Biostats Lecture 3: Introduction to Probability, Random Variables, and Distributions.

Public Health 783

Ralph Trane
University of Wisconsin–Madison

Fall 2019

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Learning Objectives

1. Familiarize ourselves with concepts within probability theory
2. Introduce Random Variables, and hint at why this concept will be useful
3. Introduce distributions, and take a look at certain properties of certain distributions

Definitions

We talk about probabilities in regards to outcomes of an experiment.

Two definitions. We will mainly be working with the second, but the first is included for completeness.

1. If all outcomes are equally likely:

$$P(\text{event}) = \frac{\text{number of outcomes that result in event}}{\text{total number of possible outcomes}}$$

2. In general: the long run proportion of times the event occurs if the experiment is repeated an *infinite number of times*

Example: roll a die

Let $A = \{\text{roll is a 4}\}$, $B = \{\text{roll is a 1 or 5}\}$, and $C = \{\text{roll is even}\}$.

What is

- $P(A)$, $P(B)$, and $P(C)$?
- $P(A \text{ OR } B)$, and $P(A \text{ OR } C)$?
- $P(B \text{ and } C)$?

Example: Framingham Heart Study

What is the probability of developing Coronary Heart Disease over ten years?

You either develop the disease, or you do not, so.... $\frac{1}{2}$?

Of course not! Why doesn't definition 1 work here? Outcomes not equally likely!

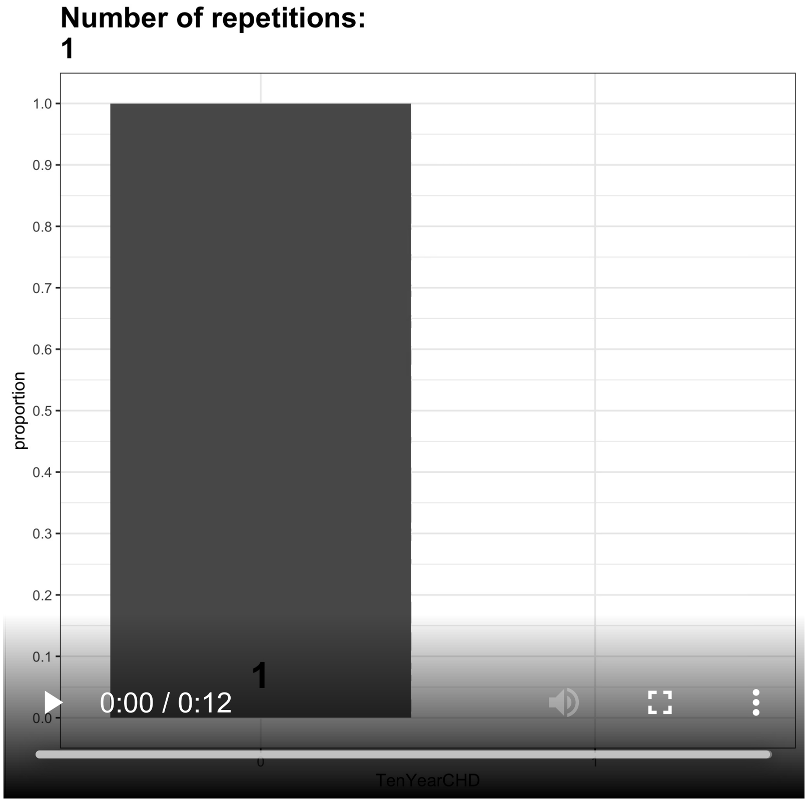
So instead, think about what would happen if we "repeat the experiment an infinite number of times"...

Before we can do that, specify what "the experiment" is.

Experiment = a randomly chosen individual develops CHD over a ten year period.

Example: Framingham Heart Study

Can't really do that... BUT we can pretend the FHS sample IS the population, and randomly choose individuals from this "population".



Example: Framingham Heart Study

So from the above, we would estimate $P(\text{develops CHD over ten years}) = 0.15.$

In this case, we can actually get the exact probability (because we "know" the entire population):

TenYearCHD	n	percent
0	3596	0.8481132
1	644	0.1518868
Total	4240	1.0000000

In general, we would estimate the probability in this way: the proportion of the sample with the attribute of interest.

Example: Framingham Heart Study

Moral of the story: probabilities dictate the results of sampling (when done right).

I.e. using probability theory, we can find out what to expect from sampling.

This allows us to judge if differences in samples are "as expected", or "out of the ordinary".

Random Variables

Why "Random Variables"?

We introduce *random variables* to

- formalize the notion of an experiment
- simplify notation
- have a rigorous way of discussing probabilities

What is a "Random Variable"?

A *random variable* is a variable tied to the outcome of an experiment.

The value of it is unknown and uncertain before the experiment is conducted.

Conducting the experiment results in a *realization* of the RV.

Distinguish between discrete and continuous RVs.

Examples:

1. $X =$ flip of a coin. Possible outcomes: heads and tails. Discrete RV.
2. $X =$ sex of randomly chosen individual. Possible outcomes: male and female. Discrete RV.
3. $X =$ educational level of randomly chosen individual from the FHS. Possible outcomes: 1, 2, 3, 4, NA. Discrete RV.
4. $X =$ height of randomly chosen 783 student. Possible outcomes: any number greater than 0. Continuous RV.

Random Variables

Talk about probabilities of different outcomes:

1. $X =$ flip of a coin. What is $P(X = \text{heads})$?
2. $X =$ sex of randomly chosen individual. What is $P(X = \text{male})$?
3. $X =$ educational level of randomly chosen individual from the FHS. What is $P(X \in \{1{,}3{,}4\})$? (X is either 1,3, or 4)
4. $X =$ height of randomly chosen 783 student. What is $P(X \geq 180\text{cm})$?

To calculate these probabilities, we specify the *distribution* of the random variable.

Distributions

The distribution specifies the probabilities of all possible outcomes. For discrete RVs, specify probability of every possible outcome.

Example: X follows

X = x	**P(X = x)**
1	0.2
3	0.5
7	0.1
8	0.2

$$P(X = 3) = ?$$

$$P(X \in \{1, 8\}) = ?$$

$$P(X = 9) = ?$$

$$P(X \text{ is even}) = ?$$

General properties of distributions:

- all probabilities are between 0 and 1
- the sum of all probabilities must be 1

Expected Value and Variance/Standard Deviation

Expected value of random variable: 'long run average'. I.e. if we observe the outcome of the random variable 'an infinite number of times', $E(X)$ is the average.

Variance: 'long run variance'

For discrete random variables:

$$E(X) = \sum_{i=1}^{n} P(X = x_i) \cdot x_i$$
$$\text{Var}(X) = \sum_{i=1}^{n} P(X = x_i) \cdot (x_i - E(X))^2$$

Expected Value and Variance/Standard Deviation

Example:

X follows

X = x	**P(X = x)**
1	0.2
3	0.5
7	0.1
8	0.2

$$\begin{aligned} E(X) &= 1 \cdot 0.2 + 3 \cdot 0.5 + \\ &\qquad 7 \cdot 0.1 + 8 \cdot 0.2 \\ &= 4 \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= 0.2 \cdot (1-4)^2 + 0.5 \cdot (3-4)^2 + \\ &\qquad 0.1 \cdot (7-4)^2 + 0.2 \cdot (8-4)^2 \\ &= 6.4 \end{aligned}$$

What is $E(X)$ and $\text{Var}(X)$?

How do we find $\text{SD}(X)$? $\text{SD}(X) = \sqrt{\text{Var}(X)} = \sqrt{6.4} \approx 2.53.$

Random Variables

Expected Value and Variance/Standard Deviation

Useful Properties:

X and Y are random variables, a is a constant (i.e. some fixed number).

$$\begin{aligned} E(a \cdot X) &= aE(X) \\ E(a) &= a \\ E(X+Y) &= E(X) + E(Y) \end{aligned}$$

$$\begin{aligned} \text{Var}(a \cdot X) &= a^2 \text{Var}(X) \\ \text{Var}(a) &= 0 \\ \text{Var}(X+Y) &= \text{Var}(X) + \text{Var}(Y) \\ &\quad (\text{IF } X \text{ and } Y \text{ are independent}) \end{aligned}$$

What is $\text{Var}(X-Y)$? $\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y)$. Don't forget this!

The Bernoulli Distribution

X follows a Bernoulli distribution if it only has two potential outcomes, success (often denoted 1) or failure (often denoted 0).

We write $X \sim \text{Bernoulli}(p)$ (read: X follows a Bernoulli distribution with probability of success p.)

By definition, $P(X = 1) = p$. So, $P(X = 0) = 1 - p$

- X_i is the sex of a subject i (male, female)
- X_i is the disease status of subject i (diseased, healthy)

If $X \sim \text{Bernoulli}(p)$, then

$$E(X) = p,$$

and

$$\text{Var}(X) = p \cdot (1 - p).$$

The Binomial Distribution

Y follows a Binomial distribution if it is the sum of n independent Bernoulli variables with same probability of success p.

In other words, the number of successful trials out of n Bernoulli trials.

In math: if $X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$ are independent, and $Y = X_1 + X_2 + \ldots + X_n$, then $Y \sim \text{Binomial}(n, p)$.

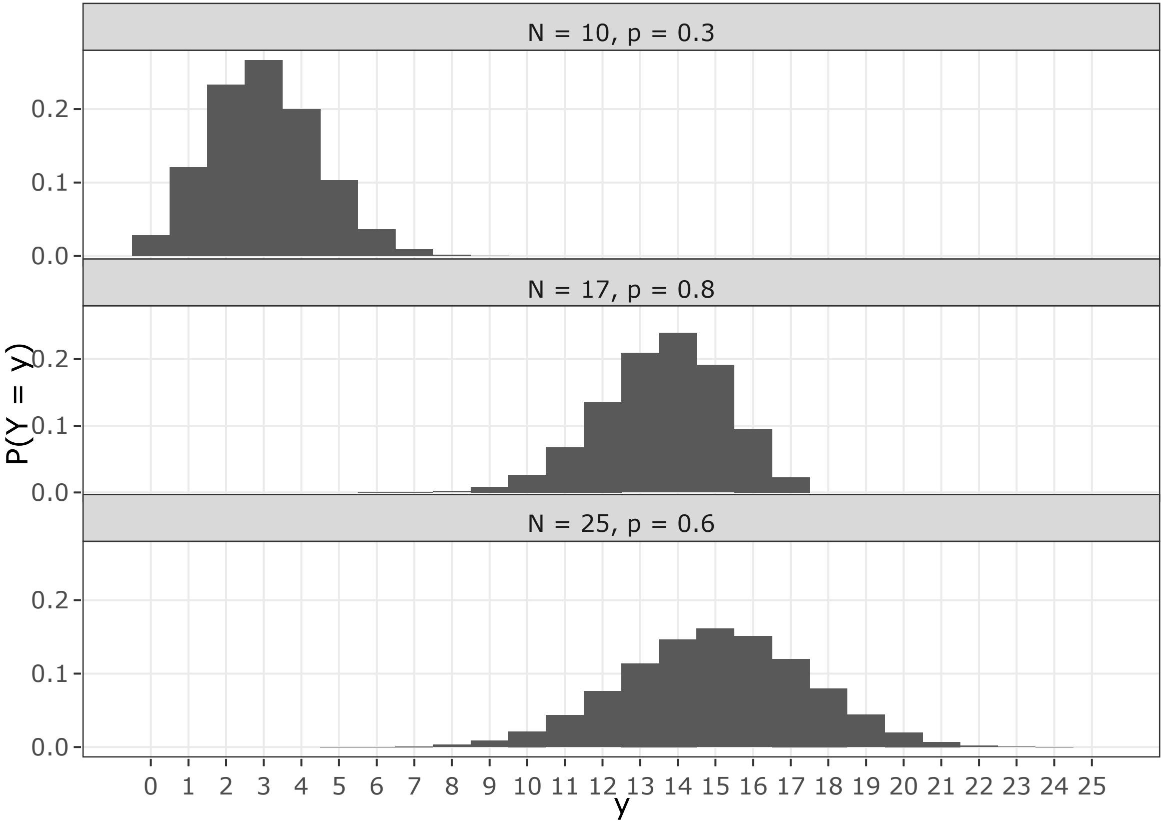
We call n the size parameter, p probability of success.

Possible values of Y? $0, 1, 2, \ldots, n$.

What is $E(Y)$? $\text{Var}(Y)$?

The Binomial Distribution

For a few different values of n and p, the Binomial distribution has the following forms:



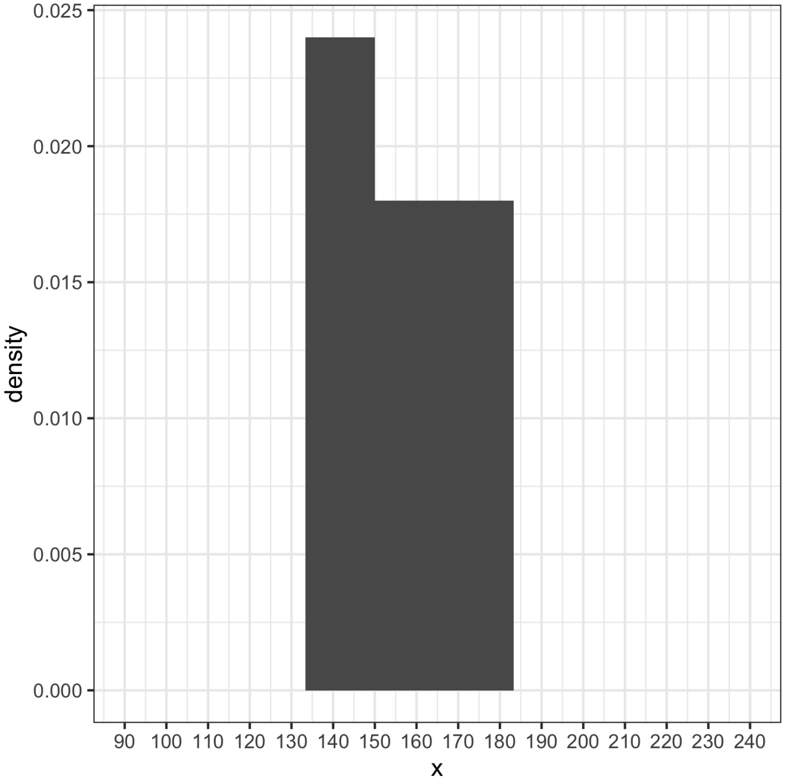
Continuous Random Variables & Distributions

For a continuous variable, can we specify the probability of every single possible outcome? No, because number of outcomes is uncountable!

Instead, specify a curve.

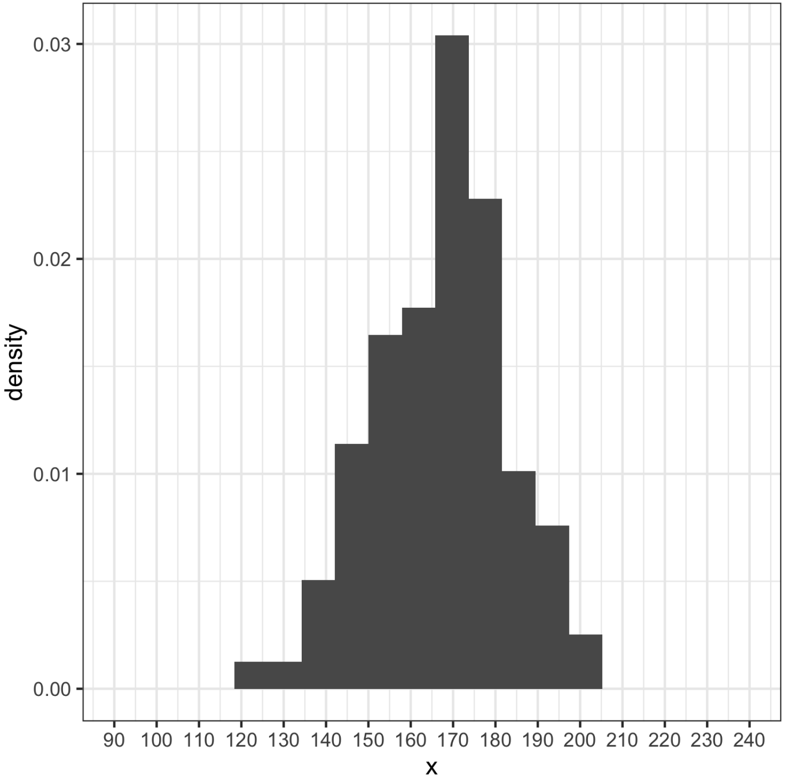
Continuous Random Variables & Distributions

Observe the height of 10 individuals, draw a histogram with 10 bins.



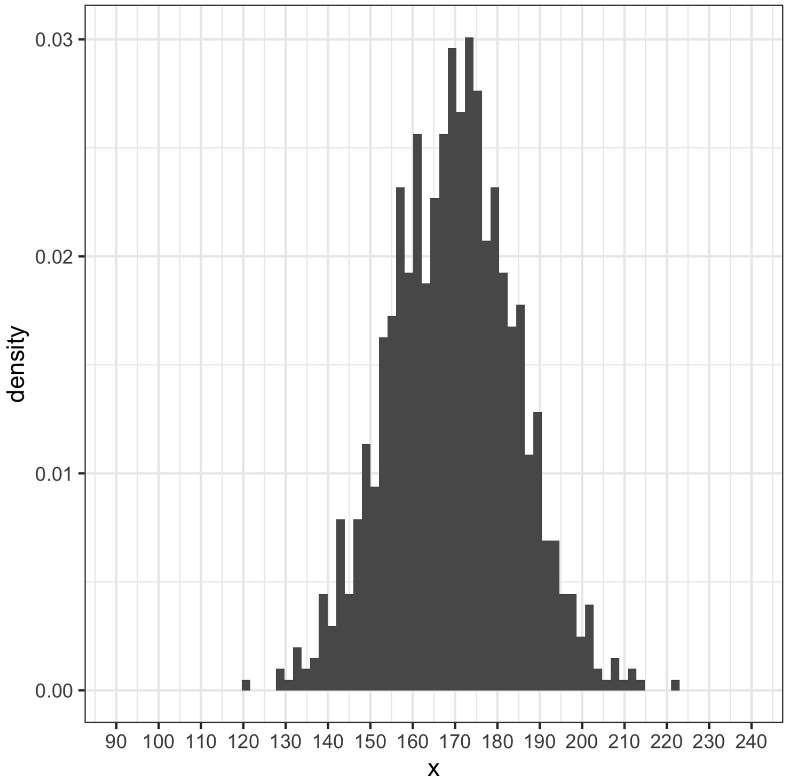
Continuous Random Variables & Distributions

Observe the height of 100 individuals, draw a histogram with 20 bins.



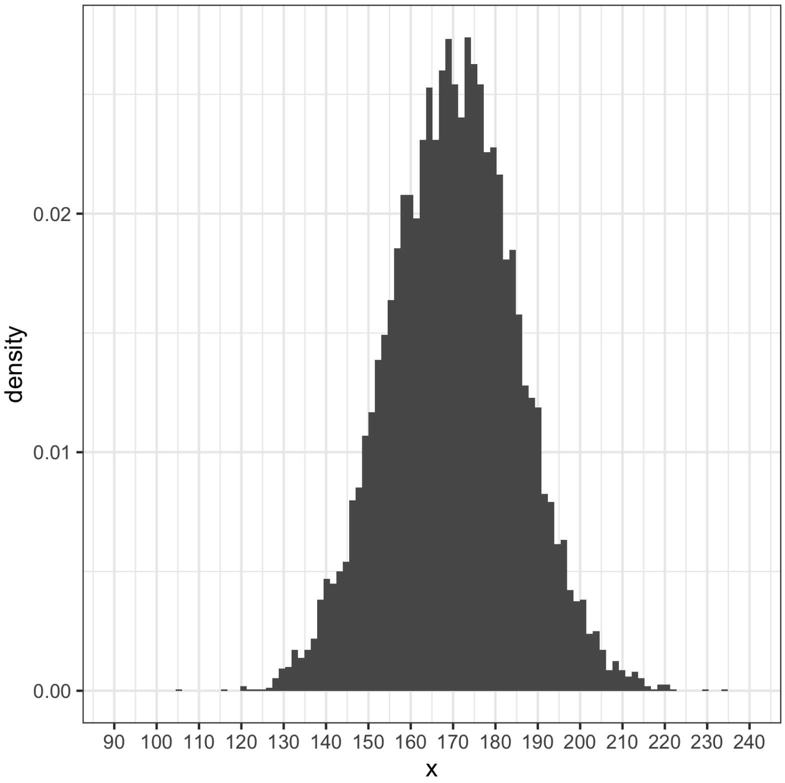
Continuous Random Variables & Distributions

Observe the height of 1000 individuals, draw a histogram with 75 bins.



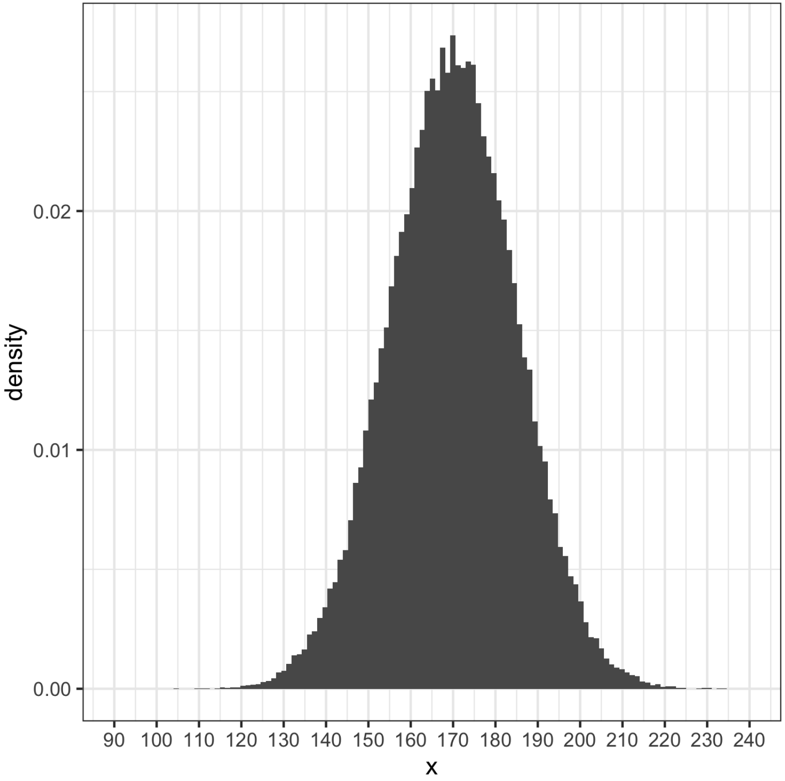
Continuous Random Variables & Distributions

Observe the height of 10000 individuals, draw a histogram with 100 bins



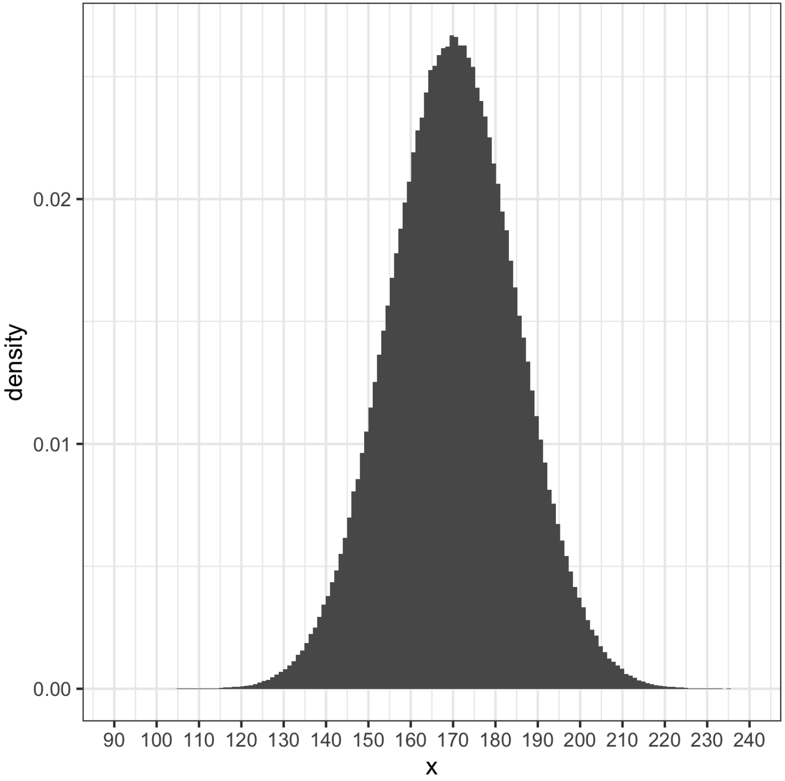
Continuous Random Variables & Distributions

Observe the height of 100000 individuals, draw a histogram with 125 bins.



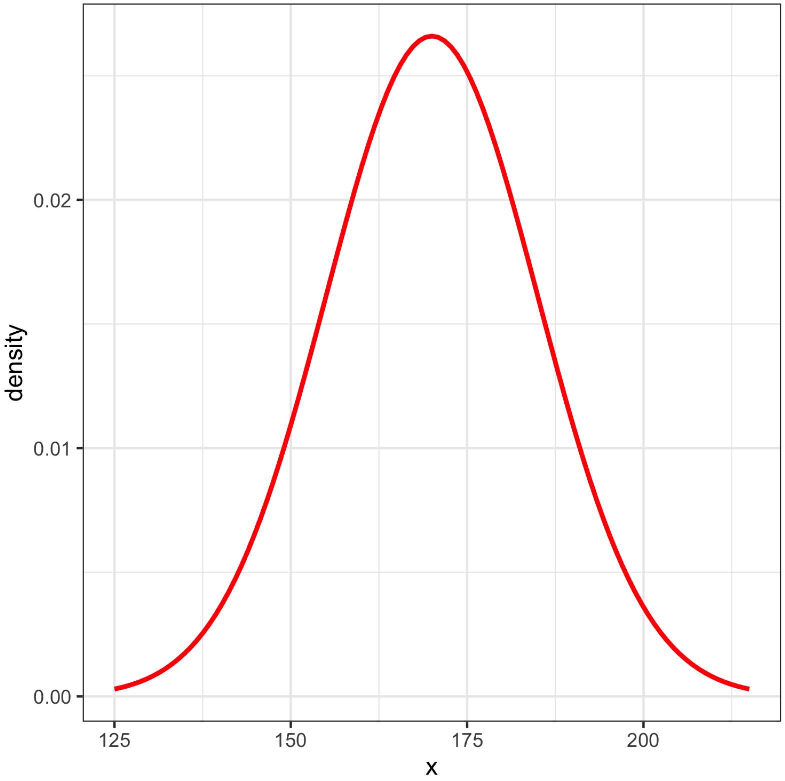
Continuous Random Variables & Distributions

Observe the height of 1000000 individuals, and 150 bins.



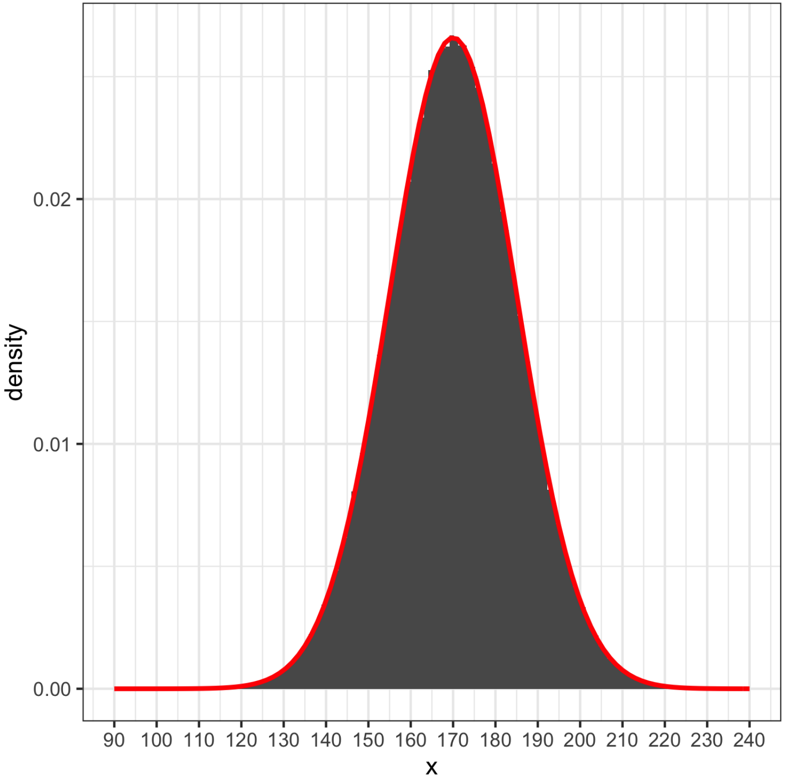
Continuous Random Variables & Distributions

The data here was simulated from a normal distribution with mean 170 and variance 225. This distribution looks like this:



Continuous Random Variables & Distributions

If we overlay it:



In words: the distribution of a continuous RV is the curve that appears when a histogram with narrow bars of many, many, many observations is drawn.

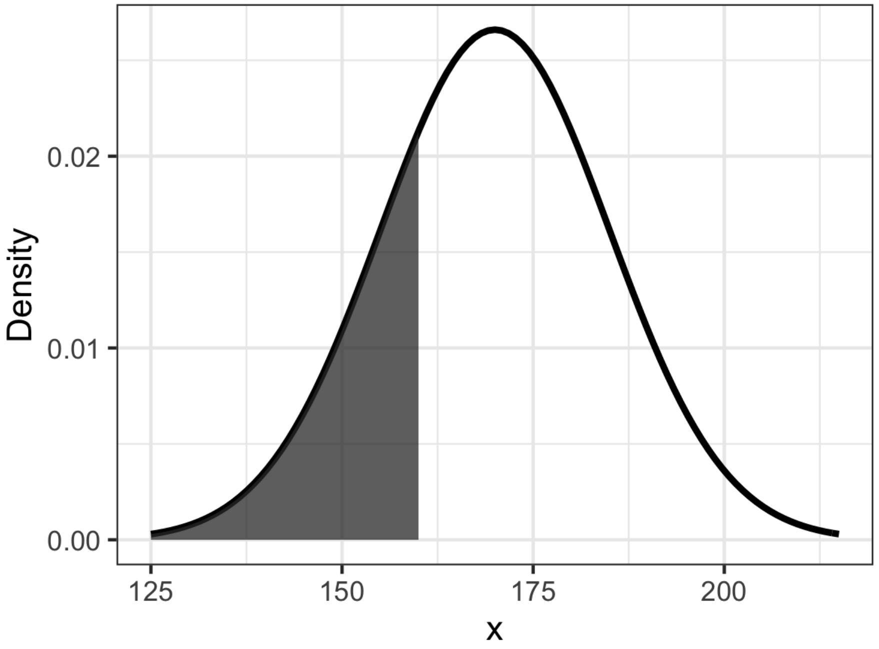
Continuous Random Variables & Distributions

Probabilities from a curve

Probability = area under the curve.

What is $P(X \leq 160)$?

It is the shaded area on the following figure



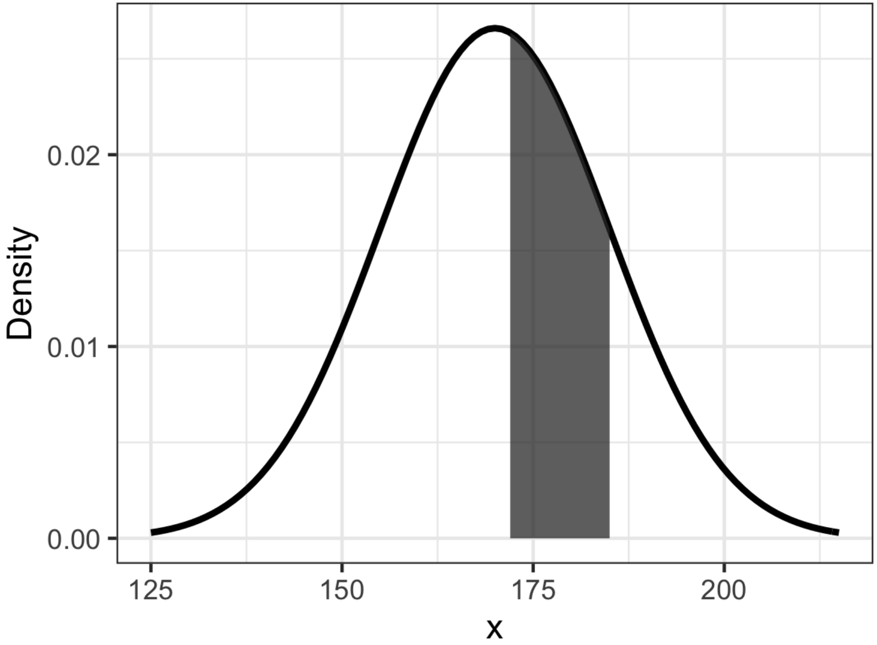
Continuous Random Variables & Distributions

Probabilities from a curve

Probability = area under the curve.

What is $P(172 < X < 185)$?

It is the shaded area on the following figure



The Normal Distribution

The Normal Distribution (also known as the Gaussian Distribution) is a continuous distribution.

It is specified using two parameters: the mean μ, and the variance σ^2. If X follows a normal distribution with mean μ, and variance σ^2, we write $X \sim N(\mu, \sigma^2)$.

Continuous Random Variables & Distributions

The Normal Distribution

Properties

The Normal Distribution has a quite a few really nice properties. If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then

1. the sum $X + Y$ is also normally distributed, and if they are independent $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$,
2. $X + a \sim N(\mu_X + a, \sigma_X^2)$, and $a \cdot X \sim N(a \cdot \mu_X, a^2 \sigma_X^2)$ where a is some constant,
3. $\frac{X - \mu_X}{\sigma_X} \sim N(0, 1)$. $N(0, 1)$ is called the *standard normal distribution*.

The Normal Distribution

Say we observe 10000 realizations of random variables X and Y. We will now take a look at $X+Y$, $X-Y$, $\frac{X-\bar{X}}{\text{SD}(X)}$, and $\frac{X-\bar{X}}{\text{SD}(X)}$.

First, this is the data.

Show entries Search:

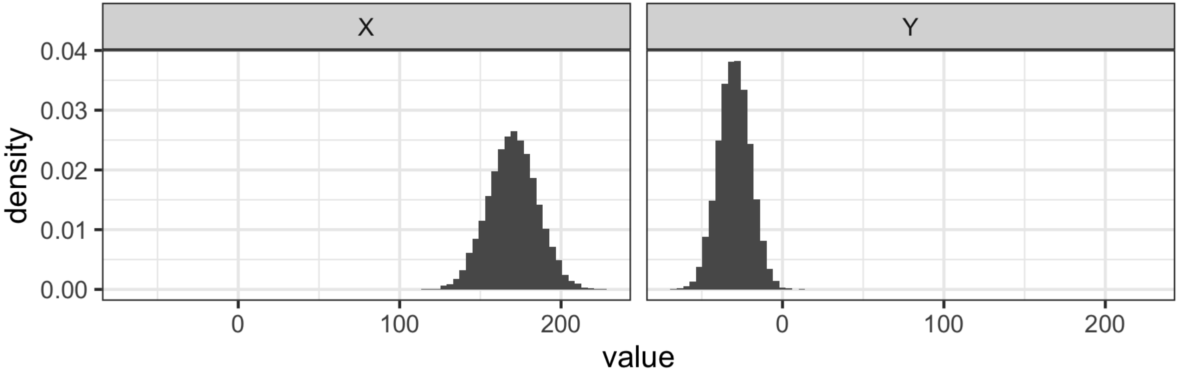
	X	Y
1	151.13	-36.27
2	181.15	-55.63
3	207.92	-29.75
4	161.46	-35.03
5	173.13	-26.31

Showing 1 to 5 of 10,000 entries

Continuous Random Variables & Distributions

The Normal Distribution

Let's take a look at the variables X, and Y.



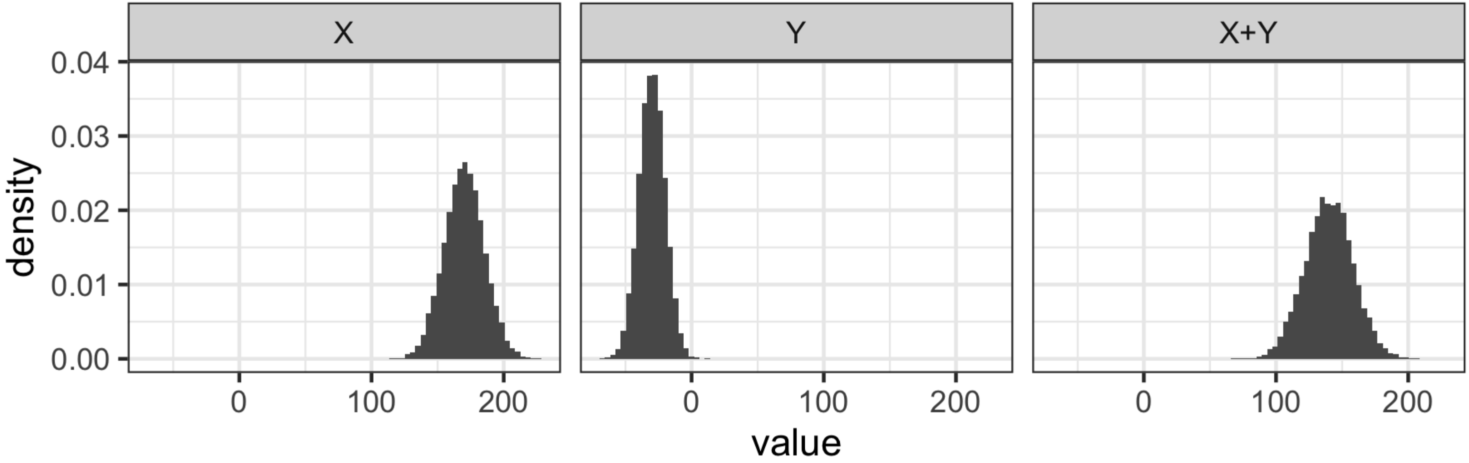
Continuous Random Variables & Distributions

The Normal Distribution

Variable	mean	var	emp_mean	emp_var
X	170	225	170.097	229.860
Y	-30	100	-29.943	99.514

The Normal Distribution

How does $X + Y$ compare?



Continuous Random Variables & Distributions

The Normal Distribution

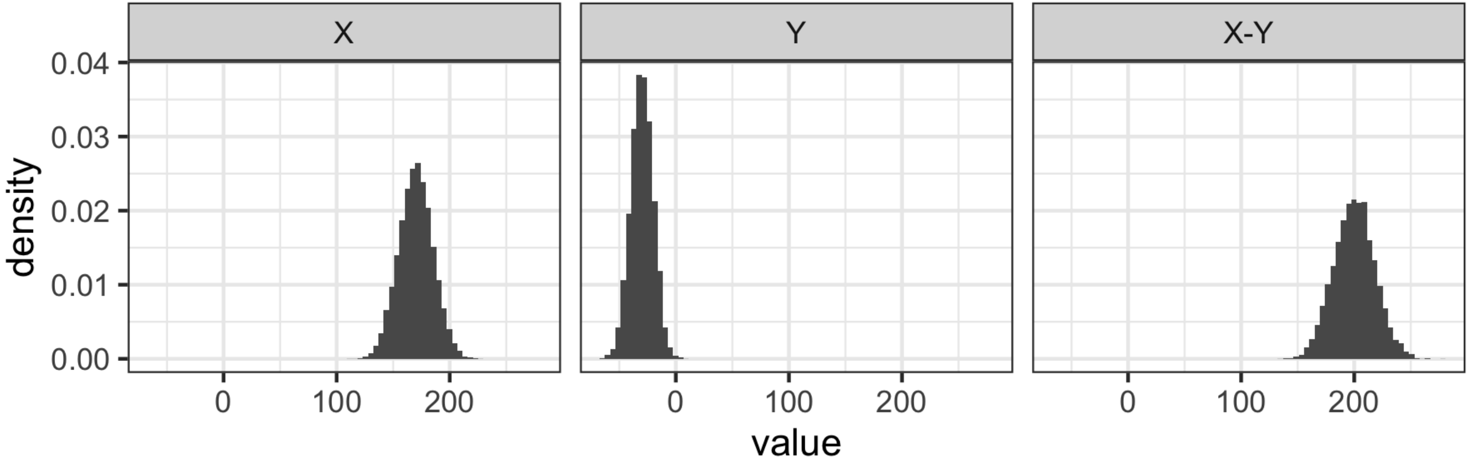
How does $X + Y$ compare?

Variable	mean	var	emp_mean	emp_var
X	170	225	170.097	229.860
Y	-30	100	-29.943	99.514
X+Y	140	325	140.154	330.717

Continuous Random Variables & Distributions

The Normal Distribution

How does $X - Y$ compare?



Continuous Random Variables & Distributions

The Normal Distribution

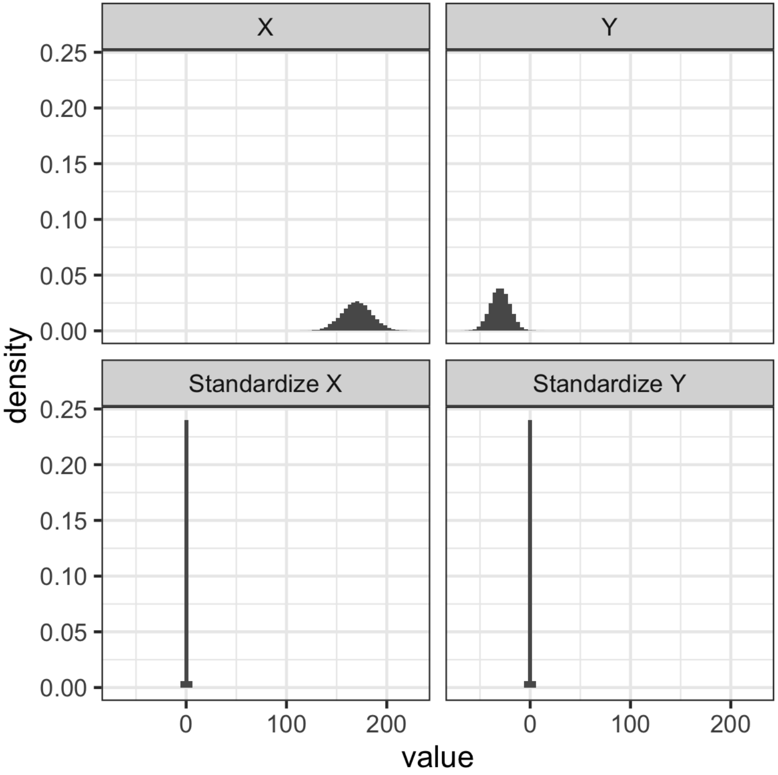
How does $X - Y$ compare?

Variable	mean	var	emp_mean	emp_var
X	170	225	170.097	229.860
Y	-30	100	-29.943	99.514
X-Y	200	325	200.040	328.030

Continuous Random Variables & Distributions

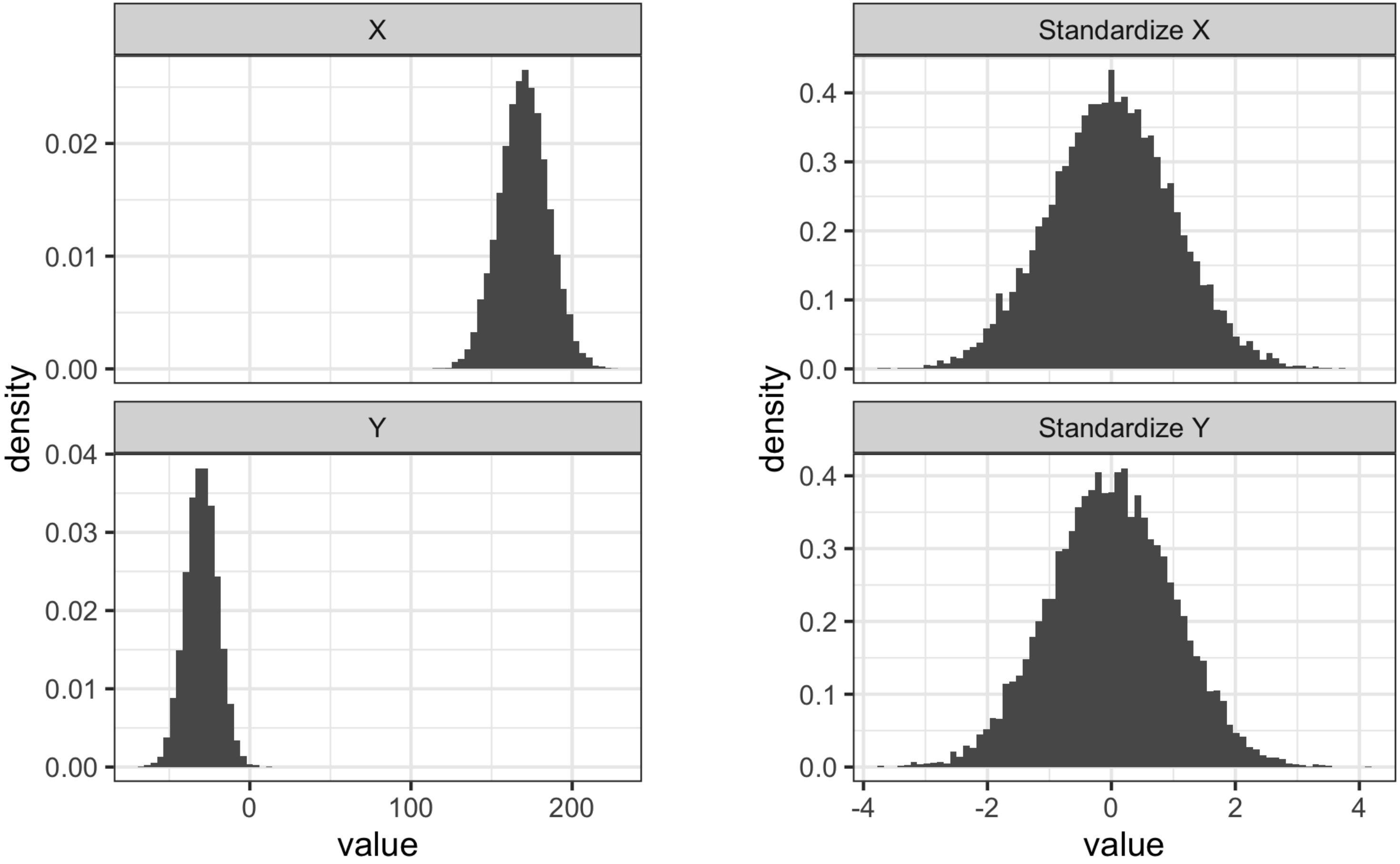
The Normal Distribution

And what if we standardize, i.e. subtract the mean, and divide by the standard deviation?



Continuous Random Variables & Distributions

The Normal Distribution



Continuous Random Variables & Distributions

The Normal Distribution

And what if we standardize, i.e. subtract the mean, and divide by the standard deviation?

Variable	mean	var	emp_mean	emp_var
X	170	225	170.097	229.860
Y	-30	100	-29.943	99.514
Standardize X	0	1	0.000	1.000
Standardize Y	0	1	0.000	1.000