Gov 50: 16. Random Variables

Matthew Blackwell

Harvard University

Fall 2018

1. Today's agenda

2. Why probability?

3. Random variables and probabilities distributions

4. Summarizing distributions

5. Famous distributions

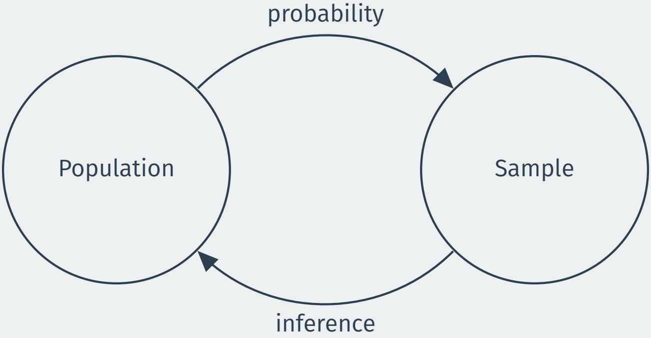
1/ Today's agenda

Where are we? Where are going?

- Learned the basics of probability.
  - Addition rule
  - Conditional probability
  - Independence
- Now, random variables and probability distributions.

2/ Why probability?

Learning about populations



- **Probability**: formalize the uncertainty about how our data came to be.
- **Inference**: learning about the population from a set of data.

Why probability?

- Statistical inference is a **thought experiment**.
- Probability is the logic of these though experiments.
- Suppose men and women were paid the same on average, but there was chance variation from person to person.
  - How likely is the observed wage gap in this hypothetical world?
  - What kinds of wage gaps would we expect to observe in this hypothetical world?
- Probability to the rescue!

The lady tasting tea

- **Thought experiment** posed by statistician R.A. Fisher.
  - “a genius who almost single-handedly created the foundations for modern statistical science” (also a racist/eugenicist)
- Setup of thought experiment:

  Your friend asks you to grab a tea with milk for her before meeting up and she says that she prefers tea poured before the milk. You stop by Tealuxe and ask for a tea with milk. When you bring it to her, he complains that it was prepared milk-first.

- You are skeptical that she can really tell the difference, so you devise a test:
  - Prepare 8 cups of tea, 4 milk-first, 4 tea-first
  - Present cups to friend in a **random** order
  - Ask friend to pick which 4 of the 8 were milk-first.

Assuming we know the truth

- Friend picks out all 4 milk-first cups correctly!
- Statistical thought experiment: how often would she get all 4 correct **if she were guessing randomly**?
  - Only one way to choose all 4 correct cups.
  - But 70 ways of choosing 4 cups among 8.
  - Choosing at random $\approx$ picking each of these 70 with equal probability.
- Chances of guessing all 4 correct is $\frac{1}{70} \approx 0.014$ or 1.4%.
- $\rightsquigarrow$ the guessing hypothesis might be implausible.

- You've done your first hypothesis test and calculated your first p-value!

3/ Random variables and probabilities distributions

What are random variables?

- Probability so far is about "events" and "outcomes"
- What's the connection to our data?

Random Variable

A random variable (r.v.) assigns a numeric value to each outcome in the sample space.

- r.v.s are numeric representation of uncertain events $\rightsquigarrow$ we can use math!
- We'll think about each observation in our data frame as a r.v.

Examples

- Random trial: Tossing a coin 3 times
  - one possible outcome: HTH
  - but not a random variable because it's not numeric.
- Random variable: X = number of heads in the five tosses
  - $HTH \rightsquigarrow X = 2$
- Same space might have many different r.v.s
  - Y = number of tails
  - $Z = 1$ if any of the 3 flips are heads.

Types of random variables

- **Discrete r.v.**: X can take on a finite (or countably infinite) number of values.
  - Number of heads in 5 coin flips
  - Trump approval or not.
  - Number of battle deaths in a civil war
- **Continuous r.v.**: X can take on 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.

Randomness and probability distributions

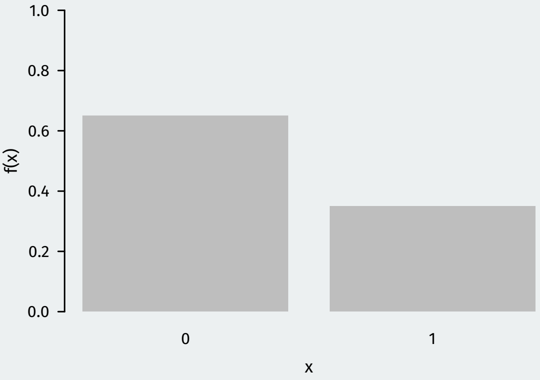
- How are r.v.s **random**?
  - Uncertainty over events/outcomes $\rightsquigarrow$ uncertainty over value of X.
  - We'll use probability to formalize this uncertainty.
- Easiest way to think about the randomness and distributions: sampling.
- Randomly select 1 person from US registered voters.
- Let $X = 1$ if the person supports Trump, $X = 0$ otherwise.
  - $\mathbb{P}(X = 1) =$ the share of people that support Trump in the population.
  - $\mathbb{P}(X = 0) =$ the share of people that don't support Trump in the population.
- Let Y be the age of the respondent.
  - $\mathbb{P}(Y > 65)$ is the share of registered voters over 65.

Probability distribution

- The **probability distribution** of a r.v. gives the probability of all of the possible values of the r.v.
- **Cumulative distribution function**: $F(x) = \mathbb{P}(X \leq x)$
    - Can recover probability of any interval.
    - $\mathbb{P}(X > x) = 1 - F(x)$
    - $\mathbb{P}(a < X \leq b) = F(b) - F(a)$

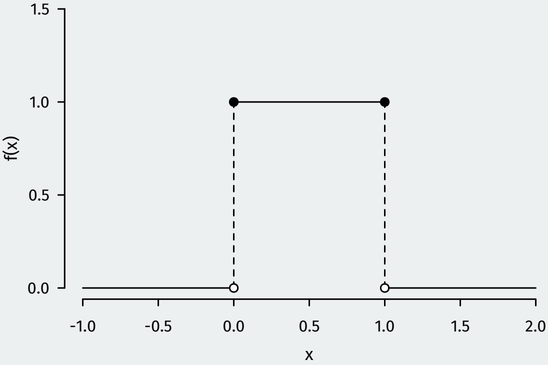
Probability mass functions

- For discrete r.v.s, **probability mass function** gives probability of each possible value, $f(x) = \mathbb{P}(X = x)$.
  - Like a bar plot for the population shares of each value.

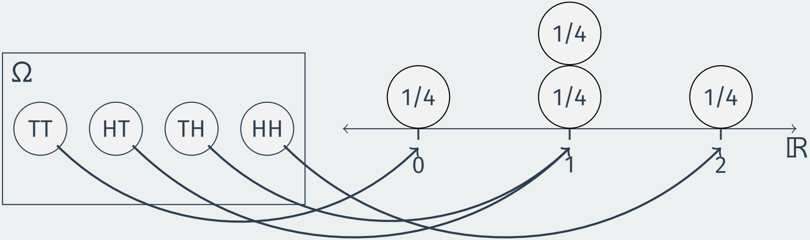


Probability density functions

- For continuous r.v.s, **probability density function** gives density of probability around a given point.
  - Like a “infinite” histogram $\rightsquigarrow$ so many bins that things look smooth.



Inducing probabilities



- Let X be the number of heads in two coin flips.

outcome	prob.	X
TT	1/4	0
HT	1/4	1
TH	1/4	1
HH	1/4	2

x	$\mathbb{P}(X = x)$
0	1/4
1	1/2
2	1/4

4/ Summarizing distributions

How can we summarize distributions?

- Probability distributions describe the uncertainty about r.v.s.
  - Problem: can involve complex formulas that are hard to work with.
- In this class, we'll focus on two summaries of the probability distribution.

1. **Central tendency**: where the center of the distribution is.
   - We'll focus on the mean/expectation.
2. **Spread**: how spread out the distribution is around the center.
   - We'll focus on the variance/standard deviation.

- With real data, we are going to try and infer these values from data on a r.v.

Expectation

- Natural measure of central tendency is the **expected value** (a/k/a the **expectation** or **mean**) of X.
- If X is age of randomly selected registered voter, then mean of X is the average age in the population of registered voters.
- Write it as $\mathbb{E}(X)$ or sometimes just μ (mu).
- For discrete $X \in \{x_1, x_2, \dots, x_k\}$ with k levels:

$$\mathbb{E}[X] = \sum_{j=1}^{k} x_j \mathbb{P}(X = x_j)$$

  - Weighted average of the values of the r.v. weighted by the probability of each value occurring.

Properties of the expected value

Let X and Y be r.v.s and a and b be constants.

1. $\mathbb{E}(a) = a$
2. $\mathbb{E}(aX) = a\mathbb{E}(X)$
3. $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$
4. $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$

Variance

- The **variance** measures the spread of the distribution:

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

- If X is the age of a randomly selected registered voter, $\mathbb{V}[X]$ is the variance of age in the population.
- Weighted average of the squared distances from the mean.
    - Larger deviations ($+$ or $-$) $\rightsquigarrow$ higher variance
- The **standard deviation** is the (positive) square root of the variance: $\sigma_X = \sqrt{\mathbb{V}[X]}$.

Properties of variances

1. If b is a constant, then $\mathbb{V}[b] = 0$.
2. If a and b are constants, $\mathbb{V}[aX + b] = a^2\mathbb{V}[X]$.
3. In general, $\mathbb{V}[X + Y] \neq \mathbb{V}[X] + \mathbb{V}[Y]$.

5/ Famous distributions

Probability distributions

- Like last slide, we can infer probability distributions from underlying probability trials.
- Easier: rely on common distributions that are well-studied.
  - Common distributions have underlying probability trials that we often just say in words.
- Three types of r.v.s we'll think about in this class:
  - Bernoulli, binomial, and normal.
  - Others in the book, but we won't focus on them.

Bernoulli r.v.

- **Bernoulli r.v.**: X can take on one of two possible values (usually 0 and 1).
  - a/k/a binary r.v. or dummy r.v.
  - Discrete random variable.
- Example: Trump approval for a respondent:
  - $\Omega = \{\text{approve, don't approve}\}$.
  - Random variable converts this into a number:

$$X = \begin{cases} 1 \text{ if approve} \\ 0 \text{ if don't approve} \end{cases}$$

- Probability distribution of Bernoulli r.v. summarized by the probability of $X = 1$.
  - Why? $\mathbb{P}(X = 0) = 1 - \mathbb{P}(X = 1)$
  - We use $p = \mathbb{P}(X = 1)$ be the probability of "success" ($X = 1$).
  - Infinite number of possible Bernoulli r.v.s: one for each value of p.

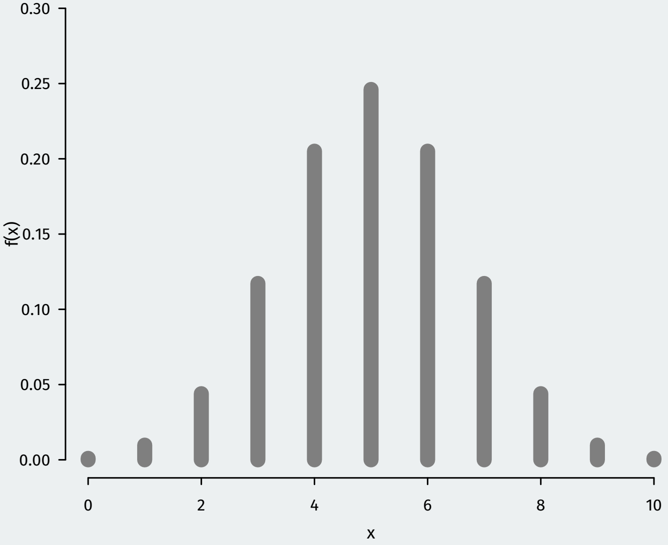
Binomial distribution

- **Binomial r.v.**: X takes on any integer between 0 and n.
  - Number of heads in n independent coin flips with probability p of heads.
  - "Binomial with n trials and probability of success p"
- Example:
  - Randomly select 10 people from the population, $X =$ how many of them support Trump?
  - If the population support for Trump is p, then X is binomial with $n = 10$ trial and probability of success p.
- Probability mass function: x

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad \text{where} \quad \binom{n}{k} = n!\,/(k!\,(n-k)!\,)$$

- Equivalent to the sum of n Bernoulli r.v.s each with probability p.
- $\rightsquigarrow \mathbb{E}[X] = np$ and $\mathbb{V}[X] = np(1 - p)$

Binomial distribution (n=10,p=0.5)



Binomials in R

- Binomial pmf $\mathbb{P}(X = x)$ in R (`size` = n and `prob` = p):

```
dbinom(5, size = 10, prob = 0.5)
```

```
## [1] 0.246
```

- Binomial cdf $\mathbb{P}(X \leq x)$ in R:

```
pbinom(5, size = 10, prob = 0.5)
```

```
## [1] 0.623
```

- We can **simulate** data from this distribution using `rbinom()`:

```
rbinom(n = 10, size = 10, prob = 0.5)
```

```
##  [1]  5  3  4 10  5  5  2  6  3  6
```

Example

- Suppose we knew (magically) that Donald Trump had a population approval rating of 42%.
  - Equivalent, 0.42 of the population approves of Trump.
- Draw a random sample of 1000 and $X =$ number of respondents that support Trump.
  - X is Binomial with size 1000 and probability of success 0.42
- What if drew lots of samples of size 1000? What would the distribution look like?
  - $\rightsquigarrow$ what if drew a lot of samples of X?

Simulations

```r
sims <- 10000

draws <- rbinom(sims, size = 1000, prob = 0.42)
length(draws)
```

```
## [1] 10000
```

```r
mean(draws)
```

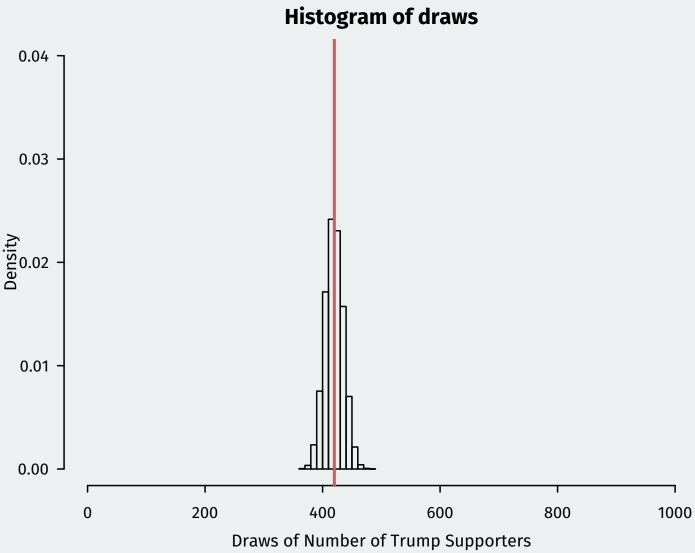
```
## [1] 420
```

```r
## convert to sample proportions
head(draws/1000)
```

```
## [1] 0.465 0.439 0.454 0.451 0.404 0.413
```

```r
hist(draws, freq = FALSE, xlim = c(0, 1000), ylim = c(0, 0.04),
     xlab = "Draws of Number of Trump Supporters")
abline(v = 420, col = "indianred", lwd = 2)
```

Histogram of draws



Next time

- Properties of sums and means in large samples.
- Normal distribution and the central limit theorem!