Probability II

Day 7 AM

Sarah Moore & J. Seawright

Northwestern University

Math Camp 2023

Random Variables

The concept of random variables is one of the most important and fundamental aspects to modeling quantitative social science.

A random variable is any concept whose outcome can be modeled by some *rule* or *function*. The potential range of values of a given random variable and the values' associated probability structure is the probability distribution.

Probability Distributions, in general

Probability distributions can be modeled as equations. The type of distribution function associated with a given probability structure will depend on whether a random variable is discrete or continuous. More on this to come.

What are some examples of random variables that we might encounter in the social sciences?

Probability Distributions and Additional Terminology

Probability distributions and their associated random variables have a range of possible values where $Pr > 0$, this range of possible values is known as the distributional *support*.

For example, in a Bernoulli distribution, often used for measuring the likelihood of success in a single trial, the support is $\{0, 1\}$ where 0 is failure and 1 is success.

Discrete Random Variables

Random variables (r.v.) take on different functional forms given the types of quantities or values that they represent.

Discrete r.v. model countable, distinct values.

Examples of discrete r.v. include the number of children in a classroom, the number of crosswalks in a neighborhood, or the number of degrees that a person holds.

https://tinyurl.com/2p94ax59

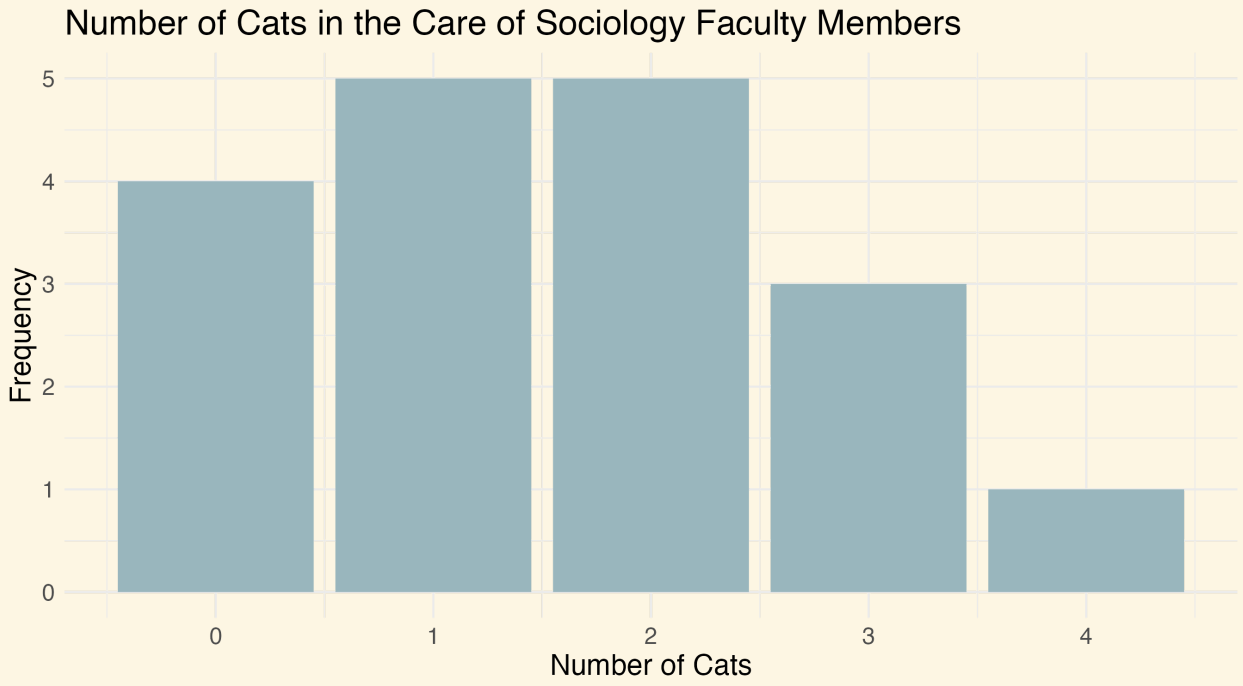
Discrete Random Variables

The associated probability distribution of a discrete r.v. is the probability mass function (p.m.f.).

Discrete r.v. are graphically shown via histograms, which are bar chart representations of frequencies over a range of values.

Histograms

Let's say we surveyed faculty members of the Sociology Dept. for the number of cats each person has in their household. The following graph shows the frequency of each value given by the surveyed individuals.



Frequency Tables

We can also display this in a table:

Table: Number of Cats Across Faculty Members in the Sociology Department

Number of Cats	Frequency
0	4
1	5
2	5
3	3
4	1

Given what we know about the distribution of values, what is the probability that any faculty member cares for 1 cat? *Assume that these responses account for all the faculty in the department.*

Probability Mass Functions (p.m.f.)

A probability mass function (p.m.f.) is the way to model the probability associated with each potential value k of a discrete r.v. X.

Remember that the probabilities are bounded as $[0, 1]$ and that all the total probability of all values in a sample space Ω sum to 1.

Let's think this through given the motivating example.

p.m.f. Example

We know a few things right off the bat from the histogram and the table.

First, the range of possible values, i.e. the sample space, is 0 to 4.

Second, we know that our n, or the total observations, is 18.

From here it is pretty easy to construct a probability mass function. We merely calculate the probability for each observed value.

p.m.f. Example

$$p(x) = \begin{cases} \frac{4}{18} & \text{if } k = 0 \\ \frac{5}{18} & \text{if } k = 1, 2 \\ \frac{3}{18} & \text{if } k = 3 \\ \frac{1}{18} & \text{if } k = 4 \\ 0 & \text{otherwise} \end{cases}$$

What is the probability that any faculty member cares for 1 cat?

Notice that we can include the probabilities of 1 and 2 together on the same line because they are equal, this does not mean that is the sum of those probabilities.

Continuous Random Variables

Continuous r.v. can take on an uncountable, infinite range of values.

Examples of continuous r.v. include annual rainfall in Chicago, gross domestic product, or the amount of oil imported to a country on a monthly basis.

The functional form of a continuous r.v. is given by the probability density function (p.d.f.).

https://tinyurl.com/2p94ax

Continuous Random Variables

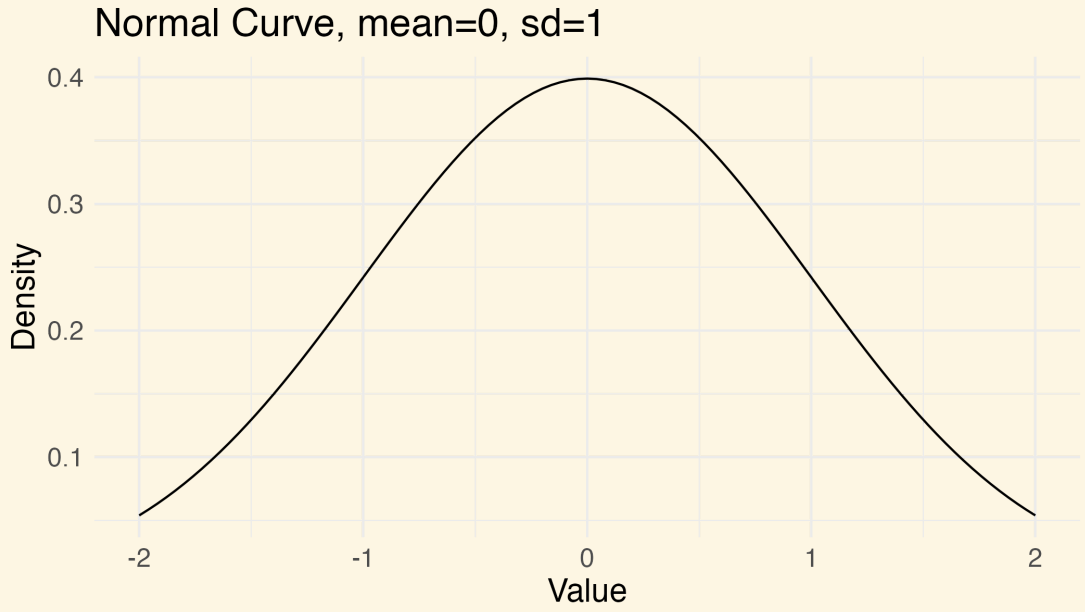
Given that a continuous r.v. has an infinite range of possible values, it is not possible to calculate the probability at any specific value. Instead, we can obtain the probability for a range of values given a density curve.

A density curve is a smoothed line that displays a density function.

Density Curve

For example, the Bell Curve (formally the Standard Normal distribution) is one of the most commonly known density curves.

Let's say you have some data that you have standardized with mean 0 and standard deviation 1 (we will get to estimating central tendencies soon!), this is how the density curve would look.



Probability Density Function (p.d.f.)

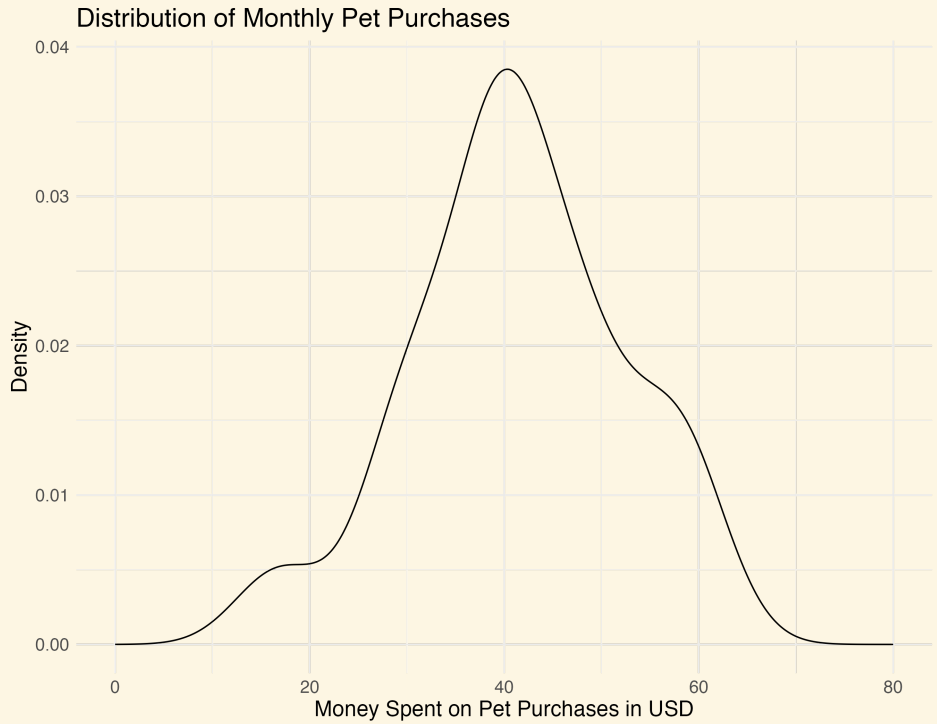
Given that any given value of a continuous r.v. X has a zero probability, we think differently about what a probability function models for a continuous r.v.

In a probability density function (p.d.f.) of a continuous r.v. we define the probability over a range of potential values.

https://online.stat.psu.edu/stat414/lesson/14/14.1

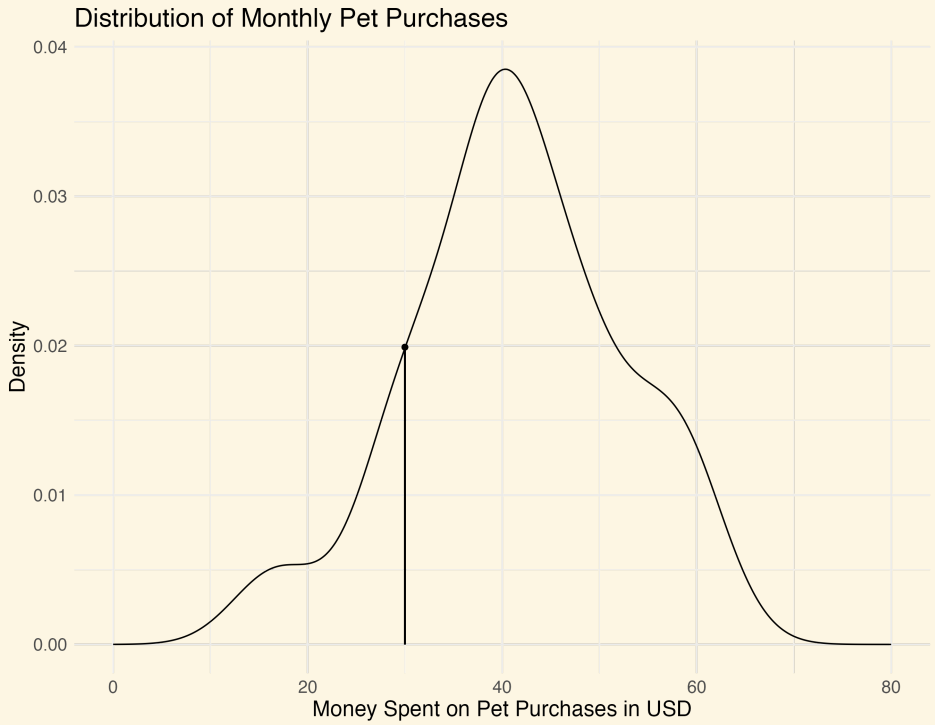
Money Spent on Pets Every Month

Let's say we can represent the amount that people spend on their cats with some probability function f(x).

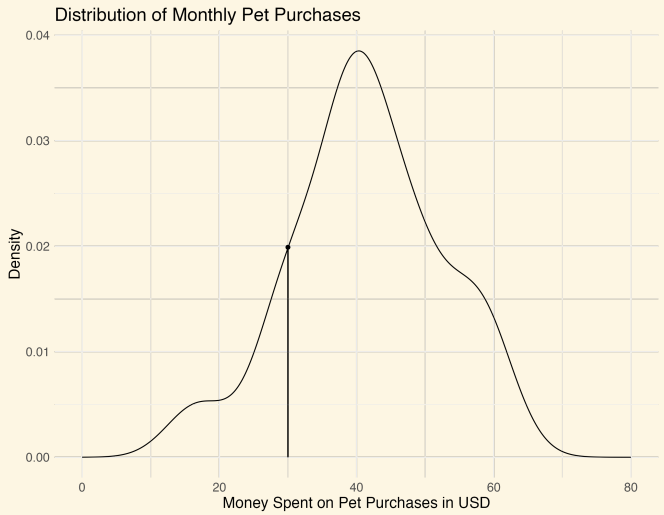


Money Spent on Pets Every Month

What's the probability that someone spends 30 dollars monthly on pet purchases?



Let's think about what that implies for finding the value "under the curve" at a given point in a function.

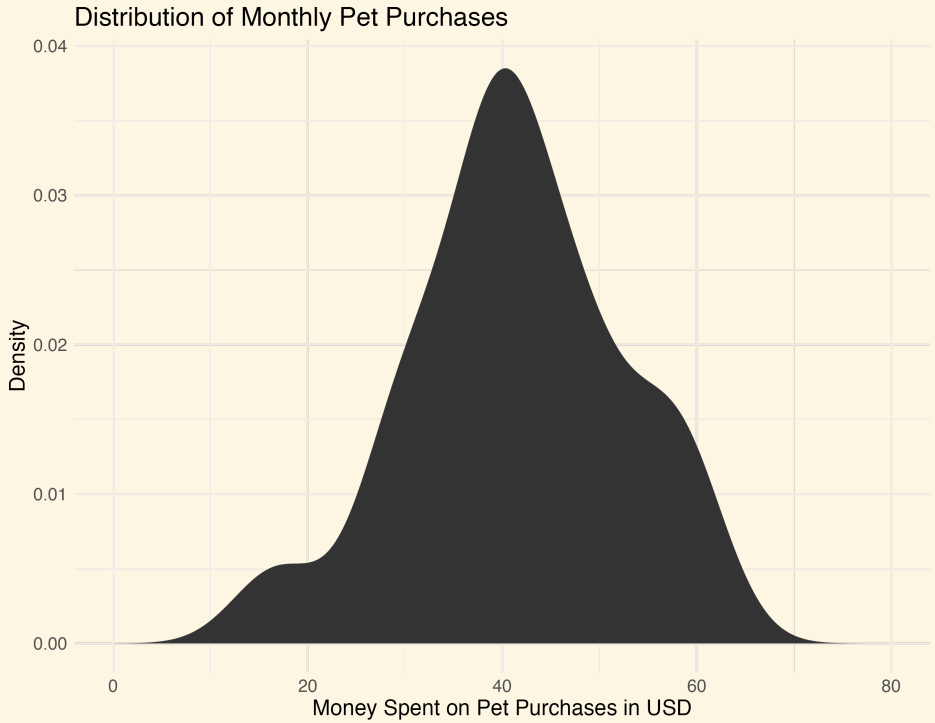


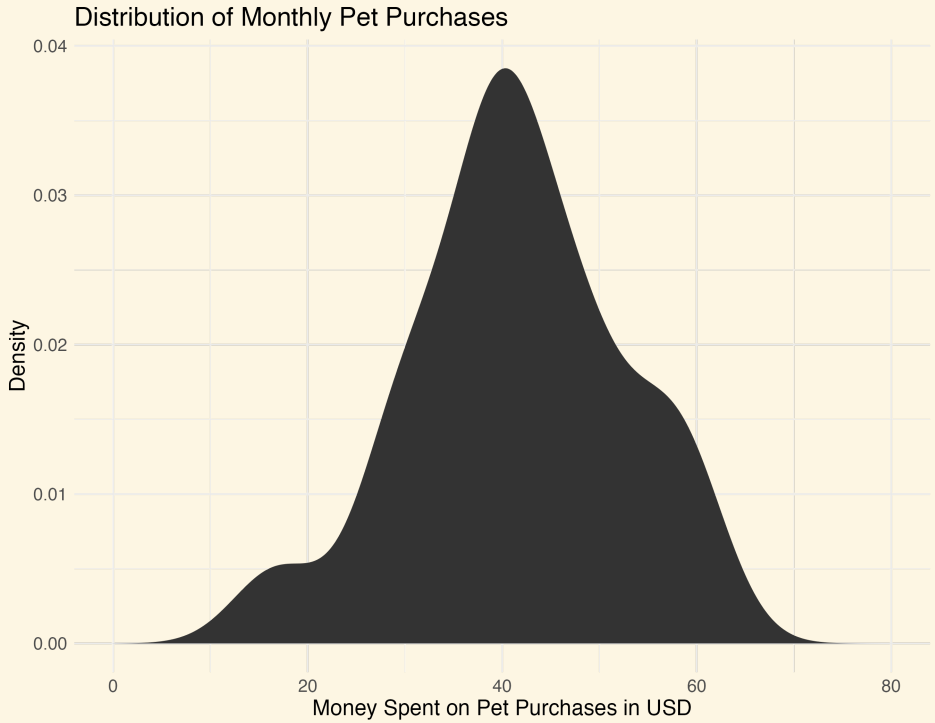
$$\int_{30}^{30} f'(x)dx$$

$$= f(x)|_{30}^{30} = f(x = 30) - f(x = 30)$$

p.d.f. Example, cont'd.

Remember that the law of total probability states that the probability over the entire sample space is equal to 1. This means that the entire area under our probability curve is 1.




Distribution of Monthly Pet Purchases
Density
0.04
0.03
0.02
0.01
0.00
0
20
40
60
80
Money Spent on Pet Purchases in USD

p.d.f.

Here is the takeaway: We can model probability via functions. Calculating probability at given values then becomes an issue of finding a function's location and behavior at those values.

Discussion

Why do we need to know the difference between discrete r.v. and continuous r.v.?

A lot of how social science models things is dependent on the types of variables modeled and how we measure them.

For example, how we model the probability of a vote choice between two candidates (a discrete variable) is different than how we would model the difference in vote share between two candidates (a continuous variable).

Understanding the underlying probability structure of different variable types will be an important foundation for developing more advanced modelling skills down the road.

Examples of Some Important Distributions in Social Science

Bernoulli

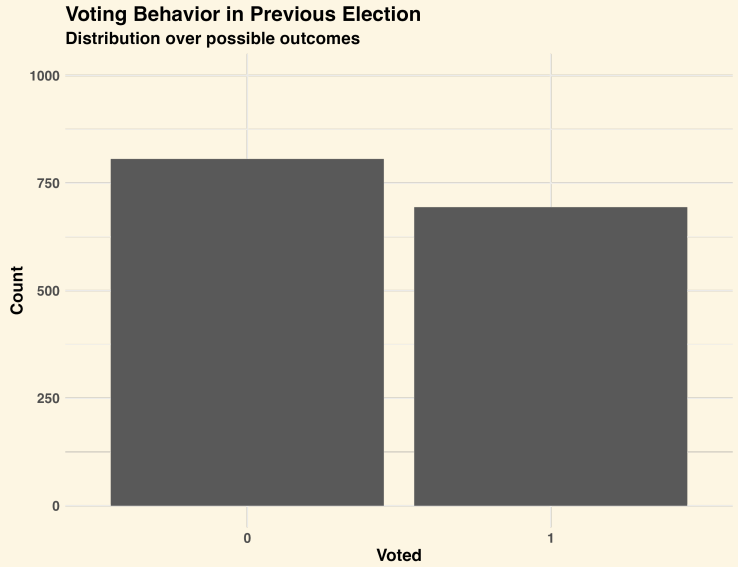
- Discrete distribution, therefore we use the probability mass function.
- Probability of failure (0) or success (1) for a single trial. Think of a dummy variable or simple indicator variable.
- Generalizes to the Binomial distribution, the number of k successes over n trials.

Bernoulli Binomial Example

Event that a person voted in a previous election.

```
set.seed(2)
voters <- tibble(voted =
                   rbinom(1500, 1
                          prob = (
head(voters)
```

```
## # A tibble: 6 × 1
##   voted
##   <int>
## 1     0
## 2     1
## 3     1
## 4     0
## 5     1
## 6     1
```



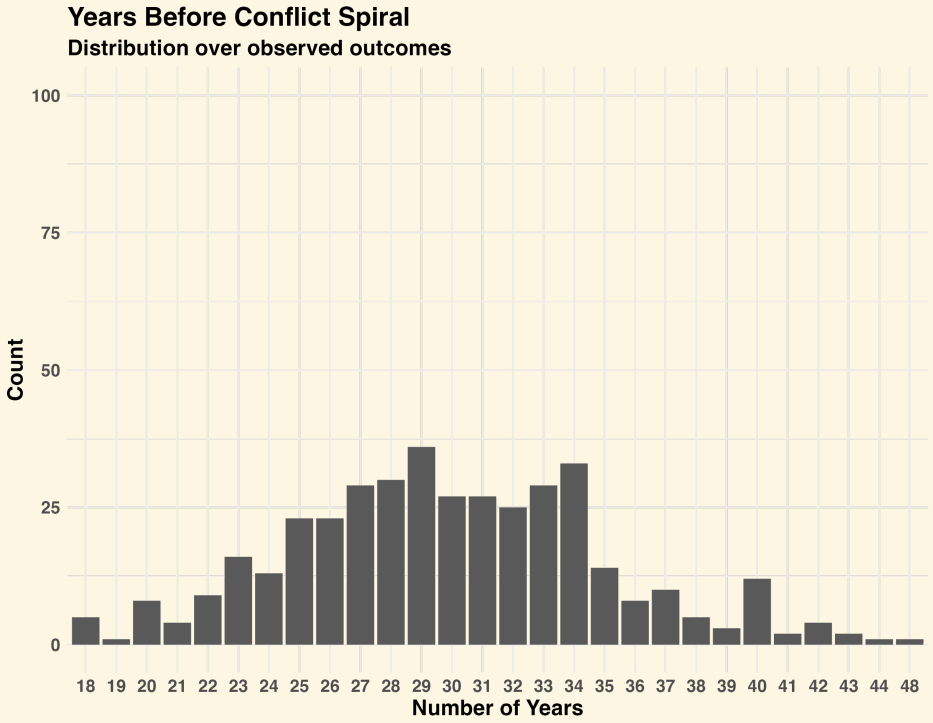
Poisson Distribution

- Discrete distribution often used for counts.
- Probability of an event happening within a given interval.
- Has support over $(0, \infty)$ for all integers (can only be whole numbers).

```
set.seed(2)
peace <- tibbl

head(peace)
```

```
## # A tibble: 6 × 1
##   years
##   <int>
## 1    25
## 2    24
## 3    23
## 4    30
## 5    26
## 6    28
```

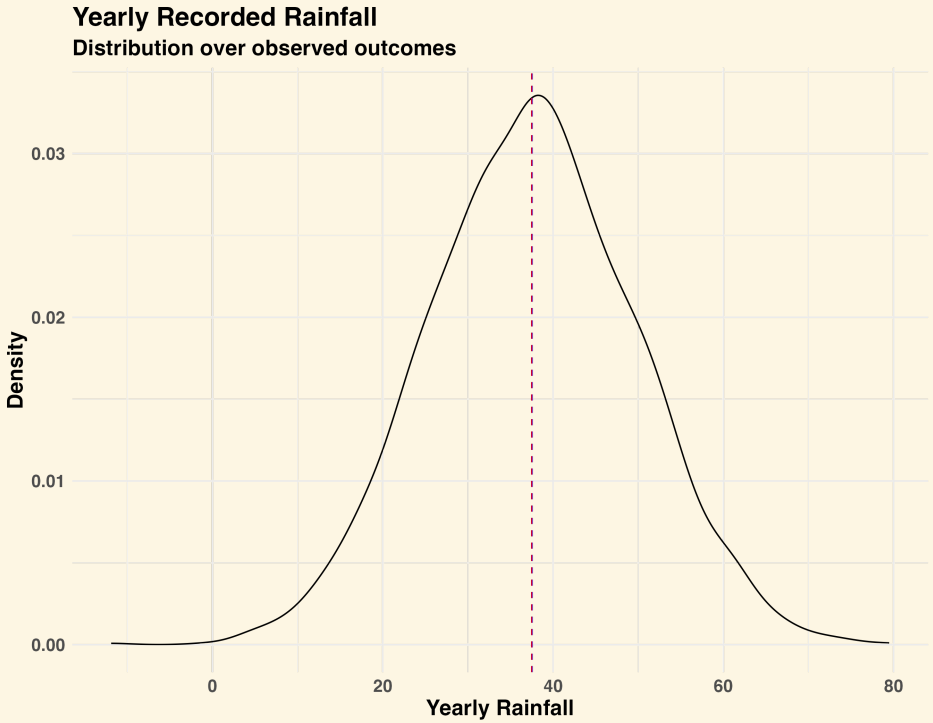


Normal Distribution

- Continuous distribution
- Mean and median are equal
- Has support over $(-\infty, \infty)$.

```r
set.seed(2)
rain_chicago <
head(rain_chic
```

```
## # A tibble: 6 × 1
##    rain
##   <dbl>
## 1  26.2
## 2  39.2
## 3  56.1
## 4  23.4
## 5  36.0
## 6  38.6
```



Continuing Examples of other RVs

1) Discrete

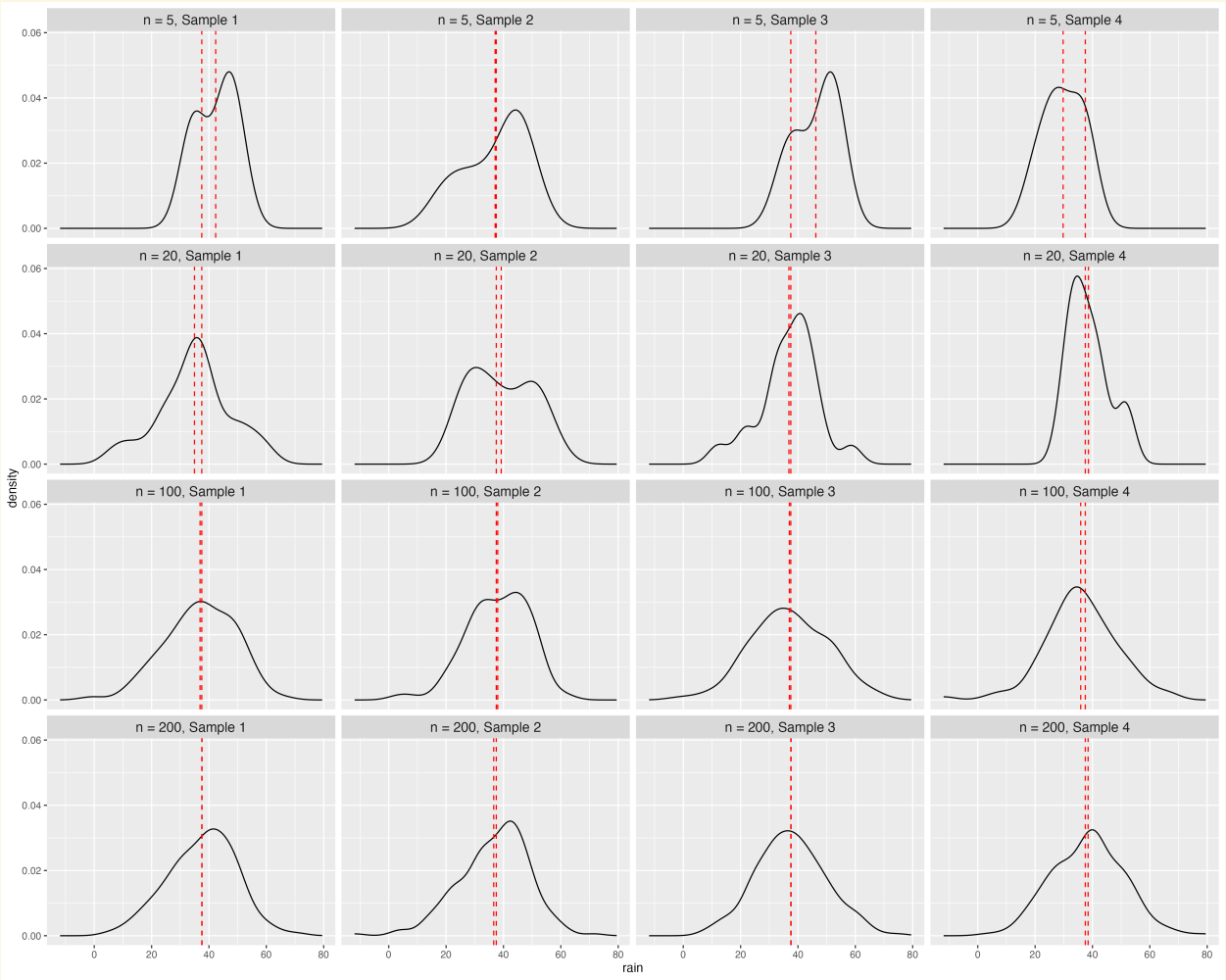
2) Continuous

Summary Quantities

- General summary statistics, e.g., mean, median, come from general properties of probability distributions.

Probability $\rightarrow$ Statistics

Law of Large Numbers



Central Limit Theorem