Computer Science

CSC380: Principles of Data Science

Probability 4

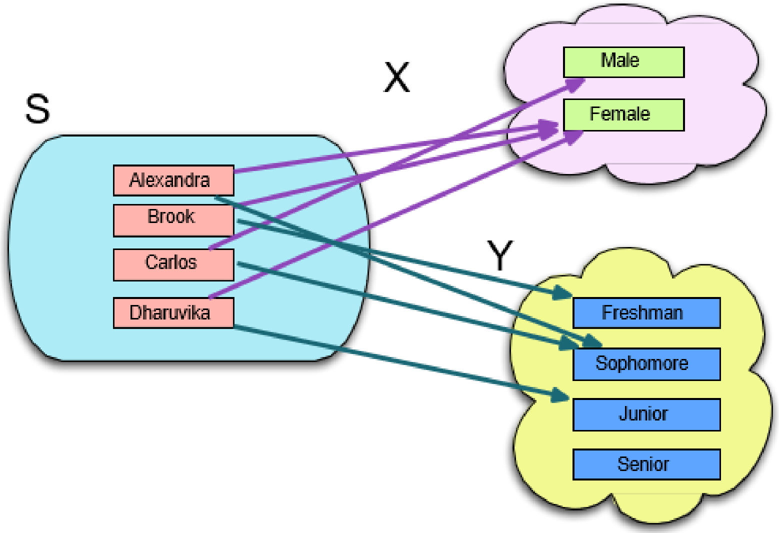
Xinchen Yu

Outline

- Multivariate Random Variables
  - Joint distribution vs. Marginal distribution
  - Independence of RVs
- Expectation and Variance Revisited
  - Covariance, correlation
- Example multivariate RVs
- Law of Large Numbers
- Central Limit Theorem

Multivariate Random Variables

Multivariate RVs: example



- X: people -> their genders
- Y: people -> their class year
- We'd like to answer questions such as: does X and Y have a correlation?
  - I.e., is a student in higher class year more likely to be male?
- We call (X, Y) a random vector, or a multivariate RV, and will study its *joint* distribution

Joint distribution of discrete RVs

- The joint PMF (probability mass function) of discrete random variables X, Y:

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

Examples

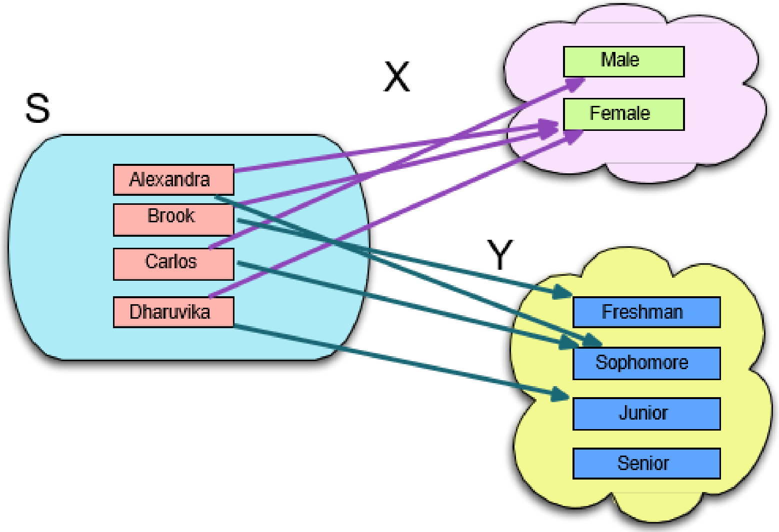
Alexandra

$$P(X = \text{Fem}, Y = \text{Soph}) = \frac{1}{4}$$

Dharuvika

$$P(X = \text{Fem}, Y = \text{Jun}) = \frac{1}{4}$$

…



Joint distribution of discrete RVs

- X: # of cars owned by a randomly selected household
- Y: # of computers owned by the same household
- Joint pmf shown with a table

x	y = 1	2	3	4
1	0.1	0	0.1	0
2	0.3	0	0.1	0.2
3	0	0.2	0	0

- Probability that a randomly selected household has ≥ 2 cars and ≥ 2 computers?
  - $P(X \geq 2, Y \geq 2) = 0.5$

Marginal distributions

Given joint distribution of (X, Y), need distribution of one of them, say X.

- Named the *marginal distribution* of X.

	y			
x	1	2	3	4
1	0.1	0	0.1	0
2	0.3	0	0.1	0.2
3	0	0.2	0	0

- How to find $P(X = x)$?
- Using law of total probability:

$$f_1(x) = \sum_y f(x, y)$$

- This operation is called *marginalization* ('marginalizing out variable Y', or variable elimination)

Marginal distributions

x	y = 1	2	3	4	Total
1	0.1	0	0.1	0	0.2
2	0.3	0	0.1	0.2	0.6
3	0	0.2	0	0	0.2
Total	0.4	0.2	0.2	0.2	1.0

f_1: *marginal distribution* of X

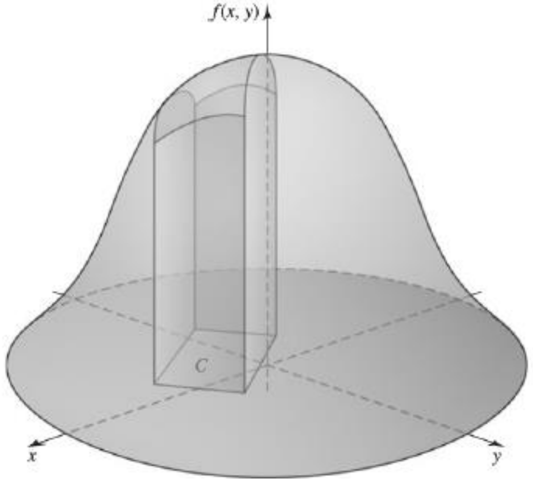
f_2: *marginal distribution* of Y

$$f_1(X = 1) = \sum_y f(1, y) = 0.1 + 0 + 0.1 + 0 = 0.2$$

Joint distribution of continuous RVs

- Any continuous random vector (X,Y) has a *joint probability density function* (PDF) $f(x, y)$, such that for all C,

$$P\big((X, Y) \in C\big) = \iint\limits_{C} f(x, y)\, dx\, dy$$



$f(x, y)$: represent a 2D surface

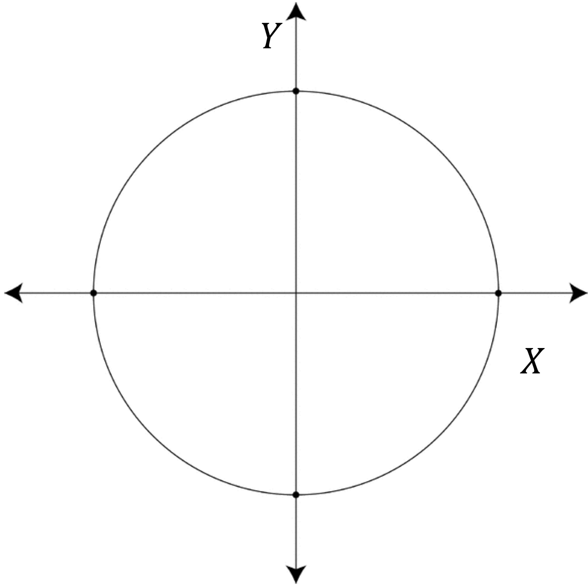
double integral: the *volume* under the surface

Properties:

- f is nonnegative
- $\iint_{R^2} f(x, y)\, dx\, dy = 1$ ($R^2 =$ the whole x-y plane)
  - $P\big((X, Y) \in R^2\big) = 1$

Example: dartboard

- Dartboard with center (0,0) and radius 1; dart lands uniformly at random on the board



- What is the joint PDF of (X, Y)?

- Fact: the PDF is

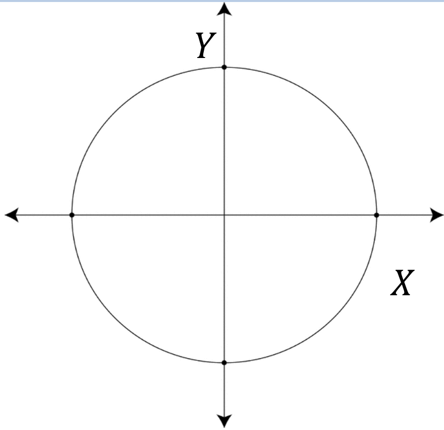
$$f(x,y) = \begin{cases} c, x^2 + y^2 \leq 1 \\ 0, \text{otherwise} \end{cases}$$

- This is called “the Uniform distribution over the unit disk”

Example: dartboard

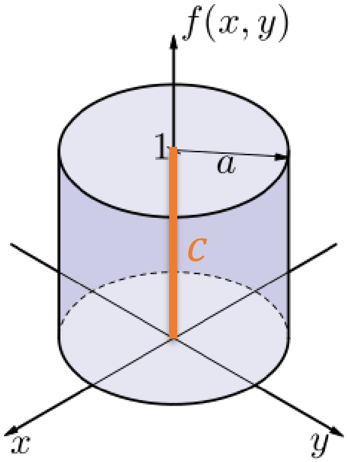
The PDF of X, Y is

$$f(x,y) = \begin{cases} c, x^2 + y^2 \leq 1 \\ 0, \text{otherwise} \end{cases}$$



Can we find c?

Observe: volume under $f(x,y)$ is πc (cylinder) which must also be 1



Therefore, $c = 1/\pi$

Marginal distribution of continuous RV

Given joint distribution of continuous RV (X, Y), how to find X's PDF f_1?

Fact (marginalization) $f_1(x) = \int_R f(x, y)\, dy$

Replacing summation with integration in the continuous case ('marginalizing / integrating out variable Y')

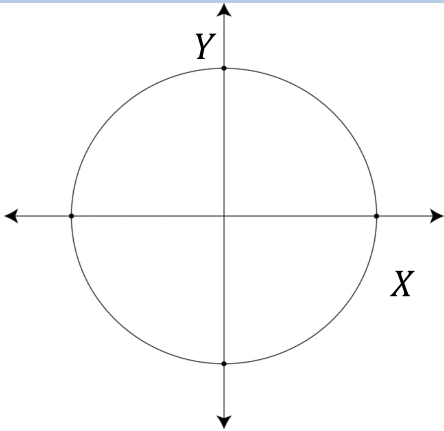
How about Y's PDF f_2?

- Marginalize out X

Example: dartboard

The PDF of X, Y is

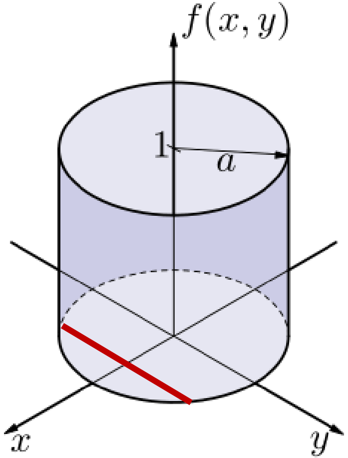
$$f(x,y) = \begin{cases} \frac{1}{\pi}, x^2 + y^2 \leq 1 \\ 0, \text{otherwise} \end{cases}$$



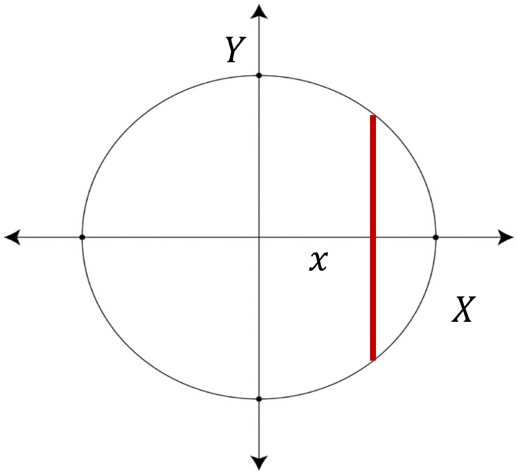
What is the marginal distribution over X?

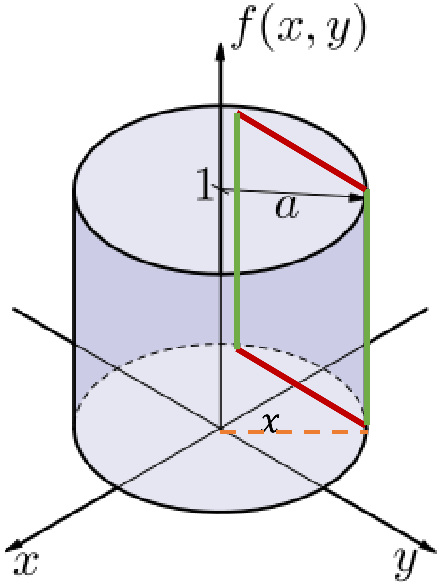
$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)\, dy$$

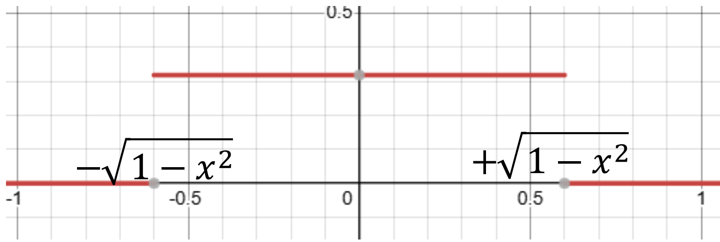
How to find this integral?



Example: dartboard



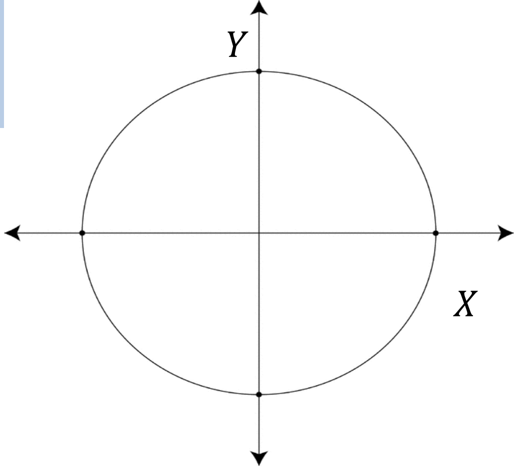




For a fixed $x \in [-1, 1]$, we can think of $f(x)$ is the area of the slice:

- height: $\frac{1}{\pi}$, width: $2 \cdot \sqrt{1-x^2}$
- $f_1(x) = \frac{2}{\pi} \cdot \sqrt{1-x^2}$

Example: dartboard



- In summary,

$$f(x) = \begin{cases} \frac{2}{\pi} \cdot \sqrt{1 - x^2}, x \in [-1,1] \\ 0, \qquad \text{otherwise} \end{cases}$$

X's distribution is NOT $\mathrm{Uniform}([-1,1])$!

Actually makes sense: X closer to 1 is harder to be hit

Joint distribution of more than 3 RVs

- We can consider the joint distribution of more than 3 random variables,
  - E.g. (A,B,C), A = gender, B = class year, C = blood type
- Discrete RVs: can still define joint PMFs

a	b	c	$P(A=a, B=b, C=c)$
0	0	0	0.06
0	0	1	0.09
0	1	0	0.08
0	1	1	0.12
1	0	0	0.06
1	0	1	0.24
1	1	0	0.10
1	1	1	0.25

Marginalization

a	b	c	$P(A=a,B=b,C=c)$
0	0	0	0.06
0	0	1	0.09
0	1	0	0.08
0	1	1	0.12
1	0	0	0.06
1	0	1	0.24
1	1	0	0.10
1	1	1	0.25

Given the joint distribution of (A, B, C)

- What is the distribution of A?
  - Need to find $P(A=0)$ and $P(A=1)$

$$P(A=0) = \sum_{b,c} P(A=0, B=b, C=c)$$

Marginalization: summing over irrelevant variables

- What is the joint distribution of (A, B)?
  - Need to find $P(A=0, B=0), \ldots, P(A=1, B=1)$

$$P(A=0, B=0) = \sum_{c} P(A=0, B=0, C=c)$$

Marginalization for continuous RVs

Suppose joint PDF of (A, B, C) is $f(a, b, c)$

- What is the PDF of A?

$$f_A(a) = \iint_{R^2} f(a, b, c)\ db\ dc$$

- What is the joint PDF of (A, B)?

$$f_{A,B}(a, b) = \int_R f(a, b, c) dc$$

Marginalization: summing over irrelevant variables

- These operations generalize to joint PDFs of more RVs..