

Probability and Random Variables

Raguvir Kunani

Data 100, Discussion 7

July 17, 2019

- exam was fair in length & difficulty

- Video Walkthrough 6/10
- More informative ref sheet
- More term-specific practice exams
- More resources on new topics
- Disproportionate weighting of topics (SQ2)
- more partial credit
- alt answer & explanations for exam solutions

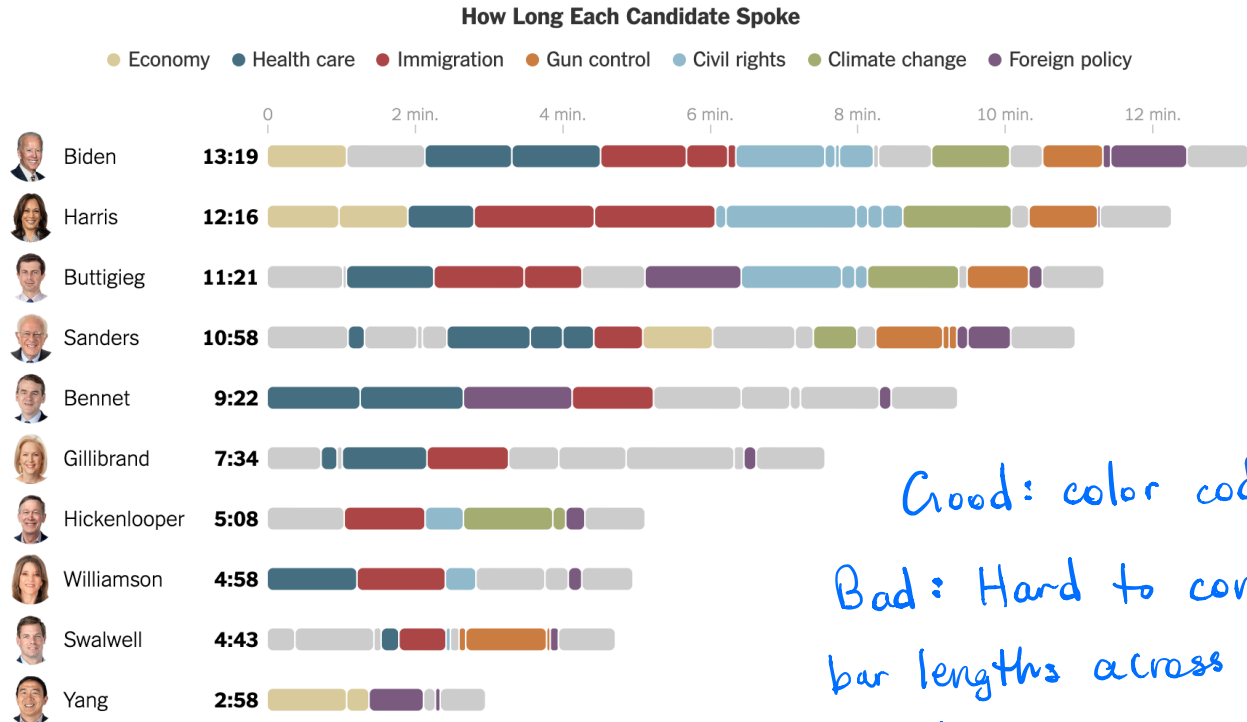
How was the midterm for you?

It's ok if the midterm didn't go so well for you. **There is still a lot of room for you to redeem yourself, so don't give up!** The material for the rest of the course is *slightly* disjoint from the previous material so you have a fresh start.

Vent about the midterm to me! No hard feelings, you can say whatever you want. This is an open space for you to share your feelings.

Data 100 in the News: Data Visualization

One visualization from the 2020 Democratic Presidential debate:



Each bar segment represents the length of a candidate's response to a question.

Good: color coded
Bad: Hard to compare
bar lengths across
candidates

What are some good and bad things in the visualization above?

Probability Foreword

You all have seen probability in some way before. It's scary for a lot of people, and I understand that feeling. I felt the same way too when I took Data 100.

However, I encourage you to clear your mind about any past stressful memories you may have about probability. Our goal in this class is *not* to slaughter you with artificially difficult probability questions.

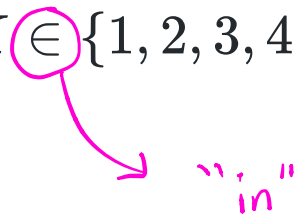
We just want to build a solid foundation in probability to facilitate learning random variables, which are very important tools for modeling and estimation.

Random Variables

A **random variable** (RV) is a variable that can take on many different values (as any variable would). A random variable *also* has a certain probability of taking on each of its values.

For example, let X be a RV denoting the outcome of a roll of a fair, six-sided die. Then, we can say that X can take on values 1, 2, 3, 4, 5, 6 each with probability $\frac{1}{6}$ since the die is fair.

The fancy math way of saying this is $X \in \{1, 2, 3, 4, 5, 6\}$ and $P(X = x) = \frac{1}{6}$ for all $x \in X$.

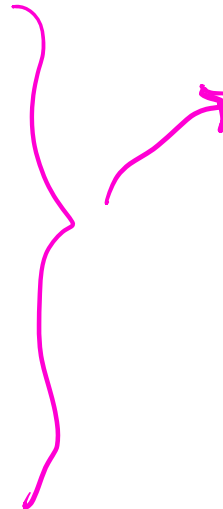


Probability Distributions

A **probability distribution** of a RV specifies the probability of each value a RV can take on. You can think of a probability distribution as a table with 2 columns: value and probability of the value.

In our die example, this is the probability distribution of X :

Value	Probability
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$



notice the probabilities
add up to 1!

x	$f(x)$
0	0
1	1
2	2
\vdots	\vdots

$\Rightarrow f(x) = x$

Probability Mass Functions

Probability distributions can sometimes be summarized with closed-form functions (these functions are called **probability mass functions** (PMFs)).

One example of a PMF:

$$P(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

this is actually a PDF (since the normal distribution is continuous)

Don't worry, you will never have to worry about something this complicated. But does this look familiar?

Common Random Variables and their PMFs

X is called a Bernoulli RV if its PMF follows the $Bern(p)$ distribution:

p : prob. of success

$$P(X = 1) = p$$

coin flip: $Bern(\frac{1}{2})$

$$P(X = 0) = 1 - p$$

X is called a Binomial RV if its PMF follows the $Binom(n, p)$ distribution:

same p
as Bern

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

"n choose k"

10 coin flips: $Binom(10, \frac{1}{2})$

Expectation and Variance

These are the key elements of RVs you need to know. The formulas for them can be hard to grasp at first, but if we look at each formula piece by piece, they aren't so bad.

$$E[X] = \sum_{x \in X} x \cdot P(X = x)$$

weight each probability
by the value

$$Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

distance
from
average

algebra

Expectation and Variance of Common RVs

If $X \sim \text{Bern}(p)$, what is $E[X]$ and $\text{Var}[X]$?

$$E[X] = \sum_a a \cdot P(X=a) = 1(p) + 0(1-p) = p$$

$$\text{Var}[X] = E[X^2] - E[X]^2 = \left[1^2(p) + 0^2(1-p) \right] - p^2 = p - p^2 = p(1-p)$$

\uparrow
 $E[X^2]$

If $X \sim \text{Binom}(n, p)$, what is $E[X]$ and $\text{Var}[X]$?

$$X = \sum_{i=1}^n X_i, \quad X_i \sim \text{Bern}(p)$$

$$E[X] = E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n p = np$$

$$\text{Var}[X] = \text{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \text{Var}[X_i] = \sum_{i=1}^n p(1-p) = np(1-p)$$

this assumes
the X_i are
independent.

Can you see
where that

assumption
is used?