Key Ideas from Probability Theory

ISTA 410 / INFO 510 - Bayesian Modeling and Inference

University of Arizona School of Information

January 27, 2021

Outline

Last time:

- Describing distributions with PMFs, PDFs, and CDFs
- Using SciPy to compute PDFs and draw random samples

Outline for today:

- Joint distribution of several variables
- Conditional probability and independence
- Marginal distributions and marginalization

Joint probability

Joint probability

We can talk about the joint probability of two events:

$$P(A \cap B) = \text{probability of } A \textit{ and } B$$

or relatedly, joint probability distribution of two random variables, X, Y, which assigns probabilities to (sets of) ordered pairs

$$(x, y) \in S_X \times S_Y$$

where $S_.$ refers to the sample space of that random variable.

Joint probability

When X, Y are both discrete, you can think of the joint PMF as a table:

X \Y	0	1	2	3
1	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$
2	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{10}$
3	$\frac{1}{30}$	$\frac{1}{30}$	0	$\frac{1}{10}$

Joint distribution defined on pairs (x, y)

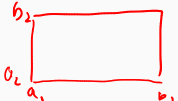
Joint probability

For continuous random variables, we have a joint PDF $p(x, y)$ with the property that

$$\Pr(A) = \iint_A p(x, y)dxdy$$

where A is a subset of the product sample space $S_X \times S_Y$; that is, a set of ordered pairs (x, y).

$$\int_{a_1}^{b_1}\int_{a_2}^{b_2} p(x,y)\,dx\,dy$$ — not always the case

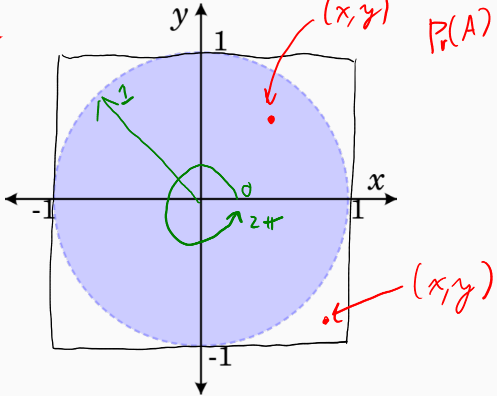


Events in a joint distribution

Note: not every event in a joint probability distribution can be written as a product of events in each variable.

Let X, Y be independent and uniformly distributed on $[-1, 1]$, and A the event that (x, y) falls inside the unit disk:

A = not a product of an event in X and an event in Y.



$Pr(A) = ?$
$= \pi/4$

Changing variables

Let's say, to make the integral easier, we did want to express A as a simple product of events. What would we do? Change variables.

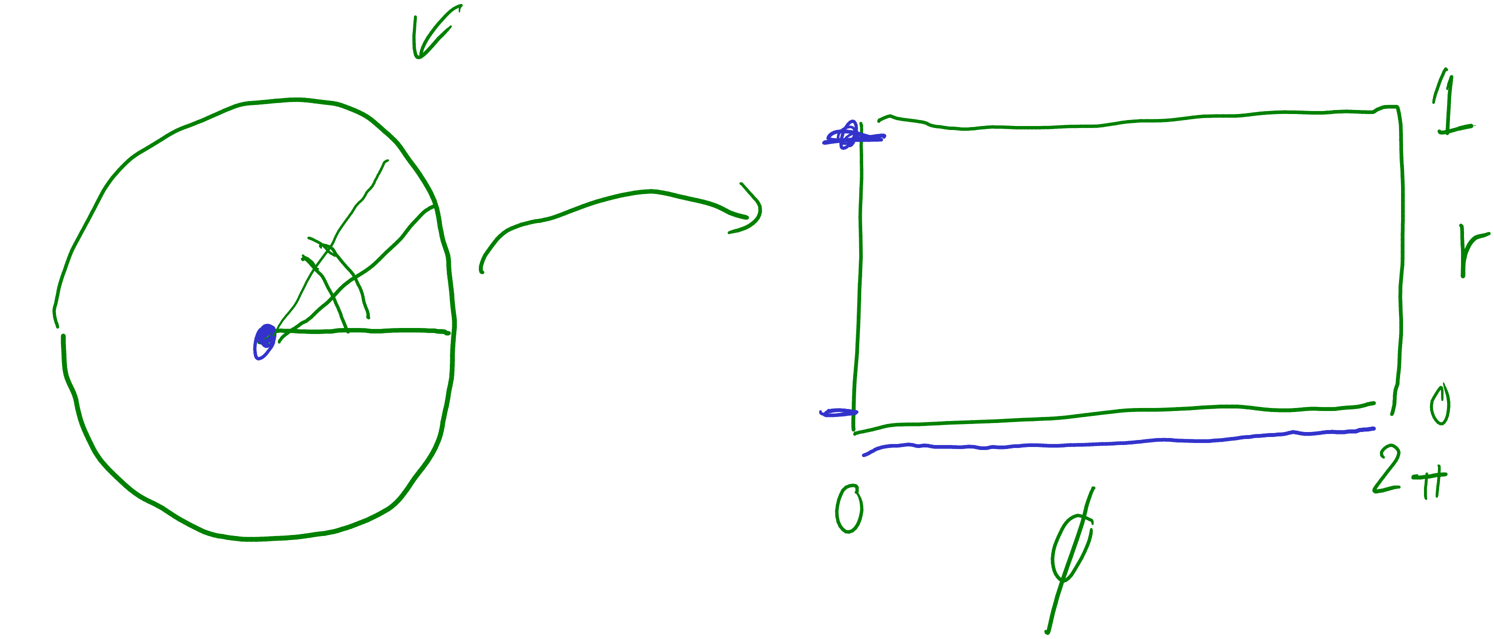
radius

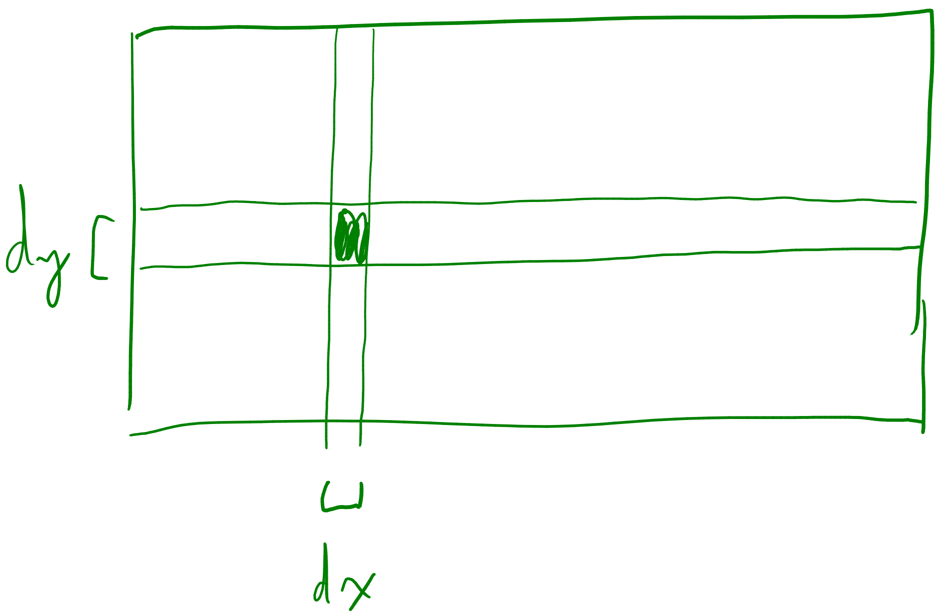
If (r, ϕ) are the distance from 0 and angle from the positive x axis respectively, then A is just the event $r \leq 1$. So, we can try to integrate:

$$\int_0^1 \int_0^{2\pi} p(r, \overset{\phi}{\cancel{\theta}})drd\overset{\phi}{\cancel{\theta}}$$

but we need to make an adjustment to account for geometric factors.

- Original PDF: $p(x, y) = \frac{1}{4}$
- If we just use the same PDF, $\Pr(A) = \pi/2$; obviously wrong!


1
r
0
0
2π
φ


dy
dx

Changing variables

To get it right, we need to think of the function that transforms between the two sets of variables:

- $x(r, \phi) = r \cos \phi$
- $y(r, \phi) = r \sin \phi$

The *Jacobian* of this transformation is the matrix of partial derivatives:

$$J = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \phi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \phi} \end{pmatrix}$$

Changing variables

The correction for changing variables is the absolute value of the determinant of the Jacobian:

$$J = \begin{pmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{pmatrix}$$

$$\det J = r(\cos^2\phi + \sin^2\phi) = \boxed{r}$$

so the corrected integral is

$$\int_0^1 \int_0^{2\pi} p(r,\theta)\, r\, dr d\theta = \int_0^1 \int_0^{2\pi} \frac{r}{4} dr d\theta = \frac{\pi}{4}$$

which agrees with geometric intuition.

When you might use this

This change-of-variables calculation isn't something you'll need to do all that often, but:

- sometimes, it will make sense to apply a distribution to a transformed parameter
- common use case: apply a prior distribution to $\log \theta$ instead of θ, especially for "scale" parameters like variances

Conditional probability and independence

Conditional probability and independence

If the probability of an event represents our knowledge about that event, we should be able to "update" this knowledge by incorporating observations:

$$\Pr(E|H) = \text{“probability of E given H”}$$

E and H are said to be *independent* if $\Pr(E|H) = \Pr(E)$.

Multiplication rule for probabilities

The multiplication or *chain rule* for probabilities of intersections of events is:

$$\Pr(E \cap H) = \Pr(E|H)\Pr(H) = \Pr(H|E)\Pr(E)$$

Intuitive interpretation:

- the probability that E and H both happen is the probability that H happens times the probability that E happens if we assume H happened

  $P(H)$
  $\times P(E|H)$

- the probability that E and H both happen is the probability that E happens times the probability that H happens if we assume E happened

  $P(E)$
  $\times P(H|E)$

Independence

This leads to an alternative characterization of independence for events; two events are independent if:

$$\mathrm{Pr}(E \cap H) = \mathrm{Pr}(E)\mathrm{Pr}(H)$$

Often this is taken as the starting definition of independence.

More relevant for random variables: two RVs described by PMFs or PDFs are independent if the joint PMF/PDF factors:

$$p(x, y) = p(x)p(y)$$

(This can be used either to write down a joint PDF for independent variables, or to argue independence)

Pairwise vs. mutual independence

One of the homework problems deals with the issue of pairwise or mutual independence:

- pairwise independence of $A_1, A_2, A_3, \ldots$: given any two i, j, $\Pr(A_i \cap A_j) = \Pr(A_i)\Pr(A_j)$.
- mutual independence of $A_1, A_2, A_3, \ldots$: given any subset $i_1, i_2, \ldots, i_n$, $\Pr(A_{i_1} \cap \ldots \cap A_{i_n}) = \Pr(A_{i_1}) \ldots \Pr(A_{i_n})$

Example from the homework

$A = \{$Alex & Betty have the same birthday$\}$

$B = \{$Betty & Carlos have the same birthday$\}$

$C = \{$Alex & Carlos have the same birthday$\}$

C is ind. of A

C is ind. of B

C is not independent of $(A \cap B)$

Marginalization

Marginal distributions

If we have a joint distribution $p(x, y)$ of two variables, we can also ask about the distributions of the individual variables: what are $p(x)$ and $p(y)$?

These are the *marginal distributions*, and the answer is easy if X and Y are independent, of course. But in general, to get $p(x)$ we must average over the possible values of y and vice versa.

Y

Example: marginalizing over a discrete variable

Slight modification of BDA exercise 1.1: let θ be a random variable with $\Pr(\theta = 0) = 0.25, \Pr(\theta = 1) = 0.75$. Then let y be a random variable with a distribution conditional on the value of θ:

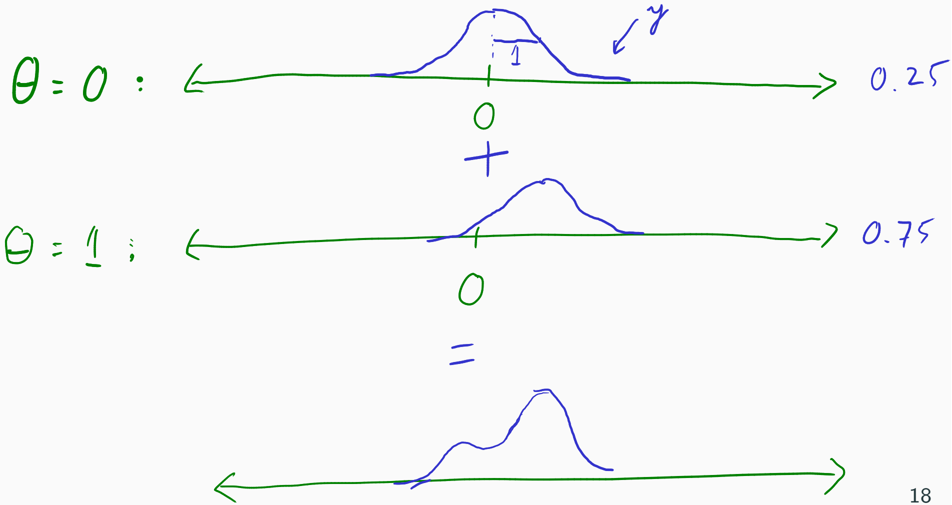
$$y \sim \mathrm{Normal}(\theta, 1)$$

so, y is normally distributed with fixed standard deviation, but its mean depends on the value of θ.

What is the marginal distribution of y?

Example: marginalizing over a discrete variable

The joint distribution is a distribution defined on two copies of the real line:



Example: marginalizing over a discrete variable

$$N(0,1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2}\right)$$

So the marginal distribution of y is the weighted sum:

$$p(y) = 0.25N(0,1) + 0.75N(1,1)$$

or, more explicitly:

$$p(y) = \underbrace{\boxed{\frac{1}{\sqrt{2\pi}}}}_{\text{normalizing const}}\left(\frac{1}{4}\underbrace{\exp\left(-\frac{x^2}{2}\right)}_{\text{mean } 0} + \frac{3}{4}\underbrace{\exp\left(-\frac{(x-1)^2}{2}\right)}_{\text{mean } 1}\right)$$

Marginal distribution

In general, you get the marginal by summing/integrating out the "unwanted" variable:

$$p(x) = \sum_i p(x, y_i)$$

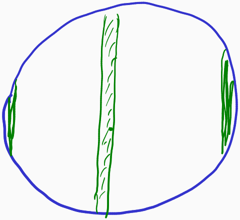
$$p(x) = \int p(x, y) dy$$

(limits of the integral) = range of y

Example: marginalizing over a continuous variable

One more example. Choose a point (x, y) uniformly at random from a unit disk. What is the marginal distribution of the x coordinate?

- Although it looks like the density function is a constant, the coordinates are not really independent
- Intuitively: x near ± 1 unlikely because there's not much area there in the disk



density $\propto \underline{1}$

Example: marginalizing over a continuous variable

For a given x, $p(x)$ is given by integrating over y:

$$p(x) = \underbrace{\frac{1}{\pi}}_{\text{normalizing constant}} \int_{-1}^{1} \underbrace{\mathbb{1}_{[-\sqrt{1-x^2},\sqrt{1-x^2}]}}_{\mathbb{1}_{[-\sqrt{1-x^2},\sqrt{1-x^2}]}(y) \,=\, \text{indicator function}} dy$$

$$\underbrace{[-\sqrt{1-x^2}, \sqrt{1-x^2}]}_{= 1 \text{ on this interval, } 0 \text{ elsewhere.}}$$

Example: marginalizing over a continuous variable

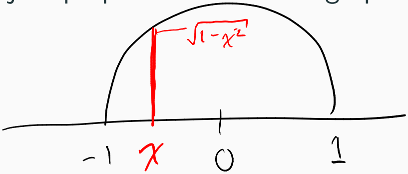
For a given x, $p(x)$ is given by integrating over y:

$$p(x) = \frac{1}{\pi}\int_{-1}^{1} \mathbb{1}_{[-\sqrt{1-x^2},\sqrt{1-x^2}]} dy$$

$$p(x) = \frac{2}{\pi}\int_{0}^{\sqrt{1-x^2}} dy = \frac{2}{\pi}\sqrt{1-x^2}$$

marginal density of x.

(unsurprisingly this is just proportional to a the graph of a semicircle!)



Bayes' theorem

The theorem that gives Bayesian statistics its name is a seemingly trivial rearrangement of the chain rule:

$$\Pr(E \cap H) = \underline{\Pr(E|H)\Pr(H) = \Pr(H|E)\Pr(E)}$$

to

$$\underbrace{\Pr(H|E)}_{\text{inference}} = \frac{\overbrace{\Pr(E|H)}^{\text{model}}\Pr(H)}{\Pr(E)}$$

The significance comes when we assign interpretations to H and E of "hypothesis" and "evidence" respectively.

The cookie problem

Suppose we have two bowls of cookies.[1] Bowl 1 has 30 vanilla and 10 chocolate cookies; Bowl 2 has 20 of each.

We select a bowl at random and, without looking at which one we picked, pull a cookie at random from it. The cookie is vanilla.

What is the probability that our randomly selected bowl was Bowl 1?

[1] This example is from *Think Bayes* by Allen Downey.

The cookie problem

Summary

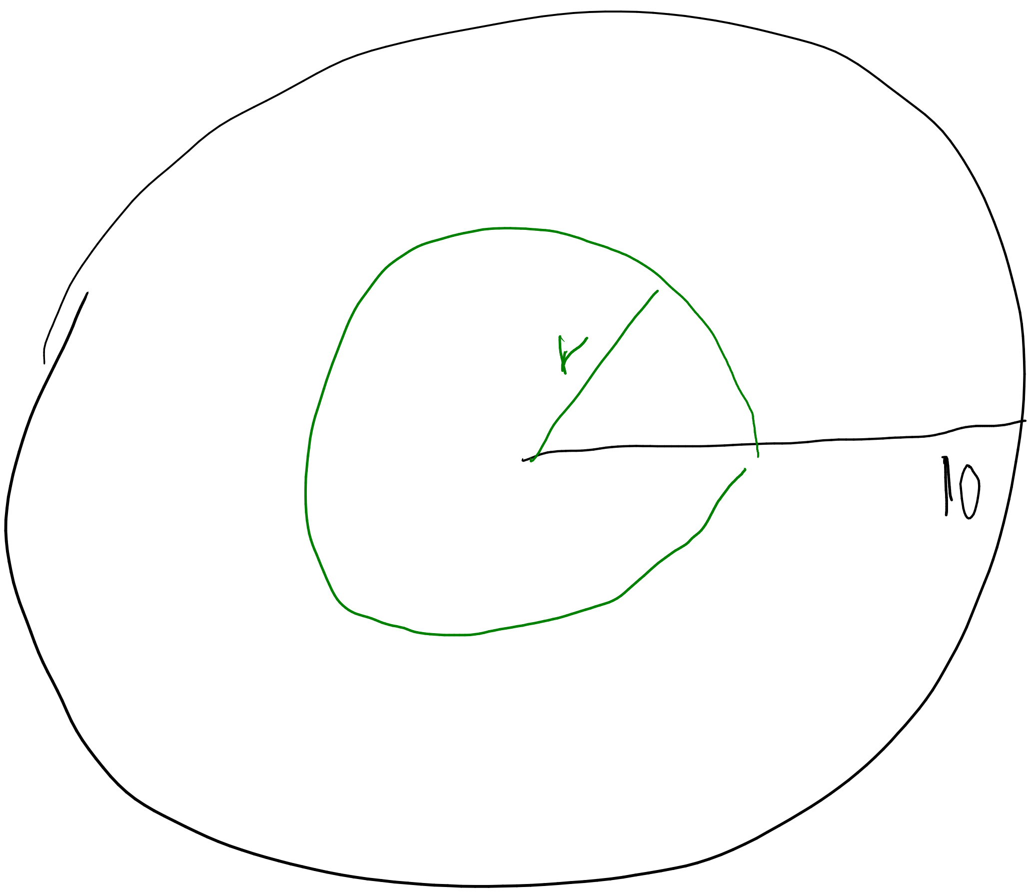
Today:

- Joint, conditional, and marginal distributions
- Marginalization
- Changes of variables
- ~~Bayes' theorem~~

Next week:

- Defining some models and making inferences

- If you need extra time on HW0, just ask
- Read sections 2.1-2.3 of BDA



Pr(dart falls within r of center)

= F(r)