Probability Cheatsheet v2.0

Janet Matsen's Statistics notes prepended to an excellent existing Statistics cheat sheet:
Original compiled by William Chen (http://wzchen.com) and Joe Blitzstein, with contributions from Sebastian Chiu, Yuan Jiang, Yuqi Hou, and Jessy Hwang. Material based on Joe Blitzstein's (@stat110) lectures (http://stat110.net) and Blitzstein/Hwang's Introduction to Probability textbook (http://bit.ly/introprobability). Please share comments, suggestions, and errors at http://github.com/wzchen/probability_cheatsheet.

Last Updated December 20, 2015

Cards in a deck: 52.
Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King

Problem solving tips

Recognize complicated versions of simple formulas. E.g.
$P(A \mid B) = P(A \cap B)/P(B)$ with an A or B like $(A \cup B^c)$
At least is a trigger for $1 -$ probability of all items. E.g. probability you will roll at least one 6 in 3 die rolls
$= 1 - P(\text{no 6s rolled}) = 1 - (5/6)^3$.

Probability Formulas

Misc.

$$\Phi = \Omega^c \text{ (Lecture 4)}$$
$$\cap E_i = (\cup E_i^c)^c \text{ (Lecture 4)}$$
$$A \cup A^c = \Phi$$

DeMorgan's laws, pg 26: You can distribute/factor a complement if you switch the $\cup$ or $\cap$

$$(A \cup B)^c = A^c \cap B^c$$
$$\text{so } P((A \cup B)^c) = P(A^c \cap B^c)$$
$$(A \cap B)^c = A^c \cup B^c$$

Goal: you have one complement in your "and" space and don't want it. $A \cap B^c = A - (A \cap B)$ Use Venn diagrams to see. (in pg 37 proof)
What about $A \cup B^c$? ???

$$P(A \cap B) + P(A^c \cap B) = P(B) \text{ (in prob 2.4.5 pg 37)}$$
$$P(A \cup B) + P(A^c \cup B) = 1 \text{ (logically b/c } P(A) + P(A^c) = 1$$

Mutually exclusive and/or exhaustive

Mutually exclusive E_1, E_2 are *mutually exclusive* if $P(E_1 \cup E_2 \cup E_3 \cup \ldots) = P(E_1) + P(E_2) + P(E_3) + \ldots$ (Lec 2, L&M 2.3).
Note that then $P(E_i \cap E_j) = 0$

Mutually exclusive and exhaustive

For $E_1, E_2, \cdots$:

$$E_i \cap E_j = \Phi,\ P(E_i \cap E_j) = 0, \text{ and } \Omega = E_1 \cup E_2 \cup \ldots$$

Mutually exclusive + exhaustive = partition

The fundamental laws of set algebra

(Wikipedia)
Commutative Laws

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

Associative Laws

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

Distributive laws

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

Basic probability formulae

$$\Omega = E \cup E^c$$
$$E \cap E^c = \Phi$$
$$\text{so } P(E^c) + P(E) = P(\Omega) = 1, or P(E^c) = 1 - P(E)$$
$$\text{and} P(D) \leq 1, \text{ since all probabilities are non-negative}$$
$$E \cup F = E \cup (E^c \cap F)$$
$$\text{and E and } E^c \text{ are disjoint, so you can break them apart}$$
$$P(E \cup F) = P(E) + P(E^c) \cap F$$
$$\text{so } P(E \cup F) + P(E \cap F) =$$
$$P(E) + P(E^c \cap F) + P(E \cap F) = P(E) + P(F)$$
$$\text{restated: } P(E \cup F) + P(E \cap F) = P(E) + P(F)$$

Law of total probability See Lec2, L&M 2.3
For events E, F of a partition, you can build F or P(F) from the overlaps of the events E with P.

$$F = U_i(F \cap E_i)$$
$$P(F) = \sum_{n=1} P(F \cap E_i)$$

Special case: If E_i is the ith outcome in a countable Ω, $F \cap E_i = Ei$ or $F \cap E_i = \Phi$ and $P(F) = \sum_{i \in F} P(E_i)$.

The inclusion and exclusion formula (The overlapping tiles sum)
If you want to get the $\cup$ of some events, sum all the areas, subtract out the singly over lapping bits, add back the tripply overlapping bits, subtract off the quadruply overlapping bits, etc. Go as far as the number of Es you have. All cups on the left side and caps on the right.

$$\text{2 items: } P(D \cup E) = P(D) + P(E) - P(D \cap E)$$

$$\text{3 items: } P(C \cup D \cup E) = P(C) + P(D) + P(E) - P(C \cap D) - P(D \cap E) - P(E \cap C) + P(C \cap D \cap E)$$

Review ways of breaking apart $\cup$ & and $\cap$
You can relate cap and cup by Theorem 2.3.6 (pg 28):
$P(A \cup B) = P(A) + P(B) - P(A \cup B)$, which can be re-arranged.

$\cup$ covers two areas. If they are independent you can separate them with pluses. If you know about their $\cap$ you can deduce their $\cup$.

$\cap$ is the

Conditional probability

Relating conditional probability & intersection L&M pg 34, L5.
If P(B) >0:

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

How to remember if you got the | and $\cap$ in the right order: visualize the and inside the circle.

This gives you a way to relate intersections and conditional probabilities: $P(A \cap B) = P(A \mid B)P(B)$ (pg. 34)

Some more tricks Recognize the relation between $A \cup B$ and $A \cap B$ even if there is something like $(A \cup B) \cap (A \cup B)^c$. You can pull the c inside and write the $P(A) + P(B)$ type statement.

A+B and A-B without P involved Generally avoid adding sets. Adding and subtracting probability of sets is great though.
With that warning, The definition of $A - B$ is $A \cap B^c$. Note that $P(A - B) \neq P(A) - P(B)$. Wikipedia uses \for $-$ in sets.
E suggests always avoiding + when doing math of sets. It's fine to have + when probability is involved, but adding sets doesn't make a lot of sense. You could call $\cup$ addition, but that is confusing.

The chain rule

If you don't want to use intersections, you can use only conditional probabilities to calculate an intersection. L&M pg 43. If
$P(E_1 \cap E_2 \cap \cdots \cap E_n) > 0$,
$P(E_1 \cap E_2 \cap \cdots \cap E_n)$
$= P(E_1)P(E_2 \mid E1)P(E_3 \mid E1 \cap E2) \ldots P(E_n \mid E_1 \cap E_2 \cap \cdots \cap E_n)$.
Note you could put $E_1, E_2, \ldots$ in any order you want (commutative rule). Also note that there is no cool rul for $P(A \cup B \cup C)$

E.g. $P(3 \text{ face cards in a row}) = P(F_1 \cap F_2 \cap F_3) = P(F_1) \cdot P(F_2 \mid F_1) \cdot P(F_3 \mid F_1 \cap F_2) = \frac{12}{52} \cdot \frac{11}{51} \cdot \frac{10}{50}$

Review: if you want to break apart unions, you think of the tiles overlapping and subtracting/adding sections intersections back. If you want to break apart intersections, you can use the chain rule to break it into conditional probabilities.

Covered later: If the events are independent then you can break them apart: $P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1) \cdot P(E_2) \cdot \ldots P(E_n)$.
E.g. odds of getting three odd rolls in a row with a die:
$P(odd_1 \cap odd_2 \cap \cdots \cap odd_3) = P(odd_1) \cdot P(odd_2) \cdot P(odd_3) = (1/2)^3$.

Law of total probability

Lecture 5, L&M pg 43
If you have mutually exclusive and exhaustive events H_i, and $i = 1, \ldots k$. Mutually exclusive & exhaustive means $\sum_{i=1}^k P(H_i) = 1$.
Note also $H_i \cap H_j = \Phi$, $P(H_i \cap H_j) = 0$ Since
$D = (D \cap H_1) \cup (D \cap H_2) \cup \cdots \cup (D \cap H_k)$

$$P(D) = \sum_{j=1}^{k} P(D \mid H_j)P(H_j)$$

We are summing across outcomes, and weighting the sums by the likelyhood of the sub-outcomes.

Bayes' formula

L&M p48, Lecture 5 Assume H_i, $i = 1, \ldots, k$ as above (mutually exclusive and exhaustive), and D with $P(D) > 0$.

$$P(H_j \mid D) = \frac{P(H_j \cap D)}{P(D)} = \frac{P(D \mid H_j)P(H_j)}{\sum_{i=1}^{k} P(D \mid H_j)P(H_j)}$$

P(have antigen A | have antigen B) $= P(AB)/(P(B) + P(AB))$

Example (J's Lecture 1). Coin C_1 has probability 1/4 of giving heads. Coin C_2 has probability 1/2 of giving heads. What is $P(C_1 \mid H)$? Bayes theorem.

$$\begin{aligned} P(C_1 \mid H) &= P(C_1 \cap H)/P(H) \\ &\text{use } P(C_1 \cap H) = P(H \mid C_1) \cdot P(C_1) \\ &\text{use } P(H) = P(H \mid C_1) \cdot P(C_1) + P(H \mid C_2) \cdot P(C_2) \\ &= \frac{P(H \mid C_1) \cdot P(C_1)}{P(H \mid C_1) \cdot P(C_1) + P(H \mid C_2) \cdot P(C_2)} \\ &= \frac{(1/4) \cdot (1/2)}{(1/4) \cdot (1/2) + (3/4) \cdot (1/2)} = \frac{1}{4} \end{aligned}$$

If you get heads again, what is your probability of coin 1? Call your $P(C_1)$ from above $P^*(C_1)$. Now $P(H_2)$ (probability of a 2nd heads) is:

$$\begin{aligned} P(C_1 \mid H_2) &= P(C_1 \cap H_2)/P(H_2) \\ &\text{use } P(C_1 \cap H_2) = P(H_2 \mid C_1) \cdot P^*(C_1) \\ &\text{use } P(H_2) = P(H_2 \mid C_1) \cdot P^*(C_1) + P(H_2 \mid C_2) \cdot P^*(C_2) \\ &= \frac{P(H_2 \mid C_1) \cdot P^*(C_1)}{P(H_2 \mid C_1) \cdot P^*(C_1) + P(H_2 \mid C_2) \cdot P^*(C_2)} \\ &= \frac{(1/4) \cdot (1/4)}{(1/4) \cdot (1/4) + (3/4) \cdot (3/2)} = \frac{1}{10} \end{aligned}$$

We could have done it without Bayes: Instead use $P(HH)$ which is P(two heads).

$$\begin{aligned} P(C_1 \mid HH) &= P(C_1 \cap HH)/P(HH) \\ &\text{use } P(C_1 \cap HH) = P(HH \mid C_1) \cdot P(C_1) \\ &\text{use } P(HH) = P(HH \mid C_1) \cdot P(C_1) + P(HH \mid C_2) \cdot P(C_2) \\ &\text{use } P(HH \mid C_1) = P(H_2 \cap H_1 \mid C_1) \\ &= \frac{P(HH \mid C_1) \cdot P(C_1)}{P(HH \mid C_1) \cdot P(C_1) + P(HH \mid C_2) \cdot P(C_2)} \\ &\text{use Conditional Independence: } P(H_2 \cap H_1 \mid C_1) = \frac{1}{4} \cdot \frac{1}{4} \\ &\text{use Conditional Independence: } P(H_2 \cap H_1 \mid C_2) = \frac{3}{4} \cdot \frac{3}{4} \\ &= \frac{(1/4)^2 \cdot (1/2)}{(1/4)^2 \cdot (1/2) + (3/4)^2 \cdot (1/2)} = \frac{1}{10} \end{aligned}$$

Independent and disjoint are not equivalent

The definition of independence is $P(A \cap B) = P(A) \cdot P(B)$. The definition of mutually exclusive is $P(A \cap B) = 0$. If A and B have nonzero probability then these can't both be true.

Independent: Two events A and B are said to be independent if $P(A \cap B) = P(A) \cdot P(B)$ (pg 53) Note that if A and B are independent, A^c and B^c are independent. pg54 ex 2.5.2. I don't believe you can visualize independence in a Venn diagram.

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(A) \cdot P(B)$
$P(A \mid B) = P(A \cap B)/P(B) = P(A) \cdot P(B)/P(B) = P(A)$

Example: (Friday 10/9 lecture) Two coins: C_1 gives heads 1/4 of the time and C_2 gives heads 3/4 of the time. You get one of the two coins randomly and toss it twice. Are the two coin tosses independent?
Answer: given the coin, they are independent. Allows for
$P(H_1 \cap H_2 | C_1) = \left(\frac{1}{4}\right)^2$.
Separate question: without knowing what coin you have,
$P(H_1 \cap H_2) = \frac{1}{2} \cdot \left(\frac{1}{4}\right)^2 + \frac{1}{2} \cdot \left(\frac{3}{4}\right)^2 = \frac{10}{32} = \frac{5}{16}$.
But if you calculate the probabilities separately:
$P(H_1) = \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{3}{4} = \frac{1}{2}$.
If H_1 and H_2 were independent, $P(H_1) = P(H_2)$ and
$P(H_1 \cap H_2) = P(H_1) \cdot P(H_2)$.
Since $P(H_1) \cdot P(H_2) = P(H_1)^2 = \frac{1}{4}$ and $P(H_1 \cap H_2) = 5/16$ we know
$P(H_1 \cap H_2) \neq P(H_1) \cdot P(H_2)$

Mutually exclusive: Two events A and B are said to be mutually exclusive if $P(A \cap B) = 0$ (pg 22) and $P(A \cup B) = P(A) + P(B)$ (E confirmed.) **Disjoint**: mutually exclusive sets.
$P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B)$
$P(A \mid B) = P(A \cap B)/P(B) = 0/B = 0$
If you assume mutual exclusivity when it isn't true, you get too high of a number for $P(A \cup B)$. Vizualize the Venn diagram for that: need to have $A \cap B = \emptyset$
If you are adding probabilities, they must be for mutually exclusive events.

Independence of More than Two events: Events $A_1, A_2, \ldots A_n$ are said to be *independent* if for every set of indices $i_1, i_2, \ldots, i_n$ between 1 and n, inclusive, pg 59, Sect 2.5, Lec6

$$P(A_{i1} \cap A_{i2} \cap \cdots \cap A_{ik}) = P(A_{i1}) \cdot P(A_{i2}) \ldots P(A_{ik})$$

Note that the k is different than the n. You have to get the doubles and the triples, etc. For the sets A, B, C, D, you have 11:
$(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)$,
$(A, B, C), (B, C, D), (A, C, D), (A, B, D), (A, B, C, D)$

If you only know single probabilities, how do you decompose $P((A_1 \cap A_2) \cup (A_3 \cap B_4))$? You can't break apart the cups that requires mutual exclusivity, so convert it to
$P(A_1 \cap A_2) + P(A_3 \cap A_4) + P((A_1 \cap A_2) \cap (A_3 \cap A_4))$
Trickier: how do you simplify $P(A_2 \cup A_3) \cap A_4$ if you only know A_2, A_3, and A_4 are independent? Factor out the $P(A_4)$ and do the same thing. $[P(A_2) + P(A_3) - P(A_2 \cap A_3)] \cdot P(A_4)$

Conditional Independence (I don't believe we covered this in class; from E) A and B are conditionally independent given C if
$P(A \cap B|C) = P(A|C)P(B|C)$. E.g. the Bayes coin flip example where you can say P(Heads 1 $\cap$ Heads 2 | Coin 1) =
$P(\text{Heads } 1|Coin_1) \cdot P(\text{Heads } 1|Coin_1) = 2 \cdot P(\text{Heads}|Coin_1)$
Conditional independence does not imply independence, and independence does not imply conditional independence.

Combinatorics

Multiplication Rule

for counting ordered sequences If A can be performed in m different ways and operation B in n different ways, the sequence (operation A, operation B) can be performed in $m * n$ different ways. pg 68

Permutations

for when all objects are distinct. The number of permutations of length k that can be formed from a set of n distinct elements, repetitions NOT allowed, is denoted by the symbol ${}_nP_k$ where ${}_nP_k = n(n-1)(n-2)\ldots(n-k-1) = \frac{n!}{(n-k)!}$ (pg 74)

$${}_nP_k = \frac{n!}{(n-k)!}$$

Corollary: the number of ways to permute an entire set of n distinct objects in ${}_nP_n = n!$ (pg 74)
When all objects are NOT distinct. The number of ways to arrange n objects, n_1 being of one kind, n_2 of a second kind, ..., and n_r of an r^{th} kind is

$$\text{not distinct: } \frac{n!}{n_1!n_2 \ldots !n_r!}$$

Note: ratios like this are called *multinomial coefficients* because the general term in the expansion of $(x_1 + x_2 + \cdots + x_n$ is $\frac{n!}{n_1!n_2!\ldots n_r!} x_1^{n1} x_2^{n2} \ldots x_r^{nr}$ (pg 81)

Sterling's formula for evaluating big factorials (pg 77):
$log_{10}(n!) = log_{10}(\sqrt{2 * \pi}) + (n + \frac{1}{2}) \cdot log_{10}(n) - n * log_{10}(e)$

Combinations

(Order doesn't matter.) The number of ways to form combinations of size k from a set of n **distinct** objects, repetitions not allowed, is denoted by the symbols $\binom{n}{k}$ or ${}_nC_k$, where (pg 86)

$$\binom{n}{k} = {}_nC_k = \frac{n!}{k!(n-k)!}$$

This is the number of permutations ($\frac{n!}{(n-k)!}$) divided by the number of ways to order those k selections (so divide by k). (Leaving $(n-k)$ items out of the selection. Order doesn't matter so reduce the number of sets by $k!$)
Note: $\binom{n}{k} = \binom{n}{n-k}$ (Lec 10/12/2015)

Classic example: How many unique hand shakes in a room of 8 people? ${}_8C_2$ applies

What about when objects are **not** distinct but you still want combinations?

Combinations/permutations on tests: First decide whether the objects are distinguishable. If they are distinguishable try to think of it as a permutation problem. If not, try to think of it as a combination problem. E.g. ways to order 4 blue chips and 4 red chips.

Differentiating permutations from permutations with repeats from combinations: For the set $(0, 1, 2, 2)$, the number of permutations of length two are ${}_nP_k = \frac{2!}{(4-1)!}$. The number of permutations if they are not distinct...?? The number of combinations:

Note: ALL CHAPTER 2 PROBLEMS ASSUME EQUAL PROBABILITY FOR DIFFERENT OUTCOMES.

Binomial and Hypergeometric

Note: this is for objects that are indistinguishable. You can distinguish a red fish from a blue fish but you can't tell whether a red fish's name is "Mary" or "Gary".

Binomial distribution/"probability"

Binomial Distribution ("Probability" in 10/12 lecture). A series of **independent trials**, each resulting in one of **two** possible outcomes, "success", or "failure". So yes replacement if it keeps trials independent, and can't have three possible outcomes.

$$P(k \text{ successes}) = \binom{n}{k} p^k (1-p)^k, \text{ for } k = 0, 1, \ldots, n$$

Note this is the number of ways of getting k indistinguishable objects in n tries. Then you multiply by the probability, p^k, of getting that number k, and the probability $(1-p)^k$ of getting the rest.
P(2 heads out of three for a coin that gives heads 3/4 of the time) = $(3) * (0.75^2) * (0.25^1)$

Hypergeometric Distribution

Hypergeometric Distribution For **no replacement**. Draw n chips from an urn that has r red chips, w white chips, and $r + w = N$. n is the total number of chips you will draw; k/n will be the fraction you draw that are red. Draw chips without replacement. Note k must be < r or the problem doesn't make sense and something like $\binom{r}{k} = \binom{1}{5}$ =?!?! can happen.

$$P(\text{k red chips chosen}) = \frac{\binom{r}{k}\binom{w}{n-k}}{\binom{N}{k}}$$

$$\frac{\text{(ways to draw k red from r)} * \text{(ways to draw n-k non-red from r)}}{\text{total number of combinations of chips you can draw}}$$

Note: this is the same as pg 111

$$= \frac{\binom{n}{k}({}_rP_k)({}_wP_{n-k})}{{}_NP_n}$$

Note that if you sample a very small fraction of the fish pond you will recover the binomial distribution (because replacement vs. no replacement isn't so important).

Say you have $1, 2, 3, \ldots, t$ types of objects with numbers $n_1, n_2, \ldots, n_t$ in an urn. If you want to chose k_1 objects of type 1, k_2 objects of type 2, ... k_t objects of type t. Then the number of objects is $N = n_1 + n_2 + \cdots + n_t$, the number you are selecting is $n = k_1 + k_2 + \cdots + k_t$, and the formula is q 3.2.35, pg 118

$$\frac{\binom{n_1}{k_1}\binom{n_2}{k_2}\cdots\binom{n_t}{k_t}}{\binom{N}{n}}$$

This is different than the formula I derived. I was multiplying the probabilities separately, and updating N each time. Is this an overestimation?

Random Variables

R.Var.	pdf	cdf
discrete	pmf (m = mass)	cdf
continuous	pdf (d = density)	cdf

A probability mass function (pmf) is a function that gives the probability that a discrete random variable is exactly equal to some value. A probability mass function differs from a probability density function (pdf) in that the latter is associated with continuous rather than discrete random variables; the values of the latter are not probabilities as such: a pdf must be integrated over an interval to yield a probability.

Discrete probability function

Suppose S is a finite or countably infinite sample space. Let p be a real-valued function defined for each element of S such that: pg 119
(a)

$$0 < p(s) \text{ for each } s \in S$$

(b)

$$\sum_{s \in S} p(s) = 1$$

Then p is said to be a *discrete probability function*

Once $p(s)$ is defined for all s, then you can say the probability of any event A is the sum of the probabilities of the outcomes comprising A:

$$P(A) = \sum_{s \in S} p(s)$$

This function satisfies the probability axioms of Section 2.3.
Note: you can still have an infinite number of outcomes in the sample space, as long as the probability of all outcomes sums to one. E.g. probability of getting heads on an odd numbered coin toss has an infinite number of events but if you sum the two sums (got it on odd, got it on even), you get a sum of 1. pg 120

Discrete random variable

3 Lecture Axioms: 10/14: $\Omega \to \mathcal{X} \subseteq$ R
(The sample space is $\mathcal{X}$, which is real.) Pg 119 has a similar set of facts, that are presented a little differently.

$$P(X = x) \geq 0, \quad P(X = x) > 0 \text{ if } x \in \mathcal{X}$$
$$P(\Omega) = 1, \text{ so } \sum_{x \in \mathcal{X}} P(x = \mathcal{X}) = 1$$

E_1, E_2, ..., E_k are disjoint (non-overlapping).

$$P(E_1 \cup E_2 \cup \cdots \cup E_k) = \sum_{i=1} P(E_i); P(x \in B) = \sum_{\mathcal{X} \in B} P(x = \mathcal{X})$$

We have done this. E.g. P(X is 2, 3, or 4) = $P(X = 2) + P(X = 3) + P(X = 4)$

A function whose domain is a sample space S and whose values form a finite or countably infinite set of real numbers is called a **discrete random variable**. We denote random variables by uppercase letters, often X or Y. page 123
Example: sum of the values of two die faces. E.g. value of die 1 is X_1, value of die 2 is X_2, value of sum is $X = X_1 + X_2$

The Probability Density Function. Associated with every discrete random variable X is a *probability density function* (or *pdf*), denoted $p_x(k)$ where

$$p_X(k) = P(s \in S \mid X(s) = k)$$

That's often written as $p_X(k) = P(X = k)$.
Note: the binomial distribution is such an example: (So is hypergeometric.)

$$p_x(k) = P(k \text{ successes}) = \binom{n}{k} p^k (1-p)^k, \text{ for } k = 0, 1, \ldots, n$$

Cumulative Distribution Function

(Discrete) You might want
$P(x \leq X \leq t) = P(X \leq t) - P(X \leq s - 1)$ pg 127
Cumulative distribution function: Let X be a discrete random variable. For any real number t, the probability that X takes on a value $\leq t$ is the *cumulative distribution* (cdf) of X [written $F_X(t)$].

$$F_X(t) = P(s \; in S \mid X(s) \leq t)$$

or more simply

$$F_X(t) = P(X \leq t)$$

then 10/19 lecture

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

Continuous random variables

A probability function P on a set of real numbers S is called **continuous** if there exists a function $f(t)$ such that for any closed interval $[a, b] \subset S$, $P([a, b]) = \int_a^b f(t)dt$.
Kind of obvious, but all values of the function must be more than zero.

Continuous probability density functions

Density functions are **not a probabiliy**. (11/4 lecture)

(pg 135) Let Y be a function from a sample space S to the real-numbers (takes values of real numbers; put in something and you get out a real number). The function Y is called a *continuous random variable* if there exists a function $f_Y(y)$ such that for any real numbers a and b with $a < b$:

$$P(a \leq Y \leq b) = \int_a^b f_Y(y)dy$$

The function $f_Y(y)$ is the ***probability density function (pdf)*** for Y. Think of "density" as corresponding to how much height there is over the x-axis when you plot $f_Y(y)$ versus y.

The value of f can be greater than 1. See 10/28 lecture. If $f_X(x) = c$ for $0 \leq x \leq 1/2$ then the area has to = 1 so $c = 2$. So $p_X(x) = 2$.

Continuous cumulative distribution functions

As in the discrete case, the *cumulative distribution function (cdf)* is defined by $F_Y(y) = P(Y \leq y)$: (pg 136)

$$F_Y(y) = P(Y \leq y) = \int_{-\inf}^{y} f_Y(t)dt$$

Also written as (Definition 3.4.3: pg 137)

$$F_Y(y) = \int_{-\inf}^{y} f_Y(r)dr = P(s \in S \mid Y(s) \leq y) = P(Y < y)$$

The cdf in this case is an integral of $f_Y(y)$. Note that now we have $f_Y(t)$ or $f_Y(r)$, not $f_Y(y)$. We are still integrating over the x-axis (y) but we can't have y in dy and in the integration limits.
Also note the derivative of the cdf is the pdf:

$$\frac{d}{dy} F_Y(y) = f_Y(y)$$

Indepenence

If X, Y are independent, $P_{X,Y}(x, y) = P_X(x) \cdot P_Y(y)$ for all x, y. Note there are cases when $E(XY) = E(X)E(Y)$ but this does **not** imply independence. (See Lec 11/4)

Expected Values

Let X be a discrete random variable with probability function $p_X(k)$. The *expected value* of X is denoted $E(X)$ (or sometimes μ or μx and is given by: pg 140
Discrete:

$$E(X) = \mu = \mu x = \sum_{\text{all } k} k * p_X(k)$$

Random:

$$E(Y) = \mu = \mu y = \int_{-\inf}^{\inf} y * f(y)dy$$

binomial: pg 141, 10/16 TA lecture

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

hypergeometric: for selecting n balls from r red balls and w white ballspg 143, 10/16 TA lecture

$$E(X) = \frac{rn}{r+w}$$

turns out the same if you substitute the proportion of red balls:

$$E(X) = n \cdot p \text{ for } p = \frac{r}{r+w}$$

The median

If X is a discrete random variable, the median, m, is that point for which $P(X < m) = P(X > m)$. In the event that $P(X \leq m) = 0.5$ and $P(X \geq m')$, the median is defined to be the arithmetic average, $(m + m')/2$.
If Y is a continuous random variable, its median is the solution to the integral equation $\int_{-\inf}^{m} f(y)dy = 0.5$
If a random variable's pdf is symmetric, both μ and m will be equal.

Variance

(pg 157; Sect3.6) The variance of a random variable is the expected value of its squared deviations from μ. Visually, it is the moment of inertia. Always positive. -E.T. MT2 review

Discrete: $Var(X) = \sigma^2 = E[(X-\mu)^2] = \sum_{\text{all } k} (k-\mu)^2 \cdot f_X(k)$

Continuous: if Y is continuous with pdf $f_Y(y)$:
$\text{Var}(Y) = \sigma^2 = E[(Y-\mu)^2] = \int_{-\infty}^{\infty} (y-\mu)^2 \cdot f_Y(y)dy$

Discrete or continuous: $Var(W) = E(W^2) - (E(W))^2 = E(W^2) - \mu^2$

Variances always add (independence not required). And,
$Var(aX + bY) = a^2 \cdot Var(X) + b^2 \cdot Var(Y)$ (page 189)

Standard Deviation: σ. Square root of variance.

Misc.

$F_X(x) = P(X \leq x)$
$p_X(x) = P(X = x) = P(X \leq x) - P(X < x)$ (discrete) (lec 10/28)

PDFs can have values greater than 1.

Multiple random variables

Double sigma notation: If you have two $\sum$s, you have a 2D grid of values. (simple video). If the two dimensions start and stop at numbers (or maybe ∞) then you have a value in every position of the grid. There can be cases, however, where the values of the two variables are interdependent (example).

Example with coefficients that depend on both indices (hasn't come up yet): Consider real numbers a_{ij}, where i ranges from 1 to 3, and j, from 1 to 2. We thus have the following equality:

$$\sum_{i=1}^{3}\sum_{j=1}^{2} a_{ij} = \sum_{i=1}^{3}(a_{i1} + a_{i2}) = (a_{11} + a_{12}) + (a_{21} + a_{22}) + (a_{31} + a_{32}).$$

Vocab:
joint density - pdf for a joint distribution.
univariate distribution - pdf of distribution with exactly 1 random variable
marginal distribution -

If you have a joint distribution with 2 random variables, the two marginal distributions are univariate distributions

See Elizabeth's single-page PDF for lots of formulas. Key ones: 10/28 lecture

One Dimension: $P(a \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$
Two Dimensions: $P(a_1 < x \leq a_2, b_1 < y \leq b_2) =$
$F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1)$
Marginal CDFs: If you have $F_{X,Y}(a, b)$,
$F_X(a) = P(X \leq a, y < \infty) = F_{X,Y}(a, \infty)$
See Elizabeth's PDFs for a lot more formulae.

Expected values for functions of random variables

NOTE: Expectations by default don't have X, Y in them.
(pg 150; Lecture 10/14) Suppose X is a discrete random variable with pdf $p_X(k)$. Let $g(X)$ be a function of X. Then the expected value of the random variable $g(X)$ is given by:

$$E[g(X)] = \sum_{\text{all k}} g(k) \cdot p_X(k)$$

Provided that $\sum_{\text{all k}} |g(k)| \cdot p_X(k) < \text{inf}$

If Y is a discrete random variable with pdf $f_Y(y)$, and if $g(Y)$ is a continuous function, then the expected value of the random variable $g(Y)$ is:

$$E[g(Y)] = \int_{-\text{inf}}^{\text{inf}} g(y) \cdot f_Y(y)dy$$

Provided that $E[g(Y)] = \int_{\text{inf}}^{\text{inf}} |g(y)| \cdot f_Y(y)dy < \text{inf}$

These two definitions highlight the difference in nomenclature. See page 127 for definition of discrete cdf. See page 135 for definition of continuous cdf

R.Var.	x-axis	pdf notation	cdf notation
discrete X	k	$p_X(k)$ or $P(X = k)$	$F_X(k)$ or $P(X \leq k)$
continuous Y	y	$f_Y(y)$	$F_Y(y)$ or $P(Y \leq y)$

R.Var.	pdf	cdf
discrete X	$p_X(x) = P(X = x)$ $= P(X \leq x) - P(X < x)$	$F_X(x) = \sum_{z \leq x} p_X(z)$
continuous Y	$p_X(x) = 0$ (at one point)	$F_Y(y) = \int_{-\infty}^{y} f_Y(y)$

NOTE: Need to have intuitive understnading that Ps are probabilities and Fs are cdfs. Can't just assume capitol letters are cdfs. E.g. if $Y = X^3$ then the Ps in $P(Y = y) = P(X^3 = y) = P(X = y^{1/3})$ are probability, not cdf. Note that there is no subscript in this case.

Say your random variable is the number on a die face. Say your function of that random variable is the square of that number. If you want the expectation of the square, you can sum the products of the squares with the probabilities of the random variable values.

Theorem (10/14 Lecture & similar on pg 150):
let $Y = g(X)$. Since:

$$\mathbb{E}(Y) = \sum_y yP(Y = y)$$
$$= \sum_y y \sum_{x:g(x)=y} P(X = \mathcal{X})$$

note: above is a sum over the y values,

and a sum over the xs that can give those y values.

There can be multiple xs that give a particular y.

$$= \sum_y \sum_{x:g(x)=y} g(x)P(X = x)$$

Drop y sum b/c now everything is in terms of x

You are summing over all the events in a set that

are indexed by y. But when you take the sum over y,

you say let's sum over all the different outcomes.

$$= \sum_{x \in \mathcal{X}} g(x)P(X = x) = \sum_{x \in \mathcal{X}} g(x)P(\mathcal{X})$$

This is a sum over all the xs in the whole set of outcomes ($\mathcal{X}$)

Note that this works for $g(Y) = Y^2$. $E(Y^2) = \int_{-\text{inf}}^{\text{inf}} Y^2 \cdot f_Y(y)dy$.
10/21 review lecture
Loop over all the values the random variable takes (all xs in $\mathcal{X}$), and sum (probability of that x)*(the function applied to that x).

Addition of functions of random variables Adding sums of expectations always works; doesn't require independence. (10/14 lecture)

$$\begin{aligned}\mathbb{E}(g_1(x) + g_2(x)) &= \sum_{x \in \mathcal{X}} (g_1(x) + g_2(x))p(x) \\ &= \sum_{x \in \mathcal{X}} g_1(x)p(x) + \sum_{x \in \mathcal{X}} g_2(x)p(x) \\ &= \mathbb{E}(g_1(x)) + \mathbb{E}(g_2(x))\end{aligned}$$

Simplest case: $E(X + Y) = E(X) + E(Y)$, valid even if X is not statistically independent of Y.

Example $\mathbb{E}(a(x) + b)$:

$$\begin{aligned}\mathbb{E}(a(x) + b) &= \mathbb{E}(a(x)) + \mathbb{E}(b) \\ &= a\mathbb{E}(x) + b\end{aligned}$$

Example $\mathbb{E}(x(x-1))$:

$$\begin{aligned}\mathbb{E}(x(x-1)) &= \mathbb{E}(x^1 - x)) \\ &= \mathbb{E}(x^2) - \mathbb{E}(x) \\ &\text{this might be easier to evaluate}\end{aligned}$$

Example $\mathbb{E}(X^2 + Y^2)$: HW 5, Q 3.9.10

$$\mathbb{E}(X^2 + Y^2) = \mathbb{E}(X^2) + \mathbb{E}(Y^2)$$

Examples

Different ways shapes can come up:
Probability of landing below some x value within a weird 2D shape.
10/19 lecture
If you have a diamond with points on the X and Y axes, but the top/bottom is cur off, what is probability of landing at some x? (What is $f_X(x)$?) You can't just derive something because X is less probable as you go out to the points. Use the classic trick: start with the CDF.
v $F_X(x) = P(X \leq x) = 1 - F_X(x) = P(X \geq x)$

Probability of a subset of events.
* E.g. $Y > 2X$. If you have a $f_{X,Y}(x, y)$ and you want to know what the probability is that $Y > 2X$ you can draw the shape (which should be smaller than the normal ranges of X and Y) and integrate both variables.
Integrate one using its two obvious limits and the other one using the relationship to the other variable. (E.g. Example 3.7.4 pg 165).
Or if the distribution is uniform for the range of X, Y your $P(X, Y$ in some subset of X and Y values) = fraction of the area.

A shape imposing restrictions on allowable values of X or Y.
(Indicator function multiplied on)
* E.g. $f_{X,Y}(x, y) = 1/x$, for $0 \leq y \leq x \leq 1$ (problem 3.7.20b). If you want $f_X(x)$ or $f_Y(y)$ you have to use the $y < x$ relationship in the integral bounds. Be careful with the integral bounds when getting cdf Fs and expectations, too!
* Or, problem 3.7.22 has $f_{X,Y}(X, Y) = 2e^{-x}e^{-y}$ but it is only valid for $y > x$ and 0 otherwise. Here the indicator function is ruining the independence. I.e. the allowable shape is affecting the sample space. This will affect the CDF (e.g. integrate y from 0 to x), and the expectation. It will also affect the marginal probabilities for either X or Y. Use the $y < x$ bound (or $x > y$ bound) in each integration.

A shape that represents a function of one or more random variables.
** E.g. area of triangle formed by (X,0), (0,Y), (0,0) for X, Y uniform on [0,1]. In this case the shape is only representing an aggregation of the variables. It does *not* represent restricting values of either (e.g. $X < Y$). Then $W = (1/2)XY$. X,Y uniform $\rightarrow$
$E(XY/2) = (1/2) \cdot \mathbb{E}(X)\mathbb{E}(Y)$. What is $f_{X,Y}(x,y)$ in this case? Independence allows us to know $f_{X,Y}(x,y) \propto f_X(x)f_Y(y)$ since the limits don't affect each other (both valid over [0,1]).

A shape that represents a function of one or more random variables.
NO. this isn't a shape. It is a line. * $Y = X^2$ Here we could draw a line $Y = X^2$. If we assume X is uniformly distributed, $f_Y(y) = c \cdot y^2$.
** Now say X is not uniform. Instead $f_X(x) = 3x^2$ for $0 < x < 1$ then $F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(X \leq y^{1/2}) = F_X(y^{1/2})$. We don't have $F_X(x)$ but can get it; it is x^3. Then
$F_Y(y) = F_X(y^{1/2}) = (y^{1/2})^3 = y^{2/3}$ and $f_Y(y) = \frac{2}{3}y^{1/3}$ We can get $\mathbb{E}(Y)$ two ways now.
(Way 1:) This is always true: $E(Y) = E(X^2)$. We can use independence so $E(Y) = E(X^2) = (E(X))^2$. And
$E(X) = \int_0^1 x \cdot f_X(x) = \int_0^1 x \cdot 3x^2 = 3/4$. Square that to get 9/16.
(Way 2:) Or, $E(g(x)) = \int_0^1 g(x) \cdot f_X(x) = \int_0^1 x^2 \cdot 3x^2 dx$

– ?? A shape restricting values of one or more variable(s). Say

Joint Density

$F_{X,Y}(a,b) = P(x \leq a \cap y \leq b)$.
The $\cap$ is often implicit: $P(x \leq a, y \leq b)$ (10/28 lecture)

To get marginal: $F_X(a) = P(x \leq a) = P(X \leq a, y \leq \text{inf})$ (10/28 lecture)
$= F_{X,Y}(a, \text{inf})$

If you are given $f_X(x)$ and $f_Y(y)$ you can't know the joint distribution *unless* you are told they are independent. There are many univariate distributions that can lead to a particular joint distribution if the two variables are not independent. (11/3 TA lecture)

Independence of random variables. (page 175)
Basic: Events E_1 and E_2 are independent iff
$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$

For Random Variables: $E_1 \equiv \{X \in A\}$, $E_2 \equiv \{Y \in B\}$. Then if for *all* subsets of A and B, $P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$ But if you had to check for every single A nd B it wouldn't be very useful.
So, **Discrete Case**: 10/30 lecture

$$\text{Take } A = \{X\}, B = \{Y\}$$
$$X, Y \text{ independent: } P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$
$$P(X \in A, Y \in B) = \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} p_{X,Y}(x,y)$$
$$= \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} p_X(x) \cdot p_Y(y)$$
$$= (\sum_{x \in \mathcal{A}} p_X(x))(\sum_{y \in \mathcal{B}} \cdot p_X(y))$$
$$= P(X \in A) \cdot P(Y \in B)$$

10/30 Independence facts (10/30 lecture)
Probabilities don't depend on each other:
$P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y)$
CDFs don't depend on each other (pg 174):
$F(X \leq x, Y \leq y) = F(X \leq x) \cdot F(Y \leq y)$
Also: $f_{X,Y}(x,y) \propto g_1(x)g_2(y)$ **BUT** limits that tie X and Y can make this break. Recall the 10/30 example: $0 < y < x < 1$ for $f_{X,Y}(x,y) = cxy$. cxy looks separable; be careful!

notation example:
$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = g_1(x_1)g_2(x_2)\ldots g_n(x_n)$.

Example: if X_1, X_2, and X_3 are independent random variables each with pdf $f_{X_i}(x_i) = 4x_i^3$. Then
$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = (4x_1^3)(4x_2^3)(4x_3^3) = 4^3x_1^3x_2^3 4x_3^3$

Example: (10/30 lecture)
$f_{X,Y} = c \cdot x \cdot y$ for $0 < x < 1$ and $0 < y < 1$. These are independent because you can pull the joint probability apart into single probabilities.
But if you have $f_{X,Y} = c \cdot x \cdot y$ for $0 < y < x < 1$ they are not independent. You can tell right away because the bounds are dependent on each other. You can also draw the $y = x$ line and shade the $0 < y < x < 1$ region to see.
Erick: "Another way to say it is to remember that the function itself is c x y in some region and zero otherwise, and so the function itself is not a product."

memoryless property. You are waiting for a phone call. The probability of the phone ringing in the next minute is constant and doesn't depend on how long you have been waiting. More formally: X = phone call time (?)
$P(X \geq s + t \mid x \geq s) = P(x \geq t)$

Combining random variables Example from 10/26 TA review lecture: Say you have 50 random variables that are uniform on $[0,1]$. $f_X1 = f_X2 = \ldots f_X50$ and independent. The joint probability density function of these random variables is the product of the PDFs of the individual random variables. The cdf is
$F_X(x) = \{0 \text{ if } x < 0, x \text{ if } x \in [0,1], 1 \text{ if } x > 1\}$. Define $Y \equiv \max(X_i)$. Since they are independent, we know

$$P(Y \leq y) = P(X_1 \leq y \text{ and } X_2 \leq y \text{ and } \ldots \text{ and } X_{50} \leq y)$$
$$\text{independence allows multipication}$$
$$= P(X_1 \leq y) \cdot P(X_2 \leq y) \cdot \cdots \cdot P(X_{50} \leq y)$$
$$= F_{X_1}(y) \ldots F_{X_{50}}(y) = y^{50}$$
$$P(Y \leq y) = y^{50}$$

differentiate to convert from F_Y to f_Y: $f_Y(y) = 50y^{49}$

Example: Getting the expectation of one variable from a joint distribution (10/30 lecture)
If you have $f_{X,Y}(x,y)$, and want $E(X)$, you can get $f_X(x) = \int_y f_{X,Y}(x,y)dy$.
Then You can get $E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)dx$.
You can also get $F_{X,Y} = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v)dudv$
Then you can get $F_X(x) = F_{X,Y}(x, y = \infty)$

Transforming/Combining Random Variables

(11/2 TA lecture) X and Y are independent: $f_X(x) = \frac{1}{2\pi}exp(-x^2/2)$, $f_Y(y) = \frac{1}{2\pi}exp(-y^2/2)$.
Question: If $U = X^2 + Y^2$, what is $F_U(u)$?
Answer: $f_{X,Y}(x,y) = f_X(x)f_Y(y) = \frac{1}{2\pi}exp((-1/2)(x^2+y^2))$

Location/Scale

$Y = X + b$**:**

$$F_Y(y) = P(Y \leq y)$$
$$= P(X \leq y - b)$$
$$= F_X(y - b)$$
$$\text{differentiate}$$
$$\mathbf{f_Y(y) = f_X(y - b)}$$

?? Do you see why we are subtracting c, not adding it?
$f_{x+c}(t) = f_X(t - c) = P(X = t - c) = P(X + c = t)$ (11/2 TA lecture)

Scaling Variables

$Y = aX$: and $a > 0$.

$$F_Y(y) = P(Y \leq y)$$
$$= P(aX \leq y)$$
$$= P(X \leq y/a)$$
$$= F_X(y/a)$$
$$\text{differentiate}$$
$$\mathbf{f_Y(y) = (1/a) \cdot f_X(y/a)}$$

Location and Scale

$Y = aX + b$: (and $a > 0$).

$$F_Y(y) = P(Y \leq y)$$
$$= P(aX + b \leq y)$$
$$= P(X \leq (y - b)/a)$$
$$= F_X((y - b)/a)$$
$$\text{differentiate}$$
$$\mathbf{f_Y(y) = (1/a) \cdot f_X((y - b)/a)}$$

Visualize the uniform distribution over $[0,1]$. When you have $W = aX + B$, you scale down so the height is now $1/a$ over $[0,a]$. Then you shift it to $1/a$ over $[b, b+a]$. (10/30 lecture; last page)

Example: (10/30 lecture) (??) Linear function of a normal distribution is a normal distribution. The mean and variance drop out. (??)

Tricks for adding and multiplying random variables

you can get mixures of fs, Ps, and Fs and still make the connection. Start with converting things to Fs, then use Ps, then move back to fs if desired. See dog eared notebook page of MT2 practice problems. And Lecture 10/30. This works great for location/scale. I'm not sure how applicable it is for R.V. addition.

example: add two variables. X and Y are uniform on $[0,1]$. Find pdf of $Z = X + Y$. Note that your range for Z is [0, 2]. Expect lots of probability around $Z = 1$, and zero at $Z = 0$ and $Z = 1$

$$F_Z(z) = P(Z \leq z) = P(X + Y \leq z)$$
$$\text{use conditional probability \& independence of X \& Y}$$
$$= \int_{-\infty}^{\infty} P(X + Y \leq z \mid X = x) \cdot f_X(x)dx$$

Covariance

(11/4Lecture) Covariance is a measure of how much two random variables change together.
The sign of the covariance therefore shows the tendency in the linear relationship between the variables.

The magnitude of the covariance is not easy to interpret.
Variance is a special case of the covariance when the two variables are identical.
The normalized version of the covariance, the correlation coefficient, however, shows by its magnitude the strength of the linear relation.

Recall that Var (variance; σ^2) is $E[(X-\mu)^2]$. And that $E[(X-\mu)] = 0$. (pg 156)

$\text{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)]$
If you multiply it out: $\text{Cov}(X,Y) = (\mathbb{E}(XY)) - (\mathbb{E}X)(\mathbb{E}Y)$ (pg 189)

If X and Y are independent, $\text{Cov}(X,Y) = 0$. Of course you can have $Cov(X,Y) = 0$ without independence (page 189).

$\text{Var}(X+Y) = \mathbb{E}((X+Y)^2) - \mathbb{E}((X+Y))^2$
$= Var(X) + 2Cov(X,Y) + Var(Y)$. If X, Y are independent, $Var(X+Y) = Var(X) + Var(Y)$.

Covariance versus Variance: A covariance refers to the measure of how two random variables will change together and is used to calculate the correlation between variables. The variance refers to the spread of the data set how far apart the numbers are in relation to the mean, for instance.

Cov$(X,X) =$ **Var**(X). Midterm 2 practice problem.
Because $\text{Cov}(X,X) = (\mathbb{E}(XX)) - (\mathbb{E}X)(\mathbb{E}X)$. We can solve for $\mathbb{E}(XX)$ using the fact that correlation = 1 and
$\rho(X,X) = \text{cov}(X,Y)/\sqrt{\text{var}(X)\text{var}(X)} = \text{cov}(X,X)/\text{var}(X)$.
$1 = \rho(X,X) = \text{cov}(X,X)/\text{var}(X)$ so **cov**$(X,X) =$ **var**(X).

Correlation

(Lec. 11/4) Correlation: $\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Cov}(X)\text{Cov}(Y)}}$ Bounded by $-1 \le \rho \le 1$
When is it 1 or -1? When $\text{Var}(X-bY) = 0$. Var = 0 when the probability mass is all at one point, not spread out.
This happens when X is a linear function of Y.

Expectation, Variance, Covariance, and Correlation

Scaling & shifting:
$E(aX+b) = aE(X) + b$
$Var(aX+b) = a^2Var(X)$ (b doesn't matter. Squared is b/c Var = σ^2; σ scales w/ a so you get a^2)
$Cov(aX+b, Y) = a \cdot Cov(X,Y)$ (Has units, so keep factor of a. b doesn't matter) Also: $Cov(aX+b, cY+d) = a \cdot c \cdot Cov(X,Y)$
$Corr(aX+b, Y) = Corr(X,Y)$ Use properties above to see. Also, dimensionless so a can't be there.

Cov, and Corr with itself: Note that you can't have E or Var of something w/ itself: univariate.
$Cov(X,X) = Var(X)$ From the definition of Variance.
$Corr(X,X) = 1$

Cov, and Corr with an independent var. Say X, Y are independent.
$E(XY) = E(X)E(Y)$ Can get there from a g(X,Y) idea. g(X,Y) is xy. Then $E(XY) = \int\int g(X,y) \cdot f_{X,Y}(x,y) = \int\int xy \cdot f_{X,Y}(x,y) = \int x \cdot f_X(x) \int y \cdot f_Y(y)$.
$Cov(X,Y) = 0$ and $Cov(aX, bY) = 0$

$Corr(X,Y) = 0$

With constants:
$E(a) = aE(1) = a$
$Var(a) = a^2Var(1) = 0$ ($= E[(a - E(a))^2] = E[(a-a)^2] = E[0] = 0$)
$Cov(a,Y) = a \cdot Cov(1,Y) = a \cdot 0 = 0$ (E((1 - E(1))) = 0. Also Y isn't dependent on a.
$Corr(a,Y) = 0$ But it doesn't really make sense to talk about corr w/ constant. Lec 11/4

Bernoulli

Wikipedia: The probability distribution of a random variable which takes the value 1 with success probability of p and the value 0 with failure probability of q=1-p. It can be used to represent a coin toss where 1 and 0 would represent "head" and "tail" (or vice versa), respectively. In particular, unfair coins would have $p \neq 0.5$.
Things based on Bernoulli:

- The **inter-arrival time**
- The **negative binomial distribution** (with stopping parameter n and success parameter p).
- The **binomial distribution** counts the number of events in a number of trials.

Bernoulli relation to Bionomial

"A success/failure experiment is also called a Bernoulli experiment or Bernoulli trial; when n = 1." (Wikipedia)

Later: "The Bernoulli distribution is a special case of the binomial distribution, where n = 1. Symbolically, X B(1, p) has the same meaning as X Bern(p). Conversely, any binomial distribution, B(n, p), is the distribution of the sum of n Bernoulli trials, Bern(p), each with the same probability p." (Wikipedia)
Also, recall that as n gets large but p is fixed $\frac{X-np}{\sqrt{np(1-p)}}$ is approx N(0,1). Below, if Bernoulli n gets large and p get small you get Poisson.

Bernoulli relation to Geometric

(11/20 Lec). Independent trials, each with probability P(success) = p.
W = number of trials to first success.

$$P(W > k) = P(0 \text{ successes in k trials})$$
$$= (1-p)^k$$

Now to get W=k we subtract b/c discrete (like differentiating)

Note: subtracting the bigger number of events from the smaller

$$P(W = k) = P(W > k-1) - P(W > k)$$
$$= (1-p)^{k-1} - (1-p)^k$$
$$= (1-p)^{k-1}[1-(1-p)]$$
$$\mathbf{P(W = k) = (1-p)^{k-1} \cdot p}$$

i.e. Geom(P), mean (1/p)

Geometric also has forgetting property: $P(W > k+m | W > k) = \frac{P(W>k+m)}{P(W>k)} = \frac{(1-p)^{k+m}}{(1-p)^k} = (1-p)^m = P(W > m)$

Geometric and exponential are the discrete/continuous analogs. Expo w/ parameter λ has mean $1/\lambda$. Geom(p) has mean $1/p$. Don't do continuity correction on exponential. (11/20 Lec) $\mathbb{E}(W) = 1/p$ and Var(W)=$\frac{1-p}{p^2}$

Number of failures before 1st success: (11/20 Lec)
W* = number of failures before 1st success. Same pdf, just slightly different notation. W* = W - 1, where
$P(W = k) = (1-p)^{k-1}p, k = 1, 2, \ldots$
Then $P(W* = k) = P(W = k+1) = (1-p)^k p$
$\mathbf{P(W^* = k) = (1-p)^k p}$,
$\mathbb{E}(W^*) = 1/p - 1 = \frac{1-p}{p}$ Var(W*)=$\frac{1-p}{p^2}$ = Var(W)

Number of failures before r^{th} success: (11/20 Lec)
W_r^* = number of failures before r^{th} success. $W_r^* = 0, 1, 2, \ldots$
$\mathbb{E}(W_r^*) = r/p - r = r(1-p)/p$
$\text{Var}(W_r^*) = r(1-p)/p^2 = \text{Var}(W_r)$
Subtracting shifts mean but not variance.

Poisson process

TA definitions:

- *E didn't like this* A random set s c[0,T] is distributed according to PP(λ) if s = $\{t_1, t_2, \ldots, t_n\}$ and n is Po(λT). Longer time and higher intensity lead to more events. There are no other requirements. ????
- *There was another but it used vocabulary/concepts we haven't done*

TA notes: Note that the busses coming at random times are modeled as Poisson processes. Can take any bus therefore some aggregation of these. Sums of independent Poisson are Poisson. Unions of Poisson are Poisson. (TA 11/30 lecture)
Things based on Poisson process:

- exponential

Internet tidbids:

- Several important probability distributions arise naturally from the Poisson process - the Poisson distribution, the exponential distribution, and the gamma distribution.
- In some sense, the Poisson process is a continuous time version of the Bernoulli trials process

"Events happening randomly and independently over time at rate λ" (11/18 Lec)
In a very small time interval of length h:
*P(0 events) = $1 - \lambda h$
*P(1 events) = λh
*P(> 1 event) ≈ 0
Divide time from 0 to s up into K bits of length h. K is large, h is small. $Kh \equiv s$ (they add up).
The number of events is binomial. And, in each interval event or not is Bern(λh).
$\text{Bin}(K, \lambda h) \approx \text{Poisson}(K\lambda h) \cong \text{Poisson}(\lambda s)$. (11/18 Lec)
(Her s appears to be equal to book's T.)

B(20,000, 1/2,000) $\approx$ Poisson (20,000/2000) = Poisson (10). (11/20 Lec)

Birthday example
First note that P(0 pairs) = $(365 \cdot 364 \cdot \ldots 335)/365^{31} = 0.2695$
Without Poisson:
1a. Derive probability of pairs. For 31 students, there are $\binom{31}{2} = 465$ pairs.
2a. Expected number of pairs: 465/365 = 1.274
With Poission:
* Note large n, small p. n = 465, p = 1/365. So λ = np = 465/365.
* Note that pairs are not quite independent. (see notes).
But, P(X=0) = $e^{-\lambda} = e^{-}1.274 = 0.279$. Close to true answer above.
If you want $P(X > 2)$, do
$1 - P(X=0) - P(X=1) = 1 - \mu^0 * e^{-\mu}/0! - \mu^1 * e^{-\mu}/1! = 0.364$
Note that we don't have time in the framework of the problem; we aren't waiting for busses.
See notes for P(3 people share). n is large and p is smaller, so Poisson a better approximation.

Adding Poisson. If s_1 and s_2 are non-overlapping time intervals and $X_1 = Po(\lambda s_1)$ and $X_2 = Po(\lambda s_2)$ then $X_1 + X_2$ = total number of events in times s = $s_1 + s_2$ then $X_1 + X_2 = Po(\lambda s)$. And if you sum up n X_is as Y, by CLT $\frac{Y - n\mu}{\sqrt{n\mu}}$ is approx N(0,1). (11/20 Lec)
?? What if your lambdas are different?

Poisson Distribution

Like binomial with $\lambda = np$ and $n \to \infty$ (TA 11/30 lecture)
One of the distributions that you can derive from the Poisson process. – Erick.
The Poisson distribution is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event.

PMF: $\frac{\lambda^k}{k!}e^{-\lambda}$
$f(k;\lambda) = \Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!}$
$\lambda = \mathrm{E}(X) = \mathrm{Var}(X)$ (Wikipedia)
Mean = variance. $Var(X) = \mu$ (Lec. 11/18)

If Binomial n is large but p is small such that $np = \mu$ then
$E(X) = np = \mu$ and $\mathrm{Var}(X) = n \cdot p(1-p) = \mu(1-p) \approx \mu$ (Lec. 11/18)

Wikipedia: A **discrete** probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.
Can get the average number of events per time; that's λ. Then you can calculate the probability of getting, say, 10 events in that time span.

Relation to Bernoulli: If the events are rare, each event is Bernoulli.
Notes:
* distribution on nonnegative integers (11/18 Lec)
* doesn't make sense to scale/shift the Poission (11/18 Lec)

Book's 3 formulas for Poisson (pg 232):
$p_X(k) = e^{-np}\frac{(np)^k}{k!}$
$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$
$p_X(k) = e^{-\lambda T}\frac{(\lambda T)^k}{k!}$
Note that $np = \lambda$
We derived $P(X = k) = \frac{\mu^k}{k!}e^{-\mu}$ in 11/18 Lecture. From Binomial as N large but p small.

Adding Poissons

(11/25 TA Lecture) If $X \sim Po(\lambda_X)$ and $Y \sim Po(\lambda_Y)$ and X and Y are independent then $X + Y \sim Po(\lambda_X + \lambda_Y)$. Written again on 12/2 as: $X \sim Po(\mu)$, $Y \sim Po(\nu)$, $X + Y \sim Po(\mu + /nu)$

(11/20 Lec) If you have two non-overlapping time intervals S_1 and S_2 ($S = S_1 + S_2$) then they are independent because they are non overlapping. Say $X_1 = Po(\lambda S_1)$ and $X_2 = Po(\lambda S_2)$. Then $X_1 + X_2$ events in time $S = S_1 + S_2$ is Po(λS).
In general for independent $X_i \sim Po(\lambda S_i)$, $\sum_1^k X_i$ is Poisson with mean $\lambda(S_1 + S_2 + \ldots S_k)$.
Similarly if $X_1 + \cdots + X_n$ are independent Po(μ), $Y = \sum_1 nX_i$ is Poisson with mean $n\mu$ and variance $n\mu$. So b the Central Limit Theorem $\frac{Y - n\mu}{\sqrt{(n\mu)}}$ is approx N(0,1).

Adding Poissons and knowing the total number

TA mumble on 12/7: If someone gives us info about a um of 2 R.V.s and asks us to say something about 1 R.V. then we use Binomial, not Poisson.
(Bus example from 11/25). If you know the number of red busses and blue busses is n, what is the probability that the number of red busses is k?

Each process is Po($\lambda_i t$) and independent. The total is Po($\sum \lambda_i t$).

$$\begin{aligned}
P(\#\text{ of red} = k | \#\text{ of red \& blue} = n) &= P(R = k | R + B = n) \\
&= \frac{P(R = k \cap B + R = n)}{P(R + B = n)} \\
&= \frac{P(R = k \cap B = n - k)}{P(R + B = n)} \\
&= \frac{P(R = k)P(B = n - k)}{P(R + B = n)} \\
&= \frac{[\frac{exp(-\lambda t)(-\lambda t)^k}{k!}][\frac{\exp(-\mu t)(-\mu t)^{n-k}}{(n-k)!}]}{\frac{\exp(-(\lambda+\mu)t)((\lambda+\mu)t)^n}{(n)!}}
\end{aligned}$$

lots cancels out

$$\begin{aligned}
&= \frac{n!}{k!(n-k)!}\frac{\lambda^k \mu^{n-k}}{(\lambda+\mu)^n} \\
&= \binom{n}{k}\frac{\lambda}{\lambda+\mu}(\frac{\mu}{\lambda+\mu})^k \\
&= Bin(n, \frac{\lambda}{\lambda+\mu})
\end{aligned}$$

Note: this binomial dist was also derived on 12/2
Addition: $\sum_x P_X(x \mid W = w) = \sum_x \frac{P(X=x, W=w)}{P(W=w)} = \frac{P(W=w)}{p(W=w)} = 1$

Special case of next 1 bus:
Given next bus is R or B, P(red) = $\lambda/(\lambda+\mu) = 3/(3+2) = 0.6$
Now say there is Red at 3/hr, Blue at 2/hr, and green at 4/hr. P(next bus green) = 4/(4+2+3) = 4/9. Don't need to re-derive the binomial.

If you want to know the probability that 10 of the next 100 buses are Red, use the Binomial theorem. If you want to know the probability that < 10 of the next 100 buses are Red, use the Central Limit Theorem approximation of the binomial. This was in HW9.
REVIEW IT

Given exactly 1 bus in the next t hours, <u>when</u> does it occur? Use CDF

$$\begin{aligned}
F_T(s) &= P(T \le s | \text{one event in}(0,t), \text{process rate } \lambda) \\
&= \frac{P(1 \text{ event in (0,s) and 0 in (s, t))}}{P(1 \text{ in (0,t))}} \\
&= \frac{\frac{\exp(-\lambda s)(\lambda s)^1}{1!}\frac{\exp(-\lambda(t-s))(\lambda(t-s))^0}{0!}}{\exp(-\lambda t)(\lambda t)^1/1!} \\
&= \frac{\exp(-\lambda s)(\lambda s)\exp(-\lambda(t-s))}{\exp(-\lambda t)(\lambda t)} \\
&= s/t
\end{aligned}$$

To get the density, differentiate. $f_T(s) = 1/$ for $0 \le s \le t$ and 0 otherwise. This was re-derived on 12/2.

Conditioning on total # with one process

(Bus example from 11/25). Green buses come at rate 4/hour.
What is P(4 green buses in next 15 minutes | 4 in next hour)?
= P(4 green buses in next 15 minutes $\cap$ 0 in 3/4 hour after)/P(4 in next hour).
Use Po(λ) with t = 1/4, t = 3/4, and t = 1.

$$\begin{aligned}
&= \frac{e^{-4(1/4)}(4\cdot 1/4)^4/4!}{e^{-4\cdot(3/4)}} \\
&= (1/4)^4.
\end{aligned}$$

(Computer bug example from 11/25).
Assume total number of bugs in a piece of software is Po(μ).
W = total # of bugs. X = total # of bugs detected, Y = # of bugs not detected yet.

$$\begin{aligned}
P(X = x, Y = y) &= P(X = x, W = x + y) \\
&= P(X = x \mid W = x + y) \cdot P(W = x + y)
\end{aligned}$$

(used conditional probability)

= Bin(x+y, p) times Po(w=x+y)

fill in Binomial and Poisson:

$$= \binom{x+y}{y} p^x (1-p)^y \cdot \frac{e^{-\mu}\mu^{x+y}}{(x+y)!}$$

simplify, but expand exponential: $\mu p + \mu(1-p) = \mu$

$$\begin{aligned}
&= \frac{1}{x!y!}(\mu p)^x(\mu(1-p))^y \exp(-(\mu p + \mu(1-p))) \\
&= \frac{(\mu p)^x e^{-\mu p}}{x!}\frac{(\mu(1-p))^y e^{-\mu(1-p)}}{y!}
\end{aligned}$$

that's a Poisson(x) and a Poisson(Y)

It factorizes & ranges are independent. Independent!

$$= P(X) \cdot P(Y)$$

where $X \sim Po(\mu p)$, and $Y \sim Po(\mu(1-p))$

You aren't likely to have less bugs just because you found one. How this is different from the bus problems: you knew the total of R and B = n, or that there <u>were</u> going to be 4 buses total but you don't know when they will come. *Ask Erick his summary?*

Questions: can I jump straight to Binomial for these Poissons?

Exponential

Waiting time for Poisson Process. Expo(λ)
"The exponential distribution (a.k.a. negative exponential distribution) is the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate." (Wikipedia)
In many respects, the geometric distribution is a discrete version of the exponential distribution. Constant rate and memoryless.

PDF and CDF.
PDF: $\lambda\exp(-\lambda x)$. Note: x = s = time.
CDF: $1 - \exp(-\lambda x)$. Note: x = s = time.
mean = expectation = λ, variance = $1/\lambda^2$

(11/18 Lec) **Waiting times in a Poisson process with rate λ.** pdf: $f_T(s) = \lambda\exp(-\lambda s)$ on $s \ge 0$ and 0 otherwise.

Comes from T being time to next event, $P(T > S)$ = P(0 events in time S) = P((Poisson mean λs) =0) = $exp(\lambda s)$.
<u>Falls right out of Poisson Distribution with k = 0.</u>
Has forgetting/memoryless property.
$P(T > t + s | T > t) = \frac{P(T>t+s \cap T>t)}{P(T>t)}$. That's vanilla conditional probability. $= \frac{P(T>s)}{P(T>t)} = \frac{exp(-\lambda(t+s))}{exp(-\lambda t)} = \exp(-\lambda s) = P(T > s)$

Min of 2 Exponentials

(11/30 TA Lec) If $Y_1 \sim expo(2/hr)$ and $Y_2 \sim expo(4/hr)$ and Y_1 and Y_2 are independent then $\min(Y_1, Y_2) \sim expo(6/hr)$

$P(Y_{min} > y)$

(11/30 TA Lec)

$$\begin{aligned} P(Y_{min} > y) &= P(Y_1 > y \text{ and } Y_2 < y) \\ &= P(Y_1 > y)P(Y_2 > Y) \\ &= exp(-\frac{4}{\text{hr}}y)exp(-\frac{2}{\text{hr}}y) \\ &= exp(-\frac{6}{\text{hr}}y) \end{aligned}$$

Adding Exponentials

(11/25 Lec, 26.5 of Wk8) The time to the first event is T_1. This is $X_1 =$ Expo(λ) with mean $= \mathbb{E}(T_n) = 1/\lambda$ and variance $1/\lambda^2$.
Now string a bunch together. T_n is the time to the n^{th} event, X_i is the time between the i^{th} and $i-1^{th}$ event. See diagram in 11/25 notes.
$T_n = \sum_1^n X_i$ where X_i are independent (? because times don't overlap).
$\mathbb{E}(T_n) = n\mathbb{E}(T_i) = n(1/\lambda)$
$\text{Var}(T_n) = n\text{Var}(X_i) = n/\lambda^2$
Want to get the PDF, but let's start with the CDF.

$$\begin{aligned} P(T_n \le t) &= P(\text{at least n events in (0,t)}) \\ &= \sum_{j=n}^{\infty} P(\text{J events in (0, t)}) \\ &= \sum_{j=n}^{\infty} \frac{e^{-\lambda t}(\lambda t)^j}{j!} \end{aligned}$$

differentiate (in principle) gives the gamma distribution

$$\begin{aligned} f_{T_n} &= \frac{d}{dt}(F_{T_n}(t)) \\ f_{T_n} &= G(n, \lambda) \text{ n is shape, } \lambda \text{ is rate} \end{aligned}$$

Gamma Distribution: probability of at least n events/successes in time t. (??) (Wikipedia:) a two-parameter family of continuous probability distributions. The common exponential distribution and chi-squared distribution are special cases of the gamma distribution.

Adding exponentials and knowing the total

(12/4 Lecture) $X \sim expo(\lambda)$, $Y \sim expo(\lambda)$. W = X + Y = time for both of them to happen. Note that X and W are not independent. Would be gamma with param 2 but just think of them as sum of 2 exponentials.
Previously derived: $(X|W = w)$ is $U[0, w]$. (PDF = constant 1/w).
So $P(X \le x \mid W = w) = x/w$ for $0 \le x \le w$. (The cdf)

Now let $V = \frac{X}{X+Y} = \frac{X}{W}$ and consider V and W.

$$\begin{aligned} P(V \le v \mid W = w) &= P(X/W \le v \mid W = w) \\ &= P(X \le vw \mid W = w) \\ &= \frac{vw}{w} \\ &= v \text{ for } 0 \le vw \le w \end{aligned}$$

differentiating:

$$f_V(v \mid W = w) = 1$$

W, V are independent, because neither function nor range depends on w.
Do again using conditional probability. (12/7 TA lecture)

$X, Y \sim expo(2)$. Define $W \equiv X + Y$ and look at $f_X \mid W(x, w)$.

$$f_{X|W}(x, w) \equiv \frac{f_{X,W}(x, w)}{f_W(w)}$$

$f_W(w)$: sum of 2 exponentials is gamma.

$$\begin{aligned} &= \frac{f_{X,Y}(x, w - x)}{4w \cdot \exp(-2w)} \\ &= \frac{2 \cdot \exp(-2w)2 \cdot \exp(-2w)}{4w \cdot \exp(-2w)} \\ &= \frac{1}{w} \text{ for } x < w, 0 < x \end{aligned}$$

still have to pay attention to range.

$X \mid W$ is U$[0, W]$.
It is a valid pdf for x since $\int_0^w 1/w dx = w/w = 1$

Conditional Statements & Random Variables

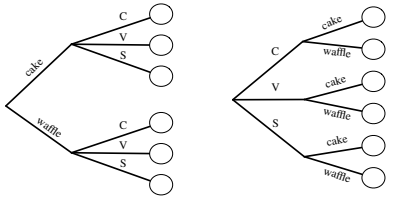
Continuous: $f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
Note: $f_{X|Y}(x \mid y) = f_{X|Y}(x) = f_X(x \mid Y = y)$
If X doesn't depend on Y: **(independence)**
discrete:
$P_X(x \mid Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_X(x)p_Y(y)}{p_Y(y)} = p_X(x)$
continuous:
$f_X(x \mid Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x)$
If we find $f_{X|Y}(x \mid y)$ does not depend on y, they X and Y are independent.

Conditional Expectation

Elizabeth: $E(Y) = E(E(Y \mid X))$. Erick:
$E(Y) = E(E(Y \mid X)) = E_X(E_Y(Y \mid X))$. $Y \mid X$ is described by $f_{Y|X}$
Integrate Y out first.

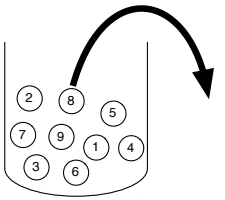
Counting

Multiplication Rule



Let's say we have a compound experiment (an experiment with multiple components). If the 1st component has n_1 possible outcomes, the 2nd component has n_2 possible outcomes, ..., and the rth component has n_r possible outcomes, then overall there are $n_1 n_2 \dots n_r$ possibilities for the whole experiment.

Sampling Table



The sampling table gives the number of possible samples of size k out of a population of size n, under various assumptions about how the sample is collected.

	Order Matters	Not Matter
With Replacement	n^k	$\binom{n+k-1}{k}$
Without Replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Naive Definition of Probability

If all outcomes are equally likely, the probability of an event A happening is:

$$P_{\text{naive}}(A) = \frac{\text{number of outcomes favorable to } A}{\text{number of outcomes}}$$

Thinking Conditionally

Independence

Independent Events A and B are independent if knowing whether A occurred gives no information about whether B occurred. More formally, A and B (which have nonzero probability) are independent if and only if one of the following equivalent statements holds:

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

Conditional Independence A and B are conditionally independent given C if $P(A \cap B|C) = P(A|C)P(B|C)$. Conditional independence does not imply independence, and independence does not imply conditional independence.

Unions, Intersections, and Complements

De Morgan's Laws A useful identity that can make calculating probabilities of unions easier by relating them to intersections, and vice versa. Analogous results hold with more than two sets.

$$\begin{aligned} (A \cup B)^c &= A^c \cap B^c \\ (A \cap B)^c &= A^c \cup B^c \end{aligned}$$

Joint, Marginal, and Conditional

Joint Probability $P(A \cap B)$ or $P(A, B)$ – Probability of A and B.

Marginal (Unconditional) Probability $P(A)$ – Probability of A.

Conditional Probability $P(A|B) = P(A, B)/P(B)$ – Probability of A, given that B occurred.

Conditional Probability *is* Probability $P(A|B)$ is a probability function for any fixed B. Any theorem that holds for probability also holds for conditional probability.

Probability of an Intersection or Union

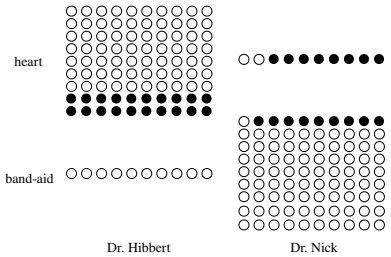
Intersections via Conditioning

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

Unions via Inclusion-Exclusion

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

Simpson's Paradox



It is possible to have

$$P(A \mid B, C) < P(A \mid B^c, C) \text{ and } P(A \mid B, C^c) < P(A \mid B^c, C^c)$$
$$\text{yet also } P(A \mid B) > P(A \mid B^c).$$

Law of Total Probability (LOTP)

Let $B_1, B_2, B_3, ...B_n$ be a *partition* of the sample space (i.e., they are disjoint and their union is the entire sample space).

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n)$$
$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n)$$

For **LOTP with extra conditioning**, just add in another event C!

$$P(A|C) = P(A|B_1, C)P(B_1|C) + \cdots + P(A|B_n, C)P(B_n|C)$$
$$P(A|C) = P(A \cap B_1|C) + P(A \cap B_2|C) + \cdots + P(A \cap B_n|C)$$

Special case of LOTP with B and B^c as partition:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$
$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Bayes' Rule

Bayes' Rule, and with extra conditioning (just add in C!)

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

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

We can also write

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

Odds Form of Bayes' Rule

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

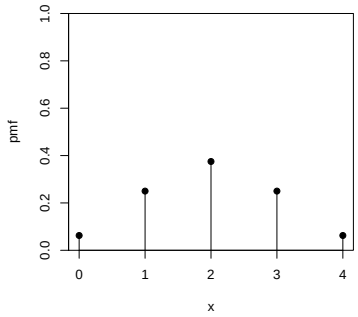
The *posterior odds* of A are the *likelihood ratio* times the *prior odds*.

Random Variables and their Distributions

PMF, CDF, and Independence

Probability Mass Function (PMF) Gives the probability that a *discrete* random variable takes on the value x.

$$p_X(x) = P(X = x)$$

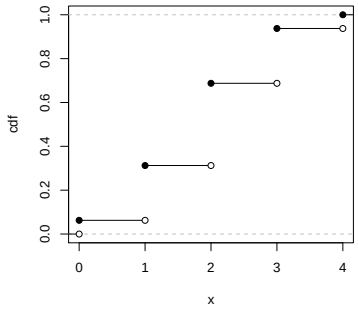


The PMF satisfies

$$p_X(x) \geq 0 \text{ and } \sum_x p_X(x) = 1$$

Cumulative Distribution Function (CDF) Gives the probability that a random variable is less than or equal to x.

$$F_X(x) = P(X \leq x)$$



The CDF is an increasing, right-continuous function with

$$F_X(x) \to 0 \text{ as } x \to -\infty \text{ and } F_X(x) \to 1 \text{ as } x \to \infty$$

Independence Intuitively, two random variables are independent if knowing the value of one gives no information about the other. Discrete r.v.s X and Y are independent if for *all* values of x and y

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Expected Value and Indicators

Expected Value and Linearity

Expected Value (a.k.a. *mean*, *expectation*, or *average*) is a weighted average of the possible outcomes of our random variable. Mathematically, if $x_1, x_2, x_3, \ldots$ are all of the distinct possible values that X can take, the expected value of X is

$$E(X) = \sum_i x_i P(X = x_i)$$

X	Y	$X+Y$
3	4	7
2	2	4
6	8	14
10	23	33
1	–3	–2
1	0	1
5	9	14
4	1	5
...	...	...
$\frac{1}{n}\sum_{i=1}^n x_i$	+ $\frac{1}{n}\sum_{i=1}^n y_i$	= $\frac{1}{n}\sum_{i=1}^n (x_i + y_i)$
$E(X)$	+ $E(Y)$	= $E(X+Y)$

Linearity For any r.v.s X and Y, and constants a, b, c,

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

Same distribution implies same mean If X and Y have the same distribution, then $E(X) = E(Y)$ and, more generally,

$$E(g(X)) = E(g(Y))$$

Conditional Expected Value is defined like expectation, only conditioned on any event A.

$$E(X|A) = \sum_x xP(X = x|A)$$

Indicator Random Variables

Indicator Random Variable is a random variable that takes on the value 1 or 0. It is always an indicator of some event: if the event occurs, the indicator is 1; otherwise it is 0. They are useful for many problems about counting how many events of some kind occur. Write

$$I_A = \begin{cases} 1 & \text{if } A \text{ occurs,} \\ 0 & \text{if } A \text{ does not occur.} \end{cases}$$

Note that $I_A^2 = I_A, I_A I_B = I_{A \cap B}$, and $I_{A \cup B} = I_A + I_B - I_A I_B$.

Distribution $I_A \sim \text{Bern}(p)$ where $p = P(A)$.

Fundamental Bridge The expectation of the indicator for event A is the probability of event A: $E(I_A) = P(A)$.

Variance and Standard Deviation

$$\text{Var}(X) = E\left(X - E(X)\right)^2 = E(X^2) - (E(X))^2$$

$$\text{SD}(X) = \sqrt{\text{Var}(X)}$$

Continuous RVs, LOTUS, UoU

Continuous Random Variables (CRVs)

What's the probability that a CRV is in an interval? Take the difference in CDF values (or use the PDF as described later).

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

For $X \sim \mathcal{N}(\mu, \sigma^2)$, this becomes

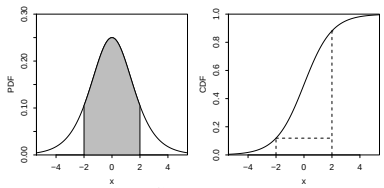
$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

What is the Probability Density Function (PDF)? The PDF f is the derivative of the CDF F.

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

A PDF is nonnegative and integrates to 1. By the fundamental theorem of calculus, to get from PDF back to CDF we can integrate:

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



To find the probability that a CRV takes on a value in an interval, integrate the PDF over that interval.

$$F(b) - F(a) = \int_{a}^{b} f(x)dx$$

How do I find the expected value of a CRV? Analogous to the discrete case, where you sum x times the PMF, for CRVs you integrate x times the PDF.

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

LOTUS

Expected value of a function of an r.v. The expected value of X is defined this way:

$$E(X) = \sum_x xP(X = x) \text{ (for discrete } X)$$

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \text{ (for continuous } X)$$

The **Law of the Unconscious Statistician (LOTUS)** states that you can find the expected value of a *function of a random variable*, $g(X)$, in a similar way, by replacing the x in front of the PMF/PDF by $g(x)$ but still working with the PMF/PDF of X:

$$E(g(X)) = \sum_x g(x)P(X = x) \text{ (for discrete } X)$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx \text{ (for continuous } X)$$

What's a function of a random variable? A function of a random variable is also a random variable. For example, if X is the number of bikes you see in an hour, then $g(X) = 2X$ is the number of bike wheels you see in that hour and $h(X) = \binom{X}{2} = \frac{X(X-1)}{2}$ is the number of *pairs* of bikes such that you see both of those bikes in that hour.

What's the point? You don't need to know the PMF/PDF of $g(X)$ to find its expected value. All you need is the PMF/PDF of X.

Universality of Uniform (UoU)

When you plug any CRV into its own CDF, you get a Uniform(0,1) random variable. When you plug a Uniform(0,1) r.v. into an inverse CDF, you get an r.v. with that CDF. For example, let's say that a random variable X has CDF

$$F(x) = 1 - e^{-x}, \text{ for } x > 0$$

By UoU, if we plug X into this function then we get a uniformly distributed random variable.

$$F(X) = 1 - e^{-X} \sim \text{Unif}(0, 1)$$

Similarly, if $U \sim \text{Unif}(0, 1)$ then $F^{-1}(U)$ has CDF F. The key point is that for any continuous random variable X, we can transform it into a Uniform random variable and back by using its CDF.

Moments and MGFs

Moments

Moments describe the shape of a distribution. Let X have mean μ and standard deviation σ, and $Z = (X - \mu)/\sigma$ be the *standardized* version of X. The kth moment of X is $\mu_k = E(X^k)$ and the kth standardized moment of X is $m_k = E(Z^k)$. The mean, variance, skewness, and kurtosis are important summaries of the shape of a distribution.

Mean $E(X) = \mu_1$

Variance $\text{Var}(X) = \mu_2 - \mu_1^2$

Skewness $\text{Skew}(X) = m_3$

Kurtosis $\text{Kurt}(X) = m_4 - 3$

Moment Generating Functions

MGF For any random variable X, the function

$$M_X(t) = E(e^{tX})$$

is the **moment generating function (MGF)** of X, if it exists for all t in some open interval containing 0. The variable t could just as well have been called u or v. It's a bookkeeping device that lets us work with the *function* M_X rather than the *sequence* of moments.

Why is it called the Moment Generating Function? Because the kth derivative of the moment generating function, evaluated at 0, is the kth moment of X.

$$\mu_k = E(X^k) = M_X^{(k)}(0)$$

This is true by Taylor expansion of e^{tX} since

$$M_X(t) = E(e^{tX}) = \sum_{k=0}^{\infty} \frac{E(X^k)t^k}{k!} = \sum_{k=0}^{\infty} \frac{\mu_k t^k}{k!}$$

MGF of linear functions If we have $Y = aX + b$, then

$$M_Y(t) = E(e^{t(aX+b)}) = e^{bt}E(e^{(at)X}) = e^{bt}M_X(at)$$

Uniqueness *If it exists, the MGF uniquely determines the distribution.* This means that for any two random variables X and Y, they are distributed the same (their PMFs/PDFs are equal) if and only if their MGFs are equal.

Summing Independent RVs by Multiplying MGFs. If X and Y are independent, then

$$M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX})E(e^{tY}) = M_X(t) \cdot M_Y(t)$$

The MGF of the sum of two random variables is the product of the MGFs of those two random variables.

Joint PDFs and CDFs

Joint Distributions

The **joint CDF** of X and Y is

$$F(x, y) = P(X \leq x, Y \leq y)$$

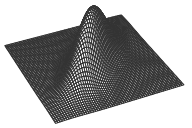
In the discrete case, X and Y have a **joint PMF**

$$p_{X,Y}(x, y) = P(X = x, Y = y).$$

In the continuous case, they have a **joint PDF**

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

The joint PMF/PDF must be nonnegative and sum/integrate to 1.



Conditional Distributions

Conditioning and Bayes' rule for discrete r.v.s

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

Conditioning and Bayes' rule for continuous r.v.s

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

Hybrid Bayes' rule

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

Marginal Distributions

To find the distribution of one (or more) random variables from a joint PMF/PDF, sum/integrate over the unwanted random variables.

Marginal PMF from joint PMF

$$P(X = x) = \sum_y P(X = x, Y = y)$$

Marginal PDF from joint PDF

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$

Independence of Random Variables

Random variables X and Y are independent if and only if any of the following conditions holds:

- Joint CDF is the product of the marginal CDFs
- Joint PMF/PDF is the product of the marginal PMFs/PDFs
- Conditional distribution of Y given X is the marginal distribution of Y

Write $X \perp\!\!\!\perp Y$ to denote that X and Y are independent.

Multivariate LOTUS

LOTUS in more than one dimension is analogous to the 1D LOTUS. For discrete random variables:

$$E(g(X,Y)) = \sum_x \sum_y g(x,y)P(X=x,Y=y)$$

For continuous random variables:

$$E(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)dxdy$$

Covariance and Transformations

Covariance and Correlation

Covariance is the analog of variance for two random variables.

$$\text{Cov}(X,Y) = E\left((X-E(X))(Y-E(Y))\right) = E(XY) - E(X)E(Y)$$

Note that

$$\text{Cov}(X,X) = E(X^2) - (E(X))^2 = \text{Var}(X)$$

Correlation is a standardized version of covariance that is always between -1 and 1.

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Covariance and Independence If two random variables are independent, then they are uncorrelated. The converse is not necessarily true (e.g., consider $X \sim \mathcal{N}(0,1)$ and $Y = X^2$).

$$X \perp\!\!\!\perp Y \longrightarrow \text{Cov}(X,Y) = 0 \longrightarrow E(XY) = E(X)E(Y)$$

Covariance and Variance The variance of a sum can be found by

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$$

$$\text{Var}(X_1+X_2+\cdots+X_n) = \sum_{i=1}^{n}\text{Var}(X_i) + 2\sum_{i<j}\text{Cov}(X_i,X_j)$$

If X and Y are independent then they have covariance 0, so

$$X \perp\!\!\!\perp Y \Longrightarrow \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$

If $X_1, X_2, \ldots, X_n$ are identically distributed and have the same covariance relationships (often by **symmetry**), then

$$\text{Var}(X_1+X_2+\cdots+X_n) = n\text{Var}(X_1) + 2\binom{n}{2}\text{Cov}(X_1,X_2)$$

Covariance Properties For random variables W, X, Y, Z and constants a, b:

$$\begin{aligned}
\text{Cov}(X,Y) &= \text{Cov}(Y,X)\\
\text{Cov}(X+a,Y+b) &= \text{Cov}(X,Y)\\
\text{Cov}(aX,bY) &= ab\text{Cov}(X,Y)\\
\text{Cov}(W+X,Y+Z) &= \text{Cov}(W,Y) + \text{Cov}(W,Z) + \text{Cov}(X,Y)\\
&\quad + \text{Cov}(X,Z)
\end{aligned}$$

Correlation is location-invariant and scale-invariant For any constants a, b, c, d with a and c nonzero,

$$\text{Corr}(aX+b, cY+d) = \text{Corr}(X,Y)$$

Transformations

One Variable Transformations Let's say that we have a random variable X with PDF $f_X(x)$, but we are also interested in some function of X. We call this function $Y = g(X)$. Also let $y = g(x)$. If g is differentiable and strictly increasing (or strictly decreasing), then the PDF of Y is

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

The derivative of the inverse transformation is called the **Jacobian**.

Two Variable Transformations Similarly, let's say we know the joint PDF of U and V but are also interested in the random vector (X,Y) defined by $(X,Y) = g(U,V)$. Let

$$\frac{\partial(u,v)}{\partial(x,y)} = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix}$$

be the **Jacobian matrix**. If the entries in this matrix exist and are continuous, and the determinant of the matrix is never 0, then

$$f_{X,Y}(x,y) = f_{U,V}(u,v)\left|\left|\frac{\partial(u,v)}{\partial(x,y)}\right|\right|$$

The inner bars tells us to take the matrix's determinant, and the outer bars tell us to take the absolute value. In a 2×2 matrix,

$$\left|\left| \begin{matrix} a & b \\ c & d \end{matrix} \right|\right| = |ad-bc|$$

Convolutions

Convolution Integral If you want to find the PDF of the sum of two independent CRVs X and Y, you can do the following integral:

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x)f_Y(t-x)dx$$

Example Let $X, Y \sim \mathcal{N}(0,1)$ be i.i.d. Then for each fixed t,

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

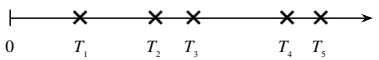
By completing the square and using the fact that a Normal PDF integrates to 1, this works out to $f_{X+Y}(t)$ being the $\mathcal{N}(0,2)$ PDF.

Poisson Process

Definition We have a **Poisson process** of rate λ arrivals per unit time if the following conditions hold:

1. The number of arrivals in a time interval of length t is $\text{Pois}(\lambda t)$.
2. Numbers of arrivals in disjoint time intervals are independent.

For example, the numbers of arrivals in the time intervals $[0,5]$, $(5,12)$, and $[13,23)$ are independent with $\text{Pois}(5\lambda)$, $\text{Pois}(7\lambda)$, $\text{Pois}(10\lambda)$ distributions, respectively.



Count-Time Duality Consider a Poisson process of emails arriving in an inbox at rate λ emails per hour. Let T_n be the time of arrival of the nth email (relative to some starting time 0) and N_t be the number of emails that arrive in $[0,t]$. Let's find the distribution of T_1. The event $T_1 > t$, the event that you have to wait more than t hours to get the first email, is the same as the event $N_t = 0$, which is the event that there are no emails in the first t hours. So

$$P(T_1 > t) = P(N_t = 0) = e^{-\lambda t} \longrightarrow P(T_1 \le t) = 1 - e^{-\lambda t}$$

Thus we have $T_1 \sim \text{Expo}(\lambda)$. By the memoryless property and similar reasoning, the interarrival times between emails are i.i.d. $\text{Expo}(\lambda)$, i.e., the differences $T_n - T_{n-1}$ are i.i.d. $\text{Expo}(\lambda)$.

Order Statistics

Definition Let's say you have n i.i.d. r.v.s $X_1, X_2, \ldots, X_n$. If you arrange them from smallest to largest, the ith element in that list is the ith order statistic, denoted $X_{(i)}$. So $X_{(1)}$ is the smallest in the list and $X_{(n)}$ is the largest in the list.

Note that the order statistics are *dependent*, e.g., learning $X_{(4)} = 42$ gives us the information that $X_{(1)}, X_{(2)}, X_{(3)}$ are ≤ 42 and $X_{(5)}, X_{(6)}, \ldots, X_{(n)}$ are ≥ 42.

Distribution Taking n i.i.d. random variables $X_1, X_2, \ldots, X_n$ with CDF $F(x)$ and PDF $f(x)$, the CDF and PDF of $X_{(i)}$ are:

$$F_{X_{(i)}}(x) = P(X_{(i)} \le x) = \sum_{k=i}^{n}\binom{n}{k}F(x)^k(1-F(x))^{n-k}$$

$$f_{X_{(i)}}(x) = n\binom{n-1}{i-1}F(x)^{i-1}(1-F(x))^{n-i}f(x)$$

Uniform Order Statistics The jth order statistic of i.i.d. $U_1, \ldots, U_n \sim \text{Unif}(0,1)$ is $U_{(j)} \sim \text{Beta}(j, n-j+1)$.

Conditional Expectation

Conditioning on an Event We can find $E(Y|A)$, the expected value of Y given that event A occurred. A very important case is when A is the event $X = x$. Note that $E(Y|A)$ is a *number*. For example:

- The expected value of a fair die roll, given that it is prime, is $\frac{1}{3}\cdot 2 + \frac{1}{3}\cdot 3 + \frac{1}{3}\cdot 5 = \frac{10}{3}$.
- Let Y be the number of successes in 10 independent Bernoulli trials with probability p of success. Let A be the event that the first 3 trials are all successes. Then
$$E(Y|A) = 3 + 7p$$
since the number of successes among the last 7 trials is $\text{Bin}(7,p)$.
- Let $T \sim \text{Expo}(1/10)$ be how long you have to wait until the shuttle comes. Given that you have already waited t minutes, the expected additional waiting time is 10 more minutes, by the memoryless property. That is, $E(T|T>t) = t + 10$.

Discrete Y	Continuous Y
$E(Y) = \sum_y yP(Y=y)$	$E(Y) = \int_{-\infty}^{\infty} yf_Y(y)dy$
$E(Y\|A) = \sum_y yP(Y=y\|A)$	$E(Y\|A) = \int_{-\infty}^{\infty} yf(y\|A)dy$

Conditioning on a Random Variable We can also find $E(Y|X)$, the expected value of Y given the random variable X. This is *a function of the random variable* X. It is *not* a number except in certain special cases such as if $X \perp\!\!\!\perp Y$. To find $E(Y|X)$, find $E(Y|X=x)$ and then plug in X for x. For example:

- If $E(Y|X=x) = x^3 + 5x$, then $E(Y|X) = X^3 + 5X$.
- Let Y be the number of successes in 10 independent Bernoulli trials with probability p of success and X be the number of successes among the first 3 trials. Then $E(Y|X) = X + 7p$.
- Let $X \sim \mathcal{N}(0,1)$ and $Y = X^2$. Then $E(Y|X=x) = x^2$ since if we know $X = x$ then we know $Y = x^2$. And $E(X|Y=y) = 0$ since if we know $Y = y$ then we know $X = \pm\sqrt{y}$, with equal probabilities (by symmetry). So $E(Y|X) = X^2, E(X|Y) = 0$.

Properties of Conditional Expectation

1. $E(Y|X) = E(Y)$ if $X \perp\!\!\!\perp Y$

2. $E(h(X)W|X) = h(X)E(W|X)$ (**taking out what's known**) In particular, $E(h(X)|X) = h(X)$.
3. $E(E(Y|X)) = E(Y)$ (**Adam's Law**, a.k.a. Law of Total Expectation)

Adam's Law (a.k.a. Law of Total Expectation) can also be written in a way that looks analogous to LOTP. For any events $A_1, A_2, \ldots, A_n$ that partition the sample space,

$$E(Y) = E(Y|A_1)P(A_1) + \cdots + E(Y|A_n)P(A_n)$$

For the special case where the partition is A, A^c, this says

$$E(Y) = E(Y|A)P(A) + E(Y|A^c)P(A^c)$$

Eve's Law (a.k.a. Law of Total Variance)

$$\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X))$$

MVN, LLN, CLT

Law of Large Numbers (LLN)

Let $X_1, X_2, X_3 \ldots$ be i.i.d. with mean μ. The **sample mean** is

$$\bar{X}_n = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

The **Law of Large Numbers** states that as $n \to \infty$, $\bar{X}_n \to \mu$ with probability 1. For example, in flips of a coin with probability p of Heads, let X_j be the indicator of the jth flip being Heads. Then LLN says the proportion of Heads converges to p (with probability 1).

Central Limit Theorem (CLT)

Approximation using CLT

We use $\dot{\sim}$ to denote *is approximately distributed*. We can use the **Central Limit Theorem** to approximate the distribution of a random variable $Y = X_1 + X_2 + \cdots + X_n$ that is a sum of n i.i.d. random variables X_i. Let $E(Y) = \mu_Y$ and $\text{Var}(Y) = \sigma_Y^2$. The CLT says

$$Y \dot{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$$

If the X_i are i.i.d. with mean μ_X and variance σ_X^2, then $\mu_Y = n\mu_X$ and $\sigma_Y^2 = n\sigma_X^2$. For the sample mean $\bar{X}_n$, the CLT says

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \dot{\sim} \mathcal{N}(\mu_X, \sigma_X^2/n)$$

Asymptotic Distributions using CLT

We use $\xrightarrow{D}$ to denote *converges in distribution to* as $n \to \infty$. The CLT says that if we standardize the sum $X_1 + \cdots + X_n$ then the distribution of the sum converges to $\mathcal{N}(0,1)$ as $n \to \infty$:

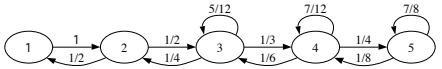
$$\frac{1}{\sigma\sqrt{n}}(X_1 + \cdots + X_n - n\mu_X) \xrightarrow{D} \mathcal{N}(0,1)$$

In other words, the CDF of the left-hand side goes to the standard Normal CDF, Φ. In terms of the sample mean, the CLT says

$$\frac{\sqrt{n}(\bar{X}_n - \mu_X)}{\sigma_X} \xrightarrow{D} \mathcal{N}(0,1)$$

Markov Chains

Definition



A Markov chain is a random walk in a **state space**, which we will assume is finite, say $\{1, 2, \ldots, M\}$. We let X_t denote which element of the state space the walk is visiting at time t. The Markov chain is the sequence of random variables tracking where the walk is at all points in time, $X_0, X_1, X_2, \ldots$. By definition, a Markov chain must satisfy the **Markov property**, which says that if you want to predict where the chain will be at a future time, if we know the present state then the entire past history is irrelevant. *Given the present, the past and future are conditionally independent.* In symbols,

$$P(X_{n+1} = j|X_0 = i_0, X_1 = i_1, \ldots, X_n = i) = P(X_{n+1} = j|X_n = i)$$

State Properties

A state is either recurrent or transient.

- If you start at a **recurrent state**, then you will always return back to that state at some point in the future. ♪*You can check-out any time you like, but you can never leave.* ♪
- Otherwise you are at a **transient state**. There is some positive probability that once you leave you will never return. ♪*You don't have to go home, but you can't stay here.* ♪

A state is either periodic or aperiodic.

- If you start at a **periodic state** of period k, then the GCD of the possible numbers of steps it would take to return back is $k > 1$.
- Otherwise you are at an **aperiodic state**. The GCD of the possible numbers of steps it would take to return back is 1.

Transition Matrix

Let the state space be $\{1, 2, \ldots, M\}$. The transition matrix Q is the $M \times M$ matrix where element q_{ij} is the probability that the chain goes from state i to state j in one step:

$$q_{ij} = P(X_{n+1} = j|X_n = i)$$

To find the probability that the chain goes from state i to state j in exactly m steps, take the (i,j) element of Q^m.

$$q_{ij}^{(m)} = P(X_{n+m} = j|X_n = i)$$

If X_0 is distributed according to the row vector PMF $\vec{p}$, i.e., $p_j = P(X_0 = j)$, then the PMF of X_n is $\vec{p}Q^n$.

Chain Properties

A chain is **irreducible** if you can get from anywhere to anywhere. If a chain (on a finite state space) is irreducible, then all of its states are recurrent. A chain is **periodic** if any of its states are periodic, and is **aperiodic** if none of its states are periodic. In an irreducible chain, all states have the same period.

A chain is **reversible** with respect to $\vec{s}$ if $s_i q_{ij} = s_j q_{ji}$ for all i, j. Examples of reversible chains include any chain with $q_{ij} = q_{ji}$, with $\vec{s} = (\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M})$, and random walk on an undirected network.

Stationary Distribution

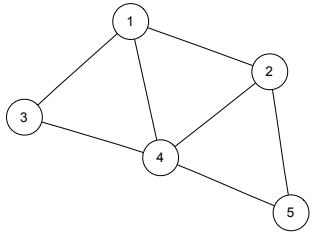
Let us say that the vector $\vec{s} = (s_1, s_2, \ldots, s_M)$ be a PMF (written as a row vector). We will call $\vec{s}$ the **stationary distribution** for the chain if $\vec{s}Q = \vec{s}$. As a consequence, if X_t has the stationary distribution, then all future $X_{t+1}, X_{t+2}, \ldots$ also have the stationary distribution.

For irreducible, aperiodic chains, the stationary distribution exists, is unique, and s_i is the long-run probability of a chain being at state i. The expected number of steps to return to i starting from i is $1/s_i$.

To find the stationary distribution, you can solve the matrix equation $(Q' - I)\vec{s}' = 0$. The stationary distribution is uniform if the columns of Q sum to 1.

Reversibility Condition Implies Stationarity If you have a PMF $\vec{s}$ and a Markov chain with transition matrix Q, then $s_i q_{ij} = s_j q_{ji}$ for all states i, j implies that $\vec{s}$ is stationary.

Random Walk on an Undirected Network



If you have a collection of **nodes**, pairs of which can be connected by undirected **edges**, and a Markov chain is run by going from the current node to a uniformly random node that is connected to it by an edge, then this is a random walk on an undirected network. The stationary distribution of this chain is proportional to the **degree sequence** (this is the sequence of degrees, where the degree of a node is how many edges are attached to it). For example, the stationary distribution of random walk on the network shown above is proportional to $(3, 3, 2, 4, 2)$, so it's $(\frac{3}{14}, \frac{3}{14}, \frac{3}{14}, \frac{4}{14}, \frac{2}{14})$.

Continuous Distributions

Uniform Distribution

Let us say that U is distributed Unif(a, b). We know the following:

Properties of the Uniform For a Uniform distribution, the probability of a draw from any interval within the support is proportional to the length of the interval. See *Universality of Uniform* and *Order Statistics* for other properties.

Example William throws darts really badly, so his darts are uniform over the whole room because they're equally likely to appear anywhere. William's darts have a Uniform distribution on the surface of the room. The Uniform is the only distribution where the probability of hitting in any specific region is proportional to the length/area/volume of that region, and where the density of occurrence in any one specific spot is constant throughout the whole support.

Normal Distribution

Let us say that X is distributed $\mathcal{N}(\mu, \sigma^2)$. We know the following:

Central Limit Theorem The Normal distribution is ubiquitous because of the Central Limit Theorem, which states that the sample mean of i.i.d. r.v.s will approach a Normal distribution as the sample size grows, regardless of the initial distribution.

Location-Scale Transformation Every time we shift a Normal r.v. (by adding a constant) or rescale a Normal (by multiplying by a constant), we change it to another Normal r.v. For any Normal $X \sim \mathcal{N}(\mu, \sigma^2)$, we can transform it to the standard $\mathcal{N}(0,1)$ by the following transformation:

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

Standard Normal The Standard Normal, $Z \sim \mathcal{N}(0,1)$, has mean 0 and variance 1. Its CDF is denoted by Φ.

Exponential Distribution

Let us say that X is distributed Expo(λ). We know the following:

Story You're sitting on an open meadow right before the break of dawn, wishing that airplanes in the night sky were shooting stars, because you could really use a wish right now. You know that shooting stars come on average every 15 minutes, but a shooting star is not "due" to come just because you've waited so long. Your waiting time is memoryless; the additional time until the next shooting star comes does not depend on how long you've waited already.

Example The waiting time until the next shooting star is distributed Expo(4) hours. Here $\lambda = 4$ is the **rate parameter**, since shooting stars arrive at a rate of 1 per 1/4 hour on average. The expected time until the next shooting star is $1/\lambda = 1/4$ hour.

Expos as a rescaled Expo(1)

$$Y \sim \text{Expo}(\lambda) \to X = \lambda Y \sim \text{Expo}(1)$$

Memorylessness The Exponential Distribution is the only continuous memoryless distribution. The memoryless property says that for $X \sim \text{Expo}(\lambda)$ and any positive numbers s and t,

$$P(X > s + t | X > s) = P(X > t)$$

Equivalently,

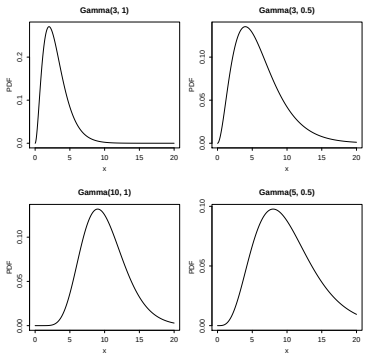
$$X - a | (X > a) \sim \text{Expo}(\lambda)$$

For example, a product with an Expo(λ) lifetime is always "as good as new" (it doesn't experience wear and tear). Given that the product has survived a years, the additional time that it will last is still Expo(λ).

Min of Expos If we have independent $X_i \sim \text{Expo}(\lambda_i)$, then $\min(X_1, \ldots, X_k) \sim \text{Expo}(\lambda_1 + \lambda_2 + \cdots + \lambda_k)$.

Max of Expos If we have i.i.d. $X_i \sim \text{Expo}(\lambda)$, then $\max(X_1, \ldots, X_k)$ has the same distribution as $Y_1 + Y_2 + \cdots + Y_k$, where $Y_j \sim \text{Expo}(j\lambda)$ and the Y_j are independent.

Gamma Distribution

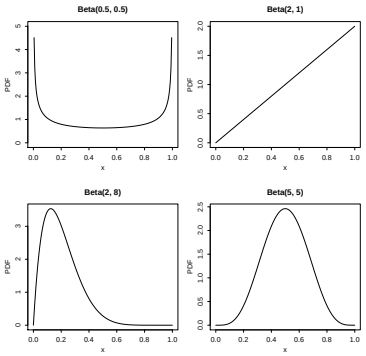


Let us say that X is distributed Gamma(a, λ). We know the following:

Story You sit waiting for shooting stars, where the waiting time for a star is distributed Expo(λ). You want to see n shooting stars before you go home. The total waiting time for the nth shooting star is Gamma(n, λ).

Example You are at a bank, and there are 3 people ahead of you. The serving time for each person is Exponential with mean 2 minutes. Only one person at a time can be served. The distribution of your waiting time until it's your turn to be served is Gamma(3, $\frac{1}{2}$).

Beta Distribution



Conjugate Prior of the Binomial In the Bayesian approach to statistics, parameters are viewed as random variables, to reflect our uncertainty. The *prior* for a parameter is its distribution before observing data. The *posterior* is the distribution for the parameter after observing data. Beta is the *conjugate* prior of the Binomial because if you have a Beta-distributed prior on p in a Binomial, then the posterior distribution on p given the Binomial data is also Beta-distributed. Consider the following two-level model:

$$X | p \sim \text{Bin}(n, p)$$
$$p \sim \text{Beta}(a, b)$$

Then after observing $X = x$, we get the posterior distribution

$$p | (X = x) \sim \text{Beta}(a + x, b + n - x)$$

Order statistics of the Uniform See *Order Statistics*.

Beta-Gamma relationship If $X \sim \text{Gamma}(a, \lambda)$, $Y \sim \text{Gamma}(b, \lambda)$, with $X \perp\!\!\!\perp Y$ then

- $\frac{X}{X+Y} \sim \text{Beta}(a, b)$
- $X + Y \perp\!\!\!\perp \frac{X}{X+Y}$

This is known as the **bank–post office result**.

χ^2 (Chi-Square) Distribution

Let us say that X is distributed χ^2_n. We know the following:

Story A Chi-Square(n) is the sum of the squares of n independent standard Normal r.v.s.

Properties and Representations

$$X \text{ is distributed as } Z_1^2 + Z_2^2 + \cdots + Z_n^2 \text{ for i.i.d. } Z_i \sim \mathcal{N}(0,1)$$

$$X \sim \text{Gamma}(n/2, 1/2)$$

Discrete Distributions

Distributions for four sampling schemes

	Replace	**No Replace**
Fixed # trials (n)	Binomial (Bern if $n = 1$)	HGeom
Draw until r success	NBin (Geom if $r = 1$)	NHGeom

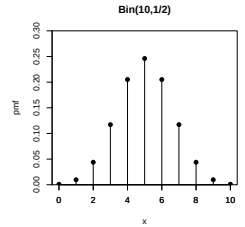
Bernoulli Distribution

The Bernoulli distribution is the simplest case of the Binomial distribution, where we only have one trial ($n = 1$). Let us say that X is distributed Bern(p). We know the following:

Story A trial is performed with probability p of "success", and X is the indicator of success: 1 means success, 0 means failure.

Example Let X be the indicator of Heads for a fair coin toss. Then $X \sim \text{Bern}(\frac{1}{2})$. Also, $1 - X \sim \text{Bern}(\frac{1}{2})$ is the indicator of Tails.

Binomial Distribution



Let us say that X is distributed Bin(n, p). We know the following:

Story X is the number of "successes" that we will achieve in n independent trials, where each trial is either a success or a failure, each with the same probability p of success. We can also write X as a sum of multiple independent Bern(p) random variables. Let $X \sim \text{Bin}(n, p)$ and $X_j \sim \text{Bern}(p)$, where all of the Bernoullis are independent. Then

$$X = X_1 + X_2 + X_3 + \cdots + X_n$$

Example If Jeremy Lin makes 10 free throws and each one independently has a $\frac{3}{4}$ chance of getting in, then the number of free throws he makes is distributed Bin(10, $\frac{3}{4}$).

Properties Let $X \sim \text{Bin}(n, p), Y \sim \text{Bin}(m, p)$ with $X \perp\!\!\!\perp Y$.

- **Redefine success** $n - X \sim \text{Bin}(n, 1 - p)$
- **Sum** $X + Y \sim \text{Bin}(n + m, p)$
- **Conditional** $X | (X + Y = r) \sim \text{HGeom}(n, m, r)$
- **Binomial-Poisson Relationship** Bin(n, p) is approximately Pois(λ) if p is small.
- **Binomial-Normal Relationship** Bin(n, p) is approximately $\mathcal{N}(np, np(1-p))$ if n is large and p is not near 0 or 1.

Geometric Distribution

Let us say that X is distributed Geom(p). We know the following:

Story X is the number of "failures" that we will achieve before we achieve our first success. Our successes have probability p.

Example If each pokeball we throw has probability $\frac{1}{10}$ to catch Mew, the number of failed pokeballs will be distributed Geom($\frac{1}{10}$).

First Success Distribution

Equivalent to the Geometric distribution, except that it includes the first success in the count. This is 1 more than the number of failures. If $X \sim \text{FS}(p)$ then $E(X) = 1/p$.

Negative Binomial Distribution

Let us say that X is distributed $\text{NBin}(r, p)$. We know the following:

Story X is the number of "failures" that we will have before we achieve our rth success. Our successes have probability p.

Example Thundershock has 60% accuracy and can faint a wild Raticate in 3 hits. The number of misses before Pikachu faints Raticate with Thundershock is distributed $\text{NBin}(3, 0.6)$.

Hypergeometric Distribution

Let us say that X is distributed $\text{HGeom}(w, b, n)$. We know the following:

Story In a population of w desired objects and b undesired objects, X is the number of "successes" we will have in a draw of n objects, without replacement. The draw of n objects is assumed to be a **simple random sample** (all sets of n objects are equally likely).

Examples Here are some HGeom examples.

- Let's say that we have only b Weedles (failure) and w Pikachus (success) in Viridian Forest. We encounter n Pokemon in the forest, and X is the number of Pikachus in our encounters.
- The number of Aces in a 5 card hand.
- You have w white balls and b black balls, and you draw n balls. You will draw X white balls.
- You have w white balls and b black balls, and you draw n balls without replacement. The number of white balls in your sample is $\text{HGeom}(w, b, n)$; the number of black balls is $\text{HGeom}(b, w, n)$.
- **Capture-recapture** A forest has N elk, you capture n of them, tag them, and release them. Then you recapture a new sample of size m. How many tagged elk are now in the new sample? $\text{HGeom}(n, N-n, m)$

Poisson Distribution

Let us say that X is distributed $\text{Pois}(\lambda)$. We know the following:

Story There are rare events (low probability events) that occur many different ways (high possibilities of occurences) at an average rate of λ occurrences per unit space or time. The number of events that occur in that unit of space or time is X.

Example A certain busy intersection has an average of 2 accidents per month. Since an accident is a low probability event that can happen many different ways, it is reasonable to model the number of accidents in a month at that intersection as $\text{Pois}(2)$. Then the number of accidents that happen in two months at that intersection is distributed $\text{Pois}(4)$.

Properties Let $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$, with $X \perp\!\!\!\perp Y$.

1. **Sum** $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
2. **Conditional** $X|(X+Y=n) \sim \text{Bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$
3. **Chicken-egg** If there are $Z \sim \text{Pois}(\lambda)$ items and we randomly and independently "accept" each item with probability p, then the number of accepted items $Z_1 \sim \text{Pois}(\lambda p)$, and the number of rejected items $Z_2 \sim \text{Pois}(\lambda(1-p))$, and $Z_1 \perp\!\!\!\perp Z_2$.

Multivariate Distributions

Multinomial Distribution

Let us say that the vector $\vec{X} = (X_1, X_2, X_3, \ldots, X_k) \sim \text{Mult}_k(n, \vec{p})$ where $\vec{p} = (p_1, p_2, \ldots, p_k)$.

Story We have n items, which can fall into any one of the k buckets independently with the probabilities $\vec{p} = (p_1, p_2, \ldots, p_k)$.

Example Let us assume that every year, 100 students in the Harry Potter Universe are randomly and independently sorted into one of four houses with equal probability. The number of people in each of the houses is distributed $\text{Mult}_4(100, \vec{p})$, where $\vec{p} = (0.25, 0.25, 0.25, 0.25)$. Note that $X_1 + X_2 + \cdots + X_4 = 100$, and they are dependent.

Joint PMF For $n = n_1 + n_2 + \cdots + n_k$,

$$P(\vec{X} = \vec{n}) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

Marginal PMF, Lumping, and Conditionals Marginally, $X_i \sim \text{Bin}(n, p_i)$ since we can define "success" to mean category i. If you lump together multiple categories in a Multinomial, then it is still Multinomial. For example, $X_i + X_j \sim \text{Bin}(n, p_i + p_j)$ for $i \neq j$ since we can define "success" to mean being in category i or j. Similarly, if $k = 6$ and we lump categories 1-2 and lump categories 3-5, then

$$(X_1 + X_2, X_3 + X_4 + X_5, X_6) \sim \text{Mult}_3(n, (p_1 + p_2, p_3 + p_4 + p_5, p_6))$$

Conditioning on some X_j also still gives a Multinomial:

$$X_1, \ldots, X_{k-1} | X_k = n_k \sim \text{Mult}_{k-1}\left(n - n_k, \left(\frac{p_1}{1-p_k}, \ldots, \frac{p_{k-1}}{1-p_k}\right)\right)$$

Variances and Covariances We have $X_i \sim \text{Bin}(n, p_i)$ marginally, so $\text{Var}(X_i) = np_i(1-p_i)$. Also, $\text{Cov}(X_i, X_j) = -np_ip_j$ for $i \neq j$.

Multivariate Uniform Distribution

See the univariate Uniform for stories and examples. For the 2D Uniform on some region, probability is proportional to area. Every point in the support has equal density, of value $\frac{1}{\text{area of region}}$. For the 3D Uniform, probability is proportional to volume.

Multivariate Normal (MVN) Distribution

A vector $\vec{X} = (X_1, X_2, \ldots, X_k)$ is Multivariate Normal if every linear combination is Normally distributed, i.e., $t_1X_1 + t_2X_2 + \cdots + t_kX_k$ is Normal for any constants $t_1, t_2, \ldots, t_k$. The parameters of the Multivariate Normal are the **mean vector** $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)$ and the **covariance matrix** where the (i,j) entry is $\text{Cov}(X_i, X_j)$.

Properties The Multivariate Normal has the following properties.

- Any subvector is also MVN.
- If any two elements within an MVN are uncorrelated, then they are independent.
- The joint PDF of a Bivariate Normal (X, Y) with $\mathcal{N}(0,1)$ marginal distributions and correlation $\rho \in (-1, 1)$ is

$$f_{X,Y}(x,y) = \frac{1}{2\pi\tau} \exp\left(-\frac{1}{2\tau^2}(x^2 + y^2 - 2\rho xy)\right),$$

with $\tau = \sqrt{1-\rho^2}$.

Distribution Properties

Important CDFs

Standard Normal Φ

Exponential(λ) $F(x) = 1 - e^{-\lambda x}$, for $x \in (0, \infty)$

Uniform(0,1) $F(x) = x$, for $x \in (0, 1)$

Convolutions of Random Variables

A convolution of n random variables is simply their sum. For the following results, let X and Y be *independent*.

1. $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2) \longrightarrow X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
2. $X \sim \text{Bin}(n_1, p)$, $Y \sim \text{Bin}(n_2, p) \longrightarrow X + Y \sim \text{Bin}(n_1 + n_2, p)$. $\text{Bin}(n, p)$ can be thought of as a sum of i.i.d. $\text{Bern}(p)$ r.v.s.
3. $X \sim \text{Gamma}(a_1, \lambda)$, $Y \sim \text{Gamma}(a_2, \lambda)$ $\longrightarrow X + Y \sim \text{Gamma}(a_1 + a_2, \lambda)$. $\text{Gamma}(n, \lambda)$ with n an integer can be thought of as a sum of i.i.d. $\text{Expo}(\lambda)$ r.v.s.
4. $X \sim \text{NBin}(r_1, p)$, $Y \sim \text{NBin}(r_2, p)$ $\longrightarrow X + Y \sim \text{NBin}(r_1 + r_2, p)$. $\text{NBin}(r, p)$ can be thought of as a sum of i.i.d. $\text{Geom}(p)$ r.v.s.
5. $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ $\longrightarrow X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Special Cases of Distributions

1. $\text{Bin}(1, p) \sim \text{Bern}(p)$
2. $\text{Beta}(1, 1) \sim \text{Unif}(0, 1)$
3. $\text{Gamma}(1, \lambda) \sim \text{Expo}(\lambda)$
4. $\chi^2_n \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$
5. $\text{NBin}(1, p) \sim \text{Geom}(p)$

Inequalities

1. **Cauchy-Schwarz** $|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$
2. **Markov** $P(X \geq a) \leq \frac{E|X|}{a}$ for $a > 0$
3. **Chebyshev** $P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$ for $E(X) = \mu, \text{Var}(X) = \sigma^2$
4. **Jensen** $E(g(X)) \geq g(E(X))$ for g convex; reverse if g is concave

Formulas

Geometric Series

$$1 + r + r^2 + \cdots + r^{n-1} = \sum_{k=0}^{n-1} r^k = \frac{1-r^n}{1-r}$$

$$1 + r + r^2 + \cdots = \frac{1}{1-r} \text{ if } |r| < 1$$

Exponential Function (e^x)

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n$$

Gamma and Beta Integrals

You can sometimes solve complicated-looking integrals by pattern-matching to a gamma or beta integral:

$$\int_0^\infty x^{t-1}e^{-x}\,dx = \Gamma(t) \qquad \int_0^1 x^{a-1}(1-x)^{b-1}\,dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Also, $\Gamma(a+1) = a\Gamma(a)$, and $\Gamma(n) = (n-1)!$ if n is a positive integer.

Euler's Approximation for Harmonic Sums

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \approx \log n + 0.577\ldots$$

Stirling's Approximation for Factorials

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Miscellaneous Definitions

Medians and Quantiles Let X have CDF F. Then X has median m if $F(m) \geq 0.5$ and $P(X \geq m) \geq 0.5$. For X continuous, m satisfies $F(m) = 1/2$. In general, the ath quantile of X is $\min\{x : F(x) \geq a\}$; the median is the case $a = 1/2$.

log Statisticians generally use log to refer to natural log (i.e., base e).

i.i.d r.v.s Independent, identically-distributed random variables.

Example Problems

Contributions from Sebastian Chiu

Calculating Probability

A textbook has n typos, which are randomly scattered amongst its n pages, independently. You pick a random page. What is the probability that it has no typos? **Answer:** There is a $\left(1 - \frac{1}{n}\right)$ probability that any specific typo isn't on your page, and thus a $\boxed{\left(1 - \frac{1}{n}\right)^n}$ probability that there are no typos on your page. For n large, this is approximately $e^{-1} = 1/e$.

Linearity and Indicators (1)

In a group of n people, what is the expected number of distinct birthdays (month and day)? What is the expected number of birthday matches? **Answer:** Let X be the number of distinct birthdays and I_j be the indicator for the jth day being represented.

$$E(I_j) = 1 - P(\text{no one born on day } j) = 1 - (364/365)^n$$

By linearity, $\boxed{E(X) = 365\,(1 - (364/365)^n)}$. Now let Y be the number of birthday matches and J_i be the indicator that the ith pair of people have the same birthday. The probability that any two specific people share a birthday is 1/365, so $\boxed{E(Y) = \binom{n}{2}/365}$.

Linearity and Indicators (2)

*This problem is commonly known as the **hat-matching problem**.* There are n people at a party, each with hat. At the end of the party, they each leave with a random hat. What is the expected number of people who leave with the right hat? **Answer:** Each hat has a $1/n$ chance of going to the right person. By linearity, the average number of hats that go to their owners is $\boxed{n(1/n) = 1}$.

Linearity and First Success

*This problem is commonly known as the **coupon collector problem**.* There are n coupon types. At each draw, you get a uniformly random coupon type. What is the expected number of coupons needed until you have a complete set? **Answer:** Let N be the number of coupons needed; we want $E(N)$. Let $N = N_1 + \cdots + N_n$, where N_1 is the draws to get our first new coupon, N_2 is the *additional* draws needed to draw our second new coupon and so on. By the story of the First Success, $N_2 \sim \text{FS}((n-1)/n)$ (after collecting first coupon type, there's $(n-1)/n$ chance you'll get something new). Similarly, $N_3 \sim \text{FS}((n-2)/n)$, and $N_j \sim \text{FS}((n-j+1)/n)$. By linearity,

$$E(N) = E(N_1) + \cdots + E(N_n) = \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} = \boxed{n\sum_{j=1}^{n}\frac{1}{j}}$$

This is approximately $n(\log(n) + 0.577)$ by Euler's approximation.

Orderings of i.i.d. random variables

I call 2 UberX's and 3 Lyfts at the same time. If the time it takes for the rides to reach me are i.i.d., what is the probability that all the Lyfts will arrive first? **Answer:** Since the arrival times of the five cars are i.i.d., all 5! orderings of the arrivals are equally likely. There are 3!2! orderings that involve the Lyfts arriving first, so the probability that the Lyfts arrive first is $\boxed{\frac{3!2!}{5!} = 1/10}$. Alternatively, there are $\binom{5}{3}$ ways to choose 3 of the 5 slots for the Lyfts to occupy, where each of the choices are equally likely. One of these choices has all 3 of the Lyfts arriving first, so the probability is $\boxed{1/\binom{5}{3} = 1/10}$.

Expectation of Negative Hypergeometric

What is the expected number of cards that you draw before you pick your first Ace in a shuffled deck (not counting the Ace)? **Answer:** Consider a non-Ace. Denote this to be card j. Let I_j be the indicator that card j will be drawn before the first Ace. Note that $I_j = 1$ says that j is before all 4 of the Aces in the deck. The probability that this occurs is 1/5 by symmetry. Let X be the number of cards drawn before the first Ace. Then $X = I_1 + I_2 + \ldots + I_{48}$, where each indicator corresponds to one of the 48 non-Aces. Thus,

$$E(X) = E(I_1) + E(I_2) + \ldots + E(I_{48}) = 48/5 = \boxed{9.6}.$$

Minimum and Maximum of RVs

What is the CDF of the maximum of n independent Unif(0,1) random variables? **Answer:** Note that for r.v.s $X_1, X_2, \ldots, X_n$,

$$P(\min(X_1, X_2, \ldots, X_n) \geq a) = P(X_1 \geq a, X_2 \geq a, \ldots, X_n \geq a)$$

Similarly,

$$P(\max(X_1, X_2, \ldots, X_n) \leq a) = P(X_1 \leq a, X_2 \leq a, \ldots, X_n \leq a)$$

We will use this principle to find the CDF of $U_{(n)}$, where $U_{(n)} = \max(U_1, U_2, \ldots, U_n)$ and $U_i \sim \text{Unif}(0,1)$ are i.i.d.

$$\begin{aligned} P(\max(U_1, U_2, \ldots, U_n) \leq a) &= P(U_1 \leq a, U_2 \leq a, \ldots, U_n \leq a) \\ &= P(U_1 \leq a)P(U_2 \leq a)\ldots P(U_n \leq a) \\ &= \boxed{a^n} \end{aligned}$$

for $0 < a < 1$ (and the CDF is 0 for $a \leq 0$ and 1 for $a \geq 1$).

Pattern-matching with e^x Taylor series

For $X \sim \text{Pois}(\lambda)$, find $E\left(\frac{1}{X+1}\right)$. **Answer:** By LOTUS,

$$E\left(\frac{1}{X+1}\right) = \sum_{k=0}^{\infty}\frac{1}{k+1}\frac{e^{-\lambda}\lambda^k}{k!} = \frac{e^{-\lambda}}{\lambda}\sum_{k=0}^{\infty}\frac{\lambda^{k+1}}{(k+1)!} = \boxed{\frac{e^{-\lambda}}{\lambda}(e^{\lambda} - 1)}$$

Adam's Law and Eve's Law

William really likes speedsolving Rubik's Cubes. But he's pretty bad at it, so sometimes he fails. On any given day, William will attempt $N \sim \text{Geom}(s)$ Rubik's Cubes. Suppose each time, he has probability p of solving the cube, independently. Let T be the number of Rubik's Cubes he solves during a day. Find the mean and variance of T. **Answer:** Note that $T|N \sim \text{Bin}(N, p)$. So by Adam's Law,

$$E(T) = E(E(T|N)) = E(Np) = \boxed{\frac{p(1-s)}{s}}$$

Similarly, by Eve's Law, we have that

$$\begin{aligned} \text{Var}(T) &= E(\text{Var}(T|N)) + \text{Var}(E(T|N)) = E(Np(1-p)) + \text{Var}(Np) \\ &= \frac{p(1-p)(1-s)}{s} + \frac{p^2(1-s)}{s^2} = \boxed{\frac{p(1-s)(p+s(1-p))}{s^2}} \end{aligned}$$

MGF – Finding Moments

Find $E(X^3)$ for $X \sim \text{Expo}(\lambda)$ using the MGF of X. **Answer:** The MGF of an Expo(λ) is $M(t) = \frac{\lambda}{\lambda - t}$. To get the third moment, we can take the third derivative of the MGF and evaluate at $t = 0$:

$$\boxed{E(X^3) = \frac{6}{\lambda^3}}$$

But a much nicer way to use the MGF here is via pattern recognition: note that $M(t)$ looks like it came from a geometric series:

$$\frac{1}{1 - \frac{t}{\lambda}} = \sum_{n=0}^{\infty}\left(\frac{t}{\lambda}\right)^n = \sum_{n=0}^{\infty}\frac{n!}{\lambda^n}\frac{t^n}{n!}$$

The coefficient of $\frac{t^n}{n!}$ here is the nth moment of X, so we have $E(X^n) = \frac{n!}{\lambda^n}$ for all nonnegative integers n.

Markov chains (1)

Suppose X_n is a two-state Markov chain with transition matrix

$$Q = \begin{matrix} \\ 0 \\ 1 \end{matrix}\begin{matrix} \begin{matrix} 0 & 1 \end{matrix} \\ \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} \end{matrix}$$

Find the stationary distribution $\vec{s} = (s_0, s_1)$ of X_n by solving $\vec{s}Q = \vec{s}$, and show that the chain is reversible with respect to $\vec{s}$. **Answer:** The equation $\vec{s}Q = \vec{s}$ says that

$$s_0 = s_0(1-\alpha) + s_1\beta \text{ and } s_1 = s_0(\alpha) + s_0(1-\beta)$$

By solving this system of linear equations, we have

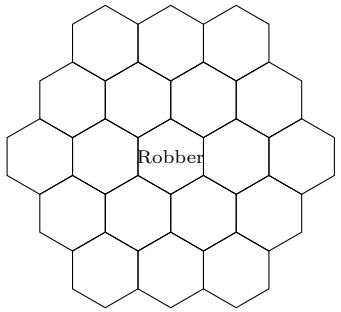
$$\boxed{\vec{s} = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)}$$

To show that the chain is reversible with respect to $\vec{s}$, we must show $s_i q_{ij} = s_j q_{ji}$ for all i, j. This is done if we can show $s_0 q_{01} = s_1 q_{10}$. And indeed,

$$s_0 q_{01} = \frac{\alpha\beta}{\alpha+\beta} = s_1 q_{10}$$

Markov chains (2)

William and Sebastian play a modified game of Settlers of Catan, where every turn they randomly move the robber (which starts on the center tile) to one of the adjacent hexagons.



(a) Is this Markov chain irreducible? Is it aperiodic? **Answer:** Yes to both. The Markov chain is irreducible because it can get from anywhere to anywhere else. The Markov chain is aperiodic because the robber can return back to a square in $2, 3, 4, 5, \ldots$ moves, and the GCD of those numbers is 1.

(b) What is the stationary distribution of this Markov chain? **Answer:** Since this is a random walk on an undirected graph, the stationary distribution is proportional to the degree sequence. The degree for the corner pieces is 3, the degree for the edge pieces is 4, and the degree for the center pieces is 6. To normalize this degree sequence, we divide by its sum. The sum of the degrees is $6(3) + 6(4) + 7(6) = 84$. Thus the stationary probability of being on a corner is $3/84 = 1/28$, on an edge is $4/84 = 1/21$, and in the center is $6/84 = 1/14$.

(c) What fraction of the time will the robber be in the center tile in this game, in the long run? **Answer:** By the above, $\boxed{1/14}$.

(d) What is the expected amount of moves it will take for the robber to return to the center tile? **Answer:** Since this chain is irreducible and aperiodic, to get the expected time to return we can just invert the stationary probability. Thus on average it will take $\boxed{14}$ turns for the robber to return to the center tile.

Problem-Solving Strategies

Contributions from Jessy Hwang, Yuan Jiang, Yuqi Hou

1. **Getting started.** Start by *defining relevant events and random variables*. ("Let A be the event that I pick the fair coin"; "Let X be the number of successes.") Clear notion is important for clear thinking! Then decide what it is that you're supposed to be finding, in terms of your notation ("I want to find $P(X = 3|A)$"). Think about what type of object your answer should be (a number? A random variable? A PMF? A PDF?) and what it should be in terms of.

   Try simple and extreme cases. To make an abstract experiment more concrete, try *drawing a picture* or making up numbers that could have happened. Pattern recognition: does the structure of the problem resemble something we've seen before?

2. **Calculating probability of an event.** Use counting principles if the naive definition of probability applies. Is the probability of the complement easier to find? Look for symmetries. Look for something to condition on, then apply Bayes' Rule or the Law of Total Probability.

3. **Finding the distribution of a random variable.** First make sure you need the full distribution not just the mean (see next item). Check the *support* of the random variable: what values can it take on? Use this to rule out distributions that don't fit. Is there a *story* for one of the named distributions that fits the problem at hand? Can you write the random variable as a function of an r.v. with a known distribution, say $Y = g(X)$?

4. **Calculating expectation.** If it has a named distribution, check out the table of distributions. If it's a function of an r.v. with a named distribution, try LOTUS. If it's a count of something, try breaking it up into indicator r.v.s. If you can condition on something natural, consider using Adam's law.

5. **Calculating variance.** Consider independence, named distributions, and LOTUS. If it's a count of something, break it up into a sum of indicator r.v.s. If it's a sum, use properties of covariance. If you can condition on something natural, consider using Eve's Law.

6. **Calculating $E(X^2)$.** Do you already know $E(X)$ or $\text{Var}(X)$? Recall that $\text{Var}(X) = E(X^2) - (E(X))^2$. Otherwise try LOTUS.

7. **Calculating covariance.** Use the properties of covariance. If you're trying to find the covariance between two components of a Multinomial distribution, X_i, X_j, then the covariance is $-np_ip_j$ for $i \neq j$.

8. **Symmetry.** If $X_1, \ldots, X_n$ are i.i.d., consider using symmetry.

9. **Calculating probabilities of orderings.** Remember that all $n!$ ordering of i.i.d. continuous random variables $X_1, \ldots, X_n$ are equally likely.

10. **Determining independence.** There are several equivalent definitions. Think about simple and extreme cases to see if you can find a counterexample.

11. **Do a painful integral.** If your integral looks painful, see if you can write your integral in terms of a known PDF (like Gamma or Beta), and use the fact that PDFs integrate to 1?

12. **Before moving on.** Check some simple and extreme cases, check whether the answer seems plausible, check for biohazards.

Biohazards

Contributions from Jessy Hwang

1. **Don't misuse the naive definition of probability.** When answering "What is the probability that in a group of 3 people, no two have the same birth month?", it is *not* correct to treat the people as indistinguishable balls being placed into 12 boxes, since that assumes the list of birth months {January, January, January} is just as likely as the list {January, April, June}, even though the latter is six times more likely.

2. **Don't confuse unconditional, conditional, and joint probabilities.** In applying $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, it is *not* correct to say "$P(B) = 1$ because we know B happened"; $P(B)$ is the *prior* probability of B. Don't confuse $P(A|B)$ with $P(A, B)$.

3. **Don't assume independence without justification.** In the matching problem, the probability that card 1 is a match and card 2 is a match is not $1/n^2$. Binomial and Hypergeometric are often confused; the trials are independent in the Binomial story and dependent in the Hypergeometric story.

4. **Don't forget to do sanity checks.** Probabilities must be between 0 and 1. Variances must be ≥ 0. Supports must make sense. PMFs must sum to 1. PDFs must integrate to 1.

5. **Don't confuse random variables, numbers, and events.** Let X be an r.v. Then $g(X)$ is an r.v. for any function g. In particular, X^2, $|X|$, $F(X)$, and $I_{X>3}$ are r.v.s. $P(X^2 < X|X \geq 0)$, $E(X)$, $\text{Var}(X)$, and $g(E(X))$ are numbers. $X = 2$ and $F(X) \geq -1$ are events. It does not make sense to write $\int_{-\infty}^{\infty} F(X)dx$, because $F(X)$ is a random variable. It does not make sense to write $P(X)$, because X is not an event.

6. **Don't confuse a random variable with its distribution.** To get the PDF of X^2, you can't just square the PDF of X. The right way is to use transformations. To get the PDF of $X + Y$, you can't just add the PDF of X and the PDF of Y. The right way is to compute the convolution.

7. **Don't pull non-linear functions out of expectations.** $E(g(X))$ does not equal $g(E(X))$ in general. The St. Petersburg paradox is an extreme example. See also Jensen's inequality. The right way to find $E(g(X))$ is with LOTUS.

Distributions in R

Command	What it does
`help(distributions)`	shows documentation on distributions
`dbinom(k,n,p)`	PMF $P(X = k)$ for $X \sim \text{Bin}(n, p)$
`pbinom(x,n,p)`	CDF $P(X \leq x)$ for $X \sim \text{Bin}(n, p)$
`qbinom(a,n,p)`	ath quantile for $X \sim \text{Bin}(n, p)$
`rbinom(r,n,p)`	vector of r i.i.d. $\text{Bin}(n, p)$ r.v.s
`dgeom(k,p)`	PMF $P(X = k)$ for $X \sim \text{Geom}(p)$
`dhyper(k,w,b,n)`	PMF $P(X = k)$ for $X \sim \text{HGeom}(w, b, n)$
`dnbinom(k,r,p)`	PMF $P(X = k)$ for $X \sim \text{NBin}(r, p)$
`dpois(k,r)`	PMF $P(X = k)$ for $X \sim \text{Pois}(r)$
`dbeta(x,a,b)`	PDF $f(x)$ for $X \sim \text{Beta}(a, b)$
`dchisq(x,n)`	PDF $f(x)$ for $X \sim \chi^2_n$
`dexp(x,b)`	PDF $f(x)$ for $X \sim \text{Expo}(b)$
`dgamma(x,a,r)`	PDF $f(x)$ for $X \sim \text{Gamma}(a, r)$
`dlnorm(x,m,s)`	PDF $f(x)$ for $X \sim \mathcal{LN}(m, s^2)$
`dnorm(x,m,s)`	PDF $f(x)$ for $X \sim \mathcal{N}(m, s^2)$
`dt(x,n)`	PDF $f(x)$ for $X \sim t_n$
`dunif(x,a,b)`	PDF $f(x)$ for $X \sim \text{Unif}(a, b)$

The table above gives R commands for working with various named distributions. Commands analogous to `pbinom`, `qbinom`, and `rbinom` work for the other distributions in the table. For example, `pnorm`, `qnorm`, and `rnorm` can be used to get the CDF, quantiles, and random generation for the Normal. For the Multinomial, `dmultinom` can be used for calculating the joint PMF and `rmultinom` can be used for generating random vectors. For the Multivariate Normal, after installing and loading the `mvtnorm` package `dmvnorm` can be used for calculating the joint PDF and `rmvnorm` can be used for generating random vectors.

Recommended Resources

- Introduction to Probability Book (`http://bit.ly/introprobability`)
- Stat 110 Online (`http://stat110.net`)
- Stat 110 Quora Blog (`https://stat110.quora.com/`)
- Quora Probability FAQ (`http://bit.ly/probabilityfaq`)
- R Studio (`https://www.rstudio.com`)
- LaTeX File (`github.com/wzchen/probability_cheatsheet`)

Please share this cheatsheet with friends!
`http://wzchen.com/probability-cheatsheet`

Table of Distributions

Distribution	PMF/PDF and Support	Expected Value	Variance	MGF
Bernoulli Bern(p)	$P(X=1)=p$ $P(X=0)=q=1-p$	p	pq	$q+pe^t$
Binomial Bin(n,p)	$P(X=k)=\binom{n}{k}p^k q^{n-k}$ $k\in\{0,1,2,\dots n\}$	np	npq	$(q+pe^t)^n$
Geometric Geom(p)	$P(X=k)=q^k p$ $k\in\{0,\ 1,\ 2,\ \dots\}$	q/p	q/p^2	$\frac{p}{1-qe^t}$, $qe^t<1$
Negative Binomial NBin(r,p)	$P(X=n)=\binom{r+n-1}{r-1}p^r q^n$ $n\in\{0,\ 1,\ 2,\ \dots\}$	rq/p	rq/p^2	$(\frac{p}{1-qe^t})^r$, $qe^t<1$
Hypergeometric HGeom(w,b,n)	$P(X=k)=\binom{w}{k}\binom{b}{n-k}/\binom{w+b}{n}$ $k\in\{0,1,2,\dots,n\}$	$\mu=\frac{nw}{b+w}$	$\left(\frac{w+b-n}{w+b-1}\right)n\frac{\mu}{n}(1-\frac{\mu}{n})$	messy
Poisson Pois(λ)	$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$ $k\in\{0,\ 1,\ 2,\ \dots\}$	λ	λ	$e^{\lambda(e^t-1)}$
Uniform Unif(a,b)	$f(x)=\frac{1}{b-a}$ $x\in(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$
Normal $\mathcal{N}(\mu,\sigma^2)$	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x\ -\ \mu)^2/(2\sigma^2)}$ $x\in(-\infty,\infty)$	μ	σ^2	$e^{t\mu+\frac{\sigma^2 t^2}{2}}$
Exponential Expo(λ)	$f(x)=\lambda e^{-\lambda x}$ $x\in(0,\infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$, $t<\lambda$
Gamma Gamma(a,λ)	$f(x)=\frac{1}{\Gamma(a)}(\lambda x)^a e^{-\lambda x}\frac{1}{x}$ $x\in(0,\infty)$	$\frac{a}{\lambda}$	$\frac{a}{\lambda^2}$	$\left(\frac{\lambda}{\lambda-t}\right)^a$, $t<\lambda$
Beta Beta(a,b)	$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ $x\in(0,1)$	$\mu=\frac{a}{a+b}$	$\frac{\mu(1-\mu)}{(a+b+1)}$	messy
Log-Normal $\mathcal{LN}(\mu,\sigma^2)$	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-(\log x-\mu)^2/(2\sigma^2)}$ $x\in(0,\infty)$	$\theta=e^{\mu+\sigma^2/2}$	$\theta^2(e^{\sigma^2}-1)$	doesn't exist
Chi-Square χ^2_n	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}$ $x\in(0,\infty)$	n	$2n$	$(1-2t)^{-n/2}$, $t<1/2$
Student-t t_n	$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1+x^2/n)^{-(n+1)/2}$ $x\in(-\infty,\infty)$	0 if $n>1$	$\frac{n}{n-2}$ if $n>2$	doesn't exist