Big Data and Security

Jeffrey Borowitz, PhD

Lecturer

Sam Nunn School of International Affairs

Probability Part 2:
Combining Random Variables

Multivariate Distributions

- You can have multiple random variables described together
- They have a joint distribution f (x, y) For rolling two dice, the distribution is:

	1	2	3	4	5	6	Subtotal
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
Subtotal	1/6	1/6	1/6	1/6	1/6	1/6	

Georgia Tech

A More Complicated Distribution

- What if every time one die was different than the other by 3 or more, we rerolled?

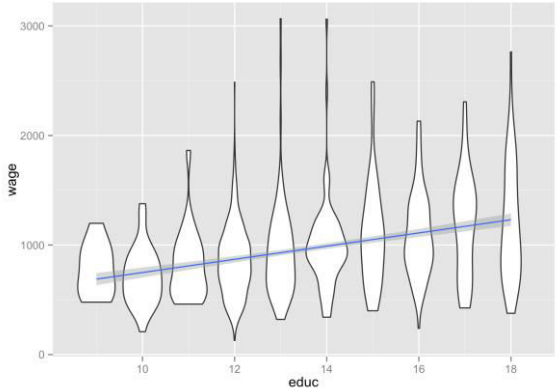
x/y	1	2	3	4	5	6	Subtotal
1	1/24	1/24	1/24	0	0	0	1/8
2	1/24	1/24	1/24	1/24	0	0	1/6
3	1/24	1/24	1/24	1/24	1/24	0	5/24
4	0	1/24	1/24	1/24	1/24	1/24	5/24
5	0	0	1/24	1/24	1/24	1/24	1/6
6	0	0	0	1/24	1/24	1/24	1/8
Subtotal	1/8	1/6	5/24	5/25	1/6	1/8	

- **Marginal distribution** of second die roll f (y)? The bottom row. . .
- Conditional on the first roll being 5, what is the **conditional distribution** of y given that x = 5 (denoted f (y|x = 5))? The fifth column to the right. . .

Georgia Tech

A More Realistic Distribution

- The distribution of wages for individuals, conditional on years of schooling completed
  - What is the conditional distribution of wages given 12 years of education?
  - What is the marginal distribution of wages?



What The Heck?

I want you guys to have a nice simple idea of conditional distributions to go back to when we think about more complex topics.

Definitions

- The initial two rolls of dice were independent, which means that the table entries equal the product of the two marginal subtotal column values (i.e. 1/6 × 1/6 = 1/36)
  - This is super intuitive, really: if two things don't depend on each other, than the probability of both events happening is the probability of one times the probability of the other.
- Correlation
  - The correlation between X and Y is the expected amount by which X is above average when Y is above average:

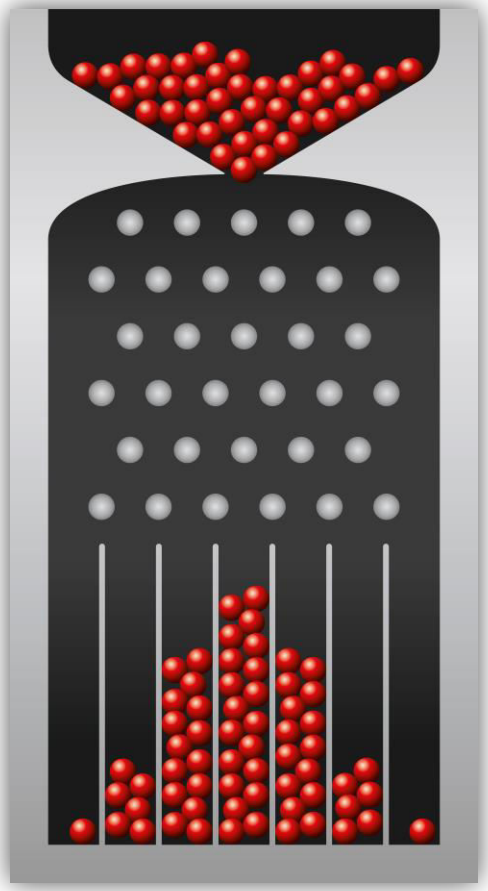
$$E[(X - E[x])(Y - E[Y])]$$

  - If X is high when Y is high, then they are positively correlated (e.g. flu search terms and flu incidence)
  - if X is high when Y is low, then they are negatively correlated (e.g. hours spent studying and rainfall)

More Things About Random Variables

- You can have functions of random variables, which are also random variables
  - The sum of the rolls of two dice
  - 2 times the number that rolls on a single die
  - Sum of heads
- Some straightforward math rules apply that tell you how these component random variables depend on their individual constituents

Sum of Coin Flips



Lesson Summary

- Random variables can have dependencies
    - Examples include constrained dice and wages/education
- Functions of random variables are also random variables
- With sum of coin flips, we're starting to see how random variables relate to data