Lecture Notes 9/28.

Erik Cheng
Date 8.to 17.

1. Announcements (write on board)
   - Proj due TODAY:
     - Pls no work on it during lab - esp. since lab concepts will be impt. for MT.
   - Midterm is 10/6 (next Friday)
     - Fill out feedback form or email me w/ questions or topics you'd like reviewed.

→ Topics for today: Probability, Sampling (Terminology), Quick programming recap (bools, ctrl statements, join, group, pivot)

2. Basic Probability:
   - Fundamental idea: $P(A) = \frac{\#\text{ of events where } A \text{ is fulfilled}}{\#\text{ of events possible.}}$
   - Range is 0 to 1: why?

$$\frac{\text{nothing satisfies A}}{\text{events}} \leq P(A) \leq \frac{\text{everything satisfies A}}{\text{events.}}$$

0 to 1 ⟷ 0% to 100%.

A) This directly leads to addition rule.

$$P(A) = P(A_1) + P(A_2)$$

If A can happen in exactly one of 2 ways.
ie, A_1 & A_2 are mutually exclusive events.

- Example:

Deck has 4 suits: Heart, Club, Spade, Diamond
H C S D

Draw one card, Probability that card is ~~red~~? Written in terms of colors,
↳ card is heart or diamond

P(card is H or D) = P(Card is red) = P(Card is heart) + P(Card is Diamond).

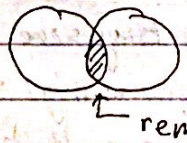
Counter example: (don't need to cover this unless asked)

P(of a dice roll, the result is odd - or less than 4)
≠ P(roll is odd) + P(roll is <4)

Because odd & <4 are compatible events!

optional bonus.

* If you want to calculate P(roll is odd or <4),
= P(roll is odd) + P(roll <4) − P(roll is odd & <4)



removing this overlap.

$$= \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{4}{6} = \frac{2}{3}$$

or 1 2 3 4 5 6 remove double counts = 3↑ + 3↑ − 2↑ = 4 total

* Make sure to note $P(A) \geq P(A_1) + P(A_2)$

b/c the more ways A can happen, the more likely it is to happen.

B) Multiplication Rule

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

- In words: Probability of two events is equal to probability of one of them, multiplied by the probability of the second given the first one has happened.

- Why does it work?

We want both A & B. This means for events where A&B are satisfied, we know A&B are satisfied (ofc).

By thinking about one being 'fixed', we subset to events where that one is true. Then, we further subset events that also satisfy the second, thus giving us the subset that satisfy both.

- Example: tickets from a jar.

1 Green, 2 Red, 3 Yellow, 4 Blue.

$P(\text{ticket 1} = Y, \text{ticket 2} = G)$?

$= P(\text{ticket 1} = Y) \cdot P(\text{ticket 2} = G \mid \text{ticket 1 was } Y)$

$= \frac{3}{10} \cdot \frac{1}{9} = \frac{3}{90} = \frac{1}{30}$

optional: note $= P(T_2 = G) \cdot P(T_1 = Y \mid T_2 = G)$

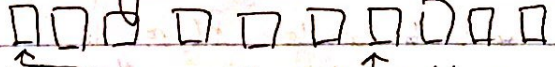
$= \frac{1}{10} \cdot \frac{3}{9} = \frac{3}{90} = \frac{1}{30}$

- Why is $P(T_2 = G) = \frac{1}{10}$ and $P(T_1 = Y) = \frac{3}{10}$?

Shouldn't i in T_i matter?

- With no extra information, asking $P(T_i = C)$ where C is one of the colors listed is the same given any i & fixed C.

Imagine shuffling tix:



asking whether → is Y or ← is Y gives the same result.

It is only after getting some information that we'd change our statement of probability: still 3 possible ways to get Yellow & 10 places it could be, in $\Rightarrow P(T_7 = Y) = \frac{3}{10}$, as opposed to

Still 3 possible ways for Y to be in Position 7, but only 9 possible card choices left: $\frac{3}{9}$

With replacement:

$P(T_1 = Y, T_2 = G)$

$= P(T_1 = Y)\, P(T_2 = G \mid T_1 = Y)$ By Multiplication Rule.

$= P(T_1 = Y)\, P(T_2 = G)$ $P(T_2 = G)$ is not affected at all by T_1's result.

$= \frac{3}{10} \cdot \frac{1}{10} = \frac{3}{100}$

What about asking for NOT A? $1 - P(A)$.

- Motivation: We want # of outcomes that don't fulfill A.

$= \#\text{total outcomes} - \#\text{of outcomes that fulfill A}$

$\#$ of outcomes that don't fulfill A.

$$P(\text{not } A) = \frac{\#\text{outcomes that don't fulfill A}}{\#\text{total outcomes}}$$

$$= \frac{\#\text{total outcomes} - \#\text{outcomes that DO fulfill A}}{\#\text{total outcomes}}$$

$$= 1 - \frac{\#\text{outcomes that DO fulfill A}}{\#\text{total outcomes}}$$

$$= 1 - P(A)$$

Example: $P(\text{Card drawn is not a heart})$

$= 1 - P(\text{card is a heart})$

$= 1 - \frac{1}{4} = \frac{3}{4}$

Notice this is the same result as

$P(\text{Card drawn is club, spade, or diamond})$

$= P(\text{Card is club}) + P(\text{Card is spade}) + P(\text{card is diamond})$

This step is addition rule, and a card drawn can't be multiple suits.

$= \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$.

Get the same result!

Super Important: When asked ≥ 1, interpret as the opposite of none!

Eg 1: ≥ 1 H in 10 coinflips is $1 - P(10 \text{ Tails})$

Eg 2: ≥ 1 even result in 10 dierolls is $1 - P(10 \text{ odds})$

Eg 1 worked out: $1 - (\frac{1}{2})(\frac{1}{2})(\frac{1}{2})(\frac{1}{2})(\frac{1}{2})(\frac{1}{2})(\frac{1}{2})(\frac{1}{2})(\frac{1}{2})(\frac{1}{2})$

$1 - (\frac{1}{2})^{10}$

↑ How did we get this? Multiplication rule.

$P(C_1 = T \,\&\, C_2 = T) = P(C_1 = T) \cdot P(C_2 = T \mid C_1 = T)$

$= P(C_1 = T) \cdot P(C_2 = T)$ b/c C_1 gives no info about C_2

Because no coin tosses give any info about any other coin toss, we multiply all 10 probabilities.

2. Sampling - getting elements from a population to determine something about a population

- Deterministic - Not random, elements are picked.
- Probability sample: You know probability that an element will enter your sample.

Tix from jar again: (let's say ticket A drawn ⇒ person A is sampled)

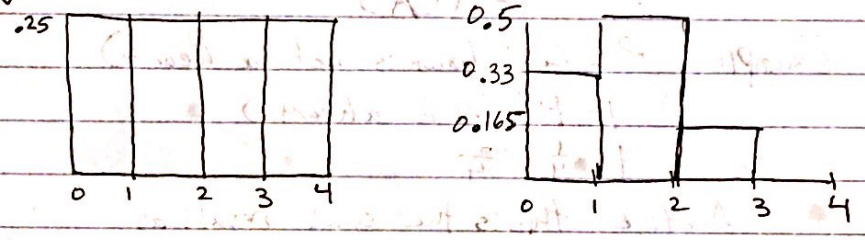
1G, 2R, 3Y, 4B.

So B has $\frac{4}{10}$ chance of entering sample.

- Pitfall: Sample of convenience.
  - If you conduct an internet poll for computer literacy skills, is that an accurate poll of the general population?

- Probability vs. empirical distribution.

Eg.



Probability dist.

as $n \to \infty$, empirical dist converges to probability dist.

Similarly, $\frac{\#\text{event A}}{\text{total events}} \longrightarrow$ true probability

All the events converging to their probability means that their collection look like the probability dist.

- Parameter vs. Statistic

# assoc. w/ pop.	# calculated from a sample.
"true val"	estimated val
fixed (eg, prop. of a certain age)	depends on your sample collected.

Some random programming stuff.

- Booleans: $a < b < c \equiv a<b$ and $b<c$.
  bools can be treated as 0s & 1s in Python.
  np.count_nonzero(array_of_bools). is useful.

- for ␣ in ____: ; for i in np.arange(3): → 0, 1, 4
  evaluation ; k = i**2
  ; print(k)

Typically don't put a return statement in evaluation block: will kill for loop.

if ____:
  evaluation ←

Returns can totally be put here.
You'll practice these in lab.

- Join: T1.join(Col from T1, T2, Col from T2)
  - Remember order matters!
  - Gotcha: if T1 < T2, you grab first rows from T2.

★ Discussion time

- Group: "Collecting" together tables by "collapsing" repeated values in a column.

Imagine this:

Table1 =

Label	Num.
A	5
A	10
B	2
B	4
C	3

Table1.group('Label', some_func) →

Label	Num some_func
A	some_func(array of vals from rows w/ A)
B	some_func(array of vals from rows w/ B)
C	some_func(array of vals from rows w/ C)

some_func gives a single number representing the array specified abov
eg. sum, mean, min, max. Also user defined fxns that take array & retur

Pivot. Either 2 or 4 args.

T2=

Label 1	Label 2	Val 1	Val 2.
A	X	2	3
B	Y	4	5
A	Z	6	7
B	Y	8	9
C	Z	10	11.

Pivot takes arg 1 as column names
arg 2 as row names

T2.pivot('Label 1', 'Label 2') → Default: count.

Label 2	A	B	C
X	1	0	0
Y	0	2	0
Z	1	0	1

T2.pivot('Label 2', 'Label 1', 'Val 1', np.mean)

Label 1	X	Y	Z
A	2	0	6
B	0	6	0
C	0	0	10

Pivot is a good way to display stats relevant to two parameters: each parameter is from a column in table.