Lecture Notes

ENEE 324 – Engineering Probability

P.S. Krishnaprasad

Fall, 2002

PROBABILITY: a bit of etymology

Wahr -true, genuine
Wahrhaft - truthful, true
Wahrhaftigkeit - truthfulness
Wahrheit - truth, fact
Wahrnehmen - to perceive, observe, notice
Wahrnehmung - perception
Wahrsagen - to prophesy, tell fortunes
Wahrsager -
Wahrsagerin - soothsayer, fortune-teller
Wahrsagung - prophecy
Wahrscheinlich - probable, probably
Wahrscheinlichkeit - probability
Wahrscheinspruch - verdict
Wahrscheinlichkeitstheorie - probability theory

– Collins Gem German-English Dictionary

ENEE 324 Engineering Probability Lecture 1

Basic Concepts

This is a course on *modeling uncertainty*. Uncertainty is all about us – the outcome of a coin toss, the life-time of an electric light bulb, Nyquist-Johnson fluctuations in the measured value of current through a resistor connected to a heat bath, the chance of rain at noon on a weekday, are all valid examples of uncertain/chance/random phenomena. Yet, through study of specific contexts, and by carrying out carefully repeated experiments, it is possible to get a handle on uncertainty, sufficiently to be able to make useful predictions. A great deal of science and engineering is concerned with making predictions in the face of uncertainty. Probability theory provides the language, the techniques, and as a consequence the mathematical models that enable us to do this.

There are other ways to approach uncertainty, but probability theory is quite possibly the most wide-ranging and successful means to do this. Probability theory offers a coherent conceptual system to understand and cope with uncertainty.

Modern technology makes extensive use of probability theory. Some examples include: (a) algorithms used to route messages/data in a communication/computing network; (b) techniques used to project the yield in acceptable quality silicon wafers in a semiconductor manufacturing plant; (c) the error-correcting codes used in compact disc players; (d) performance analysis and design of a service system using the theory of queues (waiting lines).

Everyday use of the language of probability is based on built-up intuition that people have. Sometimes such intuition can prove unreliable or ill-defined. One can build correct intuition by solving certain "toy problems", such as card-shuffling. It is useful and advisable to develop a systematic approach to probability. In particular, the models of probability have to be tested for "consistency" against data (observed in experiments).

Often, costly and sensitive decision-making processes depend on probability models. Some examples: (a) the decision by a "wild-catter" to drill or not to drill for oil in a particular parcel of optioned land; (b) the decision to launch a space-shuttle based on forecasts of weather patterns; (c) the decision to attempt circum-navigation of the globe in a hot-air balloon; (d) the decision to attempt maiden voyage of a grand ocean liner in sea-lanes known to be populated by ice-bergs. The *risks* involved in such decision processes must be quantified so that an experienced and competent human can make *rational* choices. Lloyd's of London quantifies such risks all the time. How? The answer lies in probabilistic concepts.

Probability can also be used to answer (approximately) questions in fields where one normally does not expect to have to deal with uncertainty. An example of this is the *Buffon's needle* problem: suppose a needle is "tossed at random" onto a plane ruled with parallel lines a distance L apart, where by a "needle" we mean a line segment of length $l \leq L$. What is the probability of the needle intersecting one of the parallel lines?

We present a systematic approach to this important subject by beginning with fundamental concepts.

Very often, one thinks of a problem involving uncertainty as being associated to an *experiment* $\mathcal{E}$. If $\mathcal{E}$ is repeatable, so much the better. There is a whole school of thought, that insists on attaching probabilities only to repeatable experiments, known as *the frequentists*. Yet, there are problems involving uncertainty where no natural experiments can be suggested to model or deduce the uncertainty. For instance, despite having a large body of solid geophysical knowledge and experience, a geophysicist, when called upon to offer what he/she thinks of as the "likelihood" of a cataclysmic earthquake on the eastern sea-board by the year 2000, may appear to "pick a percentage out of the hat". What is going on here is that the number offered is a measure of the scientist's conviction – an example of *subjective* probability. (There is a history of raging arguments between subjectivists and frequentists. After all, is it not the goal of science to be objective and stamp out all that carries the taint of prejudice/subjectivity? We will meet on our journey, representatives of both camps,–Richard von Mises, Bruno de Finetti, John Maynard Keynes, Leonard Savage, Ronald A. Fisher,...). Whether the probabilities that we discuss below are based on (repeatable) experiments or based on an

expert's conviction, the rules for working with probabilities are the same. These rules serve as a foundation for mathematical modeling of uncertainty. At a fundamental level, these are based on the language of *set theory* and *Boolean algebra*.

First, we need a set Ω, called the *sample space*. The elements of this set are the (exhaustive list of) possible *outcomes* of an experiment $\mathcal{E}$. With reference to $\mathcal{E}$, we will have the notion that Ω is a universal set, i.e., all possible outcomes of $\mathcal{E}$ are accounted for in Ω.

Examples

(i) $\mathcal{E}$ = single coin toss, $\Omega = \{H, T\}$

(ii) $\mathcal{E}$ = roll of a single die, $\Omega = \{1, 2, 3, 4, 5, 6\}$

(iii) $\mathcal{E}$ = coin toss until a first head, $\Omega = \{H, TH, TTH, ...\}$

(iv) $\mathcal{E}$ = mark a random dot on a ruler of length L. Here we take $\Omega = [0, L]$.
(Note, this is an easy experiment to repeat and there are different ways to repeat it, either via independent trials or dependent trials.)

(v) $\mathcal{E}$ = survey of all computers that are up or down at 11:00 a.m. Here Ω can be taken as simply a list of all IP addresses with a tag UP or DOWN

In example (i) and (ii) the sample space Ω is a finite set. In (v) it is finite but large (in Maryland campus)! In (iii) it is countably infinite. In (iv) it is uncountably infinite. In the beginning we will concentrate on situations wherein Ω is finite or countably infinite (a discrete sample space).

An *event* A (associated to an experiment) is simply a set of possible outcomes, i.e. a subset of Ω. The collection of all possible events is denoted as 2^Ω and is called the power set of Ω.

Examples (associated to above experiments)

(i) head occurs: $A = \{H\}$

(ii) even number occurs: $A = \{2, 4, 6\}$

(iii) first head occurs in at most 3 tosses: $A = \{H, TH, TTH\}$

(iv) mark within halfway point: $A = [0, 0.5L]$
(v) only one computer is down: $A = \{u_1\ d_2\ u_3, d_1\ u_2\ u_3, ...\}$

Events cannot be discussed in isolation. Thus if the event A occured, then event A^c, the complement of A, did not occur. Thus we are also thinking about A^c even as we speak of A. (Remark: We also denote A^c as $\bar{A}$.) In fact we are thinking about a whole algebra of events constructed out of the operations of set intersection and set union, respectively mirroring the logical connectives AND and OR.

We state below, the elements of set theory relevant to probablity calculations: **A set is a collection of objects.**

The set of outcomes of rolling a die, $\Omega = \{1, 2, 3, 4, 5, 6\}$.
For each experiment $\mathcal{E}$ we need to define Ω. $\emptyset$ denotes the empty set.

(1) $A \sqsubset B$ means A is a subset of B. Then, $a \in A \Rightarrow a \in B$
(2) $A \cup B = C$ means $c \in C \Rightarrow c \in A$ <u>or</u> $c \in B$ (or both).
(3) $A \cap B = C$ means $c \in C \Rightarrow c \in A$ <u>and</u> $c \in B$.
(4) $\bar{A} = C$ means $c \in C \Rightarrow c \notin A$.
(5) $A \times B = C$ denotes the Cartesian product

The cartesian product of sets means $c \in C \Leftrightarrow c = (a, b)$ where $a \in A$ $b \in B$. Note that c is an <u>ordered</u> pair.

Using these basic operations, one builds more "complicated" events from elementary events. Given an experiment $\mathcal{E}$ with sample space Ω, any member of 2^Ω could be an event, in principle. In practice, one may limit oneself to a subcollection $\mathcal{A} \subseteq 2^\Omega$.

How to choose $\mathcal{A}$?

Basic ground rules for $\mathcal{A}$ (=Boolean algebra)

$A \in \mathcal{A} \Rightarrow \bar{A} \in \mathcal{A}$
$\Omega \in \mathcal{A}$
$\emptyset \in \mathcal{A}$
$A, B \in \mathcal{A}$ then $\quad A \cup B \in \mathcal{A}, \quad A \cap B \in \mathcal{A}$
We think of $\mathcal{A}$ as a collection of *interesting events*.

Example:

$\Omega = \{1, 2, 3, 4, 5, 6\}$
$\mathcal{A} = \{\emptyset, \Omega, A_1, A_2\}$, where $A_1 = \{1, 3, 5\}, \quad A_2 = \{2, 4, 6\}$.

We define probability in a manner that *agrees* with experimental observations – mimics frequencies, and consistently for all $A \in \mathcal{A}$.

Definition: Relative frequency of event A,
$f_A \stackrel{\triangle}{=} \frac{n_A}{n}$ where $n = \#$ repetitions/trials of $\mathcal{E}$ and $n_A = \#$ occurrences of A in n trials.
Check Properties

(i) $0 \leq f_A \leq 1$
(ii) $f_A = 1$ *iff* A occurs every time in the n trials/repetitions. In particular $f_\Omega = 1$.
(iii) $f_A = 0$ *iff* A never occurs in the n trials. In particular $f_\emptyset = 0$.
(iv) If A and B are disjoint, i.e. $A \cap B = \emptyset \Rightarrow f_{A \cup B} = f_A + f_B$. In particular $f_A = 1 - f_{\bar{A}}$.
(v) As $n \to \infty$, $f_A(n) \to P(A)$ **(??)**
For probability, turn these properties into **axioms**.
Given an experiment $\mathcal{E}$, sample space Ω, and collection of *interesting events* $\mathcal{A}$, a probability law or probability measure is a function, (Here the term *measure* used in the same way as a measure of length, or area, or volume.)

$P : \mathcal{A} \to [0, 1]$, satisfying
(a) $0 \stackrel{\leq}{=} P(A) \stackrel{\leq}{=} 1$
(b) $P(\Omega) = 1$
(c) $A \cap B = \emptyset \Rightarrow \; P(A \cup B) = P(A) + P(B)$ (addition rule).

Prove that

$P(\emptyset) = 0$
$P(A) = 1 - P(\bar{A})$

Some basic properties of probabilities

(1) $A \subseteq B \Rightarrow P(A) \leq P(B)$
Proof:
Let $C = \{y \in B : y \notin A\}$
Then $B = C \cup A$, and $C \cap A = \emptyset$
Thus: $P(B) = P(C \cup A) = P(C) + P(A)$
Since $P(C) \geq 0$, the result follows. $\square$

(2) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Proof:
$A \cup B = (A \cap B) \cup (A \cap \bar{B}) \cup (\bar{A} \cap B)$ is a decomposition into disjoint sets. By the addition of axiom of probability,

$$\begin{aligned} P(A \cup B) &= P(A \cap B) + P(A \cap \bar{B}) + P(\bar{A} \cap B) \\ &= \Big(P(A \cap B) + P(A \cap \bar{B})\Big) \\ &\quad + \Big(P(\bar{A} \cap B) + P(A \cap B)\Big) - P(A \cap B) \\ &= P\Big((A \cap B) \cup (A \cap \bar{B})\Big) \\ &\quad + P\Big((\bar{A} \cap B) \cup (A \cap B)\Big) - P(A \cap B) \end{aligned}$$

(by the addition axiom)

$$\begin{aligned} &= P\Big(A \cap (B \cup \bar{B})\Big) + P\Big((\bar{A} \cup A) \cap B\Big) - P(A \cap B) \\ &= P(A \cap \Omega) + P(\Omega \cap B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \qquad \square \end{aligned}$$

In the above, we have made use of the distributive law.

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

One builds probability laws by recognizing what the equally *likely* events are in a given experiment. The idea of equally likely outcomes draws on symmetry. No special status is given to any particular outcome. One then applies the axioms. Finite sample space problems are key to building intuition.

What are the elementary events in experiment (iv) above? Assuming they are equally likely, what is their common probability?

ENEE 324 Engineering Probability Lecture 2

Counting

For the case of rolling a single *fair* die, let $\Omega = \{1, 2, 3, 4, 5, 6\}$ and let

$$\mathcal{A} = 2^{\Omega} = \{\emptyset, \Omega, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{1, 2\}, \{2, 3\}...\}$$

There are $2^6 = 64$ members in $\mathcal{A}$, and only 6 of these are elementary events. Declare that all singletons (elementary events) are *equally likely.* [recall fairness assumption]

Since $P(\Omega) = 1$ and Ω is the union of 6 singletons, all equally likely, it follows from the addition axiom that,

$$\text{Probability of a singleton} = \tfrac{1}{6}.$$

The probabilities of all the other events in $\mathcal{A}$ can be determined from this one fact ! We simply apply the addition axiom.

In a deck of *well-shuffled* cards, the probability of drawing the heart $= \frac{1}{52}$.

Example 1: Toss a coin repeatedly until the first head. In each toss,

$$P\{H\} = p \ : \ P\{T\} = 1 - p = q.$$

$\Omega = \{1, 2, 3, ...\}$ = sample space of # tosses needed until first head.
Assume $p \neq 0$.

$$\begin{aligned} p_j &= p\{j \text{ tosses until first head}\} \\ &= q^{j-1} \cdot p \qquad j = 1, 2, ... \end{aligned}$$

Where did this come from?

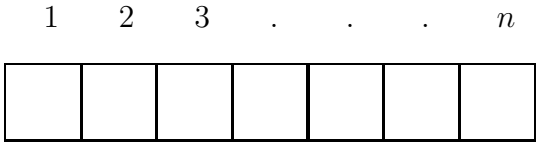
$$\begin{aligned}
\sum_{j=1}^{\infty} q^{j-1}p &= p \cdot (1 + q + q^2 + ...) \\
&= p \cdot \lim_{n\to\infty} (1 + q + q^2 + ... + q^n) \\
&= p \cdot \lim_{n \to \infty} \frac{1 - q^n}{1 - q} \\
&= \frac{p}{p} \qquad \text{(because } q < 1\text{)} \\
&= 1
\end{aligned}$$

Example 2: Lifetime of computer memory chip satisfies: "proportion of chips whose lifetime exceeds t decreases exponentially at the rate α."
Here $\alpha > 0$.

$$\begin{aligned}
\Omega &= (0, \infty) \\
P[(t, \infty)] &= e^{-\alpha t} && t > 0 \\
P[(0, \infty)] &= e^{-0\alpha} && = \text{ 1 as it should be.} \\
P[(r, s)] &= P[(r, \infty)] - P[(s, \infty)] && = e^{-\alpha r} - e^{-\alpha s} \text{ , } r < s
\end{aligned}$$

Some combinatorics

(a) Given n distinct things, how many ways can we permute them?
Think of this as filling n marked cells



Fill cell 1 in any of n ways.
Fill cell 2 in any of $(n-1)$ ways (with the remaining $(n-1)$ things).
Fill cell 3 with any of $(n-2)$ ways.

Fill cell n in (1) way.

Total number of ways of filling cells is $nPn = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1$
We call this $n!$.

(b) Given n distinct things, how many different **permutations** of r things can we make from these n things?

Treat the problem as one of filling r out of n cells.

Proceeding as before we get

$$\begin{aligned} nPr &= n(n-1)(n-2)...(n-r+1) \\ &= \frac{n!}{(n-r)!} \end{aligned}$$

SAMPLING
(c) Given n distinct things, how many **combinations** of r things out of these n things can we make?
Denote this yet to be determined quantity as nCr.
Combinations ignore order. Thus,

$$\begin{aligned} nCr \cdot r! &= nPr \\ &= \frac{n!}{(n-r)!} \quad . \\ \text{H}ence \quad nCr &= \frac{n!}{(n-r)!r!} \end{aligned}$$

The sampling is said to be **random** if all of these combinations are equally

likely. So the probability of a particular combination being picked up in a random sample is $\frac{(n-r)!r!}{n!}$
It is common to use the notation $\binom{n}{r}$ instead of nCr. These integers have a long history. **Newton's binomial expansion** says

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

Proof: $(a+b)^n$ is an expression, homogeneous of degree n. Hence each term in $(a+b)^n$ will be of the form $a^k b^{n-k}$. How many are of this form? $\binom{n}{k}$ $\square$

Identitites

$$\begin{aligned} \text{(i)} \quad \binom{n}{r} &= \binom{n}{n-r} \\ \text{(ii)} \quad \binom{n}{r} &= \binom{n-1}{r-1} + \binom{n-1}{r} \end{aligned}$$

Single out an object – a_k, say.
Numbers of choices of r objects out of n objects = (number of choices that exclude a_k) + (number of choices that include a_k) = $\binom{n-1}{r} + \binom{n-1}{r-1}$

Example 3: (sample without replacement)

Total of N items.
Choose n at random without replacement.
This will yield $\binom{N}{n}$ possible samples.

If the N items are made up of r_1 blues and r_2 reds, $r_1 + r_2 = N$, then the probability of choosing *exactly* s_1 blues and $(n - s_1)$ reds, (here $s_1 \leq n$ and

$s_1 \le r_1$, $(n - s_1) \le r_2$), is given by

$$\frac{\binom{r_1}{s_1}\binom{r_2}{n-s_1}}{\binom{N}{n}}$$

We call this the **hypergeometric law**
Where did this come from?
Answer: Think of each sample as equally likely and count how many there are favorable to the event of interest.

Example 4: (inspection for quality control) A batch of 100 manufactured items is checked by an inspector, who examines 10 items selected at random. If *none* of the 10 items is defective, the batch of 100 is accepted. Otherwise, the batch is subject to further inspection. What is the probability that a batch containing 10 defectives is accepted?

Solution: Number of ways of selecting 10 items of a batch of 100 is

$$N = \binom{100}{10}.$$

All such samples are equally likely.

A = event that the batch is accepted by the inspector. Then A occurs if all 10 items of the selected sample belong to the set of 90 non-defectives.

Number of combinations (samples) favorable to A is:

$$\begin{aligned} N(A) &= \binom{90}{10} \\ P(A) &= \frac{N(A)}{N} \\ &= \frac{\binom{90}{10}}{\binom{100}{10}} \qquad = \frac{90!}{10!80!}\,\frac{10!90!}{100!} \\ &\approx (1 - \frac{1}{10})^{10} \approx \frac{1}{e} \end{aligned}$$

□

Example 5: What is the probability that two cards picked randomly from a full deck are aces?

Solution

$$\begin{aligned} N &= 52 \text{ cards} \\ n &= 4 \quad \text{aces.} \end{aligned}$$

There are $\binom{52}{2}$ equally likely picks.

$N(A) = \binom{4}{2}$ ways are favorable to getting 2 aces.

$$P(A) = \frac{N(A)}{N} = \frac{\binom{4}{2}}{\binom{52}{2}} = \frac{6 \cdot 2}{52 \cdot 51} = \frac{6}{26 \cdot 51} = \frac{1}{221} \quad \square$$

Theorem: Given a population of n elements, let $n_1, n_2, ...n_k$ be positive integers such that $n_1 + n_2 + ...n_k = n$. Then there are precisely

$$N = \frac{n!}{n_1!n_2!...n_k!}$$

ways of partitioning the population into k sub-populations of the prescribed sizes and *order*.

Proof: Order of sub-populations matters.

$$(n_1 = 4, n_2 = 2, n_3, ..., n_k) \neq (n_1 = 2, n_2 = 4, n_3, ..., n_k).$$

Order *within* sub-populations does not matter.

$$
\begin{aligned}
N &= \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-\sum_{i=1}^{k-2} n_i}{n_{k-1}} \\
&= \frac{n!}{n_1!(n-n_1)!}\,\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}\,\frac{(n-n_1-n_2)!}{n_3!(n-n_1-n_2-n_3)!} \\
&\quad \cdots\,\frac{(n-\sum_{i=1}^{k-2} n_i)!}{n_{k-1}!n_k!} \\
&= \frac{n!}{n_1!n_2!...n_k!} \qquad \square
\end{aligned}
$$

Example 6: What is the probability that each of 4 bridge players holds an ace?

$$
\begin{aligned}
n &= 52 \\
n_1 &= n_2 = n_3 = n_4 = 13
\end{aligned}
$$

From the theorem, there are $\frac{n!}{n_1!n_2!n_3!n_4!}$ equally likely deals.

There are $4! = 24$ ways of giving an ace to each player.

Remaining 48 cards can be dealt in $\frac{48!}{12!12!12!12!}$ ways.

Thus these are $24 \cdot \frac{48!}{(12!)^4}$ distinct deals favorable to the desired event.

$$
\begin{aligned}
P(\text{event}) &= 24 \cdot \frac{\frac{48!}{(12!)^4}}{\frac{52!}{(13!)^4}} \\
&\approx 0.105
\end{aligned}
$$

Use Stirling's formula to get this approximation.

Stirling's Formula (following Feller)
Let $a_n = \frac{n!}{(n)^n}$ $n = 1, 2, ...$

$$\begin{aligned}
\frac{a_{n+1}}{a_n} &= \frac{(n+1)!}{(n+1)^{n+1}} \Big/ \frac{n!}{(n)^n} \\
&= \frac{(n+1)n!}{(n+1)^n(n+1)} \frac{(n)^n}{n!} \\
&= \frac{1}{\left(1+\frac{1}{n}\right)^n}
\end{aligned}$$

Let $b_n = n! \left(\frac{e}{n}\right)^n = a_n e^n$

$$\begin{aligned}
log_e \frac{b_{n+1}}{b_n} &= 1 + log_e \frac{a_{n+1}}{a_n} \\
&= 1 - n\, log_e \left(1 + \frac{1}{n}\right) \\
&= 1 - n\left(\frac{1}{n} - \frac{1}{2n^2} + \frac{1}{3n^3} - \cdots\right) \\
&= \frac{1}{2n} - \frac{1}{3n^2} + \cdots
\end{aligned}$$

Let $\beta_n = n! \left(\frac{e}{n}\right)^{n+\frac{1}{2}}$

$$\begin{aligned}
log_e \frac{\beta_{n+1}}{\beta_n} &= 1 - \left(n + \frac{1}{2}\right) log_e \left(1 + \frac{1}{n}\right) \\
&= -\frac{1}{12n^2} + \frac{1}{12n^3} - \cdots < 0.
\end{aligned}$$

Hence $\beta_{n+1} < \beta_n$. We have shown that β_n is a montone decreasing sequence which is bounded below by 0. Thus $\beta = \lim_{n\to\infty} \beta_n$ exists.

In other words,

$$\beta_n = n! \left(\frac{e}{n}\right)^{n+\frac{1}{2}} \;\to\; \beta \qquad \text{a constant}$$

So we can take $n! \sim \beta \cdot (n)^{n+\frac{1}{2}} e^{-(n+1/2)}$. Verify that $\beta = \sqrt{2\pi e}$.

Thus

$$\boxed{n! \sim \sqrt{2\pi}(n)^{n+\frac{1}{2}} e^{-n}} \qquad \text{(I)}$$

There is a slightly better one.

$$\boxed{n! \sim \sqrt{2\pi}(n)^{n+\frac{1}{2}} e^{-n+\frac{1}{12n}}} \qquad \text{(II)}$$

Formulas (I) and (II) are respectively the first and second approximations of Stirling.

ENEE 324 Engineering Probability Lecture 3

Conditioning

In a chance experiment $\mathcal{E}$, occurrence of event A can be influenced by that of event B. For instance, event $A = flooding$, always occurs following event $B = dam - burst$; event $A = flooding$ may have only a small likelihood of occurrence following event $C = light\ shower$.

Interdependence of events influences probabilities. Probabilities computed after obtaining data on one event can be different (higher or lower than) the probabilities computed before such data was available.

In weather forecasting, forecasts for a *Wednesday* made on *Tuesday, 8:00 AM* and *Tuesday, 8:00 PM* differ - additional observations are available during the intervening 12 hour period.

In a medical setting, the presence of a disease in a patient would increase the probability of certain symptoms in the patient. Some symptoms may be present even in the absence of disease. For example, certain symptoms are shared by allergies and by the common cold. In medical diagnosis, a doctor seeks to determine the probability of a certain disease being present given that certain symptoms are observed. This probability may be higher than when the symptoms are not observed. Thus one could say that the observation of a symptom influences the likelihood of a diagnosis of a disease. But this does not imply a causal relationship. Symptoms do not cause diseases!

Data *conditions* probabilities. In fact, practically all probabilities are *conditional probabilities*. We now give a formal definition.

Definition: Let $\mathcal{E}$ be a chance experiment with associated sample space Ω and Boolean algebra $\mathcal{A}$ of interesting events (*i.e.*, subsets of Ω). Given events $A, B \in \mathcal{A}$, the conditional probability of A given B denoted $P(A \mid B)$ is defined to be

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

provided $P(B) \neq 0$ $\square$

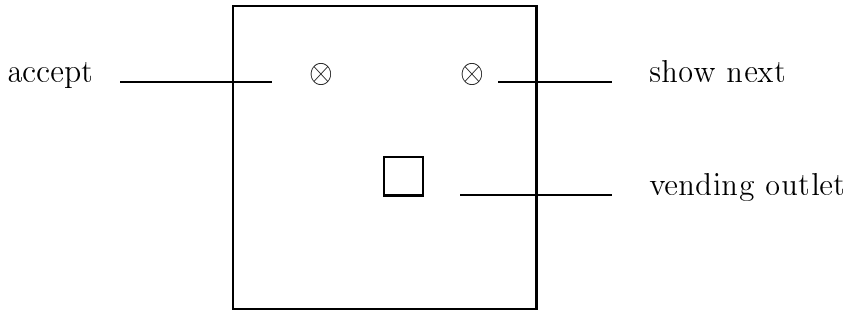
Note: If $P(B) = 0$, then $P(A \mid B)$ is *undefined*.

What is the justification for making such a definition? A bit of counting helps. Suppose experiment $\mathcal{E}$ has n equally probable/likely elementary events. Suppose n_A is the number of such elementary events favorable to the occurrence of event A. Suppose n_B is the number of elementary events favorable to the occurrence of event B. Then $P(B) = n_B/n$. If B actually occurs then the outcomes have to be one of n_B possibilities. Now, for A to occur, one looks at a subset of these that favor A, and these are $n_{(A \cap B)}$ of these. So it makes sense to say,

$$\begin{aligned} P(A \mid B) &= \frac{n_{A \cap B}}{n_B} \\ &= \frac{n_{A \cap B}/n}{n_B/n} \\ &= \frac{P(A \cap B)}{P(B)} \end{aligned}$$

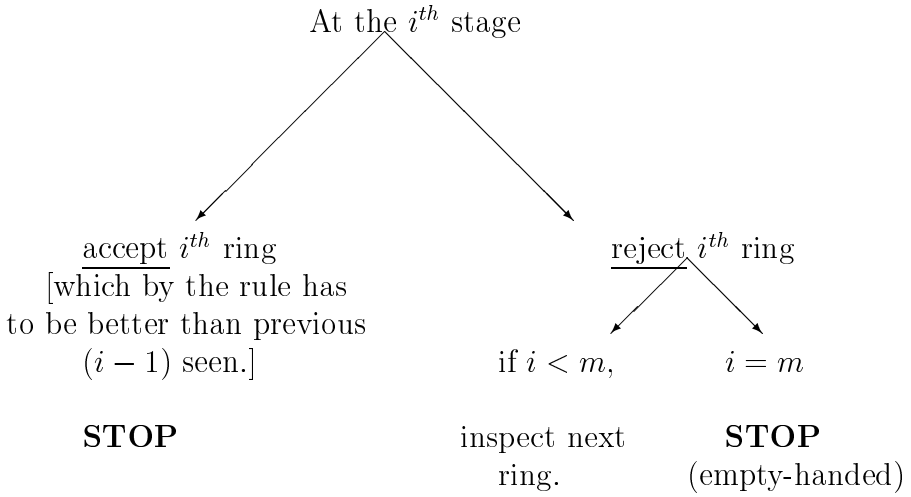
So our definition makes sense.

Example (optimal choice)



Consider a robot merchant that displays diamond rings one at a time to a player. There is a total of m rings. The order of presentation is random. The diamond rings are of differing quality. The player follows the rule: *Never accept a ring inferior to those previously rejected.* The player can press the

show next button or the **accept** button, until there are no rings remaining to be shown.



Question: Suppose the player selects the i^{th} ring. What is the probability of this being the best of all m rings? [This is a prototypical problem of deciding when to commit to a particular course of action or choice.]

Solution:

$$\begin{aligned} B &: = \text{event that the last of } i \text{ inspected} \\ &\quad\ \ \text{rings is the best of those inspected} \\ A &: = \text{event that the } i^{th} \text{ ring} \\ &\quad\ \ \text{is the best of all } m \text{ rings} \end{aligned}$$

We are interested in $P(A \,|\, B)$.
Clearly $A \subset B$. Hence $A \cap B = A$.
Thus,

$$\begin{aligned} P(A \,|\, B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A)}{P(B)} \end{aligned}$$

But

$$P(B) = \frac{(i-1)!}{i!} = \frac{1}{i}$$

Why? $(i-1)!$ is the number of permutations of i distinct things, leaving one, "the best ring," fixed in the i^{th} place.

$$P(A) = \frac{(m-1)!}{m!} = \frac{1}{m}$$

Why? $(m-1)!$ is the number of permutations of m distinct things, leave one, "the best ring," fixed in the i^{th} place. Thus,

$$P(A \mid B) = \frac{1/m}{1/i} = \frac{i}{m}$$

Late commitment is more likely to give you the best deal. □

Example: Toss 2 fair dice, producing the outcome (X, Y). Here, $X, Y \in \{1, 2, 3, 4, 5, 6\}$. Consider the events,

$$\begin{aligned} A &= \{(X,Y) \mid X + Y = 10\} \\ B &= \{(X,Y) \mid X > Y\} \end{aligned}$$

What is the probability $P(A \mid B)$?

$$\begin{aligned} B &= \{(2,1),\ (3,1),\ (3,2),\ (4,1),\ (4,2),\ (4,3), \\ &\quad\ (5,1,),\ (5,2),\ (5,3),\ (5,4),\ (6,1),\ (6,2), \\ &\quad\ (6,3),\ (6,4),\ (6,5)\} \\ n_B &= 15 \end{aligned}$$

Conditioning on B means one can *reduce* the sample space from the full set Ω of all 36 ordered pairs (X, Y) to the smaller subset B.
Within B, there is only one outcome $(6, 4)$ yielding $6 + 4 = 10$, favorable to A. So $n_{(A \cap B)} = 1$.

$$\begin{aligned} P(A \mid B) &= \frac{n_{A \cap B}}{n_B} \\ &= \frac{1}{15}. \end{aligned}$$

But $A = \{(6,4),\ (4,6),\ (5,5)\} \Rightarrow P(A) = \frac{3}{36} = \frac{1}{12}$

Also, $\dfrac{P(A \cap B)}{P(B)} = \dfrac{n_{A\cap B}/n}{n_B/n} = \dfrac{1/36}{15/36} = \dfrac{1}{15}$, as we expect.

Thus, $P(A)$ is different from $P(A \mid B)$.

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{1/36}{1/12} = \frac{1}{3}.$$

Properties of Conditional Probability

(1) $0 \leq P(A \mid B) \leq 1$.

Proof: $\emptyset \subset A \cap B \subset B$. Hence, $P(\emptyset) \leq P(A \cap B) \leq P(B)$.
It follows that $0 \leq \dfrac{P(A \cap B)}{P(B)} \leq 1$ $\square$.

(2) $A \cap B = \emptyset$. Then $P(A \mid B) = 0$.

(3) $B \subset A$, then $P(A \mid B) = 1$.

Proof: $B \subset A \Rightarrow B \cap A = B$. Thus,

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(B)}{P(B)} \\ &= 1 \qquad\qquad \square \end{aligned}$$

(4) Given $A_1, A_2, \cdots A_k$ disjoint, and $A = \cup_{i=1}^{k} A_i$. Then, $P(A \mid B) = \sum_{i=1}^{k} P(A_i \mid B)$.

Proof: $A \cap B = \left(\cup_{i=1}^{k} A_i\right) \cap B = \cup_{i=1}^{k} (A_i \cap B)$.
Since A_i are disjoint, $A_i \cap B$ are also disjoint. Thus,

$$P(A \cap B) = P((\cup_{i=1}^{k} A_i) \cap B)$$

$$
\begin{aligned}
& = \sum_{i=1}^{k} P(A_i \cap B). \\
\text{Hence } P(A \mid B) & = \sum_{i=1}^{k} \frac{P(A_i \cap B)}{P(B)} \\
& = \sum_{i=1}^{k} P(A_i \mid B) \qquad \square
\end{aligned}
$$

We have shown that conditional probability of a disjoint union is the sum of the conditional probabilities. This demonstrates the parallel to the addition axiom for probabilities.

Total probability formula: Suppose $\cup_{i=1}^{k} B_i = \Omega, \quad B_i \cup B_j = \emptyset$. (We call this a partition of Ω.)

Then, $P(A) = \sum_{i=1}^{k} P(A \mid B_i) P(B_i)$

Proof:

$$
\begin{aligned}
A & = A \cap \Omega \\
& = A \cap (\cup_{i=1}^{k} B_i) \\
& = \cup_{i=1}^{k} (A \cap B_i)
\end{aligned}
$$

It follows that,

$$
\begin{aligned}
P(A) & = P\left(\cup_{i=1}^{k} (A \cap B_i)\right) \\
& = \sum_{i=1}^{k} P(A \cap B_i) \quad (\text{because } (A \cap B_i) \cap (A \cap B_j) = \emptyset) \\
& = \sum_{i=1}^{k} P(A \mid B_i)\, P(B_i)\,. \qquad \square
\end{aligned}
$$

Now,

$$\begin{aligned} P(B_i \,|\, A) &= \frac{P(A \cap B_i)}{P(A)} \\ &= \frac{P(A \,|\, B_i)P(B_i)}{P(A)} \\ &= \frac{P(A \,|\, B_i)P(B_i)}{\sum_{j=1}^{k} P(A \,|\, B_j)P(B_j)} \end{aligned}$$

Bayes' Formula (an inversion formula)

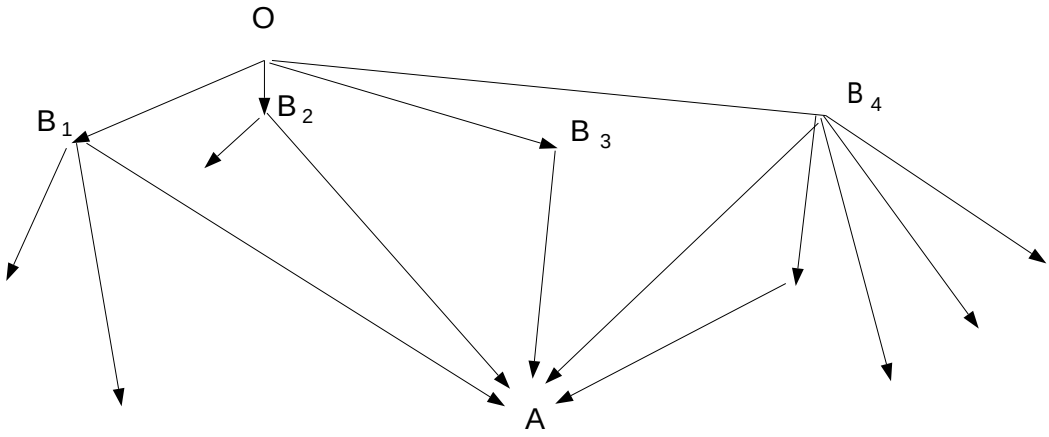
$$\boxed{P(B_i \,|\, A) = \frac{P(A \,|\, B_i)\, P(B_i)}{\sum_{j=1}^{n} P(A \,|\, B_j)\, P(B_j)}}$$

This formula has its origins in two very famous 18^{th} century papers by Reverend Thomas Bayes.

(*i*) "An essay toward solving a problem in the doctrine of chance," Philosophical Transactions of the Royal Society, 1763, pp 370-418 (reprinted in *Biometrika* 45:293-315, 1958).
(*ii*) "A letter on asymptotic series ...," Philosophical Transactions of the Royal Society, 1763, pp 269-271.

This is the most important formula in our subject. Various other versions prove to be versatile in telling us how to update or evolve probabilities using data. This is somewhat like Newton's $m\ddot{x} = f$, telling us how to evolve particle motions.

Example (Hiking): Hiker leaves O, choosing one of the roads OB_1, OB_2, OB_3, OB_4 at random. At *each subsequent fork*, he again chooses a road at random. What is the probability of the hiker arriving at point A?


O
B 1
B 2
B 3
B 4
A

$$
\begin{aligned}
P(B_k) &= \frac{1}{4}, \quad k = 1, 2, 3, 4 \\
P(A \mid B_1) &= \frac{1}{3} \\
P(A \mid B_2) &= \frac{1}{2} \\
P(A \mid B_3) &= 1 \\
P(A \mid B_4) &= \frac{2}{5} \\
P(A) &= \frac{1}{4} \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot \frac{2}{5} \\
&= 67/120 \quad \text{(by total probability formula).}
\end{aligned}
$$

We can ask a related question. If the hiker arrives at A, what is the probability that he passed through B_2? This is just $P(B_2 \mid A)$ and Bayes' formula gives us

$$
\begin{aligned}
P(B_2 \mid A) &= \frac{P(A \mid B_2)\, P(B_2)}{\sum_{k=1}^{4} P(A \mid B_k)\, P(B_k)} \\
&= \frac{1/2 \cdot 1/4}{67/120} \\
&= \frac{120}{8 \cdot 67} \\
&= \frac{15}{67}. \quad \square
\end{aligned}
$$

ENEE 324 Engineering Probability Lecture 4

Applications of Bayes' Theorem

Example: There are 10 urns, 9 of which are of type I and 1 of type II. Urn of type I carries 2 white balls and 2 black balls. Urn of type II carries 5 white balls and 1 black ball.

If a ball drawn randomly from a randomly chosen urn turns out to be *white*, then what is the probability that the chosen urn is of type II? This is a model of an inference problem.

Solution

$A :=$ ball drawn is white
$B_1 :=$ urn is of type I
$B_2 :=$ urn is of type II
B_1 and B_2 are disjoint events and define a partition $\Omega = B_1 \cup B_2$.

$$
\begin{array}{rcl}
P(B_2|A) & = & \frac{P(A|B_2)\ P(B_2)}{P(A|B_1)\ P(B_1)+P(A|B_2)\ P(B_2)} \\
P(B_1) & = & 9/10\ ;\ P(B_2) \qquad = \quad 1/10 \qquad \text{Prior probabilities} \\
P(A|B_1) & = & 2/4 \qquad = \quad 1/2 \\
P(A|B_2) & = & 5/6
\end{array}
$$

$$
\begin{aligned}
P(B_2|A) &= \frac{5/6 \cdot 1/10}{1/2 \cdot 9/10 + 5/6 \cdot 1/10} \\
&= \frac{5}{27+5} = \frac{5}{32} \qquad \square
\end{aligned}
$$

Statistical Independence; The idea that two phenomena have nothing to do with each other has a key role in probability theory.
Definition We say that in an experiment $\mathcal{E}$, two events A and B are *statistically independent* if,

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

Imagine a long series of trials, each of which involves carrying out two experiments $\mathcal{E}_1$ and $\mathcal{E}_2$, where only $\mathcal{E}_1$ leads to A_1 and only $\mathcal{E}_2$ leads to A_2.

If n = total number of trials, $n(A_1 \cap A_2)$ = number of trials leading to occurence of A_1 and A_2, then

$$\begin{aligned} P(A_1 \cap A_2) &\sim \frac{n(A_1 \cap A_2)}{n} \\ P(A_2) &\sim \frac{n(A_2)}{n} \\ P(A_1) &\sim \frac{n(A_1)}{n}. \end{aligned}$$

On the other hand

$$\begin{aligned} P(A_1 \cap A_2) &\sim \frac{n(A_1 \cap A_2)}{n} \\ &= \frac{n(A_1 \cap A_2)}{n(A_2)} \cdot \frac{n(A_2)}{n} \\ &\sim P(A_1) \cdot P(A_2) \end{aligned}$$

The following example illustrates statistical independence and related subtleties. Throw two dice resulting in the outcomes (X, Y).
Let A_1 : event that X is odd
A_2 : event that Y is odd
A_3 : event that $X + Y$ is odd.

Clearly, A_1 and A_2 are independent.

$$\begin{array}{rcl} P(A_1) & = & \frac{1}{2} \qquad\qquad = \quad P(A_2) \\ P(A_3) & = & \text{Prob } \{\text{X odd and Y even}\} \\ & & +\text{Prob } \{\text{X even and Y odd}\} \\ & = & \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2} \\ & = & \frac{1}{2} \end{array}$$

$$\begin{array}{rcl} P(A_3|A_1) & = & \text{Prob } \{Y \text{ even}\} \\ & = & \frac{1}{2} \\ P(A_3|A_2) & = & \text{Prob } \{X \text{ even}\} \\ & = & \frac{1}{2} \end{array}$$

$$\Rightarrow \quad P(A_3|A_1) = P(A_3) = P(A_3|A_2)$$

Thus A_3 and A_1 are independent *and* A_3 and A_2 are independent. □

Definition: Given events $A_1, A_2, \ldots, A_n$, we say these are *mutually independent* if:

$$\begin{array}{rcl} P(A_i \cap A_j) & = & P(A_i)\cdot P(A_j) \\ P(A_i \cap A_j \cap A_k) & = & P(A_i)P(A_j)P(A_k) \\ & \vdots & \\ P(A_1 \cap A_2 \cap \cdots \cap A_n) & = & P(A_1)P(A_2)\cdots P(A_n). \end{array}$$

In the previous example, the events A_1, A_2, A_3 are *not* mutually independent, even though they are pairwise independent, because

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

but

$$P(A_1)P(A_2)P(A_3) = \left(\tfrac{1}{2}\right)^3.$$

Probability Trees: When an experiment is of a *sequential nature*, it is often convenient, especially for purposes of calculation, to represent the experiment graphically by a *probability tree*. It is a *rooted tree* and the vertices represent *outcomes/events* of the experiment. The edges are labelled by the *conditional probabilities* required to *descend* from a given vertex to an adjacent one. The probability associated with the event corresponding to a vertex is obtained by taking under consideration the product of the probabilities labelling the edges forming the unique path between the vertex, and the root of the tree.

Example: Flipping a coin three times:

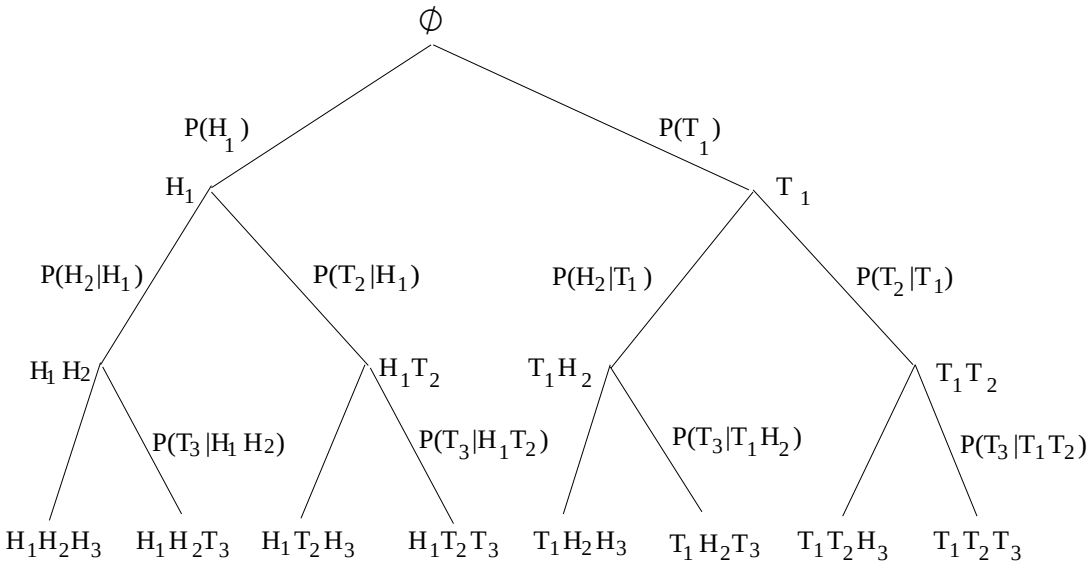


fig. 1

Corresponding Pascal's Triangle

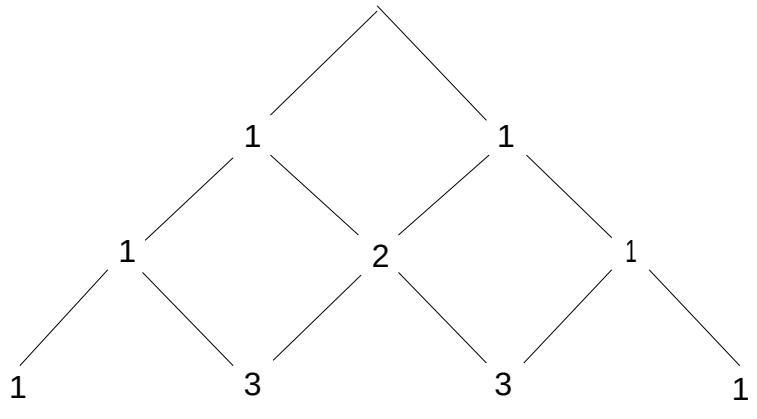


fig. 2

Probability trees may also be infinite. We give an example below.

Example: Player A flips a fair coin. If the outcome is a head, he wins; if the outcome is a tail, player B flips. If B's flip is a head, he wins; if not, player A flips the coin again. This process is repeated (*ad infinitum*, if necessary) until somebody wins. What is the probability that A wins?

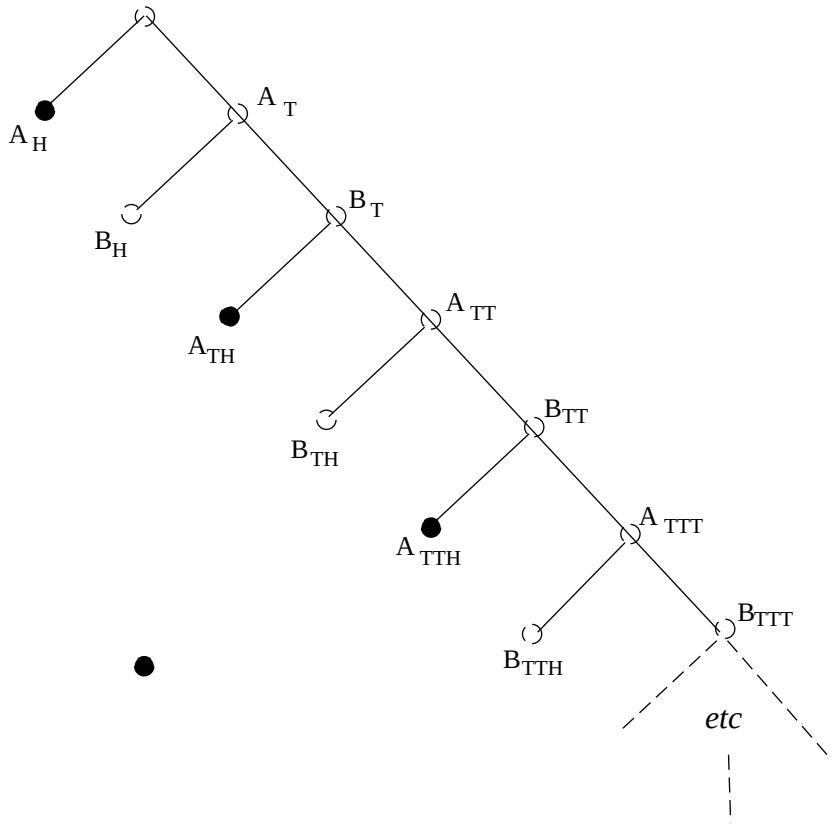


fig. 3

For the probability tree above, the darkened vertices correspond to the *elementary events* for which A wins. Since the probability represented by each branch of the tree is $1/2$, we have:

$P\{A \text{ wins}\}$ calculated via sampling with replacement

$$
\begin{aligned}
&= P\{A_H\} + P\{A_{TH}\} + P\{A_{TTH}\} + \cdots \\
&= \frac{1}{2} + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^5 + \cdots \\
&= \frac{1}{2}\left[1 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^6 + \cdots\right] \\
&= \frac{1}{2}\,\frac{1}{1-\left(\frac{1}{2}\right)^2} \\
&= \frac{1}{2}\,\frac{1}{1-1/4} \\
&= \frac{2}{3}
\end{aligned}
$$

There is a *big advantage for A* to flip first.

Gambler's Ruin – (Application of Total Probability Law)

Example: (1) Toss coin. Call correctly, win 1 dollar. Call wrongly, loose 1 dollar.

Payoff Matrix

Call \ Toss	Head	Tail
Head	1	−1
Tail	−1	1

Fig. 4

Initial Capital = x dollars and x is a positive integer.

STRATEGY	PLAY UNTIL EITHER :	
	↙	↘
	Win m Dollars (*i.e.* has a total of m dollars)	Lose Shirt (**RUIN**)

Question: What is the probability $p(x)$ of ruin?

$$\begin{array}{lcl} A & = & \text{RUIN} \\ B_1 & = & \text{Win first call} \;=\; p \\ B_2 & = & \text{Lose first call} \;=\; (1-p) \end{array}$$

$$\begin{array}{rcll} P(A) & = & P(A|B_1)\,P(B_1) + P(A|B_2)\cdot P(B_2) & \\ p(x) & = & p(x+1)\cdot\frac{1}{2} + p(x-1)\,\frac{1}{2} & 1 \le x \le m-1 \\ & = & p(x+1)\,p + p(x-1)\cdot(1-p) & \end{array}$$

$$\text{B.C.}\begin{cases} p(0) = 1 \\ p(m) = 0 \end{cases}$$

$$\begin{array}{rcll} p(x) & = & C_1 + C_2 x & \text{is the solution} \\ C_1 & = & 1 & C_1 + C_2 m = 0 \end{array}$$

Hence:

$$\boxed{p(x) = 1 - x/m} \qquad 0 \le x \le m$$

$$\boxed{\text{If } p \neq 1/2 \text{ the solution is } \textit{not} \text{ linear}}$$

Example [Matching]:

n distinct items to be matched against n distinct cells. What is the probability of at least 1 match?

Solution:

$A_k :=$ event that k^{th} item is matched (we don't care about the rest)
$P^{(n)} =$ Probability of at least 1 match

$$
\begin{aligned}
&= P(\cup_{k=1}^{n} A_k) \\
&= \sum_{i=1}^{n} P(A_i) - \sum_{i<j=2}^{n} P(A_i \cap A_j) \\
&\qquad + \sum_{i<j<k=3}^{n} P(A_i \cap A_j \cap A_k \cdots) \\
&\qquad + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots A_n) \\
&= P_1 - P_2 + P_3 \cdots \pm P_n
\end{aligned}
$$

$$
P(A_{i_1} \cap A_{i_2} \cdots \cap A_{i_m}) = \frac{(n-m)!}{n!}
$$

$$
\begin{aligned}
P_m &= \sum_{a \le i_1 < i_2 < \cdots < i_m \le n} P(A_{i_1} \cap A_{i_2} \cdots \cap A_{i_m}) = \binom{n}{m} \frac{(n-m)!}{n!} \\
&= \frac{n!}{(h-m)!m!} \frac{(n-m)!}{n!} = \frac{1}{m!} \\
P^{(n)} &= 1 - \tfrac{1}{2!} + \tfrac{1}{3!} - \tfrac{1}{4!} \cdots + (-1)^{n+1} \tfrac{1}{n!}
\end{aligned}
$$

Special Cases: Number of permutations of n things in which there is at least 1 match $= P^{(n)} \cdot n!$.

$$
\begin{aligned}
n = 3 \qquad P^{(n)} n! &= 6 \times \left(1 - \tfrac{1}{2} + \tfrac{1}{6}\right) \\
&= \underline{\underline{4}} \\
n = 4 \qquad P^{(n)} n! &= 24 \left(1 - \tfrac{1}{2} + \tfrac{1}{6} - \tfrac{1}{24}\right) \\
&= \underline{\underline{15}}
\end{aligned}
$$

Problem: Given any n events, $A_1,\ A_2, \cdots A_n$ prove that the probability of exactly $m \le n$ events occurring is

$$
P = P_m - \binom{m+1}{m} P_{m+1} + \binom{m+2}{m} P_{m+2} \cdots \ \pm \binom{n}{m} P_n
$$

where

$$
P_k = \sum_{1 \le i_1 < i_2} , \cdots i_k \le n P\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right)
$$

Good Example of Bayesian Inference

Sometimes the application of Bayes' theorem may yield results that appear counter-intuitive.

Example: A laboratory test is developed to detect *mononucleosis* (mono, for short). The probability that a person selected at random has mono is 0.005. If a person has mono, 95% of the time he test will be positive. If a person does not have mono, the test will be positive only 4% of the time. These circumstances are described by the *binary channel* shown in *Figure 5*.

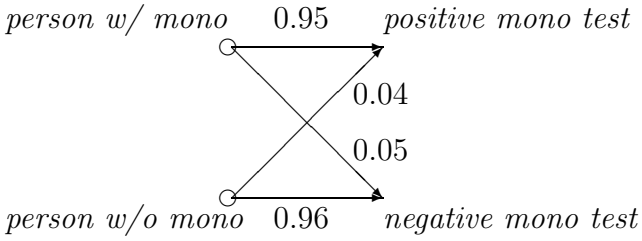


Fig. 5

What is the probability that a person has mono conditioned on the fact that his test came out positive?

$$\begin{aligned} M &= \text{person has mono} \\ T &= \text{positive mono test} \end{aligned}$$

$$\begin{array}{lll} \text{prior probabilities} & & \text{conditional probabilities} \\ \left\{ \begin{array}{l} P(M) \\ P(\bar{M}) \end{array} \right. & \begin{array}{l} = \\ = \end{array} \begin{array}{l} 0.005 \\ 0.995 \end{array} & \left\{ \begin{array}{l} P(T|M) = 0.95 \\ P(T|\bar{M}) = 0.04 \end{array} \right. \end{array}$$

Then, by Bayes' theorems,

$$\begin{aligned}
\text{a posteriori probability} \\
\{P(M|T) &= \frac{P(T|M)P(M)}{P(T|M)P(M) + P(T|\bar{M})P(\bar{M})} \\
&= \frac{0.95 \times 0.005}{0.95 \times 0.005 + 0.04 \times 0.995} \\
&= \frac{0.00475}{0.00475 + 0.0398} \\
&= \frac{0.00475}{0.04455} \\
&= \underline{\underline{0.107}} \qquad !
\end{aligned}$$

Thus the test might give rise to too many false alarms. How to improve? Bring *down the probability* $P(T|\bar{M})$ from 0.04. Improve the test.

A useful form of Bayes' theorem is obtained by conditioning in more than one event.

Let $H :=$ hypothesis (e.g. *a disease* event),
Let $E :=$ evidence of data (e.g. *image data* event), and
Let $C :=$ context (e.g. *age group*). Then,

$$P(H|E \cap C) = \frac{P(E|H \cap C) \cdot P(H|C)}{P(E|C)}$$

To see this, observe that the r.h.s. above

$$\begin{aligned}
&= \frac{P(E \cap H \cap C)}{P(H \cap C)} \cdot \frac{P(H \cap C)}{P(C)} \cdot \frac{P(C)}{P(E \cap C)} \\
&= \frac{P(E \cap H \cap C)}{P(E \cap C)} \\
&= \frac{P\left(H \cap (E \cap C)\right)}{P(E \cap C)}
\end{aligned}$$

$$
\begin{aligned}
&= P(H|E \cap C) \\
&= l.h.s
\end{aligned}
$$

Engineering Probability Lecture 5
Random Variables

In many chance experiments the outcomes are only *indirectly* known through some measurement or *observable*. It is a bit like getting a read-out from an instrument. The read-out function does not produce the same value every time you do the experiment. This is the essence of the random or chance nature of the experiment. We call such observable functions *random variables*.

Definition: Given an experiment $\mathcal{E}$ with sample space Ω, a random variable associated to the experiment is a function $X : \Omega \to \mathbb{R}$. [Initially, we confine attention to real-valued observables.]

The following picture captures the main idea.

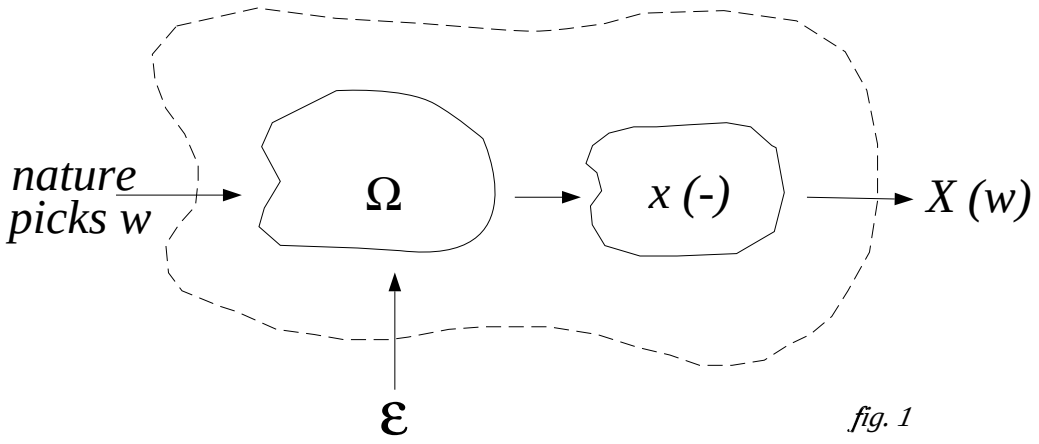


fig. 1

If the experimenter can access only the value of X, it is as if there is a (dotted) box as in *figure 1* and nothing inside the box is directly accessible.

Example 1:
$\mathcal{E}$:= insert a light bulb in a socket, switch on the light, wait till it burns out. Record when this happens

Ω := possible dates and times of burn out.

$X : \Omega \to \mathbb{R}_+$ = non-negative real numbers

$\omega \longmapsto X(\omega)$ = lifetime of bulb

$= \omega -$ (time when the bulb was switched on) $\square$

Example 2:
$\mathcal{E}$:= Pair of coin tosses

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

Suppose we make up X and Y as follows

$$X(\omega) = \begin{cases} \frac{1}{2} & \text{if } \omega\epsilon\{HH, TT\} \\ -\frac{1}{2} & \text{if } \omega\epsilon\{TH, HT\} \end{cases}$$

$$Y(\omega) = \begin{cases} 1 & \text{if } \omega\epsilon\{HH\} \\ 0 & \text{if } \omega\epsilon\{HH, HT\} \\ \frac{1}{2} & \text{if } \omega\epsilon\{TH, TT\} \end{cases}$$

Only X is a random variable, and Y is not. Why? Y is **not** a function. So it cannot be a random variable. $\square$

To avoid mixing up a function and its value, we reserve uppercase letters for functions that are random variables. A value $X(\omega)$ is denoted as x.

Example 3: Suppose we have a sequence of n tosses of a given coin. For each toss we have $\Omega = \{H, T\}$. For the i^{th} toss, let $X_i : \Omega \to \mathbb{R}$ be defined by

$$X_i(\omega) = \begin{cases} 1 & \omega = H \\ 0 & \omega = T \end{cases}$$
$$i = 1, 2, \ldots, n.$$

Thus we have n random variables associated to the entire sequence of coin tosses.

We can take the entire sequence of experiments as *one giant experiment* $\widetilde{\mathcal{E}}$ with sample space

$$\begin{aligned}\widetilde{\Omega} &= \{HH\cdots H, \quad HTH\cdots H, \cdots\} \\ &= \text{set of sequences in H and T, of length n.}\end{aligned}$$

Then $X : \widetilde{\Omega} \rightarrow \mathbb{R}$ may be defined as,

$$\begin{aligned}X(\omega) &= \text{total number of times H came up} \\ &= \sum_{i=1}^{n} X_i(\omega_i)\end{aligned}$$

where ω_i = outcome of just the i^{th} coin toss and X_i as before. Since X aggregates the X_i, it is "less informative" than the collection of X_i. □

Returning to *fig. 1*, one can stay entirely on the outside of the dotted line by defining the events in the range of X.

Definition: Let $\mathcal{E}$ be an experiment, Ω the sample space and $X : \Omega \rightarrow \mathbb{R}$ a random variable. Let R_X = range of X = set of values taken by $X(\omega)$ as ω varies in Ω. One can work with an algebra of events, denoted as $\mathcal{R}_X$, analogous to the Boolean algebra of interesting events (subsets of Ω) introduced before. Whenever $B \epsilon \mathcal{R}_X$, *i.e.* $B \subset R_X$ is such that,

$$A = \{\omega \epsilon \Omega \mid \quad X(\omega) \epsilon B\} \stackrel{\triangle}{=} X^{-1}(B) \epsilon \mathcal{A},$$

one can define the probability of occurrence of B to be simply

$$P(A) = P\left(X^{-1}(B)\right).$$

See *fig. 2*

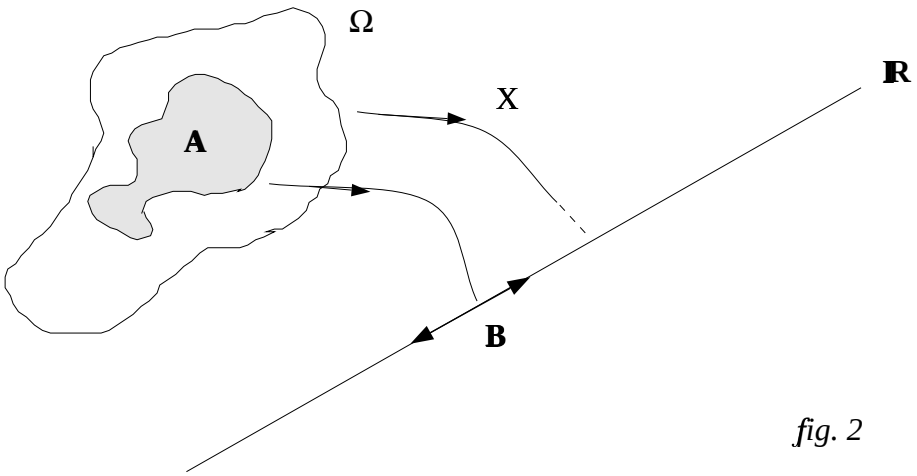


fig. 2

To keep track of how A is related to B, we write:

$$\boxed{P_X(B) \stackrel{\triangle}{=} P(A) = P\left(X^{-1}(B)\right)}$$

Definition: A random variable X is said to be discrete if R_X is a finite or countable set.

Example 4: Let N_t = number of α particles emitted by a gram of Radium in a time interval $[0, t]$. Then N_t is a discrete random variable.

For a discrete random variable X with $R_X = \{x_1, x_2, x_3 \cdots\}$, one associates a *probability mass function* given by,

$$p(x_i) = \text{Prob}\,\{X = x_i\}.$$

It is standard to use the short-hand p_i for $p(x_i)$. The probability mass function is simply the sequence $\{p_1, p_2, p_3, \cdots\}$.

Note that $p_i \geq 0$ and $\displaystyle\sum_{i=1}^{\infty} p_i = 1$

Example 5: (**Geometric** Distribution) Consider an experiment $\mathcal{E}$ in which one tosses a coin repeatedly until it turns up 'Head.' Assume that the successive tosses are independent and the probability of a 'Head' in a single toss

$= p$ (coin may be biased, so p need not be $= 1/2$). Then

$$\Omega = \{H,\ TH,\ TTH,\ TTTH, \cdots\}$$

is a space of sequences. Consider the random variable,

$$\begin{array}{rcl} X : \Omega & \rightarrow & \mathbb{R} \\ \omega & \longmapsto & \text{length of } (\omega) \\ \text{Then } R_X & = & \{1, 2, 3, \ldots\} \end{array}$$

The probability mass function in this case is given by,

$$\begin{array}{rcl} p_k & = & \text{Prob}\{X = k\} \\ & = & \text{Prob}\{\text{first}(k-1)\text{tosses come up TAIL and } k^{th}\text{comes up HEAD}\} \\ & = & (1-p)^{k-1} \cdot p \qquad\qquad k = 1, 2, 3, \cdots \end{array}$$

The probability mass function $\{p_1, p_2, \ldots\}$ is a geometric sequence and hence X is called a *geometric random* variable. $\square$

NOTE: Since $\left(1 + r + r^2 + \cdots + r^{N-1}\right) = \dfrac{1 - r^N}{1 - r}$, it follows that,

$$\begin{array}{rcl} \displaystyle\sum_{k=1}^{\infty} p_k & = & lim_{N\to\infty} \displaystyle\sum_{k=1}^{N} p_k \\ & = & lim_{N\to\infty} \displaystyle\sum_{k=1}^{N} (1-p)^{k-1} p \\ & = & p \ \ lim_{N\to\infty} \dfrac{\left(1 - (1-p)^N\right)}{(1-(1-p))} \\ & = & p \cdot \dfrac{1}{1-1+p} \cdot \qquad (\text{since}(1-p) < 1) \\ & = & 1 \qquad \text{as it should be.} \end{array}$$

Example 6: (**Binomial** Random Variable) Consider an experiment involving n successive, independent coin tosses, with a coin as in Example 5. The sample space $\Omega = \{HH\ldots H, HTHT\ldots H, \ldots\}$ is a space of 2^n sequences. The random variable X is defined by $X(\omega) =$ number of head in ω. Clearly,

$$\begin{aligned} R_X &= \{0, 1, 2, \ldots, n\}, \text{and} \\ p_k &= \text{Prob}\{X = k\} \\ &= \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, 2, \ldots, n. \quad \square \end{aligned}$$

What happens when $n \to \infty,\ p \to 0,$ but $np \to a$ in Example 6?

$$\begin{aligned} \frac{p_k}{p_{k-1}} &= \frac{n!}{k!(n-k)!} \frac{(k-1)!(n-k+1)!}{n!} \frac{p^k(1-p)^{n-k}}{p^{k-1}(1-p)^{n-k+1}} \\ &= \frac{n-k+1}{k} \frac{p}{1-p} \\ &= \frac{np-(k-1)p}{k(1-p)} \to \frac{a}{k} \\ &\qquad \text{as} \quad n \to \infty,\ p \to \infty,\ np \to a \end{aligned}$$

Thus

$$\begin{aligned} p_k \to \left(\frac{a}{k} \cdot \frac{a}{k-1} \cdots \frac{a}{1}\right) p_0 &= \frac{a^k}{k!} p_0 \\ p_0 &= (1-p)^n \sim \left(1 - \frac{a}{n}\right)^n \to e^{-a} \text{as} \quad n \to \infty. \end{aligned}$$

Thus the sequence $p_k \to e^{-a} \dfrac{a^k}{k!}$ $\square$.

Definition: The random variable X with $R_X = \{0, 1, 2, \cdots\}$ and probability mass function given by,

$$\begin{aligned} p_k &= \text{Prob}\ (x = k) \\ &= e^{-a} \frac{a^k}{k!} \qquad k = 0, 1, 2, \cdots \end{aligned}$$

is called a **Poisson** random variable. We have shown that the binomial $\rightarrow$ Poisson.

Example 7: (Lottery) How many lottery tickets should I buy to make the probability of winning at least ϵ?

Solution: In a lottery, out of a total of N tickets, there are M winning tickets. A purchase is a Bernoulli trial with probability $p = \frac{M}{N}$. A set of n purchases is a sequence of n Bernoulli trials, with Prob (holding k winning tickets) $= \frac{a^k}{k!}e^{-a}$ (approximately) where $a = np = \frac{nM}{N}$. We are asking to choose n

$$\begin{array}{rrcl} \text{so that,} & \epsilon & \leq & 1 - P(0) = 1 - e^a, \\ \text{equivalently,} & e^{-a} & \leq & 1 - \epsilon, \text{or } \; e^{-\frac{nM}{N}} \leq 1 - \epsilon, \\ \text{equivalently,} & \frac{-nM}{N} & \leq & ln(1-\epsilon), \\ \text{equivalently,} & \frac{nM}{N} & \geq & -\ln(1-\epsilon), \\ \text{equivalently,} & n & \geq & -\frac{N}{M} ln(1-\epsilon) \end{array} \qquad \square$$

Cumulative Distribution Function

As already discussed, it is possible to work with R_X instead of Ω. Similarly, instead of probabilitites defined on an algebra $\mathcal{A}$ of interesting events, one can work with an equivalent concept of the cumulative distribution function.

Definition: Let X be a random variable. The cumulative distribution function associated to X denoted as the c $\cdot$ d $\cdot$ f$\cdot$ $F_X(\cdot)$ is defined by

$$F_X(x) = \text{Prob}\,\{\omega \epsilon \Omega : X(\omega) \leq x\} \quad \square$$

Knowing the c $\cdot$ d $\cdot$ f$\cdot$ we can determine interesting probabilities.
Suppose $x_1 < x_2$
Prob $\{\omega \epsilon \Omega : x_1 < X(\omega) \leq x_2\}$ = Prob $\{\omega \epsilon \Omega : X(\omega) \leq x_2\}$−Prob $\{\omega \epsilon \Omega : X(\omega) \leq x_1\}$

This follows from the disjoint union,

$$\{\omega : X(\omega) \leq x_2\} = \{\omega : X(\omega) \leq x_1\} \cup \{\omega : x_1 < X(\omega) \leq x_2\}$$

For a discrete random variable X with sample space Ω and range

$$R_X = \{x_1, x_2, x_3, \cdots\}$$

where the x_i's are ordered $x_k < x_{k+1} \qquad k = 1, 2, \ldots,$
and probability mass function given by,

$$p_k = P(X = x_k) = \text{Prob}\{\omega : X(\omega) = x_k\},$$

the cumulative distribution function is given by,

$$F_X(x) = \sum_{k=1}^{\infty} p_k U(x - x_k)$$

Here $U(x - x_k)$ is the unit step function

$$U(x - x_k) \quad = \quad \begin{cases} 0 & x < x_k \\ 1 & x \geqslant x_k \end{cases}$$

This follows directly from the definition. The resulting picture of the c · d · f· is that of a staircase function.

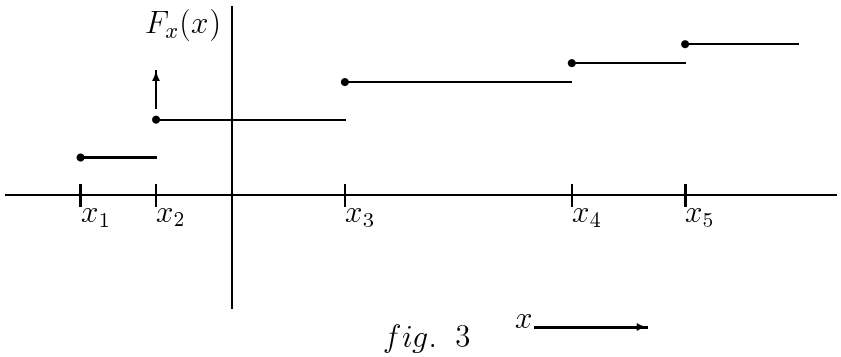


fig. 3

The jump at $x = x_k$ is p_k.

Irrespective of whether a random variable is of the discrete variety or something else, the very definition of the c · d · f· leads to some basic properties.

$$
\begin{array}{lrcll}
(i) & F_X(x) & \geqslant & 0 & \forall x \epsilon \mathbb{R} \\
(ii) & x \overset{lim}{\to} -\infty \ F_X(x) & = & 0 & \\
 & x \overset{lim}{\to} +\infty \ F_X(x) & = & 1 & \\
(iii) & x, y \text{ such that } x & \leq & y \Rightarrow F_X(x) \leq F_X(y) & \\
 & & & \text{(monotone increasing property)} & \\
(iv) & F_X \text{ is } \underline{\text{right continuous}}, \text{ i.e.} & & & \\
 & F_X(x) & = & h \overset{lim}{\downarrow} 0_+ \ F_X(x+h) &
\end{array}
$$

Only the last in the list above requires some extra concepts which we consider beyond our scope.

The random variable in Example 1, the lifetime T of a bulb, it <u>not</u> a discrete random variable. So, we do not speak of a probability mass function in this case. Instead, one can start directly from a cumulative distribution function as a given (from physics or from experimental data). The following c · d · f· seems natural,

$$
\begin{array}{rrcll}
 & F_T(t) & = & \begin{cases} 1 - e^{-\lambda t}\,, & t \geqslant 0 \\ 0 & & t < 0 \end{cases} & \\
\text{Thus} & P(T > t) & = & 1 - P(T \leq t) & t > 0 \\
 & & = & 1 - (1 - e^{-\lambda t}) & \\
 & & = & e^{-\lambda t} &
\end{array}
$$

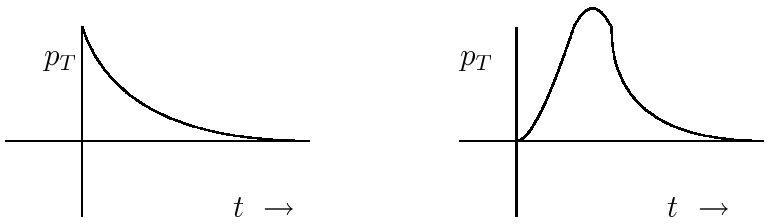
Here $\lambda > 0$ is a parameter. The formula above implies that *long lifetimes are highly unlikely*. It implies more. What is the probability that the bulb will last an extra time δ, given that it has lasted until time a ? This is,

$$
\begin{aligned}
&P\{T > a+\delta | T > a\} \\
&\qquad = \frac{P(\{T > a+\delta\} \cap \{T > a\})}{P\{T > a\}} \\
&\qquad = \frac{P\{T > a+\delta\}}{P\{T > a\}} \\
&\qquad = \frac{e^{-\lambda(a+\delta)}}{e^{-\lambda a}} \\
&\qquad = e^{-a\delta} \qquad\qquad \square
\end{aligned}
$$

The above formula suggests that your (or bulb's) present age is immaterial to how much longer you (or the bulb) will live. This *memory-less property* may be unrealistic and one may want a different one. Notice that in the present instance, one can differentiate F_T to obtain (fixing-up things at 0), the *density*

$$
\begin{aligned}
p_T(t) &= \frac{dF_T(t)}{dt} \\
&= \begin{cases} \lambda e^{-\lambda t} & t \geqslant 0 \\ 0 & t < 0 \end{cases}
\end{aligned}
$$

The graph of p_T is as in *figure 4.a.* We call T an **exponential** randon vaiable.



(a) (b)

fig. 4

A better choice of density might be as in *fig. 4(b)*. We now are led to the following:

Definition: A random variable X is said to be *continuous* if the associated cumulative distribution can be expressed as

$$F_X(x) = \int_{-\infty}^{x} p_X(y)dy$$

for a suitable piecewise continuous function p_X which we will call the *probability density function* ($\mathrm{p \cdot d \cdot f \cdot}$) of X. It follows that

$$p_X(x) = \frac{dF_X}{dx}$$

and by the properties of the $\mathrm{c \cdot d \cdot f \cdot}$, we conclude

$$\begin{aligned} p_X(x) &\geqslant 0 \qquad\qquad x \epsilon \mathbb{R} \\ \int_{-\infty}^{\infty} p_X(y)dy &= 1 \\ \mathrm{Prob}\{a < X \leq b\} &= \int_a^b p_X(y)dy \end{aligned}$$

NOTE: Even for discrete random variables, if one is willing to work with Dirac delta functions, we can use $\delta(x - x_k) = \frac{d}{dx}U(x - x_k)$ to write

$$p_X(x) = \sum_{k=1}^{\infty} p_k \delta(x - x_k)$$

This is a convenient mnemonic but not essential.

Example 8: (**Uniform** Distribution) Suppose a point ξ is "marked at random" in an interval $[a, b]$. This means that the probability of the mark falling in the sub-interval $[\xi',\ \xi''] \subset [a, b]$ *does not depend* on the location of $[\xi', \xi'']$, just the length of the subinterval $= \xi'' - \xi'$.

Let $P(s)$ denote the probability of falling into a subinterval of length x. Then,

$$P(x + t) = P(s) + P(t)$$

by hypothesis (and axion of addition). This is true for all $s,\ t$ such that the subintervals are in $[a,\ b]$.

Essentially, one function satifies the above functional equation: $P(s) = k \cdot s$ and $k = \frac{1}{b-a}$ because $P(b-a) = 1$. Thus

$$\begin{aligned} P(\xi' < \xi \leq \xi'') & \\ &= \frac{\xi'' - \xi'}{b-c} \\ &= \int_{\xi'}^{\xi''} \frac{1}{b-a} dx \end{aligned}$$

Thus the random variable ξ is continuous with density

$$p_\xi(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & x \notin [a,b] \end{cases} \qquad \square$$

Example 9: (Service System) Suppose customers arrive into a service system according to the law:

$$\begin{aligned} N_t &= \text{number of arrivals in the time} \\ &\quad \text{interval } [0,t] \qquad \text{(counter)} \\ \text{Prob}\{N_t = n\} &= e^{-\lambda t} \frac{(\lambda t)^n}{n!} \qquad n = 0, 1, 1, \ldots \end{aligned}$$

The parameter λ is called the *rate* of the arrival process. Let $t_n =$ time instant of n^{th} arrival. Let $T =$ time to the *next* arrival.

$$\text{Prob}\{T \leq \delta\} = 1 - P\{T > \delta\}$$

Proceeding on the assumption that the time t_n is as good an origin of time as 0 for the Poisson counter,

$$\begin{aligned} P\{T > \delta\} &= P\{N_\delta = 0\} \\ &= e^{-\lambda\delta} \\ \text{So } P\{T \leq \delta\} &= 1 - e^{-\lambda\delta} \end{aligned}$$

The interarrival time random variable T is what we call an **exponential** random variable, short for *exponentially distributed*. One can prove (later) that the *assumption* above is correct.

Example 1 0: (**Gaussian** random variable) X is a Gaussian random variable if it has a density

$$p_X(x) \;=\; \frac{1}{c}\exp\left(\frac{-(x-\mu)^2}{2\delta^2}\right) \qquad -\infty < x < \infty$$

where $\mu \epsilon \mathbb{R}$ and $\delta > 0$ are parameters.

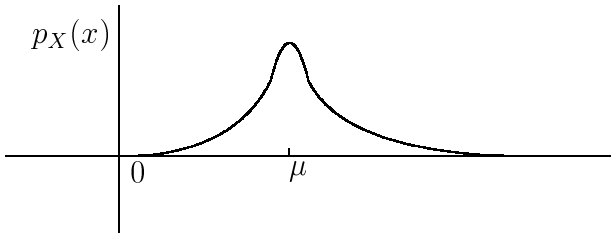


fig. 5

What do these parameters signify? μ is the point about which p_X is symmetric. σ measures the spread of the density function. Greater σ is greater is the spread. What is c? c has to ensure that

$$\int_{-\infty}^{\infty} p_X dx = 1.$$

$$\begin{aligned}
\text{Thus} \qquad c \;&=\; \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
&=\; \int_{-\infty}^{\infty} \exp\left(\frac{-y^2}{2\sigma^2}\right) dy \qquad (\text{change variable} \quad y = x - \sigma) \\
&=\; \sqrt{2}\sigma \int_{-\infty}^{\infty} \exp(-z^2) dz \qquad (\text{change} \;\; z = \tfrac{y}{\sqrt{2}\sigma})
\end{aligned}$$

What is $\int_{-\infty}^{\infty} e^{-z^2} dz$? First denote it as **I**. Then

$$\begin{aligned} \mathbf{I}^2 &= \int_{-\infty}^{\infty} e^{-z^2} dz \int_{-\infty}^{\infty} e^{-w^2} dw \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-(z^2+w^2)} dz dw, \end{aligned}$$

a double integral on the (z, w) plane.

Do a change of variable, $(z, w) \rightarrow (r, \theta)$ where $z = r\cos(\theta), \quad w = r\sin(\theta)$. Then $z^2 + w^2 = r^2, \quad dzdw = rdrd\theta$ and,

$$\begin{aligned} \mathbf{I}^2 &= \int_0^{\infty}\int_0^{2\pi} e^{-r^2} r dr d\theta \\ &= 2\pi \int_0^{\infty} e^{-r^2} r dr \\ &= \frac{2\pi}{2} \int_0^{\infty} e^{-y} dy \qquad\qquad (y = r^2) \\ &= \pi \\ \text{Hence } \mathbf{I} &= \sqrt{\pi} \end{aligned}$$

Thus, $c = \sqrt{2}\ \sigma\ \sqrt{\pi} = \sqrt{2\pi}\sigma$, so the Gaussian density is

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

We also refer to X as a *normal* random variable denoted as,

$$X \sim N(\mu, \sigma^2) \qquad\qquad \square$$

Example 11: (**Cauchy** random variable, also sometimes referred to as **Lorentzian**) A random variable X is said to be a Cauchy random variable if

$$p_X(x) = \frac{1}{\pi} \frac{1}{1 + (x - x_0)^2} \qquad -\infty < x < \infty$$

with graph symmetrical about x_0 (see *figure 6*).

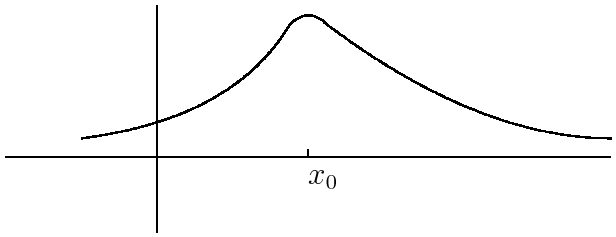


fig. 6

The graph decays slower than that of the Gaussian as $x \to \infty$. □

We now introduce a new concept of quantifying an uncertain or random function. This involves the process of *averaging.*

Suppose a chance experiment, repeated n times, produces the observations $x_1, x_2, x_3, \cdots, x_n,$ of a random variable X. Consider the average

$$Avg(X) = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

It has the properties:

$$\begin{array}{lrcl}
\text{(i)} & \text{Suppose} \quad X & \geqslant & 0, \text{i.e. } R_X \subset [0, \infty), \text{then} \\
 & Avg(X) & \geqslant & 0. \\
\text{(ii)} & Avg(cX) & = & cAvg(X) \\
\text{(iii)} & Avg(X+Y) & = & Avg(X) + Avg(Y) \\
\text{(iv)} & Avg(1) & = & 1
\end{array}$$

the left hand side, 1 denotes the "random variable" which always takes the value 1.

If X is a discrete random variable with $R_X = \{\alpha_1, \alpha_2, \alpha_3, \cdots\}$, then

$$Avg(X) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$
\begin{aligned}
&= \sum_{i=1}^{\infty} \alpha_i \; r_i \\
\text{where } r_i &= (\text{number of times } \alpha_i \text{ occurs})/n \\
&= \text{relative frequency of } \alpha_i
\end{aligned}
$$

Recall that for large number of trials $r_i \to p_i$ by the frequentist intepretation of p_i. We are led to substituting p_i for r_i in the above expression for average, and hence,

Definition: For a discrete random variable X with $R_X = \{\alpha_1, \alpha_2, \alpha_3, \cdots\}$ and probability mass function given by,

$$
p_i = \text{Prob } \{X = \alpha_i\} \qquad\qquad i = 1, 2, 3, \cdots
$$

the **expectation** of X is

$$
E(X) = \sum_{i=1}^{\infty} \alpha_i \; p_i \qquad\qquad \square
$$

All of the properties of the average Avg are satisfied by the expectation. For a continuous random variable X with density p_X,

$$
E(X) = \int_{-\infty}^{\infty} x p_X(x) dx.
$$

Example 11 (Some expectations)
(a)

$$
\begin{aligned}
X &\sim \text{Binomial } (n, p) \\
R_X &= \{0, 1, 2, \ldots, n\} \\
p_k &= \text{Prob } (X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, 2, \ldots, n
\end{aligned}
$$

$$
E(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}
$$

$$
\begin{aligned}
&= \sum_{k=1}^{n} \frac{k\,n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-1-k+1)!} p^k (1-p)^{n-k} \\
&= np \sum_{r=0}^{n-1} \frac{(n-1)!}{r!(n-1-r)!} p^r (1-p)^{n-1-r} \\
&= np(p+(1-p))^{n-1} \\
&= np
\end{aligned}
$$

$$
\begin{aligned}
\text{Thus } p &= \frac{E(X)}{n} \\
&= E\left(\frac{X}{n}\right) \\
&= E \text{ (relative frequency)}
\end{aligned}
$$

(b)

$$
\begin{aligned}
X &\sim \text{Poisson}(a) \\
R_X &= \{0, 1, 2, \ldots\} \\
p_k &= \text{Prob } (X = k) \\
&= e^{-a} \frac{(a)^k}{k!} \qquad k = 0, 1, 2, \ldots \\
E(X) &= \sum_{k=0}^{\infty} k \cdot e^{-a} \frac{a^k}{k!} \\
&= e^{-a} \cdot a \sum_{k=1}^{\infty} \frac{(a)^{k-1}}{(k-1)!} \\
&= e^{-a} \cdot a \cdot e^{a} \\
&= a.
\end{aligned}
$$

(c)

For a Poisson arrival process with $N_t = \#$ arrivals in $[0,t]$ satisfying

$$\begin{aligned}
\text{Prob}(N_t = n) &= e^{-\lambda t}\frac{(\lambda t)^n}{n!} \qquad n = 0,1,2,3,\ldots \\
E(N_t) &= \lambda t \\
\text{Hence } \lambda &= \frac{E(N_t)}{t} \\
&= E\left(\frac{N_t}{t}\right)
\end{aligned}$$

Thus we see a justification for calling λ the arrival rate.

(d)

$$\begin{aligned}
X &\sim \text{Uniform } ([a,b]) \\
R_X &= [a,b] \\
p_X(x) &= \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & x \notin [a,b] \end{cases} \\
\text{Then,} \quad E(X) &= \int_a^b \frac{x}{b-a}dx \\
&= \frac{1}{b-a}\frac{x^2}{2}\,|_a^b \\
&= \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2} \\
&= \text{center of range}
\end{aligned}$$

(e)

$$\begin{aligned}
X &\sim N(\mu,\sigma^2) \\
E(X) &= \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx \\
&= \int_{-\infty}^{\infty} (x-\mu+\mu)\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(x-\mu)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx \\
&\quad +\mu \\
&= 0+\mu \quad \text{because the integrand in the first integral is odd} \\
E(X) &= \mu
\end{aligned}
$$

(f)

$$X \sim \text{Cauchy/Lorentzian}$$

Then $E(X)$ does not exist! Why?

ENEE 324H Lecture 6 (Inequalities)

Estimating probabilities is necessary where analytic formulas are hard to find. Finding good estimates (upper and lower bounds) is an art. But there are some basic estimates derivable from first principles.

1. Markov inequality

Let X be a non-negative random variable. Let u denote the unit step function

$$u(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

Let $a > 0$. Then it is easy to see that

$$u(X-a) \leq \frac{X}{a}.$$

Hence $$E(u(X-a)) \leq \frac{E(X)}{a}$$

But $$E(u(X-a))$$

$$= 0 \cdot P\{\omega : X(\omega) < a\} + 1 \, P\{\omega : X(\omega) \geq$$

$$= P(X \geq a)$$

Thus $$\boxed{P(X \geq a) \leq \frac{E(X)}{a}}$$

Remark (a) Since $\{\omega : X(\omega) > a\} \subseteq \{\omega : X(\omega) \geq a\}$

it follows that

$$P\{\omega : X(\omega) > a\} \leq P\{\omega : X(\omega) \geq a\} \leq \frac{E(X)}{a}$$

Remark (b) If the assumption of non-negativity of X is not applicable, one can still write,

$$P\{\omega : |X(\omega) - \mu| \geq a\} \leq E\left(\frac{|X-\mu|}{a}\right)$$

where $\mu \in \mathbb{R}$ is arbitrary and $a > 0$.
This observation leads to the next inequality.

2. Chebyshev inequality

Let Y be any real-valued random variable.

Let $X = \mu + (Y - E(Y))^2$

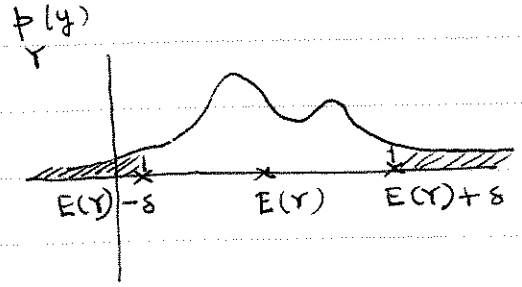
Set $a = \delta^2$ for $\delta > 0$.

Then by Markov's inequality,

$$P\left((Y - E(Y))^2 \geq \delta^2\right) \leq \frac{E\left((Y - E(Y))^2\right)}{\delta^2}$$

equivalently,

$$\boxed{P(|Y - E(Y)| \geq \delta) \leq \frac{\mathrm{Var}(Y)}{\delta^2}}$$



$P(|Y-E(Y)| \geq \delta)$

= area under density curve marked by hatch lines

= tail probability

Chebyshev's inequality estimates the tail ~~probability~~,

3. Convex functions and Jensen's inequality.

$f : \mathbb{R} \to \mathbb{R}$ is convex if

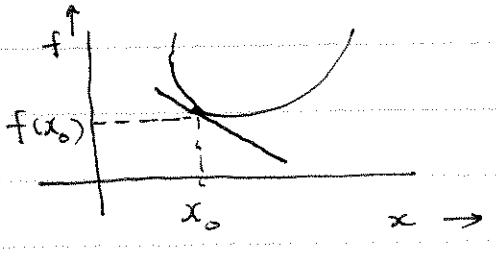
$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

for $\alpha \in [0,1]$.

From this, it follows that the derivative $f'(x)$ (if it exists), is increasing with x, and for any fixed x_0, there exists a constant λ such that

$$f(x) \geq f(x_0) + \lambda (x - x_0).$$

The line with slope λ, passing through $(x_0, f(x_0))$ is called the supporting line at the point $(x_0, f(x_0))$ as in the adjoining figure.



Let $x_0 = E(X)$ for a random variable X

Then,

$$E(f(X)) \geq E(f(x_0) + \lambda(X - x_0))$$

$$= E(f(E(X))) + E(\lambda(X - E($$

$$= f(E(X)) + \lambda E(X) - \lambda E(X)$$

$$= f(E(X))$$

Thus for a convex function f,

$$\boxed{E(f(X)) \geq f(E(X))}$$

4. Chernoff's inequality.

Let X be any random variable. Given $\varepsilon > 0$, define a new random variable dependent on X,

$$Y_\varepsilon = \begin{cases} 1 & \text{if } X \geq \varepsilon \\ 0 & \text{if } X < \varepsilon \end{cases}.$$

For any t, it follows that

$$e^{tX} \geq e^{t\varepsilon} Y_\varepsilon$$

Hence $$E(e^{tX}) \geq E(e^{t\varepsilon} Y_\varepsilon)$$

$$= e^{t\varepsilon} E(Y_\varepsilon)$$

$$= e^{t\varepsilon} P(X \geq \varepsilon).$$

(Here we assume existence of relevant expectations)

Hence,

$$P(X \geq \varepsilon) \leq e^{-t\varepsilon} E(e^{tX})$$

The free parameter t in the above inequality, can be used to obtain a tighter estimate,

$$\boxed{P(X \geq \varepsilon) \leq \inf_{t \geq 0} e^{-t\varepsilon} E(e^{tX})}$$

Here "infimum" stands for the greatest lower bound

Example Suppose X is Gaussian with mean 0 and variance 1. Then the density of X is

$$p_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \qquad -\infty < x < \infty$$

$$E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx$$

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

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

$$= e^{t^2/2} \cdot 1$$

Thm $\quad P(X \geq \varepsilon) \leq \inf_{t \geq 0} e^{-t\varepsilon + t^2/2}$

$$= e^{-\varepsilon^2/2}$$

ENEE 324H Lecture 6
Inequalities

Estimating probabilities is necessary where analytic formulas are hard to find. Finding good estimates (upper and lower bounds) is an art. But there are some basic estimates derivable from first principles.

1. **Markov inequality**
   Let X be a non-negative random variable. Let u denote the unit step function

$$u(x) = \begin{cases} 1 & x \geqslant 0 \\ 0 & x < 0 \end{cases}$$

   Let $a > 0$. Then it is easy to see that

$$u(X - a) \leq \frac{X}{a}$$

   Then

$$E\left(u(X-a)\right) \leq \frac{E(X)}{a}$$

   But

$$\begin{aligned} E\left(u(X-a)\right) &= 0 \cdot P\{\omega : X(\omega) < a\} + 1 \cdot P\{\omega : X(\omega) \geqslant a\} \\ &= P(X \geqslant a) \end{aligned}$$

   Thus

$$\boxed{P(X \geqslant a) \leq \frac{E(X)}{a}}$$

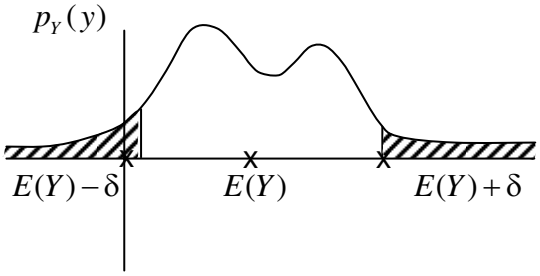
   Remark (a) Since $\{\omega : X(\omega) > a\} \subseteq \{\omega : X(\omega) \geqslant a\}$
   it follows that

$$\begin{aligned} P\{\omega : X(\omega) > a\} &\leq P\{\omega : X(\omega) \geqslant a\} \\ &\leq \frac{E(X)}{a} \end{aligned}$$

Remark (b) If the assumption of non-negativity of X is not applicable, one can still write

$$P\{\omega : |X(\omega) - \mu| \geqslant a\} \leq E\left(\frac{|X-\mu|}{a}\right)$$

where $\mu \in \mathbb{R}$ is arbitrary and $a > 0$. This observation leads to the next inequality.



$P(|Y - E(Y)| \geqslant \delta)$

= area under density curve marked by hatch lines

= tail probability

Figure 1: Chebyshev's inequality estimates the tail-probability

2. **Chebyshev inequality**
Let Y be any real-valued random variable.
Let $X = \mu + (Y - E(Y))^2$
Set $a = \delta^2$ for $\delta > 0$.
Then by Markov's inequality,

$$P((Y - E(Y))^2 \geqslant \delta^2) \leq \frac{E\left((Y - E(Y))^2\right)}{\delta^2}$$

equivalently,

$$\boxed{P(|Y - E(Y)| \geqslant \delta) \leq \frac{Var(Y)}{\delta^2}}$$

3. **Convex functions and Jensen's inequality**
$f : \mathbb{R} \to \mathbb{R}$ is <u>convex</u> if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \text{for } \alpha \in [0,1].$$

From this, it follows that the derivative $f'(x)$ (if it exists) is increasing with x, and for any fixed x_0, there exists a constant λ such that

$$f(x) \geqslant f(x_0) + \lambda(x - x_0)$$

The line with slope λ, passing through $(x_0, f(x_0))$ is called the supporting line at the point $(x_0, f(_0))$ as in *Figure 2*.

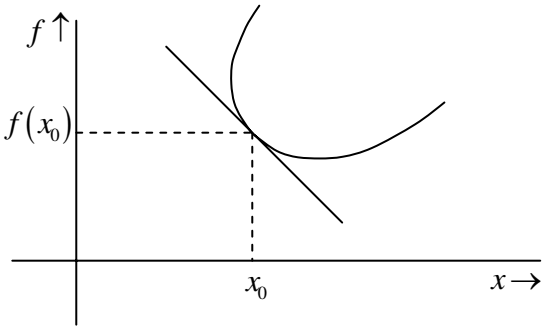


Figure 2

Let $x_0 = E(X)$ for a random variable X. Then,

$$\begin{aligned} E(f(X)) &\geqslant E(f(x_0) + \lambda(X - x_0)) \\ &= E(f(E(X))) + E(\lambda(X - E(X)) \\ &= f(E(X)) + \lambda E(X) - \lambda E(X) \\ &= f(E(X)) \end{aligned}$$

Thus, for a convex function f,

$$\boxed{E(f(X)) \geqslant f(E(X))}$$

4. **Chernoff's inequality**
Let X be any random variable. Given $\epsilon > 0$, define a new random variable dependent on X,

$$Y_\epsilon \;=\; \{ \begin{array}{ll} 1 & if \quad x \geqslant \epsilon \\ 0 & if \quad X < \epsilon \end{array}$$

For any t, it follows that

$$e^{tX} \geqslant e^{t\epsilon}\; Y_\epsilon$$

Hence

$$\begin{aligned} E(e^{tX}) &\geqslant E(e^{t\epsilon}\; Y_\epsilon) \\ &= e^{t\epsilon}\; E(Y_\epsilon) \\ &= e^{t\epsilon}\; P(X \geqslant \epsilon) \end{aligned}$$

(Here we assume existence of relevant expectations.) Hence,

$$P(X \geqslant \epsilon) \leq e^{-t\epsilon}\; E(e^{tX})$$

The free parameter t in the above inequality can be used to obtain a tighter estimate,

$$\boxed{P(X \geqslant \epsilon) \leq \inf_{t \geqslant 0}\; e^{-t\epsilon}\; E(e^{tX})}$$

Here"infimum" stands for the greatest lower bound.

Example: Suppose X is Gaussian with mean 0 and variance 1. Then the density of X is

$$\begin{aligned}
P_X(x) &= \frac{1}{\sqrt{2\pi}} exp\left(\frac{-x^2}{2}\right) \quad -\infty < x < \infty \\
E(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}}\ e^{-x^2/2}\ dx \\
&= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{1\pi}}\ e^{\left(tx-\frac{x^2}{2}-\frac{t^2}{2}\right)}\ dx \\
&= e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\ e^{-\frac{(x-t)^2}{2}}\ dx \\
&= e^{t^2/2}
\end{aligned}$$

Then

$$\begin{aligned}
P(x \geqslant \in) &\leqq \inf_{t \geqslant 0}\ e^{-t\epsilon + t^2/2} \\
&= e^{-\epsilon^2/2}
\end{aligned}$$

ENEE 324H Engineering Probability

Lecture 10

Limit Theorems

Let $X_1, X_2, \ldots$ be a sequence of n independent, identically distributed random variables associated to a chance experiment $(\mathcal{E}, \Omega, A, P)$.

Let $E(X_i) = a \qquad i = 1, 2, 3, \ldots$

Let $Var(X_i) = \sigma^2 \qquad i = 1, 2, 3, \ldots$

Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ for each integer $n > 0$

denote the arithmetic average of the random variables $X_1, X_2, \ldots, X_n$.

Then $E(\bar{X}_n) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right)$

$$= \frac{1}{n} \sum_{i=1}^{n} E(X_i)$$

$$= a$$

$$Var(\bar{X}_n) = E\left[(\bar{X}_n - E(\bar{X}_n))^2\right]$$

$$= E\left[\left(\frac{1}{n} \sum_{i=1}^{n} X_i - a\right)^2\right]$$

$$= E\left[\left(\frac{1}{n} \sum_{i=1}^{n} (X_i - a)\right)^2\right]$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i)$$

$$= \frac{1}{n^2} \, n \, \sigma^2 = \frac{\sigma^2}{n}$$

Here we have used the calculation:

V and W independent, then

$$\begin{aligned} \mathrm{Var}(V+W) &= E\left[\left((V+W) - E(V+W)\right)^2\right] \\ &= E\left[\left((V - E(V)) + (W - E(W))\right)^2\right] \\ &= E\left[(V - E(V))^2\right] + E\left[(W - E(W))^2\right] \\ &\quad - 2\,E\left[(V - E(V)) \cdot (W - E(W))\right] \\ &= \mathrm{Var}(V) + \mathrm{Var}(W) - 2\,\mathrm{cov}(V, W) \\ &= \mathrm{Var}(V) + \mathrm{Var}(W) \end{aligned}$$

($\because$ V, W independent)

Now recall Chebyshev's inequality,

$$P\left(|V - E(V)| > \varepsilon\right) \leq \frac{\mathrm{Var}(V)}{\varepsilon^2}$$

Let $V = \bar{X}_n$ $\quad E(V) = a$. $\quad \mathrm{Var}(V) = \sigma^2/n$

$$\Rightarrow \quad P\left(|\bar{X}_n - a| > \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

$$1 - P(|\bar{X}_n - a| \le \varepsilon) \le \frac{\sigma^2}{n\varepsilon^2}$$

$$\Rightarrow \quad P(|\bar{X}_n - a| \le \varepsilon) \ge 1 - \frac{\sigma^2}{n\varepsilon}$$

Given a, σ and a choice of ε we can make $P(|\bar{X}_n - a| \le \varepsilon)$ as close to 1 as we desire by picking n large enough. We have proved, in effect,

Theorem (Weak Law of Large Numbers - WLLN)

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with $E(X_i) = a$ and $Var(X_i) = \sigma^2$. Then, given $\delta > 0$ and $\varepsilon > 0$ there is a positive integer n such that

$$a - \varepsilon \le \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \le a + \varepsilon$$

with probability greater than $1 - \delta$

Proof: Previous pages + pick $n > \frac{\sigma^2}{\delta\varepsilon^2}$ ▨

Example

n consecutive Bernoulli trials with

$p = P(A)$ = probability of occurence of event A in each of those trials. (we are not changing coins). The trials are independent.

Let $X_i = \begin{cases} 1 & \text{if } A \text{ occurs in } i^{th} \text{ trial} \\ 0 & \text{if } A \text{ does not occur } " \ " \end{cases}$

$E(X_i) = p \qquad Var(X_i) = p(1-p)$.

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

$$= \frac{n(A)}{n}$$

= random variable associated to relative frequency of A.

$$P\left\{ \left| \frac{n(A)}{n} - p \right| \leq \varepsilon \right\} > 1 - \frac{p(1-p)}{n\varepsilon^2}$$

from WLLN

We say $\frac{n(A)}{n} \to p$ as $n \to \infty$ in probability

Definition (Weak convergence of probability distributions)

Let X_n, $n=1,2,\cdots$ be a sequence of discrete random variables. Let range

$$R_{X_n} = \{0, 1, 2, \cdots\}$$

for each random variable X_n. Denote the probability mass function of X_n by

$$p_n(k) = P\{X_n = k\} \quad k = 0, 1, 2, \cdots$$

for each n.

Then the sequence $\{p_n(\cdot)\}$ of probability mass functions is said to converge weakly to the distribution $\{p(\cdot)\}$ if

$$\lim_{n\to\infty} p_n(k) = p(k) \quad \text{for } k = 0, 1, 2, \cdots$$

Theorem Let $F_n(z) = E(z^{X_n})$ = probability generating function of X_n. Then $\{p_n(\cdot)\}$ converges to $\{p(\cdot)\}$ weakly iff

$\lim_{n\to\infty} F_n(z) = F(z)$ uniformly in $\{z : |z| \le \rho\}$ $0 < \rho < 1$.

where $F(z) =$ Probability generating function $\{p(\cdot)\}$. ☒

There is an analogous result for continuous random variables.

Definition Let X_n, $n=1,2,\ldots$ be a sequence of continuous random variables with $R_{X_n} = \mathbb{R}$. Let $p_{X_n}(\cdot)$ denote the density of X_n.

We say that the sequence $\{p_{X_n}(\cdot)\}$ converges weakly to the density $p(\cdot)$ iff

$$\int_{x'}^{x''} p_{X_n}(y)\,dy \to \int_{x'}^{x''} p(y)\,dy$$

as $n \to \infty$, for $[x', x''] \subseteq \mathbb{R}$ any interval.

Theorem Let $\phi_{X_n}(t) = E(e^{itX_n})$ denote the characteristic function of X_n and let $\phi(t) =$ Fourier transform $(p(\cdot))$. Then

$p_{X_n}(\cdot) \to p$ weakly (as $n \to \infty$) iff

$$\lim_{n\to\infty} \phi_{X_n}(t) = \phi(t) \quad \text{uniformly}$$

in every interval $t' \le t \le t''$ □

Proofs of weak convergence theorems and discussion of uniform convergence of functions of complex variables are omitted.

The central limit

Definition Let X_k, $k = 1, 2, \ldots$ be a sequence of random variables, with

$$E(X_k) = a_k < \infty$$

$$\mathrm{Var}(X_k) = \sigma_k^2 < \infty$$

Let $$S_n^* = \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}}$$

where $$S_n = \sum_{k=1}^{n} X_k$$

We say that $\{X_k\}$ satisfies a central limit theorem (CLT) if

$$\lim_{n\to\infty} P\{x' \le S_n^* \le x''\}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{x'}^{x''} e^{-y^2/2}\, dy \qquad \square$$

Remark Various sequences of random variables may or may not satisfy CLT. One has to put suitable hypotheses such that CLT holds

THEOREM (CLT)

Suppose $\{X_k\}_{k=1}^{\infty}$ is an independent sequence

Suppose

(Lyapunov condition) $$\lim_{n\to\infty} \frac{1}{B_n^3} \sum_{k=1}^{n} E|X_k - \lambda_k|^3 = 0$$

where $B_n^2 = \text{Var}(S_n) = \sum_{k=1}^{n} \sigma_k^2$.

Then CLT holds.

Proof: We need to prove uniform convergence of characteristic functions.

$$S_n^* = \sum_{k=1}^{n} \frac{X_k - a_k}{B_n}$$

$E(X_k - a_k) = 0 \qquad \text{Var}(X_k - a_k) = \sigma_k^2$

$$g_k(t) = \phi_{X_k - a_k}(t) = E\left(e^{it(X_k - a_k)}\right)$$

$$= 1 - \frac{\sigma_k^2 t^2}{2} + R_k(t)$$

where $R_k(t)$ = remainder term in power series

and $$|R_k(t)| \leq c\,|t|^3 E\left(|X_k - a_k|^3\right)$$

Let $$\eta_k = (X_k - a_k) / B_k$$

Characteristic function of η_k is

$$f_{kn}(t) = g_k(t/B_n)$$

$$= 1 - \frac{\sigma_k^2 t^2}{2B_n^2} + R_k\left(\frac{t}{B_n}\right)$$

$$|R_k(\tfrac{t}{B_n})| \le c\,|t|^3\, E\,\frac{|X_k - a_k|^3}{B_n^3}$$

$$S_n^* = \eta_1 + \eta_2 + \eta_3 + \cdots + \eta_n$$

$\varphi_{S_n^*}(t)$ = characteristic function of S_n^*

$$= f_n(t)$$

$$= \prod_{k=1}^{n} f_{kn}(t) \qquad \left(\because \frac{X_k - a_k}{B_n} \text{ indep. seq.}\right)$$

$$\ln(f_n(t)) = \sum_{k=1}^{n} \ln(f_{kn}(t))$$

$$\sim \sum_{k=1}^{n}\left[-\frac{\sigma_k^2}{2B_n^2}t^2 + R_k(\tfrac{t}{B_n})\right]$$

Lyapunov condition $\Leftrightarrow$

$$\left|\sum_{k=1}^{n} R_k(t/B_n)\right| \le c\,|t|^3\,\frac{1}{B_n^3}\sum_{k=1}^{n} E\,|X_k - a_k|^3 \to 0$$

as $n \to \infty$

uniformly in $[t', t'']$

$$\Rightarrow \quad \ln\left(f_n(t)\right) \sim -\frac{t^2}{2B_n^2}\sum_{k=1}^{n}\sigma_k^2$$

$$= -\frac{t^2}{2} \qquad \text{as } n\to\infty$$

$$\Rightarrow \quad f_n(t) \to e^{-t^2/2} \qquad \text{as } n\to\infty$$

∎

Special Case

If $\{X_n\}$ is i.i.d and

$\alpha_{\not\not} = \not\not E|X_k - a_k|^3 < \infty$ then

NOTE

$a_k = a$

(∵ iid)

$$B_n^2 = n\sigma^2$$

$$\lim_{n\to\infty} \frac{1}{B_n^3}\sum_{k=1}^{n} E\,|X_k - a_k|^3$$

$$= \lim_{n\to\infty} \frac{\alpha}{\sqrt{n}\,\sigma^3}$$

$$= 0$$

So Lyapunov condition holds $\Rightarrow$ CLT

Thus, for a sequence of i.i.d random variables, CLT holds if the mean, variance and 3rd central moment are finite. In that case arithmetic averages of such random variables can be analyzed via the Gaussian.