[Y].$$

Thus $E[XY] - E[X]E[Y] = 0$ if X and Y are independent

So we have the following: If X and Y are independent random variables, with finite expectations, then

$$\boxed{Var(X+Y) = Var(X) + Var(Y)}$$

In general we shall call the term

$$E[(X - E(X))(Y - E(Y)]$$

as the Covariance of X and Y as it seems to measure the expected joint variation of X and Y from their mean values. In symbols

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

$$\therefore Cov(X,Y) = E[XY] - E[X]E[Y]$$

Thus:

$$\boxed{\text{If } X \text{ and } Y \text{ are independent, then } Cov(X,Y) = 0}$$

The converse is not true. We shall provide examples but before that we state, that for any two random variables X and Y

$$\boxed{Var(X+Y) = Var(X) + Var(Y) + 2\,Cov[X,Y]}$$

- Note Covariance is symmetric in X and Y. Thus

$$\boxed{Cov(X,Y) = Cov(Y,X)}$$

Here we give two examples. One for the discrete case and other for the continuous case to show that $E[XY] = E[X]E[Y]$ is a necessary condition for X and Y to be independent, but not sufficient.

Eg 1: (from ~~Math. Stat~~ Freund's Math. Stat, by Miller & Miller)

Consider two discrete random variables X and Y, taking the values $-1, 0, +1$.

y \ x	-1	0	1	$P(Y=y)$
-1	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{2}{3}$
0	0	0	0	0
1	$\frac{1}{6}$	0	$\frac{1}{6}$	$\frac{1}{3}$
$P(X=x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1 → Total prob.

The above probability table gives the joint distribution of X and Y.

We have $E(X) = -1 \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = 0$

& $E(Y) = -1 \times \frac{2}{3} + 0 \times 0 + 1 \times \frac{1}{3} = -\frac{1}{3}$

$$E(X,Y) = (-1).(-1)\frac{1}{6} + (-1).(0) \times \frac{1}{3} + (-1)(-1)\frac{1}{6}$$

$$+ 1(-1).\frac{1}{6} + (1.1).\frac{1}{6} = 0$$

$$\therefore \boxed{E(X,Y) - E(X)E(Y) = 0}$$

Now consider $x=-1$, $y=-1$, then ~~$f(x,y) = \frac{1}{6}$~~

$$f_X(x,y) = f_X(-1,-1) = \frac{1}{6}$$

Now $f_X(x) = f_X(-1) = \frac{1}{3}$

$$f_Y(y) = f_Y(-1) = \frac{2}{3}$$

$$\therefore \boxed{f_X(x) f_Y(y) = \frac{2}{9} \neq \frac{1}{6} = f_X(x,y)} \Rightarrow X \text{ and } Y \text{ are not independent.}$$

E.g 2: Let X & Y be two continuous random variables given as

$$X = \sin 2\pi U$$

$$Y = \sin \cos 2\pi U$$

where U is a uniformly distributed random variable in $(0,1)$; i.e.

$$f_U(u) = \begin{cases} 1 & \text{if } u \in (0,1) \\ 0 & \text{otherwise.} \end{cases}$$

If X is known then $U = \frac{1}{2\pi} \sin^{-1} X$, and thus Y is known.

Now $E[Y] = \int_0^1 \cos 2\pi u \, du = 0$ & $E[X] = \int_0^1 \sin 2\pi u \, du = 0$

Now $E[XY] = \int_0^1 \sin 2\pi u \cos 2\pi u \, du = 0$

Note that,

$$XY = \sin 2\pi U \cos 2\pi U = g(U)$$

$$\therefore E[XY] = E[g(u)] = \int_0^1 g(u) f_u(u) \, du.$$

Here $E[XY] - E[X]E[Y] = 0$, though X and Y are not independent. □

Let us call two random variables X and Y, un-correlated if $\text{cov}(X,Y) = 0$.

From the above discussion we have learned the following

> If X and Y are independent $\Rightarrow$ X and Y are uncorrelated
> But the converse is not true

We can normalize the correlation measure, i.e. bring in the value of $\text{cov}(X,Y)$ with in the interval. This leads to the introduce the notion of a Pearson's correlation coefficient, denoted by σ_{XY} and denoted given as.

Thus we can write

$$\boxed{\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}}$$

Let us now focus on the following result, from Rohatgi [Statistical Inference Dover Pub. 2003.

8.3
Theorem ~~8.3~~ : If $E(X^2) < \infty$ & $E(Y^2) < \infty$, then cov(XY) exists and X and Y have have finite mean, then.

- $[\text{Cov}(XY)]^2 \leq \sigma_X^2 \sigma_Y^2$

ie $\boxed{[\text{Cov}(XY)]^2 \leq \text{Var}(X)\,\text{Var}(Y)}$. $\longrightarrow$ (#)

Proof: Set $E(X) = \mu_1$ & $E(Y) = \mu_2$.

$$\therefore \quad a(x-\mu_1)(y-\mu_2) \leq \frac{(x-\mu_1)^2 + (y-\mu_2)^2}{2}$$

$\therefore$ In terms of random variables, then

$$(X-\mu_1)(Y-\mu_2) \leq \frac{(X-\mu_1)^2 + (Y-\mu_2)^2}{2}.$$

$$\therefore \quad E\left[(X-\mu_1)(Y-\mu_2)\right] \leq \frac{1}{2}\text{Var}(X) + \frac{1}{2}\text{Var}(Y)$$

$$\therefore \quad E\left[(X-\mu_1)(Y-\mu_2)\right] \leq \frac{1}{2}\left[\text{Var}(X) + \text{Var}(Y)\right].$$

$$\therefore \quad \text{Cov}(X,Y) \leq \frac{1}{2}\left[\sigma_X^2 + \sigma_Y^2\right], \quad \text{where } \sigma_X^2 = \text{Var}(X) \text{ \& } \sigma_Y^2 = \text{Var}(Y)$$

Thus if $\sigma_X^2 < +\infty$ and $\sigma_Y^2 < +\infty$, then Cov(X,Y) is bounded above, and is also finite.

Let a & $b \in \mathbb{R}$, then

$$E\left[a(X-\mu_1) + b(X-\mu_2)\right]^2 = a^2\sigma_X^2 + 2ab\,\text{Cov}(X,Y) + b^2\sigma_Y^2 \longrightarrow (\odot)$$

Let us assume that either $\sigma_X^2 = 0$ or $\sigma_Y^2 = 0$, then (#) holds automatica

Now let us assume that $\sigma_X^2 > 0$ & $\sigma_Y^2 > 0$. The equation ⊙ holds for any a & $b \in \mathbb{R}$. Thus set

$$a = -\frac{Cov(X,Y)}{\sigma_X^2}$$

Now from ⊙ → $\frac{(Cov(X,Y))^2}{\sigma_X^4} \cdot \sigma_X^2 + b^2 \sigma_Y^2 - 2b \frac{Cov(X,Y)^2}{\sigma_X^2} \geq 0$ → ⊙⊙

The above expression is non-negative since

$$E\left[a(X-\mu_1) + b(X-\mu_2)\right]^2 \geq 0$$

Put $b = 1$ in ⊙⊙, and then we get

$$\frac{(Cov(X,Y))^2}{\sigma_X^2} + \sigma_Y^2 \geq 2\frac{Cov(X,Y)^2}{\sigma_X^2}$$

$$\Rightarrow \quad \frac{Cov(X,Y)^2}{\sigma_X^2} \leq \sigma_Y^2$$

$$\Rightarrow \quad \boxed{(Cov(X,Y))^2 \leq \sigma_X^2 \sigma_Y^2}$$

This completes the proof. □

Theorem 8.4: $-1 \leq \rho_{XY} \leq +1$, if $\sigma_X^2 > 0$, $\sigma_Y^2 > 0$ and finite.

Proof: From Theorem 8.3, $(Cov(X,Y))^2 \leq \sigma_X^2 \sigma_Y^2$

$$\Rightarrow \quad \left(\frac{Cov(X,Y)}{\sigma_X \sigma_Y}\right)^2 \leq 1$$

$$\Rightarrow \quad \rho_{XY}^2 \leq 1$$

$$\Rightarrow \quad -1 \leq \rho_{XY} \leq 1. \quad \square.$$

Section 8.2: Bivariate Normal Distribution

Can a normal distribution be defined in a joint manner for random variables X and Y. The answer turns out to be yes and before we proceed let us say little more about covariance. Sometimes one uses the symbol σ_{XY} to denote $\text{Cov}(X,Y)$. In this formalism, we have $\text{Var}(X) = \text{Cov}(X,X) = \sigma_{XX}$. Thus given any two random variables X and Y we call the following matrix Σ, as the variance given as

$$\Sigma = \begin{bmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{YX} & \sigma_{YY} \end{bmatrix}$$

as the variance-covariance matrix and often an helpful tool to represent variance. Let $a, b \in \mathbb{R}$, then

$$\begin{aligned} \text{Var}(aX+bY) &= a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\,\text{Cov}(X,Y) \\ &= a^2\sigma_{XX} + b^2\sigma_{YY} + 2ab\sigma_{XY} \\ &= \langle w, \Sigma w \rangle \end{aligned}$$

where $w = \begin{bmatrix} a \\ b \end{bmatrix}$. Note that as $\text{Var}(aX+bY) \geq 0$, we have $\langle w, \Sigma w \rangle \geq 0$, $\forall w \in \mathbb{R}^2$. Thus Σ is a positive semidefinite matrix. Another important idea is that of the moment generating function of more than one random variable. This is defined as follows

$$m_{X,Y}(t_1, t_2) = E\left[e^{t_1 X + t_2 Y} \right]$$

Note that if X and Y are independent, then

$$\begin{aligned} m_{XY}(t_1, t_2) &= E\left[e^{t_1 X} e^{t_2 Y} \right] \\ &= E\left[e^{t_1 X}\right] E\left[e^{t_2 Y}\right] \\ &= m_X(t_1)\, m_Y(t_2). \end{aligned}$$

Let us now see how do we compute the expectations in this case.

We have

$$E[X] = \frac{\partial m_{X,Y}(t_1,t_2)}{\partial t_1}\Big|_{(t_1,t_2)=(0,0)}$$

$$E[Y] = \frac{\partial m_{X,Y}(t_1,t_2)}{\partial t_2}\Big|_{(t_1,t_2)=(0,0)}$$

$$E[XY] = \frac{\partial^2 m_{X,Y}(t_1,t_2)}{\partial t_1 \partial t_2}\Big|_{(t_1,t_2)=(0,0)}.$$

I am not providing the details here which I believe can be checked easily by the reader. Further the above expressions can be written down by even thinking in an intuitive way. Till now the most important learning of the previous section is the following:

X and Y are independent $\Rightarrow$ X and Y are uncorrelated

However the converse is not true. Now let us ask the following question

Under what situation we can have
X and Y are independent $\Leftrightarrow$ X and Y are uncorrelated
if and only if

The answer is as follows: If X and Y jointly follow the bivariate normal distribution then the above assertion is true. So we shall start with the bivariate bivariate normal density.

The bivariate normal distribution has a joint pdf given by

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}}\, e^{-\frac{1}{2(1-\rho_{XY}^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho_{XY}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}}$$

where $-\infty < x < +\infty$, $-\infty < y < +\infty$, and $\sigma_X > 0$, $\sigma_Y > 0$ $\rho \in [-1, +1]$ are finite constants and further μ_X & μ_Y are also ~~finite~~ ~~constant~~ real-valued constants.

We shall first show that it is a pdf.

- $f_{X,Y}(x,y) \geq 0$, $\forall (x,y) \in \mathbb{R}^2$ is clear.
- We will now have to show that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx = 1$$

To begin with let us substitute

$$u = \frac{x-\mu_X}{\sigma_X} \quad \& \quad v = \frac{y-\mu_Y}{\sigma_Y}$$

Hence

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho_{XY}^2}}\, e^{-\frac{1}{2}\cdot\frac{1}{1-\rho_{XY}^2}\left(u^2 - 2\rho_{XY}uv + v^2\right)}\,du\,dv$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho_{XY}}}\, e^{-\frac{1}{2(1-\rho_{XY}^2)}\left[(u-\rho_{XY}v)^2 + (1-\rho_{XY}^2)v^2\right]}\,du\,dv$$

(Check it yourself)

- Again let us substitute keeping v fixed.

$$w = \frac{u-\rho_{XY}v}{\sqrt{1-\rho_{XY}^2}}, \quad \therefore dw = \frac{du}{\sqrt{1-\rho^2}}$$

$$\therefore \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,dy\,dx = \left(\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-w^2/2}\,dw\right)\left(\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-v^2/2}\,dv\right) = 1$$

(Think why??)

(Think about the standard normal distribution).

Thus f_{XY} is a pdf.

Let us now compute the mgf of the ~~bi-v~~ bivariate normal distribution

$$m_{X,Y}(t_1,t_2) = E\left[e^{t_1X+t_2Y}\right]$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1x+t_2y} f_{XY}(x,y)\,dy\,dx$$

Set $u = \frac{x-\mu_X}{\sigma_X}$ & $v = \frac{y-\mu_Y}{\sigma_Y}$, hence

$$m_{X,Y}(t_1,t_2) = e^{t_1\mu_X + t_2\mu_Y}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{e^{t_1\sigma_X u + t_2\sigma_Y v}}{2\sqrt{1-\rho_{XY}^2}}\, e^{-\frac{1}{2}\frac{1}{(1-\rho_{XY}^2)}\left(u^2-2\rho_{XY}uv+v^2\right)}\,dv\,du$$

$$\therefore m_{X,Y}(t_1,t_2) = e^{t_1\mu_X+t_2\mu_Y}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{2\sqrt{1-\rho_{XY}^2}}\, e^{-\frac{1}{2(1-\rho_{XY}^2)}\left[u^2-2\rho_{XY}uv+v^2-2(1-\rho^2)t_1\sigma_X\right.}$$
$$\left.-2(1-\rho^2)t_2\sigma\right.$$

$$= \frac{e^{t_1\mu_X+t_2\mu_Y}}{2\sqrt{1-\rho_{XY}^2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho_{XY}^2)}\left[\left[u-\rho_{XY}v-(1-\rho_{XY}^2)t_1\sigma_X\right)^2+(1-\rho^2)\left(v-\rho_{XY}t_1\sigma_X\right.\right.}$$
$$\left.\left.-t_2\sigma_Y\right)\right.$$
$$-(1-\rho_{XY}^2)\left(t_1\sigma_X^2+t_2\sigma_Y\right.$$
$$\left.+2\rho_{XY}t_1t_2\sigma_X\right.$$
$$dv\,du$$

Let $w = \frac{u-\rho_{XY}v-(1-\rho_{XY}^2)t_1\sigma_X}{\sqrt{(1-\rho_{XY}^2)}}$

$$z = v-\rho_{XY}t_1\sigma_X - t_2\sigma_Y$$

$$\therefore m_{X,Y}(t_1,t_2) = e^{t_1\mu_X+t_2\mu_Y}\, e^{\left[\frac{1}{2}\left[t_1^2\sigma_X^2+2\rho_{XY}t_1t_2\sigma_X\sigma_Y+t_2^2\sigma_Y^2\right]\right]}$$
$$\times\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{2\pi}e^{-w^2/2-z^2/2}\,dw\,dz$$

$$\therefore m_{X,Y}(t_1,t_2) = e^{t_1\mu_X+t_2\mu_Y+\frac{1}{2}\left(t_1^2\sigma_X^2+2\rho_{XY}t_1t_2\sigma_X\sigma_Y+t_2^2\sigma_Y^2\right)}$$

[Note that $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{2\pi}e^{-w^2/2-z^2/2}\,dw\,dz = 1$]

↓

Think why

Set $h = t_1\mu_X+t_2\mu_Y+\frac{1}{2}\left(t_1^2\sigma_X^2+2\rho_{XY}t_1t_2\sigma_X\sigma_Y+t_2^2\sigma_Y^2\right)$

$$\frac{\partial m_{X,Y}(t_1,t_2)}{\partial t_1} = \left(\mu_X + t_1\sigma_X^2 + \rho_{XY} t_2 \sigma_X \sigma_Y\right) e^{h}$$

$$\frac{\partial m_{X,Y}(t_1,t_2)}{\partial t_2} = \left(\mu_Y + t_2\sigma_Y^2 + \rho_{XY} t_1 \sigma_X \sigma_Y\right) e^{h}$$

$$\frac{\partial^2 m_{XY}(t_1,t_2)}{\partial t_1 \partial t_2} = \rho_{XY}\sigma_X\sigma_Y e^{h} + \cancel{\left(\mu_X + t_2\sigma_Y^2 + \rho_{XY} t_1 \sigma_X \sigma_Y\right)}$$
$$\left(\mu_Y + t_2\sigma_Y^2 + \rho_{XY} t_1 \sigma_X \sigma_Y\right)$$
$$\left(\mu_X + t_1\sigma_X^2 + \rho_{XY} t_2 \sigma_X \sigma_Y\right)^*$$

$$\therefore \left.\frac{\partial m_{X,Y}(t_1,t_2)}{\partial t_1}\right|_{(t_1,t_2)=(0,0)} = \mu_X = E(X)$$

$$\therefore \left.\frac{\partial m_{X,Y}(t_1,t_2)}{\partial t_2}\right|_{(t_1,t_2)=(0,0)} = \mu_Y = E(Y)$$

$$\left.\frac{\partial^2 m(t_1,t_2)}{\partial t_1 \partial t_2}\right|_{(t_1,t_2)=(0,0)} = \cancel{\mu} \rho_{XY}\sigma_X\sigma_Y + \mu_X\mu_Y$$

$$\therefore \text{Cov}(X,Y) = \sigma_{XY} = E[XY] - E[X]E[Y]$$
$$= \rho_{XY}\sigma_X\sigma_Y + \mu_X\mu_Y - \mu_X\mu_Y$$

$$\therefore \text{Cov}(X,Y) = \rho_{XY}\sigma_X\sigma_Y$$

$$\therefore \text{Correlation coefficient} = \rho_{XY}.$$

We leave it to the reader to check that $\text{Var}(X) = \sigma_X^2$ & $\text{Var}(Y) = \sigma_Y^2$

Now we come to the main result of this section.

Theorem 8.5 : If X and Y are two random variables which jointly follow the bivariate normal distribution, Then X and Y are independent $\Leftrightarrow$ X and Y are uncorrelated.

Proof: If X and Y are independent, then X and Y are uncorrelated is a known fact.

For the converse, let X and Y be uncorrelated

$$\Rightarrow \quad \mathrm{Cov}(X,Y) = 0$$

$$\Rightarrow \quad \rho_{XY} = 0$$

Thus

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}$$

$$= \left[\frac{1}{\sqrt{2\pi}\,\sigma_x} e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2}\right]\left[\frac{1}{\sqrt{2\pi}\,\sigma_y} e^{-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2}\right]$$

$$= f_X(x)\, f_Y(y),$$

where $X \sim N(\mu_x, \sigma_x^2)$ & $Y \sim N(\mu_y, \sigma_y^2)$.

Of course one can easily check that when $\rho_{x,y} = 0$, the marginal distributions are $f_X(x)$ and $f_Y(y)$. This shows that X and Y are independent random variable $\square$.

— x —

Section 8-3: Conditional Expectation

Conditional Expectation is a natural outcome from the fact that we have defined the notion of a conditional density. Let us define it formally

Conditional Expectation of the r.v.'s X and Y

Let (X,Y) be a two-dimensional random-vector. Let $Z = g(X,Y)$ be a function of the two random variables.

Then the conditional expectation of $g(X,Y)$, given $X = x$ is given as

$$E\left[g(X,Y) \mid X = x\right] = \int_{-\infty}^{\infty} g(x,y)\, f_{Y|X}(y|x)\, dy$$

This formula holds good if X and Y are jointly continuous

If X and Y are jointly discrete we have

$$E\left[g(X,Y) \mid X = x\right] = \sum_{j} g(x, y_j)\, f_{Y|X}(y_j|x).$$

E.g: Let $f_{X,Y}$ be the joint density of two continuous random variable given as

$$f_{X,Y}(x,y) = \begin{cases} x+y, & \text{if } x \in (0,1) \text{ \& } y \in (0,1) \\ 0 & \text{otherwise.} \end{cases}$$

How shall we now proceed to compute $E[Y \mid X = x]$ (i.e. $g(X,Y) = Y$)

First we compute

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{x+y}{f_X(x)}, \qquad x, y \in (0,1).$$

$$f_X(x) = \int_0^1 (x+y)\, dy = x + \tfrac{1}{2} \quad \therefore \quad \boxed{\frac{x+y}{x+\frac{1}{2}} = f_{Y|X}(y,x)}$$

[Here $g(X,Y) = Y$. In fact $E[g(X,Y) \mid X = x]$ is a function of X]

Thus

$$E[Y|X=x] = \int_0^1 y\, f_{Y|X}(y|x)\, dy$$

$$= \int_0^1 y\, \frac{x+y}{x+1/2}\, dy$$

$$= \frac{1}{x+1/2}\left[\int_0^1 yx\, dy + \int_0^1 y^2\, dy\right]$$

$$= \frac{1}{x+1/2}\left[\frac{x}{2} + \frac{1}{3}\right]. \qquad \square$$

For simplicity, consider $E[g(Y)|x]$, which in general a function of x. Let us denote it as

$$\phi(x) = E[g(Y)|x]$$

$\therefore$ $\phi(x)$ can also be written as

$$\phi(X) = E[g(Y)|X]$$

$\therefore$ For a given $\omega \in \Omega$

$$\phi(X(\omega)) = E[g(Y)|X(\omega)]$$

$\therefore$ If $X(\omega) = x$, then we have

$$\phi(x) = E[g(Y)|x]$$

Note that in general $\phi(X) = E[g(Y)|X]$ can be viewed as a random variable.

So our learning here is as follows:

> Conditional Expectation $E[g(Y)|X]$ is a random variable.

It might be sometimes relevant to ask, "What is the expectation of $\phi(X)$". The answer may surprise you. We do the calculation for $g(Y) = Y$. Assume $f_X(x) > 0, \forall x$

Set $E[\phi(X)] = E[E[Y|X]]$

$$= E[\phi(X)] = \int_{-\infty}^{\infty} \phi(x) f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} E[Y|x]\ f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} y f_{Y|X}(y|x)\, dy \right] f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\ f_{Y|X}(y|x)\, f_X(x)\, dy\, dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y \cdot \frac{f_{X,Y}(x,y)}{f_X(x)} \cdot f_X(x)\, dy\, dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\ f_{X,Y}(x,y)\, dy\, dx$$

$$= E[Y].$$

$$\therefore \boxed{E[E[Y|X]] = E[Y]}$$

(The reader can prove it for the discrete case).

We shall now define the notion of conditional variance of Y when the random variable X takes a particular value say x.

$$\mathrm{Var}(Y|X=x) = E[Y^2|X=x] - [E[Y|X=x]]^2.$$

In way we discussed before the case of the conditional expectation, the conditional variance can also be viewed as a r.v. $\mathrm{Var}(Y|X)$, whose value at any $\omega \in \Omega$, is computed as

$$\boxed{\mathrm{Var}(Y|X)(\omega) = \mathrm{Var}(Y|X=X(\omega))}$$

$X(\omega)$ can of course take the value of x.

We end our discussion with the following interesting result.

Theorem 8.4 Let X and Y be random variables. Then

$$\mathrm{Var}(Y) = E[\mathrm{Var}(Y|X)] + \mathrm{Var}[E[Y|X]).$$

Proof: $E[\mathrm{Var}(Y|X)] = E[E[Y^2|X] - (E[Y|X])^2]$

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

$$= E[Y^2] - E[E[Y|X]^2]$$

$\therefore$ In fact $E[E[Y^2|X]] = E[Y^2]$, can be proved from the proof style of showing $E[E[Y|X]] = E[Y]$

$\therefore$ $E[\mathrm{Var}(Y|X)] = E[Y^2] - (E[Y])^2 - E[(E[Y|X])^2] + (E[Y])^2$

$\therefore$ $E[\mathrm{Var}(Y|X)] = \mathrm{Var}(Y) - [E[(E(Y|X))^2] - (E(E(Y|X)))^2]$

$$= \mathrm{Var}(Y) - \mathrm{Var}(E[Y|X])$$

$\therefore$ $\mathrm{Var}(Y) = E[\mathrm{Var}(Y|X)] + \mathrm{Var}(E[Y|X])$.

—x—x—

Lecture 9: Transformation of Random Variables & their Distributions

Suppose we have a random variable X and another random Y is defined as $Y = g(X)$. If we assume just for the sake of it that X is a continuous random variable with a p.d.f. Sometimes specially in statistics, we need to know the density of Y. We can also have more than one random variables, $X_1, X_2, \ldots X_n$ and suppose we know their joint pmf/pdf. Then we might want to know the pmf/pdf of

$$Y = g(X_1, X_2, \ldots, X_n)$$

In this chapter we shall show how can we compute $f_Y(y)$. There are several approaches and as we will see what kind of technique to use depends on the problem at hand.

Section 1: Distribution function Approach

One of the simplest approaches is to compute the distribution function of Y and then differentiate it to get the p.d.f

We have $F_Y(y) = P(Y \leq y) = P(g(x_1 \ldots x_n) \leq y)$

$\therefore$ Then $f_Y(y) = \dfrac{dF_Y}{dy}$

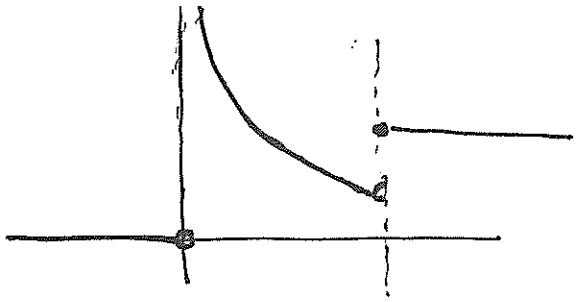
E.g. 1. Let $X \sim \text{Uniform}(0,1)$, then if $Y = g(X) = X^2$. We shall find the distribution of Y.

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = \int_{\{x: x^2 \leq y\}} f_X(x)\, dx$$

$$= \int_0^{\sqrt{y}} dx \qquad (0 < y < 1)$$

$$= \sqrt{y}, \qquad \text{when } 0 < y < 1.$$

$$\therefore F_Y(y) = \begin{cases} 0, & y \le 0 \\ \sqrt{y}, & 0 < y < 1 \\ 1, & y \ge 1 \end{cases}$$



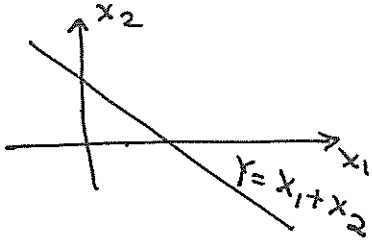
$$\therefore f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}}, & 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

E.g. 2: If X_1 is a continuous r.v. & X_2 is a continous r.v. whose joint density is given as

From Miller & Miller Freund's Mathematical

$$f_{X_1,X_2}(x_1, x_2) = \begin{cases} 6e^{-3x_1 - 2x_2}, & x_1 > 0, \; x_2 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Find the distribution of $Y = X_1 + X_2$ and also its probability density



$$F_Y(y) = P(Y \le y)$$

$$= P(X_1 + X_2 \le y)$$

$$= P(X_1 \le y - X_2, \; X_2 \le y)$$

$$= \int_0^y \int_0^{y-x_2} 6e^{-3x_1 - 2x_2} \, dx_1 \, dx_2$$

$$= 1 + 2e^{-3y} - 3e^{-2y}. \quad \text{(when: } (y > 0)\text{)}$$

Thus $$f_Y(y) = 6(e^{-2y} - e^{-3y}), \quad \text{for } y > 0$$

& $f_Y(y) = 0$, elsewhere.

Section 2: Transformation Approach : Single r.v.

So in this section we consider the case $Y = u(X)$, and given a probability mass / density of X, we would like to know what is the probability distribution of Y.

Consider the following simple example from Miller and Miller [ ]

Eg 1: X be the r.v. denoting the number of heads in a toss of ~~four coins~~ a fair coin four times.
Find the distribution of $Y = \frac{1}{1+X}$

[e.g. from Miller and Miller
Freund's Mathematical Statistics]

The distribution of X is given as

x	$f_X(x)$
0	$1/16$
1	$4/16$
2	$6/16$
3	$4/16$
4	$1/16$

Here $X \sim B(4, \frac{1}{2})$

So Y takes the values, $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}$ corresponding to $X=0, X=1, X=2, X=3, X=4$, respectively

y	$g_Y(y)$
1	$1/16$
$1/2$	$4/16$
$1/3$	$6/16$
$1/4$	$4/16$
$1/5$	$1/16$

So we have $g_Y(y) = f_X\left(\frac{1}{y} - 1\right)$. But observe that

$X = \frac{1}{y} - 1$.

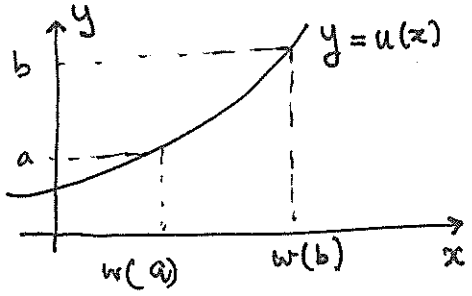
For the continuous case however we need to rely on the following result.

Theorem 9.1: Let X be a random variable, which is continuous one with p.d.f $f_X(x)$. Let $y = u(x)$ be a function of x, which is differentiable and monotone (either increasing or decreasing) over ~~Range(X)~~ the set $\{x \in \text{Range } x : f_X(x) \neq 0\}$. Then for these values of x, the function u has an inverse w, for which $x = w(y)$. Then the probability density ~~is given~~ of $Y = u(X)$ is given as

$$g_Y(y) = f_X[w(y)]\,|w'(y)|, \text{ provided } u'(x) \neq 0$$

and $g_Y(y) = 0$, elsewhere. We assume u and w are differentiable

Proof: Let us first consider the case when $y = u(x)$ is increasing



Now

$$P[a < Y < b] = P[w(a) < x < w(b)]$$

$$= \int_{w(a)}^{w(b)} f_X(x)\,dx$$

Now substitute in the integral, $x = w(y)$. $\therefore$ ~~$w(a)$ when~~

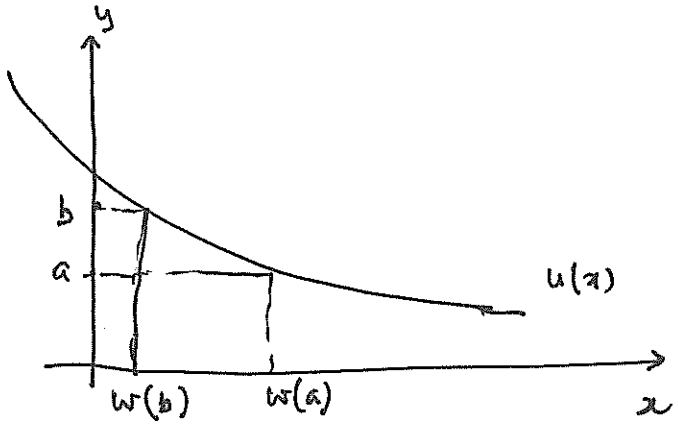
$\therefore$ When $x = w(a)$, we have $y = a$, & when $x = w(b)$, we have $y = b$ and $dx = w'(y)\,dy$.

$$\therefore \quad P[a < X < b] = \int_a^b f_X(w(y))\,w'(y)\,dy$$

Thus the probability density is given by the function

$$g_Y(y) = f_X(w(y))\,w'(y).$$

Let us consider now u to be decreasing



In this case we have

$$\therefore \quad P[a < Y < b] = P[w(b) < X < w(a)]$$

$$= \int_{w(b)}^{w(a)} f_X(x)\,dx$$

$$= \int_b^a f_X(w(y))\,w'(y)\,dy$$

$$= -\int_a^b f_X(w(y))\,w'(y)\,dy$$

In this case we have

$$g_Y(y) = -f_X(w(y))\,w'(y)\,dy$$

In fact. $w'(y) = \dfrac{dx}{dy} = \dfrac{1}{\frac{dy}{dx}}$ i.e. $w'(y) = \dfrac{1}{u'(x)}$

In fact when u is increasing $w'(y)$ is positive and when u is ~~dere~~ decreasing $-w'(y)$ is positive. Hence compactly we write

$$g_Y(y) = f_X[w(y)]\,|w'(y)|$$

or $\quad g_Y(y) = f_X(u^{-1}(y))\left|\dfrac{dx}{dy}\right|$; (Here $w \equiv u^{-1}$)

$$\boxed{g_Y(y) = f_X(u^{-1}(y))\left|\frac{dx}{dy}\right|}$$

, $u'(x) \neq 0$ & $x = u^{-1}(y)$.

E.g 12: Let $X \sim N(0,1)$. Then find the density of

$$Z = u(X) = X^2.$$

Note that the function $z = x^2$, is decreasing when $x < 0$ and increasing when $x > 0$. So conditions of the Theorem 9.1 are not met.

The key to this is to take an additional step and have a random variable on whose range the given transformation

This brings us to the case what happens if $u(x)$ behaves differently on different parts of $\{x \in Ran(X) : u(x) \neq 0\}$.

Assume that we can partition the set $\{x \in Ran(X) : u(x) \neq 0\}$ into a finite partion $\{A_n\}_{n=1}^{k}$, ie.

$$\bigcup_{n=1}^{k} A_n = \{x \in Ran(X) : u(x) \neq 0\}$$

& $A_i \cap A_j = \phi, \ \forall\ i \neq j, \quad i, j = 1, \ldots, k.$

So each A_n, the function u is either increasing or decreasing.

Thus define

$$u_n(x) = \begin{cases} u(x), & x \in A_n \\ 0, & x \text{ is otherwise.} \end{cases}$$

$\therefore$ u_n has a unique inverse in A_n. Thus

$y = u_n(x)$ ~~$y = g_n(x)$~~, $x = u_n^{-1}(y)$, for $x \in A_n$

Then we can get the pdf on the individual A_n and then sum

i.e. $$g_Y(y) = \sum_{n=1}^{k} f_X(u_n^{-1}(y)) \left| \frac{d}{dy} u_n^{-1}(y) \right|$$

~~$g_Y(y) = \sum_{n=1}^{k} f_X$~~

~~$g_Y(y) = \sum_{n=1}^{k} f_X(g_n u_n^{-1}(y)) \left| \frac{d}{dy} u_n^{-1}(y) \right|$~~

So in our particular case we have

$$\{x \in \mathbb{R} : x^2 \neq 0\} = \underbrace{\{x \in \mathbb{R}, x^2 < 0\}}_{A_1} \cup \underbrace{\{x \in \mathbb{R} : x^2 > 0\}}_{A_2}$$

So on $(-\infty, 0)$ we have $x_1 = -\sqrt{y}$ & on $(0, \infty)$ we have $x = +\sqrt{y}$. Here $y > 0$. as $y = x^2 > 0$.

$$\therefore g_Y(y) = f_X(-\sqrt{y}) \frac{1}{2\sqrt{y}} + f_X(\sqrt{y}) \frac{1}{2\sqrt{y}}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-y/2} \frac{1}{2\sqrt{y}} + \frac{1}{2\sqrt{2y}} \cdot \frac{1}{\sqrt{2\pi}} e^{-y/2}$$

$$= \frac{1}{2}\left(\frac{1}{\sqrt{2\pi y}} e^{-y/2}\right) + \frac{1}{2}\left(\frac{1}{\sqrt{2\pi y}} e^{-y/2}\right)$$

$$\therefore \boxed{g_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2}, \quad y > 0}$$

Of course $g_Y(y) = 0$, if $y \leq 0$.

In fact g_Y is often said to be the pdf of a χ^2-distri random variable with degrees of freedom 1. (more on this later).

obsv: $Y \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2})$ is call χ^2-with one d of free

Section 3: Distribution of the sum and difference of two r.v.

In this section we will be concerned with the distribution of the sum and differences of two random variables. These notions would play an important role in sampling theory. So we begin by presenting the analysis in forms of theorems, so that the reader appreciates the importance of these results. We present these results from "Introduction to the Theory of Statistics" by Mood, Graybill, Boyes.

Theorem 9.1: Let X and Y be two continuous random variables, with joint pdf $f_{X,Y}(x,y)$. Let $Z = X+Y$ & $V = X-Y$, then we have

A) $$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x)\,dx = \int_{-\infty}^{\infty} f_{X,Y}(z-y, y)\,dy$$

B) $$f_V(v) = \int_{-\infty}^{\infty} f_{X,Y}(x, x-v)\,dx = \int_{-\infty}^{\infty} f_{X,Y}(v+y, y)\,dy$$

We will just prove the first part of case A). The rest can be proved in an analogous way.

Proof:

$$F_Z(z) = P(Z \le z)$$

$$= P(X+Y \le z)$$

$$= \iint_{x+y \le z} f_{X,Y}(x,y)\,dx$$

$$= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{z-x} f_{X,Y}(x,y)\,dy\right]dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{z}\left[f_X(x, u-x)\,du\right]dx \quad \left[\text{By substituting } y = u-x\right]$$

$$= \int_{-\infty}^{z}\left[\int_{-\infty}^{\infty} f_{X,Y}(x, u-x)\,dx\right]du$$

$$\therefore \quad f_Z(z) = \frac{dF_Z(z)}{dz} = \frac{d}{dz}\int_{-\infty}^{z}\left[\int_{-\infty}^{\infty} f_{X,Y}(x, u-x)\,dx\right]du$$

$$= \int_{-\infty}^{\infty} f_{X,Y}(x, z-x)\,dx \quad \left[\text{Use the fundamental theorem of Calculus but for the improper integrals.}\right]$$

Corollary 9.1: Let X and Y are now anumed to be independent, and $Z = X+Y$.

Then.

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)\, f_Y(z-x)\, dx \quad * = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy$$

Section 4: The distribution of products and ratio

Theorem 9.2: Let X and Y are continuous random variables with joint p.d.f $f_{X,Y}(x,y)$ and let $Z = XY$ and $U = \frac{X}{Y}$.

Then;

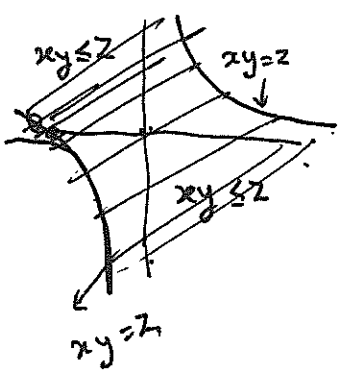
a) $$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|x|} f_{X,Y}\left(x, \frac{z}{x}\right) dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{|y|} f_{X,Y}\left(\frac{z}{y}, y\right) dy$$

b) $$f_U(u) = \int_{-\infty}^{\infty} |y|\, f_{X,Y}(uy, y)\, dy$$

We shall only prove the first part of a), the rest can be similarly worked out by the reader.

$$F_Z(z) = P(Z \le z) = \iint_{xy \le z} f_{X,Y}(x,y)\, dx\, dy$$



$$= \int_{-\infty}^{0} \left[\int_{z/x}^{\infty} f_{X,Y}(x,y)\, dy \right] dx + \int_{0}^{\infty} \left[\int_{-\infty}^{z/x} f_{X,Y}(x,y)\, dy \right]$$

■

Let us now substituting $u = xy$, we have

$$F_Z(z) = \int_{-\infty}^{0} \left[\int_{z}^{-\infty} f_{X,Y}\left(x, \frac{u}{x}\right) \frac{du}{x} \right] dx + \int_{0}^{\infty} \left[\int_{-\infty}^{z} f_{X,Y}\left(x, \frac{u}{x}\right) \frac{du}{x} \right] dx$$

$$= \int_{-\infty}^{z} \left[\int_{-\infty}^{0} \frac{1}{-x} f_{X,Y}\left(x, \frac{u}{x}\right) dx \right] du + \int_{-\infty}^{z} \left[\int_{0}^{\infty} \frac{1}{x} f_{X,Y}\left(x, \frac{u}{x}\right) dx \right] du$$

$$= \int_{-\infty}^{z} \left[\int_{-\infty}^{\infty} \frac{1}{|x|} f_{X,Y}\left(x, \frac{u}{x}\right) dx \right] du.$$

Hence

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \int_{-\infty}^{\infty} \frac{1}{|x|} f_{X,Y}\left(x, \frac{u}{x}\right) dx.$$

<u>Example:</u> Let X and Y are independent uniform random variables in $(0,1)$. Compute the distribution of $Z = XY$.

We have $f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|x|} f_{X,Y}\left(x, \frac{z}{x}\right) dx$

$= \int_{-\infty}^{\infty} \frac{1}{|x|} f_X(x) f_Y\left(\frac{z}{x}\right) dx$, by independence of X and Y.

$= \int_{0}^{1} \frac{1}{x} dx$. Now $f_Y\left(\frac{z}{x}\right) = 1$, if $0 \leq \frac{z}{x} \leq 1$

& $f_X(x) = 1$ if ~~$0 \leq x \leq 1$~~ $0 < x < 1$

$$\therefore \quad f_Y\left(\frac{z}{x}\right) = I_{(0,1)}\left(\frac{z}{x}\right)$$

$$f_X(x) = I_{(0,1)}(x)$$

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

Indicator function or characteristic function

Now $I_{(0,1)}(x)\, I_{(0,1)}\left(\frac{z}{x}\right)$

$= I_{(0,1)}(z)\, I_{(z,1)}(x)$ (Try to establish this)

$$\therefore \quad f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|x|} I_{(0,1)}(x)\, I_{(0,1)}\left(\frac{z}{x}\right) dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{|x|} I_{(0,1)}(z)\, I_{(z,1)}(x)\, dx$$

$$= I_{(0,1)}(z) \int_{-\infty}^{\infty} \frac{1}{|x|} I_{(z,1)}(x)\, dx$$

$$= I_{(0,1)}(z) \int_{z}^{1} \frac{1}{x}\, dx$$

$$\therefore \quad \boxed{f_Z(z) = -\log z\, I_{(0,1)}(z)}$$

i.e. $\boxed{f_Z(z) = -\log z, \quad z \in (0,1)}$

$f_Z(z) = 0$, otherwise.

Section 5: Further Transformation Techniques

Suppose now we have the following tranformation.

$$Y = u(X_1, X_2).$$

How are we going to get the pmf or pdf in this case.

Our key idea is the following. Suppose we keep x_2 fixed and vary x_1, then y becomes a function of x_1 Suppose under this fixed x_2, the function $u(x_1, x_2)$ is increasing or decreasing and thus has an inverse, i.e. $x_1 = w_1(y, x_2)$. So be written as a function of y, ie.

using Theorem 9.1 we write

$$g_Y(y, x_2) = f_{x_1, x_2}(x_1, x_2) \left| \frac{\partial x_1}{\partial y} \right|$$

or we may have that $u(x_1, x_2)$ satisfies the requirements of Theorem 9.1 if x_1 is kept fixed and thus we will have

$$g_Y(x_1, y) = f_{x_1, x_2}(x_1, x_2) \left| \frac{\partial x_2}{\partial y} \right|.$$

Whatever be the case, ~~we can from~~ $g_Y(x_1, y)$ or $g_Y(y, x_2)$ can be integrated out with respect to x_1 or x_2, to get the marginal distribution of Y.

Let us see an example

Let x_1 and x_2 be two random variables which are independent and each follow uniform distribution in $(0,1)$. Let us set $Y = X_1 + X_2$. Compute the pdf of Y. Here Y varies from 0 to 2 as X_1 and X_2 vary from 0 to 1.

If x_2 is kept fixed we have

$$\therefore \quad x_1 = y - x_2.$$

$$\therefore \quad g_Y(y, x_2) = I_{(0,1)}(y - x_2)\, I_{(0,1)}(x_2)\, |1|$$

$$= I_{(0,1)}(y - x_2)\, I_{(0,1)}(x_2)$$

$$\therefore \quad g_Y(y) = \int_{-\infty}^{\infty} g_Y(y, x_2)\, dx_2$$

$$= \int_{-\infty}^{\infty} I_{(0,1)}(y - x_2)\, I_{(0,1)}(x_2)\, dx_2$$

When $y - x_2 \in (0,1)$, then $0 < y - x_2 < 1$

$\therefore \quad x_2 < y$ & $y < x_2 + 1$. ~~Now if we set~~ Also

~~Set $y \in (0,1)$, then~~ if $y > 1$, & $y < 2$, then $y - x_2 < 1$

$\Rightarrow y - 1 < x_2 < 1$

Now observe that

$$I_{(0,1)}(y - x_2)\, I_{(0,1)}(x_2)$$

$$= I_{(0,y)}(x_2)\, I_{(0,1)}(y) + I_{(y-1,1)}(x_2)\, I_{(1,2)}(y)$$

$$\therefore \quad g_Y(y) = \int_{-\infty}^{\infty} \left[I_{(0,y)}(x_2)\, I_{(0,1)}(y) + I_{(y-1,1)}(x_2)\, I_{[1,2)}(y)\, dx_2 \right]$$

$$= I_{(0,1)}(y) \int_0^y I_{(0,y)}(x_2)\, dx_2 + I_{[1,2)}(y) \int_{y-1}^{1} I_{(y-1,1)}(x_2)\, dx_2$$

$$= y\, I_{(0,1)}(y) + I_{[1,2)}(y)(2 - y)$$

$$\therefore \quad g_Y(y) = \begin{cases} y & \text{if} \quad 0 < y < 1 \\ 2 - y & \text{if} \quad 1 \le y < 2 \end{cases}$$

& $g_Y(y) = 0$, elsewhere.

Now what happens if we have

$$Y_1 = u_1(X_1, X_2)$$

$$Y_2 = u_2(X_1, X_2)$$

Then if we know the joint distribution of X_1 & X_2 can we find the joint distribution of Y_1 & Y_2. The following theorem tells us how to do it.

Theorem 9.2:

Let X_1 and X_2 are continuous random variables, with joint pdf $f_{X_1,X_2}(x_1, x_2)$. If the transformations $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ are bijective mappings over the set

$$\left\{ (x_1, x_2) \in \text{Range}(x_1) \times \text{Range}(x_2) : f_{X_1,X_2}(x_1, x_2) \neq 0 \right\}.$$

Then over the above set we can uniquely write

$$x_1 = w_1(y_1, y_2)$$

$$x_2 = w_2(y_1, y_2)$$

Then the joint distribution $Y_1 = u_1(x_1, x_2)$ & $Y_2 = u_2(x_1, x_2)$ is given as

$$g_{Y_1,Y_2}(y_1, y_2) = f\left[w_1(y_1, y_2), w_2(y_1, y_2) \right] |\det J|$$

where J is the Jacobian matrix of the map

~~$\phi(x_1, x_2) = \begin{pmatrix} w_1(y_1, y_2) \\ w_2(y_1, y_2) \end{pmatrix}$~~

$$\phi(y_1, y_2) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} w_1(y_1, y_2) \\ w_2(y_1, y_2) \end{pmatrix}$$

given as $J = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{bmatrix}$.

Jacobian matrix is the derivative of the map $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$

Let us provide some interesting examples.

Eg (A): Let X_1 & X_2 are independent standard normal variables.

Let $Y_1 = X_1 + X_2$ & $Y_2 = \frac{X_1}{X_2}$

Find $g_{Y_1, Y_2}(y_1, y_2)$.

Here $y_1 = u_1(x_1, x_2) = x_1 + x_2$.

& $y_2 = u_2(x_1, x_2) = x_1/x_2$.

Then,

$$x_1 = w_1(y_1, y_2) = \frac{y_1 y_2}{1 + y_2}$$

$$x_2 = w_2(y_1, y_2) = \frac{y_1}{1 + y_2}$$

& $$J = \begin{bmatrix} \frac{y_2}{1+y_2}, & \frac{y_1}{(1+y_2)^2} \\ \frac{1}{1+y_2}, & -\frac{y_1}{(1+y_2)^2} \end{bmatrix}$$

$$\therefore \det J = -\frac{y_1}{(1+y_2)^2} \quad \therefore |\det J| = \frac{|y_1|}{(1+y_2)^2}$$

$$\therefore g_{Y_1, Y_2}(y_1, y_2) = \frac{1}{2\pi}\left[e^{-\frac{1}{2}\left[\frac{(y_1 y_2)^2}{(1+y_2)^2} + \frac{y_1^2}{(1+y_2)^2}\right]} \right] \frac{|y_1|}{(1+y_2)^2}$$

$$\therefore g_{Y_1, Y_2}(y_1, y_2) = \frac{1}{2\pi} \frac{|y_1|}{(1+y_2)^2} \left[e^{-\frac{1}{2}\frac{\left[(1+y_2^2)\, y_1^2\right]}{\left[1+y_2\right]^2}} \right].$$

Question: Can you find the marginal distribution of Y_2.

E.g (B): Let $X_1 \sim \text{Gamma}(n_1, \lambda_1)$

$X_2 \sim \text{Gamma}(n_2, \lambda_2)$

and X_1 and X_2 are independent. Then find the joint distribution of Y_1 & Y_2 with

$$Y_1 = X_1 + X_2 \quad \& \quad Y_2 = \frac{X_1}{X_2}. \quad \therefore \blacksquare$$

Find the marginal distribution of Y_1.

Our first approach would be to compute first the joint distribution of $g_{Y_1,Y_2}(y_1, y_2)$ and then compute out the marginal,

Here

$$x_1 = w_1(y_1, y_2) = \frac{y_1 y_2}{1 + y_1 y_2}$$

$$x_2 = w_1(y_1, y_2) = \frac{y_1}{1 + y_2}$$

$$\therefore |\det J| = \frac{y_1}{(1+y_2)^2}$$ (note y_1 & y_2 are positive as $x_1 \in (0,\infty)$ & $x_2 \in (0,\infty)$)

For $y_1 \in (0,\infty)$ & $y_2 \in (0,\infty)$ we have

$$\therefore g_{Y_1,Y_2}(y_1, y_2) = \frac{y_1}{(1+y_2)^2} \frac{1}{\Gamma(n_1)} \frac{1}{\Gamma(n_2)} \lambda^{n_1+n_2} \left(\frac{y_1 y_2}{1+y_2}\right)^{n_1-1} \left(\frac{y_1}{1+y_2}\right)^{n_2-1} e^{-\lambda(x_1+x_2)}$$

$$\therefore g_{Y_1,Y_2}(y_1, y_2) = \frac{y_1}{(1+y_2)^2} \frac{1}{\Gamma(n_1)} \cdot \frac{1}{\Gamma(n_2)} \lambda^{n_1+n_2} \left(\frac{y_1 y_2}{1+y_2}\right)^{n_1-1} \left(\frac{y_1}{1+y_2}\right)^{n_2-1} e^{-\lambda y_1}$$

$$= \frac{\lambda^{n_1+n_2}}{\Gamma(n_1)\Gamma(n_2)} y_1^{n_1+n_2-1} e^{-\lambda y_1} \frac{y_2^{n_1-1}}{(1+y_2)^{n_1+n_2}}$$

Now $B(n_1, n_2) = \dfrac{\Gamma(n_1)\Gamma(n_2)}{\Gamma(n_1+n_2)} \Rightarrow B(n_1,n_2)\Gamma(n_1+n_2) = \Gamma(n_1)\Gamma(n_2)$

$$\therefore \Rightarrow g_{Y_1,Y_2}(y_1, y_2) = \underbrace{\left(\frac{\lambda^{(n_1+n_2)}}{\Gamma(n_1+n_2)} \cdot y_1^{n_1+n_2-1} e^{-\lambda y_1}\right)}_{g_{Y_1}(y_1)} \times \underbrace{\left(\frac{1}{B(n_1,n_2)} \frac{y_2^{n_1-1}}{(1+y_2)^{n_1+n_2}}\right)}_{g_{Y_2}(y_2)}$$

Thus Y_1 and Y_2 are independent random variables, with

$$\boxed{Y_1 \sim \text{Gamma}(n_1+n_2, \lambda)}$$

If $n_1 = n_2 = 1$, Compute $g_{Y_2}(y_2)$.

Section 6: The mgf technique.

When we have $Y = u(x_1, x_2)$, then, sometimes it is useful to compute the mgf of Y and see if it matches with mgf of any distribution known to us. This is truly effective when x_1 & x_2 are independent. In fact this works pretty well even if we have a larger number of independent variables.

For example if we have $x_1, x_2, \ldots x_n$ as independent random variables, then, if

$$Y = x_1 + \cdots + x_n$$

Then $m_Y(t) = \prod_{i=1}^{n} m_{x_i}(t)$. (Prove it in the assignment)

Consider again for example that $x_1, x_2 \ldots x_n$ are independent random variables with

$$x_i \sim \text{Poisson}(\lambda_i)$$

$$\therefore \quad m_{x_i}(t) = e^{\lambda_i (e^t - 1)}.$$

Let $Y = x_1 + \cdots + x_n$.

$$\therefore m_Y(t) = \prod_{i=1}^{n} m_{x_i}(t)$$

$$= \prod_{i=1}^{n} e^{\lambda_i (e^t - 1)}$$

$$= e^{(\lambda_1 + \cdots + \lambda_n)(e^t - 1)}$$

Then $Y \sim \text{Poisson}(\lambda_1 + \cdots + \lambda_n)$.

Now consider another example.

Let X_1, X_2 are standard normal variables.

Find the joint distribution of Y_1 and Y_2

where $Y_1 = X_1 + X_2$ & $Y_2 = X_2 - X_1$

$$\therefore m_{Y_1 Y_2}(t_1, t_2) = E\left[e^{t_1 Y_1 + t_2 Y_2}\right]$$

$$= E\left[e^{t_1(X_1 + X_2) + t_2(X_2 - X_1)}\right]$$

$$= E\left[e^{(t_1 - t_2)X_1 + (t_1 + t_2)X_2}\right]$$

$$= E\left[e^{(t_1 - t_2)X_1} e^{(t_1 + t_2)X_2}\right]$$

$$= E\left[e^{(t_1 - t_2)X_1}\right] E\left[e^{(t_1 + t_2)X_2}\right]$$

($\because X_1$ & X_2 are indepdt.)

$$\therefore m_{Y_1 Y_2}(t_1, t_2) = m_{X_1}(t_1 - t_2)\, m_{X_2}(t_1 + t_2)$$

$$= e^{\frac{1}{2}(t_1 - t_2)^2} e^{\frac{1}{2}(t_1 + t_2)^2}$$

$$= e^{\left(t_1^2 + t_2^2\right)}$$

$$= e^{2\frac{t_1^2}{2}} e^{2\frac{t_2^2}{2}}$$

$$= e^{t_1^2} e^{t_2^2}$$

$$= m_{Y_1}(t_1)\, m_{Y_2}(t_2)$$ (See next page)

Now $\boxed{m_{Y_1, Y_2}(t_1, t_2) = m_{Y_1}(t_1)\, m_{Y_2}(t_2)}$. We shall establish this fact. when X_1 & X_2 are independent

Let $Y_1 = u_1(x_1, x_2)$

$Y_2 = u_2(x_1, x_2)$.

$$\therefore m_{Y_1, Y_2}(t_1, t_2) = E\left[e^{t_1 Y_1 + t_2 Y_2}\right]$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1 u_1(x_1, x_2) + t_2 u_2(x_1, x_2)} f_{X_1, X_2}(x_1, x_2)\, dx_1 dx_2$$

$$m_{Y_1,Y_2}(t_1, t_2) = e^{2t_1^2/2}\, e^{2t^2/2} \longrightarrow (*)$$

Now set $t_2 = 0$

$$m_{Y_1,Y_2}(t_1, 0) = E\left[e^{t_1 Y_1}\right] = m_{Y_1}(t_1)$$

But $m_{Y_1,Y_2}(t_1, 0) = e^{2t_1^2/2}$

$$\therefore \quad m_{Y_1}(t_1) = e^{2t_1^2/2}$$

$$\Rightarrow \quad Y_1 \sim N(0, 2)$$

Similarly setting $t_1 = 0$, we have ~~$Y_2 \sim N(0,2)$~~

i.e. $m_{Y_2}(t_2) = e^{2t_2^2/2}$

$$\therefore \quad Y_2 \sim N(0, 2).$$

$$\therefore \quad m_{Y_1,Y_2}(t_1, t_2) = m_{Y_1}(t_1)\, m_{Y_2}(t_2).$$

Thus Y_1 & Y_2 are both independent and identically distributed normal random variable with mean 0 and variance 2

— x —

End of the Probability Part

Lecture 10: Random Sampling and Sampling Distribution

Sec 1: Sampling

Statistics concerns itself with the study of several aspects of a population, in which chance plays a role. The population can be human or otherwise. In statistics we are more concerned with target population in the study rather than whole population. Suppose we are looking for the sleeping patterns or sleeping hours of all adults in Kanpur in the age group 30-45. Then that particular segment is the target population. In most cases the target population is large enough so that collecting individual data is unviable and thus we need to seek a sub-population or a subset of a population called a "sample" on which we will carry out exhaustive studies/measurements. However to prevent bias creeping into the formation of a sample it is always advisable to randomize the procen of ~~hand~~ drawing a sample.

Let there be a target population then we can consider two approaches to draw a random sample from it. The first one is sampling with replacement, where the sampled members of the population are returned back, while in the second approach they are not, since if we sample with replacement from a population of 100 a sample of each 10, each time, then in ten such draws we have exhausted the population

So from a theoretical framework the theory of sampling or random sampling we consider a random sampling with replacement. Thus once a value is noted it becomes eligible again to be considered.

Consider any random variable X, (say hieght, weight, etc) which represents a characteristic of a population. Let X, follows a distribution with pmf/pdf, given as $f_X(x)$. Thus by the population we shall now mean $f_X(x)$, the pmf/pdf.

A random sample is a collection of $n \in \mathbb{N}$, identically and independently (iid) distributed random variables each having pmf/pdf $f_{X_i}(x) = f_X(x)$, $\forall i = 1, 2, \ldots n$. The random sample is represented as

$$X_1, X_2, \ldots X_n \quad (n: \text{size of the sample})$$

Thus for each i, $X_i : \Omega \longrightarrow \mathbb{R}$, a random variable. Now what is this Ω? $\omega \in \Omega$ is a collection n-individuals drawn from the population. Thus $X_i(\omega) = x_i$, is the value of characteristic of the i-th member chosen. So ω is in "some sense", the chosen physical sample. What happens is the following. Suppose $X \sim N(\mu, \sigma^2)$ but μ and σ^2 are not known though we do know that our characteristic follows normal distribution. By the vehicle of random sampling we will estimate these parameters.

Let $X_1, \ldots X_n$ be a sample of size n. The following are two important parameter or statistic associated with the random sample are the sample mean and sample variance, which are themselves, random variable given as.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

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

Why we have divided $(n-1)$ will be clear soon. It will dealt in more detail when we will study point estimation.

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} E(X_i), \quad \text{Suppose } E(X_i) = \mu$$

Since $X_1, \ldots X_n$ are iid r.v.'s we have $E(X_i) = \mu$, $\forall i$

$$\therefore E(\bar{X}) = \mu.$$

$$\text{While } Var(\bar{X}) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum Var(X_i).$$

Denoting $Var(X_i) = \sigma^2$ we have

$$\text{Var}(\bar{x}) = \sigma_{\bar{x}}^2 = \frac{n}{n^2}\sigma^2 = \frac{1}{n}\sigma^2$$

$$\therefore \text{ Std. dev of } \bar{x} = \sigma_{\bar{x}} = \sqrt{\text{Var}(\bar{x})} = \frac{\sigma}{\sqrt{n}}$$

$\frac{\sigma}{\sqrt{n}}$ is often referred to as the standard error associated with the sample

Our aim in this chapter is to know the distribution of $\bar{x}$ and s^2 depending on the density of the population from which the sample is drawn. The following result is important enough to be stated as a theorem.

Theorem 10.1: Let $x_1, \ldots, x_n$ be a random sample from a population with density $f_x(\cdot)$, Then if $n > 1$, we have

$$E[s^2] = \sigma^2$$

where σ^2 is the population variance. Further

$$\text{Var}(s^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right), \text{ for } n > 1$$

where $\mu_4 = E[X^4]$, the 4th population moment.

Proof: We shall only compute the mean. We know that

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

We shall first show that

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

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

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

$$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2. \quad \left(\because \sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0\right)$$

$$E[S^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} E\left[(x_i-\mu)^2\right] - \frac{n}{n-1} E\left[(\bar{x}-\mu)^2\right]$$

$$= \frac{n\sigma^2}{n-1} - \frac{n}{n-1}\,\mathrm{Var}(\bar{x})$$

$$= \frac{n\sigma^2}{n-1} - \frac{n}{n-1}\,\frac{\sigma^2}{n}$$

$$= \frac{n\sigma^2}{n-1} - \frac{\sigma^2}{n-1}$$

$$= \frac{(n-1)\sigma^2}{(n-1)} = \sigma^2$$

$$\therefore \boxed{E[S^2] = \sigma^2}$$

So the sample mean and sample variance acts as a kind of estimator of the ~~aria~~ population mean and variance. However to compute $E(\bar{x})$ & $E(S^2)$, we need to know the distribution of $\bar{x}$ and S^2 when, the population distribution. ~~The probability distributions~~ is known. The probability distribution of $\bar{x}$ and S^2 is known as sampling distributions. This is what we discuss next.

Lec 2: Sampling Distributions

To begin with we will consider a population which obeys the normal distribution. Then we will seek in such a case the distribution of $\bar{x}$ and S^2.

Theorem 10.2: Let $X_1 \ldots X_n$ be a random sample of n iid random variables, with mean $\bar{X}$. Let each $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2 \ldots n$

Then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Proof: We will use the mgf technique

$$m_{\bar{X}}(t) = E\left[e^{t\bar{X}}\right]$$

$$= E\left[e^{t\frac{\sum_{i=1}^{n} X_i}{n}}\right]$$

$$= E\left[e^{\sum_{i=1}^{n} \frac{tX_i}{n}}\right]$$

$$= E\left[\prod_{i=1}^{n} e^{\frac{tX_i}{n}}\right]$$

$$= \prod_{i=1}^{n} E\left[e^{\frac{tX_i}{n}}\right], \text{ by independence}$$

$$= \prod_{i=1}^{n} m_{X_i}\left(\frac{t}{n}\right)$$

$$= \prod_{i=1}^{n} e^{\left(\frac{\mu t}{n} + \frac{1}{2}\frac{\sigma^2 t^2}{n^2}\right)}$$

$$= e^{\left(\mu t + \frac{1}{2}\sigma^2 \frac{t^2}{n}\right)}$$

$$\therefore \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$ □

Now we shall focus on how to find the distribution of

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

when we draw a random sample of size n from a normal population.

A very special type of probability distribution that will play a crucial role in computing the probability distribution of S^2 is the so called chi-square distribution a χ^2- distribution.

A random variable or rather a continuous random variable X is said to be a χ^2 random variable (chi-square r.v.) if X has the density

$$f_X(x) = \frac{1}{\Gamma(k/2)} \left(\frac{1}{2}\right)^{k/2} x^{k/2-1} e^{-\frac{1}{2}x}, \quad x \in (0, \infty)$$

$$f_X(x) = 0, \quad \text{otherwise.}$$

This a Gamma distribution with $r = k/2$ & $\lambda = 1/2$. k is called the degrees of freedom of the χ^2 - random variable X.

So using our information about the Gamma distribution we can write. If $X \sim \chi^2(k)$, (i.e χ^2 with k-degrees of freedom)

$$E(X) = \frac{r}{\lambda} = \frac{k/2}{1/2} = k$$

$$Var(X) = \frac{r}{\lambda^2} = \frac{(k/2)}{(1/2)^2} = 2k$$

$$m_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^r, \quad t < \lambda$$

$$= \left(\frac{\frac{1}{2}}{\frac{1}{2} - t}\right)^{k/2}; \quad t < \frac{1}{2}, \text{ for}$$

$$= \left[\frac{1}{1-2t}\right]^{k/2}; \quad t < 1/2$$

☞ We now state the following important Theorem

Theorem 10.3 : Let X_i, $i=1,2,\ldots k$ are continuous random variables and $X_i \sim N(\mu_i, \sigma_i^2)$ and are independent. Then

$$U = \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$$

follows χ^2 distribution with k-degrees of freedom.

Proof: We will again use the mgf technique which validates its powers. Let us write

$$Z_i = \left(\frac{X_i - \mu_i}{\sigma_i}\right)$$

then $Z_i \sim N(0,1)$, i.e. $Z_1 \ldots Z_k$ are iid r.v.s

$$\therefore \quad U = \sum_{i=1}^{k} Z_i^2 .$$

Hence

$$\begin{aligned} m_U(t) &= E\left[e^{tU}\right], \\ &= E\left[e^{t\sum_{i=1}^{k} Z_i^2}\right] \\ &= E\left[\prod_{i=1}^{k} e^{tZ_i^2}\right] \\ &= \prod_{i=1}^{k} E\left[e^{tZ_i^2}\right] \quad \text{(by independence)} \end{aligned}$$

Now

$$\begin{aligned} E\left[e^{tZ_i^2}\right] &= \int_{-\infty}^{\infty} e^{tZ_i^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)z^2} dz \end{aligned}$$

Set $\left(\sqrt{1-2t}\right)z = v \Rightarrow v^2 = (1-2t)z^2$, for $t < \frac{1}{2}$

& $dv = \sqrt{1-2t}\, dz \Rightarrow dz = \frac{dv}{\sqrt{1-2t}}$.

$$\therefore \quad E\left[e^{tz_i^2}\right] = \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\, dv$$

$$= \frac{1}{\sqrt{1-2t}}$$

$$\therefore \quad m_U(t) = \prod_{i=1}^{k} E\left[e^{tz_i^2}\right] = \left(\frac{1}{1-2t}\right)^{k/2}, \quad t < 1/2.$$

Thus U follows χ^2 with k-degrees of freedom.

Thus if $X_1 \ldots X_n$ is a random sample of size n, where for each i, $X_i \sim N(\mu, \sigma^2)$, then

$$U = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma}\right)^2 \longrightarrow (A)$$

follows a χ^2 distribution with n-degrees of freedom.

Question !! In the expression of U as given in (A). If either μ or σ^2 or both are unknown, will you call U a statistic?

We shall now state the following Theorem without proof

Theorem 10.4 Let $X_1, \ldots X_n$, be a random sample drawn from a standard normal distribution.

i) Then $\bar{X} \sim N\left(0, \frac{1}{n}\right)$

ii) $\bar{X}$ and $\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$ are independent

iii) $\sum_{i=1}^{n} (x_i - \bar{x})^2$ has a χ^2-distribution with $(n-1)$ degrees of freedom.

Proof: Part i) is already proved in Theorem 10.2. We shall just prove part iii) here as needed for our purpose

To prove iii) observe that

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n\bar{x}^2 \quad \text{(The reader should prove it)}$$

$$\therefore \quad m_{\sum_{i=1}^{n} x_i^2}(t) = m_{\sum_{i=1}^{n}(x_i-\bar{x})^2}(t)\; m_{n\bar{x}^2}(t) \quad \text{(Again the reader should prove it)}$$

$$\therefore \quad m_{\sum (x_i-\bar{x})^2}(t) = \frac{m_{\sum x_i^2}(t)}{m_{n\bar{x}^2}(t)}$$

Now as $x_i \sim N(0,1)$, by using Theorem 10.3 we know that $\sum x_i^2 \sim \chi^2$ with n degrees of freedom

$$\therefore \quad m_{\sum x_i^2}(t) = \left[\frac{1}{(1-2t)}\right]^{n/2}; \quad t < 1/2$$

Since $\bar{x} \sim N(0, \frac{1}{n})$; thus $Y = n\bar{x}^2$ follows a χ^2 distribution with degrees of freedom 1 as $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ (The student/reader should be able to prove this)

$$\therefore \quad m_{n\bar{x}^2}(t) = \left[\frac{1}{(1-2t)}\right]^{1/2}; \quad t < 1/2$$

$$\therefore \quad m_{\sum (x_i-\bar{x})^2}(t) = \left(\frac{1}{1-2t}\right)^{\frac{n-1}{2}}; \quad t < \tfrac{1}{2}$$

$$\therefore \quad \sum (x_i - \bar{x})^2 \sim \chi^2 \text{ distribution with } (n-1) \text{ degrees of freedom.} \; \square$$

So how do we write Theorem 10.4, if we write have $x_1, \ldots, x_n$ a random sample ~~fr~~ from a normal distribution with mean μ and σ^2, ♦

Note that we just set $Z_i = \dfrac{X_i - \mu}{\sigma}$

$$\therefore \quad \bar{Z} = \frac{(\bar{X} - \mu)}{\sigma}$$

and hence $\dfrac{\bar{X} - \mu}{\sigma} \sim N\left(0, \frac{1}{n}\right)$, i.e what we get from ~~(10.4) 4.1) Theorem~~ i) of Theorem ~~2.4.1~~ 10.4

Now $$\sum_{i=1}^{n} \left(Z_i - \bar{Z}\right)^2 = \sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)^2}{\sigma^2}$$

$$\therefore \quad \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2 \text{ with } (n-1) \text{ degrees of}$$

freedom from iii)' in Theorem 10.4

Now $$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

Hence $$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)^2}{\sigma^2}$$

Thus $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2$ with $(n-1)$ degrees of freedom. So how do we find the p.d.f of S^2

Set $$U = \frac{(n-1)S^2}{\sigma^2}$$

We have seen that $U \sim \chi^2$ with $(n-1)$ degrees of freedom. Set $S^2 = Y$.

$$\therefore U = \frac{(n-1)\,Y}{\sigma^2} \quad \& \quad Y = \frac{\sigma^2 U}{(n-1)}.$$

From standard transformation technique, (since the transformation is linear), we have for $n>1$ & $y>0$ $(\because U>0)$

$$f_{S^2}(y) = f_U\left(\frac{(n-1)y}{\sigma^2}\right)\left|\frac{dU}{dy}\right|$$

$$= \frac{1}{\Gamma\left(\frac{n-1}{2}\right)}\left(\frac{1}{2}\right)^{\frac{n-1}{2}}\left[\frac{(n-1)y}{\sigma^2}\right]^{\frac{n-1}{2}-1} e^{-\frac{1}{2}\frac{(n-1)y}{\sigma^2}} \frac{(n-1)}{\sigma^2}$$

$$= \frac{1}{\Gamma\left(\frac{n-1}{2}\right)}\left(\frac{1}{2}\right)^{\frac{n-1}{2}} \frac{(n-1)^{\frac{n-3}{2}}}{\sigma^{(n-3)}} \cdot y^{\frac{n-3}{2}} e^{\frac{(n-1)y}{2\sigma^2}} \cdot \frac{(n-1)}{\sigma^2}$$

$$= \frac{1}{\Gamma\left(\frac{n-1}{2}\right)}\left[\frac{(n-1)}{(2\sigma^2)}\right]^{\frac{(n-1)}{2}} y^{\frac{n-3}{2}} e^{\frac{(n-1)y}{2\sigma^2}}$$

$$\therefore \boxed{f_{S^2}(y) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)}\left(\frac{n-1}{2\sigma^2}\right)^{\frac{n-1}{2}} y^{\frac{n-3}{2}} e^{\frac{(n-1)y}{2\sigma^2}}} \quad \text{for } y>0,\ n>1$$

of course $f_{S^2}(y)=0$, for $y \le 0$.

Our next object of focus is the F-distribution named after the celebrated statistician Ronald Fisher. This distributions plays quite an important role in statistics. Let us see how an F-distribution pdf is constructed. Let $U \sim \chi^2$ with m degrees of freedom and $V \sim \chi^2$ with n degrees of freedom. Let U and V be independent, then their joint density is given as

$$f_{U,V}(u,v) = \frac{1}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)2^{\frac{(m+n)}{2}}} u^{\frac{m-2}{2}} v^{\frac{n-2}{2}} e^{-\frac{1}{2}(u+v)},$$

where ~~$0 \le u \le \infty$, $0 \le v \le \infty$~~. $0<u<\infty$, $0<v<\infty$

Let us consider transform

$$X = \frac{\frac{U}{m}}{\frac{V}{n}}$$

We shall call such a random variable X a F-variable.

Consider the transformation

$$X = \frac{\frac{U}{m}}{\frac{V}{n}} \quad \& \quad Y = V$$

We seek to compute $f_{X,Y}(x,y)$, since we know $f_{U,V}(u,v)$. Now

Jacobian $\quad U = \frac{m}{n} XY$

$$V = Y$$

$$\therefore \text{ Jacobian matrix } J = \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{bmatrix} = \begin{bmatrix} \frac{m}{n} y & \frac{m}{n} x \\ 0 & 1 \end{bmatrix}$$

$$\therefore \det J = \frac{m}{n} y \Rightarrow |\det J| = \frac{m}{n} y \quad (\text{as } m>0, n>0, y>0)$$

$$\therefore f_{X,Y}(x,y) = \frac{m}{n} y \cdot \frac{1}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right) 2^{\frac{m+n}{2}}} \left(\frac{m}{n} xy\right)^{\frac{m-2}{2}} y^{\frac{(n-2)}{2}} e^{-\left[\frac{m}{n} xy + y\right]/2}$$

where, $0 < x < \infty, 0 < y < \infty$

$$\therefore f_X(x) = \int_0^\infty f_{X,Y}(x,y)\,dy$$

$$= \frac{1}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right) 2^{\frac{m+n}{2}}} \left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{(m-2)}{2}} \int_0^\infty y^{\frac{m+n-2}{2}} e^{-\frac{1}{2}\left[\frac{m}{n} x + 1\right] y}\,dy$$

Set $\frac{1}{2}\left[\frac{m}{n} x + 1\right] y = t \Rightarrow \frac{dt}{\frac{1}{2}\left[\frac{m}{n} x + 1\right]} = dy$

$$f_X(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \left(\frac{m}{n}\right)^{m/2} \frac{x^{\frac{(m-2)}{2}}}{\left[\frac{m}{n}x+1\right]^{\frac{(m+n)}{2}}} \cdot \frac{1}{B\left(\frac{m}{2},\frac{n}{2}\right)}$$

Note that

$$\int_0^\infty y^{\frac{m+n-2}{2}} e^{-t}\, dt = \Gamma\left(\frac{m+n}{2}\right)$$

↓ (after substitution)

$\therefore$

$$\boxed{f_X(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{m/2} \frac{x^{(m-2)/2}}{\left[\frac{m}{n}x+1\right]^{\frac{(m+n)}{2}}}} \quad x \in (0,\infty)$$

↓

F-distribution (We say $X \sim F$ with m & n degrees of freedom)

F-variable $X = \frac{U}{m} \Big/ \frac{V}{n}$ is often called variance ratio.

But how is this discussion relevant to sampling distributions. This is how it is.

Let $x_1 \ldots x_{m+1}$ be a sample of size $(m+1)$ from a normal distribution with mean μ_1 & variance σ^2 & $Y_1 \ldots Y_{n+1}$ is a sample of size $(n+1)$ drawn from a normal distribution with mean μ_2 and variance σ^2, then we know that

$$\frac{1}{\sigma^2}\sum_{i=1}^{m+1}(x_i-\bar{x})^2 \sim \chi^2 \text{ with } m \text{ degrees of freedom}$$

and

$$\frac{1}{\sigma^2}\sum_{i=1}^{n+1}(Y_i-\bar{Y})^2 \sim \chi^2 \text{ with } n \text{ degrees of freedom}$$

~~The statistic~~

Consider the statistic $T = \dfrac{\frac{1}{\sigma^2}\sum_{i=1}^{m+1}(x_i-\bar{x})^2 \Big/ m}{\frac{1}{\sigma^2}\sum_{i=1}^{n+1}(x_i-\bar{x})^2 \Big/ n}$

$$\therefore \quad T = \frac{\frac{1}{m}\sum_{i=1}^{m+1}(x_i-\bar{x})^2}{\frac{1}{n}\sum_{i=1}^{n+1}(x_i-\bar{x})^2}$$

Then $T \sim F$ with $m \& n$ degrees of freedom.

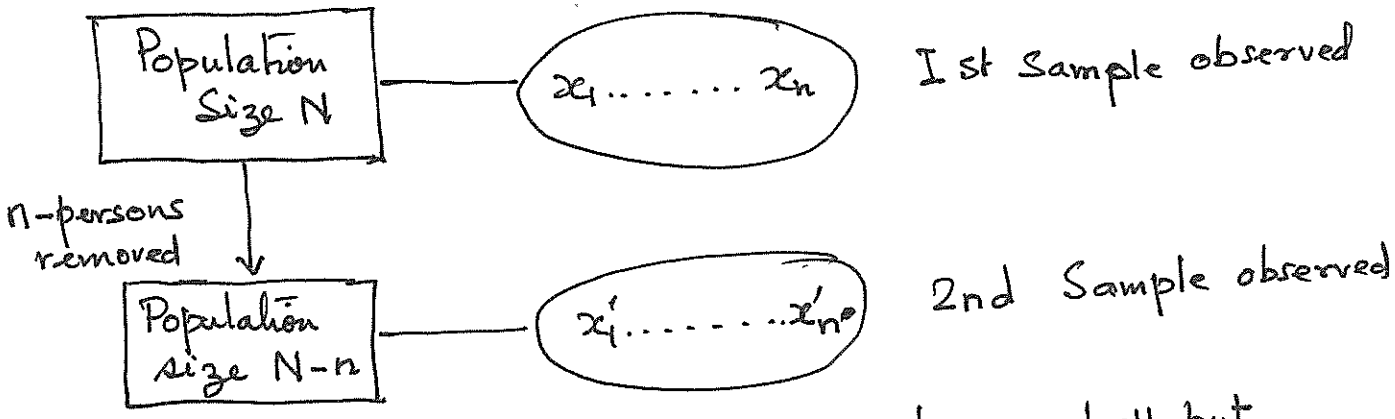
— x —

Note: Though the material here is standard, a large part of the discussion here is based on Chapter 6 of Mood, Graybill and Boes, "Introduction to the Theory of ~~Statistics~~ Statistics" 1974, Mc. Graw Hill.

Section 3: Sampling Without Replacement

Till now we have studied sampling with replacement, ie a value which appeared in the first sample can also appear in the second, third or any other sample. However when we ~~don't~~ do, so at every step we reduce the population size and thus bring in dependency among sample observation

Diagramatically



Think of the urn-model, where you draw a ball but do not return to the urn

Given a population of size N, a sample where each member can appear only once ~~in a si~~ and if we consider $n < N$ members in a sample, then there are $\binom{N}{n}$ possible samples that can be drawn from such a population. What would be the standard error or $\sqrt{Var(\bar{X})}$ in this case. So we have to compute this quantity. Note that in such a scenario we can always ask the questions, that what is the probability that the k-th member of the population is the j-th member of the sample given that k-th member of the population will is i-th member of the sample. Note that

$$Var(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(X_i, X_j)$$

So if we consider sampling without replacement, then

$$\mathrm{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}, \quad \text{if } i \neq j$$

Now observe that

$$\mathrm{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j).$$

~~Now considering that we draw the sample from a discrete distribution~~

$$E(X_i, X_j) = \sum_{k=1}^{m}\sum_{l=1}^{m} \xi_l \xi_k \, P(X_i = \xi_k \,\&\, X_j = \xi_l)$$

$$= \sum_{k=1}^{m}\sum_{l=1}^{m} \xi_l \xi_k \, P(X_i = \xi_k) P(X_j = \xi_l | X_i = \xi_k)$$

$$= \sum_{k=1}^{m} \xi_k P(X_i = \xi_k) \sum_{l=1}^{n} \xi_l P(X_j = \xi_l | X_i = \xi_k$$

Now we have used the summation formula because we have a sample population of size N and hence, the sample observations are values of a random variable from a discrete distribution. Note this idea of sampling without replacement makes sense only in that case.

$$P(X_j = \xi_l \mid X_i = \xi_k) = ??$$

How to compute this. We have to now make clear certain issues, before we proceed. Let us assume that values of the characteristic of population members are m-values given as $\xi_1 \ldots \xi_m$. We will assume that n_j of the N population members take the value ξ_j (Think of how you construct the histogram)

$$\therefore \quad P(X_i = \xi_j) = \frac{n_j}{N}$$

Now in our case thus $\mathbb{P}(X_i = \xi_k) = \frac{n_k}{N}$

Further,

$$\mathbb{P}(X_j = \xi_\ell \mid X_i = \xi_k) = \frac{n_\ell}{N-1} \quad \text{if } k \neq \ell$$

$$\mathbb{P}(X_j = \xi_\ell \mid X_i = \xi_k) = \frac{n_\ell - 1}{N-1} \quad \text{if } k = \ell.$$

$$\therefore \sum_{i=1}^{m} \xi_\ell \, \mathbb{P}(X_j = \xi_\ell \mid X_i = \xi_k)$$

$$= \sum_{\ell \neq k} \xi_\ell \frac{n_\ell}{N-1} + \xi_k \frac{n_k - 1}{N-1}$$

$$\therefore E(X_i, X_j) = \sum_{k=1}^{m} \xi_k \frac{n_k}{N} \left(\sum_{\ell \neq k} \xi_\ell \frac{n_\ell}{N-1} + \xi_k \frac{n_k - 1}{N-1} \right)$$

$$= \sum_{k=1}^{m} \xi_k \frac{n_k}{N} \left(\sum_{\ell=1}^{m} \xi_\ell \frac{n_\ell}{N-1} - \frac{\xi_k}{N-1} \right)$$

To proceed further we need more information. But for now we have

$$E(X_i, X_j) = \frac{1}{N(N-1)} \left(\sum_{k=1}^{m} \sum_{j=1}^{\ell} \xi_k \xi_\ell \, n_\ell n_k - \sum_{k=1}^{m} \xi_k^2 \eta_k \right)$$

Observe that if we look at the sum of all n-observed values of the sample and call it τ. we have

$$\tau = X_1 + X_2 + \cdots + X_n$$

So the observed value

$$\hat{\tau} = x_1 + x_2 + \cdots + x_n$$

$$\hat{\tau} = \sum_{i=1}^{m} n_j \xi_j$$

Hence

$$\hat{c}^2 = \left(\sum_{j=1}^{m} \eta_j \xi_j\right)^2$$

$$= \sum_{k=1}^{m}\sum_{l=1}^{m} \xi_k \xi_l \eta_k \eta_l$$

Also observe that the ~~sample~~ population variance is calculated as

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

$\therefore$ The observed value of σ^2 which we denote as $\hat{\sigma}^2$ is given as

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \hat{\mu}^2$$

where $\hat{\mu} = \frac{\sum x_i}{N}$, is the observed population mean

~~$\hat{\sigma}^2 = \frac{1}{N}\sum_{k=1}^{m} (\xi_k^2 \eta_k)$~~

$$\hat{\sigma}^2 = \frac{1}{N}\left(\sum_{k=1}^{m} \eta_k \xi_k^2\right) * - \hat{\mu}^2$$

$$\Rightarrow \quad \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{N}\left(\sum_{k=1}^{m} n_k \xi_k^2\right)$$

Thus

$$E(X_i X_j) = \frac{\hat{c}^2}{N(N-1)} - \frac{1}{N(N-1)}\left(\sum_{k=1}^{m} n_k \xi_k^2\right)$$

$$= \frac{\hat{c}^2}{N(N-1)} - \frac{(\hat{\sigma}^2 + \hat{\mu}^2)}{N-1}$$

$$= \frac{N^2\hat{\mu}^2}{N(N-1)} - \frac{(\hat{\sigma}^2 + \hat{\mu}^2)}{N-1}$$

$$E(x_i x_j) = \frac{N\hat{\mu}^2}{N-1} - \frac{(\hat{\sigma}^2 + \hat{\mu}^2)}{N-1}$$

$$= \frac{-\hat{\sigma}^2 + (N-1)\hat{\mu}^2}{(N-1)}$$

$$= -\frac{\hat{\sigma}^2}{N-1} + \hat{\mu}^2$$

$$\therefore \quad Cov(x_i, x_j) = E(x_i x_j) - E(x_i)E(x_j)$$

$$= -\frac{\hat{\sigma}^2}{N-1} + \hat{\mu}^2 - \hat{\mu}^2$$

$$= -\frac{\hat{\sigma}^2}{N-1}$$

$$\boxed{Cov(x_i, x_j) = -\frac{\hat{\sigma}^2}{N-1}} \longrightarrow \hat{\sigma} = \text{observed population mean.}$$

Now we have to compute $Var(\bar{x})$, and thus

$$Var(\bar{x}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(x_i, x_j)$$

$$\rightarrow \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n}$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} Var(x_i) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{i \neq j} Cov(x_i, x_j)$$

$$= \frac{1}{n^2} \cdot n\hat{\sigma}^2 + \frac{1}{n^2} n(n-1) \left[-\frac{\hat{\sigma}^2}{N-1} \right]$$

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

$$\boxed{Var(\bar{x}) = \frac{\hat{\sigma}^2}{n} \left(1 - \frac{n-1}{N-1} \right)}$$

The term

$$\left[1 - \frac{n-1}{N-1}\right]$$

is called the finite population correction.

The above discussion is largely based on Chapter 7 of the Mathematical Statistics and Data Analysis. by John. A. Rice (Cengage Learning: 2007).

Lecture 11: Parametric Estimation

In this lecture we are going to devise methods to find estimators of population parameters satisfying certain criterion. A population parameter is estimated using what is called a sample statistic. A sample statistic is a function of a random sample. Let us consider a population $f_X(\cdot;\theta)$, where θ denotes the population parameter. If θ is known then, the distribution of X is completely known. The idea of estimation comes in when θ is not known. In fact θ can be a vector or a scalar. If $X \sim$ Poisson (λ), then $\theta = \lambda$, a scalar, where as if $X \sim N(\mu, \sigma^2)$, then $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ with $\theta_1 = \mu$ & $\theta_2 = \sigma^2$.

For the moment assume $\theta \in \mathbb{R}$. Then having a random sample $x_1 \ldots x_n$ of size n from $f_X(\cdot, \theta)$, we construct a statistic symbolically given as

$$T = l(x_1, x_2 \ldots x_n)$$

We say that the random variable T is an unbiased estimator of θ if

$$E(T) = \theta.$$

Now the expectation of T is calculated based on the sampling distribution of T.

The expression $E(T) - \theta$ is called the bias in the estimation if $E(T) - \theta \neq 0$.

If $\theta \in \mathbb{R}^k$ say i.e $\theta = (\theta_1 \ldots \theta_k)$, then we can have k different statistics, $T_i = l_i(x_1 \ldots x_n)$, $i = 1 \ldots k$ which can acts as estimator of θ.

Note that $\bar{X}$, the sample mean & S^2 the sample variance have been proved in the last chapter as unbiased estimators of population mean and population variance respectively.

Such an approach to get an approximation for θ is often called point ~~estimate~~ estimation. The statistic T when viewed as a random variable is called an <u>estimator</u> and the specific value of T is called an <u>estimate</u>. Sometimes $\hat{\theta}$ is denoted as an estimate of θ. There are several approaches to find an estimator. We would also sometimes refer $\hat{\theta}$ as an estimator of θ corresponding to the estimate $\hat{\theta}$.

Section 1: Method of Moments

Let $X_1 \dots X_n$ be a random sample of size n drawn from a population $f_X(x; \theta_1 \dots \theta_k)$, where $\theta \in \mathbb{R}^k$, with $\theta = (\theta_1 \dots \theta_k)$. Let us recall the r-th moment of X, i.e

$$\mu'_r = E[X^r]$$

$$\therefore \quad \mu'_r = \int x^r f_X(x, \theta_1 \dots \theta_k)\, dx$$

$\therefore$ We can write

$$\mu'_r = \mu'_r(\theta_1 \dots \theta_k)$$

The r-th sample moment

$$M'_r = \frac{X_1^r + X_2^r + \dots + X_n^r}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

To find $\theta = (\theta_1 \dots \theta_k)$, set up the k-equations, in $\theta_1 \dots \theta_k$

$$\mu'_j(\theta_1 \dots \theta_k) = M'_j, \quad j = 1 \dots k$$

Suppose $\hat{\theta}_1, \hat{\theta}_2 \dots \hat{\theta}_k$ are the solutions to the equations. If say $\hat{\theta}_1, \hat{\theta}_2, \dots \hat{\theta}_k$, is unique, then we can consider $\hat{\theta}_1, \dots \hat{\theta}_k$ to be the estimates of $\theta_1 \dots \theta_k$.

Eg 1: $X_1 \ldots X_n$ be a random sample from a exponential population

$$f_X(x,\theta) = \theta e^{-\theta x}; \quad \theta > 0, \quad x \in (0,\infty).$$

Thus we will solve one equation

$$\therefore \quad M_1' = \mu_1' = \mu_1'(\theta) = \frac{1}{\theta}$$

$$\therefore \quad \hat{\theta} = \frac{1}{M_1'} \quad \text{So if we take a particular sample } X_1 = x_1, \ldots X_n = x_n, \text{ then } \hat{\theta} = \frac{n}{x_1 + \cdots + x_n}$$

So $\hat{\theta}$ is estimated by $\frac{1}{\bar{X}} = \frac{n}{X_1 + \cdots + X_n}$

$$\therefore \quad \hat{\Theta} = l\hat{\theta}(X_1 \ldots X_n) = \frac{n}{X_1 + \cdots + X_n}$$

$\hat{\Theta}$ is the estimator.

Eg. 2: $X_1 \ldots X_n$ be a random sample of size n, drawn from a normal population with mean μ and variance σ^2

$\therefore \theta = (\mu, \sigma^2)$.

$$M_1' = \mu_1' = \mu$$

$$M_2' = \mu_2' = \sigma^2 + \mu^2$$

μ is estimated by $M_1' = \bar{X}$. ie $\bar{X}$ is the unbiased estimator of μ.

σ^2 is estimated by the statistic

$$\hat{\Theta}_2 = M_2' - \mu^2$$

$$\therefore \hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$$

Now $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\left[\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum X_i + n\bar{X}^2\right]$

$$= \frac{1}{n}\sum X_i^2 - 2\bar{X}\frac{\sum_{i=1}^{n} X_i}{n} + \frac{n\bar{X}^2}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\bar{X}^2 + \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$$

$$\therefore \quad \hat{\Theta} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$\therefore$ $\hat{\Theta}$ is the estimator of σ^2 and note that through the method of moments we do not have S^2 as the estimator of σ^2. Thus $\hat{\Theta}$ is not an un-biased estimator of σ^2. Thus method of moments need not give us unbiased estimators always.

Section 2: Maximum Likelihood Estimator

The technique of maximum likelihood function is based on using the fact that $X_1 \ldots X_n$ are iid random variables.

Let us first look at the simple case where $\theta \in \mathbb{R}$. Let us draw a sample of size n, from a population $f_X(x;\theta)$. The ~~likelihood~~ likelihood function in this case is defined as

~~$L(\theta) = f_{X_1}(x_1,\theta) \cdot f$~~

$$L(x,\theta) = L(\theta) = f_{X_1 \ldots X_n}(x_1 \ldots x_n, \theta)$$

$$L(\theta) = f_{X_1}(x_1,\theta), f_{X_2}(x_2,\theta), \ldots f_{X_n}(x_n,\theta) \rightarrow (\text{By independence})$$

$$\therefore L(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i,\theta)$$

We compute $\max_{\theta \in A} L(\theta)$, where A is the set in which θ is restricted to belong. We find that value of θ, which maximizes $L(\theta) = L(\theta; x_1 \ldots x_n)$, for a chosen sample. Then if $\hat{\theta}$ is the maximizer for a chosen sample i.e. $\hat{\theta} = \hat{\theta}(x_1 \ldots x_n)$, then $\hat{\theta}(x_1 \ldots x_n)$ is ~~or~~ called an maximum likelihood ~~estimator~~ estimate of θ for the chosen sample values, $X_1 = x_1, \ldots, X_n = x_n$

So the estimator of θ is

$$\hat{\Theta} = \hat{\theta}(X_1 \ldots X_n).$$

thus $\hat{\Theta}$ is called the maximum likelihood estimator of θ

Noting that $\ln$ (i.e. $\log_e$) is an increasing function. We know that if $L(\theta) > 0$, $\forall \theta \in A$, then if $\hat{\theta}$ is the minimizer, we have

$$L(\theta) \leq L(\hat{\theta}), \quad \forall \theta \in A$$

$$\Rightarrow \ln L(\theta) \leq \ln L(\hat{\theta}) \quad \forall \theta \in A$$

Thus $\hat{\theta}$ is also the maximizer of $\ln L(\theta)$, over A. In most cases the set A is an open interval. Thus to compute $\hat{\theta}$ we can use the equation

$$\frac{\partial}{\partial \theta} \ln L(\hat{\theta}) = 0 \longrightarrow (*)$$

Then also check the second order condition $\left.\frac{\partial^2}{\partial \theta^2} \ln \theta\right|_{\theta=\hat{\theta}} < 0$, where $\left.\frac{\partial}{\partial \theta} \ln \theta\right|_{\theta=\hat{\theta}} = 0$, $\hat{\theta}$ solves $(*)$. Such a $\hat{\theta}$ will provide an estimator or maximum likelihood estimator using the expression

$$\hat{\theta} = \hat{\theta}(x_1 \ldots x_n).$$

Now suppose $\theta \in \mathbb{R}^k$, and we draw a random sample of size n $x_1 \ldots x_n$, from a population described by $f_X(x;\theta) = f_X(x, \theta_1, \ldots \theta_k)$.

In this case we have the likelihood function as

$$\begin{aligned} L(x, \theta_1 \ldots \theta_k) &= f_{x_1 \ldots x_n}(x_1 \ldots x_n, \theta_1 \ldots \theta_k) \\ &= f_{x_1}(x_1, \theta_1 \ldots \theta_k) \cdot f_2(x_2, \theta_1 \ldots \theta_k), \ldots, f_n(x_n, \theta_1 \ldots \theta_k) \\ &= \prod_{i=1}^{n} f_{x_i}(x_i, \theta_1 \ldots \theta_k). \end{aligned}$$

$\therefore$ Similarly, as before the maximum likelihood estimation problem is given as.

$$\max_{\theta \in K} L \text{ or } L(x, \theta_1 \ldots \theta_k), \text{ where } K \subset \mathbb{R}^k$$

or

$$\max_{\theta \in K} \ln L(x, \theta_1 \ldots \theta_k), \text{ where } K \subset \mathbb{R}^k$$

~~E.g~~ Now if $\hat{\theta}$ is the maximizer, then as before we have $\theta = (\theta_1, \ldots \theta_k)$

$$\hat{\Theta}_i = \hat{\theta}_i(X_1 \ldots X_n), \quad i = 1 \ldots k$$

So there are k, maximum likelihood estimates. We shall now provide two examples. In the first case where $\theta \in \mathbb{R}$, and in the second case $\theta \in \mathbb{R}^2$.

E.g 3: Let $X_1 \ldots X_n$ be a random sample drawn from an exponential population $f_X(\cdot, \theta)$. and is given as

$$f_X(x, \theta) = \theta e^{-\theta x}, \quad x > 0, \quad \theta > 0.$$

$$\therefore L(\theta, x) = L(\theta) = \prod_{i=1}^{n} f_{X_i}(x, \theta)$$

$$= \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

$$= \prod_{i=1}^{n} (\theta)^n e^{-\theta \sum_{i=1}^{n} x_i}$$

$$\therefore \quad \ln L(\theta) = n \ln \theta + \left(-\theta \sum_{i=1}^{n} x_i\right).$$

[The maximum likelihood estimation problem (MLE)
$\max_{\theta > 0} \ln L(\theta)$ (P1)]

Thus $$\frac{\partial}{\partial \theta} \ln L(\theta) = n \cdot \frac{1}{\theta} - \sum_{i=1}^{n} x_i = 0$$

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

$$\therefore \quad \frac{1}{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X}$$

or $$\theta = \frac{1}{\bar{X}}.$$

Now $$\frac{\partial^2}{\partial \theta^2} \log L(\theta) = -\frac{n}{\theta^2} = -n\bar{X}^2 < 0$$

Thus $\theta = \frac{1}{\bar{X}}$ maximizes (strictly) problem P1).

$\therefore \hat{\Theta}_{(M)} = \frac{1}{\bar{X}}$ is the mle estimator of θ

(6) mle: Short form for "maximum likelihood estimator"

Now for the case when there are more than one parameter, i.e. $\theta \in \mathbb{R}^k$, we first have to find $\vec{\theta} = (\hat{\theta}_1 \dots \hat{\theta}_k)$ satisfying

$$\frac{\partial \ln L_n(\theta_1 \dots \theta_k)}{\partial \theta_j} = 0; \quad \text{for } j = 1 \dots k.$$

Once we find a $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2 \dots \hat{\theta}_k)$, we compute the Hessian matrix of $\nabla^2 \ln L_n(\theta_1 \dots \theta_k)$, and see if this negative definite at $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$. In the above expression we took the liberty of writing $L_n(x, \theta_1 \dots \theta_k) = L(\theta_1 \dots \theta_k)$.

When $\theta \in \mathbb{R}^2$, i.e. $\theta = (\theta_1, \theta_2)$, then we have to check the following. First find $(\hat{\theta}_1, \hat{\theta}_2)$ satisfying

$$\left.\begin{aligned} \frac{\partial}{\partial \theta_1} \ln L_n(\theta_1, \theta_2) &= 0 \\ \frac{\partial}{\partial \theta_2} \ln L_n(\theta_1, \theta_2) &= 0 \end{aligned}\right\}$$

Then the Hessian matrix of $\ln L_n(\theta_1, \theta_2)$ is given as

$$\nabla^2 \ln L_n(\theta_1, \theta_2) = \begin{bmatrix} \frac{\partial^2}{\partial \theta_1^2} \ln L_n & , & \frac{\partial^2}{\partial \theta_2 \partial \theta_1} \ln L_n \\ \\ \frac{\partial^2}{\partial \theta_1 \partial \theta_2} \ln L_n & , & \frac{\partial^2}{\partial \theta_2^2} \ln L_n \end{bmatrix}$$

To show that $\nabla^2 \ln L_n(\theta_1, \theta_2)$ is positive definite we have

$\frac{\partial^2}{\partial \theta_1^2} \ln L_n < 0$ and $\det\left[\nabla^2 \ln L_n(\theta_1, \theta_2)\right] > 0$ evaluated at $(\hat{\theta}_1, \hat{\theta}_2)$. Or find the eigen-values which both has to be negative.

Note: If we have a diagonal matrix, which is 2x2, then to check whether it is negative definite, just see if the diagonal elements are negative as they are the eigenvalues.

E.g: Let us consider a random sample from a normal population, with mean μ and variance σ^2.

$$\therefore L_L(x, \mu, \sigma^2) = \prod_{i=1}^{n} f_{X_i}(x, \mu, \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma(\sqrt{2\pi})} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

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

where $\sigma > 0$, $\infty < \mu < -\infty$

$$\therefore \ln\left(L_L(x, \mu, \sigma^2)\right) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$\therefore$ The maximum likelihood estimation problem; ie MLE problem is

$$\max_{\substack{\sigma>0 \\ -\infty<\mu<+\infty}} \ln\left(L_L(x, \mu, \sigma^2)\right) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$\therefore$ $\frac{\partial}{\partial\mu}\ln L_L(x, \mu, \sigma^2) = 0$ & $\frac{\partial}{\partial\sigma^2}\ln L_L(x, \mu, \sigma^2) = 0$

$\Downarrow$ $\Downarrow$

$$\therefore \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \quad \& \quad -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\therefore \hat{\mu} = \frac{1}{n}\sum x_i = \bar{x}$$

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

$\longrightarrow$ The MLE estimators?

To confirm that these are MLE estimators we have to establish the fact that these are maximizers. The simplest way to realize this is that $\ln L_L(x, \mu, \sigma^2)$ is a concave function of (μ, σ^2) and hence the critical point is a global maximizer. But this is beyond the scope of this course. Hence we check the usual way stated in the previous page.

The Hessian matrix $\nabla^2 \ln L(x, \mu, \sigma^2)$ evaluated at $(\hat{\mu}, \hat{\sigma}^2)$ is given as

$$\nabla^2 \ln L(x, \hat{\mu}, \hat{\sigma}^2) = \begin{bmatrix} -\frac{n}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^4}\left[\frac{n}{2} - n^2\right] \end{bmatrix}$$

Thus $-\frac{n}{\sigma^2} < 0$ & $\frac{1}{\sigma^4}\left[\frac{n}{2} - n^2\right] < 0$, hence

$$\nabla^2 \ln L(x, \hat{\mu}, \hat{\sigma}^2)$$

to be negative definite. Thus $\hat{\mu}$ & $\hat{\sigma}^2$ are the maximizers of the mle problem.

Detail computation of $\nabla^2 \ln L(x, \hat{\mu}, \hat{\sigma}^2)$:

$$\frac{\partial}{\partial \mu} \ln L(x, \mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$$

$$\therefore \frac{\partial^2}{\partial \mu^2} \ln L(x, \hat{\mu}, \hat{\sigma}^2) = -\frac{n}{\hat{\sigma}^2}$$

&

$$\frac{\partial}{\partial \sigma^2}\frac{\partial}{\partial \mu} \ln L(x, \hat{\mu}, \hat{\sigma}^2) = \frac{\partial}{\partial \sigma^2 \partial \mu} \ln L(x, \hat{\mu}, \hat{\sigma}^2)$$

$$= -\hat{\sigma}^{-4} \sum_{i=1}^{n} (x_i - \hat{\mu})$$

$$= -\hat{\sigma}^{-4} \left[\sum_{i=1}^{n} (x_i - \bar{x})\right]$$

$$= \hat{\sigma}^{-4} \left(\sum_{i=1}^{n} x_i - n\bar{x}\right) = 0$$

$$\frac{\partial \ln L(x, \mu, \sigma^2)}{\partial \sigma^2} \cdot \frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$\therefore \frac{\partial}{\partial \mu} \cdot \frac{\partial}{\partial \sigma^2} \ln L(x, \hat{\mu}, \hat{\sigma}^2) = \frac{1}{2\hat{\sigma}^2} 2 - \left(\sum_{i=1}^{n} (x_i - \hat{\mu})\right) = 0$$

$$\frac{\partial}{\partial \sigma^2}\left(\frac{\partial}{\partial \sigma^2} \ln L(x, \hat{\mu}, \hat{\sigma}^2)\right) = \frac{n}{2}\hat{\sigma}^{-4} - n^2\hat{\sigma}^{-4} = \frac{1}{\hat{\sigma}^4}\left[\frac{n}{2} - n^2\right] < 0.$$

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

We would like to note that second order condition only tells us that the ~~note~~ estimators that we have got are ~~really~~ only local minimizer albeit strict one. To establish that they are global we need to use ~~convexity~~ Concavity. The second order check also establishes the concavity of functions here.

A crash course on Concavity Let I be an interval in $\mathbb{R}$ and $f: I \to \mathbb{R}$.

The function $f: I \to \mathbb{R}$ is called <u>concave</u> if for any x, y in I & $\lambda \in [0,1]$

$$f(\lambda y + (1-\lambda)x) \geq \lambda f(y) + (1-\lambda) f(x). \longrightarrow (A)$$

Note: If x & y are in I; then for any $\lambda \in [0,1]$, $\lambda y + (1-\lambda)x \in I$

Thus if f is differentiable, then from (A), by Taylor's Theorem.

$$f(x) + \lambda f'(x)(y-x) + o(\lambda) \geq \lambda f(y) + (1-\lambda) f(x), \quad \text{for all } \lambda \in (0,1)$$

where $\frac{o(\lambda)}{\lambda} \to 0$ as $\lambda \to 0$.

$$\therefore \quad \lambda f'(x)(y-x) + o(\lambda) \geq \lambda (f(y) - f(x))$$

$$\therefore \text{ As } \lambda \to 0, \quad f'(x)(y-x) \geq f(y) - f(x)$$

If $f'(x) = 0 \Rightarrow f(x) \geq f(y) \Rightarrow x$ is a global ~~maximum~~ maximizer

Let $f: I \times I \to \mathbb{R}$, then f is concave if for any $(x_1, x_2) \in I \times I$ and $(y_1, y_2) \in I_1 \times I_2$ & $\lambda \in (0,1)$

$$f(\lambda y_1 + (1-\lambda) x_1, \ \lambda y_2 + (1-\lambda) x_2) \geq \lambda f(y_1, y_2) + (1-\lambda) f(x_1, x_2)$$

Again using Taylor's Theorem in 2-dimensions we have $\forall (x_1, x_2)$ & (y_1, y_2)

$$\frac{\partial f}{\partial x_1}(y_1 - x_1) + \frac{\partial f}{\partial x_2}(y_2 - x_2) \geq f(y_1, y_2) - f(x_1, x_2).$$

$$\therefore \quad \frac{\partial f}{\partial x_1} = 0 \ \& \ \frac{\partial f}{\partial x_2} = 0, \quad f(x_1, x_2) \geq f(y_1, y_2), \ \forall (y_1, y_2) \in I \times I$$

$\Rightarrow$ (x_1, x_2) is the global maximizer.

E.g: $-\alpha (x-\mu)^2 = f(x)$, $\alpha > 0$, $\mu \in \mathbb{R}$ is concave in x

$f(x) = -\ln x$, $x > 0$, i.e $x \in \mathbb{R}_{++}$ is concave in x. (1D)

~~If we fix x, then~~ f is concave on I if $f''(x) \leq 0$, $\forall x$ on I

f is concave on $I \times I$ if $\nabla^2 f$ is negative definite on $I \times I$

In our study in the example the log likelihood functions can be shown to be concave.

Section 3 Fisher Information & Cramer-Rao Inequality

The maximum likelihood estimation technique, is very popular, and the reason for this is follows. Given a sample observation,

$X_1 = x_1, \ldots, X_n = x_n$ the likelihood function $L(x, \theta)$

provides us the "likelihood" or the frequency of the occurence of the observation $x_1, x_2, \ldots, x_n$. Since the parameter θ is unknown Suppose we ask the question: What is the distribution for which the observations $x_1 \ldots x_n$ occur the maximum number of time? This can be done by finding the θ, which maximizes the joint pdf, since it is the parameter θ, which specifies the distribution we shall now see how the likelihood function can be used to build some important measures related to, estimators.

We shall begin with the notion of Fisher Information, but let us first write down the two basic assumptions we need. We will study only the single parameter case.

Assumptions

A) The pdf $f_X(x, \theta)$ is twice differentiable as a function of the parameter θ.

B) We can differentiate with respect to θ under the integral sign for the integral $\int f_X(x,\theta)\,dx$. In fact we will assume that we can differentiate twice under the integral sign with respect to θ.

$$\frac{\partial}{\partial \theta}\int_A f_X(x,\theta)\,dx = \int_A \frac{\partial}{\partial \theta} f_X(x,\theta)\,dx$$

C) The parameter θ belongs to an open interval.

$A \subseteq \mathbb{R}$.

The Fisher information function is built on the idea of a Fisher score function.

Let $X_1 \ldots X_n$ be a random sample from a population $f_X(x, \theta)$.

The Fischer score function for the ith observation is given as the r.v. $F(X_i, \theta)$

whose value for the observation $X = x_i$ is given as $\rightarrow$ $F(x_i, \theta) = \dfrac{\partial}{\partial \theta} \ln f_X(x_i, \theta)$

$$\therefore \quad F(x_i, \theta) = \frac{1}{f_X(x_i, \theta)} \cdot \frac{\partial}{\partial \theta} f_X(x_i, \theta)$$

$$\therefore \quad E\left[F(X_i, \theta)\right] = \int_{-\infty}^{\infty} \frac{1}{f_X(x_i, \theta)} \frac{\partial}{\partial \theta} f_X(x_i, \theta)\, f_X(x_i, \theta)\, dx$$

$$= \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f_X(x_i, \theta)\, dx$$

$$= \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f_X(x_i, \theta)\, dx = \frac{\partial}{\partial \theta}(1) = 0.$$

$$\therefore \quad E\left[F(X_i, \theta)\right] = 0$$

The Total Fisher score function for a sample of size n is given as the r.v.

~~$F_n(\theta) = \sum_{i=1}^{n}$ [illegible]~~

$$\boxed{F_n(\theta) = \sum_{i=1}^{n} F(X_i, \theta)} \qquad \left[\text{Note } E(F_n(\theta)) = 0\right]$$

The ~~fisth~~ Fisher information about the i-th observation is denoted as $I(\theta)$ and is given as

~~$I_\theta = V$~~ $\boxed{I_i(\theta) = \text{Var}\left(F(X_i, \theta)\right)}$

$$\therefore \quad I_i(\theta) = \mathrm{Var}\left(F(x_i,\theta)\right)$$

$$= E\left[\left(F(x_i,\theta)\right)^2\right] \qquad \left(\because E\left(F(x_i,\theta)\right) = 0\right)$$

$$= E\left[\left(\frac{\partial}{\partial\theta}\ln f(x_i,\theta)\right)^2\right]$$

$$= \int_{-\infty}^{\infty}\left[\frac{\partial}{\partial\theta}\ln f(x_i,\theta)\right]^2 f_X(x,\theta)\,dx$$

$$= \int_{-\infty}^{\infty}\frac{1}{f(x,\theta)}\cdot\frac{1}{f_X^2(x_i,\theta)}\left[\frac{\partial}{\partial\theta}f_X(x_i,\theta)\right]^2 f_X(x,\theta)\,dx$$

$$= \int_{-\infty}^{\infty}\frac{1}{f_X(x_i,\theta)}\left[\frac{\partial}{\partial\theta}f_X(x_i,\theta)\right]^2 dx$$

Usually in texts one writes

$$\boxed{I_i(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln f_X(x_i,\theta)\right)^2\right]}$$

For a random sample of size n, the Fisher information is given as.

Eg1: Consider a random sample drawn from of size n drawn from an exponential population, $f_X(x,\theta)$, given as

$$f_X(x,\theta) = \theta e^{-\theta x}; \quad x>0, \ \theta>0$$

Thus

$$F(x_i,\theta) = \frac{\partial}{\partial\theta}\ln\left(\theta e^{-\theta x_i}\right)$$

$$= \frac{1}{\theta e^{-\theta x_i}}\left[\frac{\partial}{\partial\theta}\left(\theta e^{-\theta x_i}\right)\right]$$

$$= \frac{1}{\theta e^{-\theta x_i}}\left[e^{-\theta x_i} + \theta e^{-\theta x_i}(-x_i)\right]$$

$$\therefore \quad \bar{f}(x_i, \theta) = \frac{1}{\theta}\left[1 - \theta x_i\right]$$

$$= \frac{(1-\theta x_i)}{\theta}$$

$$\therefore \quad \bar{f}(x_i, \theta) = \left[\frac{1-\theta x_i}{\theta}\right]$$

$$\therefore \quad I(\theta) = E\left[\left(\bar{f}(x_i, \theta)^2\right)\right]$$

$$= \int_0^\infty \frac{(1-\theta x_i)^2}{\theta^2} \cdot \theta e^{-\theta x_i} dx$$

$$= \int_0^\infty \frac{(1-\theta x_i)^2}{\theta^2} \cdot \theta e^{-\theta x_i} dx$$

$$= \frac{1}{\theta} \int_0^\infty (1-\theta x_i)^2 e^{-\theta x_i} dx$$

$$= \frac{1}{\theta} \int_0^\infty (1 - 2\theta x_i + \theta^2 x_i^2) \cdot e^{-\theta x_i} dx$$

$$= \frac{1}{\theta} \int_0^\infty e^{-\theta x_i} dx - \frac{2}{\theta} \int_0^\infty x_i \theta e^{-\theta x_i} dx + \int_0^\infty x_i^2 \theta^2 e^{-\theta x_i} dx.$$

$$= \frac{1}{\theta^2} \int_0^\infty \theta e^{-\theta x_i} dx - \frac{2}{\theta} E(x_i) + E(x_i^2)$$

$$= \frac{1}{\theta^2} - \frac{2}{\theta} \cdot \frac{1}{\theta} + \left[\mathrm{Var}(x_i) + (E(x_i))^2\right]$$

$$= \frac{1}{\theta^2} - \frac{2}{\theta^2} + \left[\frac{1}{\theta^2} + \frac{1}{\theta^2}\right]$$

$$= \frac{1}{\theta^2} - \frac{2}{\theta^2} + \left[\frac{2}{\theta^2}\right] = -\frac{1}{\theta^2} + \frac{2}{\theta^4}$$

$$= \frac{1}{\theta^2} > 0.$$

$$= \frac{2}{\theta^4} - \frac{1}{\theta^2} = \frac{1}{\theta^2}\left[\frac{2}{\theta^2} - 1\right]$$

So the Fisher information for the whole sample is given as

$$I_n(\theta) = \text{Var}\left(F_n(\theta)\right)$$

$$= \text{Var}\left(F(x_1,\theta) + \cdots + F(x_n,\theta)\right)$$

Since $x_1 \ldots x_n$ are independent, $F(x_1,\theta), \ldots, F(x_n,\theta)$ are independent. Thus

$$I_n(\theta) = \text{Var}\left(F_{\theta}(x_1,\theta)\right) + \cdots + \text{Var}\left(F(x_n,\theta)\right)$$

$$= \sum I_{\theta}(\theta) I(x_i$$

$$\therefore \quad I_n(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

But each x_i's are having the same distribution.

$$\therefore \quad \boxed{I_n(\theta) = n\, I_i(\theta)}$$

The Cramer-Rao Inequality

A random sampling is called regular if its Fischer information is continuous, strictly positive ~~and bounded~~ and bounded for all θ in the given range.

For example if we consider a random sample ~~from~~ $x_1 \ldots x_n$ from an exponential distribution, then

$$I_n(\theta) = n\, I_i(\theta)$$

$$= \frac{n}{\theta^2} > 0$$

$I_n(\theta)$ is of course continuous on $\theta > 0$. However, $I_n(\theta)$ is not bounded above, as a function of θ. But as $I_n(\theta)$ is continuous and positive we can consider the sampling from an exponential distribution is not regular.

Theorem 11.1: Cramer-Rao Bound (Actually a lower-bound)

Let $T = \hat{\theta}(X_1 \dots X_n)$ be a statistic which is an estimator of θ of the population $f(x, \theta)$. Let the bias $b_n(\theta) = E[\hat{\theta}] - \theta$ be continuously differentiable. Then

$$\mathrm{Var}(\hat{\theta}) \geq \frac{(1 + b_n'(\theta))^2}{I_n(\theta)}$$

Proof:

$$\theta + b_n(\theta) = E[\hat{\theta}] = \int_{\mathbb{R}^n} \hat{\theta}_n(x_1 \dots x_n) f_{X_1 \dots X_n}(x_1 \dots x_n \theta)\, dx_1 \dots dx_n$$

Now differentiation under the integral sign we have.

$$1 + b_n'(\theta) = \int_{\mathbb{R}^n} \hat{\theta}_n(x_1 \dots x_n) \frac{\partial f_{X_1 \dots X_n}(x_1 \dots x_n, \theta)}{\partial \theta} dx_1 \dots dx_n$$

$$= \int_{\mathbb{R}^n} \hat{\theta}(x_1 \dots x_n) \frac{f_{X_1 \dots X_n}(x_1 \dots x_n, \theta)}{f_{X_1 \dots X_n}(x_1 \dots x_n, \theta)} \cdot \frac{\partial}{\partial \theta} f_{X_1 \dots X_n}(x_1 \dots x_n, \theta)\, dx_1 \dots dx_n$$

$$= \int_{\mathbb{R}^n} \hat{\theta}(x_1 \dots x_n) F_n(\theta) f_{X_1 \dots X_n}(x_1 \dots x_n, \theta)\, dx_1 \dots dx_n$$

$$F_n(\theta) = \sum_{i=1}^{n} F(x_i, \theta) = \sum \frac{1}{f_X(x_i, \theta)} \frac{\partial}{\partial \theta} f_X(x_i, \theta)$$

$$= \sum_{i=1}^{n} \frac{1}{f_{X_i}(x_i \theta)} \frac{\partial}{\partial \theta} f_{X_1 \dots X_n}(x_1 \dots x_n, \theta)$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f_X(x_i, \theta)$$

$$= \frac{\partial}{\partial \theta} \left[\sum_{i=1}^{n} \ln f_{X_i}(x_i, \theta) \right]$$

$$= \frac{\partial}{\partial \theta} \left[\ln \prod_{i=1}^{n} f_{X_i}(x_i, \theta) \right] = \frac{\partial}{\partial \theta} \ln f_{X_1 \dots X_n}(x_1 \dots x_n \theta)$$

↓ (By independence).

$$\therefore \quad F_n(\theta) = \frac{1}{f_{X_1 X_2 \ldots X_n}(x_1 \ldots x_n, \theta)} \frac{\partial}{\partial \theta} f_{X_1, \ldots X_n}(x_1 \ldots x_n, \theta)$$

$$\therefore \quad 1 + b_n'(\theta) = E\left[\hat{\theta} F_n(\theta)\right] = \operatorname{cov}(\hat{\theta}, F_n(\theta))$$

Note that $\operatorname{Cov}(\hat{\theta}, F_n(\theta)) = E\left[\hat{\theta} F_n(\theta)\right] - E[\hat{\theta}]\, E[F_n(\theta)]$

$= E\left[\hat{\theta} F_n(\theta)\right]$ $\quad (\because E[F_n(\theta)] = 0$ as derived ~~earlier~~ earlier)

$\therefore$ Let $\rho_n^2 =$ Square of correlation coefficient of $\hat{\theta}$, ~~$F_n(\theta)$~~ and $F_n(\theta)$

$$\therefore \quad \rho_n^2 = \frac{\left[\operatorname{Cov}(\hat{\theta}, F_n(\theta))\right]^2}{\operatorname{Var}(\hat{\theta}) \operatorname{Var}(F_n(\theta))}$$

But $\rho_n^2 \leq 1$ as $\rho_n \in [-1, +1]$

$$\Rightarrow \quad \frac{\left[\operatorname{Cov}(\hat{\theta}, F_n(\theta))\right]^2}{\operatorname{Var}(\hat{\theta}) \operatorname{Var}(F_n(\theta))} \leq 1.$$

Hence

$$\operatorname{Var}(\hat{\theta}) \geq \frac{\left[\operatorname{Cov}(\hat{\theta}, F_n(\theta))\right]^2}{\operatorname{Var}(F_n(\theta))}$$

$$\Rightarrow \quad \boxed{\operatorname{Var}(\hat{\theta}) \geq \frac{(1 + b_n'(\theta))^2}{I_n(\theta)}} \quad \square$$

This is called the Cramer-Rao Lower bound. Now if $\hat{\theta}$ is an unbiased estimator, then $b_n(\theta) = 0$ for any θ in the given range. Thus we have

$$\boxed{\operatorname{Var}(\hat{\theta}) \geq \frac{1}{I_n(\theta)}}$$

This is called the Cramer-Rao Inequality. An unbiased estimator $\hat{\theta}$ is called efficient if

$$\boxed{\operatorname{Var}(\hat{\theta}) = \frac{1}{I_n(\theta)}}$$

Let us finish our discussion with an example by drawing a sample of size n, from a normal population $N(\mu, \sigma^2)$, where μ is unknown but σ^2 is ~~not~~ known.

We shall show that $\bar{X}_n = \bar{X} = \frac{X_1 + \cdots + X_n}{n}$ is indeed an efficient estimator of μ. $\bar{X}_n$ is known to be unbiased as we have already shown. Now $Var(\bar{X}_n) = \frac{\sigma^2}{n}$. We have

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

$$\therefore \ln f_{X_i}(x_i, \mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2}$$

In the random-variable form we have

$$\ln f_{X_i}(X_i, \mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(X_i-\mu)^2}{2\sigma^2}$$

$$F(X_i, \mu) = \frac{X-\mu}{\sigma^2}$$

$$\therefore I_n(\theta) = n I_i(\theta) = n\,Var(F(X_i, \mu)) = n\,E\left[(F(X_i, \mu))^2\right]$$

$$= n\,E\left[\left(\frac{(X_i-\mu)}{\sigma^2}\right)^2\right]$$

$$= n\,E\left[\frac{(X_i-\mu)^2}{\sigma^4}\right]$$

$$= \frac{n}{\sigma^4} E\left[(X_i-\mu)^2\right]$$

$$= \frac{n}{\sigma^4} \cdot Var(X_i)$$

$$= \frac{n}{\sigma^4} \cdot \sigma^2 = \frac{n}{\sigma^2}$$

Now $Var(\bar{X}_n) = \frac{\sigma^2}{n} = \frac{1}{I_n(\theta)}$, proving that $\bar{X}_n$ is indeed an efficient estimator.

The part on Fisher Information is partially based on Chapter 1, of the book titled: Mathematical Statistics: Asymptotic minimax theory. by A. Korostelev and O. Korosteleva, AMS-2011.

If we now consider $\hat{\sigma}^2 = \hat{\theta}(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ as the maximum likelihood estimator of σ^2 in the normal distribution, then observe that

$$b_n(\hat{\sigma}^2) = -\frac{\sigma^2}{n}$$

Thus $\frac{d}{d\sigma^2} b_n(\sigma^2) = -\frac{1}{n} = b_n'(\sigma^2)$

$\therefore \left(1 + b_n'(\sigma^2)\right) = -\frac{1}{n}$. Now we shall compute

$$F(x_i, \sigma^2) = \frac{\partial}{\partial \sigma^2} \ln f_X(x_i, \sigma^2)$$

$$= \frac{\partial}{\partial \sigma^2}\left[-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{\partial}{\partial \sigma^2}\left[-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= -\frac{1}{2}\cdot\frac{\partial}{\partial \sigma^2}\ln\sigma^2 - \frac{\partial}{\partial \sigma^2}\frac{(x_i-\mu)^2}{2\sigma^2}$$

$$= -\frac{1}{2}\frac{\partial}{\partial \sigma^2}\ln\sigma^2 - \frac{1}{2}\frac{\partial}{\partial \sigma^2}\cdot\frac{(x_i-\mu)^2}{\sigma^2}$$

$$= -\frac{1}{2}\left[\frac{\partial}{\partial \sigma^2}\ln\sigma^2 + (x_i-\mu)^2\frac{\partial}{\partial \sigma^2}\left(\frac{1}{\sigma^2}\right)\right]$$

$$= -\frac{1}{2}\left[\frac{1}{\sigma^2} - \frac{(x_i-\mu)^2}{\sigma^4}\right]$$

$$= \frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2} + \frac{1}{2\sigma^4}$$

Now Fisher information for the i-th sample

$$I_i(\theta) = \text{Var}\left(F_i(x, \theta)\right)$$

$$= \text{Var}\left(\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2} + \frac{1}{2\sigma^4}\right)$$

$$= \text{Var}\left(\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)$$

$$= \frac{1}{4}\text{Var}\left[\frac{(x_i-\mu)^2}{\sigma^2}\right]$$

$$= \frac{1}{4}\text{Var}\left[\left(\frac{x_i-\mu}{\sigma}\right)^2\right]$$

$$= \frac{1}{4}\text{Var}\left(Z_i^2\right)$$

Now $Z_i^2 \sim \chi^2$ with one degree of freedom.

Now $\text{Var}(Z_i^2) = 2$

$\therefore$ $I_i(\theta) = \frac{1}{4} \cdot 2 = \frac{1}{2}$.

Thus $I_n(\theta) = \frac{n}{2}$

Hence Cramer-Rao lower-bound tells us that,

$$\boxed{\text{Var}(\hat{\sigma}^2) \geq \frac{\left(1 - \frac{1}{n}\right)^2}{\frac{n}{2}}}$$

Lecture - 12 : Interval Estimation

Section 1: Confidence Intervals

What we have learned till now is point estimation. In order to estimate the value of a population parameter, we just draw a random sample of a reasonable size and construct a statistic which can best estimate that population parameter, in some sense, either as unbiased or as the best maximum likelihood estimator. But if we are drawing a sample from a probability density function the probability that a value of the statistic will coincide with population parameter is zero. This motivates us to take the ~~value~~ ~~that~~ view that it might be better to be able to construct an interval around the statistic so that the population parameter value falls in that interval with certain probability.

Let us draw a sample from the population with pdf $f_X(x,\theta)$, and let $X_1, \ldots X_n$ be a random sample of sizes n drawn from it, and let Θ be a statistic, $\Theta = \Theta(X_1 \ldots X_n)$ constructed to estimate, θ. We may for example want to know, if

$$\Theta - c < \theta < \Theta + c$$

with probability say γ; i.e.

$$P(\Theta - c < \theta < \Theta + c) = \gamma$$

Then $(\Theta - c, \Theta + c)$ forms a confidence interval associated with θ. This is more formally defined as follows.

Definition 1: If $\hat{\theta}_1$ and $\hat{\theta}_2$ are values of the random variables $\hat{\Theta}_1, \hat{\Theta}_2$, ~~such that~~ which are themselves a ~~also~~ statistics, such that.

$P(\hat{\theta}_1 < \theta < \hat{\theta}_2)$ ~~$P(\hat{\Theta}_1 < \alpha < \hat{\Theta}_2)$~~ $= 1 - \alpha$

for some specified probability $1 - \alpha$. We shall refer to the interval $(\hat{\theta}_1, \hat{\theta}_2)$ as a $(1-\alpha)100\%$ confidence interval for the population parameter θ. The value $1 - \alpha$ is called degree of confidence & θ_1 & θ_2 the upper and lower confidence limits.

limits. So if $\alpha = 0.05$, we say the degree of confidence is 0.95 and we get a 95% confidence interval. It is also important to note at the very beginning that the confidence intervals for θ need not be unique as we will just see.

Section 2: Interval Estimation of means

In this section we will study first the interval estimation of the mean μ of a normal population whose variance σ^2 is known. Next we will see how to deal the case when σ^2 is unknown. It will be important to note how $\hat{\Theta}_1$ and $\hat{\Theta}_2$ will be very different in these two scenarios.

Case 1: σ^2 is known.

Let $X_1, \ldots, X_n$ be a random sample of size n drawn from a population $N(\mu, \sigma^2)$, where σ^2 is known but μ is not. In order to estimate the $(1-\alpha)100\%$ confidence interval of μ we proceed as follows.

Note that $E(\bar{X}) = \mu$ & $Var(\bar{X}) = \frac{\sigma^2}{n}$

where $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$.

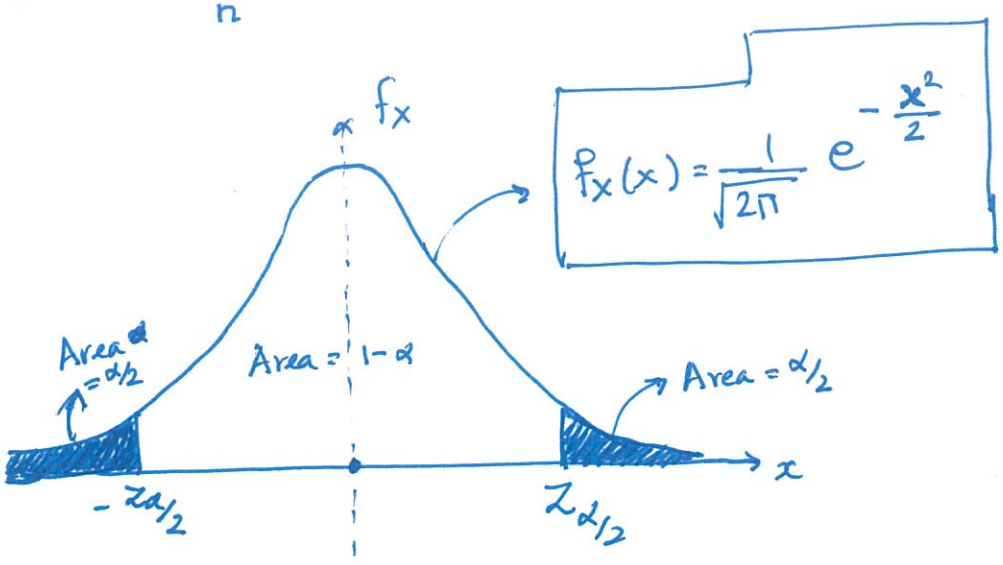


Fig 1: The point $z_{\alpha/2}$

Let us look at diagram above and it is seen how the point $Z_{\alpha/2}$ is defined; i.e.

$$\int_{Z_{\alpha/2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \alpha/2$$

$\therefore$ From diagram if $Z \sim N(0,1)$ then

$$P\left(-Z_{\alpha/2} < Z < Z_{\alpha/2}\right) = 1-\alpha \longrightarrow (A)$$

$\therefore$ Since we have drawn a normal population

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{[Check Chapter Lecture 10]}$$

$\therefore$ Set $Z = \dfrac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

Hence from (A) we have

$$P\left(-Z_{\alpha/2} < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < Z_{\alpha/2}\right) = 1-\alpha$$

$$\boxed{P\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha}$$

Here $\hat{\theta}_1 = \bar{X} - Z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ & $\hat{\theta}_2 = \bar{X} + Z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$.

Now we will again consider studying a normal population but this time we shall consider that the population variance σ^2 is unknown. In this situation we have to use sample variance S^2 in order to try to develop a statistic using it. Let us set

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Let us assume that the sample is small, i.e. $n < 30$, then it better to proceed by noting that $T \sim t$ distribution with $n-1$ degrees of freedom (see Lecture 10). To find out the $(1-\alpha)100\%$ confidence interval we must have two points q_1 & q_2, such that

$$P(q_1 < T < q_2) = 1 - \alpha$$

$$\Rightarrow P\left(q_1 < \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} < q_2\right) = 1 - \alpha$$

$$\Rightarrow P\left(\frac{S}{\sqrt{n}} q_1 < \bar{X} - \mu < \frac{S}{\sqrt{n}} q_2\right) = 1 - \alpha$$

$$\Rightarrow P\left(\bar{X} - q_2 \frac{S}{\sqrt{n}} < \mu < \bar{X} - q_1 \frac{S}{\sqrt{n}}\right) = 1 - \alpha \quad *$$

Thus in this case the confidence interval for the mean is given by

$$\left(\bar{X} - q_2 \frac{S}{\sqrt{n}} < \mu < \right)$$

$$\left(\bar{X} - q_2 \frac{S}{\sqrt{n}}, \quad \bar{X} - q_1 \frac{S}{\sqrt{n}}\right)$$

Hence the length of this interval $= (q_2 - q_1) \frac{S}{\sqrt{n}}$

or $L = (q_2 - q_1) \frac{S}{\sqrt{n}}$.

From a more practical point of view we want $(q_2 - q_1)$ to be as small as possible. However note that L is in general a random variable. However for a given sample we of size say n, we need to actually solve the optimization problem

$$\min L = (q_2 - q_1)\frac{S}{\sqrt{n}} \longrightarrow (E)$$

Subject to

$$\int_{q_1}^{q_2} f_T(q)\, dq = 1 - \alpha \longrightarrow (F).$$

where f_T is the pdf of the t-distribution with $(n-1)$ degrees of freedom. Let us remite (F) using the dummy variable t, ie.

$$\int_{q_1}^{q_2} f_T(t)\, dt = 1 - \alpha. \longrightarrow (F')$$

Since α is fixed from a theoretical point of view one can always express q_2 as a function of q_1. Let us first differentiate (F') with respect to q_1. In order to do so we need to use the Lebnitz formula for differentiating under the integral sign. We shall write it in the following box.

Lebnitz rule : For differentiating under the integral sign.

Let $a(x)$ and $b(x)$ are differentiable function. Then consider the integral

$$\int_{a(x)}^{b(x)} \varphi(x,t)\, dt$$

$$\therefore \quad \frac{d}{dx}\int_{a(x)}^{b(x)} \varphi(x,t)\, dt = \int_{a(x)}^{b(x)} \frac{\partial}{\partial x}\varphi(x,t)\, dt + \varphi(x, b(x))\frac{db}{dx} - \varphi(x, a(x))\frac{da}{dx}$$

Note that if we go back to (F'), f_T depends on t and not on q_1. Thus we will have

$$\frac{d}{dq_1} \int_{q_1}^{q_2} f_T(t)\, dt = 0$$

By Leibnitz rule we

$$\int_{q_1}^{q_2(q_1)} \frac{\partial}{\partial q_1} f_T(t)\, dt + f_T(q_2) \frac{dq_2}{dq_1} - f_T(q_1) \frac{dq_1}{dq_1} = 0$$

$\Rightarrow$ As $\frac{\partial}{\partial q_1} f_T(t) = 0$ we have from the above expression

$$f_T(q_2) \frac{dq_2}{dq_1} - f_T(q_1) = 0 \longrightarrow (G)$$

Now when we replace q_2 as a function of q_1 in L then the problem become unconstrained and to minimize L we first differentiate with respect to q_1

$$\therefore \quad \left(\frac{dq_2}{dq_1} - 1 \right) \frac{S}{\sqrt{n}} = 0$$

$$\Rightarrow \left(\frac{f_T(q_1)}{f_T(q_2)} - 1 \right) \frac{S}{\sqrt{n}} = 0$$

$$\text{or} \quad f_T(q_1) = f_T(q_2)$$

So, if q_1 is the minimizer then $f_T(q_1) = f_T(q_2)$.

Of course one solution is $q_1 = q_2$, but that will give $1 - \alpha = 0$ from (F'), which is a contradiction.

But as the graph of t distribution is symmetric we have $q_1 = -q_2$ as the solution. So from the statistical tables we have to find q_2 such that

$$P(T > q_2) = \frac{\alpha}{2}, \quad \text{where } T \sim t \ (n-1 \text{ degrees of freedom})$$

In general one writes $q_2 = t_{\alpha/2, n-1}$.

What happens if we have a large sample. Let us draw a sample from a normal distribution. Then using what is known as "central limit theorem", we can deduce that if

$$Z_n = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where n is the sample size, then as $n \to \infty$, the distribution function of Z_n converges to that of the standard normal distribution $N(0,1)$. Thus for n very large $Z_n \sim N(0,1)$ "approximately". So I leave it to the reader to deduce that the $(1-\alpha)100\%$ confidence interval of μ is given as

$$\left(\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \quad \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right)$$

Of course the next question one may ask is what about the variance. If we are studying a normal population with mean μ and variance σ^2, what is the confidence interval for σ^2, assuming of course σ^2 is not known. This is what we will study in the next section.

3. Interval Estimation for the Variance

As before let us draw a sample $X_1, \ldots, X_n$ from a normal distribution with mean μ and variance σ^2. Here the variance is unknown. In order to proceed we will introduce what is known as a pivotal quantity.

Defn 3.1 (Pivotal Quantity)

Let $X_1 \ldots X_n$ be a random sample of size n drawn from a density $f(\cdot, \theta)$; where θ as before is the parameter of the distribution. Let $Q = q(X_1 \ldots X_n, \theta)$, ie Q is a function of the sample observations and θ. If Q is a probability density that does not depend on θ; then Q is called a pivotal quantity.

E.g: In the examples of estimating mean $Z_n = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ is a pivotal quantity.

So the key idea of estimation is to find q_1 & q_2 such that

$$P[q_1 < Q < q_2] = (1-\alpha).$$

In the particular case of estimating variance; the pivotal quantity is

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

Now $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2$ with $(n-1)$ degrees of freedom.

$$\therefore P\left[q_1 < Q < q_2\right] = 1-\alpha$$

$$\Rightarrow P\left[q_1 < \frac{(n-1)s^2}{\sigma^2} < q_2\right] = 1-\alpha$$

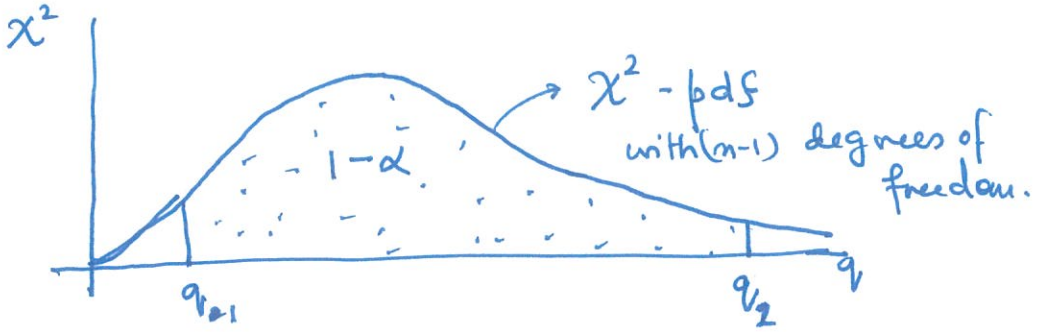
(s^2 is the observed value of the r.v. S^2)

$$\therefore \quad P\left[\frac{(n-1)s^2}{q_2} < \sigma^2 < \frac{(n-1)s^2}{q_1}\right] = 1-\alpha$$

$\therefore$ The $(1-\alpha)100\%$ confidence intervals are then given as

$$\left(\frac{(n-1)s^2}{q_2}, \quad \frac{(n-1)s^2}{q_1}\right)$$

Remember that χ^2 with $(n-1)$ degrees of freedom is not a symmetric distribution



The key idea is to find q_1 and q_2 such that

$$P[Q < q_1] = \alpha/2 \quad \& \quad P[Q > q_2] = \alpha/2.$$

This is ~~equal~~ called equal tails estimation. In practice one should figure this out from the Statistical tables related to χ^2-distribution but as before one may also obtain this by numerically solving the optimization problem.

$$\min \min L_1 = (n-1)s^2\left[\frac{1}{q_1} - \frac{1}{q_2}\right]$$

$$\text{Subject to} \quad \int_{q_1}^{q_2} f_Q(q)\,dq = 1-\alpha.$$

4. Estimation of the Confidence Interval for the difference of the means of two population

In this case we will be drawing samples from two normal populations.

Sample 1: will be drawn from $N(\mu_1, \sigma_1^2)$

Sample 2: will be drawn from $N(\mu_2, \sigma_2^2)$

We will assume that σ_1^2 and σ_2^2 are known. We want to ~~see the~~ estimate the difference of the means μ_1 and μ_2, i.e $\mu_1 - \mu_2$ to know how different these two normal populations are. Size of sample 1 is n_1

The pivotal quantity in this case is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where $\bar{X}_1$ is the mean of the first sample and $\bar{X}_2$ is the mean of the ~~smal~~ second sample. One can show that

$$Z \sim N(0,1)$$

Observe that

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

As, $\bar{X}_1$ & $\bar{X}_2$ are independent

It can be proved using any approach of finding the sampling distribution of $\bar{X}_1 - \bar{X}_2$, that

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Thus $Z \sim N(0,1)$.

Hence if we are looking for a $(1-\alpha)100\%$ ~~interval~~ confidence interval then

$$(\bar{x}_1 - \bar{x}_2) - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

If the samples are large and σ_1^2 & σ_2^2 are not known then replace them by s_1^2 and s_2^2 in the expression of Z and proceed using the central limit theorem.

In that case we have

$$(\bar{x}_1 - \bar{x}_2) - Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

However if the sample size is small ~~th~~ ie ($n_1 < 30$ or $n_2 < 30$ or both), then the confidence interval can be used if both population is assumed to have the same variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$ even ~~though~~ though its exact value is not known. Then the key is to consider the pooled sample variance

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

This is nothing but an weighted average of s_1^2 & s_2^2; i.e

$$s_p^2 = \frac{\sum_{i=1}^{n_1}(x_i^1 - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_i^2 - \bar{x}_2)^2}{n_1+n_2-2}$$

You will show in the assignments that $E(s_p^2) = \sigma^2$

Construct a new random variable

$$Y = \frac{(n_1-1)s_1^2}{\sigma^2} + \frac{(n_2-1)s_2^2}{\sigma^2} = \frac{(n_1+n_2-2)s_p^2}{\sigma^2}$$

Now $\frac{(n_1-1)}{\sigma^2}s_1^2 \sim \chi^2$ with n_1-1 degrees of freedom

and $\frac{(n_2-1)}{\sigma^2}s_2^2 \sim \chi^2$ with n_2-1 degrees of freedom.

Thus $Y \sim \chi^2$ with n_1+n_2-1 degrees of freedom. Now construct the pivotal variable

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Of course it can be shown that T follows the t distribution with n_1+n_2-2 degrees of freedom. Once this is known we can go ahead and write down the $(1-\alpha)100\%$ confidence intervals. But how did we get T.

Note that let us set

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Of course $Z \sim N(0,1)$. It can also be shown that Z and Y are independent. (Note that ~~we have~~ men one can show that $\bar{x}$ and S^2 are independent).

$$\therefore \quad T = \frac{Z}{\sqrt{\frac{Y}{n_1+n_2-2}}}$$

Hence $T \sim t$ with (n_1+n_2-2) degrees of freedom. This is known from our study of sampling distributions. The fact that $T \sim t$ with (n_1+n_2-2) degrees of freedom precisely needs the fact that Z and Y are independent

$$\therefore \quad T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Hence that $(1-\alpha)100\%$ confidence interval is given as

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2,n-1}\, S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2,(n-1)}\, S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Lecture 13 : Hypothesis Testing: (Classical Theory)

Sec1: The Basics

In many situations one is called upon to take a decision about a statistical statement, either quantitative or qualitative, in nature about parameters of some distribution representing the populations at hand. The decision will be to either accept or reject that statement, or hypothesis. Such a statement is usually called the null-hypothesis denoted as H_0. Associated with a null hypothesis is an alternative hypothesis denoted as H_1, which we accept if we reject H_0.

A simple way to think as follows. Let a population is described by a pdf $f_X(\cdot, \theta)$, and we are to test the hypothesis (null-hypo.)

$$H_0: \quad \theta = \theta_0$$

Against the alternative (say).

$$H_1: \quad \theta > \theta_0 \text{ (or say } \theta \neq \theta_0)$$

The idea is to generate a random sample from the population $f_X(\cdot, \theta)$ and based on some statistic which an estimator of θ or some related function of the observation to make an informed decision about whether to accept or reject H_0. Such a situation appears pretty often during clinical trials of drugs. Let a company X makes a new drug β for a disease for which a drug α exists and for which the average recovery rate of patients be μ_α. However the company X, claims that the new drug β, is better for which it has forecasted an average recovery rate of μ_β. If the drug β be really better than α, then

we must have to demonstrate that $\mu_\beta > \mu_\alpha$, for enough to sample (i.e. patient under trial) observations. However in medicine trials this ~~do~~ decision problem as the following hypothesis testing problem:

$$H_0 : \mu_\alpha = \mu_\beta \quad \text{(Null hypothesis)}$$

is tested against alternative

$$H_1 : \mu_\beta > \mu_\alpha.$$

Usually there are several ways to represent a hypothesis testing problem. One is called the simple representation, like

$$H_0 : \theta = \theta_0, \qquad H_1 \quad \theta = \theta_1.$$

While one can talk of a compound ~~hy~~ representation.

There are two types of error associated with it

Type-I error : Rejecting H_0 when it is true

Type-II error : Accepting H_0 when it is false, and H_1 is true.

Note: When you reject H_0 you accept H_1.

The basic structure of the testing procedure is as follows. Let the population under test is described by the pdf $f_X(\cdot, \theta)$, where $\theta \in \Omega$. Consider the hypothesis testing problem

$$H_0 : \theta_0 \in A \quad \text{against} \quad H_1 : \theta_1 \in B.$$

Note that $A \cup B = \Omega$.

Let $\mathcal{D}$ be the space of all sample observations. The test of H_0 against H_1 is based on subset C of $\mathcal{D}$ called the <u>critical region</u> of the test. It simply means one rejects H_0 if the sample observation falls in C.

~~The probability of a type-I error is given as follows~~

The rejection of H_0, given that the observed values $(x_1, x_2 \ldots x_n)$ lies in the critical region is called a non-randomized test. Given a test procedure T, we shall denote henceforth the critical region. To begin with let us define the power function of a test.

~~Let~~ Definition 13.1 Let T be a test of the null hypothesis

The power function $\pi_T(\theta)$ of T is defined to be the probability of rejecting H_0, when the distribution from which the sample is obtained is parametrized θ.

Thus

$$\pi_T(\theta) = P_\theta[\text{Reject } H_0]$$
$$= P_\theta[(x_1 \ldots x_n) \in C_T]$$

Example. 1. Let us draw a sample from a normal population with mean $\mu = \theta$, (unknown) and $\sigma^2 = 25$.

Consider the following hypothesis problem

$H_0 : \theta \leq 17$, against $H_1 : \theta > 17$.

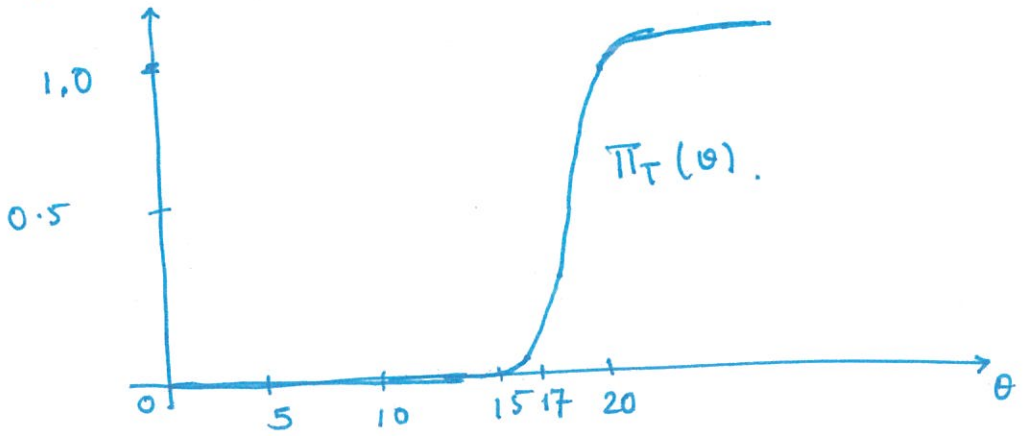
The test T is as follows: Reject if and only if $\bar{x} > 17 + \frac{5}{\sqrt{n}}$ where n is the sample size. Here

$$C_T = \left\{(x_1 \ldots x_n) : \bar{x} > 17 + \frac{5}{\sqrt{n}}\right\}$$

$$\pi_T(\theta) = P_\theta\left[\bar{x} > 17 + \frac{5}{\sqrt{n}}\right]$$
$$= P_\theta\left[\frac{\bar{x} - \mu\theta}{5/\sqrt{n}} > \frac{17 + \frac{5}{\sqrt{n}} - \theta}{5/\sqrt{n}}\right]$$
$$= 1 - \Phi\left(\frac{17 + \frac{5}{\sqrt{n}} - \theta}{\frac{5}{\sqrt{n}}}\right)$$

In this scenario where the hypothesis is composite, it is better to plot the graph of $\pi_T(\theta)$, to understand how effective is this particular test. We present is the graph from the book "Introduction to the Theory of Statistics" by Mood, Graybill and Boes, from where the above example was taken

They considered $n = 25$ and we have



So let us see what do we observe. When $\theta \leq 16$, it is clear we accept H_0. of course if $\theta \geq 19$ say we ~~de~~ clearly reject H_0 as $\pi_T(\theta^*) > 0.5$. But if $17 < \theta < 18$, we may also ~~accept H_0 as~~ reject H_0 as $\pi_T(\theta)$ is near 0.5 in some cases but may accept for $\theta > 17$ but very near it.

Following the approach of Mood, Graybill and Boes let us define the notion of the size of a test.

The Size of a Test: Let the null hypothesis be given as $H_0 : \theta \in \Omega_0$, where $\Omega_0 \subseteq \Omega$, where Ω is called the parameter space. The size of a test T associated with H_0 is defined as follows

$$\text{Size of Test} = \sup_{\theta \in \Omega_0} \pi_T(\theta)$$

$$= \sup_{\theta \in \Omega_0} P_\theta[\text{Reject } H_0]$$

- For a non-randomized test the size of test is also referred as size of the critical region.

Let us try to find out the size of the test for the previous example. Just to recall, let us state that the random sample $X_1, \ldots, X_n$ be a random sample of size n drawn from a normal population with an unknown mean μ and variance $\sigma^2 = 25$. Our test T; if we recall was as follows.

Test T: Reject H_0 if $\bar{X} > 17 + \frac{5}{\sqrt{n}}$

Here $\Omega_0 = \{\mu = \theta : \theta \leq 17\}$, (Remember we are talking about the mean μ)

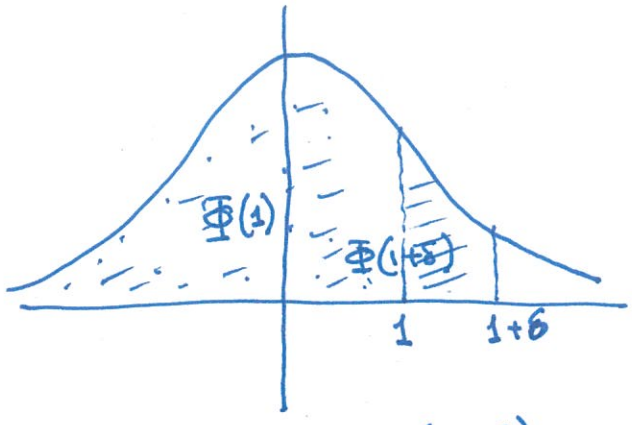
$$\pi_T(\theta) = 1 - \Phi\left(\frac{17 + 5/\sqrt{n} - \theta}{5/\sqrt{n}}\right)$$

$\therefore$ Size of the test : =

$$\sup_{\theta \leq 17}\left[1 - \Phi\left(\frac{17 + 5/\sqrt{n} - \theta}{5/\sqrt{n}}\right)\right]$$

$$= 1 + \sup_{\theta \leq 17}\left[-\Phi\left(\frac{17 + 5/\sqrt{n} - \theta}{5/\sqrt{n}}\right)\right] \Big\}^{**}$$

$$= 1 - \inf_{\theta \leq 17}\left[\Phi\left(\frac{17 + 5/\sqrt{n} - \theta}{5/\sqrt{n}}\right)\right]$$



$$\Phi(1) < \Phi(1+\delta)$$

$$\therefore \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx = \int_{-\infty}^{1+\delta} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz.$$

Observe that for $\theta = 17$,

$$\frac{17 + 5/\sqrt{n} - \theta}{5/\sqrt{n}} = 1$$

For any other $\theta < 17$

$$\frac{17 - \theta + 5/\sqrt{n}}{5/\sqrt{n}} = 1 + \left(\frac{17-\theta}{5/\sqrt{n}}\right)$$

which is greater than 1.

** we have used the fact

$$-\sup(-\phi(x)) = \inf(\phi(x))$$

Thus

$$\text{Size of the test} := 1 - \inf_{\theta \leq 17} \left[\bar{\Phi} \left(\frac{17 + 5/\sqrt{n} - \theta}{5/\sqrt{n}} \right) \right]$$

$$= 1 - \inf \Phi \left(\frac{17 + 5/\sqrt{n} - 17}{5/\sqrt{n}} \right)$$

$$= 1 - \Phi(1).$$

$$\therefore \boxed{\text{Size of test} = 1 - \Phi(1) \approx 1 \cdot 6 \quad \text{(Approx)}.}$$

Let us take a diversion and think a bit about, what is a ~~non~~ randomized test.

For a non-randomized test, we observe a sample and see if it is in the critical region, then reject it and accept it if it is the acceptance region. In fact the problem arises when the observed sample is in the boundary of these two regions. For example, consider the example we discussed above. When $\theta \in (17, 18)$, we were not completely sure what decision to take. For values of $\theta > 17$ but very near 17 we are tempted to accept H_0, while for values of θ near 18, we are more willing to reject it.

~~This~~ ~~These are preci~~ In such a kind of scenario we take the help of ~~non~~ randomized test to come to a decision.

A ~~random miz~~ randomized test T depends on the following critical function

$$\psi_T(x_1 \ldots x_n) = P\left[H_0 \text{ is rejected} \mid (x_1 \ldots x_n) \text{ is observed} \right].$$

~~Once~~ Once we have observed $(x_1 \ldots x_n)$, we first compute $\psi_T(x_1 \ldots x_n)$ and then carry out an auxilliary Bernoulli trial (a single trial with an outcome, which is either a success or a failure), which does not depend on $(x_1 \ldots x_n)$

and for which the probability of success is $\Psi_T(x_1, \ldots, x_n)$. If we observe a success in the Bernoulli trial, a then we reject H_0.

Let us see the following example from Mood, Graybill and Boes — "Introduction to the Theory of Statistics".

Example 2: Let us consider a sample of size 10, ie $x_1, x_2 \ldots x_{10}$ from a Bernoulli population described as

$$f_X(x, \theta) = \theta^x (1-\theta)^{1-x}, \quad x = 0 \text{ or } 1$$

where $x = 0$ is failure $x = 1$ is success, and θ represents the probability of success and let of course $\theta > 0$. Do not think that this Bernoulli population is just coin tossing. It could be an experiment of an archer hitting the bull's eye; or checking whether a light bulb is ~~enumerated~~ functioning or not.

Let $H_0 : \theta < 1/2$, versus $H_1 : \theta \geq 1/2$

Let us consider the following test, T:

Reject H_0; if $\sum_{i=1}^{10} x_i > 5$

Accep H_0; if $\sum_{i=1}^{10} x_i < 5$.

Now what happens if $\sum_{i=1}^{10} x_i = 5$. This is where randomization comes in. The most simple way is to decide through a coin toss. If the toss shows head (H), then reject, H_0 and if it shows tail, (T), then accept H_0. In this case how do we find the critical function. Let us partition the region of possible sample observations into three zones

$$A = \left\{ (x_1 \ldots . x_{10}) : \sum_{i=1}^{10} x_i < 5 \right\}$$

$$B = \left\{ (x_1 \ldots . x_{10}) : \sum_{i=1}^{10} x_i = 5 \right\}$$

and

$$C = \left\{ (x_1 \ldots x_{10}) : \sum_{i=1}^{10} x_i > 5 \right\}$$

$$\psi_T(x_1 \ldots x_{10}) = \begin{cases} 1 & : \text{if } (x_1 \ldots x_{10}) \in C \\ \frac{1}{2} & : \text{if } (x_1 \ldots x_{10}) \in B \\ 0 & : \text{if } (x_1 \ldots x_{10}) \in A. \end{cases}$$

This means if $(x_1 \ldots x_{10}) \in C$, H_0 will be rejected with probability 1. If $(x_1 \ldots x_{10}) \in B$ it will be rejected with probability $\frac{1}{2}$. If $(x_1 \ldots x_{10}) \in C$, then H_0 will be accepted

Section 2: Testing Simple Hypothesis against Simple Alternative.

In this section we will study the design for testing a simple hypothesis against a simple alternative. More precisely a very general way of writing, a simple hypothesis versus a simple alternative. Given a random sample $X_1, \ldots X_n$ we need to decide in which of the two populations, $f_x^0(\cdot)$ or $f_x^1(\cdot)$ this random sample comes from.

Thus we have

$$H_0 : X_i \sim f_x^0(\cdot) \quad \text{against} \quad H_1 : X_i \sim f_x^1(\cdot)$$

Suppose we just have one observation $X_1 = x_1$, then if $f_x^0(x_1) > f_x^1(x_1)$ we accept H_0 and reject it if $f_x^1(x_1) > f_x^0(x_1)$. This simple idea leads us to what is known as the likelihood ratio test.

- **Simple Likelihood Ratio Test**: Given the above H_0 versus H_1 The simple likelihood test T is given as.

$$\left.\begin{array}{l} \text{Reject } H_0 : \text{If } \lambda < k \\ \text{Accept } H_0 : \text{If } \lambda > k \end{array}\right\} (k > 0)$$

If $\lambda = k$, then either we accept H_0 or reject H_0, may be using a randomized test.

We have

$$\lambda = \lambda(x_1 \ldots x_n) = \frac{\prod_{i=1}^{n} f_X^0(x_i)}{\prod_{i=1}^{n} f_X^1(x_i)} = \frac{L_0(x_1, \ldots x_n)}{L_1(x_1, \ldots x_n)}$$

and $k \geq 0$. Thus λ is the ratio of the Likelihood functions. This leads us to the following section.

Section 3: Most powerful test and Neyman-Pearson Lemma

Consider the following testing of hypothesis problem:

$H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$

H_0: Null hypothesis

H_1: Alternative hypothesis

Corresponding to any test T, of H_0 against H_1, the power function is denoted by $\pi_T(\theta) = P_\theta(\text{Reject } H_0)$.

Now $\pi_T(\theta_0) = P_{\theta_0}(\text{Reject } H_0)$

i.e. $\pi_T(\theta_0) = P(\text{Reject } H_0 \mid H_0 \text{ is true})$

This is the probability of type-I error and this should be reasonably small.

While,

$$\pi_T(\theta_1) = P_{\theta_1}(\text{Reject } H_0 \mid H_0 \text{ is false})$$

Of course we want $\pi_T(\theta_1)$ to be large. Further we can write

$$\pi_T(\theta_1) = 1 - P(\text{Accept } H_0 \mid H_0 \text{ is false})$$

$$\Rightarrow P(\text{Accept } H_0 \mid H_0 \text{ is false}) = 1 - \pi_T(\theta_1)$$

$\downarrow$

Prob of Type-II error.

We denote β by

$$\boxed{\beta_T(\theta_1) = 1 - \pi_T(\theta_1)}$$

Ideally both $\pi_T(\theta_0)$ & $\beta_T(\theta_1)$ should be small. In practice one often fixes the value of $\pi_T(\theta_0)$ to a number say $0 < \alpha < 1$, ie $\pi_T(\theta_0) = \alpha$, often called the size of the test and thus try to find a test which minimizes $\beta_T(\theta_1)$. Such a test is often called the most powerful test, which is what we define below.

A test T^* for testing $H_0 : \theta = \theta_0$, versus $H_1 : \theta = \theta_1$, is called the most powerful test if $\pi_{T^*}(\theta_0) = \alpha$ and for any test T with $\pi_T(\theta_0) \leq \alpha$ we have

$$\boxed{\beta_{T^*}(\theta_1) \leq \beta_T(\theta_1)}$$

The following result of Neyman and Pearson shows us how to devise the most powerful test.

Theorem 1: Neyman-Pearson Lemma [Mood's, Graybill and Boes. Introduction to the theory of stat]

Let $X_1 \ldots X_n$ be a random sample of size n from a population $f_X(x,\theta)$, where θ is one of the two values θ_0 or θ_1. Let us test the following hypotheses

$H_0 : \theta = \theta_1$ versus $H_1 : \theta = \theta_2$

Further let us set $0 < \alpha < 1$, be fixed

Let $k^* > 0$ and $C^* \subset \Omega_n$ where Ω_n is the set of all possible sample values from $f_X(x,\theta)$. Assume that the following relationship holds. facts hold.

i) $P_{\theta_0}\left((x_1 \ldots x_n) \in C^*\right) = \alpha$

ii) ~~Let $\lambda = \frac{L_0(\theta_0)}{L_1(\theta_1)}$~~

Let $\lambda = \dfrac{L_0(x_1, x_2, \ldots x_n, \theta_0)}{L_1(x_1, x_2, \ldots x_n, \theta_1)} = \dfrac{L_0}{L_1}$

where $L_0 = L_1(x_1 \ldots x_n, \theta_0)$ $\Rightarrow$ Likelihood function

$L_1 = L_1(x_1 \ldots x_n, \theta_1)$

$\lambda \leq k^*$ if $(x_1, \ldots, x_n) \in C^*$

$\lambda \geq k^*$ if $(x_1^*, \ldots x_n^*) \in (C^*)^c$; $\left[(C^*)^c = \Omega_n \setminus C^*\right]$

Then the test T^* corresponding to the critical region $C_{T^*} = C^*$ is a most powerful test of size α for testing $H_0: \theta = \theta_0$ against $H_{01}: \theta = \theta_1$.

Proof: (This can be skipped in first reading).

We shall for simplicity consider any other test T, such that $\pi_T(\theta_0) = \alpha$. Let the critical region for T be given by D.

$$\therefore \underbrace{\int\int\cdots\int}_{\substack{n\text{-fold} \\ \text{integral} \\ \text{over } C^*}} L_0 \, dx = \int\int_D \cdots \int L_0 \, dx = \alpha$$

Since $\pi_{T^*}(\theta_0) = \int\int_{C^*}\cdots\int L_0 \, dx = \alpha$, where T^* is the test corresponding to C^*

Here $dx = dx_1 \ldots dx_n$.

$C^* = (C^* \cap D) \cup (C^* \cap D^c)$ and as $(C^* \cap D) \cap (C^* \cap D^c) = \phi$

we have

$$\iint\!\!\dots\!\int_{C^*} L_0\,dx = \iint\!\!\dots\!\int_{C^* \cap D} L_0\,dx + \iint\!\!\dots\!\int_{C^* \cap D^c} L_0\,dx = \alpha$$

Similarly

$$\iint\!\!\dots\!\int_{D} L_0\,dx = \iint\!\!\dots\!\int_{C^* \cap D} L_0\,dx + \iint\!\!\dots\!\int_{(C^*)^c \cap D} L_0\,dx = \alpha$$

From the above two equations we have

$$\iint\!\!\dots\!\int_{C^* \cap D^c} L_0\,dx = \iint\!\!\dots\!\int_{(C^*)^c \cap D} L_0\,dx$$

Now inside C^* we have $\lambda = \frac{L_0}{L_1} \leq k^*$ or $L_1 \geq \frac{L_0}{k^*}$

(Think why k^* has to be positive!)

$$\therefore \iint\!\!\dots\!\int_{C^* \cap D^c} L_1\,dx \geq \iint\!\!\dots\!\int_{C^* \cap D^c} \frac{L_0}{k^*}\,dx = \iint\!\!\dots\!\int_{(C^*)^c \cap D} \frac{L_0}{k^*}\,dx \geq \iint\!\!\dots\!\int_{(C^*)^c \cap D} L_1\,dx$$

$\left(\because \frac{L_0}{L_1} \geq k^* \text{ on } (C^*)^c\right)$

$$\therefore \iint\!\!\dots\!\int_{C^* \cap D^c} L_1\,dx \geq \iint\!\!\dots\!\int_{(C^*)^c \cap D} L_1\,dx$$

Now

$$\iint\!\!\dots\!\int_{C^*} L_1\,dx = \iint\!\!\dots\!\int_{C^* \cap D^c} L_1\,dx + \iint\!\!\dots\!\int_{C^* \cap D} L_1\,dx$$

$$\geq \iint\!\!\dots\!\int_{(C^*)^c \cap D} L_1\,dx + \iint\!\!\dots\!\int_{C^* \cap D} L_1\,dx = \iint\!\!\dots\!\int_{D} L_1\,dx.$$

This shows that

$$\pi_{T^*}(\theta_1) \geq \pi_T(\theta_1),$$

$$\Rightarrow \quad \beta_{T^*}(\theta_1) \leq \beta_T(\theta_1)$$

showing that T^* is the most powerful test. This proves the result.

Example: 1; Let $x_1 \ldots x_n$ be a random sample of size n from an exponential distribution given as

$$f_X(x, \theta) = \theta e^{-\theta x}, \quad \theta > 0, \quad x > 0$$

$$\therefore \quad L_0 = L(x_1 \ldots x_n \theta_0) = (\theta_0)^n e^{-(\theta \sum x_i)}, \text{ while}$$

$$L_1 = L_1(x_1 \ldots x_n \theta_1) = (\theta_1)^n e^{-(\theta \sum x_i)}.$$

Our aim is to develop the most powerful test of size α for the hypothesis testing problem

$$H_0 : \theta = \theta_0, \text{ against } H_1 : \theta = \theta_1 \quad (\theta_1 > \theta_0)$$

To find the most powerful test we have to in effect find the k^*. Let us see how we do it in this case.

The idea is reject H_0 if $\lambda \leq k^*$ (from the Neyman-Pearson lemma)

$$\text{Thus } \frac{L_0}{L_1} \leq k^* \Rightarrow \left(\frac{\theta_0}{\theta_1}\right)^n e^{\left[-(\theta_0 - \theta_1)\sum x_i\right]} \xrightarrow{(\#)} \leq k^*$$

$$\Rightarrow \quad \sum x_i \leq \frac{1}{\theta_1 - \theta_0} \log_e \left[\left(\frac{\theta_1}{\theta_0}\right)^n k^*\right]$$

This is obtained by taking logarithm on both sides of (#).

$$\text{Set } \frac{1}{\theta_1 - \theta_0} \log_e \left[\left(\frac{\theta_1}{\theta_0}\right)^n k^*\right] = k \text{ (say)}.$$

Thus

$$\frac{L_0}{L_1} \leq k^* \Rightarrow \sum_{i=1}^{n} x_i \leq k.$$

$$\therefore \quad \alpha = P_{\theta_0}\left[(x_1 \ldots x_n) \in C^*\right]$$

$$= P_{\theta_0}\left[\sum_{i=1}^{n} x_i \leq k^*\right]$$

Now since $X_i \sim \exp(\theta)$ we know from our study of sampling distribution we know that

$$\sum_{i=1}^{n} x_i \sim \text{Gamma}(n, \theta)$$

$\therefore$ If we set $Y = \sum_{i=1}^{n} x_i$, then

$$P_{\theta_0}\left[\sum_{i=1}^{n} x_i \leq k\right] = P_{\theta_0}[Y \leq k] = \int_0^k \frac{1}{\Gamma(n)} \theta_0^n y^{n-1} e^{-y\theta_0} dy = \alpha$$

Once sample values are observed, then

$$\int_0^k \frac{1}{\Gamma(n)} \theta_0^n y^{n-1} e^{-y\theta_0} dy = \alpha$$

is an equation in k from which k can be computed in fact by using the tables. Once the k is known, our test T^*:

> Reject H_0 : if $\sum x_i \leq k$.

By Neyman-Pearson's lemma, we see that T^* is the most powerful test.

—x—

> The course HS0201A formally ends here