Probability Theory

1. Background

2 notions of probability:

Probability = analysis

Probability = common notion

A few words on common notions..

2. Experiments and sample spaces

Define as experiment any sequence of events with an outcome.

Example 1: Toss of a die

Example 2: Study on deaths of cancer patients.

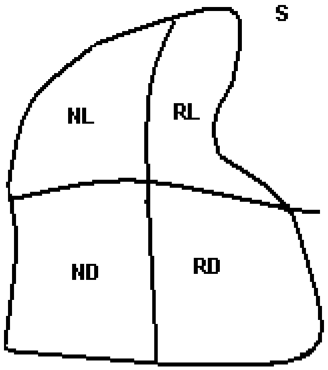
Example 3: High temperatures of day

When we are interested in an experiment, we want to somehow record its outcome, some salient aspect of outcome -- set of all possible outcomes (which has to be classified by experimenter)

Possible outcomes form $\Omega =$ sample space.

Example 4: Die toss. $\Omega = \{1, 2, 3, 4, 5, 6\}$

Example 5: Cancer patients



$=$ 4 **Outcomes**

$R =$ received treatment

$N =$ no treatment

$L =$ lived

$D =$ died

This extends to other characteristics - genetic profiles in bioinformatics

3. Events and probabilities

Example 6: High temperature measurement

Sample space $= \Omega = \{t : t$ a real number$\}$

So: Have set theory and real life situations.

If $A \subset \Omega$, A is an *event.*

Example 7: If $A = \{2, 4, 6\} \subset \{1, 2, 3, 4, 5, 6\}$

then A is an event.

Why an *event* ?

Intuitively, an event means something that has occurred, and above the event $A = \{2, 4, 6\}$ represents the *occurrence* of an even number.

Again can translate between set theory and intuitive notions of meanings of words.

Probabilist wants to assign probability a number between 0 and 1 to every event.

Thus, e.g., if $A = \{\text{event of an even roll}\} = \{2, 4, 6\}$

want $P(A) = \frac{1}{2}$ [Rationales can vary]

So: Ideally, want to assign numbers (probabilities) to subsets

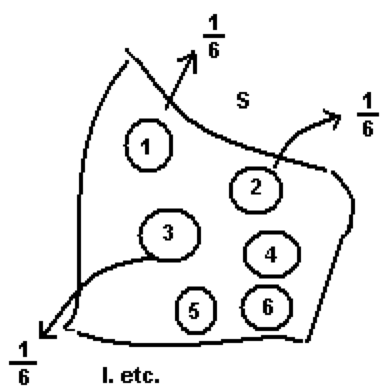
Example 8:

$$P(1) = \frac{1}{6}$$

$$P(2) = \frac{1}{6}$$

$$P(3) = \frac{1}{6}$$

$$P(6) = \frac{1}{6}$$



Thus, each component in Ω has probability $\frac{1}{6}$.
Each subset A can be obtained by adding measure of component subsets A_i.

Want $P(\Omega) = 1$

why ?

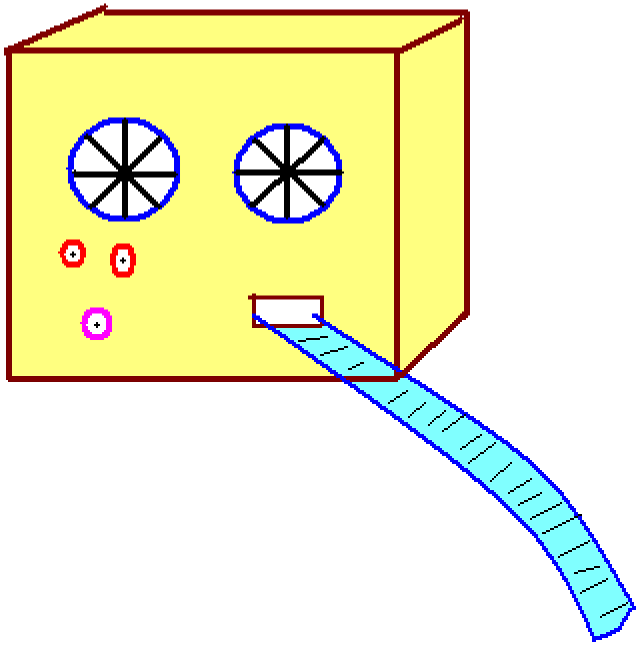
So: given a set in a sample space want probabilities...

$P(A) =$?

$P(\Omega) = 1.$

4. Probability measures

Example 9: Consider an ideal random number generator which generates a real number in $[0,1]$:



In this case:

$$\Omega = [0,1];$$

$$P(\Omega) = P([0,1]) = 1$$

Now we have:

$$P\left[0, \tfrac{1}{2}\right] \ = \ \text{proportional to likelihood of } \left[0, \tfrac{1}{2}\right] = \tfrac{1}{2}$$

$$P[a,b] = b - a.$$

What subsets can we find probability measure of?

(i) Any interval $(a,b):\ P((a,b)) = b - a$
(ii) Any finite union of disjoint intervals

$$P\left(\bigcup_{i=1}^{\infty}(a_i,b_i)\right) = \sum_{i=1}^{\infty}(b_i - a_i) \qquad (*)$$

Let's define the collection of sets whose measures are easy to calculate through formula (*):

$\mathcal{F}_0 =$
$\{$all finite unions of disjoint open intervals $(a_i, b_i)\}$

$$= \{ \bigcup_{i \in J} (a_i, b_i) \,|\, J \text{ finite}\}$$

Note it is easy to define the measure of any set in $\mathcal{F}_0$ using formula (*).

Note that $\mathcal{F}_0$ is a *field* of sets, i.e. has all the properties of a σ-field except that it is closed on only *finite* unions, not necessarily countable ones.

5. σ-Fields of subsets

The natural extension of this to the σ-field $\mathcal{F}$ of Borel sets on $[0, 1]$ can be shown to be unique, and is Lebesgue measure on $[0, 1]$.

Definition 1: If $P(\Omega) = 1$ then the measure P is called a *probability measure* on Ω, and the triple $(\Omega, \mathcal{F}, P)$ is called a *probability space.*

6. More interesting example:

Coin tossing: ∞ number of tosses

$$\Omega = \{(\text{all sequences of } H, T)\}$$

$$H = 1$$
$$T = 0$$

$\Rightarrow \quad \Omega =$ all ∞ sequences of $H's$ and $T's$

How to assign probabilities?

Let $\omega \in \Omega$, with

$$\omega = \omega_1 \omega_2 \omega_3 \ldots = 011010100...$$

Let

$$T(\omega) = .\omega_1 \omega_2 \omega_3 \ldots = .011011\ldots$$

be the corresponding dyadic expansion.

Note: decimal expansion:

$$.12345\ldots = \frac{1}{10} + \frac{2}{100} + \frac{3}{1000} + \ldots$$

$$= \frac{1}{10} + \frac{2}{10^2} + \ldots$$

dyadic expansion:

$$.01100111 = \frac{0}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{0}{2^4} + \frac{0}{2^5} + \ldots$$

Thus we work in base 2 and write numbers as
0's and 1's

Note that

$$T : \Omega \to [0,1]$$

defines $1-1$ correspondence;

$$d_1(\omega) = \omega_1 = \quad \text{first digit}$$

$$d_2(\omega) = \omega_2 \quad \text{second digit, etc.}$$

Note: decimals with first digit 0 are in $[0, \frac{1}{2})$; decmials with first digit 1 are in $[\frac{1}{2}, 1]$.

Then $A_1 = \{\omega : d_1(\omega) = 0\} \Rightarrow T(A_1) = [0, \frac{1}{2})$

$\Rightarrow A_1 = \{\omega$: first toss in corresponding sequence is a tails}

We will assign $P(A_1) = \frac{1}{2} =$ prob. of heads on first toss

$=$ Lebesgue measure of $T(A_1) = P(T(A_1))$

[note we are using the same notation P for:

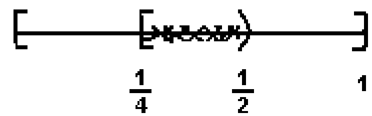
- measures of subsets of Ω = all sequences of coin tosses and for
- measures of subsets of $[0, 1]$ corresponding to subsets of Ω

We anticipate this notation will not cause problems - that

$$P(A) = P(T(A)).$$

Continuing - consider the set

$$A_2 = \{\omega : \; d_1(\omega) = 0, \; d_2(\omega) = 1\}$$
$$\Rightarrow T(A_2) = [\tfrac{1}{4}, \tfrac{1}{2}).$$



Probabilistically: would like $P(A_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

Also we have $P(T(A_2)) =$ Lebesgue measure of $T(A_2) = \frac{1}{4}$.

$$A_3 = \{\omega : d_1(\omega) = 0,\ d_2(\omega) = 1,\ d_3(\omega) = 1\}$$

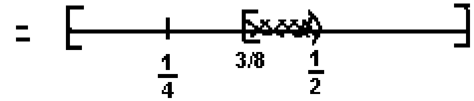
$$\Rightarrow T(A_3) = \left[\frac{3}{8}, \frac{1}{2}\right)$$

$=$ all numbers such that

$$.011 \underbrace{\ldots\ldots}_{}$$

anything

Again $P(A_3) = P(T(A_3)) = \frac{1}{8}$.



This correspondence $P(A) = P(T(A))$ clearly works for any A corresponding to a dyadic interval $T(A)$.

By using countable additivity it also works for any countable unions of sets corresponding to dyadic intervals. That is for any disjoint collection A_i sets in Ω corresponding to dyadic intervals, we must have:

$$P(\bigcup_i A_i) = P(T(\bigcup_i A_i)) = P(\bigcup_i T(A_i))$$

Since any open set (a, b) can be written as such a union, we conclude that if $T(A) = (a, b)$, then

$$P(A) = P(T(A))$$

Thus by unique extension theorem $P(A) = P(T(A))$ for any set $A \subset \Omega$ whose image $T(A)$ is a Borel set in $[0, 1]$.

$\Rightarrow$ Define probability of set $A \subset \Omega$ in coin toss space to be Lebesgue measure $P(T(A)) \subset$ [0,1]

$\Rightarrow$ Probability space $(\Omega, \mathcal{F}, P) =$ Lebesgue measure on $[0, 1]$

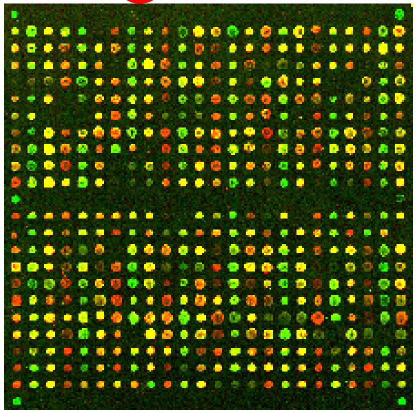
1. The span of probability

Computational biology -

A. Genomes:

- Many organisms are fully sequenced: human, mouse, chicken, yeast, viruses, microbes
- Human genes: about 3Gbp; 22,000 genes
- In humans genes represent about 1.2% of DNA
- 97% of genome considered "junk DNA" (meaning its function is yet unknown)

B. Expression of genes: when are they transcribed? Use gene expression arrays



Source: UCSC

Measure expression (transcription) of several tens of thousands of genes in a single sample.

C. Gene structures We now have 3D-structures of around 70,000 proteins (via NMR or crystallography). We have about 1,300,000 sequenced proteins.

Note: genes are up- and down-regulated (through TF control) in groups:

functional genomics - understanding basics of transcriptional regulation.

D. Hidden Markov models in computational biology

Recall:

- ∃ many genomic datasets from many organisms.
  Want to fully know genomic codes - major goal of
  computational biology.

- Needed (among others) for: drug design, medical

diagnosis, medical treatment, many other research
areas.

Initial use of HMM: Speech processing

Important characteristic for HMM - left to right ordering as a
sequence of words/sounds.

Many computational biology problems can be mapped into

corresponding speech recognition and other language
problems:

Example: protein family classification as speech recognition.

Metaphor:

Different vocalizations of the same word
$\longleftrightarrow$ finding different functional regions of proteins in the
same family

Parsing phonemes into words
$\leftrightarrow$ parsing genomic sequences into codons

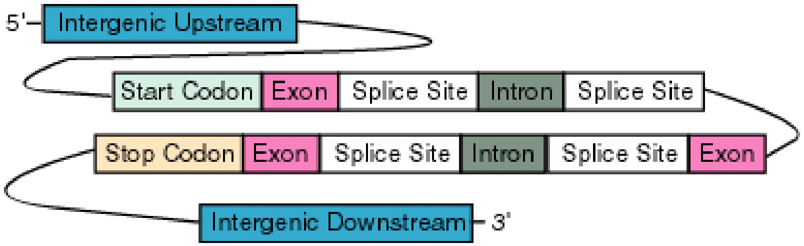
HMM as a mathematical language model
$\leftrightarrow$ HMM as a genomic sequence model

We want a structured model of sequence data; in particular of biological molecular sequences.
Input: DNA sequence $X = \{x_1, \ldots, x_n\} \in \Sigma^n$,
where $\Sigma = \{A, C, G, T\}$

Output: Labeling of x_i as belonging to an intron, exon, or an intergenic region.

Existing tools: Genie, GeneID, HMMGene, GenScan

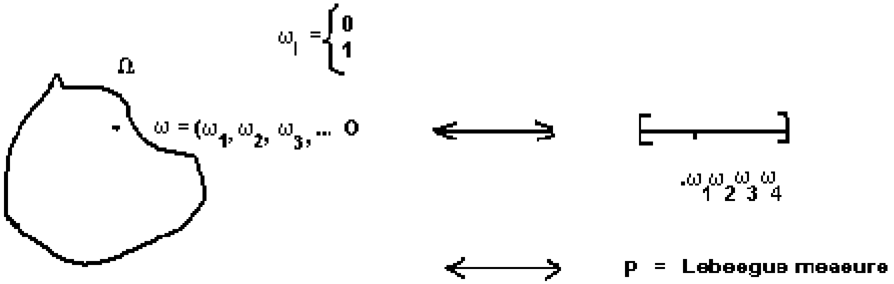
Models consist of several sub-models for different genomic regions:



2. Back to coin tossing: Some proofs

But now let's prove some things.

Recall we have identified the ∞ sequences of 0's and 1's in coin toss space with binary expansions



Recall that if $\omega = .\omega_1\omega_2\omega_3\ldots$ then $d_i(\omega) = \omega_i$.

I want to define

$$A = \left\{\omega : \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} d_i(\omega) = \frac{1}{2}\right\}.$$

$$= \left\{\omega : \quad \text{average value of the digits is} \quad \frac{1}{2}\right\}$$

$= \{\omega :$ proportion of 0's and 1's is equal asymptotically$\}$

This is the set of flip sequences where if you calculate the proportion of heads, it gets closer and closer to $\frac{1}{2}$.

Many seem like not a large set; after all, aren't there a lot of possibilities where he flips all heads or at least heads 2/3 times? NO!

We will show

$$\mathcal{P}(A) = 1$$

$$\mathcal{P}(A^c) = 0.$$

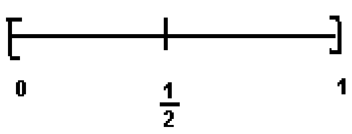
What does this say about binary expansion? It says that if $A =$ set of binary numbers where average value of the first n digits is $\frac{1}{2}$, then $m(A) = 1$. A are *normal numbers*.

Big deal?

Similarly, if $B = \{$decimal numbers where proportion of 0's approaches $\frac{1}{10}\}$, then

$$m(B) = 1.$$

In general, whatever base we're in m (normal numbers) $= 1$.



Let $A = \left\{ \omega = (\omega_1, \omega_2, \omega_3, \ldots) : \frac{1}{n} \sum_{i=1}^{n} \omega_i = \frac{1}{2} \right\}$

We wanted to show $P(A) = 1$.

Equivalently, we show

Theorem 1: *If* $A = \{\omega = \omega_1\omega_2\omega_3, \ldots : \frac{1}{n} \sum_{i=1}^{n} u_i = \frac{1}{2} \}$

($=$ "normal numbers"),

then $m(A) = 1.$

Remark: This is a special case of the *strong law of large numbers.*

Proof (optional): For each number $\omega \in [0,1]$,

$$\omega = .\omega_1\omega_2\omega_3\ldots$$

let $d_n(\omega) = \omega_n = \begin{cases} 0 \text{ or } 1 \\ \end{cases}$

Let $r_n(\omega) = 2d_n(\omega) - 1$ =

$$\begin{cases} 1 & \text{if } d_n(\omega) = 1 \\ -1 & \text{if } d_n(\omega) = 0 \end{cases}$$

Note equivalence:

$$1000110\ldots \qquad \xrightarrow{\text{avg}} \qquad \frac{1}{2}$$

$$1,-1,-1,-1,1,1,-1\ldots \qquad \xrightarrow{\text{avg}} \qquad 0$$

$$A = \left\{\omega : \ \frac{1}{n}\sum_{i=1}^{n} d_n(\omega) \to \frac{1}{2}\right\}.$$

$$
\begin{aligned}
&= \left\{\omega : \ \frac{1}{n}\sum_{i=1}^{n} \frac{r_n(\omega)+1}{2} \to \frac{1}{2}\right\} \\
&= \left\{\omega : \ \frac{1}{2n}\sum_{i=1}^{n} r_n(\omega) + \frac{1}{n}\cdot\frac{n}{2} \to \frac{1}{2}\right\} \\
&= \left\{\omega : \ \frac{1}{2n}\sum_{i=1}^{n} r_n(\omega) \to 0\right\} \\
&= \left\{\omega : \ \frac{1}{n}\sum_{i=1}^{n} r_n(\omega) \to 0\right\}
\end{aligned}
$$

But: pick $\epsilon > 0$, n an integer.

Let

$$s_n(\omega) = \sum_{i=1}^{n} r_n(\omega)$$

Now: consider

$$P(\omega : \ s_n(\omega) \ \geq \ n\epsilon \ )$$

$$\begin{aligned}
&= \quad P(\omega : \ s_n^4(\omega) \ \geq n^4 \epsilon^4) \\
&= \quad \int_{s_n^4(\omega) \geq n^4 \epsilon^4} 1 \ d\omega \\
&\leq \quad \int_{s_n^4(\omega) \geq n^4 \epsilon^4} \frac{s_n^4(\omega)}{n^4 \epsilon^4} \, d\omega \\
&\leq \quad \frac{1}{n^4 \epsilon^4} \int \ s_n^4(\omega) \ d\omega.
\end{aligned}$$

Now -- examine

$$s_n(\omega) = \sum_{i=1}^{n} r_i(\omega)$$

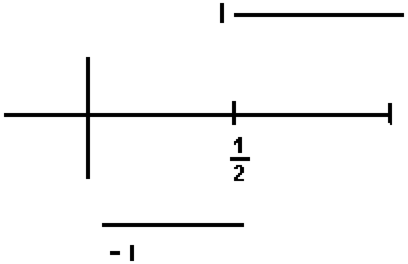
$$s_n^4(\omega) = \left(\sum_{\alpha=1}^{n} r_\alpha(\omega) \right) \left(\sum_{\beta=1}^{n} r_\beta(\omega) \right) \left(\sum_{\gamma=1}^{n} r_\gamma(\omega) \right) \left(\sum_{\delta=1}^{n} r_\delta(\omega) \right)$$

$$= \sum_{\alpha,\beta,\gamma,\delta=1}^{n} r_\alpha(\omega)\, r_\beta(\omega)\, r_\gamma(\omega)\, r_\delta(\omega)$$

$$\int s^4(\omega)d\omega = \sum_{\alpha,\beta,\gamma,\delta=1}^{n} \int d\omega\, r_\alpha(\omega) r_\beta(\omega) r_\gamma(\omega) r_\delta(\omega)$$

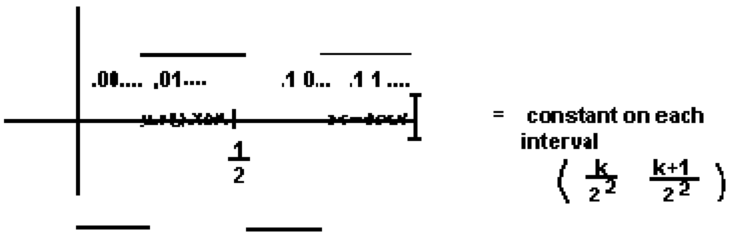
Let's look at what the r_α functions look like:

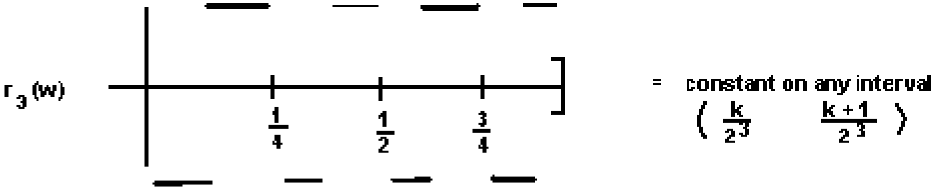
$$r_1(\omega) = \begin{cases} +1 & \text{if first digit in } \omega = 1 \\ -1 & \text{if first digit in } \omega = 0 \end{cases}$$



$r_1(w)$ = constant in each interval of size $\frac{1}{2}$

$$r_2(\omega) \;=\; \begin{cases} +1 & \text{if second digit in } \omega = 1 \\ -1 & \text{if second digit in } \omega = 0 \end{cases}$$





Now: what pops up in

$$\sum_{\alpha,\beta,\gamma,\delta=1}^{n} r_\alpha(\omega) r_\beta(\omega) r_\gamma(\omega) r_\delta(\omega)$$

(a) when $\alpha = \beta = \gamma = \delta$, get r_α^4

(b) when $\alpha = \underbrace{\beta \neq \gamma}_{\text{not equal}} = \delta$ get $r_\alpha^2 \; r_\gamma^2$

(c) when $\alpha = \beta \neq \gamma \neq \delta$ get $r_\alpha^2 \, r_\gamma \, r_\delta$

(d) when $\alpha = \beta = \gamma \neq \delta$ get $r_\alpha^3 \, r_\delta$

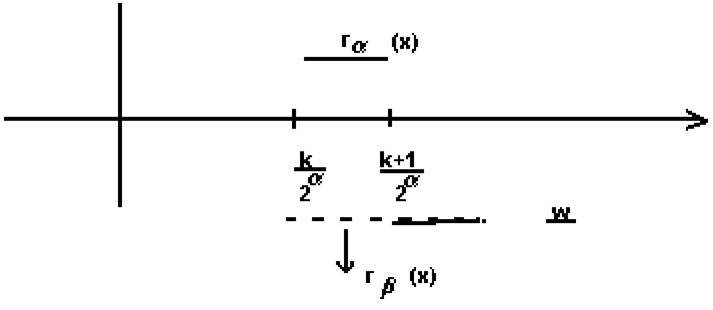
(e) when $\alpha \neq \beta \neq \gamma \neq \delta$ get $r_\alpha r_\beta r_\gamma r_\delta$

Simple case: consider

$$\int_0^1 r_\alpha \, r_\beta \, dw \qquad\qquad \alpha \neq \beta$$

assume $\beta > \alpha$

Look at any interval, $\left(\frac{k}{2^\alpha} \, , \, \frac{k+1}{2^\alpha} \right]$



Then $r_\alpha(w)$ is constant (either $+1$ or -1) on this interval. But since $\beta > \alpha$, $r_\beta(w)$ is $+1$ and -1 many times on this interval; r_β is constant on all intervals $\left(\frac{j}{2^\beta}, \frac{j+1}{2^\beta}\right)$, and there are many of these in each interval $\left(\frac{k}{2^\alpha}, \frac{k+1}{2^\alpha}\right)$.

Thus, even though r_α is constant in $\left(\frac{k}{2^\alpha},\right.$ $\left.\frac{k+1}{2^\alpha}\right)$, r_β is not, and alternates between -1 and 1 $2^{\beta-\alpha}$ times. Thus,

$$\int_{\frac{k}{2^\alpha}}^{\frac{k+1}{2^\alpha}} r_\alpha(w)\, r_\beta(w)\ dw$$

$$= r_\alpha(w) \int_{\frac{k}{2^\alpha}}^{\frac{k+1}{2^\alpha}} r_\beta(w)\, dw = 0.$$

$$\Rightarrow \int r_\alpha r_\beta = 0$$

By the same reasoning, if $\alpha \neq \beta \neq \gamma \neq \delta$,

$$\int d\omega\, r_\alpha\, r_\beta\, r_\gamma\, r_\delta = 0.$$

Similarly, the integral $\int d\omega \;\; r_\alpha^3\, r_\delta = \int d\omega\, r_\alpha\, r_\delta$

$= 0$

and $\int d\omega\, r_\alpha^2\, r_\delta\, r_\gamma = \int d\omega\, r_\delta$

$r_\gamma = 0$

But:

$$r_\alpha^4 \equiv 1$$

$$r_\alpha^2 \, r_\gamma^2 \equiv 1.$$

Now:

$$r_\alpha\, r_\beta\, r_\gamma\, r_\delta = \begin{cases} r_\alpha^4 \\ r_\alpha^2\, r_\beta^2 \\ r_\alpha^2\, r_\beta\, r_\gamma \\ r_\alpha^3\, r_\beta \\ r_\alpha\, r_\beta\, r_\gamma\, r_\delta \end{cases} \quad \longrightarrow \quad \text{intregrate two}$$

So:

$$\sum_{\alpha,\beta,\gamma,\delta=1}^{n} \int r_\alpha \, r_\beta \, r_\gamma \, r_\delta \; dw$$

$$1$$

$$= \sum_{\substack{\alpha,\beta,\gamma,\delta \\ \text{all 4 equal}}}^{\text{n}} \int r_\alpha^4 \, d\,w$$

$$+ \sum_{\substack{\alpha,\beta,\gamma,\delta \\ \text{two equal pairs}}} \int d\omega \;\; r_\alpha^2 \; r_\beta^2$$

no. times all 4 are = to I

$$= n \qquad\qquad\qquad\qquad +$$

number of times two pairs are equal

$$\alpha = \beta = \gamma = \delta = 1$$

$$\alpha = \beta \qquad \gamma = \delta$$
$$\alpha = \gamma \qquad \beta = \delta$$
$$\alpha = \delta \qquad \beta = \gamma$$

no. of chances for α=β

no. chances for γ=δ

$$= n + n \quad (n-1) \quad \cdot 3 \quad \rightarrow$$

match different components

$$\Rightarrow \sum_{\alpha,\beta,\gamma,\delta} \int d\omega \; r_\alpha \, r_\beta \, r_\gamma \, r_\delta = n + 3n(n-1)$$

$$\Rightarrow \int s_n^4(\omega)\, d\omega = n + 3n(n-1).$$

Recall $$s_n = \sum_{i=1}^{n} r_i(\omega).$$

$$\Rightarrow \; P(\omega : \; |s_n(\omega)| \geq n\epsilon) \; \leq \frac{1}{n^4\epsilon^4} \int s_n^4(\omega) d\omega$$

$$= \; \frac{n + 3n(n-1)}{n^4\epsilon^4} \leq \frac{3n^2}{n^4\epsilon^4} = \frac{3}{n^2\epsilon^4}$$

$\Rightarrow$

$$P\left(\omega : \left|\frac{1}{n}s_n(\omega)\right| \geq \epsilon\right) \leq \frac{3}{n^2\epsilon^4}$$

Let $\quad A_{nk} = \left\{\omega : \left|\frac{1}{n}\sum_{i=1}^{n} r_i(\omega)\right| \leq \frac{1}{k}\right\}$

$$P(A_{nk}) \leq \frac{3k^4}{n^2}.$$

Let $A_k = \{\omega : \omega \in A_{nk}$ for all n sufficiently large$\}$

Claim: $\omega(A_k) = 1$, since $\forall\, N$

$$A_k \supseteq \bigcap_{n=N}^{\infty} A_{nk}$$

$$
\begin{aligned}
\Rightarrow \quad & P\left(\bigcap_{n=N}^{\infty} A_{nk}\right) \\
= \quad & P\left([0,1] \sim \bigcup_{n=N}^{\infty} \tilde{A}_{nk}\right) \\
\geq \quad & P([0,1]) - \sum_{n=N}^{\infty} P(\tilde{A}_{nk}) \\
\geq \quad & 1 - \sum_{n=N}^{\infty} \frac{3k^4}{n^2} \\
= \quad & 1 - 3k^4 \sum_{n=N}^{\infty} \frac{1}{n^2}
\end{aligned}
$$

let $N \to \infty$.

$$\Rightarrow \quad P(A_k) = 1.$$

$$A = \left\{ \omega : \ \frac{1}{n} \sum_{i=1}^{n} r_n(\omega) \ \underset{n \to \infty}{\longrightarrow} 0 \right\}$$

$$A_k = \left\{ \omega : \left| \frac{1}{n} \sum_{i=1}^{n} r_n(\omega) \right| \leq 1/k \right\}$$

for n large enough $A = \bigcap A_k$

$\Rightarrow P(A) = 1.$ $\square$

3. The scope of probability: Genomic Markov Models

Hypothetical situation: choose a genome.

Model overall percentage of 2-mers (i.e., Markov statistics)

genome (available DNA)	CG	GC	TA	AT	CC GG	TT AA	TG CA	AG CT	AC GT	GA TC	G+C
Escherichia coli (4.6Mb)*	1.16	1.28	0.75	1.10	0.91	1.21	1.12	0.82	0.88	0.92	51%
Haemophilus influenzae (1.8Mb)*	1.09	1.43	0.75	0.95	1.01	1.25	1.12	0.82	0.85	0.87	38%
Neisseria gonorrhoeae (877kb)	1.32	1.26	0.63	1.05	0.99	1.50	0.99	0.67	0.83	0.89	53%
Neisseria meningitidis (2.2Mb)	1.31	1.27	0.64	1.05	0.96	1.44	1.01	0.70	0.84	0.91	52%
Rhodobacter capsulatus (1.4Mb)	1.19	1.19	0.33	1.61	0.88	1.30	1.03	0.84	0.71	1.16	67%
Rickettsia prowazekii (1.1Mb)*	0.77	1.53	0.98	0.98	1.03	1.05	1.02	1.06	0.86	0.91	29%
Helicobacter pylori (1.7Mb)*	0.93	1.56	0.73	0.86	1.17	1.37	0.97	0.97	0.67	0.87	39%
Campylobacter jejuni (1.6Mb)*	0.62	1.75	0.77	0.83	1.11	1.25	1.03	1.09	0.71	0.92	31%
Bacillus subtilis (4.2Mb)*	1.04	1.27	0.65	1.02	0.97	1.24	1.08	0.91	0.75	1.06	44%
Streptococcus pyogenes (985kb)	0.71	1.19	0.76	0.89	1.04	1.17	1.12	1.04	0.86	0.99	39%
Clostridium acetobutylicum (4.0Mb)	0.45	1.28	0.93	0.95	1.22	1.08	1.02	1.12	0.81	0.97	31%
Streptomyces coelicolor (2.4Mb)	1.14	0.97	0.51	0.93	0.88	0.82	1.00	0.95	1.14	1.25	72%
Mycobacterium leprae (1.7Mb)	1.13	1.07	0.75	1.10	0.88	1.04	1.14	0.86	1.05	1.02	58%
Mycobacterium tuberculosis (4.4Mb)*	1.18	1.07	0.58	1.24	0.86	1.05	1.11	0.80	1.05	1.08	65%
Mycoplasma genitalium (580kb)*	0.39	1.19	0.75	0.77	1.13	1.23	1.16	1.06	0.96	0.89	32%
Mycoplasma pneumoniae (816kb)*	0.82	1.14	0.77	0.71	1.12	1.30	1.08	0.96	1.02	0.81	40%
Synechocystis sp. (3.6Mb)*	0.75	1.02	0.75	1.00	1.36	1.32	1.05	0.85	0.79	0.86	48%
Deinococcus radiodurans (3.0Mb)	1.07	1.16	0.49	0.89	0.87	1.24	1.12	1.00	0.93	1.01	67%
Treponema pallidum (1.1Mb)*	1.08	1.22	0.74	0.93	0.86	1.18	1.13	0.94	0.96	0.95	53%
Borrelia burgdorferi (911kb)*	0.48	1.47	0.77	0.88	1.29	1.22	1.02	1.07	0.69	1.01	29%
Chlamydia trachomatis (1.0Mb)*	0.79	1.12	0.77	0.89	1.01	1.16	0.96	1.18	0.76	1.15	41%
Aquifex aeolicus (1.6Mb)*	0.87	0.75	0.82	0.66	1.24	1.29	0.74	1.18	0.89	1.12	43%
Methanococcus jannaschii (1.7Mb)*	0.32	1.12	0.83	0.94	1.38	1.14	1.03	1.11	0.72	1.05	31%
Methanobacterium thermoautotrophicum (1.8Mb)*	0.51	0.76	0.74	1.13	1.25	0.95	1.17	1.07	0.85	1.14	50%
Archaeoglobus fulgidus (2.2Mb)*	0.78	1.02	0.61	0.86	1.04	1.21	1.01	1.17	0.77	1.19	49%
Pyrococcus horikoshii (1.7Mb)*	0.61	0.89	0.90	0.92	1.30	1.11	0.85	1.22	0.73	1.13	42%
Pyrobaculum aerophilum (2.2Mb)*	0.97	1.15	1.07	0.93	1.10	1.18	0.86	1.06	0.83	0.90	51%
human (5.8Mb)	0.25	1.00	0.74	0.88	1.25	1.12	1.20	1.17	0.83	0.99	43%
mouse (1.1Mb)	0.22	0.95	0.72	0.80	1.19	1.08	1.24	1.26	0.88	1.01	46%
Drosophila melanogaster (4.3Mb)	0.94	1.29	0.75	0.97	1.08	1.23	1.12	0.87	0.84	0.90	41%
Caenorhabditis elegans (74Mb)	0.97	1.04	0.62	0.86	1.05	1.28	1.09	0.90	0.86	1.09	36%
yeast (12Mb)*	0.80	1.02	0.77	0.94	1.06	1.14	1.10	0.99	0.89	1.06	38%
Arabidopsis thaliana (2.0Mb)	0.72	0.93	0.74	0.90	1.03	1.13	1.11	1.04	0.91	1.11	36%
Plasmodium falciparum (947kb)	0.74	0.93	0.99	1.07	1.54	1.00	1.10	0.83	0.92	0.97	20%

* indicates complete genome

<0.50 | 0.50–0.70 | 0.70–0.78 | 0.78–1.23 | 1.23–1.30 | 1.30–1.50 | >1.50

FIG. 1. Genome signature (dinucleotide relative abundances) of complete genomes and large DNA sequence samples (>500 kb).

Source: Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA Allan Campbell, Jan Mrazek, and Samuel Karlin, Proc. Natl. Acad. Sci. USA, Vol. 96, pp. 9184–9189, August 1999

Above represent relative abundances

For a base i define ρ_i = relative abundance of i

For each successive pair ij, e.g. AG = CT, (equivalent mirror reversed) let

$\rho_{ij} =$ proportion of successive pairs which are ij

Define $R_{ij} = \frac{\rho_{ij}}{\rho_i \rho_j} =$ relative overabundance of 2-mer over expected abundance if i, j independent.

[many simple statistics can be done on the genome]

For humans:

$$\rho_A = \rho_T = .57/2 = .285$$

$$\rho_C = \rho_G = .43/2 = .215$$

$$\rho_A = .57; \quad \rho_C = .43$$

$$R_{ij} = \begin{array}{c c}
 & \begin{array}{cccc} A & C & G & T \end{array} \\
\begin{array}{c} A \\ C \\ G \\ T \end{array} & \begin{bmatrix} 1.12 & .83 & 1.17 & .88 \\ 1.2 & 1.25 & .25 & 1.17 \\ .99 & 1.00 & 1.25 & .83 \\ .74 & .99 & 1.2 & 1.12 \end{bmatrix} \\
 & \begin{array}{cccc} A & C & G & T \end{array}
\end{array}$$

$$[p_{ij}] = [R_{ij} \cdot \rho_j] = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} .319 & .178 & .252 & .251 \\ .343 & .269 & .054 & .334 \\ .282 & .214 & .268 & .236 \\ .211 & .213 & .258 & .318 \end{bmatrix} =$$

$$P$$

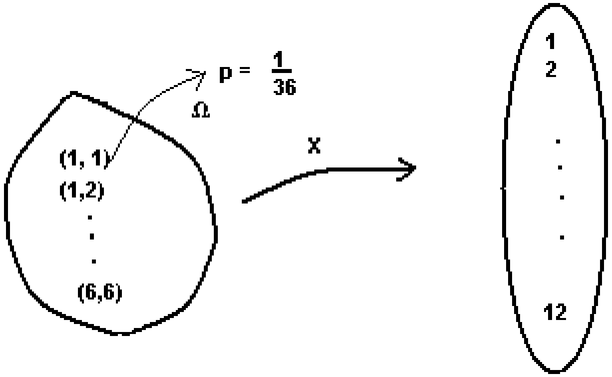
is the transition matrix for a first order Markov (background) model of the human genome.

Note that a 0^{th} order model would be

$$[\rho_{ij}^{(0)}] = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ [.285 \ .215 \ .215 \ .285] \end{array}$$

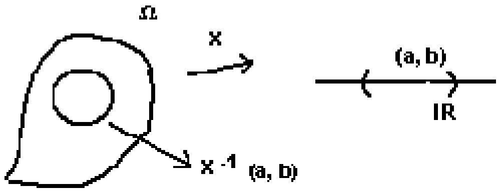
Lecture 2: Random variables and quantization

Example 1: throw 2 dice



X maps outcome to number

$X \; = \;$ **Random Variable**

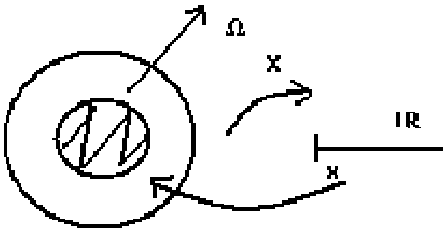


Recall: given $(\Omega, \mathcal{F}, P)$ $X : \; \Omega \; \rightarrow \mathbb{R}$ is *measurable* if

$X^{-1}(a,b) \; \in \mathcal{F}$ for all a, b (since intervals (a,b) generate all Borel sets).

Definition 1: If X is measurable from Ω to $\mathbb{R}$, then X is a *Random Variable (RV)*

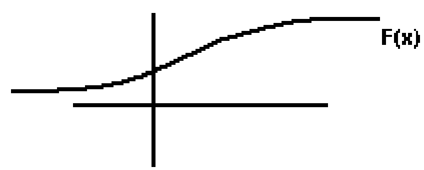
For an r.v. X:



If X is a random variable, we define

distribution function

$$F(x) := P(w : X(w) \leq x) = P(X \leq x)$$



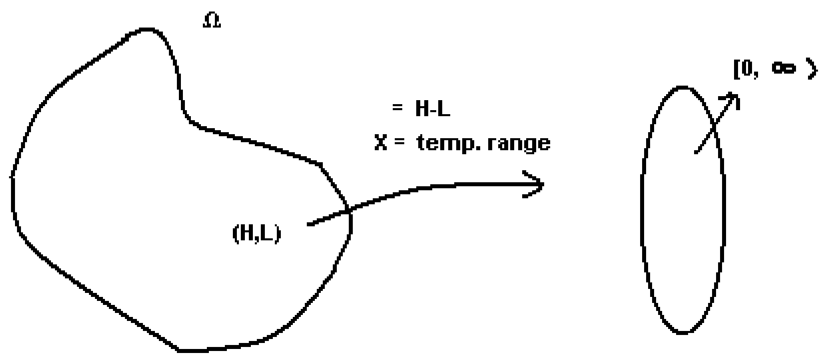
Properties of F (easily derived)

(i) $F(x) \to 1$; $x \to \infty$
$F(x) \to 0$; $x \to -\infty$

(ii) F has at most countably, many discontinuities i.e., if $x_1, x_2, \ldots$ are points of discontinuity, they can be listed in a string.

Example 2: Suppose we record high, low temperatures on a given day; form a sample space

$$\Omega = \{(H, L) : H \geq L\}$$



For each element of Ω, let

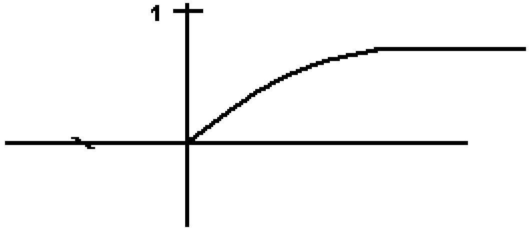
$$X(H,L) = H - L = \text{ temperature}$$

range. Might find that

$$X(H,L) =$$

$$\begin{aligned} F(x) &= P(\,(H_1 L) : (H-L) \leq x) \\ &= P(X \leq x) = \begin{cases} 1 - e^{-x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \end{aligned}$$

[i.e., it's right continuous]



can check this is a d.f.

If F has a derivative, or equivalently if F is the integral of some function $F(x) = \int_{-\infty}^{x} dx\, f(x)$, then

$$F'(x) = f(x) = \textit{density function}$$

of X.

Example 3: here density $= f(x) = \begin{cases} e^{-x} \\ 0 \end{cases}$

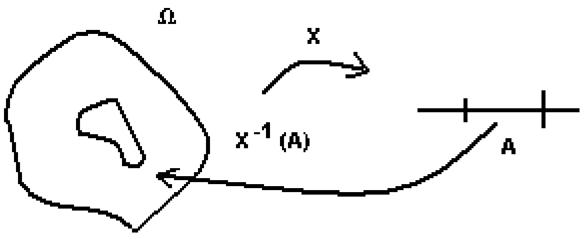
check: $$F(x) = \int_{-\infty}^{x} f(x')dx'.$$

Example 4: Normal $\quad - \frac{1}{\sqrt{2\pi}}\, e^{-x^2/z} = f(x)$

Thus, each $\quad X \;\longrightarrow\; F(x) = \int_{-\infty}^{x} dx\, f(x)$.

Now: $\quad F(x) \,=\, P(w: \; X(w) \leq x)$

Define a measure μ on Borel sets $\mathcal{B}$ in $\mathbb{R}$, with the property:



$$\mu(A) = \mathcal{P}(w : X(w) \in A)$$

Can check this is a probability measure in $\mathcal{B}$.

Now:

$$\begin{aligned} \mu(-\infty, x] &= P(w : X(w) \in (-\infty, x]) \\ &= P(w : X(w) \leq x) \\ &= P(X \leq x) = F(x). \end{aligned}$$

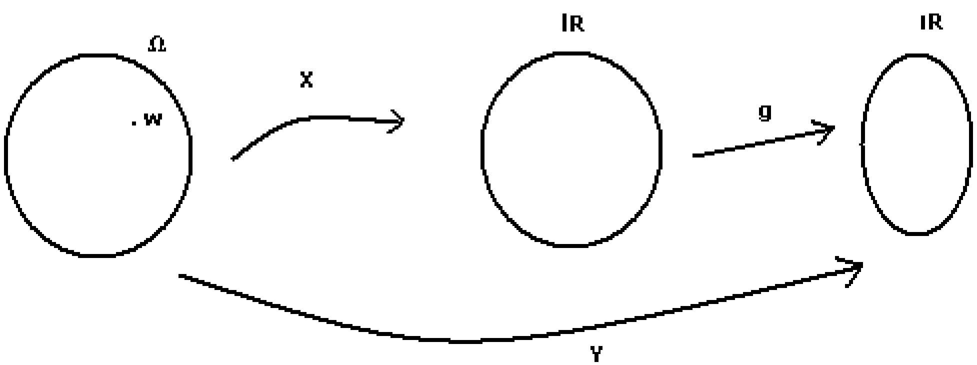
$\Rightarrow \qquad \mu(-\infty, x]$ determined by $F(x)$.

But μ is a Stieltjes measure defined by the increasing function $F(x)$, and so is totally determined by F

μ is called the *distribution* of X.

Now: Let X be a random variable, and g be a function: define a new random variable by

$$Y = g(X(w))$$



Then Y is a random variable. How to calculate d.f. of Y ?

$$
\begin{aligned}
F_Y(y) &\equiv P(Y \leq y) \\
&= P(g(X) \leq y) \\
&= P(g(X) \in (-\infty, y]) \\
&= P(X \in g^{-1}(-\infty, y])
\end{aligned}
$$

Example 5: Suppose X has d.f.

$$F(x) = \begin{cases} 1 - e^{-x} \\ 0 \end{cases}$$

$$f(x) = \begin{cases} e^{-x} \\ 0 \end{cases}$$

$$Y = g(X) = X^2$$

p.d.f. of Y is:

$$P(Y \leq y) = P(X^2 \leq y)$$

$$= \begin{cases} 0 & \text{if } y < 0 \\ P(-\sqrt{y} \leq X \leq \sqrt{y}) & y \geq 0 \end{cases}$$

$$= \begin{cases} 0 & \text{if } y < 0 \\ P(X \leq \sqrt{y}); & y \geq 0 \end{cases}$$

$$= \begin{cases} 0, & y < 0 \\ 1 - e^{-\sqrt{y}} & y > 0. \end{cases}$$

4. Algebraic integration theory.

A new formulation of measure and integration theory allows for non-commutative probability and quantum probability as generalizations of regular probability.

Key element: fundamental quantities are random variables (which is what we observe).

Example 1. Consider the sample space Ω of possible daily closing price records $(\omega_1, \ldots, \omega_{30})$ over a given month, for Hewlett-Packard corporation. We assume $0 \leq \omega_i \leq \$100$.

Thus

$$\Omega = \{\omega = (\omega_1, \ldots, \omega_{30}) : 0 \leq \omega_i \leq 100\} = [0, 100]^{30}.$$

For $A \subset \Omega$, let $P(A) = $ probability that the outcome vector **x** is in the set A. Thus P is a measure on Ω.

There are lots of possible random variables (functions on Ω):

(1) $R = R(\omega) =$ return $= \frac{\omega_{30}-\omega_0}{\omega_0}$.

(2) for given $1 \leq d \leq 30$, let

$$r_d = r_d(\omega) = \text{daily return} = \frac{x_d - x_{d-1}}{x_d} .$$

(3) $\sigma = \sigma(\omega) =$ volatility $=$ standard deviation of returns

$$= \sqrt{\frac{1}{29}\sum_{d=1}^{30}(r_d - \mu)^2}$$

with $\mu = \mu(\omega) = \frac{1}{30}\sum_{d=1}^{30} r_d$.

Many other financial metrics:

(4) Sharpe ratio $= \frac{R(\omega)}{\sigma(\omega)}$.

Common point: these are all functions on the fundamental probability (measure) space P on Ω.

Note these and all other observables are functions on Ω, i.e., random variables.

5. Expectations.

Note: we are really interested in random variables $X(\boldsymbol{\omega})$ on Ω rather than Ω itself.

Given a random variable (RV) $X(\omega) : \Omega \longrightarrow \mathbb{R}$ or $\mathbb{C}$, we define its *expectation* (or average value) to be

$$E(X) = \int_{\Omega} X(\omega) d\mu(\omega)$$

[standard def. of average of a function; recall $\mu(\Omega) = 1$].

Consider the space **B** of all bounded random variables $X(\boldsymbol{\omega})$ on Ω. Note this is a Banach space $L^\infty(\Omega)$ with norm $\|X(\omega)\| = \operatorname*{ess\,sup}_{\omega\in\Omega} X(\omega)$

[i.e. the maximum not counting sets of measure 0].

But it is also an algebra since if $X(\omega)$ and $Y(\omega)$ are bounded random variables then so is $X(\omega)Y(\omega)$.

[Note all definitions complex vector spaces also work for real vector spaces below]

Definition 2. An *algebra* **A** is a complex vector space with multiplication defined on it, i.e. for $X, Y \in$ **A**, $XY \in$ **A** is defined and satisfies

(i) $X(Y+Z) = XY + XZ$

(ii) $(Y+Z)X = YX + ZX$

Definition 3. A *Banach algebra* **B** is a Banch space with the additional structure of an algebra such that $\|XY\| \leq \|X\|\|Y\|$ for $X, Y \in$ **B**.

We will show that the structure of all random variables $X(\omega)$ on a probability space Ω will be determined by their structure as a Banach algebra, together with knowing only their expectations.

Definition 4. An *involution* on an algebra **A** is a map $X \longrightarrow X^*$ that is a conjugate linear isomorphism, i.e., for $X, Y \in$ **A** and $c \in \mathbb{C}$,

(i) $(cX)^* = \overline{c}X$
(ii) $X^{**} = X$
(iii) $(X + Y)^* = X^* + Y^*$
(iv) $(XY)^* = Y^*X^*$.

Definition 5. An *integration algebra* is a system $(\mathbf{A}, E, *)$ in which **A** is a complex associative algebra (i.e. $(XY)Z = X(YZ)$), * is an involution on **A**, and $E : \mathbf{A} \to \mathbb{C}$ is an *expectation*, i.e.

(i) $E(X^*) = \overline{E(X)}$

(ii) $E(X^*X) \geq 0$

(iii) $E(XY) = E(YX)$

(iv) $|E(X^*YX) \leq c(Y)E(X^*X)$,

where $c(Y)$ is positive and depends only on Y.

Example 2. Consider the algebra of all bounded random variables $X(\omega)$ on a probability (measure) space Ω. With the norm $\|X\| = \|X(\omega)\|_\infty$, this forms a Banach algebra **B**.

If $X = X(\omega) \in$ **B**, we can define $X^* = \overline{X}(\omega)$ (i.e. complex conjugate) to be our involution.

We an define our expectation to be

$$E(X(\omega)) = \int X(\omega) dP(\omega).$$

[can show has above properties of expectation].

Note this algebra is *commutative*, i.e. $XY = YX$.

Definition 6. The *spectrum* of **B** is the collection of all (nonzero) continuous linear functionals ϕ : **B** $\rightarrow \mathbb{C}$ which are multiplicative, i.e., such that

$$\phi(XY) = \phi(X)\phi(Y).$$

6. The algebra of random variables determines the probability structure

Theorem 2. Assume we are given a probability space Ω and any algebra **A** of bounded random variables on Ω, thus forming a natural integration algebra $(\mathbf{A}, E, *)$. Then the structure of this integration algebra uniquely determines Ω and the family of random variables **A**, up to isomorphism.

Proof: We need to show that if two measure spaces Ω_1, Ω_2 with their own specific algebras

$\mathbf{A}_1, \mathbf{A}_2$ of functions have the same integration algebra structures, so that $(\mathbf{A}_1, E_1, *_1)$ and $(\mathbf{A}_2, E_2, *_2)$ are isomorphic as algebras, then the two spaces Ω_1 and Ω_2 are equivalent as measure spaces. We also need to show that the corresponding families $\mathbf{A}_1$ and $\mathbf{A}_2$ are equivalent as families of functions on these two spaces.

So assume we have two measure spaces Ω_i with algebras of functions $\mathbf{A}_i$ on them. Assume that as integration algebras $(\mathbf{A}_i, E_i, *_i)$ are isomorphic. This means that there is a bijective isomorphic mapping $U : \mathbf{A}_1 \longrightarrow \mathbf{A}_2$, such that for $X, X_1, X_2 \in \mathbf{A}_1$,

(1) $U(a_1 X_1 + a_2 X_2) = a_1 U(X_1) + a_2 U(X_2)$
(2) $U(X_1(\omega) X_2(\omega)) = U(X_1(\omega)) M(X_2(\omega))$
(3) $E_2(MX) = E_1(X)$.
(4) $(UX)^* = U(X^*)$

We then need to show that Ω_1 and Ω_2 are equivalent as measure spaces and $\mathbf{A}_1$ and $\mathbf{A}_2$ are equivalent as families of functions on these two spaces.

To do this we will find a measure preserving mapping $T : \Omega_1 \to \Omega_2$ such that for $X \in \mathbf{A}_1$,

$$UX(\omega) = X(T\omega).$$

We will show that this mapping gives the equivalence between $(\Omega_i, \mathbf{A}_i)$ as families of measureable functions.

To find such a mapping T, first consider a set $E \subset \Omega_1$. Let

$$\chi_E(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

be the characteristic function of E. Then note that $\chi_E^2(\omega) = \chi_E(\omega)$, so

$$(M\chi_E)^2 = M(\chi_E^2) = M(\chi_E) = M\chi_E.$$

Thus $M\chi_E$ is the characteristic function of a set, call it $T(E)$.

7. Next: Quantum (free) probability.