APPLIED PROBABILITY ~~A~~ W.A. THOMSON

JOINT MARGINAL AND CONDITIONAL PROBABILITIES

We will wants to consider the joint performance of two or more probability experiments (sum as well as difference)

The probabilities of the various sums and differences are (according to the axioms) calculated by totaling the rows and columns respectively; hence these latter probabilities are called marginal probabilities

According to the multiplication rule, conditional probabilities will be calculated by dividing a joint by a marginal probability.

For example, the conditional probability of a sum of 8 given a difference of 2 is $(1/36) / (4/36) = 1/4$

Marginal and conditional probabilities do not differ in any essential way from ordinary probabilities; the word marginal emphasises that there is at least one other experiment which is simultaneously being considered while "conditional" simply makes it explicit that the certain event has been restricted

We write $p(s,d)$ for the joint ~~distribution~~ probability of the sum and difference, $p(s)$ and $p(d)$ for the marginal probabilities $p(s|d)$ and $p(d|s)$ for the conditional probability

Here $s = 2 \ldots 12$ and $d = 5, 4 \cdots -4, -5)$
Thus $p(3,2) = 0$ $p(8) = 5/36$ for the sum and $p(8|2) = 1/4$ for the conditional probability of a sum given a difference of 2

Two experiments are said to be independent if $p(C \cap B) = p(C) \cdot p(B)$ whenever C and B are possible results of the first and second experiments respectively

From the multiplication rule we see that this is the as requiring $P(B) = P(B|C)$, that is, the probability of every possible result of the second experiment is independent in the grammatical sense of what has happened in first experiment.

If the potential results and their probabilities are identical for a sequence of n independent ~~trials~~ experiments, then the sequence is said to consists of n independent trials.

In general we will say that we have a sequence of <u>n Bernoulli</u> trials if:

1. Each trial has 2 possible outcomes, success or failure
2. The probability of success is the same for each trial.
3. The n trials are independent probability experiments

P is probability of success and hence failure is $1-p$ (say q).
The 2^n potential results of a sequence of n Bernoulli trials may be represented by ordered sequences of S's and F's

As example $P(SFFFS) = p \cdot q \cdot q \cdot q \cdot p = p^2 q^3$

C:3 DISCRETE RANDOM VARIABLES C:S

3.1 Random Variables, Probability Functions, and Expectations

A random variable (r.v) $x(s)$ is a function defined for each of the simplest events of some probability experiment and having the real line as range, $x(s)$ is discrete if it's range is some countable subset $X = \{x_1, x_2 \dots\}$ of the real line

That is, to each simple event $s \in S$, there corresponds exactly one of the numbers, $x_1 x_2 \dots$, if x corresponds to s, then we write $x(s) = x$.

The probability $P_x(A)$, that x is an element of some set A in the range of x is the probability of the set of all events in S which map

AD T.

3

into A, that is $P_x(A) = P[\{s \in S : X(s) \in A\}]$

The function $P_X(x) = P(X(s) = x)$ will be called the probability function of X

Frequently, we will write x for $x(s)$ and $p(x)$ for $P_x(x)$.

Example:
If peter matches paul, peter wins a cent otherwise paul does. Let $X(s)$ be peters winnings. HT peters roll H paul rolls T.

We have $X(HH) = X(TT) = +1$ and $X(HT) = X(TH) = -1$ which $P_X(+1) = P(X = +1) = \frac{1}{2}$ and $P(X=-1) = P_X(-1) = \frac{1}{2}$

If peter and paul play game n times and if x matched result, then Peter's average winning will be:

$$\frac{x - (n-x)}{n} = 1 \cdot \frac{x}{n} + (-1)\left(1 - \frac{x}{n}\right)$$

But x/n is a measure of the probability of a match and hence, peters average winnings will be a measure of $1 \cdot P(X=1) + (-1)P(X=-1)$.

We describe this situation by saying peter will "expect" to win $1 \cdot P(X=1) + (-1)P(X=-1) = 0$ cents each time he matches coins with paul.

Thus on average, Peter should expect neither to win nor lose in a long sequence of matches. More generally we define the expectation of any discrete random variable $X(s)$ as:

$$E\,X(s) = \sum_{i=1}^{\infty} x_i \cdot P[X(s) = x_i]$$

Provided that the series converges absolutely. or:

$$E X = \sum_x x \cdot P_X(x) = \sum_x x \cdot p(x)$$

4

Theorem: The expectation of the r.v $y(s) = z[x(s)]$ is given by

$$Ey = \Sigma_x \; z(x) p_x(z)$$
$$Ex = \Sigma x \, P(x)$$

Theorem: If a and b are constants while x and y are r.v's whose expectation exist then:

$$E(ax) = a \cdot Ex$$
$$E(b+x) = b + E(x)$$
$$E(x+y) = E(x) + E(y)$$

that is, E is a linear operator

The Binomial Probability function

Probabilistic experiment observing whether a flipped thumb tack falls up or down

If x denotes the number of times, out of n flips, that the tack falls point up, then we may ask for the probability function of x.

We may ask: What is the probability $b(x;n,p)$ of exactly x successes in a sequence of n Bernoulli trials, the probability of success on a single trial being p?

The probability of x success out of n, in any particular order is $p^x(1-p)^{n-x}$ $0 \le x \le n$, Thus since the number of ways of ordering x S's and $n-x$ F's is $\binom{n}{x}$ then:

$$b(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}; \quad x = 0,1 \ldots n.$$

This is the binomial probability function

We have expected number of successes in n Bernoulli trials:

$$Ex = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}$$

Define the indicator variables $X_i = \begin{cases} 1 & \text{if } S \text{ occurs on the } i^{th} \text{ trial} \\ 0 & \text{if } F \text{ occurs on the } i^{th} \text{ trial} \end{cases}$

AP T

5.

Now $X = X_1 + \cdots + X_n$ is the total number of success in n Bernoulli trial. $EX_i = 1 \cdot p + 0 \cdot (1-p) = p$ and $EX = EX_1 + \cdots + EX_n$ so that $EX = n \cdot p$

Taking $n = 100$ and $p = \frac{1}{2}$ we see that in this sense one should expect 50 heads in 100 flips of a coin

The Poisson Probability Function

Consider collisions of a distinguished molecule of gas with many other molecules making up a gas in a steady state. We might ask for probability $P_t(x)$ of x collisions in time t.

Between collisions, the distinguished molecule will travel in a straight line at a constant speed. It is not unreasonable to assume that in a short interval of time $(t, t + \Delta t)$, the probability of a single collision is proportional to Δt while the probability of two or more collisions vanishes to an order higher than Δt.

Because the gas is in a steady state p, the constant of proportionality will be independent of t and collisions in disjoint time intervals will not influence one another

To be precise, our assumptions are:

1. Events defined on nonoverlapping time intervals are statistically independent.
2. $P_{\Delta t}(0) = 1 - p \cdot \Delta t + o(\Delta t)$
3. $P_{\Delta t}(1) = p \cdot \Delta t + o(\Delta t)$

Where $o(\Delta t)$ denotes a quantity which approaches zero faster than Δt

To evaluate $P_t(x)$ we divide interval $(0, t)$ into n equal parts

7

Theorem: If $N \to \infty$ while $\frac{D}{N} \to p$ a constant, then:

$$h(x; n, D, N) \sim b(x; n, p)$$ no hypergeometric

in the sense that the ratio of the two sides approaches 1.

Chapter 4 PROBABILITY DISTRIBUTION FUNCTIONS AND CONTINUOUS RANDOM VARIABLES

4.1 Distribution Functions:

For "error of measurement" problems of type above it is frequently possible to describe the random variation of the error magnitude X in the following manner:

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

This is the normal or Gaussian probability integral

Various probabilities of interest can be obtained from $P(X \le x)$;

$$P(X > x) = 1 - P(X \le x)$$

and $P(|X| \le a) = P(X \le a) - P(X < -a)$

A real valued random variable is a function whose domain is an event space S and whose range is the real line. Discrete r.v.s treated in chapter 3 are a special kind of real valued r.v.

Henceforth, we will understand that all r.v.'s are real valued unless stated otherwise

The function $F(x) = P(X \le x)$ will be called the distribution function (d.f.) of the random variable X.

Properties, from the nonnegativity of the probability function we have for $y > x$

$$F(y) = F(x) + P(x < X \le y) \ge F(x)$$

So that all d.f.s are monotone non decreasing

Now let $y_1 > y_2 > \dots$ be a sequence approaching x from the

8.

right and write $B_i = \{S : X \le y_i\}$ then $B_1 \supset B_2 \supset \dots$ and $\bigcap_{i=1} B_i = \{S : X \le x\}$

From the continuity property of probability we see that:

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

Thus, every d.f is continuous from the right. Examples of d.f.s which are not everywhere continuous from the left are common; the d.f of every discrete r.v will have points of discontinuity

Arguments similar to those used to establish above eqⁿ lead to:

$$\lim_{x \to \infty} F(x) = 1 \quad \text{and} \quad \lim_{x \to -\infty} F(x) = 0$$

Simple d.f's:

$$P(X \le x) = \begin{cases} 0 & x < 0 \\ 1/2 & 0 \le x < 1 \\ 1 & 1 \le x \end{cases} \quad \text{or} \quad P(X \le x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x < 1 \\ 1 & 1 \le x \end{cases}$$

Density Function

The distribution function $F(x)$ of the r.v X will be called smooth if (i) $F(x)$ is continuous on the entire real line and (ii) $F(x)$ has a continuous derivative except at a finite number of points

For smooth d.f's the density function of X is defined to be the derivative of $F(x)$ where it exists:

$$f(x) = \frac{dF(x)}{dx} = F'(x)$$

If: $$F'(x) = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x}$$

is infinite, or not a unique point, then the density is not determined there.

$$P(X \le x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x < 1 \\ 1 & 1 \le x \end{cases} \Rightarrow f(x) = \begin{cases} 0 & x < 0 \text{ or } x > 1 \\ 1 & 0 < x < 1 \\ \text{undefined} & x = 0, 1 \end{cases}$$

APT

a

In an interval (a,b) within which $F(x)$ has a continuous derivative

$P(a < x < b) = F(b) - F(a) = F'(\xi)(b-a)$

where $a < \xi < b$

The smooth distribution are an important special case of these possessing density functions. If the d.f of x is representable in the form

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)\,dy$$

where $f \geq 0$, then x is called a continuous r.v and $F(x)$ is an absolutely continuous d.f with density function $f(x)$

All densities have the properties:

i. $f(x) \geq 0$

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

iii. $P(a \le x \le b) = \int_a^b f(x)\,dx$

We define the expectation as

$$Ex = \int_{-\infty}^{\infty} x f(x)\,dx = \lim_{t \to -\infty} \int_t^0 x f(x)\,dx + \lim_{u \to \infty} \int_0^u x f(x)\,dx$$

provided that the two limits on the right are finite

The uniform Distribution

- In a calculation you get say 31.6 number might mean m(s) of a length L.
- If the usual convention for rounding have been observed then we will be confident that $31.55 \le L \le 31.65$ but the exact value of L within this interval will be uncertain

- We might be willing to consider L as a r.v L distributed on the interval $[31.55, 31.65]$ with the probability of any subinterval depending only on its length

This assumption is enough to determine the dist of L concretely

Theorem 4.1
If X is an r.v defined on the interval $[0,1]$ and if $P(x<X\leq y)$ depends only on the length $y-x$ for all $0\leq x\leq y\leq 1$, then X has the distribution

$$P(X\leq x) = F(x) \begin{cases} 0 & x\leq 0 \\ x & 0\leq x\leq 1 \\ 1 & 1\leq x \end{cases}$$

The Normal Distribution
The error magnitude X will have density function

$$n(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]}$$

An easy integration shows that

$$E\,X = \int_{-\infty}^{\infty} x\cdot n(x;\mu,\sigma^2)\,dx = \mu$$

It is not easy to prove that A, the area under the density function is 1.

Exponential Distribution
We calculated the probability $P_t(x)$ of X collisions within a fixed time interval of length t.

Now reconsider the same problem from a different point of view, let Y_x be a random time at which the x^{th} collision occurs

$$P(t<Y_x\leq t+\Delta t) = P_t(x-1)\cdot P_{\Delta t}(1)$$

$$= \frac{(\rho t)^{x-1}}{(x-1)!} e^{-\rho t}\cdot[\rho\Delta t + o(\Delta t)]$$

Therefore the density function of Y_x is:

$$\rho\,\frac{(\rho t)^{x-1}}{(x-1)!} e^{-\rho t} \qquad t\geq 0 \qquad \underline{4.18}$$

We remark that 4.18 is a density since the integral involved is a gamma function

PROBABILITY AN INTRODUCTION : GRIMMETT, WELSH

Probability function P satisfies certain conditions

1. Each event A in the event space should have probability $P(A)$ which lies between 0 and 1.
2. the event Ω that 'something happens' should have probability 1, and
3. the event $\emptyset$, that 'nothing happens' should have probability 0.

If A and B are disjoint events (so that $A \cap B = \emptyset$) then

$P(A \cup B) = P(A) + P(B)$

1.6 Conditional Probabilities

In general, if A and B are events (that is $A, B \in \mathcal{F}$) and we are given that B occurs, then in light of this information, the new probability of A may no longer be $P(A)$

In this new circumstance, A occurs if and only if $A \cap B$ occurs, suggesting that the new probability of A is 'proportional' to $P(A \cap B)$

Formal definition, if $A, B \in \mathcal{F}$ and $P(B) > 0$, then the (conditional) probability of A given B is denoted by $P(A|B)$ and defined by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad 1.9$$

Note that the constant of proportionality in (1.9) has been chosen so that the probability $P(B|B)$, that B occurs given that B occurs satisfies $P(B|B) = 1$.

1.7 Independent Events

We call two events A and B 'independent' if the occurrence of one of them does not affect the probability that the other occurs. More formally this requires that, if $P(A), P(B) > 0$ then

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B). \qquad 20$$

Writing $P(A|B) = P(A \cap B)/P(B)$ we see that the following def is appropriate

2

Events A and B of a probability space $(\Omega, \mathcal{F}, P)$ are called independent if $P(A \cap B) = P(A)P(B)$ 21

and dependent otherwise

Thus three events A, B, C are independent if and only if the following equalities hold:

$$P(A \cap B \cap C) = P(A)P(B)P(C) \qquad P(A \cap B) = P(A)P(B)$$
$$P(A \cap C) = P(A)P(C) \qquad P(B \cap C) = P(B)P(C)$$

The Partition Theory:

A partition of Ω is a collection $\{B_i : i \in I\}$ of disjoint events (so that $B_i \in \mathcal{F}$ for each i and $B_i \cap B_j = \emptyset$ if $i \neq j$) with union $\bigcup_i B_i = \Omega$

Theorem (Partition): if $(B_1, B_2 \ldots)$ is a partition of Ω such that $P(B_i) > 0$ for each i then:

$$P(A) = \sum P(A \mid B_i)P(B_i) \quad \text{for all } A \in \mathcal{F}$$

Also called law of total probability.

$$\begin{aligned} P(A) &= P(A \cap (\textstyle\bigcup_i B_i)) \\ &= P(\textstyle\bigcup_i (A \cap B_i)) \\ &= \textstyle\sum_i P(A \cap B_i) && \text{by } (9) \\ &= \textstyle\sum_i P(A \mid B_i)P(B_i) && \text{by } (19) \end{aligned}$$

Example: Tomorrow rain or snow but not both probability rain $= \frac{2}{5}$ and snow is $\frac{3}{5}$. If it rains then probability I will be late is $\frac{1}{5}$ while the corresponding probability of snow late is $\frac{3}{5}$. What is probability I'm late?

Solution

Let A be the event that I am late and B be the event that it rains. Then the pair B, B^c is a partition of sample space (since one or the other must occur)

3 Apply them to find that:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$
$$= \frac{1}{5}\cdot\frac{2}{5} + \frac{3}{5}\cdot\frac{3}{5} = \frac{11}{25}.$$

1.4 Probability measures and continuity

A sequence $A_1, A_2 \ldots$ of events in probability space $(\Omega, \mathcal{F}, P)$ is called increasing if

$$A_1 \subseteq A_2 \subseteq \ldots \subseteq A_{n-1} \subseteq A_n \subseteq A_{n+1} \ldots$$

The union: $A = \bigcup_{i=1}^{\infty} A_i$

of such a sequence is called limit of the sequence, and it is an elementary consequence of the axioms for an event space that A is an event. It is not surprising that the probability $P(A)$ of A may be expressed as $\lim_{n\to\infty} P(A_n)$ of the probabilities of the A's

Theorem: Let $(\Omega, \mathcal{F}, P)$ be a probability space. If $A_1, A_2 \ldots$ is an increasing sequence of events in $\mathcal{F}$ with limit A, then:

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

CHAPTER 2: DISCRETE RANDOM VARIABLES

2.1 Probability mass functions

Given a probability space $(\Omega, \mathcal{F}, P)$, we are often interested in situations involving some real-valued function X on Ω

· For example let $\mathcal{E}$ be the experiment of throwing a fair die once, so that $\Omega = \{1,2,3,4,5,6\}$ and suppose we gamble on outcome of $\mathcal{E}$ in such a way that our profit is

-1 if 1, 2 or 3
0 if 4
2 if 5 or 6

If the outcome is w then our profit is $X(w)$ where $X: \Omega \to \mathbb{R}$ defined by:

$X(1) = X(2) = X(3) = -1 \qquad X(4) = 0 \qquad X(5) = X(6) = 2$

The mapping X is an example of a discrete random variable

More formally, a discrete random variable X on the probability space $(\Omega, \mathcal{F}, P)$ is defined to be a mapping of $X: \Omega \to \mathbb{R}$ such that

the image $X(\Omega)$ of Ω under X is a countable subset of $\mathbb{R}$ 1.

$\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$ 2

The word discrete here refers to the condition that X takes only countably many values in $\mathbb{R}$

Condition (2): A discrete random variable X takes values in $\mathbb{R}$ but we cannot predict the actual value of X with certainty since the underlying experiment $\mathcal{E}$ involves chance; instead we would like to assess the probability that X takes the value x if and only if the result of $\mathcal{E}$ lies in that subset of Ω which is mapped onto x namely $X^{-1}(x) = \{\omega \in \Omega : X(\omega) = x\}$

Condition (2) postulates that all such subsets are events, in that they belong to $\mathcal{F}$, and are therefore assigned probabilities by P.

If X is a discrete random variable on the probability space $(\Omega, \mathcal{F}, P)$, then the image $\operatorname{Im} X$ of X is the image of Ω under X; the image of X is the set of values taken by X. The (probability) mass function p_X of X is the function which maps $\mathbb{R}$ into $[0,1]$ defined by:

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

3

Thus, $p_X(x)$ is the probability that the mapping X takes the value x. We usually abbreviate $P(X = x)$

Note that $\operatorname{Im} X$ is countable for any discrete random variable X, and

$p_X(x) = 0$ if $x \notin \operatorname{Im} X$

5

$$\sum_{x \in \operatorname{Im} X} p_X(x) = P\left(\bigcup_{x \in \operatorname{Im} X} \{ w \in \Omega : X(w) = x \} \right) \quad 5.$$
$$= P(\Omega) = 1$$

Eq^n (5) sometimes wrote as: $\sum_{x \in \mathbb{R}} p_X(x) = 1$
in light of the fact that only countable many values of x make non-zero contributions to this sum

Theorem 2.4: If $S = \{ s_i : i \in I \}$ is a countable set of ~~distinct~~ distinct real numbers and $\{ \pi_i : i \in I \}$ is a collection of numbers satisfying: $\pi_i \geq 0$ for all $i \in I$ and $\sum_{i \in I} \pi_i = 1$

(small s)

then there exists a probability space $(\Omega, \mathcal{F}, P)$ and a discrete random variable X on $(\Omega, \mathcal{F}, P)$ such that the probability mass function $p_X(x)$ of X is given by:

$p_X(s_i) = \pi_i$ for all $i \in I$

$p_X(s) = 0$ if $s \notin S$

EXAMPLES:

There are certain classes of discrete random variables which occur very frequently

- Throughout this section n is a positive integer, p is a number in $[0,1]$ and $q = 1 - p$. We never describe the underlying probability space

Bernoulli Distribution

Discrete R.V X has the bernoulli distribution with parameter p if the image of X is $\{0, 1\}$, so that X takes value 0 and 1 only and

$P(X = 0) = q \quad P(X = 1) = p$ 6

(6) asserts that the mass function of X is given by:
$p_X(0) = q, \ p_X(1) = p \quad p_X(x) = 0$ if $x \neq 0, 1$.

Poisson Distribution

X has poisson dist with parameter $\lambda (>0)$ if X has image $\{0,1,2 \dots\}$ and $p(X=k) = \frac{1}{k!}\lambda^k e^{-\lambda}$ for $k=0,1,2 \dots$

Again this gives rise to a mass function since:
$$\sum_{k=0}^{\infty} \frac{1}{k!}\lambda^k e^{-\lambda} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{1}{k!}\lambda^k = e^{-\lambda}e^{\lambda} = 1$$

Binomial Distribution

X has binomial dist with parameter n and p if X has image $\{0,1 \dots n\}$ and $P(X=k) = \binom{n}{k}p^k q^{n-k}$ for $k=0,1,2 \dots n$

Gives rise to mass function: $\sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k} = (p+q)^n = 1$

Geometric Distribution

X has geometric distribution with parameter $p(>0)$ if X has image $\{1,2,3 \dots\}$ and $P(X=k) = pq^{k-1}$ for $k=1,2 \dots$

$$\sum_{k=1}^{\infty} pq^{k-1} = \frac{p}{1-q} = 1$$

Example

Suppose that a coin is tossed n times and there is probability p that heads appears on each toss. H (head) and T (tail), Sample Space is the set Ω of all ordered sequences of length n containing H's and T's, where k^{th} entry of such a sequence represents result of kth toss. Ω is finite, we take $\mathbb{F}$ to be the set of all subsets of Ω. For each $\omega \in \Omega$ we define the probability that ω is the actual outcome by:
$$P(\omega) = p^{h(\omega)}q^{t(\omega)}$$
where $h(\omega)$ is number of heads in ω and $t(\omega)$ $n-h(\omega)$ is number of tails

Similarly for any $A \in \mathbb{F}$ $\quad P(A) \sum_{\omega \in A} P(\omega)$ for $i=1,2 \dots n$

7

We define the discrete random variable X_i by:

$$X_i(\omega) = \begin{cases} 1 & \text{if } i\text{th entry in } \omega \text{ is } H \\ 0 & \text{if } i\text{th entry in } \omega \text{ is } T \end{cases}$$

Then each X_i has image $\{0,1\}$ and mass function given by:

$$P(X_i = 0) = P(\{\omega \in \Omega : \omega_i = T\})$$

where ω_i is the ith entry in ω. Then:

$$P(X_i = 0) = \sum_{\omega : \omega_i = T} p^{h(\omega)} q^{n-h(\omega)}$$

$$= \sum_{h=0}^{n-1} \sum_{\substack{\omega : \omega_i = T \\ h(\omega) = h}} p^h q^{n-h}$$

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

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

and $P(X_i = 1) = 1 - P(X_i = 0) = p$

Hence, each X_i has Bernoulli distribution with parameter p

Let $S_n = X_1 + \dots X_n$;
more formally, $S_n(\omega) = X_1(\omega) + \dots X_n(\omega)$ (clearly S_n is the total number of heads which occur, and S_n takes values in $\{0, 1 \dots n\}$ since each X_i equals 0 or 1. Also for $k = 0, 1 \dots n$ we have:

$$P(S_n = k) = P(\{\omega \in \Omega : h(\omega) = k\})$$

$$= \sum_{\omega : h(\omega) = k} P(\omega)$$

$$= \binom{n}{k} p^k q^{n-k}$$

and so S_n has binomial dist with parameters n and p

8.

If n is very large and p is very small but np is a 'reasonable size' ($np = \lambda$ say), then distribution of S_n may be approximated by the Poisson dist with parameter λ as follows

For fixed $k \geq 0$, write $p = \lambda n^{-1}$ and suppose n is large, to find:

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

$$\approx \frac{n^k}{k!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k}$$

$$\approx \frac{1}{k!}\lambda^k e^{-\lambda}$$

Example: Single newspaper page containing 10^6 characters, suppose typesetter flips a coin before setting each character and then deliberately mis-sets this character whenever the coin comes up heads

If coin comes up heads with $p = 10^{-5}$ on each flip then equivalent to taking $n = 10^6$ and $p = 10^{-5}$, giving number S_n of deliberate mistakes has binomial distribution with parameters 10^6 and 10^{-5}.

Easier to use poisson, $\lambda = np = 10$ and so:

$$P(S_n = 10) \approx \frac{1}{10!}(10e^{-1})^{10} \approx 0.125$$

Example:

Suppose that we toss the coin of the previous example until a first head turns up, and then we stop. Sample space is now:

$\Omega = \{H, TH, T^2H, \dots\} \cup \{T^\infty\}$

where T^kH represents the outcome of k tails followed by a head and T^∞ represents an infinite sequence of tails with no head

As before $\mathcal{F}$ is the set of all subsets of Ω and P is given by

9

$P(T^k H) = pq^k$ for $k = 0, 1, 2 \ldots$

$P(T^\infty) = \begin{cases} 1 & \text{if } p=0 \\ 0 & \text{if } p>0 \end{cases}$

If Y is total number of tosses in this experiment (so that $Y(T^k H) = k+1$ for $0 \le k < \infty$ and $Y(T^\infty) = \infty$). If $p > 0$ then

$P(Y = k) = P(T^{k-1} H) = pq^{k-1}$ for $k = 1, 2 \ldots$

Showing that Y has the geometric dist. with parameter p

2.3 Functions of Discrete Random Variables

Suppose that X is a discrete r.v on probability space $(\Omega, \mathbb{F}, P)$ and that $g: \mathbb{R} \to \mathbb{R}$. It is easy to check that $Y = g(X)$ is a discrete random variable on $(\Omega, \mathbb{F}, P)$ also defined by:

$Y(\omega) = g(X(\omega))$ for all $\omega \in \Omega$

Simple examples:

If $g(x) = ax+b$ then $g(X) = aX+b$

If $g(x) = cx^2$ then $g(X) = cX^2$

If $Y = g(X)$ then mass function of Y is given by

$$P_Y(y) = P(Y=y) = P(g(X)=y) \quad \text{inverse both sides}$$
$$= P(X \in g^{-1}(y))$$
$$= \sum_{x \in g^{-1}(y)} P(X=x)$$

Since there are only countably many non-zero contributions to this sum

Thus if $Y = aX+b$ where $a \neq 0$ then

$P(Y=y) = P(aX+b=y) = P(X = a^{-1}(y-b))$ for all y

while if $Y = X^2$ then:

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

12

2.4 Expectation

For die if it were thrown a large number of times, each of possible outcome 1, 2 .. 6 would appear 1/6 of the

Average would be $\frac{1}{6}\cdot 1+\frac{1}{6}\cdot 2 \dots +\frac{1}{6}\cdot 6 = 3.5$

Which we call the mean value

If X is a discrete random variable, the expectation of X is denoted by $E(X)$ and defined by

$$E(X) = \sum_{x \in \operatorname{Im} X} x\, P(X=x)$$

Whenever this sum converges absolutely (in that $\sum_x |x P(X=x)| < \infty$)

Often write: $E(X) = \sum_x x P(X=x) = \sum_x x\, p_X(x)$ and the expectation of X is often called expected value or mean of X

- If X is a discrete r.v and $g: \mathbb{R} \to \mathbb{R}$ then $Y = g(X)$ is a discrete r.v also
- According to above definition we need to know the mass function of Y before we can calculate its expectation

Theorem: If X is a discrete r.v and $g: \mathbb{R} \to \mathbb{R}$ then

$$E(g(X)) = \sum_{x \in \operatorname{Im} X} g(x) P(X=x)$$

whenever this sum converges absolutely

Since $g(X)$ takes the value $g(x)$ when X takes value x, an event which has probability $P(X=x)$

Example: Suppose X is a r.v with Poisson dist parameter λ and want to find expected value of $Y = e^X$.

$E(Y) = E(e^X)$

$= \sum_{n=0}^{\infty} e^n\, P(X=n)$

11

$= \sum_{k=0}^{\infty} e^{k} \frac{1}{k!}\lambda^{k} e^{-\lambda}$

$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{1}{k!}(\lambda e)^{k}$

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

The expectation $E(X)$ of a discrete random variable X is an indicator of the 'centre' of the distribution of X

Variance is a measure of the degree of dispersion of X about its expectation $E(X)$

Formally, the variance of a discrete r.v X is defined to be the expectation of $[X - E(X)]^2$ the variance of X is usually written $var(X) = E([X-E(X)]^2)$

$var(X) = \sum_{x \in Im X} (x-\mu)^2 P(X=x)$

Where $\mu = E(X) = \sum_{x \in Im X} x P(X=x)$

If dispersion of X about its expectation is very small then $|X-\mu|$ tends to be small, giving that $var(X) = E(|X-\mu|^2)$ is small also, same for large

$var(X) = \sum_{x} \overbrace{(x^2 - 2\mu x + \mu^2)}^{(x-\mu)^2} P(X=x)$

$= \sum_{x} x^2 P(X=x) - 2\mu \sum x P(X=x) + \mu^2 \sum P(X=x)$ $(E(X) = \sum x P(X=x))$

$= E[X^2] - 2\mu^2 + \mu^2$ $(\sum P(X=x) = 1)$

$= E[X^2] - \mu^2$ where $\mu = E(X)$

$\Rightarrow var(X) = E(X^2) - E(X)^2$

12

If X has geometric dist with parameter $p = (1-q)$ then mean of X is

$$E(X) = \sum_{k=1}^{\infty} k p q^{k-1}$$

$$\text{using } \frac{p}{1-q} \qquad = \frac{p}{(1-q)^2} = \frac{1}{p}$$

and variance is $\mathrm{var}[X] = \sum_{k=1}^{\infty} k^2 p q^{k-1} - \frac{1}{p^2}$

$$\text{Here } \sum_{k=1}^{\infty} k^2 q^{k-1} = q \sum_{k=1}^{\infty} k(k-1) q^{k-2} + \sum_{k=1}^{\infty} k q^{k-1}$$

$$= \frac{2q}{(1-q)^3} + \frac{1}{(1-q)^2}$$

$$\text{gives } \mathrm{var}(X) = p\left(\frac{2q}{p^3} + \frac{1}{p^2}\right) - \frac{1}{p^2}$$

$$= q p^{-2}$$

Conditional expectation and the Partition Theorem

X discrete r.v. $P(B) > 0$. If we are given B occurs, this info affects the probability distribution of X, that is, probabilities such as $P(X=x)$ are replaced by conditional probabilities such as $P(X=x \mid B) = P(\{\omega \in \Omega : X(\omega) = x\} \cap B) / P(B)$

If X is a discrete r.v and $P(B) > 0$, then the conditional expectation of X given B is denoted by $E(X \mid B)$ and defined by

$$E(X \mid B) = \sum_{x \in \mathrm{Im}} x P(X = x \mid B)$$ whenever sum converges absolutely

Theorem: If X is a discrete R.V and $\{B_1, B_2, \ldots\}$ is a partition of the sample space such that $P(B_i) > 0$ for each i, then

$$E(X) = \sum_i E(X \mid B_i) P(B_i)$$

whenever this sum converges absolutely.

26

13

RHS of (26) equals:-

$$\sum_i \sum_x x P(\{X=x\} \cap B_i) = \sum_x x P(\{X=x\} \cap (\cup B_i))$$

$$= \sum_x x P(X=x)$$

CHAPTER 3 MULTIVARIATE DISCRETE DISTRIBUTIONS AND INDEPENDENCE

Bivariate Discrete Distributions

Let X and Y be discrete random variables on the probability space $(\Omega, \mathcal{F}, P)$. Instead of treating X and Y seperately, it is often necessary to regard the pair (X, Y) as a random vector taking values in $\mathbb{R}^2$

If X and Y are discrete random variables on $(\Omega, \mathcal{F}, P)$, the joint probability mass function p_{XY} of X and Y is the function $p_{XY}: \mathbb{R}^2 \to [0,1]$ defined by

$$p_{XY}(x,y) = P(\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\})$$

Thus $p_{XY}(x,y)$ is the probability that $X=x$ and $Y=y$ and we often write this as $p_{XY} = P(X=x, Y=y)$

If p_{XY} is the joint mass function of X and Y then it is clear that
$p_{XY} = 0$ unless $x \in \operatorname{Im} X$ and $y \in \operatorname{Im} Y$

$$\sum_{x \in \operatorname{Im} X} \sum_{y \in \operatorname{Im} Y} p_{XY}(x,y) = 1$$

The individual mass functions p_X and p_Y of X and Y may be found from p_{XY} thus $p_X(x) = P(X=x) = \sum_{y \in \operatorname{Im} Y} P(X=x, Y=y)$

$$= \sum_y p_{XY}(x,y)$$

$$= \sum_x p_{XY}(x,y)$$

14

These mass functions, given by (4) and (5) are called the marginal mass function of X and Y respectively since, if we think of (X, Y) as a randomly chosen point in the plane, then X and Y are the projections of this point onto the coordinate axis

Example Suppose that X and Y are random variables each taking values 1, 2 or 3, and that the probability that the pair $(X,Y) = (x,y)$ is given in following table for all relevant values of x and y

$x \backslash y$	1	2	3
1	$\frac{1}{12}$	$\frac{1}{18}$	0
2	$\frac{3}{18}$	0	$\frac{3}{18}$
3	$\frac{1}{6}$	$\frac{5}{18}$	$\frac{1}{12}$

Then for example $P(X=3) = P(X=3, Y=1) + P(X=3, Y=2) + P(X=3, Y=3)$
$= \frac{1}{6} + \frac{5}{18} + \frac{1}{12} = \frac{19}{36}$

Similarly $P(Y=2) = \frac{1}{18} + 0 + \frac{5}{18} = \frac{1}{3}$

Similar ideas apply to families $X = (X_1, X_2 \dots X_n)$ of discrete random variables on a probability space. For example, the joint mass function of X is a function p_X defined by:

$$p_X(x) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \quad \text{for } x = (x_1, x_2, x_3 \dots x_n) \in \mathbb{R}^n$$

Expectation in the multivariate case

If X and Y are discrete r.v on $(\Omega, \mathbb{F}, P)$ and $g: \mathbb{R}^2 \to \mathbb{R}$ then it is easy to check that $Z = g(X,Y)$ is a discrete random variable on $(\Omega, \mathbb{F}, P)$ also, defined formally by $Z(\omega) := g(X(\omega), Y(\omega))$ for $\omega \in \Omega$

Expectation of Z may be calculated directly from the joint mass function $p_{X,Y}(x,y) = P(X=x, Y=y)$ of X and Y as following theorem indicates

15.

Theorem 3.x. $E(g(X,Y)) = \sum_{x \in Im} \sum_{y \in Y} g(x,y) P(X=x, Y=y)$

Whenever this sum converges absolutely

The expectation operator E acts linearly on the set of discrete random variables. That is to say, if X and Y are discrete random variables on $(\Omega, \mathbb{F}, P)$ and $a, b \in \mathbb{R}$ then

$$E(aX+bY) = aE(X) + bE(Y)$$

Whenever $E(X)$ and $E(Y)$ exist.

Independence of discrete random variables

In a probability space events A and B are independent if $P(A \cap B) = P(A)P(B)$.

Discrete rv X and Y on $(\Omega, \mathbb{F}, P)$ are called independent if the value taken by X is independent of the value taken by Y.

That is to say, X and Y are independent if pair of events $\{w \in \Omega : X(w) = x\}$ and $\{w \in \Omega : Y(w) = y\}$ are independent for all $x, y \in \mathbb{R}$, and we normally write this condition as

$$P(X=x, Y=y) = P(X=x)P(Y=y) \text{ for all } x, y \in \mathbb{R} \quad (8)$$

r.v which aren't independent are dependent

Condition (8) may be expressed as

$$P_{X,Y}(x,y) = \left(\sum_y P_{X,Y}(x,y)\right)\left(\sum_x P_{X,Y}(x,y)\right) \text{ for all } x, y \in \mathbb{R}$$

Theorem: Discrete r.v X and Y are independent if and only if there exist functions $f, g : \mathbb{R} \to \mathbb{R}$ such that the joint mass function of X and Y satisfies

$P_{X,Y}(x,y)$ [illegible] for all $x, y \in \mathbb{R}$

16

Theorem: If X and Y are independent discrete r.v with expectation $E(X)$ and $E(Y)$ then:

$$E(XY) = E(X)E(Y)$$

$$E(XY) = \sum_{x,y} xy P(X=x, Y=y)$$

$$\sum xy P(X=x)P(Y=y) \quad \text{by independence}$$

$$\sum_x xP(X=x) \sum_y yP(Y=y)$$

$$= E(X) \, E(Y)$$

It is the existence of $E(X)$ and $E(Y)$ which allows us to interchange the summations as we have done.

Converse NOT true

Example: Suppose X has dist given by

$$P(X=-1) = P(X=0) = P(X=1) = 1/3$$

and Y is given by $Y = \begin{cases} 0 & \text{if } X=0 \\ 1 & \text{if } X \neq 0 \end{cases}$

Easy to find probability space $(\Omega, \mathcal{F}, P)$ together with two random variables having these distributions.

For example take $\Omega = \{-1, 0, 1\}$, $\mathcal{F}$ the set of all subsets of Ω, P given by $P(-1) = P(0) = P(1) = 1/3$ and $X(\omega) = \omega$, $Y(\omega) = |\omega|$. Then X and Y are dependent since $P(X=0, Y=1) = 0$

but $P(X=0)P(Y=1) = \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{9}$

On the other hand, $E(XY) = \sum_{x,y} xy P(X=x, Y=y)$

$$= (-1) \cdot \tfrac{1}{3} + 0 \cdot \tfrac{1}{3} + 1 \cdot \tfrac{1}{3} = 0$$

and $E(X)E(Y) = 0 \cdot \frac{2}{3} = 0.$

The family is called pairwise independent if X_i and X_j are independent whenever $i \neq j$.

17.

Sum of random variables

If X and Y are discrete r.v with a certain joint mass function what is the mass function $Z = X+Y$?

$Y = z - x$

Clearly Z takes the value z if and only if $X = x$ and $Y = y$ for some value of x

$$P(Z=z) = P\left(\bigcup_x (\{X=x\} \cap \{Y = z-x\})\right)$$

$$= \sum_{x \in \operatorname{Im} X} P(X=x, Y=z-x) \quad \text{for } z \in \mathbb{R}$$

If X and Y are independent then their joint mass function factorises and we obtain the following result.

Theorem: If X and Y are independent discrete r.v on $(\Omega, \mathbb{F}, P)$ the $Z = X+Y$ has mass function

$$P(Z=z) = \sum_{x \in \operatorname{Im} X} P(X=x)\, P(Y=z-x) \quad \text{for } z \in \mathbb{R}$$

In language of analysis, formula says that mass function of X and Y is the convolution of the mass functions of X and Y.

CHAPTER 5 DISTRIBUTION FUNCTIONS AND DENSITY FUNCTION

Distribution Functions

Discrete r.v may take only countably many values. This condition is too restrictive for many situations, and accordingly we make a broader definition: a random variable X on the probability space $(\Omega, \mathbb{F}, P)$ is a mapping $X: \Omega \to \mathbb{R}$ such that

(1) $\{\omega \in \Omega : X(\omega) \le x\} \in \mathbb{F}$ for all $x \in \mathbb{R}$

- We are interested in the values taken by a random variable X and the likelihood of these values

18.

It turns out the right way to do this is to fix $x \in \mathbb{R}$ and ask for the probability that X takes a value in $(-\infty, x]$; this probability exists only if the inverse image $X^{-1}((-\infty, x]) = \{\omega \in \Omega : X(\omega) \leq x\}$ of $(-\infty, x]$ lies in the event space $\mathbb{F}$, and so we postulate this is the case for all $x \in \mathbb{R}$.

Note that every discrete r.v. X is a r.v. To see this, observe that if X is a discrete random variable then:

$$\{\omega \in \Omega : X(\omega) \leq x\} = \bigcup_{\substack{y \in \operatorname{Im} X \\ y \leq x}} \{\omega \in \Omega : X(\omega) = y\}$$

which is the countable union of events in $\mathbb{F}$ and therefore belongs to $\mathbb{F}$

Whereas discrete random variables were studied through their mass functions, random variables in the broader sense are studied through their distribution functions, defined as follows

If X is a r.v. on $(\Omega, \mathbb{F}, P)$ the distribution function of X is a mapping $F_X : \mathbb{R} \to [0,1]$ defined by:

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

We often write as $F_X(x) = P(X \leq x)$ CDF

Example Suppose X is a discrete R.V. taking non negative integer values, with mass function

$$P(X = k) = p(k) \quad \text{for } k = 0, 1, 2, \ldots$$

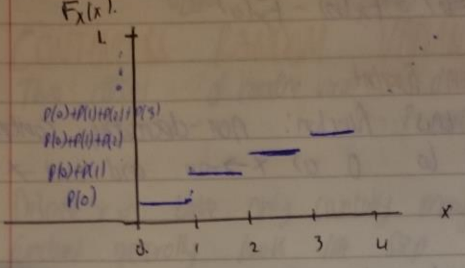
For any $x \in \mathbb{R}$ it is the case that $X \leq x$ if and only if X takes one of the values $0, 1, 2, \ldots, \lfloor x \rfloor$, where $\lfloor x \rfloor$ denotes the greatest integer not greater than x.

Hence

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ p(0) + p(1) + \ldots p(\lfloor x \rfloor) & \text{if } x \geq 0 \end{cases}$$

19

Sketch of function



Distribution function of a random variable which takes values in the non-negative integers

The distribution function F_X of a r.v X has various general and elementary properties, the first which is

$$F_X(x) \le F_X(y) \text{ if } x \le y$$

Which is to say that F_X is monotonic non-decreasing. This holds because

$$\{\omega \in \Omega : X(\omega) \le x\} \subseteq \{\omega \in \Omega : X(\omega) \le y\}$$

whenever $x \le y$, since if X takes on a value smaller than x then this value is certainly smaller than y.

Other elementary properties of $F_X(x)$ concern its behaviour when x is near $-\infty$ or $+\infty$: $F_X(x) \to 0$ as $x \to -\infty$

and $F_X(x) \to 1$ as $x \to \infty$

In first case as $x \to -\infty$ the event X is smaller than x becomes less likely, opposite for $x \to \infty$

$$P(X \le -\infty) = 0 \qquad P(X \le \infty) = 1$$

The probability $F_X(x) = P(X \le x)$ is the probability that X takes a value in the infinite interval $(-\infty, x]$

To find the probability that X takes a value in the bounded interval $(a, b]$ we proceed in following way. Let $a < b$

Then: $P(a < X \le b) = P(\{X \le b\} \setminus \{X \le a\})$

$$= P(X \le b) - P(X \le a)$$

22

Since the event $\{X \le a\}$ is a subset of the event $\{X \le b\}$
Hence $P(a < X \le b) = F_X(b) - F_X(a)$

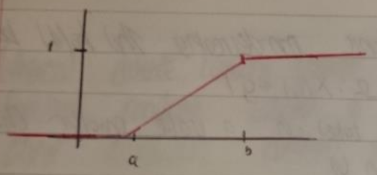
Examples of Distribution functions

General features of this function: non-decreasing, continuous from the right, tending to 0 as $x \to -\infty$ and 1 as $x \to \infty$

Uniform distribution

Let $a, b \in \mathbb{R}$ and $a < b$ the function

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \le x < b \\ 1 & \text{if } x \ge b \end{cases}$$



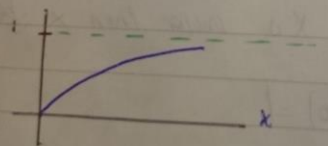
uniform distribution function

Exponential distribution:

Let $\lambda > 0$ and F be given by

$$F(x) = \begin{cases} 0 & \text{if } x \le 0 \\ 1 - e^{-\lambda x} & x > 0 \end{cases}$$

Exponential distribution with parameter λ.



Any non-negative function F, which is continuous and non decreasing and satisfies

$$\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1 \quad \text{is a}$$

distribution function

$P\frac{0.4}{460}$ or $P\frac{X<340}{X<460}$

CONTINUOUS RANDOM VARIABLES

Two classes of random variable 1. discrete r.v
2. continuous r.v.

Discrete r.v take only countably many values and their distribution functions generally look like step functions

R.V whose dist. f^n are smooth - are "continuous"

More formally, we call a r.v. X continuous if its dist. f^n F_X may be written in form of:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(u)\,du \quad \text{for all } x \in \mathbb{R}$$

for some non negative function f_X. In this case we say that X has probability density function f_X

Example: If X has exponential distribution with parameter λ then

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1-e^{-\lambda x} & \text{if } x > 0 \end{cases}$$

and

$$f_X(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \lambda e^{-\lambda x} & \text{if } x > 0 \end{cases}$$

Provided that X is a continuous r.v and F_X is well behaved we can take: $f_X(x) = \begin{cases} \frac{d}{dx} F_X(x) & \text{if this derivative exists at } x. \\ 0 & \text{otherwise} \end{cases}$

as density function of X

It is clear that the density function f_X of X satisfies:

$f_X(x) \geq 0$ for all $x \in \mathbb{R}$ $(p_X(x) \geq 0)$

$\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ $\left(\sum_x p_X(x) = 1\right)$

where brackets represent contain corresponding property of mass function p_X

However, this analogy can be dangerous, since $f_X(x)$ is NOT a probability and may well exceed 1 in value.

On the other hand, $f_X(x)$ is indeed a 'measure' of probability in the following sense:

If dx is small and positive then, roughly speaking, the probability that X is "near" to x satisfies

$$P(x < X \le x+dx) = F(x+dx) - F(x)$$
$$= \int_x^{x+dx} f_X(u)\,du$$
$$\approx f_X(x)\,dx \quad \text{for small } dx$$

So the true analogy is not between a density function $f_X(x)$ and a mass function $p_Y(x)$, but between $f_X(x)\,dx$ and $p_Y(x)$.

Values of mass function are replaced by $f_X(x)\,dx$ and the summation is replaced by the integral

Difference between discrete and continuous r.v is given in the first part of next theorem

Theorem: If X is continuous with density function f_X then

$$P(X=x) = 0 \quad \text{for all } x \in \mathbb{R}$$
$$P(a \le X \le b) = \int_a^b f_X(u)\,du \quad \text{for all } a,b \in \mathbb{R} \text{ with } a \le b$$

Proof: $P(X=x) = \lim_{\varepsilon \to 0} P(x-\varepsilon < X \le x)$

$$= \lim_{\varepsilon \to 0} \left(F_X(x) - F_X(x-\varepsilon)\right)$$

23

$$\lim_{\varepsilon \to 0} \int_{x-\varepsilon}^{x} f_X(u)\,du$$

$= 0.$

If $a \leq b$ then
$$P(a < X \leq b) = P(a \leq X \leq b) = F_X(b) - F_X(a)$$

$$= \int_a^b f_X(u)\,du$$

All r.v have a distribution function and continuous r.v have a density function and discrete have mass function

Some Common Density Functions

It is clear that any function f which satisfies
$f(x) \geq 0$ for all x
and $\int_{-\infty}^{\infty} f(x)\,dx = 1$

is the density function of some r.v. To confirm this define:
$F(x) = \int_{-\infty}^{x} f(u)\,du$
and check that F is a distribution function.

The uniform distribution on (a,b) has density function

$$f(x) = \begin{cases} (b-a)^{-1} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

The exponential distribution with parameter $\lambda > 0$ has density function:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases}$$

The normal dist with parameter μ and σ^2 $N(\mu, \sigma^2)$ has density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]} \quad \text{for } -\infty < x < \infty$$

2.4

Functions of random variables

X r.v on $(\Omega, \mathbb{F}, P)$ and suppose $g: \mathbb{R} \Rightarrow \mathbb{R}$

Then $Y = g(X)$ is a mapping from Ω into $\mathbb{R}$ defined by

$$Y(\omega) = g[X(\omega)] \quad \text{for } \omega \in \Omega$$

Assume henceforth that all quantities of the form $Y = g(X)$ are r.v's

Main question now, if we know the dist of X, then how do we find dist of $Y = g(X)$?

If X is discrete 2.18 provides answer

We consider case where X is continuous with density function f_X.

Example: If X is continuous with d.f F_X and $g(x) = ax+b$ where $a > 0$, then $Y = g(X) = aX + b$ has distribution function given by:

$$\begin{aligned} P(Y \leq y) &= P(aX + b \leq y) \\ &= P(X \leq a^{-1}(y-b)) \\ &= F_X(a^{-1}(y-b)) \end{aligned}$$

and differentiation with respect to y yields:

$$f_Y(y) = a^{-1} f_X(a^{-1}(y-b)) \quad \text{for all } y \in \mathbb{R}$$

Theorem: If X is a continuous r.v with density function f_X, and g is a strictly increasing and differentiable function from $\mathbb{R} \Rightarrow \mathbb{R}$, then $Y = g(X)$ has density function:

$$f_Y(y) = f_X(g^{-1}(y)) \frac{d}{dy}[g^{-1}(y)] \quad \text{for } y \in \mathbb{R} \qquad (32)$$

where g^{-1} is the inverse function of g

Proof: First we find distribution function of Y

$$\begin{aligned} P(Y \leq y) &= P(g(X) \leq y) \\ &= P(X \leq g^{-1}(y)) \quad \text{since } g \text{ is increasing} \end{aligned}$$

We differentiate with respect to y to obtain: (32)

If g were strictly decreasing then the same argument gives that $Y = g(X)$ has density function

$$f_Y(y) = -f_X(g^{-1}(y)) \frac{d}{dy}[g^{-1}(y)] \quad \text{for } y \in \mathbb{R}$$

Example If X has d.f F_X and $g(x) = x^2$, then $Y = g(X) = X^2$ has d.f: $P(Y \le y) = P(X^2 \le y)$

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

Hence $f_Y(y) = 0$ if $y < 0$ while for $y > 0$:

$$f_Y(y) = \frac{d}{dy} P(Y \le y) \quad \text{if derivative exists}$$

$$\frac{d}{dy}[F_X(\sqrt{y}) - F_X(-\sqrt{y})]$$

$$= \frac{1}{2\sqrt{y}}[f_X(\sqrt{y}) + f_X(-\sqrt{y})]$$

Expectation of continuous random variables

If X is a continuous r.v with d.f f_X then the expectation of X is denoted by $E(X)$ defined by:

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

whenever this integral converges absolutely (in that $\int_{-\infty}^{\infty} |x f_X(x)|\, dx < \infty$)

As in case of drv, the expectation of X is often called the expected value or mean of X.

Theorem If X is a crv. with d.f. f_X and $g: \mathbb{R} \to \mathbb{R}$ then

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

whenever this integral converges absolutely

28.

As in case of d.r.v. the mean $E(X)$ of a c.r.v X is an indication of the 'centre' of the distribution of X.

As a measure of the degree of ~~variability~~ dispersion of X about this mean we normally take the variance of X.

Example If X has uniform dist on (a,b) then
$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx$$
$$= \int_a^b x \frac{1}{b-a}\,dx$$
$$= \frac{1}{2}(a+b)$$

As example, let $Y = \sin X$
then $$E(Y) = \int_{-\infty}^{\infty} \sin x f_X(x)\,dx$$
$$\int_a^b \sin x \frac{1}{b-a}\,dx$$
$$= \frac{\cos a - \cos b}{b-a}$$

Example: If X has exponential dist. with parameter λ then
$$E(X) = \int_0^{\infty} x \lambda e^{-\lambda x}\,dx = \frac{1}{\lambda}$$
and $$E(X^2) = \int_0^{\infty} x^2 \lambda e^{-\lambda x}\,dx = \frac{2}{\lambda^2}$$
var $= \frac{1}{\lambda^2}$

Example: If X has normal dist, parameters $\mu = 0$, $\sigma^2 = 1$
$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\,dx = 0$$
by symmetry properties of the integrand. Here
$$\text{var}(X) = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\,dx = 1$$

29.

Similar integrals) show normal dist with params μ and r^2 [illegible]

Markov Chains

When the probability from going from one state to another state depends only on the current state of the system and is not influenced by additional information.

The probability model with this feature is called Markov.

Markov model

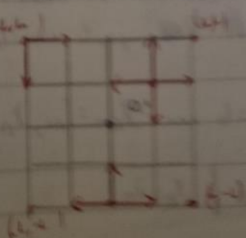
A Markov chain deals with a collection of random variables, indexed by an ordered time parameter.

It is the simplest conceivable generalisation of a sequence of independent random variables.

A Markov chain is a sequence of trials having the property that the outcome of each last trial provides enough information to predict the outcome of any future trial.

Example

Drunk wanders around square. At each step no longer remembers the direction of previous step. Each step is a unit distance in a randomly chosen direction and has equal probability ¼ for N S E W. as long as not reached edge of square. Doesn't never leave square. If at edge equally likely to go other 3 directions. At corner, equally likely to go other 2 directions. The drunk starts in middle of square. What stochastic (random) process describes the walk?



2

Solution: We define random variable X_n as
X_n = position of drunk after n^{th} step
For $n = 0, 1 \ldots$ and convention $X_0 = (0,0)$

We say that the drunkard is in state (x,y) when the current position of the drunkard is described by point (x,y).

The collection $\{X_0, X_1, \ldots\}$ of random variables is a stochastic process with discrete time-parameter and finite state space
$I = \{(x,y): x,y \text{ integer and } -L \leq (x, y) \leq L\}$
where L is the distance from middle of square to its boundary

The successive states of the drunk are not independant of each other, but the next position depends only on his current position and is not influenced by the earlier positions in his path

That is, the process $\{X_0, X_1, \ldots\}$ has so called markovian property, which says at any given time, summarises everything about the past that is relevant to the future

$X_0, X_1, \ldots$ sequence of random variables
X_n state at time $t = n$.
set of possible values of the random variables X_n finite and I.
set I called the state space of the stochastic process $\{X_0, X_1 \ldots\}$

Definition The stochastic process $\{X_n, n = 0, 1, \ldots\}$ with state space I is said to be a (discrete-time) Markov chain if it possessed the Markovian property, that is, for each time point $n = 0, 1 \ldots$ and all possible values of the states $i_0, i_1, \ldots i_{n+1} \in I$ the process has property:

$$P(X_{n+1} = i_{n+1} \mid X_0 = i_0, X_1 = i_1 \ldots X_{n-1} = i_{n-1}, X_n = i_n)$$
$$= P(X_{n+1} = i_{n+1} \mid X_n = i_n)$$

Markov 2

Markovian property says that this conditional probability depends only on the current state i_n and is not altered by knowledge of the past states $i_0, i_1 \dots i_{n-1}$.

current state summarize everything about the past that is relevant to the future

the following, we will restrict our attention to time homogeneous Markov chains. For such chains, the transition probability $P(X_{n+1} = j \mid X_n = i)$ does not depend on the value of the time parameter n and so

$$P(X_{n+1} = j \mid X_n = i) = P(X_1 = j \mid X_0 = i) \text{ for all } n.$$

$$\Rightarrow \quad p_{ij} = P(X_{n+1} = j \mid X_n = i).$$

The probabilities p_{ij} are called one step transition probabilities of the markov chain and are the same for all time points n.

They satisfy: $p_{ij} \geq 0$ for $i, j \in I$ and $\sum_{j \in I} p_{ij} = 1$ for all $i \in I$

Notation p_{ij} is <u>not</u> a joint probability, but a conditional probability. Same as $p_{ij} = p(j \mid i)$

A markov chain $\{X_n, n = 0, 1 \dots\}$ is completely determined by the probability distribution of the initial state X_0 and the one step transition probabilities p_{ij}. The application:

Choose the state variable(s) such that Markovian property holds

determine the one step transition probabilities p_{ij}.

4

Example: Two compartment A and B contain r particles. With the passage of every time unit, one of the particles is Selected at random and is removed from its compartment to the other. What stochastic process describes the contents of the compartments?

Solution: Take state of system to be number of particles in A. If compartment A contains i particles, then compartment B contains r-i particles. Define random variable X_n as

X_n = number of particles in A after n^{th} transfer

By the physical construction of the model with independent selections of a particle, the process $\{X_n\}$ satisfies the Markovian property and thus is a Markov chain.

The State space is $I = \{0, 1, \dots r\}$
The probability of going from state i to j in one step is zero unless $|i-j| = 1$.

The one step transition probability $P_{i,i+1}$ translates into the probability that the randomly selected particle belong to compartment B and $P_{i,i-1}$ translates into probability that randomly selected particle belong to compartment A

Thus for $1 \leq i \leq r-1$

$$P_{i,i+1} = \frac{r-i}{r} \quad P_{i,i-1} = \frac{i}{r}$$

amount in B, total, amount in A, total

Further $P_{01} = P_{r,r-1} = 1$ the other P_{ij} are zero

5

5 Markov

Transient Analysis of Markov Chains:

A markov chain $\{X_n, n=1,2..\}$ is completely determined by its one step transition probabilities p_{ij} and the probability distribution of initial state X_0

The transient analysis of a markov chain concerns the calculation of the so called n-step transition probabilities

The probability of going from state i to state j in the next n-transitions of the markov chain is easily calculated from the one step transition probabilities

For any $n=1,2,\dots$ the n-step transition probabilities $p_{ij}^{(n)}$ are defined by:

$$p_{ij}^{(n)} = P(X_n = j \mid X_0 = i) \quad \text{for } i,j \in I$$

Note that $p_{ij}^{(1)} = p_{ij}$. A basic result is given in the following rule:

Rule 15.1 (Chapman-Kolmogorov equation)

For any $n \geq 2$ $\quad p_{ij}^{(n)} = \sum_{k \in I} p_{ik}^{(n-1)} p_{kj}$ for all $i,j \in I$

This rule states that the probability of going from state i to state j in n transitions is obtained by summing the probabilities of the mutually exclusive events of going from state i to some state k in the first n-1 transitions and then going from state k to state j in the n^{th} transition

Using the law of conditional probabilities and invoking the Markovian property:

$$p_{ij}^{(n)} = P(X_n = j \mid X_0 = i)$$

$$= \sum_{k \in I} P(X_n = j \mid X_0 = i, X_{n-1} = k)\, P(X_{n-1} = k \mid X_0 = i)$$

$$= \sum_{k \in I} P(X_n = j \mid X_{n-1} = k)\, P(X_{n-1} = k \mid X_0 = i) = \sum_{k \in I} p_{kj}\, p_{ik}^{(n-1)}$$

where the last equality uses the assumption of time homogeneity.

6 6

It is convenient to write result of Rule in terms of matrices

Let $P = (p_{ij})$

to be the matrix having the one step transition probabilities p_{ij} as entries

If we let $P^{(n)}$ denote the matrix of the n-step transition probabilities $p_{ij}^{(n)}$.

Rule 15.1 shows that $P^{(n)} = P^{(n-1)} \times P$ for all $n \geq 2$

By iterating this formula and using the fact that $P^{(1)} = P$ we obtain:

$P^{(n)} = P \times P \ldots \times P = P^n$

This gives us the following important result:

Rule 15.2 The n-step transition probabilities $p_{ij}^{(n)}$ can be calculated as the entries in the matrix product P^n which is obtained by multiplying the matrix P by itself n times.

Example: On island the weather each day is classified as sunny, cloudy or rainy. Next day's weather depends on today's weather and today's weather only. If present day is sunny, next day will be sunny, cloudy or rainy with probabilities 0.7, 0.1 and 0.2. Transition probabilities are 0.5, 0.25, 0.25 when cloudy and 0.4, 0.3 and 0.3 when rain.

- What is ptobability it will be sunny three days from now if it is rainy today?
- What are the long run proportion of time the weather is sunny, cloudy, rainy over a long period?

8.

What is the probability distribution of the weather after many days? Intuitively, we expect that this probability distribution does not depend on the present ~~state~~ state of weather

Confirmed by following calculations:

$$P^5 = \begin{pmatrix} 0.5963113 & 0.1719806 & 0.2317081 \\ 0.5957781 & 0.1723641 & 0.2318578 \\ 0.5954788 & 0.1725744 & 0.2319419 \end{pmatrix}$$

$$P^{12} = \begin{pmatrix} 0.5960265 & 0.1721854 & 0.2317881 \\ 0.5960265 & 0.1721854 & 0.2317881 \\ 0.5960265 & 0.1721854 & 0.2317881 \end{pmatrix} = P^{13} = P^{14} = \dots$$

That is, after 12 matrix multiplications the entries agree row-row to 7 decimal places. You see that the weather after many days will be sunny, cloudy, rainy with probabilities 0.5960, 0.1722, 0.2318 respectively

It will be clear that the limiting probabilities will give the probabilities of time that weather will be sunny, cloudy, rainy over a long period

We have answered question about long run behaviour of weather by computing sufficiently high powers of P^n

4. Markov

Rule 15.3 For any two states $i, j \in I$

E(number of visits to state j over time points $t = 1 \ldots n \mid X_0 = i$)

$= \sum_{t=1}^{n} p_{ij}^{(t)}$ for $n = 1, 2 \ldots$

The proof of result is instructive: Fix $i, j \in I$ for any $t \geq 1$ let.

$$I_t = \begin{cases} 1 & \text{if } X_t = j \\ 0 & \text{otherwise (a)} \end{cases}$$

The number of visits to state j over the time points $t = 1 \ldots n$ is then given by random variable $\sum_{t=1}^{n} I_t$ using observation that

$$E(I_t \mid X_0 = i) = 1 \times P(I_t = 1 \mid X_0 = i) + 0 \times P(I_t = 0 \mid X_0 = i)$$
$$= P(X_t = j \mid X_0 = i) = p_{ij}^{(t)}$$

We obtain $E(\sum_{t=1}^{n} I_t \mid X_0 = i) = \sum_{t=1}^{n} E(I_t \mid X_0 = i) = \sum_{t=1}^{n} p_{ij}^{(t)}$ proving the desired result.

As example: What is the expected value of the number of sunny days in the coming seven days when it is cloudy today? The answer is the expected value is equal to $\sum_{t=1}^{7} p_{21}^{(t)}$ days. The value of this sum is calculated as 4.049

$0.5 + 0.575$ 7 — 4.049

10 Absorbing Markov Chains

Markov chains can also be used to analyse systems in which some states are "absorbing". Once the system reaches an absorbing state, it remains in that state permanently.

Let $\{X_n\}$ be a Markov chain with one step probabilities p_{ij}. State i is said to be absorbing state if $p_{ii} = 1$

The Markov chain $\{X_n\}$ is said to be an absorbing Markov chain if it has one or more absorbing states and the set of absorbing states is accessible from the other states

How long will it take before the system hits an absorbing state?

If there are multiple absorbing states, what is the probability that the system will end up in each absorbing state?

Example: A fair die is rolled until each of the six possible outcomes 1, 2 .. 6 has appeared. How to calculate the probability mass function of the number of rolls needed?

Solution: Lets say the system is in state i if i different outcomes have appeared so far. Define random variable X_n as the state of the system after the n^{th} roll. State 6 is taken as an absorbing state. The process $\{X_n\}$ is an absorbing Markov chain with state space $I = (1, 2 \dots 6)$

The matrix $P = p_{ij}$ of one-step transition probabilities is given by

$p_{01} = 1$ $\quad p_{ii} = \frac{i}{6}$ $\quad p_{i,i+1} = 1 - \frac{i}{6}$ for $i = 1 \dots 5$ $\quad p_{66} = 1$

and $p_{ij} = 0$ otherwise

The starting state of the process is state 0. Let the random variable R denote the number of rolls of the die needed to obtain all of the six possible outcomes. The random variable R takes on a value larger than r only if the Markov chain

4

has not visited the absorbing state 6 in the first r transitions
Hence:

$$P(R>r) = P(X_k \neq 6 \text{ for } k=1,\ldots,r \mid X_0=0)$$

However, since state 6 is absorbing, it automatically holds that $X_k \neq 6$ for any $k<r$ if $X_r \neq 6$ hence.

$$P(X_k \neq 6 \text{ for } k=1,\ldots,r \mid X_0=0) = P(X_r \neq 6 \mid X_0=0)$$

Noting that: $P(X_r \neq 6 \mid X_0=0) = 1-P(X_r=6 \mid X_0=0)$ we obtain

$$P(R>r) = 1-p_{06}^{(r)} \quad \text{for } r=1,2\ldots$$

(6

Conditional Probability And Bayes

Eg probability of getting two aces $= \frac{4}{52} \times \frac{3}{51}$

Rule can be seen as a way of understanding how the probability of an event is affected by a new piece of information

Conditional Probability:

- Sample space and a probability measure P are defined
- Let A be an event of the experiment
- The probability P(A) reflects our knowledge of the occurrence of event A before the experiment takes place
- Therefore probability P(A) sometimes referred to as the a priori probability of A or the unconditional probability of A.

- Suppose we are now told that an event B has occurred in the experiment, but we still don't know the precise outcome in the set B
- In light of this info, the set B replaces the sample space as the set of possible outcomes and consequently probability of the occurrence of event A changes.

A conditional probability now reflects our knowledge of the occurrence of the event A given that event B has occurred. Notation $P(A|B)$.

Definition 8.1: For any two events A and B with $P(B) > 0$, the conditional probability $P(A|B)$ is defined as

$$P(A|B) = \frac{P(AB)}{P(B)}$$

Here AB stands for the occurrence of both events A and B

Define relative frequency $f_n(E)$ of the occurrence of event E as $\frac{n(E)}{n}$ where $n(E)$ represents the number of times that E occurs in n repetitions of experiment

Assume, now, that in n^2 independent reps of experiment, event B occurs r times simultaneously with event A and s times without event A.

We can say that $f_n(AB) = \frac{r}{n}$ and $f_n(B) = \frac{r+s}{n}$.
If we divide $f_n(AB)$ by $f_n(B)$ we find

$$\frac{f_n(AB)}{f_n(B)} = \frac{r}{r+s}$$

Example: Someone rolled fair die twice. We know that one of the rolls turned up a face value of 6. What is probability that the other roll turned up a six aswell?

Solution: Take as sample space the set $\{(i,j) \mid i, j = 1, \dots 6\}$ where i and j denote the outcome of the first and second roll. Probability of $1/36$ assigned to each element. The event of two sixes is given by $A = \{(6,6)\}$ and the event of at least one six is given by $B = \{(1,6) \dots (5,6) (6,6), (6,5) \dots (6,1)\}$

Applying definition of conditional probability gives:

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{1/36}{11/36} = \frac{1}{11}$$

Can also be written in form of

$$P(AB) = P(A|B)P(B)$$

Independent events

In case of $P(A|B) = P(A)$, occurrence of event A is not contingent on the occurrence or nonoccurrence of event B

Events are independent if

$$P(AB) = P(A)P(B)$$

Be aware that independent events and disjoint events are completely different

If events A and B are disjoint, you calculate the probability of the union $A \cup B$ by adding the probabilities of A and B

For independent events A and B you calculate the probability of the intersection AB by multiplying probabilities of A and B.

Since $P(AB)=0$ for disjoint events A and B, independent events are typically not disjoint

The law of conditional probabilities

Rule 8: Let A be an event that can occur when one of mutually exclusive events $B_1, \ldots B_n$ occurs then:

$P(A) = P(A|B_1)\,P(B_1) + P(A|B_2)P(B_2) + \ldots P(A|B_n)P(B_n)$

Uniform Distribution

We encounter a continuous random variable that describes an experiment where the outcome is completely arbitrary except that we know it lies between certain bounds

Example, measuring recording time each hour (minutes) that particles are emmited. Outcomes lie on interval $[0,60]$ minutes. If measurements are concentrated something is wrong

Not concentrating in any way means that subintervals of the same length should have the same probability

It is clear that the probability density function associated with this experiment should be constant on $[0,60]$

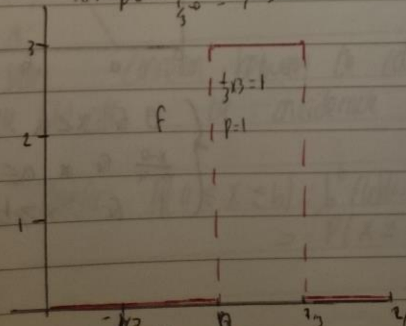
Definition A continuous random variable has a uniform distribution on the interval $[\alpha,\beta]$ if its probability density function f is given by $f(x)=0$ if x is not in $[\alpha\ \beta]$ and

$$f(x) = \frac{1}{\beta-\alpha} \quad \text{for } \alpha \leq x \in \beta$$

We denote this distribution by $U(\alpha,\beta)$

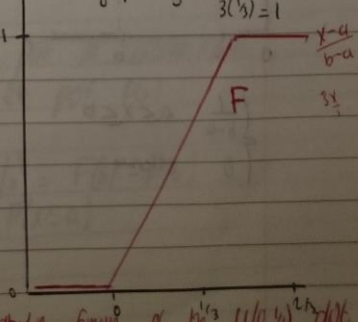
The probability density function and distribution function of a $U(0,\frac{1}{3})$ distribut:

$f(x) = \frac{1}{\beta-\alpha} = \frac{1}{\frac{1}{3}-0} = \frac{3}{1} = 3$



$F(x) = \int_0^{1/3} \frac{1}{\beta-\alpha} = \int_0^{1/3} 3 = 3x \Big|_0^{1/3}$

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



Probability density function and distribution function of the $U(0,\frac{1}{3})$ dist

A continuous random variable X is said to have a uniform density over interval (a, b) if its probability density function is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 \end{cases}$$

b > a.

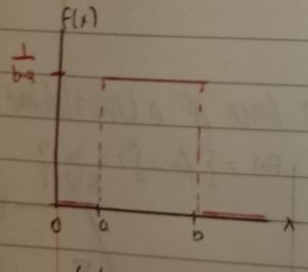
Uniform distribution provides a probability model for selecting a point at random from the interval (a, b).
Since f(x) = 0 outside interval (a, b), the random variable X must assume a value in (a, b).

Since f(x) is constant over interval (a, b) the random variable X is just as likely to be near any value in (a, b) as any other value.

An explicit expression can be given for the c.d.f
$F(x) = \int_{-\infty}^{x} f(y)\,dy$. This function also satisfies $F(x) = 0$ for $x < a$, $F(x) = 1$ for $x \geq b$

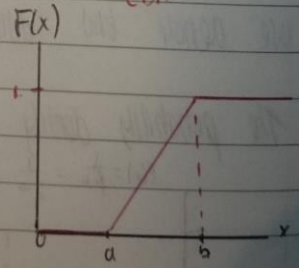
$$F(x) = \frac{b-x}{b-a} \quad \text{for } a \leq x \leq b.$$

Pdf.



$$\begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

CDF



$$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \leq x < b \\ 1 & \text{for } x \geq b \end{cases}$$

Cumulative distribution function CDF

The cdf function or just distribution function describes the probability that a real valued random variable X with a given probability distribution will be found at a value less than or equal to x.

Intuitively, it is "the area so far" function of the probability distribution

The cdf is defined by

$$F(x) = P(X \le x)$$

$F(x)$ gives the "accumulated" probability "up to x".

We can see how pdf and cdf are related

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt$$

(since "x" is used as a variable in upper limit of integration, we use some other variable "t" in the integrand)

Notice that $F(x) \geq 0$ because it's a probability

$$\lim_{x \to \infty} F(x) = \lim_{x \to \infty} \int_{-\infty}^{x} f(t)\,dt = 1$$

$$\lim_{x \to -\infty} F(x) = \lim_{x \to -\infty} \int_{-\infty}^{x} f(t)\,dt = 0$$

$$F'(x) = f(x)$$

$\Rightarrow$ states connection between the cdf and pdf in another way.
the $F(x)$ cdf is the antiderivative of the pdf $f(x)$

and therefore $P(a \le X \le b) = \int_a^b f(x)dx = F(x)\Big|_a^b = F(b) - F(a)$

$$= P(X \le b) - P(X \le a)$$

Expected Value

Expected value of y $E[Y]$

The expected value of a variable y on the pdf can be calculated by $\frac{1}{b-a}$ where b and a are the limits on the region

expected value of a uniform random variable X is $E[X] = \frac{a+b}{2}$

Proof : $E[Y] = \int_{-\infty}^{\infty} y\, f_Y(y)\, dx$

$$= \int_a^b y \left(\frac{1}{b-a}\right) dx$$

$$= \frac{1}{b-a} \int_a^b y\, dy$$

$$= \frac{1}{b-a} \left[\frac{1}{2} y^2\right]_a^b$$

$$= \frac{1}{b-a} \cdot \frac{1}{2} \left[b^2 - a^2\right]$$

$$= \frac{1}{b-a} \cdot \frac{1}{2} (b-a)(b+a)$$

$$= \frac{b+a}{2}$$

Expected value $E[Y^2]$

Suppose y has the $U(a,b)$ distribution

then the n^{th} moment of y given by:

$$E[Y^n] = \frac{1}{b-a} \int_a^b y^n dy$$

$$= \frac{1}{b-a} \frac{1}{n+1} y^{n+1} \Big|_a^b$$

$$= \frac{1}{n+1} \left(\frac{b^{n+1} - a^{n+1}}{b-a}\right)$$

$$E[Y^2] = \frac{1}{b-a} \int_a^b y^2 dy$$

$$\frac{1}{b-a} \left(\frac{y^3}{3}\right)\Big|_a^b$$

$$\frac{1}{b-a} \left[\frac{b^3}{3} - \frac{a^3}{3}\right)$$

$$\frac{1}{3} \left(\frac{b^3 - a^3}{b-a}\right) = E[Y^n]$$

Expected value Var

Var [y] is given by $E[y^2] - E[y]^2$

$$= \frac{1}{3}\left(\frac{b^3-a^3}{b-a}\right) - \left[\left(\frac{b+a}{2}\right)\right]^2$$

$$\frac{1}{3}\left(\frac{b^3-a^3}{b-a}\right) - \frac{1}{2}\left(\frac{b+a}{2}\right)$$

$$\frac{1}{3}\frac{(b-a)(b^2+ab+a^2)}{\cancel{b-a}}$$

$$\frac{b^2+ab+a^2}{3} - \left[\frac{b^2+2ab+a^2}{4}\right]$$

$$\frac{4b^2+4ab+4a^2-3b^2-6ab-3a^2}{12}$$

$$\frac{b^2-2ab+a^2}{12} = \frac{(a-b)^2}{12}$$

Applied Probability Probability Density Function PDF

Q3

One way to look at continuous random variables is that they arise by a (never ending) process of refinement from discrete random variables.

Suppose for example some equipment takes on the value of 6.283 with probability p. If we refine the number (get 4th decimal or 5th etc) then the probability p is spread over the outcomes 6.2830, 6.2831... 6.2839.

Usually this means that each of these new values is taken on with a probability that is much smaller than p - the sum of the 10 probabilities $= p$

Continuing the refinement process to more and more decimals, the ~~probability that the possible values lie in some fixed interval~~ probabilities of the possible outcomes become smaller and smaller, approaching zero.

However the probability that the possible values will lie in some fixed interval $[a, b]$ will settle down

This is closely related to the way sums converge to an integral

Definition:

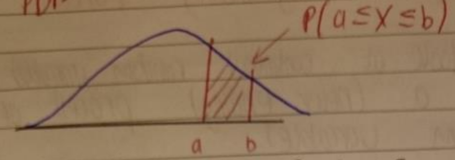
A random variable X is continuous if for some function $f: \mathbb{R} \to \mathbb{R}$ and for any numbers a and b with $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

The function f has to satisfy $f(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f(x)\,dx = 1$

We call f the probability density function (or probability density of X)

PDF



Area under a probability density function F on the interval [a,b]

Note that the probability that X lies in an interval [a,b] is equal to the area under the probability density function f of X over the interval [a,b]

If the interval gets smaller and smaller, the probability will go to zero for any positive ε

$$P(a-\epsilon \le X \le a+\epsilon) = \int_{a-\epsilon}^{a+\epsilon} f(x)dx$$

and sending ε to 0, it follows that for any a

$$P(X=a)=0$$

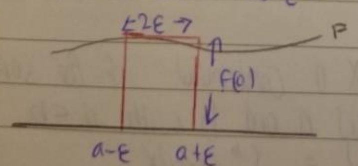
eg $P(X=5)=0$ because $p(5 \le x \le 5) = \int_5^5 f(x)dx = 5(x^2) - 5(x^2) = 0$

This implies that for continuous random variables you may be careless about the precise form of the intervals:

$$P(a \le X \le b) = P(a < X \le b) = P(a < X < b) = P(a \le X < b)$$

What does f(a) represent? See figure: note that: for small positive ε

$$P(a-\epsilon \le X \le a+\epsilon) = \int_{a-\epsilon}^{a+\epsilon} f(x)dx = 2\epsilon f(a)$$



Hence f(a) can be interpreted as a (relative) measure of how likely it is that X will be near a

Do not think of f(a) as a probability: it can be arbitrarily large

PDF ③

Example/exercise: Let f be defined by $f(x)=0$ if $x \le 0$ or $x \geq 1$, and $f(x) = 1/(2\sqrt{x})$ for $0 < x < 1$. Let X be a random variable with f as its probability density function. Compute probability X lies between 10^{-4} and 10^{-2}.

We know from integral calculus that for $0 \le a \le b \le 1$

$$\int_a^b f(x)dx = \int_a^b \frac{1}{2\sqrt{x}} dx = \sqrt{b} - \sqrt{a}$$

Hence $\int_{-\infty}^{\infty} f(x)dx = \int_0^1 (1/2\sqrt{x})dx = 1$, so f is a probability density function - nonnegativity being obvious) and

$$P(10^{-4} \le X \le 10^{-2}) = \int_{10^{-4}}^{10^{-2}} \frac{1}{2\sqrt{x}} dx$$
$$= \sqrt{10^{-2}} - \sqrt{10^{-4}} = 10^{-1} - 10^{-2} = 0.09$$

You should realise that discrete random variables do not have a probability density function f and continuous random variables do not have a probability mass function p, but both have a distribution function $F(a) = P(X \le a)$

Using the fact that for $a \le b$ the event $\{X \le b\}$ is a disjoint union of events $\{X \le a\}$ and $\{a < X \le b\}$, we can express the probability that X lies in an interval $(a,b]$ directly in terms of F for both cases:

$$P(a < X \le b) = P(X \le b) - P(X \le a) = F(b) - F(a)$$

There is a simple relation between the distribution function F and the probability density function f of a continuous random variables:

$$F(b) = \int_{-\infty}^{b} f(x)dx \quad \text{and} \quad f(x) = \frac{d}{dx}F(x)$$

holds for all x where f is continuous.

4

Both the probability density function and the distribution of a continuous random variable X contain all the probabilistic information about X; the probability distribution of X is described by either of them.

Example:

Suppose we want to make a probability model for an experiment that is described as "an object hits a disc of radius r in a completely arbitrary way". (like darts)

We are interested in the distance X between hitting the point and the centre of the disc. Since distances cannot be negative we have $F(b) = 1$ when $b > r$.

That the dart hits the disk in a completely arbitrary way we interpret as that the probability of hitting any region is proportional to area.

Because disc has area πr^2 and disc with radius b has area πb^2 we should put: $F(b) = P(X \leq b) = \frac{\pi b^2}{\pi r^2} = \frac{b^2}{r^2}$ for $0 \leq b \leq r$.

Then the probability density function f of X equal to 0 outside interval $[0, r]$ and

$$f(x) = \frac{d}{dx} F(x) = \frac{1}{r^2} \frac{d}{dx} x^2 = \frac{2x}{r^2} \text{ for } 0 \leq x \leq r.$$

Example: Let X, the thickness of a certain metal sheet be a uniform distribution on [A, B]. The pdf for X is

$$f(x) = \begin{cases} \frac{1}{B-A} & A \le x \le B \\ 0 & \text{otherwise} \end{cases}$$

The cdf is calculated:

For $x < A$, $F(x) = 0$, for $A \le x < B$ we have

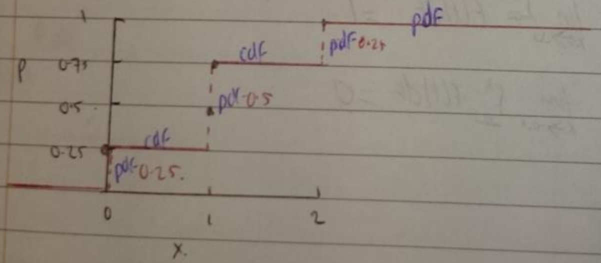
$$F(x) = \int_{-\infty}^{x} f(y)\,dy = \int_{A}^{x} \frac{1}{B-A}\,dy = \frac{1}{B-A} \cdot y \Big|_{y=A}^{y=x} = \frac{x-A}{B-A}$$

For $x \geq B$, $F(x) = 1$

Therefore entire cdf is

$$F(x) = \begin{cases} 0 & x < A \\ \frac{x-A}{B-A} & A \le x < B \\ 1 & x \geq B \end{cases}$$

Example head or tails two rolls



pdf

$p(0) = (t,t) = \frac{1}{4}$

$p(1) = (h,t) + (t,h) = 0.5$

$p(2) = (h,h) = \frac{1}{4}$

cdf

$P(0) = (t,t) = \frac{1}{4}$

$P(1) = (h,t) + (t,h) + (t,t)$

$P(2) = (h,t) + (t,h) + (t,t) + (h,h) = 1$