COSC 290 Discrete Structures

Lecture 39: Probability V

Prof. Michael Hay
Monday, Dec. 11, 2017
Colgate University

Plan for today

1. Problem Set
2. Random Variables, Expectation
3. Randomized response

Problem Set

Problem section

Please turn in the problem set.

Let's review a few questions. Which ones do you want to review?

Random Variables, Expectation

The Ferengi

Figure 1: An alien species in Star Trek notorious for extreme sexism.

http://memory-alpha.wikia.com/wiki/The_Magnificent_Ferengi_(episode)

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Tech interview question

Ferengi want boys, so every family keeps on having children until a boy is born.

- If the newborn is a girl, have another child
- If the newborn is a boy, stop

Can their strategy influence the composition of their population?

Let's think about the sample space, outcomes, and probability.

Let's draw tree diagram, or at least part of it, on board.

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Random variable

A random variable assigns a numerical value to every outcome in sample space.

Despite being called a variable, random variable X is formally a function $X : S \to \mathbb{R}$.

Example: possible outcomes in Ferengi family, and random variables B (# boys) and G (# girls).

Outcome	Probability	B	G
boy	$\frac{1}{2}$	1	0
girl,boy	$\frac{1}{4}$	1	1
girl,girl,boy	$\frac{1}{8}$	1	2
girl,girl,girl,boy	$\frac{1}{16}$	1	3
...	...	...	...

Probability mass function

We can associate a probability with a random variable as follows,

$$Pr(X = x) := Pr(\{ s \in S : X(s) = x \})$$

In other words, we can define an event as the set of outcomes s where random variable X maps s to x.

Example: probability of G and B:

Outcome	Probability	B	G
boy	$\frac{1}{2}$	1	0
girl,boy	$\frac{1}{4}$	1	1
girl,girl,boy	$\frac{1}{8}$	1	2
girl,girl,girl,boy	$\frac{1}{16}$	1	3
...	...	...	...

G	Prob.
0	$\frac{1}{2}$
1	$\frac{1}{4}$
2	$\frac{1}{8}$
3	$\frac{1}{16}$
...	...

B	Prob.
0	0
1	1
2	0
3	0
...	...

Expectation

The expected value of a random variable X, denoted $\mathbb{E}[X]$ is the average value of X, defined as

$$\mathbb{E}[X] := \sum_{s \in S} X(s) \cdot Pr(s)$$

Alternatively, let $Range(X)$ denote the range of values that X can take. The expected value can also be calculated as,

$$\mathbb{E}[X] = \sum_{x \in Range(X)} x \cdot Pr(X = x)$$

Back to the Ferengi...

Expected number of boys?

$$\mathbb{E}[B] = 1$$

Expected number of girls?

$$\mathbb{E}[G] = \sum_{g=0}^{\infty} g \cdot Pr(G = g) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{8} + 3 \cdot \frac{1}{16} + \cdots = ???$$

(We can also use Theorem 10.5 (p. 1046) from textbook.)

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

More expectations

Let's define new random variables: $T = B + G$ and $F = B/(B + G)$

What is $\mathbb{E}[T]$?

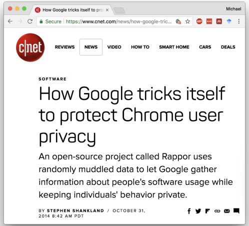
What is $\mathbb{E}[F]$? Write out an expression.

- Hint: more than half!
- $\mathbb{E}[F] = \mathbb{E}[B/(B + G)] \neq \mathbb{E}[B] / \mathbb{E}[B + G]$!

Expected fraction of boys in Ferengi family: $\approx 70\%$!

Randomized response

Using randomization to safely extract private information



Privacy through randomization

Suppose pollster wants to ask sensitive question.

Example: Do you support legalization of marijuana? Respondent may be reluctant to answer "Yes."

Randomized response (Warner, 1965)

- Pollster randomly samples respondent from population
- Respondent flips biased coin (heads with probability $p > \frac{1}{2}$). Result of coin flip hidden from pollster.
- If heads, answers truthfully.
- If tails, lies.

Indicator random variable

Let θ be fraction of population that would truthfully answer Yes to question.

Let X_i be the following indicator random variable,

$$X_i = \begin{cases} 1 & \text{if } i^{th} \text{ respondent gives randomized answer "Yes"} \\ 0 & \text{if } i^{th} \text{ respondent gives randomized answer "No"} \end{cases}$$

What is $Pr(X_i = 1)$?

(Shown on board)

$$Pr(X_i = 1) = \theta p + (1 - \theta)(1 - p)$$

What can we learn about θ?

Suppose we repeat this process with a sample of n respondents.

Let $Y := \sum_{i=1}^{n} X_i$.

What is $\mathbb{E}[Y]$?

(In other words, how many people do we expect, on average, to give a randomized answer of Yes?)

Linearity of expectations

Let X_1 and X_2 be any two random variables.

$$\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$$

and let a be some constant,

$$\mathbb{E}[aX_1] = a\mathbb{E}[X_1]$$

What can we learn about θ?

Suppose we repeat this process with a sample of n respondents. Let $Y := \sum_{i=1}^{n} X_i$.

What is $\mathbb{E}[Y]$?

- Linearity of expectations: $\mathbb{E}[Y] = \mathbb{E}[\sum_i X_i] = \sum_i \mathbb{E}[X_i]$.
- $\mathbb{E}[X_i] = Pr(X_i = 1)$

Let's rearrange and "solve" for θ:

$$\theta = \frac{\frac{\mathbb{E}[Y]}{n} - (1 - p)}{(2p - 1)}$$

Key point: if you could estimate $\frac{\mathbb{E}[Y]}{n}$, the expected fraction of sampled respondents who give randomized answer of Yes, then you have an estimate for θ, the fraction of the population who would give truthful answer of Yes.

Estimating θ

Let $\hat{\theta}$ denote the following random variable

$$\hat{\theta} := \frac{\frac{Y}{n} - (1 - p)}{(2p - 1)}$$

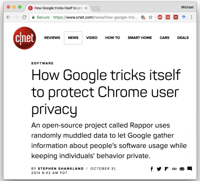
What is $\mathbb{E}\left[\hat{\theta}\right]$? $\mathbb{E}\left[\hat{\theta}\right] = \theta$ (an unbiased estimator)

How accurate is $\hat{\theta}$? We can look at the *variance* of $\hat{\theta}$, which is a measure of how much it deviates from its expected value.

$$V(\hat{\theta}) = \underbrace{\frac{\theta(1-\theta)}{n}}_{\text{error from sampling}} + \underbrace{\frac{p(1-p)}{n(2p-1)^2}}_{\text{error due to randomized answers}}$$

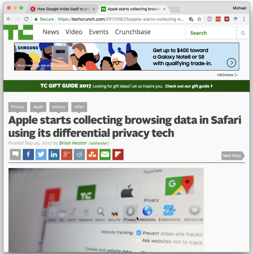
What happens when $p = 1/2$? $p = 1$? $p = 0$? (Note: You can derive this result using definition of V in book and the fact that $V(\sum_i X_i) = \sum_i V(X_i)$ when X_i are independent, which they are here.)

Using randomization to safely extract private information



Google's approach: compress user data using bloom filter, then use randomized response on each bit of bloom filter.

Apple uses similar technologies



Exercise

Consider this alternative randomized protocol.

Flip coin: if heads, answer Yes; if tails, answer truthfully.

What is $\mathbb{E}[Y]$ under this randomized model?

As before,

- assume θ fraction of the population would answer Yes truthfully.
- use linearity of expectations: $\mathbb{E}[\sum_i X_i] = \sum_i \mathbb{E}[X_i]$
- for indicator random variable $\mathbb{E}[X_i] = Pr(X_i = 1)$

Does this approach leak more/less information than previous approach?