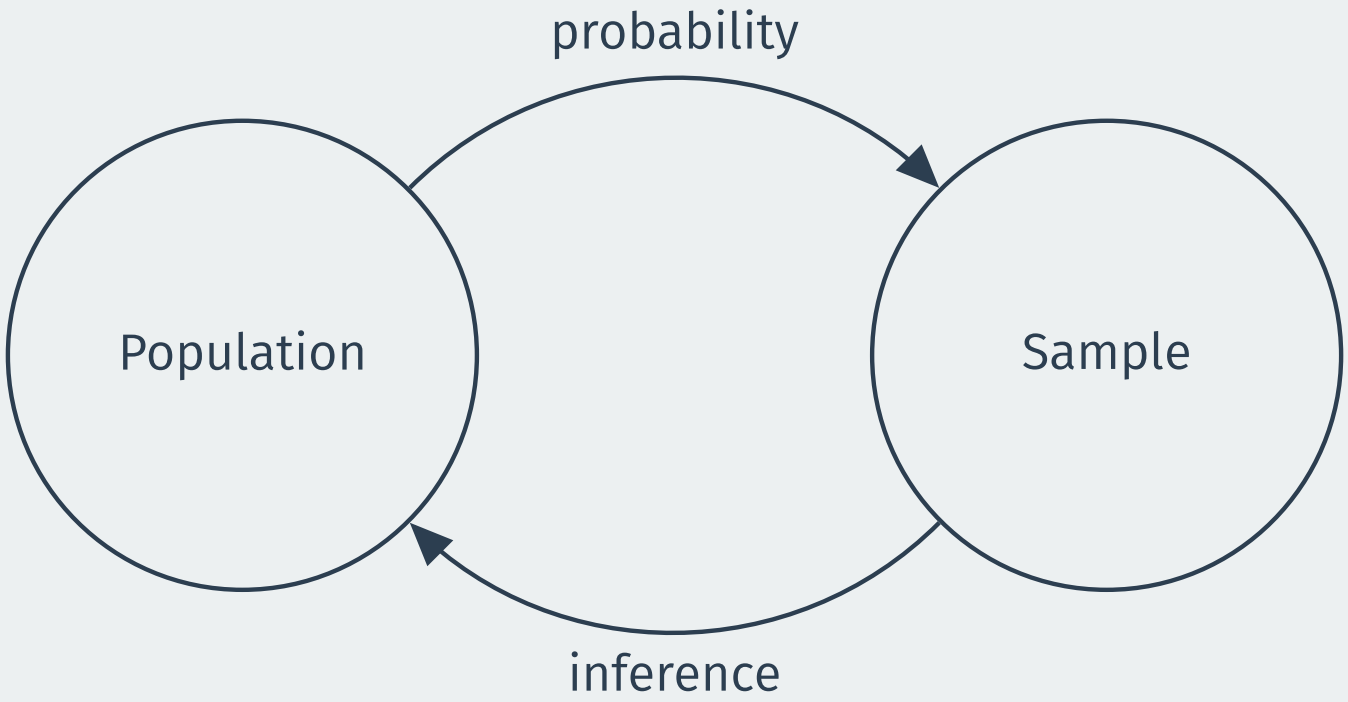
Gov 51: Probability Distributions

Matthew Blackwell

Harvard University

Remember our goal



- We want to learn about the chance process that generated our data.
- More specifically: learn about the **distribution** of the r.v.s in our data.
    - What values of the r.v. are more or less likely?

Probability distribution

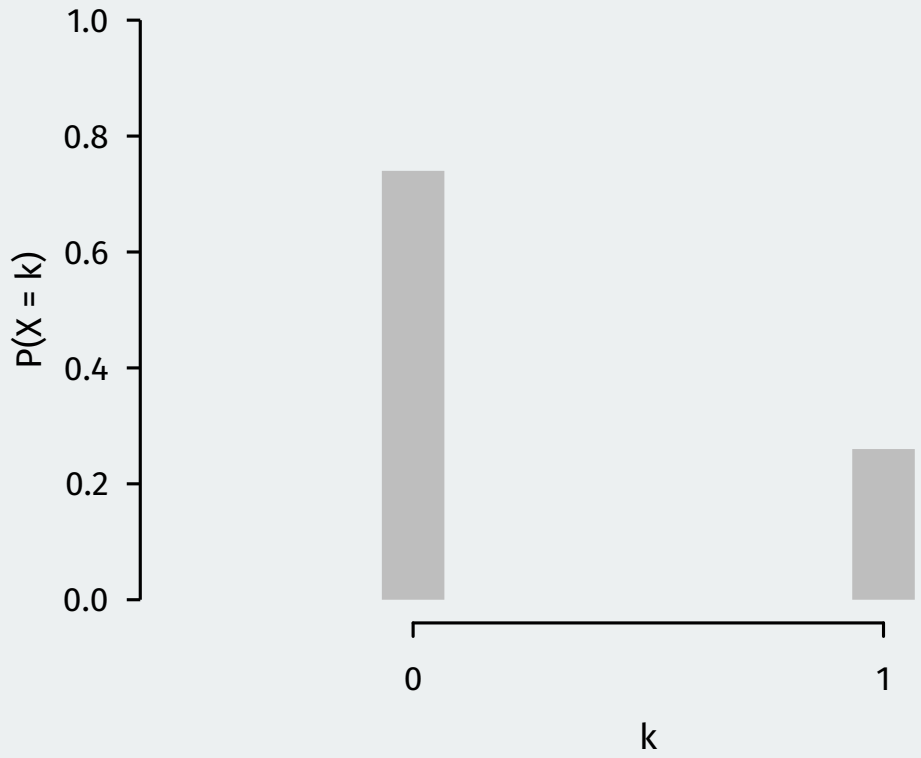
- **Probability distributions** describe the uncertainty of a random variable.
  - Functions that give the probability of different possible values of an r.v.
  - Why do we care? **Learning about populations from samples**
- Simple example: suppose we randomly sample a single U.S. adult.
  - Let X be 1 if they support Trump, 0 otherwise.
  - X is Bernoulli with some probability p
  - Learning p would give us the probability a random adult supports Trump!
- Multiple ways to represent the distribution.
  - Depends on what kind of r.v. we have.

Types of random variables

- **Discrete**: X can take a finite (or countably infinite) number of values.
  - Number of heads in 5 coin flips
  - Sampled senator is a woman ($X = 1$) or not ($X = 0$)
  - Number of battle deaths in a civil war
- **Continuous**: X can take any real value (usually within an interval).
  - GDP per capita (average income) in a country.
  - Share of population that approves of Trump.
  - Amount of time spent on a website.

Probability mass functions

- For discrete r.v.s: **probability mass function (PMF)**
  - Gives the probability of each possible value, $\mathbb{P}(X = k)$.
  - Like a bar plot for the population shares of each value.
  - Here's the PMF for the Bernoulli of drawing a woman senator:



Binomial PMFs

- PMFs expressed in mathematical formulas depending on **parameters**.
  - Binomial with n draws and probability of "success" p:

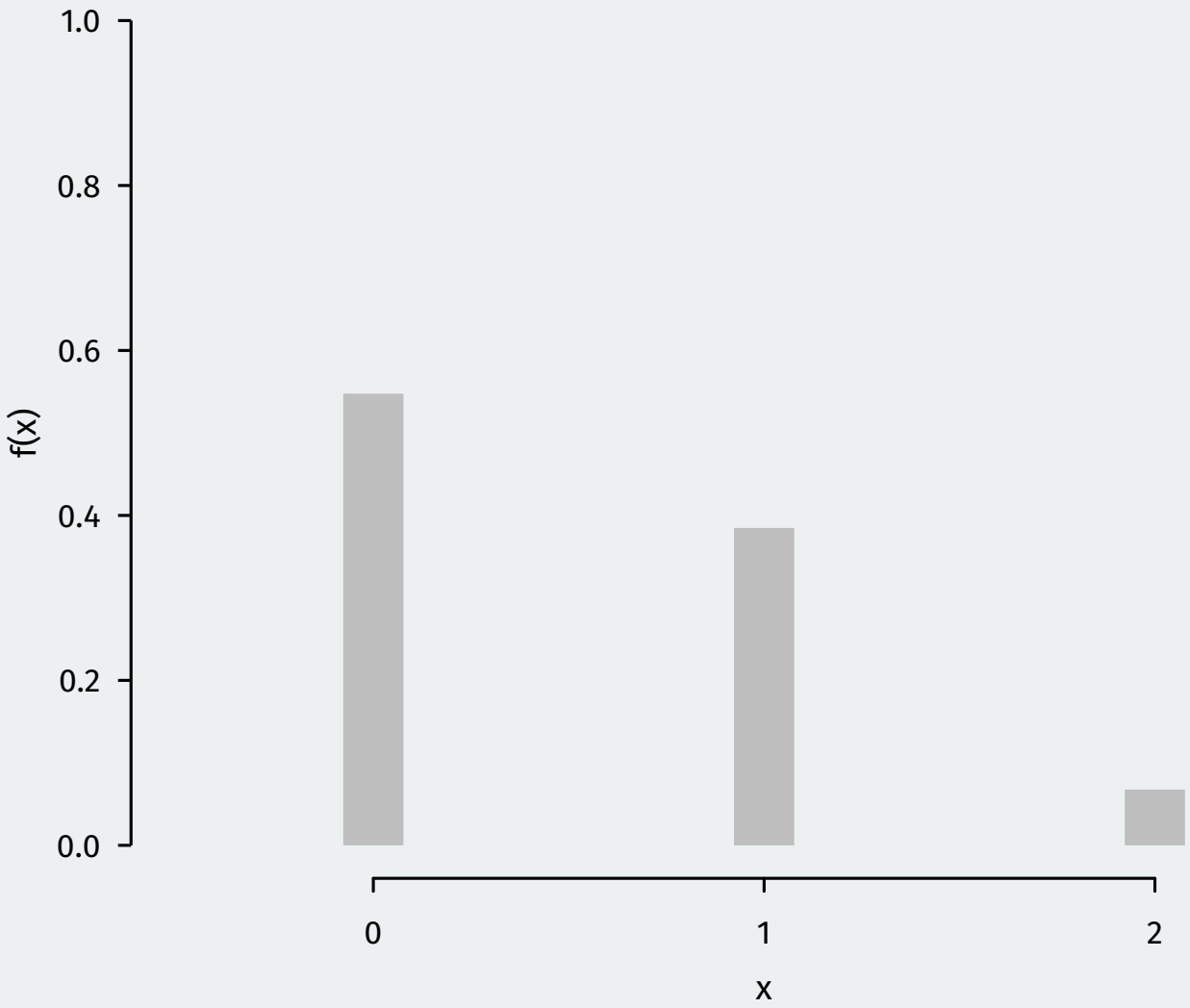
$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

  - We'll almost always use R to calculate the PMF.
- We can use the `dbinom` function to calculate the PMF of a Binomial r.v.

```
dbinom(x = c(0, 1, 2), size = 2, prob = 26/100)
```

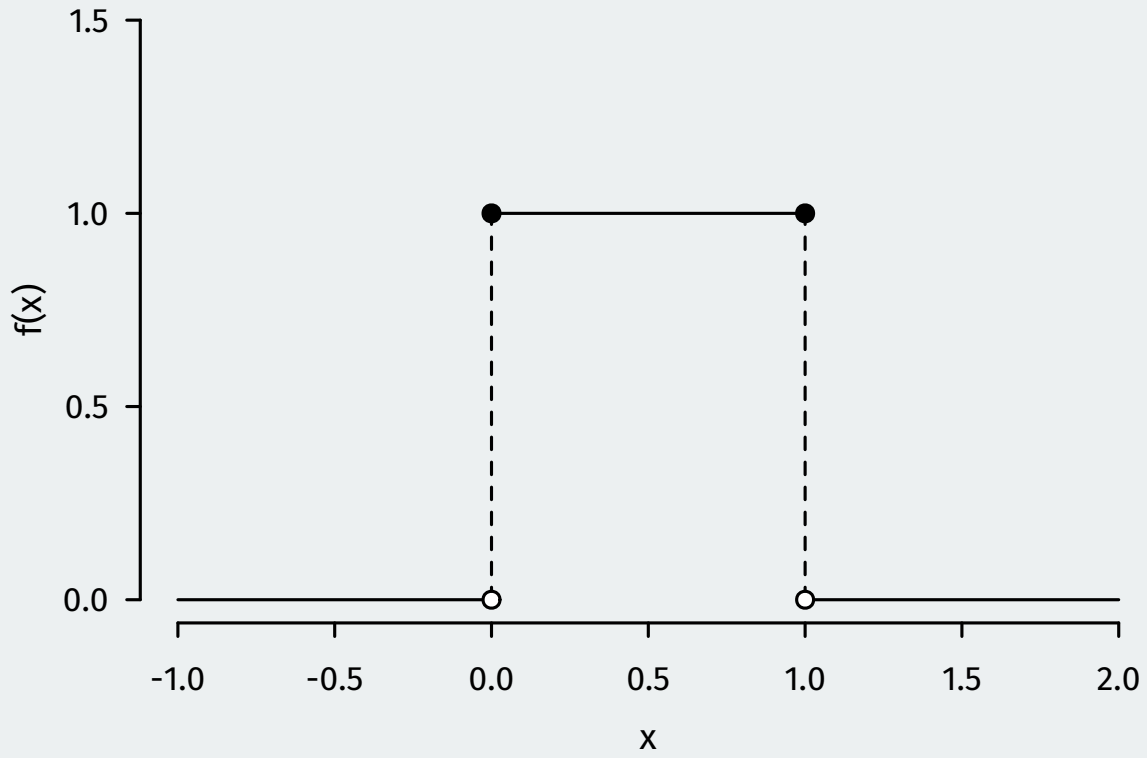
```
## [1] 0.5476 0.3848 0.0676
```

Binomial PMF plot



Probability density functions

- For continuous r.v.s, **probability density function (PDF)**
  - Gives density of probability around a given point.
  - Like a "infinite" histogram $\rightsquigarrow$ so many bins that things look smooth.
  - Area under the curve = prob. of some interval.



Cumulative distribution functions

- **Cumulative distribution function (CDF)**: $F_X(k) = \mathbb{P}(X \leq k)$
  - Returns the probability of X being at k or lower.
  - Area under the density for a continuous r.v.
  - Never decreasing as k gets bigger.
  - Drawing two women senators example: