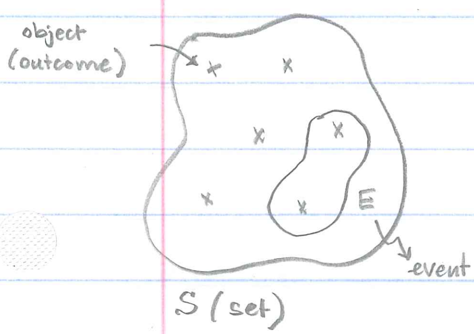
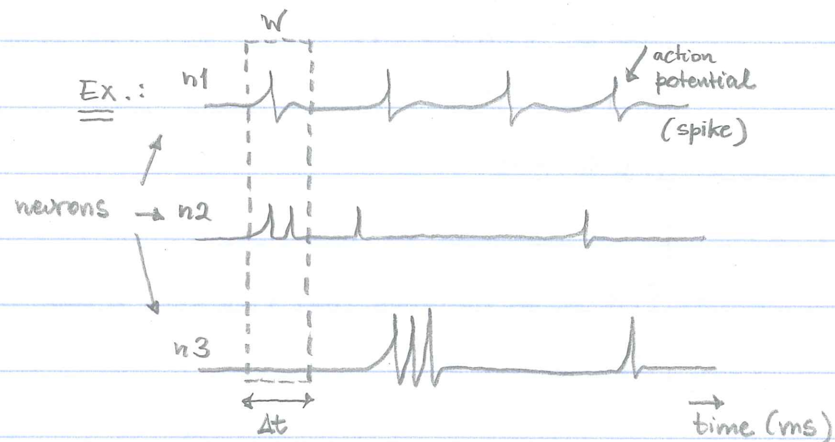
LECTURE 1

* Probability and Random Variables: A refresh

The study of probability is based on the notion of "sets", where a "set" is a collection of objects (elements), each one being potentially accessible.
In case of biomedical signals, the elements we have access to are the outcomes of an experiment or an observational process (e.g., measurement) ⇒ A collection of outcomes is an "event"



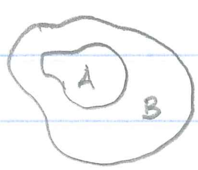


In this example, we may define: outcome $\triangleq$ spike time of neuron n_j, $j = 1,2,3$

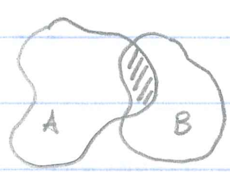
event $\triangleq$ neuron n_j spikes in window W

⇒ Events can be combined to form new events (e.g., a new event is: "neurons n_j and n_i spike both in window W") ⇒ This can be mathematically formulated as combinations of sets ⇒ Set operations are:

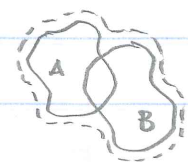
- Inclusion: $A \subset B$



Intersection: $A \cap B$



Union: $A \cup B$



Remember from the Set Theory:

$\emptyset$ = The void set is a set and $\emptyset \subset A \;\; \forall A$

$A \subset B$, $B \subset C \Rightarrow A \subset C$ (transitivity)

$A \cup B = B \cup A$ (union is commutative)

$(A \cup B) \cup C = A \cup (B \cup C)$ (union is associative)

$A \cap B = B \cap A$ (intersection is commutative)

$(A \cap B) \cap C = A \cap (B \cap C)$ (intersection is associative)

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ (intersection is distributive)

$A \subset B \Rightarrow A \cap B = A$

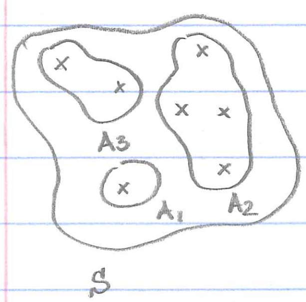
$A \cap \phi = \phi \quad \forall A$

$A \cup \phi = A \quad \forall A$

- Finally, let us recall that two sets A and B are MUTUALLY EXCLUSIVE iff $A \cap B = \emptyset$

- A collection of subsets $U \triangleq \{A_1, A_2, \dots, A_n\}$ of a set S of outcomes is a PARTITION of S iff: $A_i \cap A_j = \emptyset \quad \forall i \neq j$

$$A_1 \cup A_2 \cup \dots \cup A_n = S$$



- The COMPLEMENT $\bar{A}$ of a subset A of S is a subset of S such that:

$A \cup \bar{A} = S$ and $A \cap \bar{A} = \emptyset$

From the definition, it follows: $\bar{\bar{A}} = A$; $\bar{S} = \phi$; $\bar{\emptyset} = S$

From these notions, one can observe that - given a set S of $n > 0$ outcomes - the list of events defined on S easily exceeds n (it is actually 2^n) and is driven by the kind of information we would like to obtain from the outcomes. However, not all the events are equally likely to occur when an experiment is run (e.g., an event happens more often than others) $\Rightarrow$ We need a function that maps events to real numbers $\Rightarrow$ This is the PROBABILITY FUNCTION:

S - set of $n>0$ outcomes $\Rightarrow$ Let us define: $\Omega \triangleq \{A_1, A_2, \dots, A_{2^n}\}$ - SAMPLE SPACE

with $A_i \triangleq$ event on S and $i = 1, 2, 3, \dots, 2^n$

PROBABILITY: $P: \Omega \longrightarrow [0, 1]$ such that:

- $P(A_i) \geq 0 \quad \forall i = 1, 2, 3, \dots, 2^n$
- $P(S) = 1$
- If $A_i \cap A_j = \emptyset \Rightarrow P(A_i \cup A_j) = P(A_i) + P(A_j)$ for any i, j

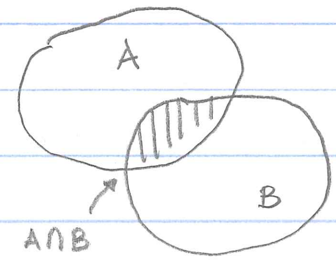
From the definition one can derive the following properties:

- $P(\bar{A}) = 1 - P(A)$
- If $A_i \cap A_j = \emptyset \Rightarrow P(A_i \cap A_j) = 0$
- $P(A_i \cup A_j) = P(A_i) + P(A_j) - P(A_i \cap A_j)$

* Conditional Probability and Bayes' Theorem

The probability of an event A, $P(A)$, can eventually be affected by a priori knowledge about the occurrance of another event (e.g., the likelihood of having a heart beat in a time window $[t, t+\Delta t]$ dramatically changes if a beat just happened before time t) $\Rightarrow$ This makes sense if one relies on the Set Theory:

A priori knowledge of event B $\overset{\text{EQUIV}}{\Longleftrightarrow}$



we limit our universe to event B only and study the probability of event A RELATIVE to B

Hence, we introduce the following definition:

CONDITIONAL PROBABILITY OF EVENT A TO EVENT B

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

Note that the definition holds only if $P(B)>0 \Rightarrow$ In this case, we can write:

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

It is important to emphasize that the conditional probability is normalized to the probability of event B $\Rightarrow$ It is NOT the mere probability of $A\cap B$

EX: Let us consider two patients (P_1 and P_2), each one of whom being either obese (o) or not with probability $P(P_i = o) = 1/2 \quad i=1,2$. If we know that at least one patient is obese (i.e., $B = \{P_1 = o \text{ or } P_2 = o\}$), what is the conditional probability of event $A = \{P_1 = o \text{ and } P_2 = o\}$?

outcomes:

P_1	P_2
o	$\bar{o}$
$\bar{o}$	o
$\bar{o}$	$\bar{o}$
o	o

$$\left.\begin{aligned} P(A) &= 1/4 \\ P(B) &= 3/4 \\ P(A\cap B) &= P(A) \end{aligned}\right\} \Rightarrow P(A|B) = \frac{1/4}{3/4} = 1/3 \neq 1/4$$

INDEPENDENCE Two events A and B are independent $\overset{\text{DEF}}{\Longleftrightarrow}$ $P(A\cap B) = P(A)\cdot P(B)$

This definition is different from assuming $A \cap B = \emptyset$. Moreover, under the assumption that $P(A)>0$ and $P(B)>0$, one can write:

$$\left.\begin{aligned} P(A|B) &= \frac{P(A)P(B)}{P(B)} = P(A) \\ P(B|A) &= \frac{P(B)P(A)}{P(A)} = P(B) \end{aligned}\right\} \text{i.e., the occurrence of one event is NOT related to the occurrence of the other}$$

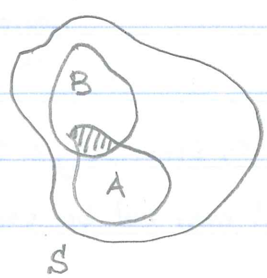
In general, $N>2$ events $A_1, A_2, A_3, \ldots, A_N$ are independent iff

$$P\left(\bigcap_{i=1}^{N} A_i\right) = \prod_{i=1}^{N} P(A_i)$$

Ex.: In the previous example, some events are independent:

$$P(P_1=0 \mid P_2=\bar{0}) = P(P_1=0 \mid P_2=0) = 1/2$$

Let us note one fact about two generic events A and B defined on the set S:



$$\left.\begin{aligned} B &= (B\cap A) \cup (B \cap \bar{A}) \\ (B\cap A) &\cap (B\cap \bar{A}) = \emptyset \end{aligned}\right\} \Rightarrow P(B) = P(B\cap A) + P(B\cap\bar{A})$$

Moreover, we know: $P(B\cap A) = P(B|A)P(A)$

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

Hence: $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$

Let us now recall: $P(A|B) = P(A\cap B)/P(B) \Rightarrow$ We can write:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} \qquad (1)$$

In (1) we estimate the conditional probability $P(A|B)$ by using the conditional probability of B given $A, \bar{A}$ and the probabilities $P(A), P(\bar{A})$. This result is powerful when it comes about testing hypotheses in clinical settings.

<u>Ex.</u>: $A = \{\text{the patient has disease } D\}$

$B = \{\text{the patient is positive to test } T\}$

T is a test for disease D but it is NOT definitive

$P(B|A) \triangleq$ probability of testing positive given the disease (a.k.a. "sensitivity")
$P(\bar{B}|\bar{A}) \triangleq$ probability of testing negative given the absence of disease (a.k.a. "specificity")

We can write:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B|A)\,P(A) + (1 - P(\bar{B}|\bar{A}))(1 - P(A))}$$

positive predictive value of the test

⇐ This value tells how reliable the test is in spotting patients with the disease

Now, even if we assume that the test is good (i.e., sensitivity $\cong 1$ and specificity $\cong 1$) still if $P(A) \cong 0$, then the test may not be definitive!

Note that in (1) the knowledge of $P(B|A)$, $P(B|\bar{A})$, and $P(A)$ must be given
⇒ Some background information is needed. The (1) can be generalized into the following

BAYES' THEOREM: Let us partition S in $n > 1$ mutually exclusive events $A_1, A_2, \ldots, A_n$, with $P(A_i) > 0$ $i = 1, 2, 3, \ldots, n$. If $P(B) > 0$ then:

$$P(A_k|B) = \frac{P(B|A_k)\,P(A_k)}{\sum_{i=1}^{n} P(B|A_i)\,P(A_i)} \qquad k = 1, 2, 3, \dots, n$$

∗ Random Variables

A "random variable" is a function $X(\cdot)$ that maps experimental outcomes onto real numbers:

$$X : S \longrightarrow \mathcal{R}$$
$$\xi \in S \longrightarrow X(\xi) \in \mathcal{R}$$

Based on this definition, an event on S can be defined in terms of random variable, e.g.:

$$A \triangleq \{\xi \in S : X(\xi) = 2\} = \{X = 2\}$$

$$B \triangleq \{\xi \in S : X(\xi) < 2\} = \{X < 2\}$$

conventional notation

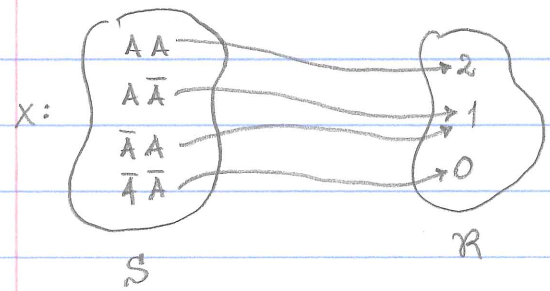
DEFINITION A random variable (RV) is a function $X : S \to \mathcal{R}$ that satisfies the following two conditions:

- $\forall x \in \mathcal{R}$, $\{X \leq x\}$ is an event on S
- $P(X = \infty) = P(X = -\infty) = 0$

For a random variable, we can give the following definition

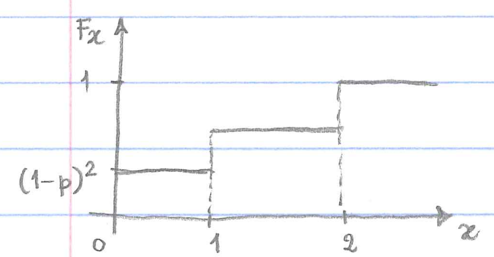
DEFINITION Cumulative Distribution Function (CDF) $\quad F_X : x \in (-\infty, \infty) \longrightarrow P(X \leq x) \in [0, 1]$

Ex.: Let us consider a binary event A (e.g., a neuron fires in response to a stimulus) and let us call $p = P(A)$. Let us repeat the experiment two times and consider the set S of outcomes. A RV can be:



$$\{X=1\} = \{A\bar{A}\,;\,\bar{A}A\}$$
$$\{X=2\} = \{AA\}$$
$$\{X=0\} = \{\bar{A}\bar{A}\}$$

$$F_x(0) = P(X \le 0) = P(X=0) = (1-p)^2$$
$$F_x(1) = P(X \le 1) = P(X=0) + P(X=1) = (1-p)^2 + 2p(1-p)$$
$$F_x(2) = P(X \le 2) = P(X=0) + P(X=1) + P(X=2) = (1-p)^2 + 2p(1-p) + p^2 = 1$$



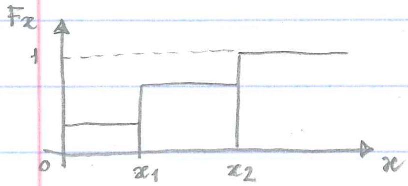
$F_x(\cdot)$ is a non-decreasing function that reaches saturation at $F_x = 1$

Based on the example, we can intuitively understand that $F_x(\cdot)$ has the following properties:

- $F_x(-\infty) = 0$; $F_x(+\infty) = 1$
- If $x_1 < x_2 \Rightarrow F_x(x_1) \le F_x(x_2)$
- If $F_x(x_0) = 0 \Rightarrow F_x(x) = 0 \quad \forall x \le x_0$
- $P(X > x) = 1 - F_x(x)$
- $P(x_1 < X \le x_2) = F_x(x_2) - F_x(x_1)$
- $P(X = x) = \lim_{\varepsilon \to 0} \left(F_x(x+\varepsilon) - F_x(x-\varepsilon) \right)$

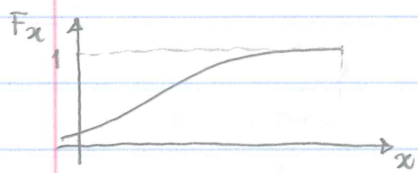
Note the notation: $F_x(x^+) \triangleq \lim_{\varepsilon \to 0} F_x(x+\varepsilon)$ $\qquad F_x(x^-) \triangleq \lim_{\varepsilon \to 0} F_x(x-\varepsilon)$

Also, we can observe:



F_x - piece-wise constant with jump discontinuity $\Rightarrow X$ is a discrete RV

In particular, we have that $F_x(x_i^-) \neq F_x(x_i^+)$ and $P(X = x_i) > 0$ for a finite number of points x_i



F_x - continuous with no jump discontinuity $\Rightarrow X$ is a continuous RV

In this case, we have $F_x(x^-) = F_x(x^+)$ $\forall x \Rightarrow P(X = x) = 0$ $\forall x$

DEFINITION Probability Density Function (pdf)

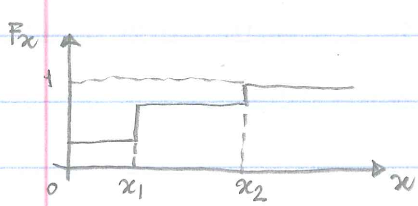
$$f_x(x) \triangleq \frac{d\, F_x(x)}{dx} \qquad (2)$$

Note that, since $F_x(\cdot)$ is a non-decreasing function of x, we have:

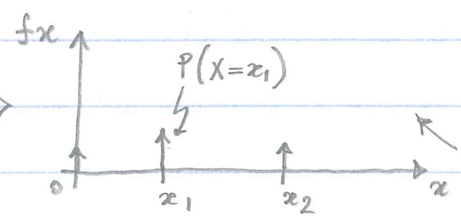
$$f_x(x) = \lim_{\Delta x \to 0} \frac{F_x(x + \Delta x) - F_x(x)}{\Delta x} \geq 0$$

Also, in case of discrete-type RV, we have that: $f_x(x) = \sum_i P(X = x_i)\, \delta(x - x_i)$

Dirac's Delta Function



$\Rightarrow$



$f_x(\cdot)$ in this case is also called "probability mass function"

By using (2), both continuous-type and discrete-type RVs satisfy the conditions:

- $F_x(x) = \int_{-\infty}^{x} f_x(u)\,du$

- $P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f_x(u)\,du$

- $\int_{-\infty}^{+\infty} f_x(u)\,du = F_x(+\infty) = 1$ (i.e., the area of $f_x(\cdot)$ is always 1)

Two quantities are typically computed to characterize a RV:

- expected value (mean) $\quad \mu_x \triangleq \int_{-\infty}^{+\infty} x f_x(x)\,dx = E(X)$

other notation

- variance $\quad \sigma_x^2 \triangleq \int_{-\infty}^{+\infty} (x-\mu_x)^2 f_x(x)\,dx = E\left((X-\mu_X)^2\right)$

Note that, in case of a discrete RV, $f_x(x) = \sum_i P(X=x_i)\,\delta(x-x_i) \Rightarrow$

$$\mu_x = \sum_i P(X=x_i)\,x_i$$

$$\sigma_x^2 = \sum_i P(X=x_i)(x_i-\mu_x)^2 \qquad (*)$$

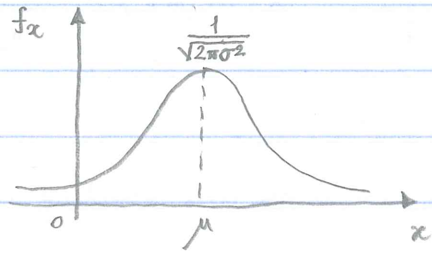
Note that the definitions given above are for RVs, i.e., for mathematical functions from S to $\mathcal{R}$. If one deals with data (e.g., n data samples are obtained by repeating an experiment n times), a definition of mean and variance can be obtained from (*) by considering:

$\underbrace{x_1, x_2, \ldots, x_n}_{\text{data samples}} \qquad P(X=x_i) = \frac{1}{n} \Rightarrow \mu_X = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_X)^2$$

Examples of relevant RVs:

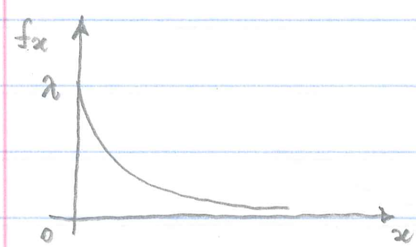
Normal (Gaussian) Distribution $\quad X \sim N(\mu, \sigma^2) \overset{\text{DEF}}{\Longleftrightarrow} f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



The importance of the Normal distribution depends on its ability to describe quantities derived from the data (e.g., sample means)

$E(X) = \mu \qquad E((X-\mu)^2) = \sigma^2$

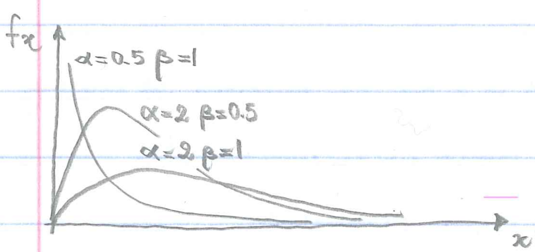
Exponential Distribution $\quad X \sim Exp(\lambda) \overset{\text{DEF}}{\Longleftrightarrow} f_x(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$



The exponential distribution is typically used to describe the waiting times between independent events occurring over nonoverlapping intervals

$E(X) = 1/\lambda \qquad E((X-\mu)^2) = 1/\lambda^2 \qquad F_x(x) = 1 - e^{-\lambda x}$

Gamma Distribution $\quad X \sim Gamma(\alpha, \beta) \overset{\text{DEF}}{\Longleftrightarrow} f_x(x) = \begin{cases} \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{-x/\beta} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$

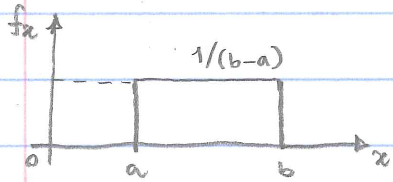


where $\Gamma(\alpha) \triangleq \int_0^\infty x^{\alpha-1} e^{-x} dx$

and $\alpha, \beta > 0$

Uniform Distribution $\quad X \sim U(a,b) \overset{DEF}{\iff} f_X(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$

where $-\infty < a < b < +\infty$

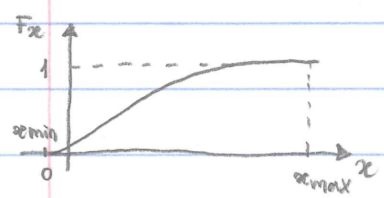


$$E(X) = \frac{a+b}{2} \qquad E((X-\mu)^2) = \frac{(b-a)^2}{12}$$

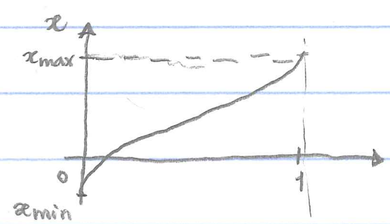
These distributions are recurrent in the processing and modeling of biosignals. A few quantities (beside mean and variance) are usually computed to characterize the behavior of these distributions:

- PERCENTILES: (QUANTILES) The u-percentiles $(0 \le u \le 1)$ of the RV X is the smallest number x_u such that: $F_X(x_u) = P(X \le x_u) = u$, i.e.:

$$x_u = F_X^{-1}(u)$$



To compute percentiles $\Longrightarrow$



$u = 0.5 \Longrightarrow x_u = F_X^{-1}(u)$ - median

- COEFFICIENT OF VARIATION $\quad CV(X) \triangleq \dfrac{\sigma}{\mu}$ - It summarizes the variation of the RV X relative to the mean

- HAZARD FUNCTION: Let us assume that X is a RV that describes the waiting time until some event occurs (e.g., time until the next heart beat). One may be interested in the probability

$$P(\underbrace{x < X \le x+h}_{\text{The waiting time is in the interval } [x, x+h]} \mid \underbrace{X > x}_{\text{No event has occurred up to time } x})$$

Remember: if we call $A \triangleq \{x < X \le x+h\}$

$B \triangleq \{X > x\}$

we can apply the definition of "conditional probability" and write:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \underset{\substack{\uparrow \\ \text{in this case}}}{=} \frac{P(A)}{1-P(\bar{B})} = \frac{F_x(x+h) - F_x(x)}{1 - F_x(x)}$$

The "hazard function" is defined as

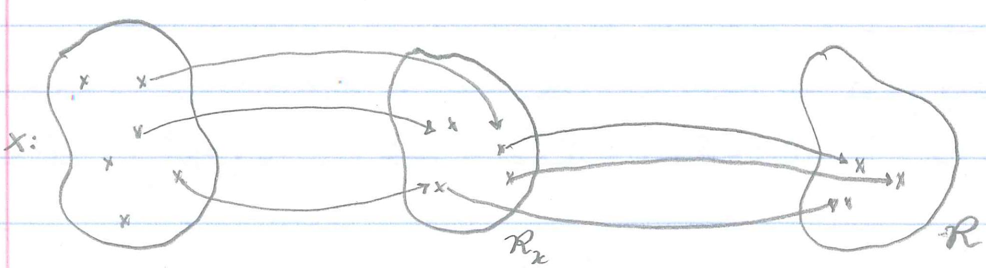
$$\lambda(x) \triangleq \lim_{h \to 0} \frac{P(x<X \le x+h \mid X > x)}{h} = \lim_{h \to 0} \frac{1}{1-F_x(x)} \cdot \frac{F_x(x+h) - F_x(x)}{h}$$

$$= \frac{f_x(x)}{1 - F_x(x)}$$

- It is the instantaneous probability that the event of interest occurs at time x

* Functions of one Random Variable

Let us assume that X is a RV defined on a set S. We can define a function g



from the support $\mathcal{R}_x$ of X on the set of real numbers: $g: \mathcal{R}_x \to \mathcal{R}$

Now, if we consider any generic outcome ξ in S we have:

$$\xi \xrightarrow{X} x = X(\xi) \xrightarrow{g} y = g(x)$$

Hence, by definition of "composite function", we have:

$$\xi \xrightarrow{g \circ X} y = g(X(\xi)) \underset{\uparrow}{=} Y(\xi)$$

$$Y \triangleq g \circ X$$

i.e., the function $g(X(\cdot))$ is a new RV defined on the set S. Moreover, since $Y \triangleq g(X(\cdot))$ is obtained from X, we can determine the distribution of Y from the distribution functions $F_x(\cdot)$ and $f_x(\cdot)$ of X

Technicalities: Note that, from the mathematical standpoint, Y is a RV only if the following conditions occur (which is the case in the applications we typically deal with):

- $\forall y \in \mathbb{R}$, $\{Y \leq y\}$ is an event. This is satisfied if the set of values $x \in \mathbb{R}_x$ such that $g(x) \leq y$ is the union or/and intersection of a countable number of intervals
- $P(Y = \pm\infty) = 0$

For example, let us consider:

$g(X) = aX + b$

For any $y \in \mathbb{R}$ we must find the values x such that $ax + b < y$. Hence:

$$F_y(y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_x\left(\frac{y-b}{a}\right) \quad \text{if } a > 0$$

$$F_y(y) = P\left(X \geq \frac{y-b}{a}\right) \underset{\uparrow}{=} 1 - F_x\left(\frac{y-b}{a}\right)$$

F_x is assumed continuous

Depending on $g(\cdot)$, the passage from F_X to F_Y may be nontrivial. For example:

$$Y = X^2 \Rightarrow \text{If } y \geq 0,\ \{Y \leq y\} = \{X^2 \leq y\} = \{-\sqrt{y} \leq X \leq +\sqrt{y}\}$$
$$\text{If } y < 0,\ \{Y \leq y\} = \{X^2 \leq y\} = \emptyset$$

Hence: $$F_Y(y) = \begin{cases} F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \text{if } y \geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad (**)$$

Similar considerations can be done for the pdf:

$Y = X^2 \Rightarrow$ We can estimate $f_Y(y)$ from $(**)$ by differentiation:

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right) & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$Y = aX + b \Rightarrow \quad f_Y(y) = \begin{cases} f_X\left(\frac{y-b}{a}\right)\frac{1}{a} & \text{if } a > 0 \\ -f_X\left(\frac{y-b}{a}\right)\frac{1}{a} & \text{if } a < 0 \end{cases}$$

In general, though, we do not need to go through $F_Y(\cdot)$ to estimate $f_Y(\cdot)$. The following theorem holds:

Theorem: X is a RV with $f_X(x) > 0\ \forall x \in (a,b)$ and $f_X(x) = 0$ otherwise

$g(x)$ is differentiable and $\frac{dg}{dx}(x) \neq 0\ \ \forall x \in (a,b)$

Then: $$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy}(g^{-1}(y)) \right| \quad \text{for } y \in \mathcal{R}_{a,b}$$

Codomain of (a,b) via $g(\cdot)$

and $f_Y(y) = 0$ otherwise

Ex: $X \sim N(\mu_x, \sigma_x^2)$

$g(X) = aX + b$

- By applying the theorem, we have:

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{|a|} \quad \text{(as we expected)}$$

and, since X is Gaussian, we have: $Y \sim N(a\mu_x + b,\ a^2\sigma_x^2)$

Finally, two results that have large practical relevance:

1) From X to U. Assume that X is a RV with cdf $F_X(x)$. The function $g(x) \triangleq F_X(x)$ is such that: $g(X) \sim U(0,1)$

2) From U to Y. Assume that U is uniform in $(0,1)$. Let us also assume that the function $F_Y(\cdot)$ is assigned. The transformation $Y = g(U)$ with $g = F_Y^{-1}(\cdot)$ is such that:

$Y = F_Y^{-1}(U)$ is a RV with $F_Y(\cdot)$ as CDF ∎

References:

Textbook: ch 3

ch 1-2 (reading)