Examples of EECS 401 Prerequisite Material (Not Necessarily Completely Inclusive)

Differentiation from First Principles

$$\frac{d}{dx}f(x) = \lim_{\delta \searrow 0^+} \frac{f(x+\delta) - f(x)}{\delta} = \lim_{\delta \searrow 0^+} \frac{f(x) - f(x-\delta)}{\delta}, \quad \text{if } f(\cdot) \text{ is continuous at } x \text{ and the limits are equal,}$$

where $\delta \searrow 0^+$ means δ approaches 0 from the right (positive δ)

Riemann Integration

If f is a continuous function on $[a, b]$, then

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \frac{a-b}{n} \sum_{i=1}^{n} f\left(a + \frac{i}{n}(b-a)\right).$$

Similarly, if δ is small, then

$$\int_a^{a+\delta} f(x)\,dx \approx \delta \cdot f(a).$$

Multivariate Integration (Especially Double Integration Limits)

If $A = \{(x, y) : \ x^2 + y^2 \le 2, \ x \ge 0, \ y \ge 1\}$, then

$$\iint_A xy\,dxdy = \int_1^{\sqrt{2}} \int_0^{\sqrt{2-y^2}} xy\,dxdy = \int_1^{\sqrt{2}} \frac{1}{2}(2-y^2)y\,dy = \frac{1}{2}(y^2 - \frac{1}{4}y^4)\Big|_1^{\sqrt{2}} = \frac{1}{2}(2-1) - \frac{1}{2}(1-\frac{1}{4}) = 1/8$$

Simple Matrix Inversion

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \quad \text{if} \quad ad \neq bc$$

Leibniz's Rule

$$\text{If } G(x) = \int_{a(x)}^{b(x)} h(x,y)\,dy \quad \text{then} \quad \frac{d}{dx}G(x) = h(x,b(x))\frac{d}{dx}b(x) - h(x,a(x))\frac{d}{dx}a(x) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x}h(x,y)\,dy$$

Geometric Series

$$\sum_{i=0}^{n} a^n = \frac{1-a^{n+1}}{1-a}, \quad \text{if } a \neq 1; \qquad \sum_{i=0}^{\infty} a^n = \frac{1}{1-a}, \quad \text{if } |a| < 1$$

Proof: $S = a^0 + a^1 + \cdots + a^n$ so $aS = a^1 + a^2 + \cdots + a^{n+1}$ so $(1-a)S = S - aS = a^0 - a^{n+1}$.

Fourier Transform
If $h(t) = e^{-at}u(t)$, for $a > 0$, then

$$H(\omega) = \int_{-\infty}^{\infty} h(t)e^{-j\omega t}\,dt = \int_0^{\infty} e^{-at}e^{-j\omega t}\,dt = \int_0^{\infty} e^{-(a+j\omega)t}\,dt = \frac{-1}{a+j\omega}e^{-(a+j\omega)t}\Big|_0^{\infty} = \frac{1}{a+j\omega}$$

Discrete-Time Fourier Transform (DTFT)
If $h_k = a^{|k|}$, for $|a| < 1$, then

$$H(\omega) = \sum_{k=-\infty}^{\infty} h_k e^{-j\omega k} = \sum_{k=-\infty}^{\infty} a^{|k|}e^{-j\omega k} = \sum_{k=-\infty}^{-1} a^{-k}e^{-j\omega k} + \sum_{k=0}^{\infty} a^k e^{-j\omega k} = \sum_{k=1}^{\infty}(ae^{j\omega})^k + \sum_{k=0}^{\infty}(ae^{-j\omega})^k$$

$$= \frac{ae^{j\omega}}{1-ae^{j\omega}} + \frac{1}{1-ae^{-j\omega}} = \frac{1-a^2}{1-2a\cos\omega + a^2}$$

Convolution: Continuous Time
If $h(t) = e^{-at}u(t)$, for $a > 0$, and $x(t) = e^{-bt}u(t)$, for $b > 0$, then

$$y(t) = (h \star x)(t) = \int_{-\infty}^{\infty} h(t-\tau)x(\tau)\,d\tau = \int_{-\infty}^{\infty} e^{-a(t-\tau)}e^{-b\tau}u(t-\tau)u(\tau)\,d\tau = \left(\int_0^t e^{-a(t-\tau)}e^{-b\tau}\,d\tau\right)u(t)$$

$$= e^{-at}\left(\int_0^t e^{-(b-a)\tau}\,d\tau\right)u(t) = e^{-at}\left(\frac{-1}{b-a}e^{-(b-a)\tau}\Big|_0^t\right)u(t) = \frac{e^{-at}-e^{-bt}}{b-a}u(t)$$

Convolution: Discrete Time
If $h_k = 2\delta_k - \delta_{k-1}$ and $g_k = \delta_k + \delta_{k+1}$, then $h_k \star g_k = \sum_{j=-\infty}^{\infty} h_j g_{k-j} = 2\delta_{k+1} + \delta_k - \delta_{k-1}$.

Ch. 2 Basic Concepts of Probability

Conceptual Framework

Every discipline has one.
Circuits: impedanceless resistors, resistanceless capacitors, ...
Linear Systems: impulse or delta functions, infinite duration sinusoids, ...
Abstractions + Math $\rightarrow$ Predictions (often agree with physical experiments despite simplifications) if not, the problem is abstractions (assumptions), not math (if done right)
Probability theory is math

Terminology

Random Experiment or Chance Experiment

A random phenomena (or experiment) having a known set of possible **outcomes**, for which the particular outcome on a given **trial** is unpredictable, and that can be (conceptually) repeated arbitrarily often under (essentially) identical circumstances.

Examples: roll a dice, disk controller receiving instruction to access a particular sector

A **trial** is a single instance of an experiment.
Repeated trials are multiple instances under identical conditions.

Result of an experiment is called the **outcome** or **sample point**, denoted ζ (zeta) by text. (I will use s).

The set of all possible outcomes is the **universal set** or (only universal in the context of this experiment) or **sample space**, and is denoted S.

Example: Coin Toss
S = {H,T}
S = {H,T,edge}
S = {H,T,edge,vaporized by meteor}
Probability theory is self-consistent for any of the above choices; whether the theory predicts reality depends on whether an appropriate choice is made.

Example: Toss 2 Die
S = {(1,1), (1,2), (2,1), . . . , (6,6)}

Example: Number of hairs on 34th birthday
S = {0,1,2,. . . }

Example: AC Voltage at random time instant
S = $[-120\sqrt{2}, 120\sqrt{2}]$

Often we are more interested in aggregate phenomena, such as winning a game of poker, than about the specific outcome (exactly which hand).

Event
An **event** is a collection (set) of outcomes.
A **simple event** or **elementary event** is a set consisting of a single outcome, e.g. A = $\{\text{H}\}$ in coin toss experiment.
A **compound event** is a set consisting of more than one outcome.

Example: Single Die
Sample space: S = $\{1,2,\ldots,6\}$
Event "even face" is $A = \{2, 4, 6\}$. Note $A \subseteq S$

Example: Toss 3 Coins
Sample space: S = $\{\text{HHH,HHT,HTH,}\ldots\text{,TTT}\}$ Q?: size = 2^3
Event "two heads" is $A = \{\text{HHT,HTH,THH}\}$

Example: AC Voltage and Heart Attack (e.g. if voltage exceeds 100V)
$A = [-120\sqrt{2}, -100) \cup (100, 120\sqrt{2}]$

Unfortunately, for uncountable sample spaces, not *all* subsets of S can be called events for a rigorous and self-consistent probability theory. Fortunately, all subsets of *practical* interest can be called events, so we won't worry about this in EECS 401.)

Clearly, to describe events we need **set theory**.

Probabilities are numbers we assign to **events** that indicate how "likely" it is that the events will occur when performing an experiment.

To be useful in practice, probabilities should agree with the "relative frequency" concept.

Experiment: roll fair die. Event: A = $\{$Roll 5$\}$

Then for a "large" number of trials N we would hope that however we define $P(A)$ it would satisfy:

$$P(A) \stackrel{?}{=} \frac{N_A}{N} = \frac{\text{\# trials } A \text{ occurred}}{\text{\# trials (total)}}.$$

So if we roll it $N = 120$ times, we expect about $N_A = 20$ trials where we roll a 5. But not exactly, because it is a random phenomena!

Properties of Relative Frequency

Since $0 \leq N_A \leq N$, we have

$$0 \leq \frac{N_A}{N} \leq 1$$

so apparently we want $0 \leq P(A) \leq 1$.

The next property is called **additivity**.

Let $A = \{$Roll 5 or Roll 6$\} = A_5 \cup A_6$ where $A_5 = \{$Roll 5$\}$ and $A_6 = \{$Roll 6$\}$. Then

$$\frac{N_A}{N} = \frac{\text{\# of 5's or 6's}}{N} = \frac{\text{\# of 5's}}{N} + \frac{\text{\# of 6's}}{N}$$

so apparently we want

$$P(A) = P(A_5 \cup A_6) + P(A_5) + P(A_6).$$

But not always! (And this is a common mistake)

Consider: A = $\{$Roll 2$\}$, B = $\{$Roll Even$\}$ = $\{$2,4,6$\}$
Then $P(A) = 1/6$, $P(B) = 3/6$, $P(A \cup B) = P(B) = 3/6 \neq P(A) + P(B) = 4/6$.

Q? what is the problem?

What we really want is

$$\boxed{P(A \cup B) = P(A) + P(B) \quad \text{if} \quad A \cap B = \phi.}$$

Unfortunately, "N large" is not mathematically precise, so the relative frequency principle alone does not provide a rigorous self-consistent probability theory. It is also impractical for complex problems (fly 1000 space shuttles?)

Next we present an axiomatic approach to probability that is mathematically rigorous, but still captures the basic idea behind relative frequency.

2.2

Axioms of Probability

Given a sample space S (a collection of outcomes for a random experiment), a **probability law** or **probability measure** P is a function that assigns to each event A a number $P[A]$ called the "probability of A" that must satisfy the following axioms.

Axiom 1: $0 \leq P[A]$ for all A (nonnegativity)

Axiom 2: $P[S] = 1$ (some outcome must occur)

Axiom 3: If $A \cap B = \phi$, then $P[A \cup B] = P[A] + P[B]$ (additivity)

Axiom 3': If $A_i \cap A_j = \phi$ for $i \neq j$, then $P[\bigcup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} P[A_i]]$ (countable additivity)

By induction from Axiom 3:

$$\boxed{P\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} P[A_i] \quad \text{if} \quad A_i \cap A_j = \phi \quad \text{for} \quad i \neq j.}$$

Note: axiom $\equiv$ assumption (but history has shown that the above yield predictions that agree with reality).

(picture)

Course Goals Include:

$\star$ Parsing problem statements (learned by examples from lecture, text, HW)

- determine sample space
- extract "given" probabilities
  (using symmetries, physical reasoning, experiments, assumptions of independence)
- translate question into a probability that is to be found

$\star\star$ Apply probability tools to determine desired information

Equally likely probability assignment

The simplest example of an initial probability assignment is the case where we assume all outcomes are **equally likely**.

Note that this is often an incorrect assumption, and a common error is to apply this probability assignment to problems where is is inapplicable!

Assumptions for equally likely probability assignment:
- Suppose sample space is **discrete** and **finite** $S = \{a_1, a_2, \ldots, a_n\}$
- Suppose each outcome is **equally likely** $P[\{a_i\}] = 1/n$

Now suppose we want to compute the probability of a compound event E, such as $E = \{a_1, a_4, a_6\} = \{a_1\} \cup \{a_4\} \cup \{a_6\}$, i.e. a union of elementary events.

Note: elementary events are always disjoint, since outcomes are inherently distinct. So $\{a_i\} \cap \{a_j\} = \phi$ for $i \neq j$.

Thus by additivity:

$$\begin{aligned} P[E] &= P[\{a_1\} \cup \{a_4\} \cup \{a_6\}] \\ &= P[\{a_1\}] + P[\{a_4\}] + P[\{a_6\}] \\ &= 1/n + 1/n + 1/n = 3/n \end{aligned}$$

More generally, *if* all outcomes are **equally likely**, then

$$\boxed{P[E] = \frac{\text{\# of outcomes in } E}{\text{total \# of outcomes in S}} \quad \text{for equally likely outcomes.}}$$

Example Roll 2 *fair* die
What is probability that the sum of dots is 8?
S = {(1,1),(1,2),(2,1),. . . ,(6,6)} Q? size = 6^2
$E = \{(2,6), (3,5), (4,4), (5,3), (6,2)\}$ (5 outcomes where sum is 8)
For fair die, all 6^2 outcomes are equally likely, so $P[E] = 5/36$

Common Pitfall setting up sample space where outcomes are *not* equally likely

Example: toss 2 fair coins. What is P[2 heads] ?

Right Way	Wrong Way
S = {HH,HT,TH,TT}	S = {0 heads, 1 head, 2 heads}
P[E] = P[HH] = 1/4	P[E] = P[2 heads] $\stackrel{?}{=}$ 1/3 no! because outcomes are not equally likely (there are 2 ways to get 1 head)

2.3

Combinatorics

Context:
- We have a set of n (distinct) elements (such as balls in an urn numbered $1, 2, \ldots, n$).
- We intend to pick k items from the set "at random."

We are interested in the probabilities of various events.

Questions:
- Do we sample with replacement, or not? (determines total number of outcomes in S)
- For the event of interest, does the *order* in which the items were drawn matter?

If we enumerate the outcomes exhaustively, then all possible outcomes are equally likely since we draw the elements "at random." So to compute probabilities using the "equally likely" formula, we must find the total number of possible outcomes in S (the denominator), as well as the number of outcomes in events of interest (the numerator).

Denominator: How many total outcomes are there? (Must consider ordering so that all outcomes are equally likely.)
- Sampling with replacement (can have $k \leq n$ or $k \geq n$):
  There are n possibilities for each draw, and k draws, so $n \cdot n \cdot \ldots \cdot n = n^k$
- Sampling without replacement (must have $k \leq n$):
  $n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = (n)_k = n!/(n-k)!$

Ordered Samples

Making an ordered sample by drawing k items with replacement

Example: Urn contains a red, green, blue, and yellow ball.
What is the probability of getting the particular sequence of draws: $E = \{\text{BBBGGY}\}$?
Answer: $P(E) = 1/n^k = 1/4^6$.
What is the probability of drawing exactly 5 blue balls in a row in a sequence of 6 draws?
$E = \{\text{BBBBBR, BBBBBG, BBBBBY, RBBBBB, GBBBBB, YBBBBB}\}$, so $P(E) = 6/4^6$.
(Enginering application: disk media that require not too many bits of same sign sequentially.)

In general $\boxed{P[E] = \frac{\text{(\# of outcomes in } E)}{n^k}}$ for an ordered sample drawn with replacement.

Making an ordered sample by drawing k items without replacement

Example: deck of cards ($n = 52$)
What is the probability of being dealt the hand (3♣, 4♣, 5♣, 6♣, 7♣) in that order?
Answer: $1/(n!/(n-k)!) = (52-5)!/52! = 47!/52! = 1/311,875,200 \approx 3.2 \cdot 10^{-9}$
What is the probability of being dealt the hand $(3, 4, 5, 6, 7)$ in that order, but any suit?
Answer: $4^5/(n!/(n-k)!) = 4^5(52-5)!/52! \approx 3.28 \cdot 10^{-6}$

In general $\boxed{P[E] = \frac{\text{(\# of outcomes in } E)}{n!/(n-k)!}}$ for an ordered sample drawn without replacement.

Unordered Samples

Making an unordered sample by drawing k items without replacement
(Think of taking balls from one urn and putting in a jar)
There are $n!/(n-k)!$ ordered samples having k items.
But each of these samples can be rearranged $k!$ different ways.

So the number of distinct unordered samples taken without replacement is $\frac{n!}{(n-k)!k!} = \left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n \\ n-k \end{array} \right)$.

Example: E = {3 heads in 5 flips of a fair coin}, what is P(E)?
How many length-5 sequences of H,T are there with 3 heads?
Think of urn with balls numbered 1 to 5. Pick 3 balls w/o replacement.
Set the corresponding positions to H, rest to T. (Order of balls unimportant!)
So number of length-5 sequences of H,T with 3 heads is $\left(\begin{array}{c} 5 \\ 3 \end{array} \right)$. So $P(E) = \left(\begin{array}{c} 5 \\ 3 \end{array} \right) / 2^5$

Making an unordered drawing of k items with replacement
Make a list of length n initialized with zeros; add a check to ith entry list each time you draw the ith item.
The number of such lists you can make is $\left(\begin{array}{c} n-1+k \\ k \end{array} \right) = \left(\begin{array}{c} n-1+k \\ n-1 \end{array} \right)$.
Caution: these lists are *not* equally likely, so the formula is rarely used for calculating probabilities.
Instead, we usually just resort to counting arguments.

Example: Wheel of fortune with 26 letters (A to Z).
What is probability of getting a vowel on each of 4 spins?
(With "replacement," order unimportant)
Answer: $5^4/26^4 \approx 0.0014$
P(4 distinct consonants in 4 spins) = $(21 \cdot 20 \cdot 19 \cdot 18)/26^4 \approx 0.31$

Formulas are simple. Be careful to pick correct case. (Examples!)

Summary

Urn: 4 Balls (Red, Green, Yellow, and Blue)

	Ordered	Unordered
With Replacement	(R,R) (R,G) (R,Y) (R,B) (G,R) (G,G) (G,Y) (G,B) (Y,R) (Y,G) (Y,Y) (Y,B) (B,R) (B,G) (B,Y) (B,B) $n^k = 4^2 = 16$	(2,0,0,0) (1,1,0,0) (1,0,1,0) (1,0,0,1) (0,2,0,0) (0,1,1,0) (0,1,0,1) (0,0,2,0) (0,0,1,1) (0,0,0,2) $\left(\begin{array}{c} n-1+k \\ k \end{array} \right) = \left(\begin{array}{c} 5 \\ 2 \end{array} \right) = 10$
Without Replacement	(R,G) (R,Y) (R,B) (G,R) (G,Y) (G,B) (Y,R) (Y,G) (Y,B) (B,R) (B,G) (B,Y) $\frac{n!}{(n-k)!} = \frac{4!}{2!} = 12$	{R,G} {R,Y} {R,B} {G,Y} {G,B} {Y,B} $\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} 4 \\ 2 \end{array} \right) = 6$

Enumeration is tedious in more complicated problems, and inapplicable to problems where the outcomes are not equally likely. So we need more general methods...

Properties of Probability Laws

(1) $\boxed{P(\overline{A}) = 1 - P(A)}$
Since $A \cap \overline{A} = \phi$, by additivity: $P(A \cup \overline{A}) = P(A) + P(\overline{A})$, but $A \cup \overline{A} = S$ and $P(S) = 1$, so $1 = P(A) + P(\overline{A})$.
(2) $\boxed{P(A) \leq 1}$
By (1): $P(A) = 1 - P(\overline{A}) \leq 1$ since $P(\overline{A}) \geq 0$.
(3) $\boxed{P(\phi) = 0}$
By (1): $P(\phi) = 1 - P(S) = 1 - 1 = 0$.
(4) $\boxed{\text{If } A_i \cap A_j = \phi \text{ for } i \neq j, \text{ then } P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)}$
Proof by induction.
(5) $\boxed{P(A - B) = P(A) - P(A \cap B)}$
$A = A \cap S = A \cap (B \cup \overline{B}) = (A \cap B) \cup (A \cap \overline{B})$ and $(A \cap B) \cap (A \cap \overline{B}) = \phi$.
So by additivity: $P(A) = P(A \cap B) + P(A \cap \overline{B})$
(6) $\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$ (exchanging $\cup$ and $\cap$)
$A \cup B = (A - B) \cup (A \cap B) \cup (B - A)$, and the three sets are disjoint. (Venn diagram)
So by (4): $P(A \cup B) = P(A - B) + P(A \cap B) + P(B - A)$.
Now apply (5): $P(A \cup B) = (P(A) - P(A \cap B)) + P(A \cap B) + (P(B) - P(B \cap A)) = P(A) + P(B) - P(A \cap B)$
(6') $\boxed{P(A \cup B) \leq P(A) + P(B)}$ (called "union bound")
Follows from (6) since $P(B - A) \geq 0$.
(7) $\boxed{\text{If } A \subseteq B \text{ then } P(A) \leq P(B)}$
By (5): $P(B - A) = P(B) - P(A \cap B) \geq 0$ (by Axiom 1). So $P(B) \geq P(A \cap B)$, but $A \cap B = A$ since $A \subseteq B$.

Tools
- $P(A) = 1 - P(\overline{A})$
- Break into disjoint events using set properties and apply additivity
- For inequalities, work towards the two known inequalities: $0 \leq P(A) \leq 1$.

Example
Roll 2 fair die.
What is probability of rolling *at least* one 4?
Use above properties to avoid exhaustive enumeration:
Let A be the event "roll a 4 on die 1" $P(A) = 1/6$
Let B be the event "roll a 4 on die 2" $P(B) = 1/6$
Let E be the event "roll a 4 on either die"
$E = A \cup B$
Not disjoint!
By (7) $P(E) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1/6 + 1/6 - 1/36 = 11/36$
Since $A \cap B$ is the event "roll a 4 on both die."
More elegant (and practical) than enumerating outcomes: $E = \{(4,1), (4,2), \ldots\} \subset S$

Summary
Defined basic events
Expressed desired event E in terms of basic event
Used set operations and probability properties to get $P(E)$

Joint Probability

If A and B are events, the **joint probability** of A and B is defined to be $\boxed{P(A \cap B).}$ One way to find joint probabilities is to use the formula $\boxed{P(A \cap B) = P(A) + P(B) - P(A \cup B).}$ More often we use conditional probabilities.

2.4

Conditional Probability

Often we want to answer questions such as "what is the probability the shuttle will fail given that the O-rings leak?"

Intuition: Dart Board (relative frequency)
$P(A) = \text{Area}(A)/\text{Area}(S)$
(bad dart thrower = "random" throws)

If A and B are two events and if $P(B) > 0$, then we define the conditional probability of A given B to be

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

To justify calling $P(\cdot|B)$ a probability law, it must satisfy the Axioms.

- $P(A|B) > 0$
  Clear from its definition
- $P(S|B) = 1$
  $P(S|B) = P(S \cap B)/P(B) = P(B)/P(B) = 1$
- If $A_1 \cap A_2 = \phi$, then $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B)$
  $P(A_1 \cup A_2|B) = P((A_1 \cup A_2) \cap B)/P(B) = P((A_1 \cap B) \cup (B \cap A_2))/P(B) = P(A_1 \cap B)/P(B) + P(A_2 \cap B)/P(B) = P(A_1|B) + P(A_2|B)$, since $A_1 \cap A_2 = \phi$ implies that $(A_1 \cap B)$ and $(A_2 \cap B)$ are disjoint.
- Similar proof for countable additivity.

Note also that $P(B|B) = P(B \cap B)/P(B) = 1$.

Example

Suppose we have a light bulb that will fail at some (unpredictable) time after time 0. So $S = [0, \infty)$.
Assume that $P[(t, \infty)] = e^{-t/5}$, i.e. the probability that the bulb fails after any given time t (in years, say) is $e^{-t/5}$.
(Later we learn that the mean lifetime of the bulb is 5 years here.)
Given that the lightbulb is still working at 4 years, what is the probability it will fail sometime after 6 years?
Let B be the event the lightbulb is still working at 4 years. $B = (4, \infty)$.
Let A be the event the lightbulb fails sometime after year 6. $A = (6, \infty)$.
$P[A|B] = P[A \cap B]/P(B) = P[(4, \infty) \cap (6, \infty)]/P[(4, \infty)] = P[(6, \infty)]/P[(4, \infty)] = e^{-6/5}/e^{-4/5} = e^{-2/5} \approx 0.67$
Compare to $P(A) = e^{-6/5} \approx 0.3$

Example

We roll two fair 20-sided die.
Given that the sum is 36, what is the probability that either die rolled a 19?
A is event either die rolled a 19.
B is event that sum of dots is 36.
Want $P(A|B)$
S is 20^2 equally likely outcomes (die1,die2)
$A \cap B = \{(17, 19), (19, 17)\}$
$B = \{(16, 20), (17, 19), (18, 18), (19, 17), (20, 16)\}$
$P[A|B] = P[A \cap B]/P(B) = (2/20^2)/(5/20^2) = 2/5.$

We can rewrite definition of conditional probability to get an equally useful formula:

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

Especially useful for *sequential experiments*.

Example: urn with 5 red balls and 3 green balls.
What is probability of drawing 2 red balls in 2 (random) draws w/o replacement?
Translate:

- D_1 = draw red ball on 1st draw, $P(D_1) = 5/8$
- D_2 = draw red ball on 2nd draw, $P(D_2|D_1) = 4/7$
- $E = D_1 \cap D_2$, $P(E) = P(D_1 \cap D_2) = P(D_2|D_1)P(D_1) = \frac{4}{7}\frac{5}{8} = 5/14$

Easier than counting from $8 \cdot 7$ outcomes in S

Chain Rule (useful for sequential experiments)
More generally: $P(A_1 \cap A_2 \cap \dots A_n) = P(A_n|A_1 \cap A_2 \cap \dots A_{n-1})P(A_{n-1}|A_1 \cap A_2 \cap \dots A_{n-2}) \dots P(A_2|A_1)P(A_1)$.

The following is another tool for computing prob. of complicated events from simple events

---

Total Probability
Suppose events $B_1, \dots, B_n$ partition S
Recall $S = \bigcup_{i=1}^n B_i$ and $B_i \cap B_j = \phi,\ i \neq j$
Also assume that $P(B_i) \neq 0$ for $i = 1, \dots, n$
Law of **total probability**:

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

Proof

$$P[A] = P[A \cap S] = P[A \cap (\cup_{i=1}^n B_i)] = P[\cup_{i=1}^n (A \cap B_i)] = \sum_{i=1}^{n} P[A \cap B_i] = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Because for $i \neq j$, $(A \cap B_i)$ and $(A \cap B_j)$ are disjoint since $B_i \cap B_j = \phi$.

Example (a sequential experiment)
Three chests, each with 2 drawers, containing gold or socks
Chest 1 $\boxed{\text{G}\|\text{G}}$ Chest 2 $\boxed{\text{G}\|\text{S}}$ Chest 3 $\boxed{\text{S}\|\text{S}\|\text{S}}$
Pick a chest at random, then pick a drawer at random from chosen chest
(This is a *sequential experiment* - a sequence of sub-experiments.)
Open drawer and find Gold!
What is the probability the other drawer also contains Gold??
Sample space: 7 outcomes (7 drawers)
Let C_i be the event "picked drawer i," $i = 1, 2, 3$ $P(C_i) = 1/3$
Let A be the event "open drawer has gold"
Let B be the event "other drawer has gold"
Want $P(B|A) = P(B \cap A)/P(A) = P(C_1)/P(A)$
Total probability: $P(A) = \sum_{i=1}^{3} P(A|C_i)P(C_i) = (1 + 1/2 + 0) \cdot (1/3) = 1/2$
So $P(B|A) = P(C_1)/P(A) = (1/3)/(1/2) = 2/3$

Bayes Rule or Bayes Theorem
Let A and B be any two events where $P(A) \neq 0$ and $P(B) \neq 0$. Then **Bayes Rule** is:

$$\boxed{P(B|A) = \frac{P(A|B)P(B)}{P(A)}} \quad \text{“Exchange order of conditioning”}$$

Proof: $P(B|A) = P(B \cap A)/P(A)$ and $P(A|B) = P(A \cap B)/P(B)$
Use commutative law and rearrange

If events $B_1, \ldots, B_n$ partition S, then combining with law of total probability:

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

- $P(B_i)$ called “a priori” probability
- $P(B_i|A)$ called “a posteriori” probability

Applications to medicine, communications, decision theory, gambling...

Classic Example from Digital Communications
We transmit a 0 or 1 using an imperfect binary channel
Given that we receive a 0, what is the probability that a 0 was actually sent?
- A_0 is event “a 0 was sent”
- A_1 is event “a 1 was sent”
- B_0 is event “rcvd. a 0”
- B_1 is event “rcvd. a 1”
- Assume $P(A_0) = P(A_1) = 1/2$ (e.g. compressed data)
- Want $P(A_0|B_0)$

Bayes rule: $P(A_0|B_0) = P(B_0|A_0)P(A_0)/P(B_0) = (1)(1/2)/P(B_0)$
Total prob.: $P(B_0) = P(B_0|A_0)P(A_0) + P(B_0|A_1)P(A_1) = (1)(1/2) + \epsilon(1/2) = (1/2)(1+\epsilon)$
Thus: $P(A_0|B_0) = 1/(1+\epsilon)$

Example
Box 1: 99 Red, 1 Green
Box 2: 102 Red, 98 Green
Pick box at random, then pick ball at random from chosen box.
Given that we chose a Red ball, what is probability we chose Box 1?
Sample space: 300 outcomes (but not very important).
B_i denotes event “picked box i”
R denotes event “picked red ball”
Want $P(B_1|R)$
Total probability: $P(R) = P(R|B_1)P(B_1) + P(R|B_2)P(B_2) = (99/100)(1/2) + (102/200)(1/2) = 3/4$
Bayes rule: $P(B_1|R) = P(R|B_1)P(B_1)/P(R) = (99/100)(1/2)/(3/4) = 33/50$

Alternate Experiment
First mix all balls into one urn
Randomly pick a ball from the urn.
Given that we chose a Red ball, what is probability it originated in Box 1?
Want $P(B_1|R)$
All balls equally likely now, so $P(R) = (99+102)/300 = 201/300$
$P(B_1 \cap R) = P(\text{red ball from Box 1}) = 99/300$
$P(B_1|R) = P(B_1 \cap R)/P(R) = (99/300)/(201/300) = 99/201$

N.B.: Conditional probabilities (and ordinary probabilities too) depend on experiment!

2.5

Independence

Two events A and B are said to be **statistically independent** (or just **independent**) iff

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

If $P(B) \neq 0$, then

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

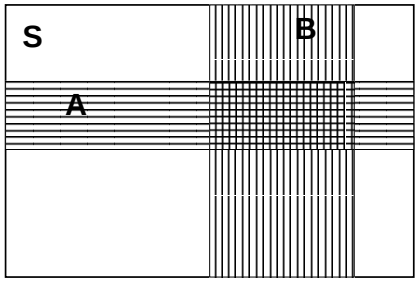
is an equivalent definition, since $P(A|B) = P(A \cap B)/P(B) = P(A)$ iff $P(A \cap B) = P(A) \cdot P(B)$
Likewise, if $P(A) \neq 0$, then

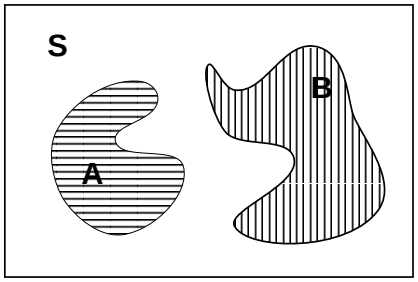
$$P(B|A) = P(B)$$

is an equivalent definition. (cf fortune cookie)

This is one of the most useful methods for finding probabilities of interesting events from known probabilities of simple events. If we can assume (because of physics, reasoning etc.) that two events A and B are *independent*, then $P(A \cap B) = P(A) \cdot P(B)$ gives us the joint probability in terms of the prob. of the individual events.

Independent Events vs Disjoint Events





Independent Events: $P(A \cap B) = P(A) \cdot P(B)$

Disjoint Events: $A \cap B = \phi$, so $P(A \cap B) = P(\phi) = 0$.

Independence of Multiple Events

We say $A_1, A_2, \ldots, A_n$ are **independent** events iff the prob. of any intersection of the A_i's is the product of the individual prob's.

- $P(A_i \cap A_j) = P(A_i)P(A_j), \forall i \neq j$ (pairwise independence)
- $P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k), \forall i \neq j, j \neq k, i \neq k$
- $\vdots$
- $P(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i)$

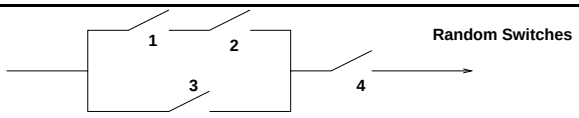
Total of $2^n - n - 1$ conditions to verify!
Pairwise independence alone *does not* imply independence.

Example

Suppose we roll a loaded die 3 times, with $P(\{\text{roll 1}\})$ = 1/5.
What is probability of rolling 3 ones?
6^3 outcomes. But *not* all equally likely now!
Let A_j = {roll 1 on jth roll}
Want P(E) where $E = A_1 \cap A_2 \cap A_3$
Since rolls are physically independent, it is reasonable to assume they are statistically independent,
so $P(E) = P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3) = (1/5)^3$

Combinations of Independent Events
If $A_1, \ldots, A_n$ are independent events, then A_i is independent of any set combination of events $A_1, \ldots, A_{i-1}, A_{i+1}, \ldots, A_n$.
Example: if A, B, and C are independent events, then

$$P[A \cap (B \cup C)] = P[A] \cdot P[B \cup C]$$



Example What is probability of a closed path from input to output?
Old way: A = {CCOC, OOCC, CCCC}, sample space has 4^2 outcomes.
If all switches are independent and equally likely to be open or closed ($p = 1/2$) then all 4^2 outcomes equally likely (explain why using independence!) and $P(A) = 3/16$ (counting method).

For system reliability applications (e.g. Russian space station oxygen system) hopefully $p >> 1/2$, so all 4^2 outcomes are *not* equally likely, so must calculate $P(A)$ another way.

Let C_i ={ith switch is closed}
- assume $P(C_i) = p$
- C_1 = {COOO, COOC, COCO, …, CCCC} (8 outcomes)
- Assume C_i's are independent events.

$$\begin{aligned} P(A) &= P[C_4 \cap (C_3 \cup (C_1 \cap C_2))] \\ &= P[C_4] \cdot P[C_3 \cup (C_1 \cap C_2)] \\ &= P[C_4] \cdot (P[C_3] + P(C_1 \cap C_2) - P[C_1 \cap C_2 \cap C_3]) \\ &= P[C_4] \cdot (P[C_3] + P(C_1)P(C_2) - P(C_1)P(C_2)P(C_3)]) \\ &= p^2 + p^3 - p^4 \end{aligned}$$

Independence let us express the probability of an interesting event in terms of individual probabilities of basic events.
Can now assess how reliable each component needs to be to ensure overall reliability of system.
Caution with assumptions: Apollo 13 - independent failures of subsystems?

Example
Suppose we have a biased coin with $P(H) = 4/5$.
Flip coin 3 times.
2^3 outcomes. But *not* all equally likely now!
What is $P[(HHT)]$ i.e. $P(H_1 \cap H_2 \cap T_3)$?
Since flips are physically independent, it is reasonable to assume they are statistically independent, so $P(H_1 \cap H_2 \cap T_3) = P(H_1)P(H_2)P(T_3) = (4/5)^2(1/5)^1$

Independent Subexperiments
As in preceding example, experiments often consist of a sequence of subexperiments; often it is reasonable to assume the subexperiments are independent. Then we assume that events associated with different sub-experiments are independent, and we do not have to verify all the conditions above.

The canonical example of a sequence of independent subexperiments is called:

Bernoulli Trials

A **Bernoulli Trial** is a random experiment with 2 outcomes called "success" and "failure" (cf H,T yes,no 1,0 etc.)
We use $p = P\{\text{ success}\}$ and $q = P\{\text{ failure}\} = 1 - p$

Bernoulli Trials a sequence of independent Bernoulli trial subexperiments

Question: what is P(k successes in n trials) ?

Example: 3 toss of a biased coin with $p = P$(heads) What is probability of getting 2 heads (exactly)?

$$\begin{aligned} P(\{HHT, THH, THH\}) &= P(\{HHT\}) + P(\{THH\} + P(\{THH\}) \\ &= P(H)P(H)P(T) + P(T)P(H)P(H) + P(T)P(H)P(H) \\ &= 3p^2(1-p)^1 = \left(\begin{array}{c} 3 \\ 2 \end{array} \right) p^2 q^{3-2} \end{aligned}$$

In general

$$P(\{k \text{ successes in } n \text{ trials}\}) = (\#\text{ of ways to get } k \text{ successes in } n \text{ trials}) \cdot p^k q^{n-k} = \boxed{\left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k}}$$

called the **Binomial Probability Law**

Note: if $p = q = 1/2$ (fair coin), then all outcomes are equally likely, so $P(k \text{ in } n) = \left(\begin{array}{c} n \\ k \end{array} \right) / 2^n$ which agrees with earlier counting method.

Binomial theorem: $(a+b)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) a^k b^{n-k}$ where $\left(\begin{array}{c} n \\ k \end{array} \right)$ is called the **Binomial coefficient**.

Example: Binary data on CD player in 8-bit words.
Probability that a given bit is flipped = p
Assume bit errors are independent.
Error-correcting code fails if 3 or more bits are flipped (hypothetically)
What is probability of failure?
$B_k = \{k \text{ bits flipped}\}$
$E = \{\text{failure}\} = B_3 \cup \ldots \cup B_8$
$P(E) = \sum_{k=3}^{8} P(B_k) = \sum k = 3^8 \left(\begin{array}{c} 8 \\ k \end{array} \right) p^k q^{8-k}$
Such calculations surely done by philips and sony engineers

Approximations
Stirling's formula: $m! \approx \sqrt{\pi m} \cdot m^m \cdot e^{-m}$ for m large
DeMoivre-Laplace approximation:

$$\left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k} \approx \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{(k-np)^2}{2npq} \right)$$

(follows from Central Limit Theorem) if n, k, $n-k$ all large and $|k - np|/\sqrt{npq} < 1$

Poisson approximation (for n large p small):

$$\left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k} \approx (np)^k e^{-np}/k!$$

Random Variables

In many engineering problems, the random quantities of interest are numerical (voltages, currents, forces, strains).
"Random variables" provide a rigorous but convenient tool for describing and analyzing numerical random phenomena.

Example

Toss biased coin 3 times. p =P(H)
Win $1 for each head and lose $1 for each tail.

Outcome s	Probability $P(\{s\})$	"Winnings" $X(s)$
HHH	p^3	3
HHT	p^2q	1
HTH	p^2q	1
THH	p^2q	1
HTT	pq^2	-1
TTH	pq^2	-1
THT	pq^2	-1
TTT	q^3	-3

The function $X(s)$ is called a random variable because it maps each possible outcome into a real number.
From the point of view of a gambler, the value of X is all that matters, not the specific outcome s.
So the properties of X characterize the problem more parsimoniously.

In fact, all events of interest (to a gambler) can be expressed in terms of the possible values of X.
Example: $W = \{$Win money$\}$
Old way: $W = \{$HHH,HHT,HTH,THH$\}$
New parsimonious way: $W = \{X > 0\} = \{s \in S : X(s) > 0\}$
$P(W) = P[X > 0] = P[(X = 1) \cup (X = 3)] = P[X = 1] + P[X = 3] = 3p^2q + p^3$
(doesn't do justice to power of random variables)

Voltage across a resistor vs state of all electrons within...

Definition

Given a sample space S, we call X a random variable if $X(s)$ is a function that assigns to each outcome $x \in S$ a real number.
Formally: $X : S \to [-\infty, \infty]$
Technical conditions:

- The event $[X \leq x] = \{s \in S : X(s) \leq x\}$ must be an event whose probability we can determine for any x.
- $P[X = +\infty] = P[X = -\infty] = 0$.

We *always* use capital letters to denote random variables, usually from end of alphabet.

This formal definition allows us to connect random variables with the Axioms of Probability etc. As we go on, we will develop tools for manipulating random variables "generically," without reference to the underlying sample space, in the same way that a linear systems course describes generic methods for analyzing signals, without reference to the physical phenomena that generated those signals.

Example
Circular dart board, uniformly distributed throws.
$S = \{(a,b) : \sqrt{a^2+b^2} \leq 30\}$
Each outcome $s \in S$ is a coordinate pair $s = (a,b)$
If $A \subseteq S$, then $P(A) = \text{area}(A)/(\pi 30^2)$
For dart player, numerical quantity of interest is distance from center:
$X(s) = \sqrt{(a-1)^2 + (b-2)^2}$
Range of X is $[0, 30]$
For a betting dart player, probability of interest is:
$P[X \leq r]$ for various r.
If $0 \leq r \leq 30$, then $P[X \leq r] = \pi r^2/(\pi 30^2) = (r/30)^2$
If $r \geq 30$, then $P[X \leq r] = 1$
If $r < 0$, then $P[X \leq r] = 0$

Example
Fair wheel of fortune
$S = [0, 360)$
If $[a,b) \subseteq S$, i.e. if $0 \leq a \leq b < 360$, then $P[[a,b)] = (b-a)/360$
Consider

$$X(s) = \frac{180}{|180 - s|}$$

Note: $P[X = \infty] = P[\{s \in S : X(s) = \infty\}] = P[\{180\}] = 0$
The outcome $X = \infty$ is "possible," but its probability is 0.
Probability 0 does not mean "impossible." Only "extremely unlikely."
The relative frequency concept does not quite explain P=0.

For $a > 1$:

$P[X \leq a] = P(\{s \in S : \frac{180}{|180-s|} \leq a\})$
$= P(\{s : |180 - s| \geq 180/a\})$
$= P(\{s : 180 - s \geq 180/a\} \cup \{s : s - 180 \leq 180/a\})$
$= P(\{s : s \leq 180(1 - 1/a)\} \cup \{s : s \geq 180(1 + 1/a)\})$
$= P(\{s : s \leq 180(1 - 1/a)\}) + P(\{s : s \geq 180(1 + 1/a)\})$
$= P([0, 180(1 - 1/a)]) + P([180(1 + 1/a), 360))$
$= (1/2)(1 - 1/a) + 1 - (1/2)(1 + 1/a)$ Thus

$$P[X \leq a] = \begin{cases} 0, & a \leq 1 \\ 1 - 1/a, & a \geq 1 \end{cases}$$

All the events of practical interest can be expressed as set combinations of the following two types of events:
- Events of the form $[X = x]$
- Events of the form $[X \leq x]$

We calculate probabilities for such events using the method of "equivalent events."

$$[X \in B] = \{s \in S : X(s) \in B\} = X^{-1}(B)$$

$$P[X \in B] = P(A), \text{ where } A = \{s \in S : X(s) \in B\}$$

Cumulative Distribution Function (CDF)

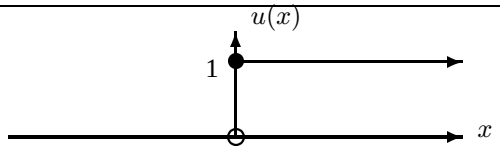
The cumulative distribution function (CDF) of a random variable X is defined to be

$$F_X(x) = P[X \leq x], \text{ for } -\infty < x < \infty$$

Read: the probability of the event that the random variable X takes on a value in the set $(-\infty, x]$
The X subscript reminds us which r.v. when more than one in a problem
The argument x is just a placeholder: $F_X(a) = P[X \leq a]$ is equally good

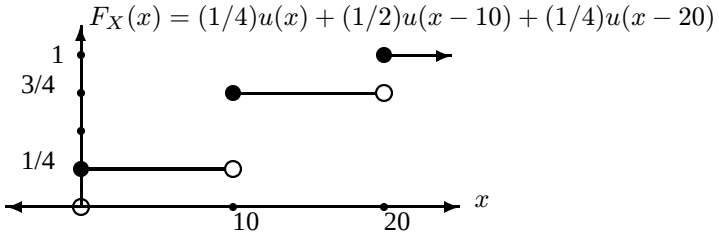
Unit step function

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



Note: $u(0) = 1$ is crucial!

Example: CDF of discrete random variable
Toss 2 fair coins
X is 10 times number of heads
$S_X = \{0, 10, 20\}$
$P[X = 0] = 1/4$
$P[X = 10] = 1/2$
$P[X = 20] = 1/4$



Role of CDF for random phenomena analogous to role of Fourier transform in linear systems theory: generic description

Properties of CDF
- $0 \leq F_X(x) \leq 1$
- $\lim_{x \to \infty} F_X(x) = 1$
- $\lim_{x \to -\infty} F_X(x) = 0$
- $F_X(x)$ is monotone nondecreasing. If $a \leq b$, then $F_X(a) \leq F_X(b)$
  Proof: $(-\infty, a] \subseteq (-\infty, b]$ if $a \leq b$.
- $F_X(x)$ is continuous from the right. $F_X(b) = F_X(b^+) = \lim_{\delta \searrow 0} F_X(b + \delta)$, (for $\delta > 0$).
- $P[a < X \leq b] = F_X(b) - F_X(a)$
  Proof: if $a \leq b$, then $(-\infty, b] = (-\infty, a] \cup (a, b]$, so $[X \leq b] = [X \leq a] \cup [a < X \leq b]$, (usual trick)
  so $P[X \leq b] = P[X \leq a] + P[a < X \leq b]$.
- $P[X = b] = F_X(b) - F_X(b^-)$, where $F_X(b^-) = \lim_{\delta \searrow 0} F_X(b - \delta)$.
  Proof: $P[b - \delta < X \leq b] = F_X(b) - F_X(b - \delta)$ and $P[X = b] = \bigcap_{\delta > 0} [b - \delta < X \leq b]$
  Thus, if cdf is continuous at b, then $P[X = b] = 0$
- $P[X > x] = 1 - F_X(x)$

If 2-5 hold, then F_X is a valid CDF

Probabilities of Other Intervals
$P[a < X < b]$?
$[a < X \leq b] = [a < X < b] \cup [X = b]$ so $P[a < X \leq b] = P[a < X < b] + P[X = b]$
Thus $P[a < X < b] = P[a < X \leq b] - P[X = b]$

Types of Random Variables
- **Discrete** random variables
  Range(X) is a finite or countably infinite set.
  Examples: $S_X = \{-3, -1, 1, 3\}$ or $S_X = \{2, 4, 8, 16, \ldots\}$ General: $S_X = \{x_1, x_2, \ldots\}$
- **Continuous** random variables take a continuous range of values.
  Formally: $P[X = x] = 0$ for all x
- **Mixed** random variables are neither of the above.
  Examples: half-wave rectified random voltage, waiting time in queues that empty (Ex. 3.6).

Discrete random variables have "stair step" CDFs
(zero slope everywhere except a countable number of jump discontinuities)
Continuous random variables have continuous CDFs: no jump discontinuities.

CDF of Discrete Random Variables
Let X be a discrete r.v. with range $S_X = \{x_1, \ldots, x_n\}$
(finite for illustration, could also be countably infinite)

$$[X \leq x] = \bigcup_{\{i \,:\, x_i \leq x\}} [X = x_i]$$

Thus

$$P[X \leq x] = \sum_{\{i : x_i \leq x\}} P[X = x_i] = \sum_{i=1}^{n} P[X = x_i] \, u(x - x_i)$$

since $u(x - x_i)$ is one if $x_i \leq x$ and zero otherwise.
Thus for discrete r.v.:

$$F_X(x) = \sum_{i=1}^{n} P[X = x_i] \, u(x - x_i)$$

cf previous examples

Sometimes we use the shorthand notation: $P(x_i) = P[X = x_i]$
This form is often called the **probability mass function** (PMF)
Note that $\sum_{i=1}^{n} P(x_i) = 1$

Height of jump discontinuity at x_i equals $P[X = x_i]$
CDF is a bid redundant for discrete r.v. since $P(x_i) = P[X = x_i]$ completely describes r.v., but included for consistency with continuous r.v.

Histogram for Discrete r.v.
Height of bar is $P[X = x_i]$

Mystery CDF
(illustrates that CDF tells all)

CDF of Uniform r.v.
Experiment: spin fair wheel of fortune.
$S = [0, 360)$
For $0 \leq a \leq b < 360$, $P([a, b)) = (b - a)/360$ (uniformly likely)
Define $X(s) = s/360$
$S_X = [0, 1)$ For $0 < x < 1$ $P[X \leq x] = P[0 \leq X \leq x] P[0 \leq s/360 \leq x] = P([0, 360x]) = x$

$$F_X(x) = \begin{cases} 0, & x \leq 0 \\ x, & 0 \leq x \leq 1 \\ 1, & 1 \leq x \end{cases}$$

For any $a \in [0, 9/10]$ $P[a < X \leq a + 1/10] = F_X(a + 1/10) - F_X(a) = (a + 1/10) - (a) = 1/10.$ independent of a
X is equally likely to fall in any interval of length 1/10 in [0,1]
Can replace 1/10 by any small positive number δ
random number generator is like spinning a wheel and normalizing to [0,1]

Recall $P[a < X \leq b] = F_X(b) - F_X(a)$
If $F_X(\cdot)$ is continuous, then by calculus:

$$P[a < X \leq b] = F_X(b) - F_X(a) = \int_{a^+}^{b+} \frac{d}{dx} F_X(x) \, dx$$

This representation is so useful and important that it has its own name. (it is even more useful than cdf)

Probability Density Function (pdf)
The probability density function (pdf) of a random variable X is defined to be

$$f_X(x) = \frac{d}{dx} F_X(x)$$

(details about nondifferentiability dealt with soon)

Example: pdf of Uniform(0,1) R.V
From wheel-of-fortune example:

$$F_X(x) = \begin{cases} 0, & x \leq 0 \\ x, & 0 \leq x \leq 1 \\ 1, & 1 \leq x \end{cases} \quad \text{so} \quad f_X(x) = \begin{cases} 0, & x \leq 0 \\ 1, & 0 \leq x \leq 1 \\ 0, & 1 \leq x \end{cases}$$

equally likely to take any value between 0 and 1...

Properties of pdf
- $f_X(x) \geq 0 \; \forall x$
- $\int_{-\infty}^{\infty} f_X(x) \, dx = 1$
  Proof: $\int_{-\infty}^{\infty} f_X(x) \, dx = \lim_{T\to\infty} \int_0^T f_X(x) \, dx + \lim_{T\to\infty} \int_{-T}^{\infty} f_X(x) \, dx$
  $= \lim_{T\to\infty} F_X(T) - F_X(0) + \lim_{T\to\infty} F_X(0) - F_X(-T) = 1 - F_X(0) + F_X(0) - 0 = 1$
- $F_X(x) = \int_{-\infty}^{x^+} f_X(t) \, dt$
- $P[a < X \leq b] = \int_{a^+}^{b+} f_X(x) \, dx = F_X(b) - F_X(a)$

If 1st two hold, then f_X is valid pdf

Interpretation of pdf
$P[x - \delta < X \leq x] = \int_{x-\delta}^{x} f_X(x) \, dx \approx \delta \cdot f_X(x)$ for small δ
So $f_X(x)\delta$ is approximately the probability of the event that the r.v. X takes a value in an interval near x of width δ.
Higher density = more likely: cf bell curve

---

pdf of Discrete Random Variables

Recall if X is a discrete r.v. with range $S_X = \{x_1, \ldots, x_n\}$, then

$$F_X(x) = \sum_{i=1}^{n} P[X = x_i] \, u(x - x_i)$$

so

$$f_X(x) = \sum_{i=1}^{n} P[X = x_i] \, \frac{d}{dx} u(x - x_i) = \sum_{i=1}^{n} P[X = x_i] \, \delta(x - x_i)$$

Similar to Histogram!

Dirac Delta "Function"

The Dirac Delta function is the (generalized) derivative of the unit step function

$$\delta(x) = \frac{d}{dx} u(x)$$

The unit step function is the integral of the Dirac delta:

$$u(x) = \int_{-\infty}^{x} \delta(t) \, dt$$

Properties

area is 1

sifting

derivatives three cases

Memorize: uniform, exponential, Gaussian pdf

Exponential Distribution

$$f_X(x) = ce^{-x/\mu}u(x)$$

What is c? Need $\int f_X(x)\,dx = 1$ so

$$1 = \int_0^\infty ce^{-x/\mu}\,dx = c(-\mu)\,e^{-x/\mu}\Big|_0^\infty = c\mu$$

Thus $c = 1/\mu$
This density is a popular model for the (random) lifetime of systems.
X denotes the failure time relative to some starting time (typically 0, hence X is the lifetime of the system)
μ is called the mean lifetime (units: time)

$P[3 < X < 4] = \int_3^4 f_X(x)\,dx = F_X(4) - F_X(3)$

$F_X(x) = \int_0^x \frac{1}{\mu}e^{-t/\mu}\,dtu(x) = (1 - e^{-x/\mu})u(x)$, (cdf of r.v. with exponential distribution)
so $P[3 < X < 4] = (1 - e^{-4/\mu}) - (1 - e^{-3/\mu}) = e^{-3/\mu}) - e^{-4/\mu}$
If $\mu = 3$ years, then $P[3 < X < 4] = e^{-3/3} - e^{-4/3} \approx 0.104$
(cf extended warranty on VCR)

New perspective: assume some model for a random phenomena, then compute probability of interest.
As opposed to starting with some sample space with assigned probabilities, defining a random variable X as a function on that sample space, then deriving the pdf of X.

Application of Exponential Distribution to Reliability
Suppose a system fails when *any* of its n components fail.
Let X_i denote failure time of ith component, $i = 1, \ldots, n$
Let Y denote failure time of system.
Find cdf of Y
$[Y \le t] = [X_1 \le t] \cup \cdots \cup [X_n \le t]$ (system fails before time t if any component fails before time t)
Thus $P[Y \le t] = P([X_1 \le t] \cup \cdots \cup [X_n \le t])$
Not disjoint events! But suppose we assume component failure times are independent:
$P[Y \le t] = 1 - P([X_1 > t] \cap \cdots \cap [X_n > t]) = 1 - \prod_{i=1}^n P[X_i > t]$
Thus $F_Y(t) = 1 - \prod_{i=1}^n (1 - F_{X_i}(t))$
So reliability of system related to reliability of components.
A *general* relationship independent of any underlying sample space

If each X_i has exponential cdf with mean μ, then $F_Y(t) = 1 - \prod_{i=1}^n e^{-t/\mu} = 1 - e^{-nt/\mu} = 1 - e^{-t/(\mu/n)}$
which is exponential cdf with mean μ/n, so mean lifetime reduced by factor of n

Some important discrete random variables

Bernoulli (1 with prob. p, 0 with prob. $1-p$)

$$f_X(x) = (1-p)\delta(x-1) + p\delta(x)$$

Binomial $X \sim \text{Binom}(n,p)$ (number of successes in n independent Bernoulli trials with success prob. p)

$$f_X(x) = \sum_{k=0}^{n} P[X=k]\delta(x-k) = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \delta(x-k)$$

Poisson $X \sim \text{Poisson}(\lambda)$ (useful for counting number of occurrences within a finite time window)

$$f_X(x) = \sum_{k=0}^{\infty} P[X=k]\delta(x-k) = \sum_{k=0}^{n} e^{-\lambda}\lambda^k/k!\delta(x-k)$$

λ is mean number of occurrences

Scale and Shift of Random Variable
Let X be a r.v. with cdf $F_X(x)$
Define $Y = aX + b$ for $a > 0$
Find cdf and pdf of Y
Back to definition: $F_Y(y) = P[Y \leq y] = P[aX + b \leq y] = P[X \leq (y-b)/a] = F_X((y-b)/a)$
Taking derivative w.r.t. y using chain rule to find pdf: $f_Y(y) = d/dy F_Y(y) = d/dy F_X((y-b)/a) = f_X((y-b)/a)/a$

$$F_Y(y) = F_X\left(\frac{y-b}{a}\right) \qquad f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right)$$

Analogous to scale/shift formula for Fourier transforms: signal independent; (here it is pdf independent)
Important to learn steps in such derivations, not just final formula

Gaussian Distribution (aka Normal distribution or "bell curve")
Of great importance in engineering, in part due to central limit theorem
We say X is a Gaussian r.v. iff its pdf has the form

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

μ called mean or location parameter (average value)
σ^2 called variance or scale parameter (spread)
σ called standard deviation
Shorthand to avoid writing pdf: $X \sim N(\mu, \sigma^2)$
Read: "X has a normal distribution with mean μ and variance σ^2
It is often more convenient to work with a "standardized" r.v. with mean 0 and variance 1
If $X \sim N(\mu, \sigma^2)$, then for $Z = (X - \mu)/\sigma$ we have $Z \sim N(0, 1)$, i.e. $f_Z(z) = \frac{1}{2\pi} e^{-z^2/2}$
$f_Z(z) = \frac{1}{2\pi} e^{-z^2/2}$ is called the standard normal distribution.
Proof: by scale/shift property, $f_Z(z) = \sigma f_X(z\sigma + \mu) = \frac{1}{2\pi} e^{-z^2/2}$
The transformation $Z = (X - \mu)/\sigma$ is called standardizing.

$$F_Z(z) = P[Z \leq z] = \int_{-\infty}^{z} \frac{1}{2\pi} e^{-t^2/2} \, dt$$

Unfortunately, no closed form for cdf of standard Gaussian
In engineering, it is customary to work with "the Q function:"

$$Q(z) = 1 - F_Z(z) = P[Z > z] = \int_{z}^{\infty} \frac{1}{2\pi} e^{-t^2/2} \, dt \approx \left[\frac{\pi}{(\pi - 1)z + \sqrt{z^2 + 2\pi}} \right] e^{-z^2/2}$$

Tabulated in most books, or use approximation for $z > 0$. By symmetry, $Q(-z) = 1 - Q(z)$

Calculating probabilities for Gaussian (must express in terms of Q function to use table or approximation):

$$P[a < X < b] = P[\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}] = P[\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}]$$

$$= F_Z(\frac{b-\mu}{\sigma}) - F_Z(\frac{a-\mu}{\sigma}) = Q(\frac{a-\mu}{\sigma}) - Q(\frac{b-\mu}{\sigma})$$

From table, $Q(1.96) \approx 0.025$, thus

$$P[-1.96 < \frac{X-\mu}{\sigma} < 1.96] = P[\mu - 1.96\sigma < X < \mu + 1.96\sigma] = Q(-1.96) - Q(1.96) = 1 - 0.025 - 0.025 = 0.95$$

so a Gaussian r.v. takes values within 2 standard deviations of its mean about 95% of the time

If $X \sim N(\mu, \sigma^2)$ then $F_X(x) = 1 - Q(\frac{x-\mu}{\sigma})$
Also commonly used is error function

$$\text{erf}(x) = \int_0^x \frac{2}{\sqrt{\pi}} e^{-z^2} \, dz$$

Conditional CDF
Recall $P(B|A) = P(A \cap B)/P(A)$ and $F_X(x) = P[X \le x]$
The conditional CDF of a random variable X given event A for which $P(A) \neq 0$ is defined to be

$$F_X(x|A) = F_{X|A}(x|A) = P[X \le x|A]P([X \le x] \cap A)/P(A), \text{ for } -\infty < x < \infty$$

Read: the prob. of the event that the r.v. X takes on a value in the set $(-\infty, x]$ given that event A occurred

Properties of conditional CDF (all same as ordinary cdf)
- $0 \le F_X(x|A) \le 1$
- $\lim_{x \to \infty} F_X(x|A) = 1$
- $\lim_{x \to -\infty} F_X(x|A) = 0$
- $F_X(x|A)$ is monotone nondecreasing. If $a \le b$, then $F_X(a|A) \le F_X(b|A)$
- $F_X(x|A)$ is continuous from the right: $F_X(b|A) = F_X(b^+|A)$
- $P[a < X \le b|A] = F_X(b|A) - F_X(a|A)$
- $P[X = b|A] = F_X(b|A) - F_X(b^-|A)$, where $F_X(b^-|A) = \lim_{\delta \searrow 0} F_X(b - \delta|A)$.

Conditional pdf
The conditional pdf is defined as the derivative of the conditional cdf:

$$f_X(x|A) = f_{X|A}(x|A) = \frac{d}{dx} F_X(x|A)$$

Properties of conditional pdf
- $f_X(x|A) \ge 0 \; \forall x$
- $\int_{-\infty}^{\infty} f_X(x|A) \, dx = 1$
- $F_X(x|A) = \int_{-\infty}^{x^+} f_X(t|A) \, dt$
- $P[a < X \le b|A] = \int_{a^+}^{b+} f_X(x|A) \, dx = F_X(b|A) - F_X(a|A)$

Same conventions for nondifferentiable corners and jump discontinuities

Fact: derivatives of cdfs exist except at most on a set of countably infinite points in $\mathbb{R}$

If $F_X(x|A)$ has a jump discontinuity at x_0, then "at" $x = x_0$: $f_X(x|A) = (F_X(x_0) - F_X(x_0^-)) \cdot \delta(x - x_0)$

Total probability for cdf and pdf
If $A_1, A_2, \ldots$ partition S, and if $P(A_i) \neq 0 \; \forall i$ then

$$F_X(x) = \sum_i F_X(x|A_i) P(A_i) \quad \text{since} \quad P[X \le x] = \sum_i P[X \le x|A_i] P(A_i)$$

Similarly

$$f_X(x) = \sum_i f_X(x|A_i) P(A_i)$$

Example (pdf from conditioning on events)
Company 1 makes 100Ω resistors with 10% tolerance
Company 2 makes 100Ω resistors with 5% tolerance
You buy 25% of your resistors from C_1 and 75% from C_2
Pick resistors at random from common storage bin.
What is the pdf of resistance value?
Plausible assumptions:
$f_X(x|C_1)$ is Uniform(90,110)
$f_X(x|C_2)$ is Uniform(95,105)

$$f_X(x) = f_X(x|C_1)P(C_1) + f_X(x|C_1)P(C_1)$$

Conditioning on interval (homework problem, example 3.10)

Conditioning on point $X = x$
How do we define/compute $P[A|X = x]$, i.e. prob. of snow tomorrow, given todays high temperature was 33 degrees?
$P[A|X = x] = P(A \cap [X = x])/P[X = x]$ ok for discrete r.v., but not for continuous r.v. since $P[X = x] = 0$
Define

$$P[A|X = x] = \lim_{\delta \searrow 0^+} P[A|x - \delta < X \le x] = \lim_{\delta \searrow 0^+} \frac{P(A \cap [x - \delta < X \le x])}{P[x - \delta < X \le x]}$$

provided limit is well defined

Now apply Bayes rule (assuming $P(A) \neq 0$):

$$P(A|x - \delta < X \le x) = \frac{P(x - \delta < X \le x|A)P(A)}{P[x - \delta < X \le x]} = \frac{\frac{1}{\delta}\int_{x-\delta}^{x} f_X(t|A)\, dt}{\frac{1}{\delta}\int_{x-\delta}^{x} f_X(t)\, dt} P(A) \to \frac{f_X(x|A)P(A)}{f_X(x)} \text{ as } \delta \to 0$$

Thus

$$P(A|X = x) = \frac{f_X(x|A)P(A)}{f_X(x)}$$

provided $f_X(x) \neq 0$ and pdfs sufficiently regular
Rearranging we have

$$f_X(x|A)P(A) = P(A|X = x)f_X(x)$$

so by integrating both sides over x:

$$P(A) = \int P(A|X = x)f_X(x)\, dx$$

(another version of total probability)

Functions or transformations of a r.v.
Suppose X is a r.v. with known cdf $F_X(x)$ and pdf $f_X(x)$
Define $Y = g(X)$ for some function $g : \mathbb{R} \to \mathbb{R}$
Find cdf and pdf of Y
Two approaches:
- Method of events: always works
- Plug-and-chug formula: only works for certain g functions

Note that (for all practical g) Y is a well-defined r.v., defined by $Y(s) = g(X(s))$

Transformations by method of equivalent events
General procedure to find cdf of Y when $Y = g(X)$:
For each y find $[Y \le y]$ in terms of corresponding values of X. Formally:

$$[Y \le y] = [g(X) \le y] = [X \in \{x \in \mathbb{R} : g(x) \le y\}] = [X \in g^{-1}((-\infty, y])]$$

so

$$F_Y(y) = P[Y \le y] = P[X \in \{x : g(x) \le y\}]$$

Typically we find the last prob. by integrating f_X over the set $\{x : g(x) \le y\}$, which is usually one or more intervals.
If $B_y = \{x : g(x) \le y\}$, then

$$F_Y(y) = \int_{B_y} f_X(x)\, dx = \int_{\{x:g(x)\le y\}} f_X(x)\, dx$$

Virtually always must find cdf of Y first, then differentiate to get $f_Y(y)$

Example
Suppose X is a voltage with a Uniform[-5,15] distribution
Define $Y = X^2$. What is cdf/pdf of Y?
First: Range(Y) = [0,225] so for $y < 0$, $F_Y(y) = 0$ and for $y \ge 225$, $F_Y(y) = 1$

For $y > 0$

$$[Y \le y] = [X^2 \le y] = [-\sqrt{y} \le X \le \sqrt{y}]$$

so

$$F_Y(y) = P[-\sqrt{y} \le X \le \sqrt{y}] = \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x)\, dx$$

For $\sqrt{y} \le 5$,

$$\int_{-\sqrt{y}}^{\sqrt{y}} f_X(x)\, dx = \int_{-\sqrt{y}}^{\sqrt{y}} 1/20\, dx = 2\sqrt{y}/20$$

For $5 < \sqrt{y} \le 15$,

$$\int_{-\sqrt{y}}^{\sqrt{y}} f_X(x)\, dx = \int_{-5}^{\sqrt{y}} 1/20\, dx = (\sqrt{y} + 5)/20$$

Thus

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ \sqrt{y}/10, & 0 \le y \le 25 \\ (\sqrt{y} + 5)/20, & 25 \le y \le 225 \\ 1, & 15 \le y \end{cases}$$

Taking derivative

$$f_Y(y) = \begin{cases} \frac{1}{20\sqrt{y}}, & 0 < y \le 25 \\ \frac{1}{40\sqrt{y}}, & 25 < y \le 225 \\ 0, & \text{otherwise} \end{cases}$$

A mixed random variable (introduces transformations)
Suppose X is a voltage with a Uniform[-5,15] distribution
Let Y be a new r.v. defined to be X half-wave rectified:

$$Y = g(X) = \begin{cases} 0, & X \leq 0 \\ X, & X \geq 0 \end{cases}$$

What is cdf/pdf of Y?
First: Range(Y) = [0,15] so for $y < 0$, $F_Y(y) = 0$ and for $y \geq 15$, $F_Y(y) = 1$
For $y = 0$: $F_Y(y) = P[Y \leq 0] = P[Y = 0] = P[X \leq 0] = 5/20$
For $0 < y < 15$: $F_Y(y) = P[Y \leq y] = P[Y = 0] = P[X \leq y] = (y + 5)/20$
Thus

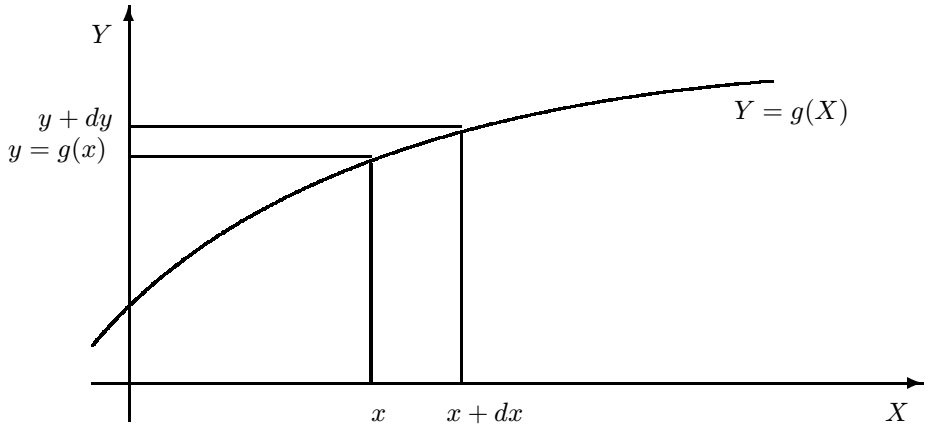
$$F_Y(y) = \begin{cases} 0, & y < 0 \\ (y+5)/20, & 0 \leq y < 15 \\ 1, & 15 \leq y \end{cases}$$

Taking derivative

$$f_Y(y) = \frac{1}{4}\delta(y) + \frac{1}{20}(u(y) - u(y-15)) = \frac{1}{4}\delta(y) + \frac{1}{20}1_{\{0<y\leq 15\}}$$

Monotonic increasing, continuous, differentiable transformations of a continuous r.v.
Suppose X is a continuous r.v. with known cdf $F_X(x)$ and pdf $f_X(x)$
Define $Y = g(X)$ for some function $g : \mathbb{R} \to \mathbb{R}$
Assume $g(x)$ is monotonic increasing, continuous, differentiable
Find cdf and pdf of Y

Intuition



Equivalent events: $[y \le Y \le y + dy] = [x \le X \le x + dx]$
Thus $P[y \le Y \le y + dy] = P[x \le X \le x + dx]$
So for small dy we have $f_Y(y)dy = f_X(x)dx$ for the value of x such that $g(x) = y$.
Denote this value $x = g^{-1}(y)$.
Inverse exists due to monotonicity of $g()$

$$f_Y(y) = \frac{f_X(x)}{dy/dx} = \left.\frac{f_X(x)}{g'(x)}\right|_{x=g^{-1}(y)}$$

where $g'(x) = dy/dx = \frac{d}{dx}g(x)$
Note $g'(x) \neq 0$ since g is monotone increasing

Alternative derivation (via equivalent events)

$$[Y \le g(x)] = [g(X) \le g(x)] = [X \le x]$$

since g monotone increasing. Thus

$$P[Y \le g(x)] = P[X \le x] \quad \text{so} \quad F_Y(g(x)) = F_X(x)$$

Differentiating using chain rule:

$$f_Y(g(x))g'(x) = f_X(x) \quad \text{or} \quad f_Y(y) = \left.\frac{f_X(x)}{g'(x)}\right|_{x=g^{-1}(y)}$$

For monotonic decreasing transformations, dy/dx is negative.
General formula for monotonic (strictly increasing or decreasing), continuous, differentiable g:

$$f_Y(y) = \left.\frac{f_X(x)}{|g'(x)|}\right|_{x=g^{-1}(y)}$$

Continuous, differentiable g with "no flat segments"
Now there can be multiple x's for which $g(x) = y$
Let n_y be the number of x's for which $g(x) = y$ (depends on y)
So there are $x_1, x_2, \ldots, x_{n_y}$ roots for which $g(x_i) = y$

$$[y \le Y \le y + dy] = [x_1 \le X \le x_1 + dx_1] \cup [x_2 \le X \le x_2 + dx_2] \cup \cdots \cup [x_n \le X \le x_n + dx_n]$$

Thus

$$P[y \le Y \le y + dy] = \sum_{i=1}^{n_y} [x_i \le X \le x_i + dx_i]$$

$$f_Y(y) = \sum_{\{x_i : g(x_i) = y\}} \frac{f_X(x_i)}{|g'(x_i)|}$$

Example: $Y = X^4$
Roots at $x = \pm \sqrt[4]{y}$ for $y \ge 0$
$g'(x) = 4x^3$
At $x = \pm \sqrt[4]{y}$, we have $|g'(x)| = 4y^{3/4}$

$$f_Y(y) = \begin{cases} \frac{f_X(\sqrt[4]{y})}{4y^{3/4}} + \frac{f_X(-\sqrt[4]{y})}{4y^{3/4}}, & y \ge 0 \\ 0, & \text{otherwise} \end{cases}$$

Sanity check: $Y = aX + b$
$|g'(x)| = |a|$
Solution to $y = ax + b$ is $x = (y - b)/a$

$$f_Y(y) = \frac{f_X((y-b)/a))}{|a|}$$

(same as before)

Linear transformation of Gaussian r.v.
Suppose $X \sim N(\mu_X, \sigma_Y^2)$
Let $Y = aX + b$

$$f_Y(y) = \frac{f_X((y-b)/a))}{|a|} = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{((y-b)/a - \mu_X)^2}{2\sigma_X^2} \right) / |a|$$

$$= \frac{1}{\sqrt{2\pi}(|a|\sigma_X)} \exp\left(-\frac{(y - (a\mu_X + b))^2}{2(a\sigma_X)^2} \right) = \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left(-\frac{(y - \mu_Y)^2}{2\sigma_Y^2} \right)$$

where $\mu_Y = a\mu_X + b$ and $\sigma_Y = |a|\sigma_X$
$Y \sim N(\mu_Y, \sigma_Y^2)$
$aX + b \sim N(a\mu_X + b, a^2\sigma_X^2)$

Thus linear transformation of a Gaussian r.v. yields a new Gaussian r.v.
(Gaussian distribution preserved under linear transformations)

Functions of a discrete r.v.
Suppose X is a discrete r.v. with range $\{x_1, x_2, \ldots\}$ and known pmf $P[X = x_i]$
Define $Y = g(X)$ for some function $g : \mathbb{R} \to \mathbb{R}$
Clearly Y is a discrete r.v. and the range of Y is $S_Y = \cup_i \{g(x_i)\}$
It is easy to find the pmf of Y
If $y \in S_Y$, then

$$P[Y = y] = P[g(X) = y] = \sum_{\{i: g(x_i) = y\}} P[X = x_i]$$

i.e. the probability Y takes the value y is the sum of the probabilities that X takes those values x_i for which $g(x_i) = y$.
If $g(\cdot)$ is monotone increasing or decreasing, then there will be only one such x value for each y, namely $x = g^{-1}(y)$

Computer generation of r.v.
(reverse of above thinking!)

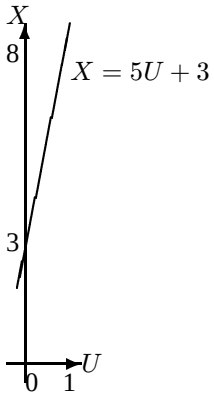
Computer generation of discrete r.v.
Recall from computer assignment:
to generate Bernoulli r.v. with $P[Y = 1] = p$ and $P[Y = 0] = 1 - p$ we used $U \sim$ Uniform$(0, 1)$ and then defined $Y = g(U)$ where

$$g(u) = \begin{cases} 1, & u < p \\ 0, & \text{otherwise} \end{cases}$$

This approach can be generalized for other discrete r.v., just split up interval...

Computer generation of continuous r.v. by "Transformation Method"

Most computing languages provide subroutines only for generating pseudo-random numbers from the Uniform$(0, 1)$ distribution.
To generate realizations of a continuous random variable X with cdf $F_X(x)$ we must first generate $U \sim$ Uniform$(0, 1)$ and then transform it by $X = g(U)$ for some function $g(\cdot)$.
How to generate $X \sim$ Uniform$(3, 8)$?
Intuitively: $X = 5U + 3$.
From scale and shift of pdf: $f_X(x) = \frac{1}{5} f_U(\frac{x-3}{5})$
Note: scaled down since "stretched out" but must integrate to 1



"Transformation Method" for generating continuous r.v. (on a computer)
To generate realizations of a continuous random variable X with cdf $F_X(x)$ we must first generate $U \sim \text{Uniform}(0,1)$ and then transform it by $X = g(U)$ for some function $g(\cdot)$.
In general we must use:

$$X = F_X^{-1}(U) = g(U) \text{ where } g(u) = F_X^{-1}(u)$$

where $F_X^{-1}(\cdot)$ is the inverse of the function F_X.
Note that since F_X is monotone increasing for continuous a r.v., so the inverse is well defined.
In practice we find this inverse by setting $F_X(x) = u$ and solving for x in terms of u. That gives us some relationship $x = g(u)$ and we subsequently use $X = g(U)$ in the computer.

Proof (that if we use the above transformation, then X will have the desired cdf F_X.

$$P[X \le x] = P[g(U) \le x] = P[F_X^{-1}(U) \le x] = P[U \le F_X(x)] = F_X(x)$$

since F_X is monotone increasing for continuous r.v.
since $F_U(u) = u$ for $U \sim \text{Uniform}(0,1)$

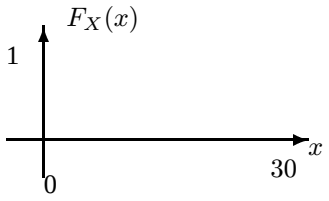
Caution
Note that

$$[F_X(u)]^{-1} = \frac{1}{F_X(u)}$$

i.e., a number raised to the -1 power is the reciprocal of the number.
But $F_X^{-1}(u)$ is the value of x that satisfies $F_X(x) = u$, which is almost never the reciprocal of $F_X(u)$

Example
Recall that for the 30-cm circular dart board example the CDF of the r.v. X (distance to center) was

$$F_X(x) = \begin{cases} 0, & x \le 0 \\ (x/30)^2, & 0 \le x \le 30 \\ 1, & 30 \le x \end{cases}$$



Set $u = F_X(x)$ and solve for x.
(We only need to consider $0 < u < 1$ since U is a Uniform(0,1) r.v.)
So $u = (x/30)^2$; thus $x = 30\sqrt{u}$
Thus, to generate a r.v. with the above CDF, use $X = 30\sqrt{U}$.
In Matlab, to generate 1000 realizations: `X = 30 * sqrt(rand(1000,1))`

(book does exponential)

Synopses of the properties of a r.v.
The pdf is a complete description of the behavior of a r.v. (any probability can be computed from it). Often, due to lack of data or no known model for a random phenomena, one must resort to simpler quantities that characterize part of the random behavior, but are less complete than the entire pdf.
Example: summarizing mean and std. dev. after an exam

Possible "summary statistics:"
- median: point a such that $\int_{-\infty}^{a} f_X(x)\,dx = \int_{a}^{\infty} f_X(x)\,dx = 1/2$.
- upper quartile: point b such that $\int_{b}^{\infty} f_X(x)\,dx = 1/4$
- mode: value of x where $f_X(x)$ is maximum
- "center of mass" or mean: $\mu = \int_{-\infty}^{\infty} x f_X(x)\,dx$
- "moment of inertia" or variance: $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\,dx$

If $f_X(x)$ is symmetric about some point x_0, then mean = median = mode = x_0

The most important of the above is usually mean and variance,
or standard deviation σ, which is square root of variance

Mean or Average (essentially equivalent concepts for discrete r.v.)
Interpretations:
If n students in class with ages $a_1, \ldots, a_n$ then average age: $\mu = \frac{1}{n}\sum_{j=1}^{n} a_j$
There are only m possible ages $x_1, \ldots, x_m$, e.g. $19, 20, \ldots, 28$
Let n_i be the number of students whose age is x_i.
Note that $\sum_{i=1}^{m} n_i = n$
Another way to compute mean is

$$\mu = \frac{1}{n}\sum_{i=1}^{m} x_i \cdot n_i = \sum_{i=1}^{m} x_i \cdot \frac{n_i}{n}$$

Take board of length $m+1$ cm, label $19, 20, \ldots, 28$ put a 1gram weight at age of each student.
Balancing point (center of mass) will be μ

So far no r.v.! Now suppose have each student put age on a slip of paper and put in a hat. Draw one out at random, call it X_1. Put it back, and draw out another one, call it X_2. etc. Take "average" $(X_1 + \ldots + X_N)/N$ for a large number of draws N
By "law of large numbers" this average will be approximately μ
In this experiment, for $i = 1, \ldots, m$ we have $P[X = x_i] = n_i/n$
So from above

$$\mu = \sum_{i=1}^{m} x_i \cdot P[X = x_i] \tag{1}$$

This is the most useful formula for computing the mean of a discrete R.V.
Example: roll fair die. What is average? $\mu = \sum_{i=1}^{6} i \cdot (1/6) = 3.5$
Note that mean does not need to be in the range of r.v.!
Think: toss die repeatedly and take long-term average...

Example
Let X be the number of Bernoulli attempts required to get one success, where $p = P$(success)
Geometric: $P[X = k] = P[k - 1$ failures followed by a success$] = (1 - p)^{k-1} p$
Mean:

$$E[X] = \sum_{k=1}^{\infty} kP[X = k] = \sum_{k=1}^{\infty} k(1-p)^{k-1} p = p \sum_{k=1}^{\infty} k(1-p)^{k-1} = p \left(\frac{1}{p^2} \right) = \frac{1}{p}$$

sensible!

Note for $|a| < 1$:

$$\sum_{k=1}^{\infty} ka^{k-1} = \frac{d}{da} \sum_{k=1}^{\infty} a^k = \frac{d}{da} \left(\frac{a}{1-a} \right) = \frac{1 - a - a(-1)}{(1-a)^2} = \frac{1}{(1-a)^2}$$

The formula $\mu = \sum_i x_i \cdot P[X = x_i]$ is perfect for any discrete r.v.,
but we also need a similar concept for continuous r.v. and mixed r.v.,
hopefully one that is consistent with the discrete r.v. interpretation.

All r.v.'s have a pdf
so express above formula in terms of pdf

The pdf of X (for discrete r.v.) is:

$$f_X(x) = \sum_i P[X = x_i] \delta(x - x_i)$$

so from (1):

$$\int_{-\infty}^{\infty} x f_X(x) \, dx = \int_{-\infty}^{\infty} x \sum_i P[X = x_i] \delta(x - x_i) \, dx = \sum_i P[X = x_i] \int_{-\infty}^{\infty} x \delta(x - x_i) \, dx = \sum_i P[X = x_i] x_i = \mu$$

This gives us a universally useful formula for the mean of any r.v. X

Expectation

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

provided

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

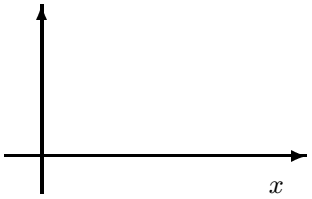
(otherwise we say the mean is *undefined* or *does not exist*)

Note: $E[X]$ is a property of the pdf $f_X(x)$, not of the values of X in a particular realization of the experiment.

If $P[X = a] = 1$, then $E[X] = a$, but the converse is not true!

Example: Uniform r.v. If $X \sim \text{Uniform}(a, b)$ then

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_a^b x \frac{1}{b-a}\, dx = \frac{1}{2}\frac{b^2 - a^2}{b-a} = \frac{b+a}{2}$$



sensible!

Symmetry Property
If $f_X(x)$ is symmetric about m, i.e. $f_X(m + x) = f_X(m - x)$ (i.e. $f_X(x) = f_X(2m - x)$) then $E[X] = m$
Proof:

$$\mu = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_{-\infty}^{\infty} x f_X(2m - x)\, dx = \int_{-\infty}^{\infty} (2m - y) f_X(y)\, dy = 2m \int_{-\infty}^{\infty} f_X(y)\, dy - \int_{-\infty}^{\infty} y f_X(y)\, dy = 2m - \mu$$

Thus $\mu = m$
Agrees with $\mu = (a + b)/2$ for Uniform(a,b)

Gaussian
The pdf of a Gaussian $X \sim N(\mu, \sigma^2)$ is symmetric about μ.
Thus $E[X] = \mu$, so we were justified in calling μ the mean!

Mean of a function of a r.v.
Suppose we want $E[g(X)]$ for some function $g : \mathbb{R} \to \mathbb{R}$
Note that $g(X)$ is a random variable
Example: roll die repeatedly, take average of *square* of number of dots
Hard way:
- Define $Y = g(X)$, now we want $E[Y] = E[g(X)]$
- First find pdf $f_Y(y)$ of Y
- Then integrate: $E[Y] = \int_{-\infty}^{\infty} y f_Y(y)\, dy$

Easy way:

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

This *always* works for any function g used in engineering
Only defined if

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

(a technical condition we won't worry about too much)

Function of discrete r.v.

$$\int_{-\infty}^{\infty} g(x) f_X(x)\, dx = \int_{-\infty}^{\infty} g(x) \sum_i P[X = x_i] \delta(x - x_i)\, dx = \sum_i P[X = x_i] \int_{-\infty}^{\infty} g(x) \delta(x - x_i)\, dx = \sum_i P[X = x_i] g(x_i)$$

so

$$E[g(X)] = \sum_i P[X = x_i] g(x_i)$$

For die example we will get $E[X^2] = (1^2 + 2^2 + \cdots + 6^2)/6 = 91/6$

Proof that the two methods are equivalent when g is monotone increasing, continuous, and differentiable.
(The proof for more general case is beyond the scope of this course.)
Recall that for g monotone increasing, continuous, and differentiable:

$$f_Y(y) = \left. \frac{f_X(x)}{g'(x)} \right|_{x = g^{-1}(y)}$$

Now make the change of variables $y = g(x)$ in the integral, noting that $dy = g'(x)dx$ and $x = g^{-1}(y)$:

$$E[Y] = \int_{-\infty}^{\infty} y f_Y(y)\, dy = \int_{-\infty}^{\infty} g(x) \frac{f_X(x)}{g'(x)} g'(x) dx = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

Scale and Shift
If $Y = aX + b$, i.e. $Y = g(X)$ where $g(x) = ax + b$ then

$$E[Y] = E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx = \int_{-\infty}^{\infty} (ax + b) f_X(x)\, dx = a \int_{-\infty}^{\infty} x f_X(x)\, dx + b \int_{-\infty}^{\infty} f_X(x)\, dx = aE[X] + b$$

thus

$$E[aX + b] = aE[X] + b$$

In particular $E[b] = b$ (the average value for a r.v. that is just a constant is the constant)
And $E[X + b] = E[X] + b$, so we can shift the mean of a r.v. by adding a constant to it

---

Linearity
If $g(X) = \sum_j g_j(X)$ then

$$E[g(X)] = E[\sum_j g_j(X)] = \sum_j E[g_j(X)]$$

So can exchange summation and expectation. But $E[g(X)h(X)] \neq E[g(X)]E[h(X)]$ in general.

Indicator Function

If

$$1_A(x) = \begin{cases} 1, & x \in A \\ 0, & \text{otherwise} \end{cases}$$

then

$$E[1_A(X)] = \int_{-\infty}^{\infty} 1_X(x) f_X(x)\, dx = \int_A f_X(x)\, dx = P[X \in A]$$

Conditional Expectation

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

for discrete r.v.:

$$E[X|A] = \sum_i x_i P[X = x_i|A]$$

Example: roll 6-sided fair die. What is $E[X|X > 2]$

$$E[X|X > 2] = \sum_{i=1}^{6} iP[X = i|X > 2] = \sum_{i=3}^{6} i(1/4) = 4.5$$

since

$$P[X = i|X > 2] = P([X = i] \cap [X > 2])/P[X > 2] = \begin{cases} 0, & i = 1, 2 \\ (1/6)/(4/6) = 1/4, & i = 3, 4, 5, 6 \end{cases}$$

Moments of r.v.
The mean only tells us the average value taken of a r.v.
It doesn't tell us other important factors such as how "spread out" the values of X are (e.g. 100Ω resistors ±?), or about asymmetry of pdf.
The moments of a r.v. tell us that information. Most important (central) moment is the variance.

Moments about the origin
The nth moment about the origin of a r.v. X is defined to be:

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

$\mu_0 = 1$
$\mu_1 = \mu = \mu_X = E[X]$, the mean of X

Central Moments
The nth central moment of a r.v. X is defined to be:

$$m_n = E[(X - \mu_X)^n] = \int_{-\infty}^{\infty} (x - \mu_X)^n f_X(x)\, dx$$

$m_0 = 1$
$m_1 = 0$
$m_2 = E[(X - \mu_X)^2] = \sigma_X^2 = \text{Var}[X]$, the variance of X
$m_3 = E[(X - \mu_X)^3]$ is called the "skewness" of X. If pdf of X is symmetric, then $m_3 = 0$

Standard Deviation
$\text{Std}[X] = \sqrt{\text{Var}[X]} = \sigma_X$
Units:
If X has units volts, then μ and σ_X also have units volts. $f_X(x)$ has units 1/volts, and $\text{Var}[X]$ has units volts^2

Gaussian
If $X \sim N(\mu, \sigma^2)$, then $\text{Var}[X] = \sigma^2$

$$\text{Var}[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\, dx = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}\, dx = \sigma^2 \int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy = \sigma^2$$

letting $y = (x - \mu)/\sigma$, since by integration by parts:

$$\int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy = 1$$

so calling σ^2 the variance was correct

Relationship between variance and moments about origin

$$\text{Var}[X] = E[(X - \mu_X)^2] = E[X^2 - 2X\mu_X + \mu_X^2] = E[X^2] - 2E[X]\mu_X + \mu_X^2 = E[X^2] - 2\mu_X^2 + \mu_X^2 = E[X^2] - \mu_X^2$$

so

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

often (but not always!) easier than applying definition directly: $\text{Var}[X] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\, dx$

Effect of shift and scale on variance

$$\text{Var}[aX+b] = E[(aX+b-E[aX+b])^2] = E[(aX+b-(aE[X]+b))^2] = E[(a(X-E[X]))^2] = a^2 E[(X-E[X])^2] = a^2 \text{Var}[X]$$

so shift has no effect on variance, but scaling by a scales variance by a^2
$\text{Var}[b] = 0$
$\text{Var}[X+b] = \text{Var}[X]$
$\text{Var}[aX] = a^2 \text{Var}[X]$

(3.7)

Although the moments of a r.v. are very useful, a finite set of moments does not in general tell you the whole pdf, so you cannot compute exact probabilities knowing only the moments. We can compute bounds on the probabilities though.

Markov Inequality
If X is a nonnegative r.v., i.e. $P[X \geq 0] = 1$ with known mean

$$P[X \geq a] \leq \frac{E[X]}{a} \quad \text{for } a > 0 \quad \text{if } P[X \geq 0] = 1$$

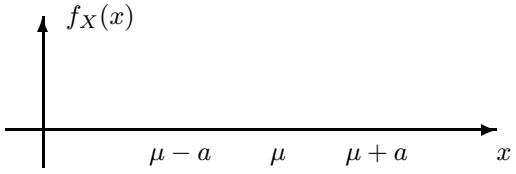
Proof:

$$E[X] = \int_0^\infty x f_X(x)\, dx = \int_0^a x f_X(x)\, dx + \int_a^\infty x f_X(x)\, dx \geq \int_a^\infty x f_X(x)\, dx \geq \int_a^\infty a f_X(x)\, dx = aP[X \geq a]$$

Example: suppose mean age in class is 20
Let X denote age of randomly selected student
$P[X \geq 25] < 20/25 = 0.8$, so no more than 20% of class can be over 25 years old

Chebyshev Inequality
Useful when mean *and variance* of a r.v. X are known
Let $Y = |X - \mu_X|^2$. Clearly $P[Y \geq 0] = 1$, so by Markov inequality $P[Y \geq a^2] \leq E[Y]/a^2$ for $a > 0$
But events $[Y \geq a^2] = [|X - \mu_X| \geq a]$ are equivalent. And $E[Y] = E[|X - \mu_X|^2] = \text{Var}[X]$

$$P[|X - \mu_X| \geq a] \leq \frac{\text{Var}[X]}{a^2} \quad \text{for } a > 0$$



Sensible: as $a \to \infty, P \to 0$
Often gives fairly loose bounds
Perhaps more useful for theoretical derivations (e.g. law of large numbers) than for practice

Expected value minimizes mean squared error
$E[(X-c)^2] \geq \text{Var}[X]$, and the minimum is achieved iff $c = E[X]$

$$E[(X-c)^2] = E[(X - \mu_X - (c - \mu_X))^2] = E[(X-\mu_X)^2] - 2E[X-\mu_X](c-\mu_X) + (c-\mu_X)^2 = \text{Var}[X] + (c-\mu_X)^2$$

How to generate all moments easily?

Characteristic Function (Fourier transform)

$$\Phi(\omega) = E[e^{j\omega X}] = \int_{-\infty}^{\infty} e^{j\omega x} f_X(x)\, dx$$

$j = \sqrt{-1}$, so $j^2 = -1$ and $-j \cdot j = 1$
The above integral always exists. Note that $|\Phi(\omega)| \leq \Phi(0) = 1$
Generate moments from derivatives of Φ:

$$E[X^n] = (-j)^n \left. \frac{d^n}{d\omega^n} \Phi(\omega) \right|_{\omega=0}$$

Proof:

$$\frac{d^n}{d\omega^n} \Phi(\omega) = \frac{d^n}{d\omega^n} \int_{-\infty}^{\infty} e^{j\omega x} f_X(x)\, dx = \int_{-\infty}^{\infty} \frac{d^n}{d\omega^n} e^{j\omega x} f_X(x)\, dx = \int_{-\infty}^{\infty} (jx)^n e^{j\omega x} f_X(x)\, dx$$

so

$$\left. \frac{d^n}{d\omega^n} \Phi(\omega) \right|_{\omega=0} = j^n \int_{-\infty}^{\infty} x^n f_X(x)\, dx = j^n E[X^n]$$

Example: X exponential with mean μ

$$\Phi(\omega) = E[e^{j\omega X}] = \int_{-\infty}^{\infty} e^{j\omega x} f_X(x)\, dx = \int_0^{\infty} e^{j\omega x} \frac{1}{\mu} e^{-x/\mu}\, dx = \int_0^{\infty} \frac{1}{\mu} e^{-x(1/\mu - j\omega)}\, dx = \frac{1}{1 - j\omega\mu}$$

$$\frac{d}{d\omega} \Phi(\omega) = \frac{j\mu}{(1 - j\omega\mu)^2} = j\mu \quad \text{at } \omega = 0$$

By induction:

$$\frac{d^n}{d\omega^n} \Phi(\omega) = \frac{(j\mu)^n n!}{(1 - j\omega\mu)^{n+1}} = j^n \mu^n n! \quad \text{at } \omega = 0$$

So $E[X^n] = \mu^n n!$
In particular $\text{Var}[X] = E[X^2] - \mu^2 = 2\mu^2 - \mu^2 = \mu^2$

Moment Generating Function (Laplace transform)

$$M(s) = E[e^{sX}] = \int_{-\infty}^{\infty} e^{sx} f_X(x)\, dx \quad (\text{Text uses } E[e^{-sX}])$$

$$E[X^n] = \left. \frac{d^n}{ds^n} M(s) \right|_{s=0} \quad (\text{for } E[e^{-sX}] \text{ multiply by } (-1)^n)$$

Advantage: avoids complex numbers. Always exists for X nonnegative for $s \leq 0$.
Disadvantage: integral may not always exist

Probability Generating Function (Z-transform)
For discrete nonnegative integer valued r.v.

$$G(z) = E[z^X] = \sum_{k=0}^{\infty} z^k P[X = k]$$

$$P[X = k] = \frac{1}{k!} \left. \frac{d^k}{dz^k} G(z) \right|_{z=0}$$

Most engineering problems deal with multiple numerical quantities.
Random vectors (or vector random variables) are a general tool for analyzing such random phenomena.
The development of random vectors completely parallels that of random variables:

- 4.1, 4.5: Definition
- 4.2, 4.5: Joint cdf, pdf
- 4.3, 4.5: Independence (new)
- 4.4, 4.5: Conditional cdf, pdf, Bayes rule, total probability
- 4.6: Transformations: $Z = g(X, Y)$
- 4.7: Expectation, Moments
- 4.7: Correlation, covariance (new)
- 4.8: Gaussian random vectors
- skip 4.9, 4.10

4.1

Random Vectors (or vector random variables)

Given a sample space S, a random vector $\underline{X} = (X_1, \ldots, X_n)$ is a n-tuple of random variables: $X_i : S \to \mathbb{R},\ i = 1, \ldots, n$.
Thus each sample point is mapped into a vector of n real numbers, or $\underline{X} : S \to \mathbb{R}^n$

The behavior of a random vector is completely describes by

- the **joint probability mass function (joint pmf)** for discrete random vectors, and
- the **joint probability density function (joint pdf)** for continuous (or mixed) random vectors.

Event Shorthand

Example: $[X = 5,\ Y < 3] = \{s \in S : X(s) = 5 \text{ and } Y(s) < 3\}$, i.e., the "," denotes "and" or intersection:
$[X = 5,\ Y < 3] = [X = 5] \cap [Y < 3]$

Joint Probability Mass Function (Joint PMF) for discrete r.v.

Example: roll fair 6-sided die.

roll	1	2	3	4	5	6	(outcomes or sample points $s \in S$)
X	0	0	3	0	0	3	3 if number of dots is a multiple of 3, and 0 otherwise
Y	0	2	0	2	0	2	2 if number of dots is even, and 0 otherwise

$S_X = \{0, 3\}$, $S_Y = \{0, 2\}$
PMF Shorthand: $p(x_i, y_j) = P[X = x_i, Y = y_j]$
$p(0, 0) = P($ roll 1 or 5 $)$ = 2/6
$p(3, 0) = P($ roll 3 $)$ = 1/6
$p(0, 2) = P($ roll 2 or 4 $)$ = 2/6
$p(3, 2) = P($ roll 6 $)$ = 1/6
Note that $\sum_{x_i, y_j} p(x_i, y_j) = 1$
Can answer any question, e.g. $P[XY > 5] = \sum_{\{(x,y): x \in S_X, y \in S_Y, xy > 5\}} P[X = x, Y = y] = P[X = 3, Y = 2] = 1/6$

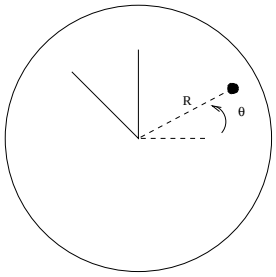
For a random vector with two components (a pair of random variables), the joint pmf can be displayed as a **2D Histogram**

Equivalent Event approach to calculating probabilities for random vectors

Example: uniformly-likely circular dart board
Sample space: $S = \{(a,b) : \ a^2 + b^2 \leq 30^2\}$
Two random variables defined on S:

- Distance from center: if $s = (a,b)$ then $R(s) = \sqrt{a^2 + b^2}$
- Angular position $\Theta(s) = \tan^{-1}(b/a)$

Random vector: $\underline{X} = (R, \Theta)$ i.e. $\underline{X}(s) = (R(s), \Theta(s))$
Induced sample space: $S_{\underline{X}} = [0, 30] \times [0, 360]$
Questions: what is $P[R \cos \Theta \leq 10]$ and $P([R \leq 20] \cap [45 \leq \Theta \leq 90])$?
Answer: find **equivalent event** (set of sample points that satisfy the conditions)



For 2nd question, area(wedge)/area(board) = $\pi 20^2/8/(\pi 30^2) = 1/18$

General formula:

$$[\underline{X} \in B] = \{s \in S : \underline{X}(s) \in B\} = \underline{X}^{-1}(B) \quad \text{so} \quad P[\underline{X} \in B] = P(A), \text{ where } A = \{s \in S : \underline{X}(s) \in B\}$$

The latter question is said to be in "product form"
In general an event of the form $A = [X_1 \in B_1] \cap [X_2 \in B_2] \cap \cdots \cap [X_n \in B_n]$ is in **product form**
Since $[X_1 \in B_1] \cap [X_2 \in B_2] \cap \cdots \cap [X_n \in B_n] = [\underline{X} \in B]$ where $B = B_1 \times B_2 \times \cdots \times B_n$

Any event of interest can be formed from (limits of) unions of events in product form. Therefore, if we know all the probabilities of the form

$$P([X_1 \le x_1] \cap [X_2 \le x_2] \cap \cdots \cap [X_n \le x_n]) = P[X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n]$$

we can compute any probability of interest
(although really we will use the above to get joint pdf and then integrate to get probabilities)

4.2

Joint Cumulative Probability Distribution Function (Joint CDF) (Joint Distribution)

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P[X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n] \text{ for } -\infty < x_i < \infty,\ i = 1, \ldots, n$$

Read: the probability of the event that the r.v. X_1 takes on a value in the interval $(-\infty, x_1]$ and the r.v. X_2 takes on a value in the interval $(-\infty, x_2]$ and ...

Properties of Joint CDF (for 2-vector (X, Y) only, generalization to n-vector is straightforward)

- $0 \le F_{X,Y}(x, y) \le 1,\ \forall x, y$
- $F_{X,Y}(\infty, \infty) = 1$
- $F_{X,Y}(-\infty, y) = F_{X,Y}(x, -\infty) = 0,\ \forall x, y$
- Monotone nondecreasing: $F_{X,Y}(x_1, y_1) \le F_{X,Y}(x_2, y_2)$ if $x_1 \le x_2$ and $y_1 \le y_2$
- $F_{X,Y}(x)$ is "right" continuous: $F_{X,Y}(x, y) = \lim_{\delta \searrow 0} F_{X,Y}(x + \delta, y) = \lim_{\delta \searrow 0} F_{X,Y}(x, y + \delta)$ (for $\delta > 0$)
- $P[x_1 < X \le x_2, Y \le y] = F_{X,Y}(x_2, y) - F_{X,Y}(x_1, y)$ (sketch)
- $P[X \le x,\ y_1 < Y \le y_2] = F_{X,Y}(x, y_2) - F_{X,Y}(x, y_1)$
- $P[x_1 < X \le x_2,\ y_1 < Y \le y_2] = F_{X,Y}(x_2, y_2) - F_{X,Y}(x_2, y_1) - F_{X,Y}(x_1, y_2) + F_{X,Y}(x_1, y_1)$
- $P[X = x, Y = y] = \lim_{\delta_x \to 0} \lim_{\delta_y \to 0} P[x - \delta_x < X \le x, y - \delta_y < Y \le y]$ = ...?

Example: circular dart board with R and Θ as defined above

$$F_{R,\Theta}(r, \theta) = P([R \le r,\ \Theta \le \theta]) = P([R \le r] \cap [\Theta \le \theta]) = \begin{cases} 0, & r < 0 \text{ or } \theta < 0 \\ (r/30)^2 \theta / 360, & 0 \le r \le 30, 0 \le \theta \le 360 \\ \frac{\theta}{360}, & r > 30, 0 \le \theta \le 360 \\ (r/30)^2, & 0 \le r \le 30, \theta > 360 \\ 1, & r > 30, \theta > 360 \end{cases}$$

Marginal CDF or Marginal Distribution
$F_X(x) = F_{X,Y}(x, \infty)$, $F_Y(y) = F_{X,Y}(y, \infty)$

$$F_{X,Y}(x, \infty) = P([X \le x] \cap [Y \le \infty]) = P([X \le x] \cap S) = P([X \le x]) = F_X(x)$$

In preceding example

$$F_\Theta(\theta) = F_{R,\Theta}(\infty, \theta) = \begin{cases} 0, & \theta < 0 \\ \frac{\theta}{360}, & 0 \le \theta \le 360 \\ 1, & \theta > 360 \end{cases}$$

which is Uniform(0,360): makes sense.

Proof of relation between probability and CDF

$$[x_1 < X \le x_2,\ y_1 < Y \le y_2] = ([x_1 < X \le x_2] \cap [Y \le y_2]) \cap [Y \le y_1]^c$$

since $P(A \cap B^c) = P(A) - P(A \cap B)$:

$$P[x_1 < X \le x_2,\ y_1 < Y \le y_2] = P[x_1 < X \le x_2,\ Y \le y_2] - P[x_1 < X \le x_2,\ Y \le y_1]$$

Applying same approach to X yields:

$$P[x_1 < X \le x_2,\ y_1 < Y \le y_2] = P[X \le x_2,\ Y \le y_2] - P[X \le x_1,\ Y \le y_2] - (P[X \le x_2,\ Y \le y_1] - P[X \le x_1,\ Y \le y_1])$$
$$= F_{X,Y}(x_2, y_2) - F_{X,Y}(x_1, y_2) - F_{X,Y}(x_2, y_1) + F_{X,Y}(x_1, y_1)$$

Discrete random vector

CDF:

$$F_{X,Y}(x, y) = \sum_{i,j} P[X = x_i, Y = y_j] u(x - x_i) u(y - y_j)$$

since

$$P[X \le x, Y \le y] = \sum_{\{i: x_i \le x\}} \sum_{\{j: y_j \le y\}} P[X = x_i, Y = y_j]$$

pdf:

$$f_{X,Y}(x, y) = \sum_{i,j} P[X = x_i, Y = y_j] \delta(x - x_i, y - y_j)$$

Computing probabilities via $P[x_1 < X \leq x_2,\ y_1 < Y \leq y_2] = F_{X,Y}(x_2, y_2) - F_{X,Y}(x_2, y_1) - F_{X,Y}(x_1, y_2) + F_{X,Y}(x_1, y_1)$ is inconvenient. Integrating a pdf would be easier

Joint Probability Density Function (pdf)
The joint probability density function (pdf) of a n-dimensional random vector $\underline{X}$ is defined to be

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \frac{d^n}{dx_1 dx_2 \cdots dx_n} F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$$

(details about nondifferentiability similar to 1D case)

Example: Joint pdf of R, Θ for circular dartboard:

$$f_{R,\Theta}(r, \theta) = \begin{cases} \frac{2r}{30^2} \frac{1}{360}, & 0 < r < 30, 0 < \theta < 360 \\ 0, & \text{otherwise} \end{cases}$$

Properties of joint pdf (for 2-vector only) (generalization straightforward)

- $f_{X,Y}(x, y) \geq 0 \ \forall x$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx\, dy = 1$
- $F_{X,Y}(x, y) = \int_{-\infty}^{x^+} \int_{-\infty}^{y^+} f_{X,Y}(x', y')\, dx'\, dy'$
- $F_X(x) = \int_{-\infty}^{x^+} \int_{-\infty}^{\infty} f_{X,Y}(x', y)\, dx'\, dy$ since $F_X(x) = F_{X,Y}(x, \infty)$
  $F_Y(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y^+} f_{X,Y}(x, y')\, dx\, dy'$
- $P[x_1 < X \leq x_2,\ y_1 < Y \leq y_2] = \int_{x_1}^{x_2^+} \int_{y_1}^{y_2^+} f_{X,Y}(x, y)\, dx\, dy$
- $P[(X, Y) \in B] = \iint_B f_{X,Y}(x, y)\, dx\, dy$

If 1st two hold, then $f_{X,Y}$ is valid joint pdf

Interpretation of joint pdf
$P[x - \delta_x < X \leq x,\ y - \delta_y < Y \leq y] \approx \delta_x \delta_y f_{X,Y}(x, y)$ for small δ's
Higher density = more likely

Example: $P[R \leq 20,\ \pi/4 \leq \Theta \leq \pi/2]$

$$= \iint_{\{(r,\theta): r\theta \leq 20\}} f_{R,\Theta}(r, \theta)\, dr d\theta = \int_0^{20} \int_{\pi/4}^{\pi/2} \frac{2r}{30^2} \frac{1}{2\pi}\, dr d\theta = \int_0^{20} \frac{2r}{30^2} \frac{\pi/2 - \pi/4}{2\pi}\, dr d\theta = \frac{1}{8} \frac{r^2}{30^2}\Big|_{r=0}^{20} = \frac{1}{18}$$

(same as before)

Marginal pdf

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} \int_{-\infty}^{x^+} \int_{-\infty}^{\infty} f_{X,Y}(x', y)\, dx'\, dy = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$$

Similarly

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx \quad \text{and in general } f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\, dx_1 \cdots dx_{i-1}\, dx_{i+1} \cdots dx_n$$

Example: Joint pdf of X, Y for circular dartboard: (we can finally formalize!)

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{\pi r^2}, & x^2 + y^2 \le r^2 \\ 0, & \text{otherwise} \end{cases} \quad \text{uniform over entire dartboard}$$

so $P[(X, Y) \in A] = \iint_A \frac{1}{\pi r^2}\, dxdy = \text{Area}(A)/(\pi r^2)$

$S_X = [-r, r]$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy = \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} \frac{1}{\pi r^2}\, dy = \begin{cases} \frac{2}{\pi r^2}\sqrt{r^2 - x^2}, & \text{for } |x| \le r \\ 0, & \text{o.w.} \end{cases}$$

(Picture)

Intuition: why peaked at center?

(skip)

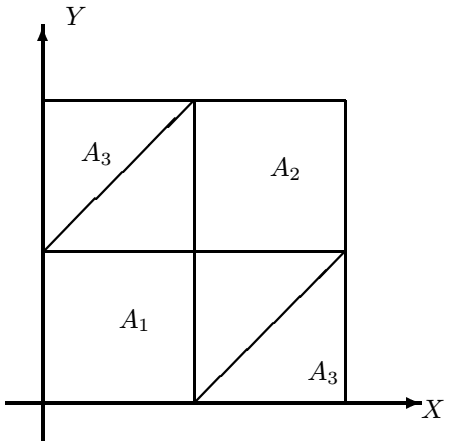
Example: a 1-meter stick breaks "randomly" into 3 pieces.
What is probability a triangle cannot be formed from 3 pieces?
Let X and Y denote break points. Assume

$$f_{X,Y}(x, y) = \begin{cases} 1, & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0, & \text{otherwise} \end{cases} \quad \text{(2d uniform pdf)}$$

$A = \{$ cannot form triangle $\}$ = $\{$ longest piece $> 1/2\} = A_1 \cup A_2 \cup A_3$
where $A_1 = [X < 1/2, Y < 1/2]$, $A_2 = [X > 1/2, Y > 1/2]$, $A_3 = [|X - Y| > 1/2]$

$$P(A_1) = \iint_{A_1} f_{X,Y}(x, y)\, dx\, dy = \text{Area}(A_1) = 1/4$$

$P(A) = 3/4$



Recall that events A and B are independent iff $P(A \cap B) = P(A)P(B)$
We want an analogous concept for random variables.

4.3

Independence of Random Variables

We say $X_1, \ldots, X_n$ are independent r.v.s iff any of the following equivalent *separability* conditions hold $\forall x_1, \ldots x_n \in \mathbb{R}$:

$$\begin{aligned}
P[X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n] &= P[X_1 \le x_1] P[X_2 \le x_2] \cdots P[X_n \le x_n] \\
F_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) &= F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n) \\
f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) &= f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n) \\
P[X_1 \in B_1, \ \ldots, X_n \in B_n] &= P[X_1 \in B_1] \cdots P[X_n \in B_n]
\end{aligned}$$

For discrete r.v., joint PMF must factor into product of marginal PMFs:

$$P[X_1 = x_1, \ldots, X_n = x_n] = P[X_1 = x_1] \cdots P[X_n = x_n]$$

Example: Joint pdf of R, Θ for circular dartboard:

$$f_{R,\Theta}(r, \theta) = \begin{cases} \frac{2r}{30^2} \frac{1}{360}, & 0 < r < 30, \ 0 < \theta < 360 \\ 0, & \text{otherwise} \end{cases}$$

$$f_R(r) = \begin{cases} \frac{2r}{30^2}, & 0 < r < 30 \\ 0, & \text{otherwise} \end{cases} \qquad f_\Theta(\theta) = \begin{cases} \frac{1}{360}, & 0 < \theta < 360 \\ 0, & \text{otherwise} \end{cases}$$

So R and Θ are independent for *circular* dartboard. (But X and Y are dependent!)
(In contrast, for *square* dartboard X and Y are independent, but R and Θ are dependent!)

Property of Independent R.V.s

If $X_1, \ldots, X_n$ are independent r.v.s, then functions of disjoint subsets of the X_i's are independent.
E.g., if $Y_1 = X_1 + X_2, \ Y_2 = X_7 e^{X_4}, \ Y_3 = \log X_5$ then Y_1, Y_2 and Y_3 are independent r.v.s.

Independence arises two ways. In some problems (such as dart board above) we know how the r.v. is defined (from sample space) and we form joint CDF and then check to see if independent or not. In other problems we are given that some of the r.v.s are independent (for example if they come from independent sub-experiments), and we use that fact to calculate other things of interest.

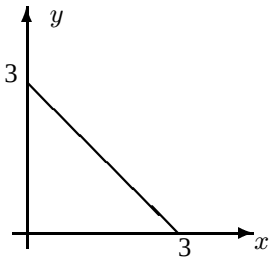
Example: a satellite has two power supplies, a primary and a backup that switches on if the primary fails.
The failure time X of primary (relative to launch time) is exponential with mean 1 year.
The failure time Y of backup (relative to engaging) is exponential with mean 1 year.
Assume lifetime of two supplies is independent.
What is prob. that satellite still works after 3 years?
Translate: {satellite still works after 3 years} = $[X + Y > 3]$

$$P[X + Y \geq 3] = \iint_{\{(x,y):x+y\geq 3\}} f_{X,Y}(x,y)\,dx\,dy = 1 - \iint_{\{(x,y):x+y<3\}} f_X(x) f_Y(y)\,dx\,dy$$

$$= 1 - \int_0^3 \int_0^{3-x} e^{-x} e^{-y}\,dx\,dy = 1 - \int_0^3 e^{-x}(1 - e^{-(3-x)})\,dx = 1 - \int_0^3 e^{-x} - e^{-3}\,dx = 4e^{-3}$$

Two aspects of region to consider:

- where is $f_{X,Y}$ nonzero,
- what subset of that is the event of interest?



also: Buffon's needle

4.4 Conditional Probability and Conditional Expectation

CDF $F_X(x) = P[X \leq x]$	$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = P[X_1 \leq x_1,\ldots,X_n \leq x_n]$
pdf $f_X(x) = \frac{d}{dx} F_X(x)$	$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \frac{d^n}{dx_1 \cdots dx_n} F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$

Conditioning on Events

Conditional CDF $F_X(x\|A) = P[X \leq x\|A]$	
Conditional pdf (given an event) $f_X(x\|A) = \frac{d}{dx} F_X(x\|A)$	
Total Probability for CDF, if $\{A_i\}$ partition S $F_X(x) = \sum_i F_X(x\|A_i) P(A_i)$	
Total Probability for pdf, if $\{A_i\}$ partition S $f_X(x) = \sum_i f_X(x\|A_i) P(A_i)$	
Total Expectation, if $\{A_i\}$ partition S $E[g(X)] = \sum_i E[g(X)\|A_i] P(A_i)$	

Point Conditioning and Events

Bayes rule for point conditioning $P(A\|X = x) = f_X(x\|A) P(A) / f_X(x)$	
Total Probability for event from pdf $P(A) = \int_{-\infty}^{\infty} P(A\|X = x) f_X(x)\, dx$	

Two random variables (NEW)

Bayes rule for two r.v.s	$f_{Y\|X}(y\|x) = f_{X\|Y}(x\|y) f_Y(y) / f_X(x)$
Total probability for two r.v.s	$f_Y(y) = \int_{-\infty}^{\infty} f_{Y\|X}(y\|x) f_X(x)\, dx = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx$
Law of iterated expectation for two r.v.s	$E[Y] = E[E[Y\|X]] = \int_{-\infty}^{\infty} E[Y\|X = x] f_X(x)\, dx$

4.4

Interval conditioning

$$F_Y(y \mid x_1 < X \le x_2) = P[Y \le y \mid x_1 < X \le x_2] = \frac{P[Y \le y,\ x_1 < X \le x_2]}{P[x_1 < X \le x_2]}$$

$$= \frac{\int_{-\infty}^{y} \int_{x_1}^{x_2} f_{X,Y}(x,y')\,dxdy'}{\int_{-\infty}^{\infty} \int_{x_1}^{x_2} f_{X,Y}(x,y')\,dxdy'} = \frac{\int_{-\infty}^{y} \int_{x_1}^{x_2} f_{X,Y}(x,y')\,dxdy'}{\int_{x_1}^{x_2} f_X(x)\,dx}$$

Define the conditional pdf

$$f_Y(y \mid x_1 < X \le x_2) = \frac{d}{dy} F_Y(y \mid x_1 < X \le x_2) = \frac{d}{dy} \frac{\int_{-\infty}^{y} \int_{x_1}^{x_2} f_{X,Y}(x,y')\,dxdy'}{\int_{x_1}^{x_2} f_X(x)\,dx} = \frac{\int_{x_1}^{x_2} f_{X,Y}(x,y)\,dx}{\int_{x_1}^{x_2} f_X(x)\,dx}$$

Note that naturally as $x_1 \to -\infty$ and $x_2 \to \infty$, $f_Y(y \mid x_1 < X \le x_2) \to f_Y(y)$.

Point conditioning and conditional pdfs

Recall that earlier we defined point-conditioning probabilities by

$$P(A|X=x) = \lim_{\delta \to 0} P(A \mid x - \delta < X \le x) = \frac{f_X(x|A)P(A)}{f_X(x)}$$

Now let A be the event $[Y \le y]$, then

$$F_{Y|X}(y|x) = P[Y \le y \mid X = x] = \lim_{\delta \to 0} P[Y \le y \mid x - \delta < X \le x] = \lim_{\delta \to 0} \frac{P[Y \le y, x - \delta < X \le x]}{P[x - \delta < X \le x]}$$

$$= \lim_{\delta \to 0} \frac{F_{X,Y}(x,y) - F_{X,Y}(x-\delta,y)}{F_X(x) - F_X(x-\delta)} = \lim_{\delta \to 0} \frac{\frac{1}{\delta}\left(F_{X,Y}(x,y) - F_{X,Y}(x-\delta,y)\right)}{\frac{1}{\delta}\left(F_X(x) - F_X(x-\delta)\right)} = \frac{\frac{d}{dx} F_{X,Y}(x,y)}{f_X(x)}$$

Define the conditional pdf

$$f_{Y|X}(y|x) = \frac{d}{dy} F_{Y|X}(y|x) = \frac{d}{dy} \frac{\frac{d}{dx} F_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{f_X(x)} \text{ so } \boxed{f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}}$$

Also useful is the rearrangement:

$$\boxed{f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x)}$$

Similarly

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

So combining yields **Bayes rule** for pdfs:

$$\boxed{f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)}}$$

"Total probability" for pdfs:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x)\,dx \text{ so } \boxed{f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x)\,dx}$$

Calculating probability from condition pdf:

$$P[Y \in B | X = x] = \int_B f_{Y|X}(y|x)\,dy$$

Conditional pdfs in general

$$f_{X_1,X_2|X_3,X_4,X_5}(x_1,x_2 \mid x_3,x_4,x_5) = \frac{f_{X_1,X_2,X_3,X_4,X_5}(x_1,x_2,x_3,x_4,x_5)}{f_{X_3,X_4,X_5}(x_3,x_4,x_5)}$$

etc.

Chain rule for pdfs

$$f(x_1,\ldots,x_n) = f(x_n|x_1,\ldots,x_{n-1})f(x_{n-1}|x_1,\ldots,x_{n-2})\cdots f(x_2|x_1)f(x_1)$$

(with lazy notation omitting subscripts)

Example (4.23 reworded)
Packets arriving at a router.
The number that arrive during any time interval $[t_1, t_2]$ is a Poisson r.v. with mean $\beta(t_2 - t_1)$
Assume that the amount of time T needed to process the packet is a r.v. with exponential distribution with mean $1/\lambda$.
(random due to routing tables, etc.)
Router has 8 packet buffer. What is P(overflow), i.e., 9 or more new packets arrive while processing "the first" packet.
Need $P[N \geq 9]$, where N is the number of (new) packets that arrive while "the first" packet is being processed.
Find the PMF of N.
N is discrete r.v. with $S_N = \{0, 1, 2, \ldots\}$
Let T be the time required to process specified packet. We are given:

$$f_T(t) = \lambda e^{-\lambda t} u(t).$$

Also:

$$P[N = k|T = t] = \frac{(\beta t)^k}{k!} e^{-\beta t}.$$

By "total prob" (eqn 4.35)

$$P[N = k] = \int_{-\infty}^{\infty} P[N = k|T = t] f_T(t)\, dt = \int_0^{\infty} \frac{(\beta t)^k}{k!} e^{-\beta t} \lambda e^{-\lambda t}\, dt$$

$$= \frac{\beta^k \lambda}{k!} \int_0^{\infty} t^k e^{-(\beta+\lambda)t}\, dt = \frac{\beta^k \lambda}{(\beta+\lambda)^{k+1}} = \left(\frac{\lambda}{\beta+\lambda}\right)\left(\frac{\beta}{\beta+\lambda}\right)^k$$

since $\int_0^{\infty} x^k e^{-ax}\, dx = k!/a^{k+1}$, so

$$P[N \geq k] = \sum_{j=k}^{\infty} \frac{\lambda}{\beta+\lambda} \left(\frac{\beta}{\beta+\lambda}\right)^j = \left(\frac{\beta}{\beta+\lambda}\right)^k$$

Note: large β means overflow, unless λ large enough to keep up.
Specific application: suppose $\beta = 2/\mu$sec (on average 2 new packets per microsecond)
How large must λ be to ensure that only 0.1% of the time will more then 8 new packets arrive while during processing?
Want $P[N \geq 9] \leq 0.001 = p$
$(\frac{\beta}{\beta+\lambda})^k \leq p$ so $\lambda \geq \beta(1/\sqrt[k]{p} - 1) = 2(1/\sqrt[9]{0.001} - 1) \approx 2.31$ per μsec.
Need $\mu \leq 1/2.31 = 0.43\mu$sec / packet.

Naive design: use $\mu = 0.5\mu$sec so that $\lambda = 2/\mu$sec.
Then $P[N \geq 9] = (1/2)^9 = 0.002$ or 0.2% of the time there will be overflow.

Independence and conditional pdfs
If X and Y are independent r.v.s, then $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ so

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y) \quad \text{equivalently} \quad F_Y(y|X \in B) = F_Y(y) \; \forall B \in \mathcal{B}$$

Similarly

$$f_{X|Y}(x|y) = f_X(x)$$

this gives us two alternate conditions for testing independence

Conditional Expectation

$$E[Y|X=x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) \, dy$$

Note that $E[Y|X=x]$ is a function that maps any value x into a number.
Call this function $g(x) = E[Y|X=x]$ Then $E[Y|X] = g(X)$ is a random variable

Independence and expectation
If X and Y are independent r.v.s, then

$$E[Y|X=x] = E[Y] \quad \text{and} \quad E[X|Y] = E[X]$$

However, the converse is *not* true
Example: circular dart board with

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi r^2}, & x^2 + y^2 \le r^2 \\ 0, & \text{otherwise} \end{cases}$$

then deriving the conditional pdf $f_{Y|X}(y|x)$ is symmetric about 0 so $E[Y|X=x] = 0$. Also the marginal pdf $f_Y(y)$ is symmetric about 0 so $E[Y] = 0$. Thus $E[Y|X] = E[Y] = 0$. But X and Y are not independent.

Law of Iterated Expectation

$$E[E[Y|X]] = E[Y]$$

$$E[E[Y|X]] = \int_{-\infty}^{\infty} E[Y|X=x] f_X(x) \, dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X}(y|x) \, dy \; f_X(x) \, dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x,y) \, dx dy = \int_{-\infty}^{\infty} y f_Y(y) \, dy = E[Y]$$

Example: mean of packets arriving (earlier)
Intuition: $E[N] = \beta E[T] = \beta/\lambda$

$$E[N] = E[E[N|T]] = \int_0^{\infty} E[N|T=t] f_T(t) \, dt = \int_0^{\infty} (\beta t) f_T(t) \, dt = \beta E[T] = \beta/\lambda$$

4.6

Functions or transformations of a random vectors
Let $X_1, \ldots, X_n$ be r.v.s whose joint pdf (perhaps via jcdf) is known.
Let $Z = g(X_1, \ldots, X_n)$ where $g : \mathbb{R}^n \to \mathbb{R}$. Want to find pdf (via cdf) of Z in terms of joint pdf of X_i's.

Method of Equivalent Events
Let

$$R_z = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : g(x_1, \ldots, x_n) \le z\} = g^{-1}((-\infty, z])$$

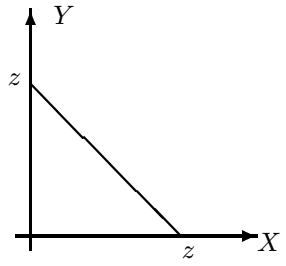
then $[Z \le z]$ is equivalent to the event $[\underline{X} \in R_z]$, so

$$F_Z(z) = P[Z \le z] = P[\underline{X} \in R_z] = \int \int \cdots \int_{R_z} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\, dx_1 \, \cdots \, dx_n$$

Not really any simple plug-and-chug formula! (except Jacobian)

Typical problems for pairs of r.v.s: $Z = g(X, Y)$.
$Z = X + Y$, $Z = X \cdot Y$, $Z = X/Y$, $Z = \max(X, Y)$, $Z = \min(X, Y)$, $Z = \sqrt{X^2 + Y^2}$, $Z = |X - Y|$, $X = R \cos \Theta$, ...

Example: $Z = X + Y$
Equivalent event is easy in this case: $[Z \le z] = [X + Y \le z] = [Y \le z - X] = [(X, Y) \in R_z]$ where $R_z = \{(x, y) \in \mathbb{R}^2 : x + y \le z\}$



$$F_Z(z) = P[Z \le z] = P[\underline{X} \in R_z] = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_{X,Y}(x, y)\, dy\, dx$$

gives cdf of Z in terms of joint pdf of X and Y

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z - x)\, dx$$

gives pdf of Z in terms of joint pdf of X and Y

Sums of Independent r.v.s If X and Y are independent r.v.s, then

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx = (f_X \star f_Y)(z)$$

which is a convolution integral. More generally, if $X_1, \ldots, X_n$ are independent r.v.s and $Z = X_1 + \cdots + X_n$, then

$$f_Z(z) = (f_{X_1} \star f_{X_2} \star \cdots \star f_{X_n})(z)$$

(well defined because convolution is a commutative and associative operator)

Convolution

$$f(x) = (g \star h)(x) = (h \star g)(x) = \int_{-\infty}^{\infty} g(t) h(x - t)\, dt = \int_{-\infty}^{\infty} h(t) g(x - t)\, dt$$

Example: sum of 2 independent exponential r.v.s is Erlang
Assume X and Y are independent and have exponential distributions with same mean μ
$f_X(x) = e^{-x/\mu}/\mu u(x)$
Let $Z = X + Y$ and find pdf of Z
Since range of X and Y is nonnegative, $S_Z = [0, \infty)$ so only need to consider $z \geq 0$

$$f_Z(z) = (f_X \star f_Y)(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx = \int_0^z \frac{1}{\mu} e^{-x/\mu} \frac{1}{\mu} e^{-(z-x)/\mu}\, dx$$

$$= \int_0^z \frac{1}{\mu^2} e^{-z/\mu}\, dx = \frac{z}{\mu^2} e^{-z/\mu} u(z)$$

so Z has an Erlang distribution.

Sums of independent r.v.s (can be shown using above convolution approach)
- $N(\mu_1, \sigma_1^2) + N(\mu_2, \sigma_2^2) \equiv N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
- $\text{Poisson}\{\lambda_1\} + \text{Poisson}\{\lambda_2\} \equiv \text{Poisson}\{\lambda_1 + \lambda_2\}$
- Binomial(n_1, p) + Binomial$(n_2, p) \equiv$ Binomial$(n_1 + n_2, p)$

Example: method of equivalent events
$Y = \max(X_1, X_2, X_3)$ where X's are assumed independent. Find pdf of Y

$$F_Y(y) = P[Y \leq y] = P[\max(X_1, X_2, X_3) \leq y] = P[X_1 \leq y, X_2 \leq y, X_3 \leq y] = F_{X_1}(y) F_{X_2}(y) F_{X_3}(y)$$

or in general:

$$= \int_{-\infty}^{y} \int_{-\infty}^{y} \int_{-\infty}^{y} f_{X_1, X_2, X_3}(x_1, x_2, x_3)\, dx_1\, dx_2\, dx_3$$

Thus by chain rule (for independent X's case):

$$f_Y(y) = \frac{d}{dy} F_Y(y) = f_{X_1}(y) F_{X_2}(y) F_{X_3}(y) + F_{X_1}(y) f_{X_2}(y) F_{X_3}(y) + F_{X_1}(y) F_{X_2}(y) f_{X_3}(y)$$

For this case the equivalent event could be found without picture. Usually a picture is needed.

4.7

Expectation, Mean, Average, or Expected Value of function of vector random variables

For function of 1 r.v.:

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

For function of multiple r.v.s:

$$E[g(X_1, \ldots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) \, dx_1 \, \cdots \, dx_n$$

provided

$$E[|g(X_1, \ldots, X_n)|] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |g(x_1, \ldots, x_n)| \, f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) \, dx_1 \, \cdots \, dx_n < \infty$$

(otherwise we say it is *undefined* or *does not exist*)

(hard way would be to let $Y = g(X_1, \ldots, X_n)$, find pdf of Y, then use $E[Y] = \int_{-\infty}^{\infty} y f_Y(y) \, dy$

Discrete random vector

$$E[g(X_1, \ldots, X_n)] = \sum_{x_1 \in S_{X_1}} \cdots \sum_{x_n \in S_{X_n}} g(x_1, \ldots, x_n) P[X_1 = x_1, \ldots, X_n = x_n]$$

Linearity Properties

$$E[X+Y] = \iint (x+y) f_{X,Y}(x,y) \, dx \, dy = \iint x f_{X,Y}(x,y) \, dx \, dy + \iint y f_{X,Y}(x,y) \, dx \, dy = \int x f_X(x) \, dx + \int y f_Y(y) \, dy$$

So $E[X+Y] = E[X] + E[Y]$. By similar argument more generally:

$$E[\sum_j g_j(X_1, \ldots, X_n)] = \sum_j E[g_j(X_1, \ldots, X_n)]$$

so we can exchange expectation and summation.
Note that linearity of expectation does not require independence!

Example (mean of Binomial(n,p))

Let $X_1, \ldots, X_n$ be independent Bernoulli random variables (1 if success, 0 if failure) with success probability p
Since X_i is a discrete r.v.: $E[X_i] = 1p + 0(1-p) = p$
Let $Y = X_1 + \cdots + X_n$ be the number of successes in n trials, so Y has Binomial(n, p) distribution
$E[Y] = \sum_{i=1}^n E[X_i] = np$, which is much easier than earlier approach of $E[Y] = \sum_{k=0}^n k P[Y = k]$

Conditional Expectation

$$E[g(X_1, \ldots, X_n)|A] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f_{X_1, \ldots, X_n}(x_1, \ldots, x_n|A) \, dx_1 \, \cdots \, dx_n$$

If $A_1, A_2, \ldots$ partition S:

$$E[g(X_1, \ldots, X_n)] = \sum_i E[g(X_1, \ldots, X_n)|A_i] P(A_i) \quad \text{(“total expectation”)}$$

Example (expectation of sum of dependent r.v.s)
Suppose n concertgoers check their hats at a coatroom. The coat checkers randomly rearrange all of the hats before returning them to the concertgoers. What is the expected number of concertgoers that get their own hat back? (Y)
Hard way: find PMF of Y, and then sum: $E[Y] = \sum_{k=0}^{n} kP[Y = k]$.
Easy way: let X_i be 1 if the ith concertgoer gets his or her own hat back, and 0 otherwise. Since $P[X_i = 1] = 1/n$,

$$E[X_i] = 1P[X_i = 1] + 0P[X_i = 0] = 1\frac{1}{n} + 0 = \frac{1}{n}$$

Since $Y = \sum_{i=1}^{n} X_i$

$$E[Y] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{1}{n} = 1 \quad \text{independent of } n \text{ !}$$

Independence
If $X_1, \ldots, X_n$ are independent r.v.s, then

$$E[g_1(X_1)g_2(X_2)\cdots g_n(X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g_1(x_1)\cdots g_n(x_n) f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\, dx_1 \, \cdots \, dx_n$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g_1(x_1)\cdots g_n(x_n) f_{X_1}(x_1)\cdots, f_{X_n}(x_n)\, dx_1 \, \cdots \, dx_n = \int_{-\infty}^{\infty} g_1(x_1) f_{X_1}(x_1)\, dx_1 \cdots \int_{-\infty}^{\infty} g_n(x_n) f_{X_n}(x_n)\, dx_n$$

$$= E[g_1(X_1)]E[g_2(X_2)]\cdots E[g_n(X_n)]$$

do not apply this formula to dependent r.v.s!

(4.7)

Moments

Marginal moments are same as before (mean and variance of a r.v.)

Now what is interesting is the moments that relating to the coupling between two r.v.s

- The **correlation** between r.v.s X and Y is defined to be $E[XY]$
- If $E[XY] = 0$, then we say X and Y are **orthogonal**
- The **covariance** between X and Y is defined to be $\mathrm{Cov}\{X, Y\} = E[(X - \mu_X)(Y - \mu_Y)]$
- The **correlation coefficient** of X and Y is $\rho_{X,Y} = \dfrac{\mathrm{Cov}\{X, Y\}}{\sigma_X \sigma_Y}$
- If $\rho_{X,Y} = 0$, then we say X and Y are **uncorrelated**
- Note that $\rho_{X,Y} = 0$ is essentially equivalent to $\mathrm{Cov}\{X, Y\} = 0$. (caution: potentially confusing terminology)

Properties of Covariance

- $\mathrm{Cov}\{X, Y\} = E[XY] - \mu_X \mu_Y$
- $\mathrm{Cov}\{X, Y\} = \mathrm{Cov}\{Y, X\}$
- $\mathrm{Cov}\{X, X\} = \mathrm{Var}[X]$
- $\mathrm{Cov}\{aX + b, Y\} = a\,\mathrm{Cov}\{X, Y\}$
- $|\,\mathrm{Cov}\{X, Y\}\,| \le \sigma_X \sigma_Y$ (called the **Schwarz Inequality**)
- Thus $|\rho_{X,Y}| \le 1$
- If X and Y are independent r.v.s, then $E[XY] = \mu_X \mu_Y$ so $\mathrm{Cov}\{X, Y\} = \rho_{X,Y} = 0$.
- The reverse is *not true* in general (uncorrelated does not ensure independence); an exception is Gaussian.
- $\mathrm{Cov}\left\{\sum_i X_i, \sum_j Y_j\right\} = \sum_i \sum_j \mathrm{Cov}\{X_i, Y_j\}$

Correlation and Linearity

Suppose $Y = aX + b$. Find $\rho_{X,Y}$. Note that $\sigma_Y^2 = a^2 \sigma_X^2$ and $\mu_Y = a\mu_X + b$, so $Y - \mu_Y = aX + b - (a\mu_X + b) = a(X - \mu_X)$

$$\rho_{X,Y} = \frac{\mathrm{Cov}\{X, Y\}}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)a(X - \mu_X)]}{\sigma_X |a| \sigma_X} = \frac{a\sigma_X^2}{\sigma_X |a| \sigma_X} = \frac{a}{|a|} = \begin{cases} 1, & a > 0 \\ -1, & a < 0 \end{cases}$$

Hölder Inequality (Proof uses Jensen inequality)

$$\text{If } 1/p + 1/q = 1 \text{ for } p > 0 \text{ and } q > 0, \text{ then: } |E[XY]| \le E[|XY|] \le (E[|X|^p])^{1/p} (E[|Y|^q])^{1/q}$$

Variance of Sum of two r.v.s

If $Z = X + Y$ then $E[Z] = E[X] + E[Y]$ by linearity

$$\mathrm{Var}[Z] = E[(Z - \mu_Z)^2] = E[(X + Y - (E[X] + E[Y]))^2] = E[(X - E[X] + Y - E[Y])^2]$$

$$= E[(X - E[X])^2] + 2E[(X - E[X])(Y - E[Y])^2] + E[(Y - E[Y])^2] = \mathrm{Var}[X] + 2\,\mathrm{Cov}\{X, Y\} + \mathrm{Var}[Y]$$

note pattern: express new things concisely in terms of old

4.8

Bivariate Gaussian r.v.s (important for signal + noise models)
We say X and Y are jointly Gaussian r.v.s iff their joint pdf has the following form:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y} \exp\left(\frac{-1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X} \right)^2 - 2\rho \left(\frac{x-\mu_X}{\sigma_X} \right) \left(\frac{y-\mu_Y}{\sigma_Y} \right) + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 \right] \right)$$

- Marginal pdfs of X and Y are Gaussian: $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$
- ρ in jpdf is indeed the correlation coefficient, and $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ are means and variances
- Entire jpdf is specified by only the two means, variances, and ρ
- If $\rho = 0$ (uncorrelated) then X and Y are independent (not true in general!)
- Linear transform of Gaussian still Gaussian. If $Z = aX + bY$ then $Z \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y)$
- Conditional pdf of X given $Y = y$ is Gaussian: $N(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}y,\ \sigma_X^2(1-\rho^2))$

Practical reasons for using Gaussian: central limit theorem, and only first and second order moments needed.

Gaussian random vectors

$$\underline{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \quad \underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \mu_{\underline{X}} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{bmatrix}$$

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{|C_{\underline{X}}|}} \exp\left(-\frac{1}{2}(\underline{x} - \mu_{\underline{X}})^T C_{\underline{X}}^{-1} (\underline{x} - \mu_{\underline{X}}) \right)$$

where $|C_{\underline{X}}|$ denotes matrix determinant, and $C_{\underline{X}}$ is the $n \times n$ **covariance matrix** of $\underline{X}$, where

$$[C_{\underline{X}}]_{ij} = \mathrm{Cov}\{X_i, X_j\}$$

In $n = 2$ case,

$$C_{\underline{X}} = \begin{bmatrix} \mathrm{Cov}\{X_1, X_1\} & \mathrm{Cov}\{X_1, X_2\} \\ \mathrm{Cov}\{X_2, X_1\} & \mathrm{Cov}\{X_2, X_2\} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

so $|C_{\underline{X}}| = (1-\rho^2)\sigma_1^2\sigma_2^2$ and

$$C_{\underline{X}}^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} \sigma_1^{-2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \sigma_2^{-2} \end{bmatrix} \quad \text{so} \quad \underline{x}^T C_{\underline{X}}^{-1} \underline{x} = \frac{1}{1-\rho^2} \left[\left(\frac{x_1-\mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1} \right) \left(\frac{x_2-\mu_2}{\sigma_2} \right) + \left(\frac{x_2-\mu_2}{\sigma_2} \right)^2 \right]$$

Thus the bivariate form is indeed the $n = 2$ special case of the general form

---

Minimum mean squared error
The constant c that minimizes this mean squared error: $E[(g(X_1, \ldots, X_n) - c)^2]$ is $c = E[g(X_1, \ldots, X_n)]$

5.1

Characteristic Function for sums of independent r.v.s

If $Z = X_1 + \cdots + X_n$ where the X_i's are independent then

$$\Phi_Z(\omega) = E[e^{j\omega Z}] = E[e^{j\omega(X_1+\cdots+X_n)}] = E[e^{j\omega X_1} \cdot \cdots \cdot e^{j\omega+X_n}] = E[e^{j\omega X_1}] \cdot \cdots \cdot E[e^{j\omega+X_n}] = \Phi_{X_1}(\omega) \cdot \cdots \cdot \Phi_{X_n}(\omega)$$

Thus char. fun. for Z is product of char. fun.'s of X_i's
Compute Φ for each r.v., multiply, then Fourier transform to get pdf of Z
Easier then n-fold convolution
For large n can still be a pain - central limit theorem can help

---

Example: Sum of (independent) Gaussians is Gaussian

$$\text{If } X \sim N(\mu, \sigma^2) \text{ then } \Phi_X(\omega) = e^{j\omega\mu - \omega^2\sigma^2/2}$$

Define $Y = \sum_{i=1}^n X_i$ where $X_i \sim N(\mu_i, \sigma_i^2)$ and X_i's are independent.
We know the 1st and 2nd moments of Y: $\mu_Y = \sum_{i=1}^n \mu_i,\ \sigma_Y^2 = \sum_{i=1}^n \sigma_i^2$ (to be shown shortly)
For the entire pdf of Y:

$$\Phi_Y(\omega) = \prod_{i=1}^n \Phi_{X_i}(\omega) = \prod_{i=1}^n e^{j\omega\mu_i - \omega^2\sigma_i^2/2} = \exp[\sum_{i=1}^n (j\omega\mu_i - \omega^2\sigma_i^2/2)] = e^{j\omega\mu_Y - \omega^2\sigma_Y^2/2}$$

Thus $Y \sim N(\mu_Y, \sigma_Y^2)$, i.e. sum of (independent) Gaussians is still Gaussian! (Even with X_i's having different moments.)
Similarly, sum of i.i.d. exponentially distributed r.v.s has Erlang dist'n
Unfortunately, in general the pdf of the sum of r.v.s may be intractable. Often it is sufficient just to look at the moments.

Sum of Independent (actually just uncorrelated) r.v.s

Let $Y = X_1 + \cdots + X_n$ where X_i and X_j are uncorrelated for $i \neq j$
Recall $E[Y] = \sum_i E[X_i]$ by linearity so

$$Y - \mu_Y = \sum_i X_i - \sum_i E[X_i] = \sum_i (X_i - E[X_i])$$

$$\text{Var}[Y] = E[(Y-\mu_Y)^2] = E[\left(\sum_i (X_i - E[X_i])\right)^2] = \sum_i E[(X_i - E[X_i])^2] + \sum_{i,i\neq j} E[(X_i - E[X_i])(X_i - E[X_j])]$$

Thus in general

$$\text{Var}[\sum_i X_i] = \sum_i \text{Var}[X_i] + \sum_{i\neq j} \text{Cov}\{X_i, X_j\}$$

Easier derivation:

$$\text{Var}[\sum_i X_i] = \text{Cov}\left\{\sum_i X_i, \sum_i X_i\right\} = \sum_i \sum_j \text{Cov}\{X_i, X_j\} = \sum_i \text{Var}[X_i] + \sum_{i\neq j} \text{Cov}\{X_i, X_j\}$$

And in particular if the r.v.s are independent (or simply uncorrelated) then $\text{Cov}\{X_i, X_j\} = 0$ for $i \neq j$ so

$$\text{Var}[\sum_i X_i] = \sum_i \text{Var}[X_i]$$

Special case:

$$\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2\,\text{Cov}\{X, Y\}$$

Example: easier derivation of variance of Binomial

Let X_i be the r.v. that takes 1 or 0 for success or failure in the ith of n Bernoulli trials.
Let $Y = \sum_{i=1}^n X_i$, so that Y has a Binomial PMF.
Since the X_i's are independent, the variance of Y is the sum of the variances of the X_i's.
If $P[X_i = 1] = p$ is the success probability, then $E[X_i] = p$ and $E[X_i^2] = p$.
So $\text{Var}[X_i] = p - p^2 = p(1-p) = pq$. Thus $\text{Var}[Y] = npq$.

5.2

Sample Mean

Often we collect repeated measurements of some random phenomena under (essentially) identical conditions.
We say $X_1, X_2, \ldots$ are **independent and identically distributed (i.i.d.)** r.v.s iff

- the X_i's are independent
- $f_{X_i}(x) = f_X(x)\ \forall x\ \forall i$ (same marginal pdf) and hence same moments: $E[X_i] = \mu_X, \text{Var}[X_i] = \sigma_X^2,\ \forall i$

The sample mean is the *average* of the X_i's:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Moments:

$$E[\hat{\mu}] = E[\frac{1}{n} \sum_{i=1}^{n} X_i] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} \sum_{i=1}^{n} \mu_X = \mu_X$$

$$\text{Var}[\hat{\mu}] = \text{Var}[\frac{1}{n} \sum_{i=1}^{n} X_i] = \frac{1}{n^2} \text{Var}[\sum_{i=1}^{n} X_i] = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}[X_i] = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_X^2 = \frac{\sigma_X^2}{n}$$

Weak Law of Large Numbers

Suppose $X_1, X_2, \ldots$ i.i.d. r.v.s with finite mean μ_X and finite variance σ_X^2. Then

$$P\left[\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu_X \right| < \epsilon \right] \to 1 \text{ as } n \to \infty \text{ for any } \epsilon > 0$$

Proof:

$$P\left[\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu_X \right| \geq \epsilon \right] \leq \frac{\text{Var}[\frac{1}{n} \sum_{i=1}^{n} X_i]}{\epsilon^2} = \frac{\sigma_X^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty \text{ for any } \epsilon > 0$$

by Chebyshev inequality.

- Do not really need i.i.d., only independence and r.v.s with the same mean μ_X and variance
- Can relax assumption of finite variance!

Strong Law of Large Numbers

Suppose $X_1, X_2, \ldots$ are i.i.d. r.v.s with finite mean μ_X and finite variance σ_X^2. Then

$$P\left[\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu_X \right] = 1$$

Originally we said that to be meaningful and useful, probabilities of events should have properties similar to relative frequencies. Now we have the above results that confirm that the theory derived under the axioms of probability predict that large-sample averages will be close to the underlying mean.

5.3

Central Limit Theorem

The random phenomena in many engineering problems is the aggregate of a multitude of small contributions,
Such as electrons in resistors, magnetic domains on magnetic tape, photons, etc.
The distribution (cdf) of the sum of many i.i.d. r.v.s is approximately Gaussian

Suppose $X_1, X_2, \ldots$ are i.i.d. r.v.s with finite mean μ_X and finite variance σ_X^2. Then

$$P\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i-\mu_X}{\sigma_X} \le z\right] \to 1-Q(z) = \int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}\,dt \text{ as } n\to\infty \text{ for any } z\in\mathbb{R}$$

Utility: approximation for finite n.
Most accurate for z near 0, hence "central"

Example

Toss coin $n = 10000$ times. If 4884 heads, do you think it is a fair coin?
Let $X_i = 1$ if head on ith toss, 0 o.w.
$E[X_i] = p = 1/2$
Var$[X_i] = p(1-p) = 1/4$

$$P[\sum_{i=1}^{10000} X_i \le 4884] = P[\sum_{i=1}^{10000}(X_i - 1/2) \le 4884 - 5000] = P[\sum_{i=1}^{10000}\frac{X_i-1/2}{\sqrt{10000}\sqrt{1/4}} \le \frac{4884-5000}{50}]$$

$$= P[\sum_{i=1}^{10000}\frac{X_i-1/2}{\sqrt{10000}\sqrt{1/4}} \le -2.32] \approx 1 - Q(-2.32) = Q(2.32) = 0.01$$

so not very likely to be a fair coin!

Proof (see Ross)

Example: how large should n be?

Computer program is supposed to generate fair coin tosses.

Want to test if $p = 1/2$.

Can't just toss 100 times and conclude it works if number of heads is 50.

Use CLT

$$M_n = \sum_{i=1}^{n} X_i$$

$$P[|M_n - 1/2| > \varepsilon] = \alpha \ll 1$$

$$P[|M_n - 1/2| \leq \varepsilon] = P[-\varepsilon \leq M_n - 1/2 \leq \varepsilon] = P[-\varepsilon \leq \frac{1}{n} \sum_i (X_i - 1/2) \leq \varepsilon]$$

$$= P[-\varepsilon\sqrt{n}/\sigma_X \leq \frac{1}{\sqrt{n}} \sum_i (X_i - 1/2)/\sigma_X \leq \varepsilon\sqrt{n}/\sigma_X] = P[-\varepsilon\sqrt{n}/\sigma_X \leq Z \leq \varepsilon\sqrt{n}/\sigma_X]$$

$$Z = \frac{1}{\sqrt{n}} \sum_i \frac{X_i - 1/2}{\sigma_X}$$

easy to show $E[Z] = 0$ Var$[Z] = 1$

$$P[-\varepsilon\sqrt{n}/\sigma_X \leq Z \leq \varepsilon\sqrt{n}/\sigma_X] = F_Z(\varepsilon\sqrt{n}/\sigma_X) - F_Z(-\varepsilon\sqrt{n}/\sigma_X) \approx 1 - Q(\varepsilon\sqrt{n}/\sigma_X) - (1 - Q(-\varepsilon\sqrt{n}/\sigma_X))$$
$$= 1 - Q(\varepsilon\sqrt{n}/\sigma_X) - Q(\varepsilon\sqrt{n}/\sigma_X) = 1 - 2Q(\varepsilon\sqrt{n}/\sigma_X)$$

$$P[|M_n - 1/2| > \varepsilon] \approx 2Q(\varepsilon\sqrt{n}/\sigma_X)$$

$$2Q(\varepsilon\sqrt{n}/\sigma_X) = \alpha = 0.01$$

from table Q(2.6) = 0.005

$$\varepsilon\sqrt{n}/\sigma_X) = 2.6$$
$$n = (2.6\sigma_X/\varepsilon)^2$$

if $\varepsilon = 0.013$ then $n = 10^4$.

6.1

Random Processes

Modem example: frequency shift keying

A random process $X(t)$ for $t \in I$ is an indexed collection of random variables.
Index set I

- Continuous-time r.p, typically $I = \mathbb{R}$ or $I = [0, \infty)$
- Discrete-time r.p, typically $I = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ or $I = \{0, 1, 2, \ldots\}$
  Discrete-time r.p. also called **random sequence**, with notation $X_n,\ n \in I$

Two useful ways of thinking:

- Fix t_0, then $X(t_0)$ is a random variable: $X(t_0, s)$ for $s \in S$
- Fix $s_0 \in S$, then $X(t, s_0)$ vs $t \in I$ is called a **realization** or **sample path** or **sample function**

Example decaying cosine with random amplitude (not a very "random" random process...)
Let Y have a Uniform(1,3) distribution
Define $X(t) = Y(1 + e^{-t} \cos t)$ for $t \in I = [0, \infty)$

Can do simple calculations:

- Variance function: $\text{Var}[X(t)] = \text{Var}[Y(1 + e^{-t} \cos t)] = \text{Var}[Y](1 + e^{-t} \cos t)^2 = \frac{1}{3}(1 + e^{-t} \cos t)^2$
- First-order cdf: $P[X(t) \leq x] = P[Y(1 + e^{-t} \cos t) \leq x] = P[Y \leq x/(1 + e^{-t} \cos t)] = F_Y(x/(1 + e^{-t} \cos t))$

The above calculations only describe the **marginal** properties of the r.p. To fully characterize a r.p. we need:

6.2

kth-order joint cdf

$$F_{X(t_1),\ldots,X(t_k)}(x_1, \ldots, x_k) = P[X(t_1) \leq x_1, \ldots, X(t_k) \leq x_k]$$

If we know this for $k = 1, 2, \ldots$ and for all $t_j \in I$ and for all $x_j \in \mathbb{R}$, $j = 1, \ldots, k$, then we can compute any statistical quantity of interest about the r.p.

Equivalently can work with kth-order joint pdfs

$$f_{X(t_1),\ldots,X(t_k)}(x_1, \ldots, x_k) = \frac{d^k}{dx_1 \cdots dx_k} F_{X(t_1),\ldots,X(t_k)}(x_1, \ldots, x_k)$$

For a **discrete-valued** r.p., use the kth-order joint pmf.

The above could be painful in general, fortunately many r.p.s have properties that simplify their statistical characterization, as follows.

Independent Increments
A r.p. $X(t)$ is said to have **independent increments** iff for any $k = 1, 2, \ldots$ and any $t_1 < t_2 < \cdots < t_k$ where $t_j \in I$, the following r.v.s are independent:

$$X(t_1),\ X(t_2) - X(t_1),\ \ldots, X(t_k) - X(t_{k-1})$$

Example: Poisson process, random walks

Markov r.p.
A r.p. $X(t)$ is said to be **Markov** iff for any $k = 2, 3, \ldots$ and any $t_1 < t_2 < \cdots < t_k$ where $t_j \in I$,

$$f_{X(t_k)}(x_k | X(t_{k-1}) = x_{k-1},\ \ldots,\ X(t_1) = x_1) f_{X(t_k)}(x_k | X(t_{k-1}) = x_{k-1})$$

conditional statistics at time t_k depend only on most recently given value of r.p.
An independent increments r.p. is also a Markov r.p., but the converse is not true in general.

Stationary r.p.
A r.p. $X(t)$ is called **strict-sense stationary** iff its kth-order joint cdfs (or pdfs or pmfs) are all time-shift invariant:

$$F_{X(t_1),\ldots,X(t_k)}(x_1, \ldots, x_k) = F_{X(t_1+\tau),\ldots,X(t_k+\tau)}(x_1, \ldots, x_k)$$

for all $k = 1, 2, \ldots$, for all $x_j \in \mathbb{R}$ for all $t_j I$
- If $I = \mathbb{R}$, must hold for all $\tau \in \mathbb{R}$
- If $I = [0, \infty)$, must hold for all $\tau > 0$

Given a segment in time, can you predict what times it came from? If so, then nonstationary.

Even then, for many r.p.s, a full statistical characterization is intractable.
So often we focus on the moments.

Mean Function

$$\mu_X(t) = E[X(t)] = \int_{-\infty}^{\infty} x f_{X(t)}(x)\, dx$$

reflects trends in the *average* behavior of r.p. over time

For studying relationship between different points in time (e.g. predicting stock market), the 2nd-order moments are more useful.

Autocorrelation Function (autos: Greek for "self")

$$R_X(t_1, t_2) = E[X(t_1)X(t_2)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X(t_1),X(t_2)}(x, y)\, dx\, dy$$

Autocovariance Function

$$C_X(t_1, t_2) = E[(X(t_1) - \mu_X(t_1))(X(t_2) - \mu_X(t_2))] = R_X(t_1, t_2) - \mu_X(t_1)\mu_X(t_2)$$

Variance Function

$$\mathrm{Var}[X(t)] = E[(X(t) - \mu_X(t))^2] = C_X(t, t)$$

If $X(t)$ is a Gaussian r.p. (the joint pdf of any finite collection of time samples is jointly Gaussian) then all joint pdfs of $X(t)$ are completely specified by the mean function and autocovariance function.

6.3 Discrete-time random processes

i.i.d. random process

We say a discrete-time r.p. X_n is an **i.i.d. random process** iff $\forall k = 1, 2, \ldots$ and $\forall x_j \in \mathbb{R}$, $j = 1, \ldots, k$

$$F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = F_X(x_1)F_X(x_2)\cdots F_X(x_k) = \prod_{i=1}^{k} F_X(x_k)$$

for some common cdf $F_X(x)$

Example: **Bernoulli r.p.**: $P[X_n = 1] = p$ and $P[X_n = 0] = 1 - p$ for $n = 1, 2, \ldots$ (and independent)

Properties

- An i.i.d. random process is strict-sense stationary!
- $E[X_n] = \mu_X$ a constant for all n
- $C_X(n_1, n_2) = \text{Cov}\{X_{n_1}, X_{n_2}\} = \sigma_X^2 \delta_{n_1 - n_2}$ where the Kronecker delta function is: $\delta_n = \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases}$

Random Walk Process (drunkard's walk) (e.g. net auto traffic by incrementing and decrementing counter)

Let D_n be the modified Bernoulli r.p. with $P[D_n = 1] = p$ and $P[D_n = -1] = 1 - p$

$$W_n = \sum_{i=1}^{n} D_i, \quad n = 1, 2, \ldots$$

W_n is called a random walk process. (Picture)

Moments

$$\mu_W(n) = E[W_n] = E[\sum_{i=1}^{n} D_i] = \sum_{i=1}^{n} (1p - 1(1-p)) = n(2p-1)$$

$\text{Var}[W_n] = \text{Var}[\sum_{i=1}^{n} D_i] = n\sigma_D^2$ where $\sigma_D^2 = E[D_n^2] - (2p-1)^2 = 1^2 p + (-1)^2(1-p) - (2p-1)^2 = 1 - (4p^2 - 4p + 1) = 4pq$

so $\text{Var}[W_n] = n4pq$ (increases with n)

Suppose $n > m$ then (trick for any sum process):

$$C_W(n, m) = \text{Cov}\{W_n, W_m\} = \text{Cov}\left\{W_m + \sum_{i=m+1}^{n} D_i, W_m\right\} = \text{Cov}\{W_m, W_m\} + \text{Cov}\left\{\sum_{i=m+1}^{n} D_i, W_m\right\} = \text{Var}[W_m]$$

Repeating for $n < m$ we see

$$C_W(n, m) = \min(n, m)\sigma_D^2$$

Note that the range of W_n is $\{-n, -n+2, \ldots, n-2, n\}$, i.e. $R_{W_1} = \{-1, 1\}$ and $R_{W_2} = \{-2, 0, 2\}$ etc.

The pmf for $k = 0, \ldots, n$:

$$P[W_n = 2k - n] = P[W_n = k - (n-k)] = P[k \text{ of the } D_i\text{'s are 1's and the other } n-k \text{ are -1's}] = \binom{n}{k} p^k (1-p)^{n-k}$$

Pick $n > m$ then

$$P[W_n - W_m = k | W_m = w_m] = P[D_{m+1} + \cdots + D_n = k | D_1 + \cdots + D_m = w_m] = P[D_{m+1} + \cdots + D_n = k]$$

Thus the pmf of $W_n - W_m$ is independent of W_m for $n > m$, so W_n has **independent increments**

Also, the pmf of the increment $W_n - W_m$

$$P[W_n - W_m = k] = P[D_{m+1} + \cdots + D_n = k] = P[D_1 + \cdots + D_{n-m} = k]$$

depends only on $n - m$ and not on m, so W_n is said to have **stationary increments**

But not stationary, since moments depend on n (cf Bernoulli and random walk)

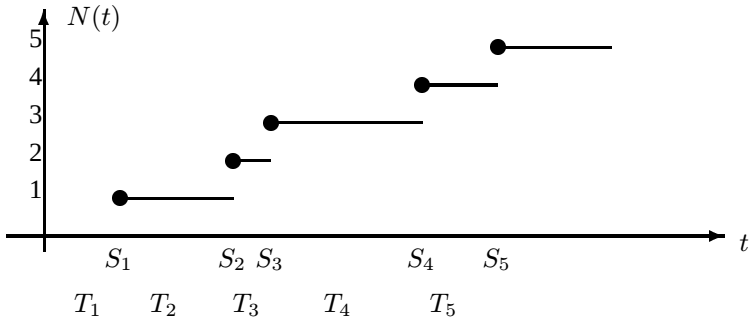
Sum process

In general a sum process formed by summing an i.i.d. r.p. will have independent increments and stationary increments.

6.4

Poisson Process

A counter increments every time a new packet arrives at a port on an asynchronous network.
The interarrival times are r.v.s $T_1, T_2, \ldots$, so the nth packet arrives at time $S_n = \sum_{i=1}^{n} T_i$
Let $N(t)$ denote the number of packets that have arrived by time t; typical sample function for $N(t)$:



We say $N(t)$ is a **Poisson process** iff

- $N(0) = 0$ (counter starts at 0)
- $N(t)$ has independent increments
- For any $0 \leq t_1 < t_2$ and $k = 0, 1, 2, \ldots$

$$P[N(t_2) - N(t_1) = k] = [\lambda(t_2 - t_1)]^k e^{-\lambda(t_2 - t_1)}/k!$$

so in particular $N(t)$ has a Poisson pmf:

$$P[N(t) = k] = (\lambda t)^k e^{-\lambda t}/k!$$

Properties:

- $P[N(t+\delta) - N(t) > 1] = 1 - P[N(t+\delta) - N(t) > 1] = 1 - e^{-\lambda\delta}((\lambda\delta)^0/0! + (\lambda\delta)^1/1!) = 1 - e^{-\lambda\delta}(1 + \lambda\delta)$
  So $P[N(t+\delta) - N(t) > 1] \to 0$ as $\delta \to 0$: only one new arrival at a time.
- $P[N(t+\delta) - N(t) > 0] = 1 - e^{-\lambda\delta}(\lambda\delta)^0/0! = 1 - e^{-\lambda\delta} \approx \lambda\delta$ for $\delta \approx 0$.
  So arrival probability roughly proportional to time interval.

Example: McDonalds opens at 8:00AM and assume arrival of customers is Poisson. with mean rate $\lambda = 3$/minute.
If by 8:30AM 90 customers have been served, what is probability that by 9:10 over 200 customers will have been served?

$$P[N(70) > 200 | N(30) = 90] = P[N(70) - N(30) > 200 - 90 | N(30) = 90] = P[N(70) - N(30) > 110] = \sum_{k=111}^{\infty} 120^k e^{-120}/k!$$

since $\lambda(t_2 - t_1) = 3 \cdot 40 = 120$

Since independent increments, joint pmf easy

Moments: $E[N(t)] = \lambda t$, hence λ is mean number of arrivals per unit time.

Autocovariance function (a trick for all independent increments r.p.s)
Choose $t_2 > t_1$:

$$C_N(t_2, t_1) = \text{Cov}\{N(t_2), N(t_1)\} = \text{Cov}\{(N(t_2) - N(t_1) + N(t_1), N(t_1)\} + \text{Cov}\{N(t_1), N(t_1)\}$$
$$= \text{Cov}\{N(t_2) - N(t_1), N(t_1)\} + \text{Cov}\{N(t_1), N(t_1)\} = 0 + \text{Var}[N(t_1)] = \lambda t_1$$

Repeating for $t_1 > t_2$ and combining:

$$C_N(t_2, t_1) = \lambda \min(t_1, t_2)$$

(Not w.s.s.)

Interarrival pdfs are exponential with mean $1/\lambda$. Partial argument for T_1, for $t \geq 0$:

$$P[T_1 > t] = P[N(t) = 0] = e^{-\lambda t} \text{ so } F_{T_1}(t) = (1 - e^{-\lambda t})u(t)$$

Arrival times S_n are Erlang since sum of i.i.d. exponentials

6.5

Stationarity

Some r.p.s are easier to handle because their statistical behavior does not change with time, in some sense.

Stationary r.p.

A r.p. $X(t)$ is called **strict-sense stationary** iff its kth-order joint cdfs (or pdfs or pmfs) are all time-shift invariant:

$$F_{X(t_1),\ldots,X(t_k)}(x_1,\ldots,x_k) = F_{X(t_1+\tau),\ldots,X(t_k+\tau)}(x_1,\ldots,x_k)$$

for all $k = 1, 2, \ldots$, for all $x_j \in \mathbb{R}$ for all $t_j I$

- If $I = \mathbb{R}$, must hold for all $\tau \in \mathbb{R}$
- If $I = [0, \infty)$, must hold for all $\tau > 0$

Given a segment in time, can you predict what times it came from? If so, then nonstationary.

Wide-sense stationary

s.s.s. can be hard to verify, so sometime we settle for time-shift invariance of the first two moments.

A r.p. $X(t)$ is called **wide-sense stationary** iff

- $E[X(t+\tau)] = E[X(t)] = \mu_X$ is independent of time t
- $C_X(t_1, t_2) = C_X(t_1 + \tau, t_2 + \tau)$ depends only on the time difference $t_2 - t_1$.
  Thus we have $C_X(t_1, t_2) = C_X(0, t_2 - t_1) = C_X(t_2 - t_1)$ where we drop the unneeded first argument.

If $X(t)$ is stationary, then $F_{X(t)}(x) = F_{X(t+\tau)}(x) \forall t, \tau, x$.

Thus the marginal moments are independent of time, and in particular $E[X(t)]$ is a constant.

Furthermore, the 2nd-order joint pdf can be written

$$f_{X(t_1),X(t_2)}(x_1, x_2) = f_{X(0),X(t_2-t_1)}(x_1, x_2)$$

so it only depends on the difference of $t_2 - t_1$, thus

$$C_X(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f_{X(t_1),X(t_2)}(x_1, x_2)\, dx_1\, dx_2 - \mu_X^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f_{X(0),X(t_2-t_1)}(x_1, x_2)\, dx_1\, dx_2 - \mu_X^2$$

depends only on the time difference

Conclusion: If $X(t)$ is s.s.s., then $X(t)$ is also w.s.s.

The converse is not true in general.

Exception: Gaussian and w.s.s. implies s.s.s., since joint pdfs of Gaussian depends only on 1st and 2nd moments

Poisson process: not stationary since $E[N(t)] = \lambda t$ depends on t

---

Example: sinusoid with random phase.

$\Theta \sim$ Uniform$(0, 2\pi)$.

$X(t) = \sin(t + \Theta)$

$\mu_X(t) = E[X(t)] = E[\sin(t + \Theta)] = \int_0^{2\pi} \sin(t + \phi)\frac{1}{2\pi}\, d\phi = 0$

$R_X(t, t+\tau) = E[X(t)X(t+\tau)] = E[\sin(t + \Theta)\sin(t + \tau + \Theta)] = \frac{1}{2}E[\cos(\tau) - \cos(2t + \tau + 2\Theta)] = \frac{1}{2}\cos(\tau)$ Autocorrelation depends only on time difference τ, so $X(t)$ is w.s.s.

Properties of autocorrelation function for w.s.s. r.p.

- Average power: $R_X(0) = E[X^2(t)]$
- Symmetry: $R_X(\tau) = R_X(-\tau)$
- $|R_X(\tau)| \leq R_X(0)$ since Schwarz Inequality: $|E[XY]| \leq E[|XY|] \leq \sqrt{E[|X|^2]E[|Y|^2]}$
- If continuous at origin, then continuous everywhere.
  Proof:

$$\begin{aligned} |R_X(\tau + \delta) - R_X(\tau)| &= |E[(X(\tau + \delta) - X(\tau))X(0)]| \leq \sqrt{E[|X(\tau + \delta) - X(\tau)|^2]E[|X(0)|^2]} \\ &= \sqrt{2(R_X(0) - R_X(\delta))R_X(0)}, \end{aligned}$$

  so if $|R_X(0) - R_X(\delta)| \to 0$ as $\delta \to 0$, then $|R_X(\tau + \delta) - R_X(\tau)| \to 0$ as $\delta \to 0$ for any τ.

- Measure of rate of change of r.p. see 6.59, p 360.

1st order stationarity does not imply s.s.s
X_n Bernoulli r.p. Let $Y(t) = X_n$ for $t \in [n, n+1)$

Often need multiple random processes, such as signal + noise.

Pairs of Random Processes

Complete statistical characterization through **joint finite-dimensional cdfs** (or pdfs or pmfs):

$$F_{X(t_1),\ldots,X(t_k),Y(s_1),\ldots,Y(s_j)}(x_1,\ldots,x_k,y_1,\ldots,y_j) = P[X(t_1) \le x_1,\ldots,X(t_k) \le x_k, Y(s_1) \le y_1,\ldots,Y(s_j) \le y_j]$$

$X(t)$ and $Y(t)$ are **jointly strict-sense stationary** r.p.s iff their finite joint cdfs are time-shift invariant, i.e. $\forall \tau \in \mathbb{R}$:

$$F_{X(t_1),\ldots,X(t_k),Y(s_1),\ldots,Y(s_j)}(x_1,\ldots,x_k,y_1,\ldots,y_j) = F_{X(t_1+\tau),\ldots,X(t_k+\tau),Y(s_1+\tau),\ldots,Y(s_j+\tau)}(x_1,\ldots,x_k,y_1,\ldots,y_j)$$

$X(t)$ and $Y(t)$ are **independent** r.p.s iff their finite joint cdfs factor into the product of their individual cdfs:

$$F_{X(t_1),\ldots,X(t_k),Y(s_1),\ldots,Y(s_j)}(x_1,\ldots,x_k,y_1,\ldots,y_j) = F_{X(t_1),\ldots,X(t_k)}(x_1,\ldots,x_k) F_{Y(s_1),\ldots,Y(s_j)}(y_1,\ldots,y_j)$$

(or jpdfs or jpmfs)

If $X(t)$ and $Y(t)$ are jointly strict-sense stationary, then they are individually strict-sense stationary. The reverse is not true in general. Exception: if $X(t)$ and $Y(t)$ are independent and individually s.s.s., then they are jointly s.s.s.

Moments of Pairs of Random Processes

- **Cross-correlation Function**

$$R_{XY}(t_1,t_2) = E[X(t_1)Y(t_2)] = \iint xy f_{X(t_1)Y(t_2)}(x,y)\,dx\,dy$$

- **Cross-covariance Function** measures linear coupling:

$$C_{XY}(t_1,t_2) = E[(X(t_1) - \mu_X(t_1))(Y(t_2) - \mu_Y(t_2))]$$

- $X(t)$ and $Y(t)$ are called **uncorrelated** r.p.s iff $C_{XY}(t_1,t_2) = 0\ \forall t_1,t_2 \in I$

Properties of Moments of Pairs of Random Processes (all for all $t_1,t_2 \in \mathbb{R}$)

- Hermitian symmetry:

$$R_{XY}(t_1,t_2) = R_{YX}(t_2,t_1), \quad C_{XY}(t_1,t_2) = C_{YX}(t_2,t_1)$$

- Autocorrelation function from cross-correlation function

$$R_X(t_1,t_2) = R_{XX}(t_1,t_2), \quad C_X(t_1,t_2) = C_{XX}(t_1,t_2)$$

- Cross-covariance / cross-correlation relationship:

$$C_{XY}(t_1,t_2) = R_{XY}(t_1,t_2) - E[X(t_1)]E[Y(t_2)]$$

- Schwarz inequality for cross-correlation:

$$|R_{XY}(t_1,t_2)| \le \sqrt{R_X(t_1,t_1)R_Y(t_2,t_2)}, \quad |C_{XY}(t_1,t_2)| \le \sqrt{\text{Var}[X(t_1)]\text{Var}[X(t_2)]}$$

- Mutual independence and cross-covariance:
  If $X(t)$ and $Y(t)$ mutually independent, then $C_{XY}(t_1,t_2) = 0$
  The converse not true in general. Exception: when $X(t)$ and $Y(t)$ are jointly Gaussian r.p.s.

We say $X(t)$ and $Y(t)$ are **jointly Gaussian random processes** iff all of their joint finite-dimensional density functions have the normal form with the appropriate mean and covariance, i.e.

$$\begin{bmatrix} X(t_1) \\ \vdots \\ X(t_k) \\ Y(s_1) \\ \vdots \\ Y(s_j) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} E[X(t_1)] \\ \vdots \\ E[X(t_k)] \\ E[Y(s_1)] \\ \vdots \\ E[Y(s_j)] \end{bmatrix}, \begin{bmatrix} C_X(t_1,t_1) & \cdots & C_X(t_1,t_k) & C_{XY}(t_1,s_1) & \cdots & C_{XY}(t_1,s_j) \\ \vdots & & \vdots & \vdots & & \vdots \\ C_X(t_k,t_1) & \cdots & C_X(t_k,t_k) & C_{XY}(t_k,s_1) & \cdots & C_{XY}(t_k,s_j) \\ C_{YX}(s_1,t_1) & \cdots & C_{YX}(s_1,t_k) & C_Y(s_1,s_1) & \cdots & C_Y(s_1,s_j) \\ \vdots & & \vdots & \vdots & & \vdots \\ C_{YX}(s_j,t_1) & \cdots & C_{YX}(s_j,t_k) & C_Y(s_j,s_1) & \cdots & C_Y(s_j,s_j) \end{bmatrix} \right)$$

for all appropriate values of the indices.

Joint wide-sense stationarity

We say $X(t)$ and $Y(t)$ are **jointly wide-sense stationary** r.p.s iff

- Each of $X(t)$ and $Y(t)$ are individually WSS, and
- Their **cross-correlation** is invariant to time shifts: $R_{XY}(t_1,t_2) = R_{XY}(t_1+\tau, t_2+\tau)\ \forall \tau \in \mathbb{R}$ and $\forall t_1, t_2 \in \mathbb{R}$
  In other words, $R_{XY}(t_1,t_2) = R_{XY}(0, t_2-t_1) = R_{XY}(t_2-t_1)$ depends only on the time difference $t_2 - t_1$, or equivalently $R_{XY}(t, t+\tau)$ is independent of t

If $X(t)$ and $Y(t)$ are jointly strict-sense stationary, then they are jointly wide-sense stationary. In general, the reverse is not true. An exception is jointly Gaussian, jointly WSS random processes.

Example: $Y(t) = AX(t) + N(t)$ where $X(t)$ and $N(t)$ are jointly WSS.
Assume $X(t)$ and $N(t)$ are independent of r.v. A.
Is $Y(t)$ w.s.s.?
$E[Y(t)] = E[AX(t) + N(t)] = E[A]E[X(t)] + E[N(t)] = \mu_A \mu_X + \mu_N$
so mean is independent of time t

$$R_Y(t, t+\tau) = E[Y(t)Y(t+\tau)] = E[(AX(t) + N(t))(AX(t+\tau) + N(t+\tau))]$$

$$= E[A^2 X(t)X(t+\tau)] + E[AX(t)N(t+\tau)] + E[AX(t+\tau)N(t)] + E[N(t)N(t+\tau)]$$

$$= E[A^2]R_X(t,t+\tau) + \mu_A R_{XN}(t,t+\tau) + \mu_A R_{XN}(t+\tau,t) + R_N(t,t+\tau) = E[A^2]R_X(\tau) + \mu_A R_{XN}(\tau) + \mu_A R_{XN}(\tau) + R_N(\tau)$$

Thus $Y(t)$ is w.s.s.

Considering the case where A is a constant, we have shown that the sum of two jointly w.s.s. r.p.s is w.s.s.

Product: $Z(t) = X(t)Y(t)$ where $X(t)$ and $Y(t)$ are w.s.s. and independent r.p.s

$$E[Z(t)] = E[X(t)Y(t)] = E[X(t)]E[Y(t)] = \mu_X \mu_Y$$

$$R_Z(t, t+\tau) = E[Z(t)Z(t+\tau)] = E[X(t)Y(t)X(t+\tau)Y(t+\tau)] = E[X(t)X(t+\tau)]E[Y(t)Y(t+\tau)]R_X(\tau)R_Y(\tau)$$

So $Z(t)$ is w.s.s.

White Noise

Consider triangular autocorrelation function as width goes to 0 but area stays constant (so height goes to ∞).
We say $N(t)$ is **white noise** if its autocorrelation function is the Dirac delta: $R_N(\tau) = \alpha\delta(\tau)$ or equivalently $R_N(t_2,t_1) = \alpha\delta(t_2 - t_1)$ for some constant α
Continuous-time generalization of i.i.d. random sequence

Binary Communications (at last a fairly real example...)
We transmit a rectangular pulse $X \cdot s(t)$, where X is 1 or 0 depending on whether we send a 1 or 0.

Received signal: $Y(t) = Xs(t) + N(t)$ where $N(t)$ is additive white Gaussian noise (AWGN) with mean zero.
Sensible receiver: $Z = \frac{1}{T} \int_0^T Y(t)\ dt$
Hopefully Z is close to 1 if a 1 is sent, and 0 if a 0 is sent.

$$E[Z|X=x] = E[\frac{1}{T}\int_0^T Y(t)\ dt|X=x] = \frac{1}{T}\int_0^T E[Y(t)|X=x]\ dt = \frac{1}{T}\int_0^T x\, s(t)\ dt = x$$

$$E[Z^2|X=x] = E[(\frac{1}{T}\int_0^T Y(t)\ dt)^2|X=x] = E[(\frac{1}{T}\int_0^T Y(t)\ dt)(\frac{1}{T}\int_0^T Y(t')\ dt')|X=x]$$

$$= E[\frac{1}{T^2}\int_0^T\int_0^T Y(t)Y(t')\ dt\ dt'|X=x] = \frac{1}{T^2}\int_0^T\int_0^T E[Y(t)Y(t')|X=x]\ dt\ dt'$$

Now

$$E[Y(t)Y(t')|X=x] = E[(xs(t)+N(t))(xs(t')+N(t'))] = x^2 s(t)s(t') + xs(t)E[N(t')] + xs(t')E[N(t)] + E[N(t)N(t')]$$

$$= x^2 s(t)s(t') + \alpha\delta(t-t')$$

so

$$E[Z^2|X=x] = \frac{1}{T^2}\int_0^T\int_0^T [x^2 s(t)s(t') + \alpha\delta(t-t')]\ dt\ dt' = x^2 + \frac{1}{T^2}\int_0^T \alpha\ dt' = x^2 + \alpha/T$$

Thus

$$\text{Var}[Z|X=x] = E[Z^2|X=x] - (E[Z|X=x])^2 = x^2 + \alpha/T - x^2 = \alpha/T$$

Note that larger α means more variance for Z, but larger time T reduces variance

Natural decision rule: choose 1 if $Z > 1/2$ and 0 otherwise.
Probability of error?

$$P[E] = P[E|X=0]P[X=0] + P[E|X=1]P[X=1] = P[Z > 1/2|X=0]P[X=0] + P[Z < 1/2|X=1]P[X=1]$$

If $X = 0$, then Z has Gaussian distribution with mean 0 and variance α/T.

$$P[Z > 1/2|X=0] = Q\left(\frac{1/2-0}{\sqrt{\alpha/T}}\right) = Q\left(\sqrt{\frac{T}{4\alpha}}\right)$$

Hence importance of Q function to EE's

Many concepts:
- Random processes: white noise - wide-sense stationary
- Moments: mean and variance
- Probability: conditional probability, total probability
- Random variables: Gaussian pdf., calculating probabilities by integrating pdf

Summary (new concepts in Ch. 4, 5, 6 over previous)

Multiple Random Variables
- Joint pdf, cdf $P[(X,Y) \in B] = \iint_B f_{X,Y}(x,y)\, dx\, dy$
- Independence: jpdf factors, simplifies calculating P and moments
- Functions of multiple r.v.s: $Z = g(X,Y)$ by method of events (cdf then pdf)
- Moments: correlation, covariance, correlation coefficient. Independence implies uncorrelated (zero covariance).

Sums of Random Variables
- Mean of sum is sum of means (always).
- Variance of sum is sum of variances, if independent, otherwise must include covariance of all cross terms.
- Sample mean is unbiased, its variance is $\sigma_X^2/n \to 0$ as $n \to \infty$
- Weak Law of Large Numbers: proof by Chebychev, P[sample mean is close to mean μ_X] goes to 1 as $n \to \infty$.
- Strong Law of Large Numbers: almost all sample means converge to mean μ_X
- Central Limit Theorem: sum of standardized r.v.s normalized by $1/\sqrt{n}$ approaches a Gaussian distribution for large n, so can calculate approximate probabilities.

Random Processes
- Simplifying properties: strict-sense stationarity, independent increments, Markov
- Moments: mean function, autocorrelation function, autocovariance function
- Wide-sense stationarity (need Ch. 7 to fully realize utility)
- Sum processes: random walk, Binomial process. have independent increments, stationary increments, calculated moments
- Poisson counting process: useful for "random" arrivals
- Pairs of random processes: independence, cross-correlation, cross-covariance
- Signal+noise and applications...

7.2

Linear Systems

$$Y(t) = (X \star h)(t) = \int_{-\infty}^{\infty} h(s) X(t-s)\, ds$$

Assume $X(t)$ is w.s.s.

$$E[Y(t)] = E[\int_{-\infty}^{\infty} h(s) X(t-s)\, ds] = \int_{-\infty}^{\infty} h(s) E[X(t-s)]\, ds = \int_{-\infty}^{\infty} h(s) \mu_X(t-s)\, ds = \mu_X \int_{-\infty}^{\infty} h(s)\, ds = \mu_X H(0)$$

independent of time t

$$R_Y(t, t+\tau) = E[Y(t) Y(t+\tau)] = E\left[\int_{-\infty}^{\infty} h(s) X(t-s)\, ds \int_{-\infty}^{\infty} h(r) X(t+\tau-r)\, dr \right]$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(s) h(r) E[X(t-s) X(t+\tau-r)]\, ds\, dr = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(s) h(r) R_X(\tau + s - r)\, ds\, dr$$

independent of t, so $Y(t)$ is also w.s.s.!

WSS Random Processes and LSI systems

- For BIBO LSI system, WSS input yields WSS output, and input and output are jointly WSS.
- Power spectral density: $S_X(\omega) = \int R_X(t) e^{-j\omega t}\, dt$. (Fourier transform of autocorrelation function.)
- For LSI system with impulse response $h(t)$ and transfer function $H(\omega) = \int h(t) e^{-j\omega t}\, dt$, the input-output relationship is $S_Y(\omega) = |H(\omega)|^2 S_X(\omega)$.

Total probability with conditioning
Recall if A_i's partition S, then

$$P(B) = \sum_i P(B|A_i)P(A_i)$$

Fact:

$$\boxed{P(B|C) = \sum_i P(B|A_i, C)P(A_i|C)}$$

Proof:

$$RHS = \sum_i \frac{P(B \cap A_i \cap C)}{P(A_i \cap C)} \frac{P(A_i \cap C)}{P(C)} = \frac{\sum_i P(B \cap A_i \cap C)}{P(C)} = \frac{\sum_i P(B \cap C|A_i)P(A_i)}{P(C)} = \frac{P(B \cap C)}{P(C)} = P(B|C)$$

Gambler's Ruin Markov Chain
Gambler starts with $X_0 = \$10$, plays game repeatedly, wins \$1 with probability p, and loses \$1 with probability $q = 1 - p$.
This problem statement implies the following:

$$P[X_{n+1} = j|X_n = k] = \begin{cases} p, & j = k+1 \\ q & j = k-1 \\ 0, & \text{otherwise.} \end{cases}$$

This is a Markov random process because PMF of the next state only depends on the previous state, not on earlier states.

Gambler must stop playing if $X_n = 0$ (ruined).
Gambler decides in advance to stop playing if $X_n = 20$.

Hitting times:

$$T_0 = \min\{n \geq 0 : X_n = 0\} \quad T_{20} = \min\{n \geq 0 : X_n = 20\}$$

Main probability of interest is $u(10)$, where:

$$u(k) = P[T_0 < T_{20}|X_0 = k]$$

i.e., what is the probability of being ruined rather than walking away with \$20?

If $p = 1/2$, then expect $u(10) = 1/2$. But most casino games have $p < 1/2$.

End conditions: $u(0) = 1$ and $u(20) = 0$.
Trick: using above total probability for $0 < k < 20$

$$P[T_0 < T_{20}|X_0 = k]$$

$$= P[T_0 < T_{20}|X_0 = k, X_1 = k+1]P[X_1 = k+1|X_0 = k] + P[T_0 < T_{20}|X_0 = k, X_1 = k-1]P[X_1 = k-1|X_0 = k]$$

thus

$$u(k) = u(k+1)p + u(k-1)q$$

Solution to this recursive equation with end conditions is (See Hoel, Port, Stone):

$$u(k) = \frac{\sum_{j=k}^{20-1}(q/p)^j}{\sum_{j=0}^{20-1}(q/p)^j} = (q/p)^k \frac{1-(q/p)^{20-k}}{1-(q/p)^{20}} \quad \text{if } q \neq p$$

Note if $p = q = 1/2$, then as expected

$$u(k) = \frac{20-k}{20} \quad \text{so} \quad u(10) = 1/2$$

But more realistic value might be $p = 0.45$. In which case $u(10) = 0.88$.
So very high odds of running out of money before doubling money!