STAT 110

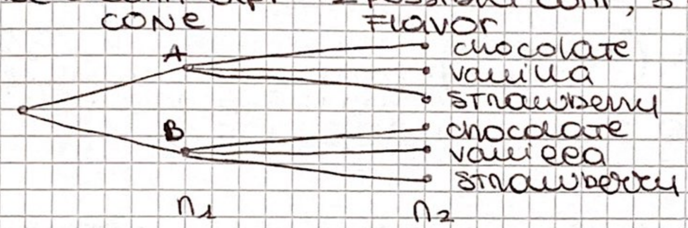
SAMPLE SPACE: set of all possible outcomes of an experiment

EVENT: subset of the sample space

NAIVE DEF: $\frac{\text{\# favorable outcomes}}{\text{\# possible outcomes}}$ IF all outcomes equally likely & finite sample

COUNTING:

- MULTIPLICATION RULE: sia expt 1 con n_1 possibili esiti per ogni possibile esito di expt 1 ci sono n_2 possibili esiti per expt 2 e così via fino ad avere n_r possibili esiti per expt r. Allora ci sono in TOTALE $n_1 \cdot n_2 \cdot \ldots \cdot n_r$ possibili esiti
- Ice cream expt: 2 possibili coni, 3 possibili gusti



TOT possibili esiti: $6 = 2 \cdot 3$

- se fai 10 scelte sequenziali, ogni volta scegliendo tra 2 possibili alternative, alla fine puoi ottenere 1024 (2^{10}) possibili esiti

COEFFICIENTE BINOMIALE

- $\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$, 0 if $k > n$

  "n choose k" / # of subsets of size k

  HOW many ways you can choose k out of the n?

  ORDER DOES NOT MATTER
- probability of a full house in poker, 5 cards hand

$$\frac{\binom{13}{1} \cdot \binom{4}{3} \cdot \binom{12}{1} \cdot \binom{4}{2}}{\binom{52}{2}}$$

full house: 3 carte dello stesso rank + 2 carte dello stesso rank

$\binom{13}{1}$: scelta del 1° rank

$\binom{4}{3}$: possibilità di prendere 3 carte delle 4 dello stesso rank

$\binom{12}{1}$: scelta del 2° rank

$\binom{4}{2}$: possibilità di prendere 2 carte delle 4 dello stesso rank

SAMPLING TABLE

	ORDER MATTERS	ORDER DOESN'T
REPLACE	n^k	$\binom{n-k+1}{k}$
DON'T REPLACE	$n \cdot (n-1) \cdot \ldots \cdot (n-k+1)$	$\binom{n}{k}$

→ They come from the multiplication rule! easy

$\binom{n}{0} = 1$, $\binom{n}{1} = n$

- example: 10 people split into team of 6, team of 4
  this means just $\binom{10}{4}$ which is $= \binom{10}{6}$ BUT
  if I say 10 people split into two teams of 5
  this means $\binom{10}{5}/2$ because you double count!
- example: how many ways are there to put k indistinguishable particles into n distinguishable boxes?

  $|\circ\circ\circ|\;|\quad|\;|\circ\circ|\;|\circ|$ → $\circ\circ\circ\,||\,\circ\circ\,|\,\circ$, dove | sono separatori

  $n = 4$, $k = 6$

  Allora $\binom{n+k-1}{k}$ → posizioni disponibili → posizioni occupabili dalle •

  che è uguale a $\binom{n+k-1}{n-1}$ → posizioni occupabili dai separatori

  (ciò dimostra ~~che~~ la casella in alto a dx)
- Definition:

  i.i.d. : independent and identically distributed

FUNCTION P

- PROBABILITY SPACE: consists of S and P where S is the sample space and P is a function the domain of which is all subsets of S. So P takes an event $A \subseteq B$ as input and returns $P(A) \in [0,1]$
- function P needs to satisfy two axioms:
  1. $P(\emptyset) = 0$; $P(S) = 1$
  2. $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$ if $A_1, \dots, A_n$ are disjointed
- proprietà
  1. $P(A^c) = 1 - P(A)$
  2. if $A \subseteq B$, then $P(A) \leq P(B)$
  3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  4. $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
  5. inclusion-exclusion rule general case:

$$P(A_1 \cup \dots \cup A_n) = \sum_{j=1}^{n} P(A_j) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \dots + (-1)^{n+1} P(A_1 \cap \dots \cap A_n)$$

cioè somma di tutte le singole P, sottrai le intersezioni doppie prese una volta sola, aggiungi le intersezioni triple prese una volta sola, sottrai quelle quadruple e così via fino ad aggiungere/sottrarre l'ultima intersezione completa

MONTMORT'S PROBLEM/MATCHING PROBLEM

deck of cards labelled 1 to n. One flips the cards one at a time and one counts from 1 to n. You win if you say the n^{th} numer and the value of that card corresponds to n.

let A_j be the event that the j^{th} card is also valued j. what's the P that AT LEAST 1 CARD matches?

$P(A_1 \cup \dots \cup A_n)$ = using inclusion-exclusion

This means that at least 1 matches

$P(A_j) = \frac{1}{n}$ since all position are equally likely for the card labeled j

$P(A_1 \cap A_2) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$

...

$P(A_1 \cap \dots \cap A_n) = \frac{(n-k)!}{n!} = \cancel{n} \cdot \frac{1}{\cancel{n}} - \frac{\cancel{n(n-1)}}{2!} \frac{1}{\cancel{n(n-1)}} +$

$\frac{\cancel{n(n-1)(n-2)}}{3!} \frac{1}{\cancel{n(n-1)(n-2)}}$

→

$$= 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \dots + (-1)^{n+1}\frac{1}{n!} \approx 1 - \frac{1}{e}$$

INDEPENDENT EVENTS

- events A,B are independent if $P(A \cap B) = P(A)P(B)$
- independence is $\neq$ from disjointness. if A,B are disjointed, if A occurs B can NOT possibly occur. if A,B are independent, if A occurs that tells us nothing about B occurring or not.
- events A,B,C are independent if: $P(A,B) = P(A)P(B)$; $P(A,C) = P(A)P(C)$; $P(B,C) = P(B)P(C)$ $P(A,B,C) = P(A)P(B)P(C)$ con la virgola che equivale all'intersezione
- Generalmente valida per n eventi $A_1, \dots, A_n$ prima a 2, poi a 3, poi a 4 fino al gruppo completo di tutte le intersezioni.

NEWTON PEPYS' PROBLEM

Having fair dice, which is more likely?

A) At least one 6 with 6 dice
B) " two 6's " 12 dice
C) " three 6's " 18 dice

$P(A) = 1 - \left(\frac{5}{6}\right)^6$, $P(B) = 1 - \left(\frac{5}{6}\right)^{12} - 12\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{11}$,

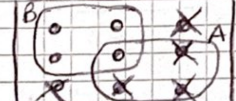
$$P(C) = 1 - \sum_{k=0}^{2} \underbrace{\binom{18}{k}\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{18-k}}_{\text{BINOMIAL PROBABILITY}}$$

so A is the most likely ($P(A) \approx 0.665$), C is the least likely (~ 0.597)

CONDITIONAL PROBABILITY

- Definition: $P(A|B) = \frac{P(A \cap B)}{P(B)}$ if $P(B) > 0$

example/interpretation (Pebbles)



B happens so B^c is irrelevant. $P(A|B)$ is $P(A \cap B)/P(B)$ because you rearrange the numbers to get the total mass of 1 again!

- Theorems:

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

$P(A_1, \dots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1,A_2)P(A_4|A_1,A_2,A_3) \dots$
$\dots P(A_n|A_1, \dots, A_{n-1})$

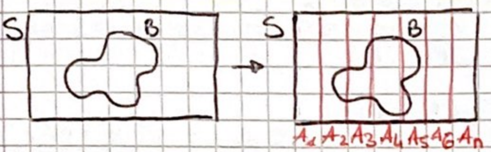
$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ → Bayes' Rule

- **Formula di Bayes**

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

solve a problem
1. try simple and extreme cases
2. break up problems into simpler pieces

LAW OF TOTAL PROBABILITY



$A_1, \dots, A_n$ is a PARTITION of S which means that $A_1, \dots, A_n$ are disjointed and $A_1 \cup \dots \cup A_n = S$

Then $P(B) = P(B \cap A_1) + \dots + P(B \cap A_n)$

$$= P(B|A_1)P(A_1) + \dots + P(B|A_n)P(A_n)$$

and the latter is the law of total probability

example: patient gets tested for disease that afflicts 1% of population. The test's outcome is positive. Test is advertised as "95%" accurate.
Given the events:
D: patient has the disease
T: patient tests positive

Suppose that "95%" accurate means:

$P(T|D) = 0.95 = P(T^c|D^c)$ → one interpretation of the 95% accuracy

But, what the patient cares about is whether they have the disease or not, so:

$$P(D|T) = \frac{P(T|D)\,P(D)}{P(T)}$$

$P(D)$ → 0.01 (1% of population)

$P(T)$ → we use law of tot probability for that

$$= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)} \simeq 0.16$$

so the partition of $P(T)$ is easy: either the patient has the disease or does not have it.

$P(T|D^c) = 0.05$ since $P(T^c|D^c) = 0.95$

0.16 means that the patient only has 16% chance to actually have the disease, given that they test positive

Common errors:
1. confusing $P(A|B)$ and $P(B|A)$
2. confusing $P(A)$ "prior" with $P(A|B)$ "posterior"
3. confusing independence with conditional independence

CONDITIONAL INDEPENDENCE

Events A,B are conditionally independent given C
if $P(A \cap B \mid C) = P(A \mid C)P(B \mid C)$
Does independence imply conditional independent and vice versa?

CONDITIONAL INDEPENDENCE GIVEN C DOES NOT IMPLY independence

Independence DOES NOT imply conditional independence

MONTY HALL'S PROBLEM / THREE DOORS PROBLEM

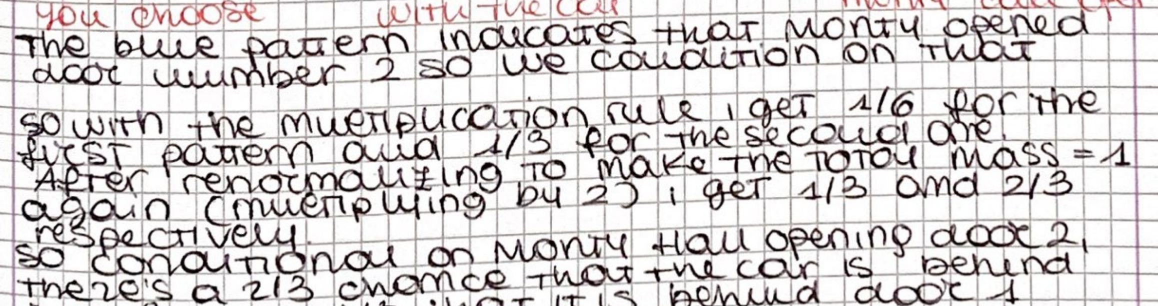
3 doors. 1 door has a car behind it, the other 2 have goats. The contestant has no idea of which door has what behind it, Monty Hall knows which door has what.
The contestant has to choose 1 door.

$\boxed{1}$ $\boxed{2}$ $\boxed{3}$

The contestant picks door 1.
Monty opens up either door 2 or 3 revealing a goat
example he opens door #2 and shows a goat.
so now the car is either behind door 1 or 3
Monty then gives the contestant the possibility to switch. Is it beneficial to switch your door?
Two more assumption: Monty always opens a goat door (obviously) and if he has a choice, he picks with equal probability (so every goat door is equally likely if you choose the car door)
SHOULD THE CONTESTANT SWITCH?

Initially all doors are $\frac{1}{3}$ (30% prob you get the car)
when Monty picks a goat door (example door #2) we know that door has a goat AND that Monty Hall opened the door. Let's see why it's relevant:

• Tree diagram:



The blue pattern indicates that Monty opened door number 2 so we condition on that

So with the multiplication rule I get 1/6 for the first pattern and 1/3 for the second one.
After renormalizing to make the total mass = 1 again (multiplying by 2) I get 1/3 and 2/3 respectively.
So conditional on Monty Hall opening door 2, there's a 2/3 chance that the car is behind door 3 and 1/3 that it is behind door 1
SO IT IS BENEFICIAL TO SWITCH

- LAW OF TOTAL PROBABILITY
we wish we knew where the car is so this is what we condition on.
let S be the event we succeed (assuming switch)
let D_j be the event that door j has the car ($j=1,2,3$)
now $P(S) = P(S|D_1)P(D_1) + P(S|D_2)P(D_2) + P(S|D_3)P(D_3)$
$= P(S|D_1)\frac{1}{3} + P(S|D_2)\frac{1}{3} + P(S|D_3)\frac{1}{3}$

so in this case we got the door right and then switched so it's bad so it's gonna be 0

those two cases are gonna work so 1

$= 0 + 1\frac{1}{3} + 1\frac{1}{3} = 2/3$

By simmetry we also have:
$P(S \mid \text{monty opens door } 2) = 2/3$

so both the conditional & unconditional probability are 2/3

SIMPSON'S PARADOX

- you have 2 doctors. the first has a better success rate for every category of possible surgery taken individually, yet the second has a better success rate overall. is that possible? YES
  the signs of inequalities can flip when you aggregate data/categories together
- example: 2 doctors, 2 types of surgery

Dr. Hibbert

	Heart	Bandaid
success	70	10
failure	20	0

Dr. Nick

	Heart	Bandaid
success	2	81
failure	8	9

100 total surgeries for both doctors with a success rate: Hibbert: 80% Nick: 83%

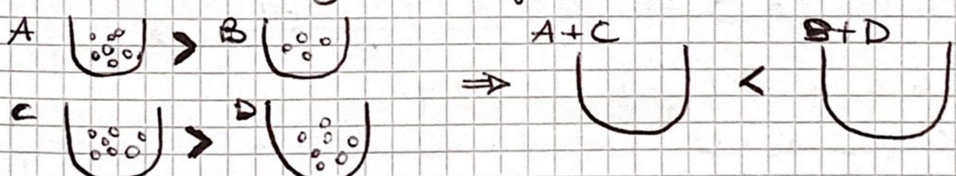
when we aggregate data the direction may flip!

- express the paradox through conditional prob.
  let A be the event that the surgery is successful
  & B be the event that we're treated by Dr Nick
  & C be the event that we have heart surgery.
  Then, $P(A|B,C) < P(A|B^c,C)$
  $P(A|B,C^c) < P(A|B^c,C^c)$
  But $P(A|B) > P(A|B^c)$

  pretty much the general setup of the Simpson's Paradox

  C is called a "confounder" which is a control since if you only condition on event B the results are misleading.
  so, sure you want to condition on event B but it's also necessary for the results to be fair that you also condition on C.

- examples: two types of jelly beans you like one type better. you have 4 jars of jelly beans, each containing both types.



then A+C < B+D is a possible scenario.

GAMBLER'S RUIN / LA ROVINA DEL GIOCATORE

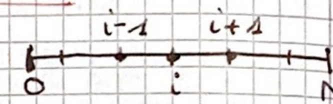
- TWO gamblers play the same game over and over betting 1\$ each time. Is any of the two ever going to bankrupt?

Let A and B be the two gamblers, betting 1\$ each round of a sequence of rounds which are independent.
For each round, let

p be $p = P(\text{A wins a certain round})$ and
q be $q = 1 - p = P(\text{B wins a certain round})$

Find the probability that A wins the entire game (so B is "ruined") assuming that A start with \$ i and B starts with \$ $(N-i)$
(so the total is N)

Try to see the recursive structure (every round is the same problem)



Strategy: condition on the first step
Let $P_i = P(\text{A wins entire game} \mid \text{A starts with } i \text{ dollars})$
Then $P_i = p P_{i+1} + q P_{i-1}$ for $1 \leq i \leq N-1$
with the extreme cases of $P_0 = 0$ $P_N = 1$
$P_i = pP_{i+1} + qP_{i-1}$ is called the DIFFERENCE EQUATION
(è il corrispondente discreto dell'eq. differenziale)
Solve it:

$P_i = pP_{i+1} + qP_{i-1}$ Guess $P_i = x^i$

$x^i = p x^{i+1} + q x^{i-1}$

$p x^2 - x + q = 0 \rightarrow x = \dfrac{1 \pm \sqrt{1-4pq}}{2p} = \begin{cases} 1 \\ q/p \end{cases}$

$P_i = A\,1 + B\left(\frac{q}{p}\right)^i$, $p \neq q$

If we put $P_0 = 0$, $P_N = 1$ so $B = -A$, $1 = A\left(1 - \frac{q}{p}\right)$

$$\Rightarrow P_i = \begin{cases} \dfrac{1 - \left(\frac{q}{p}\right)^i}{1 - \left(\frac{q}{p}\right)^N} & \text{if } p \neq q \\[2ex] i/N & \text{if } p = q \end{cases}$$

$i/N \downarrow$ lo ottiene tramite un limite (minuto 26 lez 7)

- example with numbers: let $i = N - i$ $p = 0.49$

Then
$N = 20 \rightarrow 0.40$
$N = 100 \rightarrow 0.12$
$N = 200 \rightarrow 0.02$
} chance that A wins

with SAME amount of \$ and just 0.01 difference in p

RANDOM VARIABLES

A random variable is a function from the sample space S to the real line $\mathbb{R}$.
Think of a random variable as a numerical "summary" of an aspect of the experiment.
The "randomness" comes from the experiment itself.

BERNOULLI DISTRIBUTION (Bernoulli (P)) → probability of success

A random variable X is said to have a Bernoulli distribution if X has only 2 possible values, 0 and 1, and $P(X=1)=p$, $P(X=0)=1-p$,
event

BINOMIAL (n,p) DISTRIBUTION

The distribution of the number of successes (X) in n independent Bernoulli (p) trials is called the binomial (n,p) distribution. Its distribution is given by:

with n = trials, k = # of successes and $0 \leq k \leq n$

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

and that is called the probability-mass function (PMF). Var(x): *

OBSERVATION:

if $X \sim Bin(n,p)$, $Y \sim Bin(m,p)$ with X,Y independent
Then $X+Y \sim Bin(n+m, p)$

HOW TO THINK ABOUT THE BINOMIAL DISTRIBUTION

1) X is the # of successes in n independent Bernoulli(p) trials
2) sum of indicator random variables
ex. $X = X_1 + X_2 + \ldots + X_n$ with $X_j = \begin{cases} 1 & \text{if } j\text{th trial is a success} \\ 0 & \text{otherwise} \end{cases}$
with $X_1, \ldots, X_n$ i.i.d. Bern(p) (iid = indep. & identically distributed)
3) PMF $P(X=k) = \binom{n}{k} p^k q^{n-k}$, $q = 1-p$ and $0 \leq k \leq n$

CUMULATIVE DISTRIBUTION FUNCTIONS (CDF)

CDF is a way to describe the distribution of a variable

$X \leq x$ is an event
$F(x) = P(X \leq x)$
Then F is the CDF of X

PROBABILITY MASS FUNCTION (PMF)

for discrete random variables only
(DISCRETE: possible values are $a_1, \ldots, a_n$ or $a_1, a_2, \ldots$)
The PMF is $P(X = a_j)$ for all j (cioè l'insieme di tutte le p_j, cioè probabilità che $X = a_j$)

* Var(x) given $X \sim Bin(n,p)$: $Var(X) = npq$, $q = 1-p$

- conditions PMF of X has to satisfy:

$p_j \geq 0$

$\sum_j p_j = 1$

in fact $\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1^n = 1$

by binomial theorem

- Given $X \sim Bin(n,p)$, $Y \sim Bin(m,p)$ then $X+Y \sim Bin(n+m,p)$ with X, Y independent, then use the PMF to describe X+Y

$P(X+Y=k) = \sum_{j=0}^{k} P(X+Y=k \mid X=j) P(X=j)$ so we condition on X=j with the LOTP

$= \sum_{j=0}^{k} P(Y=k-j \mid \cancel{X=j}) \binom{n}{j} p^j q^{n-j} =$ so we use the Binomial distr. itself

we cross X=j out because they are independent!

$= \sum_{j=0}^{k} \binom{m}{k-j} p^{k-j} q^{m-k+j} \binom{n}{j} p^j q^{n-j} =$

$= p^k q^{m+n-k} \underbrace{\sum_{j=0}^{k} \binom{m}{k-j}\binom{n}{j}}$

VANDERMONDE IDENTITY

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

so the PMF proves X+Y to have Bin(n+m, p) distribution

- VANDERMONDE IDENTITY

$$\sum_{j=0}^{k} \binom{m}{k-j}\binom{n}{j} = \binom{m+n}{k}$$

- NOTE: in the Binomial distribution, all trials are independent and have the same probability of success

HYPERGEOMETRIC DISTRIBUTION

We have a jar full of marbles. b of them are black and w of them are white. We pick simple random samples (which means that all samples of that size are equally likely) of size n. Find the distribution of the # of white marbles in the sample

X = # white marbles in sample

$$P(X=k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$$

→ The other n−k marbles are taken from the b

exactly k of the w marbles have to be chosen

→ we choose n out of the total w+b marbles

with $0 \le k \le w$, $0 \le n-k \le b$

so $P(X=k) = \dfrac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$ is its **PMF**

The key of the Hypergeometric distribution is that the sampling is done WITHOUT REPLACEMENT

(if the sampling was done WITH REPLACEMENT then X has a BINOMIAL DISTRIBUTION)

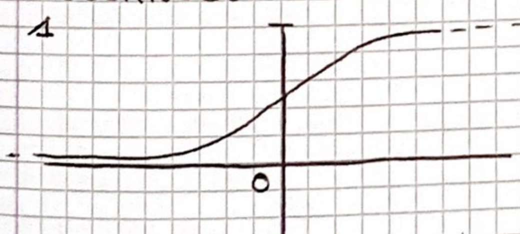
- Does the PMF sums to 1? (condition)

$$\sum_{k=0}^{w} \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}} = 1 \text{ by Vandermonde}$$

$\binom{w+b}{n}$ constant

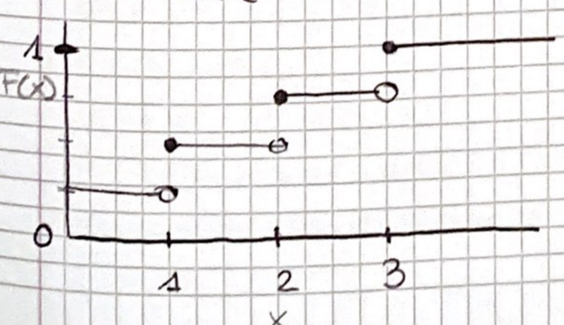
CDF: $F(x) = P(X \le x)$, as a function of real x

continuous



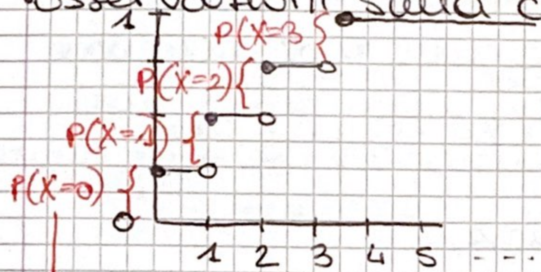
→ DOES NOT REACH 1

discrete (for x's possible values = 0, 1, 2, 3)



→ DOES REACH 1 AND STAYS THERE

• Osservazioni sulla CDF



→ THIS DOES REACH 1 AND THEN STAYS THERE FOREVER

The jump is the $P(X = \text{value})$

The jump sizes are the PMF of X

- the CDF gives you the whole distribution
- example: Find $P(1 < X \leq 3)$ using F

  $$P(X \leq 1) + \underbrace{P(1 < X \leq 3)}_{\text{THAT'S WHAT I WANT}} = P(X \leq 3)$$

  $$\Rightarrow P(1 < X \leq 3) = F(3) - F(1)$$

  General: $P(a < X \leq b) = F(b) - F(a)$
- Properties of CDF:
  1) increasing (not strictly)
  2) right continuous (so it's "continuous from the right")
  3) as $X \to -\infty$, $F(x) \to 0$
     as $X \to +\infty$, $F(x) \to 1$

INDEPENDENCE OF RANDOM VARIABLES r.v.s

X, Y are independent if $P(X \le x, Y \le y) = P(X \le x)\,P(Y \le y)$ for all x, y

In the discrete case that's equivalent to
$P(X = x, Y = y) = P(X=x)\,P(Y=y)$

AVERAGES

When you say average you usually mean the usual MEAN or EXPECTED VALUE, which are all equivalent terms.

average/mean/exp. value = $\dfrac{\text{sum of all values}}{\text{\# of values}}$

- average/mean/exp. value

$1, \dots, 6 \rightarrow \dfrac{1+2+3+4+5+6}{6} = 3.5$

$1, \dots, 100 \rightarrow \dfrac{1+\dots+100}{100} = 50.5$

$1, \dots, n \rightarrow \dfrac{1+\dots+n}{n} = \dfrac{1+n}{2}$ = result !♡

$\dfrac{1}{n}\sum_{j=1}^{n} j = \dfrac{n+1}{2}$ → arithmetic series

- weighted average (altro modo di fare la media normale)

1, 1, 1, 1, 1, 3, 3, 5

$\rightarrow \dfrac{5}{8}\cdot 1 + \dfrac{2}{8}\cdot 3 + \dfrac{1}{8}\cdot 5 = 2$ same as $\dfrac{1+1+1+\dots+5}{8}$

- average of a discrete r.v. X

expected value(X) = $E(X) = \sum_x \underbrace{x}_{\text{value}}\ \underbrace{P(X=x)}_{\text{PMF}}$

summed over x with $P(X=x) > 0$

- example

$X \sim \text{Bern}(p)$

$E(X) = 1\,P(X=1) + 0\,P(X=0) = p$ by definition

- example

$X = \begin{cases} 1, & \text{if } A \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$ } → indicator r.v.

then $E(X) = P(A) = p$ again

- example $X \sim Bin(n,p)$

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

$$= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{n-k} = np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{n-j-1}$$

with $j = k-1$

$$= np \underbrace{\sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{n-j-1}}_{1} = np$$

LINEARITY

$E(X+Y) = E(X) + E(Y)$

even if X, Y are dependent

$E(cX) = cE(X)$

with c constant

- so now we can do $X \sim Bin(n,p)$ again:

since $X \sim Bin(n,p)$ is $X \sim Ber(p)$ n times we just think about $E(X \sim Ber(p)) = p$ n times

so $n(E(X \sim Ber(p))) = np$

since $X = X_1 + \dots + X_n$ with $X_{1,\dots,n} \sim Ber(p)$

GEOMETRIC DISTRIBUTION Geom(p)

As with the binomial, we have independent Bernoulli trial each one with probability of success p so Bern(p).
We consider the number of failures before the 1^{st} success (the success is not included in the count)

- PMF: $P(X=k) = q^k p$ with $X \sim Geom(p)$
  $q = 1-p$
  $k = 0,1,2,\dots$ until the 1^{st} success

This is a valid PMF since $\sum_{k=0}^{\infty} pq^k = \frac{p}{1-q} = 1$

- EXPECTED VALUE of $X \sim Geom(p)$

$$E(X) = \sum_{k=0}^{\infty} kpq^k = p\sum_{k=1}^{\infty} kq^k = \frac{pq}{p^2} = q/p$$

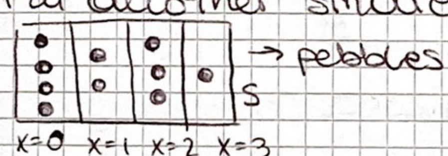
LINEARITY OF EXPECTATION

- let $T = X+Y$. show $E(T) = E(X)+E(Y)$, so

$$E(T) = \sum_t tP(T=t) \stackrel{?}{=} \sum_x xP(X=x) + \sum_y yP(Y=y)$$

let $P(T=t) = \sum_x P(T=t \mid X=x)P(X=x)$ THIS STRATEGY DOESN'T WORK

Try another strategy:



$$E(X) = \underbrace{\sum_x xP(X=x)}_{\text{GROUPED}} = \underbrace{\sum_s X(s)P(\{s\})}_{\text{UNGROUPED}}$$

X is a function so this makes sense

mass of the pebbles

then proof of linearity (discrete case)

$$E(T) = \sum_s (X+Y)(s)P(\{s\}) = \sum_s (X(s)+Y(s))P(\{s\}) =$$
$$= \sum_s X(s)P(\{s\}) + \sum_s Y(s)P(\{s\}) = E(X)+E(Y) \quad \square$$

- show $E(cX) = cE(X)$ if c const.

we prove it through dependency $X=Y$

$$E(X+Y) = E(2X) = 2E(X) = E(X)+E(Y) \quad \square$$

NEGATIVE BINOMIAL DISTRIBUTION

not negative, not binomial it's a generalisation of the geometric distribution with parameters r, p

we have independent Bern(p) trials and we want to know the number of failures until the r^{th} success.

- PMF : $P(X=n) = \binom{n+r-1}{r-1} p^r (1-p)^n$ for $n = 0, 1, 2 \ldots$

based on example

1 0 0 0 1 0 0 1 0 0 0 0 1 0 0 $\boxed{1}$ r^{th} success

$r=5$ $n=11$ total trials $=16$

- EXPECTED VALUE

$E(X) = E(X_1 + \ldots + X_r)$ with $X_j = \#$ failures between $(j-1)^{st}$ and j^{th} success

and $X_j \sim \text{Geom}(p)$

so $E(X) = E(X_1) + \ldots + E(X_r) = \frac{rq}{p}$

FIRST SUCCESS DISTRIBUTION

$X \sim FS(p)$ time until first success, counting the first success.

let $Y = X-1$ then $Y \sim \text{Geom}(p)$

so $E(X) = E(Y) + 1 = \frac{q}{p} + 1 = \frac{1}{p}$

RULE

The expected value of an indicator is the probability of the event

ST PETERSBURG PARADOX

Game: we will flip a fair coin over and over until the coin lands heads for the 1st time. If the coin lands head on the 1st trial you get \$2, if on the 2nd trial you get \$4, if on the 3rd you get \$ 8, if on the 4th you get \$ 16 and so on (it doubles every time)

you get \$ 2^x where x = # of flips of the coin until it lands head (H), including the success

How much should you be willing to pay to play this game?

Let $Y = 2^x$ (your payoff), find $E(Y)$

$$E(Y) = \sum_{k=1}^{\infty} 2^k \cdot \frac{1}{2^k} = \sum_{k=1}^{\infty} 1 = 1+1+1+\dots = \infty$$

so you should be willing to pay \$$\infty$

so what you need to think about is that in the real world you have some upper bounds (no one has ∞ money). suppose that you can't get more than a trillion dollars from the game.

bound at 1 trillion $\sim \$2^{40}$

then: $\sum_{k=1}^{40} 2^k \cdot \frac{1}{2^k} = 40$

which means \$40 is what you should be willing to pay (in the case in which if you win the 41st he flees the country, or \$41 in the case he does actually pay you)

$E(2^x) = \infty$

POISSON DISTRIBUTION

- PMF: $P(X=k) = e^{-\lambda} \frac{\lambda^k}{k!}$, $k \in \{0,1,2...\}$

  λ is the "rate" parameter , $\lambda > 0$
- proof the PMF is valid:

  $$\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \boxed{\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}} \text{ TAYLOR SERIES FOR } e^{\lambda} = e^{-\lambda} e^{\lambda} = 1$$
- EXPECTED VALUE:

  Given $X \sim Pois(\lambda)$

  $E(X)$ = value · probability of the value =

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

  $$= \lambda e^{-\lambda} \boxed{\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}} \text{ TAYLOR SERIES FOR } e^{\lambda} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$
- The Poisson distribution is often used for applications where we're counting the # of something (ex. successes) where we have a large # of trials, each of them having a small probability of success.

  examples: # of emails in an hour

  \# of choc chips in a choc chips cookie
- POISSON PARADIGM / APPROXIMATION

  Events $A_1, ..., A_n$ with $P(A_j) = p_j$ and n large, p_j's all small. events independent or "weakly dependent."

  Then the # of A_j's that occur is approximately Poisson. so it's clear that

  $$\lambda = \sum_{j=1}^{n} p_j$$
- The $Bin(n,p)$ does converge to $Pois(\lambda)$ when n is large and p is small
- observation: the probabilities of the n trials in a Pois distribution can be different
- Variance

  $Var(X) = \lambda$

  so $X \sim Pois(\lambda)$ has mean λ , variance λ

- $X \sim \text{Bin}(n,p)$, let $n \to \infty$, let $p \to 0$ so that $\lambda = np$ is held constant (so $n \to \infty$ and $p \to 0$ at the same rate). What happens to the PMF?

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \text{ fixed}$$

$$= \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!\, n^k} \lambda^k \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\to 1}$$

$$= \frac{\lambda^k}{k!} e^{-\lambda} = \text{Pois PMF at } k$$

↳ so Bin converges to Pois here

- **DISCRETE**

X

PMF $P(X=x)$

CDF $F_X(x) = P(X \le x)$

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

$Var(X) = E(X^2) - (E(X))^2$

LOTUS $E(g(X)) = \sum_x g(x) P(X=x)$

CONTINUOUS

X

PDF $f_X(x)$

CDF $F_X(x) = P(X \le x)$

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

$Var(X) = E(X^2) - (E(X))^2$

LOTUS $E(g(X)) = \sum_x$

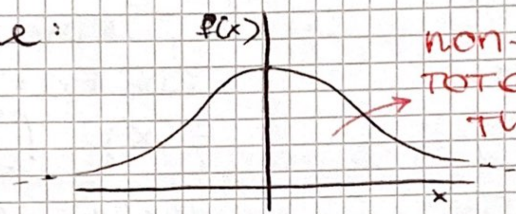
PDF - Probabiliy Density Function

- **Definition**: R.V X has PDF $f(x)$ if $P(a \leq X \leq b) =$
  $= \int_a^b f(x)\,dx$, for all a, b

so PDF is what you intepgrate to get the probability

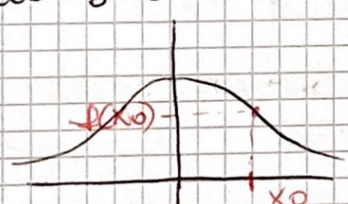
- remember if $a = b \Rightarrow \int_a^a f(x)\,dx = 0$
  the probability of any specific point is zero
- For a PDF to be valid we need two conditions to be true:
  1. $f(x) \geq 0$
  2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$

example:



non-negative and total area under the curve = 1

imagine:

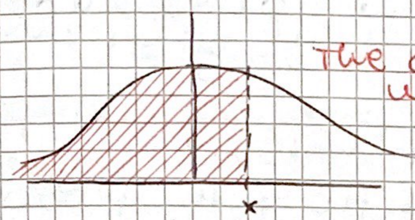


$f(x_0)$ is a density but we can think that the probability falls in a small interval like:

$$f(x_0)\,\varepsilon \approx P\left(X \in \left(x_0 - \frac{\varepsilon}{2}, x_0 + \frac{\varepsilon}{2}\right)\right) \text{ for a small } \varepsilon$$

- if X has PDF f, the CDF is: $F(x) = P(X \leq x) =$
  $= \int_{-\infty}^{x} f(t)\,dt$

so:



The area under the curve until x

- if X has CDF F (and X is a continuous rv), then the PDF is: $f(x) = F'(x)$
  by the fundamental theorem of calculus
- also: $P(a < X < b) = \int_a^b f(x)\,dx = F(b) - F(a)$

RELATION BETWEEN PDF and CDF (continuous case)

X has PDF f, then CDF: $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$

X has CDF F (X continuous r.v.), then PDF: $f(x) = F'(x)$

also: $P(a < X < b) = \int_a^b f(x)\,dx = F(b) - F(a)$

VARIANCE

$Var(X) = E(X - E(X))^2$ → This changes the unit

↳ for both cont. & discrete cases

STANDARD DEVIATION

$SD(X) = \sqrt{Var(X)}$ → This changes it back

ANOTHER WAY TO EXPRESS VARIANCE

$Var(x) = E(X^2 - 2X(EX) + (EX)^2) = E(X^2) + 2E(X)E(X) +$
$+ (EX)^2 = E(X^2) - (E(X))^2$

remember: if X is a constant, $var(X) = 0$

UNIFORM DISTRIBUTION Unif(a,b)

Given an $[a,b]$ interval, saying that x is a completely random point in $[a,b]$ means that if we split $[a,b]$ in halves, x has the same probability to be in the 1st or 2nd half since they have the same lenght

so uniform means that probability is proportional to lenght (prob $\propto$ lenght)

c d — prob of c is half the prob of d

- **PDF**: $f(x) = \begin{cases} c \text{ constant}, & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$

$$1 = \int_a^b c\, dx = c = \frac{1}{b-a}$$

- **CDF**: $F(x) = \int_{-\infty}^{x} f(t)\, dt = \int_a^x f(t)dt = \begin{cases} 0 & \text{if } x<a \\ \frac{x-a}{b-a} & \text{if } a\leq x\leq b \\ 1 & \text{if } x>b \end{cases}$

as you increase x, the probability is increasing linearly

- **$E(X)$** $= \int_a^b \frac{x}{b-a}\, dx = \frac{x^2}{2(b-a)}\Big|_a^b = \frac{1}{2(b-a)}(b-a)(b+a) =$

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

now we try to get $E(x^2)$ to get the variance
$E(x^2) = E(Y)$ given $Y=X^2$, so we need the PDF of Y, and that's a problem. let's try another way: $E(x) = \int_{-\infty}^{\infty} x^2 f_x(x)\, dx$

→ Law of the unconscious statistician (LOTUS)
$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$ and that's TRUE :)

let $U \sim$ Unif$(0,1)$, $E(U) = \frac{1}{2}$, $E(U^2) = \int_0^1 u^2 f_u(u) du$
$= \frac{1}{3}$

$\Rightarrow Var(U) = E(U^2) - (EU)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$

- **$Var(X)$** $= E(X^2) - (EX)^2$

UNIVERSALITY OF THE UNIFORM

Let $U \sim Unif(0,1)$ and F be a CDF (assume F is strictly increasing and continuous as a function). Then if we le $X = F^{-1}(U)$, then $X \sim F$ (so X is distributed according to F)

Let's prove it: $P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) =$
$= F(x)$ $\square$

so X will have CDF $= F$! :)

ALSO: if $X \sim F$, then $F(X) \sim Unif(0,1)$

a function of a r.v. is a r.v.

Be careful: $F(x) = P(X \leq x)$, then $F(X) = P(X \leq X) = 1$? FALSE!

INDEPENDENCE OF r.v.s $X_1, \dots, X_n$

- Definition: $X_1, \dots, X_n$ are independent if
$P(X_1 \leq x_1, \dots, X_n \leq x_n) = \underbrace{P(X_1 \leq x_1) \cdot \ldots \cdot P(X_n \leq x_n)}_{\text{these are all CDF's}}$ for all $x_1, \dots, x_n$

then, $P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$ is the JOINT CDF

- in the discrete case it's easier to work with PMF so $P(X_1 = x_1, \dots, X_n = x_n)$ (joint PMF) $= P(X_1 = x_1) \cdot \ldots \cdot P(X_n = x_n)$
- These two examples of full independence are stronger than just paired independence
- Example: $X_1, X_2 \sim Bern(\frac{1}{2})$ (2 coins), $X_3 = \begin{cases} 1 & \text{if the 2 coins match} \\ 0 & \text{otherwise} \end{cases}$

$\Rightarrow$ these are pairwise independent but NOT independent in fact, knowing X_1, X_2 gives us total information about X_3, knowing X_1 tells us nothing about X_2 or about X_3, knowing X_2 tells us nothing about X_1/X_3 and knowing X_1, X_3 tells us X_2 (same for $X_2, X_3 \Rightarrow X_1$)

$\Rightarrow$ pairwise independence isn't enough to have total independence

NORMAL DISTRIBUTION / GAUSSIAN

Central Limit Theorem: if you add up a bunch of i.i.d random variables the distribution is going to look like a normal distribution

- standard normal $N(0,1)$

the mean is 0 and the variance is 1

PDF: $f(z) = c\, e^{-z^2/2}$

$\underbrace{c}_{\text{normalizing constant}}$

with $c = \frac{1}{\sqrt{2\pi}}$

$E(Z)$ given $Z \sim N(0,1)$:

$E(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \underbrace{z e^{-z^2/2}}_{\text{This is an odd function}} dz = 0$ by simmetry

$Var(Z) = E(Z^2) - (\underbrace{E(Z)}_{0})^2 = E(Z^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \underbrace{z^2 e^{-z^2/2} dz}_{\text{LOTUS}}$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz$$

$$= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} z^2 e^{-z^2/2} dz = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} \underbrace{z}_{u} \cdot \underbrace{z \cdot e^{-z^2/2}}_{dv} dz \quad *$$

now $u = z$, $du = dz$

$dv = z e^{-z^2/2}$, $v = -e^{-z^2/2}$

so $* = \frac{2}{\sqrt{2\pi}} \left(\underbrace{(uv)\Big|_0^{\infty}}_{=0} + \int_0^{\infty} e^{-z^2/2} dz \right) = 1$

Notation: Φ is the standard normal's CDF

$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$

$\Phi(-z) = 1 - \Phi(z)$ by simmetry

- recap of standard normal:

$Z \sim N(0,1)$ / CDF Φ / $E(Z)=0$ / $Var(Z) = E(Z^2) = 1$,

$\left.\begin{array}{l} E(Z)=0 \text{ first moment} \\ E(Z^2)=1 \text{ second moment} \\ E(Z^3)=0 \text{ third moment} \\ \quad \dots \end{array}\right\}$ odd moments = 0

$-Z \sim N(0,1)$ ($-Z$ is still a standard normal) because by adding a constant we just shift the location

- General normal

Let $X = \mu + \sigma Z$, $\mu \in \mathbb{R}$ (mean/location), $\sigma > 0$ (st. deviation or scale)

because we just rescale everything by multiplying by a constant

Then $X \sim N(\mu, \sigma^2)$

mean, variance

$Var(X) = E((X-E(X))^2) = EX^2 - (EX)^2$

$Var(X+c) = Var(X)$, $Var(cX) = c^2 Var(X)$

$Var(X+Y) \neq Var(X) + Var(Y)$ in general (not linear)

(equal if X, Y are independent)

mean $E(X) = \mu$

variance $Var(\mu + \sigma Z) = \sigma^2 \, Var(Z) = \sigma^2$

standardization $Z = \dfrac{X-\mu}{\sigma}$

it gives us the standard normal starting from a general normal

PDF: of $X \sim N(\mu, \sigma^2)$

CDF: $P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$

dimensionless quantity

$\Rightarrow$ PDF: $\dfrac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2}$

$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

$X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$

68 - 95 - 99.7 % rule

it tells us how likely is it that a normal random variable will be a certain distance from its mean measured in terms of standard deviation, given $X \sim N(\mu, \sigma^2)$

$P(|X-\mu| \leq \sigma) \approx 0.68$

the distance is ~~more~~ less than 1 st. deviation away from the mean

$P(|X-\mu| \leq 2\sigma) \approx 0.95$

$P(|X-\mu| \leq 3\sigma) \approx 0.997$

PROVE LOTUS FOR DISCRETE SAMPLE SPACE

SHOW $E(g(X)) = \sum_x g(x)P(X=x)$

$$\underbrace{\sum_x g(x)P(X=x)}_{\text{grouped case}} = \underbrace{\sum_{s \in S} g(X(s))P(\{s\})}_{\text{ungrouped case}}$$

grouped case → This says first combine all the pebbles that have the same value of x into "super-pebbles" then average

ungrouped case → This one says take each pebble, compute the function and take the weighted average ($P(\{s\})$ are the weights)

COUPON/TOY COLLECTOR

Works like the Happy Meal, so each time you get a toy/coupon. How long will it take you to get the full set of toys/coupons

n: toy types (equally likely)

Find expected time to get the full set with time measured discretely (so how many toys you'll have to buy to have them all)

T: # of toys we have to buy to get a full set

$T = T_1 + T_2 + \dots + T_n$

T_1 = time until the 1st new toy = 1

T_2 = time until the 2nd new toy

$T_3 = \dots$

...

$T_n = \dots$

now think about:

$T_1 = 1, \quad T_2 - 1 \sim \text{Geom}\left(\frac{n-1}{n}\right)$, ...

$T_j - 1 \sim \text{Geom}\left(\frac{n-(j-1)}{n}\right)$

also:
$$E(T) = E(T_1) + E(T_2) + \dots + E(T_n)$$
$$= 1 + \frac{n}{n-1} + \frac{n}{n-2} + \dots + \frac{n}{1} =$$
$$= n\left(1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n}\right) \approx n \log n$$

for large n

Logistic Distribution

CDF: $F(x) = \frac{e^x}{1+e^x}$

$u \sim Unif(0,1)$, consider $\underbrace{F^{-1}(u) = \log \frac{u}{1-u}}_{\text{this is logistic}}$

Example with linearity & simmetry

Let X, Y, Z be i.i.d positive r.v.s. Find $E\left(\frac{X}{X+Y+Z}\right)$

Think about the simmetry since they're i.i.d.

$$E\left(\frac{X}{X+Y+Z}\right) = E\left(\frac{Y}{X+Y+Z}\right) = E\left(\frac{Z}{X+Y+Z}\right)$$

It's not that $X=Y$ or $X=Z$ but since they're i.i.d. by simmetry it's not relevant how we order them.

now by linearity

$$E\left(\frac{X}{X+Y+Z}\right) + E\left(\frac{Y}{X+Y+Z}\right) + E\left(\frac{Z}{X+Y+Z}\right) = E\left(\frac{X+Y+Z}{X+Y+Z}\right) = 1$$

but this is the same thing added 3 times

$$3E\left(\frac{X}{X+Y+Z}\right) = 1 \implies E\left(\frac{X}{X+Y+Z}\right) = \frac{1}{3}$$

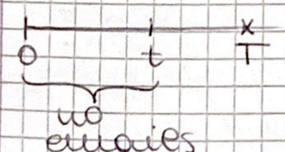
Example with Poisson

Suppose that the # of emails I get in an interval of time t is distributed $Pois(\lambda t)$ (so λt is both variance & mean). Find the PDF of T, the time ~~when~~ the 1st emails arrives

→ This connects continuous & discrete

It's easier to find the complement:

$P(T > t)$ so $1 - CDF$ (?) (cioè $P(X \geq x) = 1 - P(X \leq x)$)



$$P(T > t) = P(N_t = 0) \text{ where } N_t = \#\text{emails in } [0,t]$$

$$= e^{-\lambda t} \frac{(\lambda t)^0}{0!} = e^{-\lambda t}$$

CDF is $1 - e^{-\lambda t}$, $t > 0$ ☺

EXPONENTIAL DISTRIBUTION

one parameter (λ) that we can call rate parameter, because it's the rate at which some type of event occurs

$X \sim Expo(\lambda)$

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

valid PDF because: $\int_0^\infty \lambda e^{-\lambda x} dx = 1$

CDF: $F(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}, \quad x > 0$

if we let $Y = \lambda X$, then $Y \sim Expo(1)$, so the CDF of Y is $P(Y \leq y) = P\left(X \leq \frac{y}{\lambda}\right) = 1 - e^{-y}$. we want to find $E(Y)$ and $Var(Y)$

$$E(Y) = \int_0^\infty \underbrace{y}_{u} \underbrace{e^{-y} dy}_{dv} = \underbrace{(-ye^{-y})\Big|_0^\infty}_{=0} + \underbrace{\int_0^\infty e^{-y} dy}_{=1} = 1$$

$du = dy$

$v = -e^{-y}$

$$Var(Y) = E(Y^2) - (E(Y))^2 = \int_0^\infty y^2 e^{-y} dy - 1 = 1$$

and that's for Y.

$X = \frac{Y}{\lambda}$, so $E(X) = \frac{1}{\lambda}$, $Var(X) = \frac{1}{\lambda^2}$

MEMORYLESS PROPERTY

if you have a random variable that you can recognise as a waiting time, then this waiting time has the property that no matter how long you waited, that time is giving you no progress towards getting the phone call. It's like starting from fresh every time

$$P(X \geq s + t \mid X \geq s) = P(X \geq t)$$

t: additional time you could wait

s: time you already waited

s is irrelevant because you're starting from fresh every time

Here $P(X \geq s) = 1 - P(X \leq s) = e^{-\lambda s}$

ridondante

$$P(X \geq s+t \mid X \geq s) = \frac{P(X \geq s+t, X \geq s)}{P(X \geq s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} =$$

$$= P(X \geq t)$$

Also, given $X \sim Expo(\lambda)$, $E(X|X>a) = a + E(X-a|X>a)$
$= a + \frac{1}{\lambda}$

The exponential distr. is the only distr. with the memoryless property in the continuous time. The geometric is the only one for the discrete time.

THE MEMORYLESS PROPERTY IS A PROPERTY OF THE DISTRIBUTION, NOT OF THE R. VARIABLE ITSELF !

MOMENT GENERATING FUNCTIONS (MGF)

an MGF is another way of describing a distribution

- Definition: A r.v. X has MGF $M(t) = E(e^{tX})$, as a function of t, if this is finite on some $(-a, a)$, $a>0$

e^{tX} is a function of a r.v., which is a r.v. itself!

- why is it called moment "generating"?

$$E(e^{tX}) = E\left(\sum_{n=0}^{\infty} \frac{X^n t^n}{n!}\right) = \sum_{n=0}^{\infty} \frac{E(X^n)\, t^n}{n!}$$

$E(X^n)$ is called the n^{th} moment

1^{st} moment: mean

$1^{st} + 2^{nd}$: what we need to describe the distribution

- why is the MGF important?

1) the n^{th} moment, $E(X^n)$ is the coefficient of $t^n/n!$ in the Taylor series of M or the n^{th} derivative $M^{(n)}(0)$

2) the MGF determines the distribution (example if X and Y have the same MGF $\Rightarrow$ they have the same distribution

3) the MGF makes sums easier to handle (example if X has MGF M_X, Y has MGF M_Y, and X indep. of Y, then MGF of $X+Y$ is $E(e^{t(X+Y)})$ and $E(e^{t(X+Y)}) = E(e^{tX})E(e^{tY})$ (ONLY IF they're indep) $= M_X(t)M_Y(t)$

- $X \sim Bern(p) \Rightarrow$ MGF $M(t) = E(e^{tX}) = pe^t + q$, $q = 1-p$
- $X \sim Bin(n,p) \Rightarrow$ MGF $M(t) = (pe^t + q)^n$
- $Z \sim N(0,1) \Rightarrow$ MGF $M(t) = e^{t^2/2}$

LAPLACE'S RULE OF SUCCESSION

He phrased it in terms of what's the probability that the sun will rise tomorrow?

Let $X_1, X_2, \dots$ i.i.d $Bern(p)$ be the indicators r.v. for the sun rising on the 1st day, the second day and so on, all of this GIVEN p (meaning that given p, all of that is fine and $X_1, X_2 \dots$ are i.i.d.)

But we do not know p. How do we deal with that?

BAYESIAN POV: since p is unknown, we're going to quantify its uncertainty by treating it as a r.v. that has some distribution

then we can use Bayes' rule to adjust p's distribution given the evidence that we have

Let $p \sim Unif(0,1)$ let $S_n = X_1 + X_2 + \dots + X_n$
so $S_n | p \sim Bin(n, p)$, $p \sim Unif(0,1)$

Find posterior (after we collect data) distribution of $p | S_n$

What's the probability that the sun will rise tomorrow? $P(X_{n+1} = 1 | S_n = n)$

$$\underset{\text{PDF}}{f}(\underset{\text{continuous}}{p} | S_n = k) = \frac{P(S_n = k | p)\overbrace{f(p)}^{\text{prior}}}{P(S_n = k)} \propto$$

$$P(S_n = k) = \int_0^1 P(S_n = k | p) f(p)\, dp$$

does not depend on p

$$\underset{\text{proportional to}}{\propto} p^k (1-p)^{n-k}$$

$$f(p | \underbrace{S_n = n}_{\text{sun has risen for the last n days}}) = \underbrace{(n+1)}_{\text{to normalize the PDF}} p^n$$

$$P(X_{n+1} = 1 | S_n = n) = \int_0^1 (n+1)\, p\, p^n\, dp$$

$$= \frac{n+1}{n+2}$$

EXPONENTIAL & MGF

- $X \sim Expo(1)$, find MGF, moments

$M(t) = E(e^{tX})$ we take an expectation of some r.v. and we see it as a function of t

$= \int_0^\infty e^{tx} e^{-x} \, dx = \int_0^\infty e^{-x(1-t)} \, dx = \frac{1}{1-t}, \; t<1$

and $t<1$ is fine because we said we wanted to have some interval $(-a, a)$ in ~~which~~ which this is finite. In this case this is finite in $(-\infty, 1)$

so $M(t) = \frac{1}{1-t}$ is fine as an MGF

now find the moments:

$M'(0) = E(X)$, $M''(0) = E(X^2)$, $M'''(0) = E(X^3)$ ...

try to recognise the pattern ($\frac{1}{1-t}$ geometric series)

then $\frac{1}{1-t} = \sum_{n=0}^{\infty} t^n$ for $|t|<1$

BE CAREFUL! with the MGF you take the Taylor expansion and the moment is whatever is in front of t^n over $n!$

$\Rightarrow = \underbrace{\sum_{n=0}^{\infty} n! \frac{t^n}{n!}}_{\text{that's the pattern}} \longrightarrow$ whatever is in front of this is the moment!

$\Rightarrow E(X^n) = n!$ for all n

- $Y \sim Expo(\lambda)$, let $X = \lambda Y \sim Expo(1)$ so $Y^n = \frac{X^n}{\lambda^n}$

so $E(Y^n) = \frac{n!}{\lambda^n}$

NORMAL & MGF

let $Z \sim N(0,1)$, find all its moments

$E(Z^n) =$ 0 for n odd by simmetry

MGF: $M(t) = e^{t^2/2} = \sum_{n=0}^{\infty} \frac{(t^2/2)^n}{n!} = \sum_{n=0}^{\infty} \frac{t^{2n}}{2^n \cdot n!} \rightarrow$

$\rightarrow \sum_{n=0}^{\infty} ((2n)!) \frac{t^{2n}}{2^n \cdot n! (2n!)}$

$\Rightarrow E(Z^{2n}) = \frac{(2n)!}{2^n (n!)}$

$\hookrightarrow$ ~~~~ even numbers

example
$n=1 \Rightarrow E(Z^2) = 1$ (1)
$n=2 \Rightarrow E(Z^4) = 3$ $(1 \cdot 3)$
$n=3 \Rightarrow E(Z^6) = 15$ $(1 \cdot 3 \cdot 5)$
$n=4 \Rightarrow E(Z^8) = 105$ $(1 \cdot 3 \cdot 5 \cdot 7)$
...

$\Rightarrow E(Z^n) = 0$ for n: odd

$E(Z^{2n}) = \frac{(2n)!}{2^n (n!)}$ for 2n: even

POISSON & MGF

let $X \sim Pois(\lambda)$

MGF: $E(e^{tX}) = \sum_{k=0}^{\infty} e^{tk} \underbrace{e^{-\lambda} \frac{\lambda^k}{k!}}_{\text{poisson PMF}} = e^{-\lambda} \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k}{k!}$

$= e^{-\lambda} e^{\lambda e^t} = \underbrace{e^{\lambda(e^t - 1)}}_{\text{poisson MGF}}$

let $Y \sim Pois(\mu)$ X and Y are independent

Find the distribution of X+Y

THAT'S CALLED A CONVOLUTION

so all we do is multiply the MGF's:

MGF_X: $e^{\lambda (e^t - 1)}$

MGF_Y: $e^{\mu (e^t - 1)}$

$e^{\lambda(e^t-1)} e^{\mu(e^t-1)} = \boxed{e^{(\lambda+\mu)(e^t-1)}}$

$\Rightarrow X+Y \sim Pois(\lambda+\mu)$

$e^{(\lambda+\mu)(e^t-1)}$ is the $Pois(\lambda+\mu)$'s MGF

$E(X+Y) = \lambda + \mu$

if you have independent r.v.s joint distribution just means multiply the individual CDF's or the individual PDF's

JOINT DISTRIBUTION

- start with the simplest case: 2 r.v.s, both binary (so think about a 2 by 2 table)

X, Y Bernoulli (possibly dep/indep and same p/≠p)

	Y=0	Y=1
X=0		
X=1		

To specify the joint distribution all we have to do is put in 4 numbers that are non negative and add up to 1

- **X, Y r.v.s**

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

joint PMF (discrete case) : $P(X=x, Y=y)$

marginal CDF: this just means "take them separately"
$P(X \leq x)$ is the marginal CDF of X

marginal PMF: again, you take them separately
$P(X = x)$

Independence means that the joint distribution is the product of the two marginals

joint PDF (continuous case) : $f(x,y)$ such that
$$P((x,y)\in B) = \iint_B \underbrace{f(x,y)\,dx\,dy}$$

the PDF is what you integrate to get the probability

Independence: X, Y are independent if and only if the joint CDF is the product of the marginal CDF's
$F(x,y) = F_X(x)F_Y(y)$

Equivalent: the joint PMF is the product of the marginal PMF's
$P(X=x, Y=y) = P(X=x)\,P(Y=y)$
(DISCRETE CASE)

the joint PDF is the product of the marginal PDF's
$f(x,y) = f_X(x)\,f_Y(y)$ FOR ALL $x,y \in \mathbb{R}$
(CONTINUOUS CASE)

- Going back to the X,Y Binary example:
we can make up any four numbers we want as long as they're non-negative and add up to one
let's say:

	$Y=0$	$Y=1$
$X=0$	2/6	1/6
$X=1$	2/6	1/6

→ discrete example

Are X and Y independent?
Each number of the table is one of the joint probabilities.
$P(X=0, Y=0) = 2/6$, $P(X=0, Y=1) = 1/6$ ...
To prove that they're independent, we first need to find the marginal distributions.
How do we get those?

Getting Marginals (from joint distribution):
DISCRETE CASE: $P(X=x) = \sum_y P(X=x, Y=y)$

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

CONTINUOUS CASE: $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx$

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

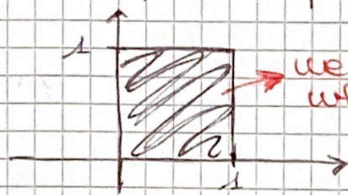
so in our example:
what's the probability of $Y=0$? You add the two cases $P(X=0, Y=0)$ and $P(X=1, Y=0)$

⇒

	$Y=0$	$Y=1$	X marginals:
$X=0$	2/6	1/6	→ $P\ X=0 = 3/6$
$X=1$	2/6	1/6	→ $P\ X=1 = 3/6$
Y marginals:	$P\ Y=0 = \frac{4}{6}$	$P\ Y=1 = \frac{2}{6}$	

and the marginals show X,Y are independent:
In fact $P(X=0, Y=0) = P(X=0)\,P(Y=0) = \frac{4}{6}\cdot\frac{3}{6} = \frac{2}{3}\cdot\frac{1}{2} = \frac{2}{6}$
which is, in fact, the value in the table's cell
so each of the joint probabilities is obtained by just multiplying two marginal probabilities

- another continuous example involving the uniform distribution.
  uniform on square $\{(x,y) \; x,y \in [0,1]\}$



we want a distribution which is normal over this square

so we want a joint PDF constant on the square and zero outside

we want to pick a completely random point (X,Y), so we want the density to be constant all over that square

the joint PDF is then $= \begin{cases} c & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$

the integral here is the area, so $c = \frac{1}{\text{area}}$ would normalise the PDF $\left(\text{since } c = \frac{1}{\text{area}} = 1\right)$

marginally, X,Y are independent uniform.

- for example X,Y on a disc $x^2+y^2 \le 1$

have a joint PDF $= \begin{cases} 1/\pi, & x^2+y^2 \le 1 \\ 0, & \text{otherwise} \end{cases}$

and X,Y are NOT independent because of the constraint $x^2+y^2 \le 1$

$\Rightarrow$ X,Y are DEPENDENT
in fact, given $X=x \Rightarrow -\sqrt{1-x^2} \le y \le \sqrt{1-x^2}$

this constraint shows us why they're dependent

with the uniform distribution, the probability is proportional to the area.

JOINT, CONDITIONAL, MARGINAL DISTRIBUTIONS

joint CDF: $F(x,y) = P(X \le x, Y \le y)$

continuous case

continuous case: joint PDF: $f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y)$

so that $P((x,y) \in A) = \iint_A f(x,y)\,dx\,dy$

marginal PDF of X: $\int_{-\infty}^{\infty} f(x,y)\,dy$ → this won't depend on y

of Y: $\int_{-\infty}^{\infty} f(x,y)\,dx$ → this won't depend on x

if you integrate once more you should get 1

$$\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} f(x,y)\,dy\right)dx = 1$$

conditional PDF of Y|X: $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} =$

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

independence: X, Y are independent if:

$$\underset{\text{PDF}}{f_X(x)}\ \underset{\text{PDF}}{f_Y(y)} = f_{X,Y}(x,y) \text{ for all } x,y$$

2D LOTUS

Let (X,Y) have joint PDF $f(x,y)$ and let $g(x,y)$ be a real-valued function of x,y

Then $\mathbf{E}\, g(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f(x,y)\,dx\,dy$

Theorem

If X, Y are independent, then $E(XY) = E(X)E(Y)$
("independence implies uncorrelated")

proof for the continuous case:

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x) f_Y(y)\,dx\,dy$$

$$= \int_{-\infty}^{\infty} y f_Y(y) \underbrace{\int_{-\infty}^{\infty} x f_X(x)\,dx}_{E(X)}\,dy$$

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

X, Y unif $(0,1)$

- example: let $M = \max(X,Y)$, $L = \min(X,Y) \Rightarrow |X-Y| = M-L$

  then $E(M-L) = \frac{1}{3}$, $E(M) - E(L) = \frac{1}{3}$

  $E(M+L) = E(M) + E(L) = \frac{1}{2} + \frac{1}{2} = 1$

  then $E(M) = \frac{2}{3}$, $E(L) = \frac{1}{3}$

CHICKEN-EGG PROBLEM (discrete case)

There are some eggs. some of them hatch, some of them don't (the eggs are independent). the number of eggs, N, is random.

$N \sim \text{Pois}(\lambda)$ eggs, each one hatches with probability p and each one is independent

let X = # of eggs that hatch, so $X|N \sim \text{Bin}(N,p)$
let Y = # of eggs that don't hatch
so $X+Y=N$

find the joint PMF of X, Y. Are they independent?

$$P(X=i, Y=j) = \sum_{n=0}^{\infty} P(X=i, Y=j \mid N=n)\, P(N=n)$$

what we wish we knew

$$= P(X=i, \cancel{Y=j} \mid N=i+j)\, P(N=i+j) = \frac{\cancel{(i+j)!}}{i!\,j!} p^i q^j \frac{e^{-\lambda}\lambda^{i+j}}{\cancel{(i+j)!}} =$$

Redundant!

$$= \left(e^{-\lambda p}\frac{(\lambda p)^i}{i!}\right)\left(e^{-\lambda q}\frac{(\lambda q)^j}{j!}\right) \Rightarrow$$

X, Y are independent
$X \sim \text{Pois}(\lambda p)$
$Y \sim \text{Pois}(\lambda q)$

Theorem If $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ are indep. Then:

$$X+Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$$

Proof (use MGFs):

MGF of $X+Y$ is $e^{\mu_1 t + \frac{1}{2}\sigma_1^2 t^2} + e^{\mu_2 t + \frac{1}{2}\sigma_2^2 t^2} =$

$$= e^{(\mu_1+\mu_2)t + \frac{1}{2}t^2(\sigma_1^2+\sigma_2^2)}$$

and that's exactly the MGF of $N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$ and the MGF determines the distribution.

- example: Find $E|Z_1 - Z_2|$ with Z_1, Z_2 iid $\sim N(0,1)$

  NOTE $Z_1 - Z_2 \sim N(0,2)$

  so $= E|\sqrt{2}\cdot Z|$ where $Z \sim N(0,1)$

  $$= \sqrt{2}\, E|Z| = \sqrt{2}\underbrace{\int_{-\infty}^{\infty} |z| \frac{1}{\sqrt{2\pi}} e^{-z^2/2}}_{\text{even fn}} dz$$

  $$= 2\sqrt{2}\int_0^{\infty} |z| \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = \sqrt{\frac{2}{\pi}}$$

MULTIVARIATE DISTRIBUTION: it's a joint distr. for more than one random variable

MULTINOMIAL DISTRIBUTION: $Mult(n, \vec{p})$, $\vec{p} = (p_1, \dots, p_k)$

$\vec{p}$ is a probability vector: $p_j \geq 0$ & $\sum_j p_j = 1$

$(X_1, \dots, X_k) = \vec{X} \sim Mult(n, \vec{p})$ if we have n objects (trials), which we are independently putting into k categories, with $p_j = P(\text{category } j)$ being the probability that any one of the objects is in category j and X_j = # objects in category j

PMF (joint): $P(X_1 = n_1, \dots, X_k = n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$

if $n_1 + \dots n_k = 1$ (the PMF is 0 otherwise)

Marginal: $\vec{X} \sim Mult_k(n, \vec{p})$. Marginal of X_j (# objects in category j): $X_j \sim Bin(n, p_j)$

$E(X_j)$ $= np_j$, **$Var(X_j)$** $= np_j(1-p_j)$

Lumping Property $\vec{X} = (X_1, X_2, \dots, X_{10}) \sim Mult(n, (p_1, \dots, p_{10}))$

people in the 1st political party — people in 2nd political party

since the first 2 are the dominant parties you want to put the other 8 together

$\vec{Y} = (X_1, X_2, \underline{X_3 + \dots + X_{10}})$ then $\vec{Y} \sim Mult(n, (p_1, p_2, \underline{p_3 + \dots + p_{10}}))$

CONDITIONAL DISTRIBUTION $\vec{X} \sim Mult(n, \vec{p})$ then given $X_1 = n_1$, $(X_2, \dots, X_k) \sim Mult_{k-1}(n - n_1, (p_2', \dots, p_k'))$

with $p_2' = P(\text{being in cat. } 2 \mid \text{not in cat. } 1) =$

$$= \frac{p_2}{1-p_1} = \frac{p_2}{p_2 + \dots + p_k}$$

$$p_j' = \frac{p_j}{p_2 + \dots + p_k}$$

to re-normalise them so that they add up to 1

COVARIANCE: Given X, Y r.v.s, then $Cov(X,Y)$ is
$$Cov(X,Y) = E\big((X-E(X))(Y-E(Y))\big)$$

Properties:
1. $Cov(X,X) = Var(X)$
2. $Cov(X,Y) = Cov(Y,X)$
3. analogue way to write it
   $Cov(X,Y) = E(XY) - E(X)E(Y)$
4. $Cov(X,c) = 0$ if c is a constant

bilinearity:
5. $Cov(cX,Y) = c\,Cov(X,Y)$
6. $Cov(X,Y+Z) = Cov(X,Y) + Cov(X,Z)$
7. $Cov(X+Y, Z+W) = Cov(X,Z) + Cov(X,W) + Cov(Y,Z) + Cov(Y,W)$

generally: $Cov\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i,j} a_i b_j Cov(X_i, Y_j)$

8. $Var(X_1+X_2) = Var(X_1) + Var(X_2) + 2Cov(X_1,X_2)$

if X_1, X_2 are independent $\Rightarrow Cov = 0$

generally: $Var(X_1+\dots+X_n) = Var(X_1)+\dots+Var(X_n) + 2\sum_{i<j} Cov(X_i, X_j)$

Theorem: If X, Y are independent, then they are uncorrelated (the covariance is zero).
Converse is false ($Cov = 0 \not\Rightarrow$ independence)

CORRELATION: $Corr(X,Y) = \dfrac{Cov(X,Y)}{SD(X)\,SD(Y)}$

with $SD = \sqrt{Var}$

equivalent $= Cov\left(\dfrac{X-E(X)}{SD(X)}, \dfrac{Y-E(Y)}{SD(Y)}\right)$

$\downarrow$ dimensionless quantity

Theorem: correlation is always between -1 and 1
$-1 \le Corr(X,Y) \le 1$
(form of Cauchy-Schwarz)

MULTINOMIAL & COVARIANCE

Given $(X_1, \dots, X_K) \sim Mult(n, \vec{p})$, find $cov(X_i, X_j)$ for all i, j

If $i = j$, $cov(X_i, X_j) = Var(X_i) = np_i(1-p_i)$

if $i \neq j$, $cov(X_1, X_2)$?

let's take $Var(X_1 + X_2) = np_1(1-p_1) + np_2(1-p_2) + 2Cov(X_1, X_2)$

let's call $cov(X_1, X_2) = c$

now $Var(X_1 + X_2) = np_1(1-p_1) + np_2(1-p_2) + 2c$

now for the lumping property:

$Var(X_1 + X_2) = n(p_1 + p_2)(1 - (p_1 + p_2))$

$\Rightarrow c = -np_1p_2 = Cov(X_1, X_2)$

Generally: $\underline{Cov(X_i, X_j) = -n p_i p_j \text{ for } i \neq j}$

covariance for the multinomial

- let I_A be the indicator r.v. of an event A
then: $I_A^2 = I_A$, $I_A^3 = I_A$, $I_A^4 = I_A \dots$

  and also: $I_A I_B = I_{A \cap B}$

- $X \sim Bin(n, p)$ with $X = X_1 + X_2 + \dots + X_n$
X_j are i.i.d $Bern(p)$

  $\underline{Var\ X_j} = E(X_j^2) - (E(X_j))^2 = p - p^2 = p\underbrace{(1-p)}_{q} = pq$

  $\Rightarrow \underline{Var\ X} = npq$ (you're adding up n of them

  (VARIANCE OF THE BINOMIAL)

- $X \sim HGeom(w, b, n)$ → w: white balls
  → b: black balls
  → n: sample of size n w/o replacement!!!

  $X = X_1 + \dots + X_n$, $X_j = \begin{cases} 1 & \text{if } j^{th} \text{ ball is white} \\ 0 & \text{otherwise} \end{cases}$

  ↳ these are DEPENDENT indicator r.v.'s

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

  $Cov(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2) = \frac{w}{w+b}\left(\frac{w-1}{w+b-1}\right) - \left(\frac{w}{w+b}\right)^2$

  (VARIANCE & COVARIANCE OF THE HYPERGEOMETRIC)

- Population is usually N
sample space is usually n

CHANGE OF VARIABLES - TRANSFORMATIONS

Remember that a function of a r.v. is a r.v. and we usually use LOTUS to get its expected value. BUT LOTUS doesn't give us the whole DISTRIBUTION so how do you get that?

Theorem: Let X be a continuous r.v. with PDF f_X and let $Y = g(X)$, where g is differentiable, strictly increasing

Then the PDF of Y's given by

$$f_Y(y) = f_X(x)\,\frac{dx}{dy} \quad \text{where } y = g(x)$$

and this is written in terms of y since we use $x = g^{-1}(y)$

Also note that $\frac{dx}{dy} = \left(\frac{dy}{dx}\right)^{-1}$

proof: let's find the CDF

CDF of Y's: $P(Y \le y) = P(g(X) \le y) =$

(the derivative of the PDF is the CDF)

$= P(X \le g^{-1}(y)) =$

same event

this is just the CDF of X evaluated here at $g^{-1}(y)$

$= F_X(g^{-1}(y)) = F_X(x)$

⇒ take the derivatives of both sides

⇒ $f_Y(y) = f_X(x)\,\frac{dx}{dy}$ (chain rule)

LOG NORMAL EXAMPLE

$Y = e^Z$, $Z \sim N(0,1)$ is standard normal

Log normal means that the log is normal (if you take the log of Y which is Z, Z is normal)

PDF: $f_Y(y) = \frac{1}{y}\frac{1}{\sqrt{2\pi}}\,e^{-\frac{(\ln(y))^2}{2}}$, $y > 0$

$\frac{dy}{dz} = e^z = y$ $\quad \left(\frac{dy}{dz}\right)^{-1}$

TRANSFORMATIONS in $\mathbb{R}^n$

$\vec{Y} = g(\vec{X})$, $g: \mathbb{R}^n \to \mathbb{R}^n$, $\vec{X} = (X_1, \dots, X_n)$ continuous

what's the joint PDF of Y ? analogous ~~xxxx~~ to the previous case:

joint PDF of $\vec{Y}$: $f_{\vec{Y}}(\vec{y}) = f_{\vec{X}}(\vec{x}) \left| \frac{d\vec{x}}{d\vec{y}} \right|$

→ Jacobian: it's the matrix of all possible partial derivatives

so $\frac{d\vec{x}}{d\vec{y}} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & & \ddots & \vdots \\ \vdots & & & \\ \frac{\partial x_n}{\partial y_1} & \cdots & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}$

so $\left| \frac{d\vec{x}}{d\vec{y}} \right|$ means: take the absolute value of the determinant

there also : $\left| \frac{d\vec{x}}{d\vec{y}} \right| = \left(\left| \frac{d\vec{y}}{d\vec{x}} \right| \right)^{-1}$

CONVOLUTION (SUMS)

we want the distribution of a sum or r.v.s
we already did the binomial, poisson & normal convolution

let $T = X+Y$, X, Y independent

- in the discrete case:

$$P(T=t) = \sum_x P(X=x)\, P(Y=t-x)$$

- in the continuous case:

$$f_T(t) = \int_{-\infty}^{\infty} f_X(x)\, f_Y(t-x)\, dx$$

since $F_T(t) = P(T \leq t) = \int_{-\infty}^{\infty} P(X+Y \leq t \mid X=x) f_X(x)\, dx$

now we plug in that $X=x$

$$\overset{\#}{=} \int_{-\infty}^{\infty} \underbrace{F_Y(t-x)}_{\text{CDF of Y evaluated at } t-x} f_X(x)\, dx$$

now take the derivative of both side and there's a theorem that says you can swap the derivative and the integral and ~~xxxxxxxxxx~~ then the formula for the continuous case is proven true

BETA DISTRIBUTION

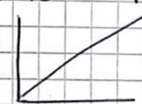
Generalisation of the uniform in the sense that it's still continuous and bounded but the PDF is not just flat

Beta(a,b), $a>0$, $b>0$

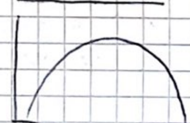
PDF: $f(x) = \underbrace{c}_{\text{constant}} \underbrace{x^{a-1}(1-x)^{b-1}}$, $0<x<1$

the question is how do we integrate this from 0 to 1 so that it's equal to 1? what is c equal to to make it $\int = 1$?

Properties:
- flexible family of continuous distribution on $(0,1)$
- if $a=b=1$ it becomes the normal

  if $a=2, b=1$ $\Rightarrow$ 

  if $a=b=\frac{1}{2}$ $\Rightarrow$ 

  if $a=b=2$ $\Rightarrow$ 

- often used as prior for a parameter in $(0,1)$
- "conjugate prior to Binomial"
- connections with other distributions

CONJUGATE PRIOR FOR BINOMIAL

We observe $X|p \sim \text{Bin}(n,p)$ and let $p \sim \text{Beta}(a,b)$ (prior) and that's because we don't know p so we treat it as a r.v.; then, after we observe X, we update our prior uncertainty p.

We want the posterior distribution $p|X$

$f(p|X=k)$ posterior PDF $= \dfrac{P(X=k|p)\,f(p)}{P(X=k)} =$ → does not depend on p

$$= \frac{\binom{n}{k} p^k (1-p)^{n-k}\, c\, p^{a-1}(1-p)^{b-1}}{P(X=k)} \propto$$

$$\propto p^{a+k-1}(1-p)^{b+n-k-1} \Rightarrow p|X \sim \text{Beta}(\underbrace{a+X}_{\text{successes}}, \underbrace{b+n-X}_{\text{failures}})$$

conjugate means that it's a family of distr.

PUZZLE

What is the next term supposed to be in this sequence?
$0, 1, 2, 720!$

One solution could be: $0, 1!, (2!)!, 3!!!, 4!!!!$ ...

STERLING'S FORMULA

$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ → That's what causes the rapid growth because it almost increases as n^n

GAMMA FUNCTION

$\Gamma(a) = \int_0^\infty x^a e^{-x} \frac{dx}{x}$, for real $a>0$

and $\Gamma(n) = (n-1)!$, for n positive integer

also $\underbrace{\Gamma(x+1) = x\Gamma(x)}_{\text{Recursive formula!}}$

and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, $\Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{\pi}}{2}$ ...

GAMMA DISTRIBUTION

↳ GAMMA-EXPONENTIAL CONNECTION

T_1 → Time of the 1st email (expo(λ))

N_t = # emails up to time t $\sim Pois(\lambda t)$

and # of arrivals in disjoint intervals are indep

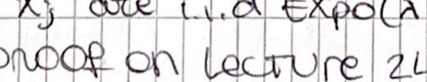
note $P(T_1 > t) = \underbrace{P(N_t = 0)}_{\text{Poisson}} = e^{-\lambda t}$

so the time you wait until the 2nd email (T_2) is also exponential and it's independent of the 1st for the memoryless property

any waiting time is all exponential (λ)

⇒ inter arrival times (times between the × of the emails arriving) are i.i.d Expo(λ)

But what's the actual time?



We know those interarrival times are Expo(λ)
but what is the actual time?

$\Rightarrow T_n$ = (Time of n^{th} arrival) $= \sum_{j=1}^{n} X_j \sim$ Gamma (n, λ)

X_j are i.i.d Expo(λ)

(proof on lecture 24 min 33)

MOMENTS (using LOTUS)

Let $X \sim \text{Gamma}(a,1)$, find $E(X^c)$, c = constant

$$\frac{1}{\Gamma(a)} \int_0^\infty x^c \underbrace{x^a e^{-x} \frac{dx}{x}}_{} = \frac{1}{\Gamma(a)} \int_0^\infty \underbrace{x^{a+c} e^{-x} \frac{dx}{x}}_{}$$

Gamma PDF

Gamma (a+c) / Gamma(a)

$$= \frac{\Gamma(a+c)}{\Gamma(a)}, \text{ if } a+c>0$$

bc Gamma function is only defined on positive numbers

let $c=1$ $\quad E(X) = \frac{\Gamma(a+1)}{\Gamma(a)} = \frac{a\Gamma(a)}{\Gamma(a)} = a$ (1st moment) mean

$$E(X^2) = \frac{\Gamma(a+2)}{\Gamma(a)} = \frac{(a+1)(a)\Gamma(a)}{\Gamma(a)} = a^2+a$$

variance is then $Var(X) = a$

Then for **Gamma (a, λ)**:

mean: $\frac{a}{\lambda}$

variance: $\frac{a}{\lambda^2}$

~~BANK AND POST OFF~~

BANK AND POST OFFICE

they are both sums of i.i.d exponential

wait at bank $X \sim \text{Gamma}(a, \lambda)$ } indep

wait at post office $Y \sim \text{Gamma}(b, \lambda)$

Find the distribution of $T = X+Y$ and $W = \frac{X}{X+Y}$ and their joint distrib.

a+b i.i.d exponentials

First of all T and W are independent

Let $\lambda = 1$ to simplify notation

joint PDF $f_{T,W}(t,w) = f_{X,Y}(x,y) \underbrace{\left| \frac{\partial(x,y)}{\partial(t,w)} \right|}_{\text{Jacobian}} =$

$$= \frac{1}{\Gamma(a)\Gamma(b)} x^a e^{-x} y^b e^{-y} \frac{1}{xy} \underbrace{\begin{vmatrix} w & t \\ 1-w & -t \end{vmatrix}}_{=-t}$$

$$= \frac{1}{\Gamma(a)\Gamma(b)} w^{a-1} (1-w)^{b-1} \underbrace{t^{a+b} e^{-t} \frac{1}{t}}_{\text{Gamma}(a+b, \lambda)}$$

multiply by ~~Γ(a)Γ(b)~~ $\Gamma(a+b)$

$$\Rightarrow \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} w^{a-1}(1-w)^{b-1} \boxed{\frac{t^{a+b} e^{-t} \, 1/t}{\Gamma(a+b)}} \text{ Gamma } (a+b, 1)$$

→ marginal PDF of w $f_W(w) = \int_{-\infty}^{\infty} f_{T,W}(t,w)\,dt =$

$$= \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, w^{a-1}(1-w)^{b-1}}_{\text{Beta}(a,b)}$$

Then this has to be the normalising constant of the Beta

so $T \sim \text{Gamma}(a+b, 1)$
$W \sim \text{Beta}(a, b)$
and they're independent

ORDER STATISTICS

Let $X_1, \dots, X_n$ be i.i.d., the order statistics are
$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, where $X_{(1)} = \min(X_1, \dots, X_n)$
$X_{(n)} = \max(X_1, \dots, X_n)$

Also if n is odd, the median is $X_{\left(\frac{n+1}{2}\right)}$
if n is even, the median is the simple average of the two middle numbers

Definitions:

Percentiles: A percentile is a measure at which that percentage of the total values are the same as or below that measure
e.g. 90% of data values are below the 90th percentile, while 10% of data values lie above the 90th percentile

Quartiles: are values that divide a data group into 4 groups containing an approximately equal number of observations. the total of 100% is split into 4 equal parts.

1st quartile = 25th percentile (0.25)
2nd quartile = 50th percentile (0.50)
3rd quartile = 75th percentile (0.75)
4th quartile = 100th percentile (1.00)

EXAMPLE OF PATTERNS IN COIN FLIPS.

Repeated four coin flips waiting for a pattern. How many flips until HT? And until HH?

Let W_{HT} = # of flips until you get HT for the 1st time
W_{HH} = # of flips until you get HH for the 1st time

Find the average: $E(W_{HT})$ and $E(W_{HH})$

First of all, note that: $E(W_{HH}) = E(W_{TT})$
$E(W_{HT}) = E(W_{TH})$

But that doesn't mean $E(W_{HT}) = E(W_{HH})$!

Now think about $E(W_{HT})$. Imagine this sequence

TTTT H, then → H: fine, we still have that partial progress
→ T: done!

So you can split the waiting time in two:

$\underbrace{TTTT\,H}_{W_1}\underbrace{HHT}_{W_2}$ W_1 = time til the 1st H
W_2 = time til the 1st T after the 1st H

W_1 and W_2 are independent bc coin is memoryless and also $W_j - 1 \sim \text{Geom}(\frac{1}{2})$

since the Geom does NOT include the success

and Geom $(\frac{1}{2})$ has expected value = 1, so if you add back the success you get

$E(W_1) = 1+1$, $E(W_2) = 1+1 \Rightarrow E(W_{HT}) = 2+2 = 4$

Now think about $E(W_{HH})$. Imagine this sequence

TTTT H, then → H: done
→ T: we lose our previous progress

So $E(W_{HH}) = E(W_{HH} \mid 1^{st}\text{ toss is H})\,\frac{1}{2} + E(W_{HH} \mid 1^{st}\text{ toss T})\,\frac{1}{2}$

probability 1st toss is H

$= \left(2\cdot\frac{1}{2} + \left(2 + E(W_{HH})\right)\frac{1}{2}\right)\frac{1}{2} + \left(1 + E(W_{HH})\right)\frac{1}{2} = 6$

if (given 1st toss is H) also the 2nd toss is H you're done

prob of this happening

you toss and you get T

then it's the same problem again

if (given 1st toss is H) you get T on the 2nd toss you wasted 2 tosses and it's the same problem again

CONDITIONAL EXPECTATION

Y DISCRETE: $E(Y|X=x) = \sum_y y\, P(Y=y|X=x)$

Y CONTINUOUS: $E(Y|X=x) = \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\, dy$ → conditional PDF

X,Y CONTINUOUS: $E(Y|X=x) = \int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y)}{f_X(x)}\, dy$ → joint PDF; → marginal PDF of X

Let $g(x) = E(Y|X=x)$, then define $E(Y|X) = g(X)$

e.g. if $g(x) = x^2$, then $g(X) = X^2$

so $E(Y|X)$ is a r.v., function of X

EXAMPLE

- X,Y i.i.d. Pois(λ). Find $E(X+Y|X) =$

$= E(X|X) + E(Y|X) = X + E(Y) = X + \lambda$

linearity

I know X & I need a prediction of X ⇒ $E(X|X) = X$

since X,Y independent knowing X tells me nothing about Y so I drop it

note that $E(h(X)|X) = h(X)$

- X,Y i.i.d. Pois(λ). Find $E(X|X+Y)$

YOU CAN'T use linearity here

Straight from the definition:

Let $T = X+Y$, find conditional PMF

$P(X=k|T=n) = \frac{P(T=n|X=k)\,P(X=k)}{P(T=n)}$

Baye's Rule

$= \frac{P(Y=n-k)\,P(X=k)}{P(T=n)}$

$= \frac{\frac{e^{-\lambda}\lambda^{n-k}}{(n-k)!} \cdot \frac{e^{-\lambda}\lambda^{k}}{k!}}{\frac{e^{-2\lambda}\, 2\lambda^{n}}{n!}}$

bc Y ~ Pois

bc X ~ Pois

sum of indep Pois is Pois so this is going to be Pois(2λ)

$= \binom{n}{k}\frac{1}{2^n}$ binomial

$\Rightarrow X|T=n \sim \text{Bin}(n, \frac{1}{2})$

$\Rightarrow E(X|T=n) = \frac{n}{2} \Rightarrow E(X|T) = \frac{T}{2}$

By definition!

Also note that $E(X|X+Y) = E(Y|X+Y)$ is true by simmetry since they're independent

And $E(X|X+Y) + E(Y|X+Y) = E(X+Y|\cancel{}\, X+Y) = \underbrace{X+Y}_{T}$

$\Rightarrow E(X|T) = T/2$

ADAM'S LAW - ITERATED E

$$E(E(Y|X)) = E(Y)$$

EXAMPLE

- Let $X \sim N(0,1)$, $Y = X^2$

  Then $E(Y|X) = E(X^2|X) = X^2 = Y$

  Then $E(X|Y) = E(X|X^2) = 0$ since if $X^2 = a \Rightarrow X = \pm\sqrt{a}$ and by simmetry $+\sqrt{a}, -\sqrt{a}$ are equally likely so if you average them you get 0.

- We have a stick 0 —— 1 and we break off a random piece. Then we throw one away and break the other one again. We want the E for the lenght of the 2nd piece.

  Let X be the 1st breaking point, $X \sim Unif(0,1)$

  Y be the 2nd breaking point, $\underline{Y|X} \sim Unif(0, X)$

  0 Y X 1

  this means: if we know $X = x$ then Y it's going to be $Unif(0,x)$

  $\underline{E(Y|X=x)} = x/2$, $E(Y|X) = X/2$

  this is going to be a fn of x

  you can think about it as a short hand for $E(Y|X=x) = x/2$

  then $E(E(Y|X)) = E(Y) = \frac{1}{4}$ for adam's law

PROPERTIES OF CONDITIONAL EXPECTATION

1. $E(h(X)Y|X) = h(X)E(Y|X)$ → bc since we know X we treat it as a constant
2. $E(Y|X) = E(Y)$ → if X, Y are independent
   non vale il v.v.
3. $E(E(Y|X)) = E(Y)$ → Generalisation of law of total probability
4. $E((Y - \underbrace{E(Y|X)}_{\text{prediction}})\, h(X)) = 0$ → and that's because the residual $Y - E(Y|X)$ is uncorrelated with $h(X)$
5. $Var(Y) = E(Var(Y|X)) + \underbrace{Var(E(Y|X))}_{\text{How far off is the prediction}}$ (Eve's Law)

CONDITIONAL VARIANCE (Definition)

$$Var(Y|X) = E(Y^2|X) - (E(Y|X))^2 = E((Y - E(Y|X))^2 \mid X)$$

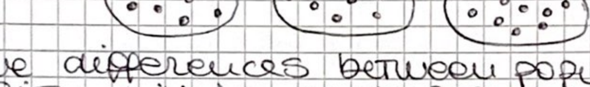
(keep in mind the property:
$Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$)

EVE'S LAW

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$$

EXAMPLE

You have some population of people which consists of 3 subpopulations. Y being the height. You want to know the mean and the variance of Y



You have differences between populations and variability within each population.

Let X be 1, 2, 3 to indicate, if you take a random person, which subpop. they're in.

So $E(Y|X=1)$ = mean for 1st population

So Eve's Law says $= \underbrace{E(Var(Y|X))}_{\text{LOOK WITHIN EACH POP. THEN AVERAGE}} + \underbrace{Var(E(Y|X))}_{\text{LOOK BETWEEN POP.S, THEN TAKE THE VAR OF THOSE}}$

EXAMPLE

Pick random city in a state. Pick random sample of n people in that city. Test for a disease.

Let X = (# people tested positive for the disease)

Q = (proportion of people in the random city with disease)

because ≠ cities have ≠ proportions ⇒ Q is a r.v.

Find $E(X)$, $Var(X)$, assuming $Q \sim Beta(a,b)$, a,b known

Also $X|Q \sim Bin(n, Q)$ (implicit) since Q has to be between 0 and 1

so $E(X) = E(E(X|Q)) = E(nQ) = \dfrac{na}{a+b}$

n is just a constant and $Q \sim Beta(a,b)$ has $E = a/a+b$

and $Var(X) = E(Var(X|Q)) + Var(E(X|Q))$

$= E(\underbrace{nQ(1-Q)}_{\text{bc Binomial}}) + Var(nQ) =$

$= E(nQ(1-Q)) + n^2 \, Var(Q) = nE(Q(1-Q)) + n^2 var(Q)$

we want this to become the exact integral of a beta PDF

now by LOTUS

$$E(Q(1-Q)) = \left(\underline{\int_0^1 q^a (1-q)^b \, dq}\right) \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

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

we put in the normalising constant for the beta (multiplying and dividing)

we divided & multiplied so that's what left

the integral simplifyed to 1

$$= \frac{ab\,\cancel{\Gamma(a+b)}}{(a+b+1)(a+b)\cancel{\Gamma(a+b)}} = \frac{ab}{(a+b+1)(a+b)}$$

$$Var(Q) = \frac{\mu(1-\mu)}{a+b+1}, \quad \mu = \frac{a}{a+b}$$

EXAMPLE

Store with a random # of clients N. Let X_j be the amount of \$ the j^{th} customer spends. X_j has mean μ, var σ^2.
Assume $N, X_1, \dots, X_N$ are indep.
Find $E(X)$, $Var(X)$ with X = total \$ spent

$$E(X) = \sum_{n=0}^{\infty} E(X|N=n)P(N=n) = \sum_{n=0}^{\infty} \mu n P(N=n) = \mu E(N)$$

or with Adam's law:

$$E(X) = E(E(X|N)) = E(\mu N) = \mu E(N)$$

since $E(X|N=n) = \mu n \Rightarrow E(X|N) = \mu N$

now with Eve's law:

$$Var(X) = E(\underbrace{Var(X|N)}_{N \text{ times the variance of 1 term}}) + Var(E(X|N))$$

$$= E(N\sigma^2) + Var(\mu N)$$

$$= \sigma^2 \underline{E(N)} + \mu^2 \underline{Var(N)}$$

so $Var(X)$ is in term of E and Var of X. We just need to know the distribution of N and we're done :)

STATISTICAL INEQUALITIES

① CAUCHY-SCHWARZ INEQUALITY

$$|E(XY)| \le \sqrt{E(X^2)E(Y^2)}$$

if X,Y are uncorrelated $\Rightarrow E(XY) = E(X)E(Y)$
so this inequality is particularly useful if they are correlated

note that $E(X^2)$ is the marginal 2nd moment of X and $E(Y^2)$ is the marginal 2nd moment of Y

if X,Y have mean 0 then $|Corr(X,Y)| = \left| \frac{E(XY)}{(EX^2 EY^2)^{1/2}} \right| \le 1$

so the interpretation is that correlation is between -1 and +1

② JENSEN'S INEQUALITY

If g is a convex function, then $E(g(X)) \ge g(E(X))$
with convex meaning: if $g''(x)$ exists then $g''(x) \ge 0$
also if h is concave ($h''(x) \le 0$) then $E(h(X)) \le h(E(X))$

convex  concave 

also convex meaning: if you take two points on the curve and connect them with a line that line is all above the curve

③ MARKOV'S INEQUALITY

$$P(|X| \ge a) \le \frac{E(|X|)}{a} \quad \text{for any } a > 0$$

note that $a I_{|X| \ge a} \le |X|$ is true
so $aE(I_{|X|\ge a}) \le E|X|$ and by the fundamental bridge, that's the same of Markov's inequality
($I_{|X|\ge a}$ is the indicator of the event $|X| \ge a$
so it can only be 0 if false, 1 if true)

what this inequality says is:
imagine you have 100 people
is it possible that at least 95% are younger than the mean of the group? yes
is it possible that at least 50% are older than twice the average age? No
is it possible that at least 33% are older than triple the average age? No
And so on

④ **CHEBYSHEV'S INEQUALITY**

$P(|X-\mu|>a) \leq \frac{Var\, X}{a^2}$ for $\mu = E(X)$, $a>0$

that can also be written as:

$P(|X-\mu| > c\, SD(X)) \leq \frac{1}{c^2}$ for $c>0$ ($a = c\,SD(X)$)

which means that (if $c=2$) the probability that X is more than 2 SD away from its mean is at MOST one quarter

---

let $X_1, X_2, \ldots$ be i.i.d., mean μ, var σ^2

let the sample mean $\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$ be the average of the first n.

what can we say about $\bar{X}_n$ as n gets large?

Law of Large Numbers (STRONG)

$\bar{X}_n \to \mu$ as $n \to \infty$ with probability 1

which means that the sample mean converges to the true mean

note that $\bar{X}_n \to \mu$ is a limit of a r.v. which matematically is a function. The whole $\bar{X}_n$ converging is an event (which, as said, has probability 1)

example: $X_j \sim Bern(p)$, then $\frac{X_1 + \ldots + X_n}{n} \to p$ with prob. 1

Law of Large Numbers (WEAK)

For any $c>0$, $P(|\bar{X}_n - \mu| > c) \to 0$ as $n \to \infty$

which means: if n is large, it's extremely likely for $\bar{X}_n$ to be close to μ (the difference is extremely little)

proof (using chebyshev):

$$P(|\bar{X}_n - \mu| > c) \leq \frac{Var(\bar{X}_n)}{c^2} = \frac{\frac{1}{n^2}(n\sigma^2)}{c^2} = \frac{\sigma^2}{\underset{\to\infty}{n}c^2} \to 0$$

so another way to write the concept is

$\bar{X}_n - \mu \to 0$ as $n \to \infty$ with probability 1

still, this tells us nothing about $\bar{X}_n$'s distribution

CENTRAL LIMIT THEOREM

$n^{1/2}\left(\frac{\bar{X}_n-\mu}{\sigma}\right) \to N(0,1)$ in distribution

which means: the distribution of $n^{\frac{1}{2}}\left(\frac{\bar{X}_n-\mu}{\sigma}\right)$ converges to the normal $N(0,1)$

subtract the mean

Equivalently: $\frac{\sum_{j=1}^{n} X_j - n\mu}{\sqrt{n}\sigma} \to N(0,1)$ in distribution

we standardise the sum

standard deviation for $\sum X_j$

BINOMIAL APPROXIMATED BY NORMAL

let $X \sim Bin(n,p)$. think of $X = \sum_{j=1}^{n} X_j$, $X_j \sim Bern(p)$ i.i.d.

so we are adding up i.i.d. r.v.s

$P(a \le X \le b) = P\left(\frac{a-np}{\sqrt{npq}} \le \frac{X-np}{\sqrt{npq}} \le \frac{b-np}{\sqrt{npq}}\right)$

standardise X

so if N is large this is going to have approximately a normal distr.

$\approx \Phi\left(\frac{b-np}{\sqrt{npq}}\right) - \Phi\left(\frac{a-np}{\sqrt{npq}}\right)$

CDF

contrast with Pois approximation of Bin

Poiss approx. : n large, p small, $\lambda = np$ (unlikely events)

Normal approx. : n large, p close to $\frac{1}{2}$ (events with p close to $\frac{1}{2}$)

(still makes sense because Pois with a large λ also looks Normal)

Also note the continuity correction (since you're using something continuous to approximate something discrete):

$P(X=a) = P\left(a-\frac{1}{2} < X < a+\frac{1}{2}\right)$

CHI-SQUARED DISTRIBUTION

$\chi^2(n)$ with n being the degrees of freedom
How is it related to the normal?
Let $V = Z_1^2 + Z_2^2 + \dots + Z_n^2$ where Z_j i.i.d. $N(0,1)$
then $V \sim \chi^2(n)$ (so V is the sum of $\underbrace{\text{squared}}_{n}$ i.i.d normals)
note that: $\chi^2(1)$ is Gamma$\left(\frac{1}{2}, \frac{1}{2}\right)$
↳ so it's just 1 squared normal
so $\chi^2(n)$ means we're adding up n Gammas$\left(\frac{1}{2}, \frac{1}{2}\right)$
which means $\chi^2(n)$ is Gamma$\left(\frac{n}{2}, \frac{1}{2}\right)$

STUDENT-T DISTRIBUTION (GOSSET)

Let $T = \frac{Z}{\sqrt{V/n}}$ where $Z \sim N(0,1)$, $V \sim \chi^2(n)$, indep.
Then $T \sim t_n$ (n: degrees of freedom)
↳ also: how many normals squared did we add up? n :)

Properties:

1. symmetric (i.e. $-T \sim t_n$ (-T has same distr.))
2. if $n=1$ ⇒ it's really just a ratio of two indep standard normals: cauchy distr and the mean doesn't exist.
3. if $n \geq 2$ ⇒ $E(T) = 0$ by symmetry but also because $E(T) = E(Z)\, E\left(\frac{1}{\sqrt{V/n}}\right) = 0$
4. Doesn't have all its moments; The odds moments are $=0$;
5. Looks like the normal distr. but heavier tailed (which means more extreme values are relatively more likely than that they would be for the normal)
6. for n large, t_n looks very much like standard normal
distribution of t_n goes to $N(0,1)$ as $n \to \infty$
(also proven by the Law of Large Numbers)

MULTIVARIATE NORMAL (MVN) DISTRIBUTION

DEFINITION: Random vector (which means that we have a vector made up of r.v s) $\vec{X} = (X_1, X_2, \ldots, X_k)$ is multivariate Normal if every linear combination (e.g. $t_1X_1 + t_2X_2 + \ldots + t_kX_k$) is Normal

EXAMPLE

let Z, W be i.i.d. $N(0,1)$ Then $(Z+2W, 3Z+5W)$ is MVN (as well as (Z,W) since they're indep.) and that is true bc $s(Z+2W) + t(3Z+5W) =$

$$= \underbrace{(s+3t)Z}_{\text{independent normal}} + \underbrace{(2s+5t)W}_{\text{independent normal}} \text{ is Normal}$$

EXAMPLE

let $Z \sim N(0,1)$, S be a random sign (so +1 or -1 with equal probability) indep. of Z.
Then Z, SZ are marginally still standard normal, but (Z, SZ) is NOT MVN
(look at $Z+SZ$: is not normal because there's no normal that equal 0 half of the time)

MGF: MGF of $\vec{X}$, $\vec{X}$ being MVN is

$$E(e^{\vec{t}'\vec{X}}) = E(e^{t_1X_1 + \ldots + t_kX_k})$$

$\vec{t}'\vec{X}$ is still normal by definition

$$= e^{t_1\mu_1 + \ldots + t_k\mu_k + \frac{1}{2}\mathrm{var}(t_1X_1 + \ldots + t_kX_k)}$$

with $\mu_j = EX_j$ (mean of the j^{th} component)

= (just to be clear)

$$\exp\left(t_1\mu_1 + \ldots + t_k\mu_k + \frac{1}{2}\mathrm{var}(t_1X_1 + \ldots + t_kX_k)\right)$$

Theorem

Within a MVN, it is true that uncorrelated implies independent.

so if $\vec{X} = \begin{pmatrix} \vec{X}_1 \\ \vec{X}_2 \end{pmatrix}$ is MVN and if every component of $\vec{X}_1$ is uncorrelated with every component of $\vec{X}_2$ (which means if their covariance is 0) then $\vec{X}_1$ and $\vec{X}_2$ are independent.

EXAMPLE

Let X, Y be i.i.d. $N(0,1)$. Then $(X+Y, X-Y)$ is MVN (it's called bivariate normal) and they're uncorrelated since

$$\mathrm{Cov}(X+Y, X-Y) = \mathrm{Var}(X) + \cancel{\mathrm{Cov}(X,Y) - \mathrm{Cov}(X,Y)} - \mathrm{Var}(Y)$$
$$= 0 \quad (\text{since } \mathrm{Var}(X) = \mathrm{Var}(Y))$$

so $(X+Y)$ and $(X-Y)$ are uncorrelated and because they're MVN we know that $(X+Y)$ and $(X-Y)$ are independent.

COSE MANCANTI

- **Leggi di De Morgan**: ① $(A\cap B)^c = A^c \cup B^c$
  ② $(A\cup B)^c = A^c \cap B^c$
- **Identità di Poincaré**: Principio di inclusione - esclusione

$$P(A_1\cup\ldots\cup A_n) = \sum_{j=1}^{n} P(A_j) - \sum_{i<j} P(A_i\cap A_j) + \sum_{i<j<k} P(A_i\cap A_j\cap A_k) - \ldots + (-1)^{n+1} P(A_1\cap\ldots\cap A_n)$$

- **Disuguaglianza di Hoeffding**:

fornisce un limite superiore alla probabilità che la somma delle r.v.s limitate indipendenti si discosti dal suo valore atteso

Siano $X_1,\ldots,X_n$ r.v.s tc: $P(a_i\le X_i\le b_i)=1$, $-\infty<a_i, b_i<+\infty$ per $i=1,\ldots,n$ e sia $S=X_1+\ldots+X_n$. Allora

$P(S-E(S)\ge t)\le e^{-\frac{2t^2}{\sum(b_i-a_i)^2}}$ per $t>0$, equivalente:

$$P(S-E(S)\ge t)\le \exp\left(-\frac{2t^2}{\sum(b_i-a_i)^2}\right) \text{ per } t>0$$

- **Teorema di De Moivre - Laplace**:

è un caso particolare del teorema del limite centrale e afferma che la distribuzione normale può essere utilizzata come approssimazione della distribuzione binomiale (sotto determinate condizioni)

In particolare il teorema mostra che la PDF del numero di successi osservati in una serie di n prove di Bernoulli indipendenti (ognuna con probabilità di successo p in quanto il singolo tentativo è binomiale) converge alla PDF della normale con media: np e standard variation: $\sqrt{np(1-p)}$ al crescere di n.