Random Variables

August 7, 2019

Example: Commute Times

- I come to campus four days per week. (Fri-Sun I work from home.)
- We will use X_1 to represent my commute time on Monday, X_2 to represent commute time on Tuesday, etc.
- We want to write an equation using $X_1, ..., X_4$ that represents my weekly commute time for going to and from campus, denoted by W.

Example: Commute Times

My weekly commute time will be

$$W = X_1 + X_2 + X_3 + X_4.$$

- Breaking W down into several component parts allows us to understand each source of randomness.
- This may be useful for modeling W.
- For example, some days I get a ride to work. Other days I might be more likely to walk or take the bus.

Example: Commute Times

Let's say I spent an average of 168 minutes to commute to and from work each week.

This tells us that

$$E(W) = 168$$

but doesn't tell us anything about sources of randomness.

Example: Commute Times

If instead we knew that

- Mondays and Wednesdays I get a ride to and from work, for a total of about 14 minutes each day.
- Tuesdays and Thursdays I walk, for a total of about 70 minutes each day.

Then

$$E(W) = E(X_1) + E(X_2) + E(X_3) + E(X_4)$$
$$= 14 + 70 + 14 + 70 = 168$$

This lets us think about my day-to-day commute times (which can vary quite a bit) *and* lets us calculate my average weekly commute time.

Linear Combinations of Random Variables

We've alluded to two important concepts:

1. A final value can sometimes be described in an equation as the sum of its parts.

2. Putting the individual average values into this equation gives the average value we would expect in total.

We want to clarify and formalize this second point.

Linear Combinations of Random Variables

A **linear combination** of two random variables X and Y describes any situation where we can write our relationship out as

$$aX + bY$$

where a and b are some fixed, known numbers.

Example: Linear Combinations of Random Variables

- For my commute time, there were four random variables (one for each day I come to campus).
- Each random variable could be written as having a fixed coefficient of 1.

Then

$$W = 1X_1 + 1X_2 + 1X_3 + 1X_4.$$

Expected Values

If X and Y are random variables and a and b are some fixed numbers, then

$$E(aX + bY) = a \times E(X) + b \times E(Y).$$

Essentially, to compute the expected value of a linear combination of random variables, we plug in the average of each individual random variable, multiply by the constants, and compute the result.

Nonlinear Combinations of Random Variables

A nonlinear combination falls into any other format. For example, we might want to know about

$$X_2^{X_1} + \frac{X_2}{X_3}.$$

In this case, we cannot just plug in the means! These settings are often far more complicated.

We will not work with nonlinear combinations of random variables in this course.

Example: Investments

- Leonard has invested $6000 in Caterpillar Inc and $2000 in Exxon Mobil Corp.
- Let X represent the change in Caterpillar's stock next month and Y represent the change in Exxon Mobil's stock next month.
- We want to write an equation that describes how much money will be made or lost in Leonard's stocks for the month.

Example: Investments

Write an equation that describes how much money will be made or lost in Leonard's stocks for the month.

Assume X and Y are not in decimal form (e.g. if Caterpillar's stock increases 1%, then X = 0.01; if it loses 1%, then X = -0.01).

Then we can write an equation for Leonard's gain as

$$\$6000 \times X + \$2000 \times Y$$

Example: Investments

If Caterpillar stock rises at 2.0% monthly and Exxon Mobil at 0.2% monthly, the expected monetary gain is

$$
\begin{aligned}
E(\$6000 \times X + \$2000 \times Y) &= \$6000 \times E(X) + \$2000 \times E(Y) \\
&= \$6000 \times 0.02 + \$2000 \times 0.002 \\
&= \$124
\end{aligned}
$$

Variability in Linear Combinations of Random Variables

- So far, we've focused on expected values for linear combinations of random variables.
- However, like all random processes, linear combinations of random variables are variable!
- Thus we must also consider variability of linear combinations of random variables.

Variability in Linear Combinations of Random Variables

For example,

- We considered the expected net gain or loss of Leonard's stock portfolio, but we did not consider the volatility of the stock market.
- The stock market has increased slowly *on average* over the past 5 years.
- However, when your money is in stocks it is entirely possible to gain or lose money very quickly.
- Getting comfortable with this variability is crucial when investing in stocks, so we may be interested in thinking about exactly how volatile the stock market actually is.

Variability in Linear Combinations of Random Variables

- As before, we use the variance and standard deviation to examine variability.
- We can learn something about the variability of Leonard's stock portfolio using the variability of each individual stock's monthly return.

Variance of Linear Combinations of Random Variables

Given random variables X and Y and known constant numbers a and b, the variance of the linear combination $aX + bY$ is

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$$

Essentially, we plug in the variances for each individual variable and square the coefficients.

Unfortunately, the intuition for this formula is not clear and the proof is outside of the scope of this course.

Variance of Linear Combinations of Random Variables

- The formula on the previous slide requires that X and Y are independent.
- If X and Y are dependent, we must modify this equation.
- This modification requires something called the *covariance*.
- However, the covariance is outside of the scope of this course and therefore we will not cover this formula.

Note that X and Y do *not* need to be independent for the expected value formula for linear combinations of random variables.

We can use this formula to calculate the variance of Leonard's monthly return. First, suppose that

	Mean $(\bar{x})$	Standard Deviation (s)	Variance (s^2)
CAT	0.0204	0.0757	0.0057
XOM	0.0025	0.0455	0.0021

Example: Investments

We can then use the formula to calculate the variance of Leonard's monthly return.

$$
\begin{aligned}
Var(\$6000 \times X + \$2000 \times Y) &= 6000^2 \times Var(X) + 2000^2 \times Var(Y) \\
&= 36000000 \times 0.0057 + 4000000 \times 0.0021 \\
&= 205200 + 8400 \\
&= 213600
\end{aligned}
$$

The standard deviation is

$$
\sqrt{213600} = \$462.1688.
$$

Example: Commute Time

Suppose my daily commute has a standard deviation of 5 on days when I get a ride and a standard deviation of 10 on days when I walk.

Let's find the variability in my weekly commute time.

Find the variability in my weekly commute time.

First, we convert the standard deviations from my commute time to variances.

$$Var(X_1) = Var(X_3) = 5^2 = 25$$
$$Var(X_2) = Var(X_4) = 10^2 = 100$$

Find the variability in my weekly commute time.

My commute time was $X_1 + X_2 + X_3 + X_4$. So the variance is

$$
\begin{aligned}
Var(W) &= Var(X_1 + X_2 + X_3 + X_4) \\
&= 1^2 Var(X_1) + 1^2 Var(X_2) + 1^2 Var(X_3) + 1^2 Var(X_4) \\
&= 1^2 \times 25 + 1^2 \times 100 + 1^2 \times 25 + 1^2 \times 100 \\
&= 25 + 100 + 25 + 100 \\
&= 250
\end{aligned}
$$

and the standard deviation is

$$
sd(W) = \sqrt{Var(W)} = \sqrt{250} = 15.811.
$$

Linear Combinations With Negatives

Note that if we have a linear combination such as

$$aX + bY = 2X - 3Y$$

Then $a = 2$ and $b = -3$. The expected value will be

$$2E(X) - 3E(Y)$$

so the negative will impact the expectation.

Linear Combinations With Negatives

However, the variance will be

$$2^2 Var(X) + (-3)^2 Var(Y) = 4Var(X) + 9Var(Y)$$

so the negative will *not* impact the variance or standard deviation.

Continuous Distributions

So we've focused on cases where the outcome of a variable is discrete.

Now we want to consider a context where the outcome is a continuous numerical variable.
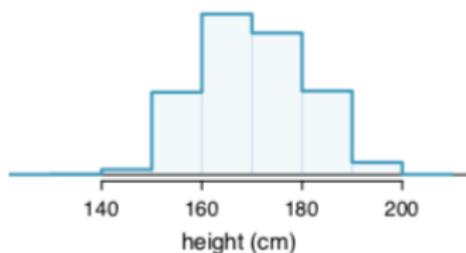
Continuous Distributions
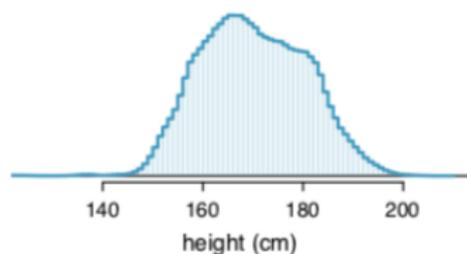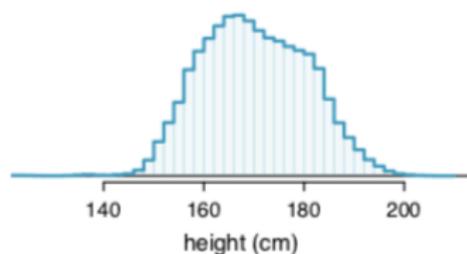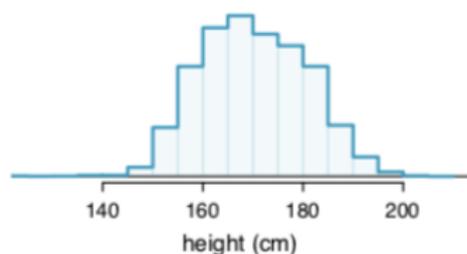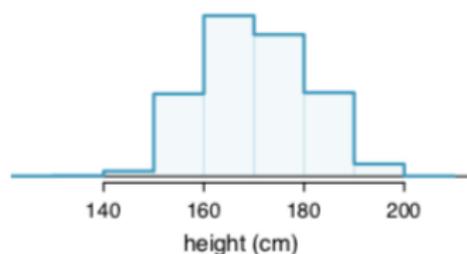
- These histograms are all of the same data.
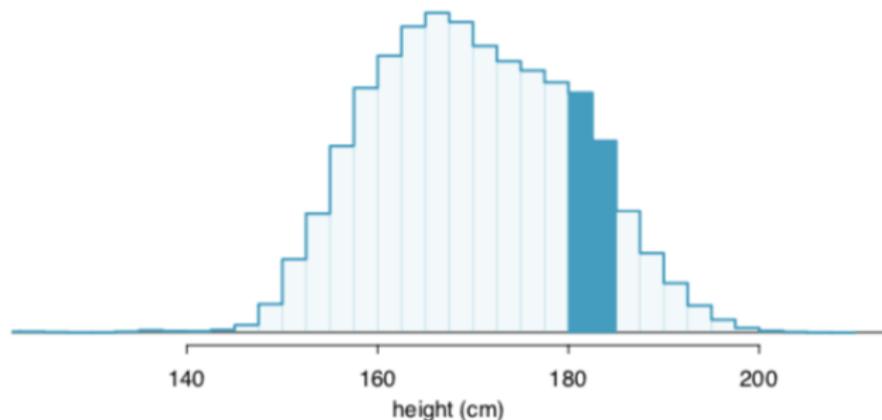- Varying bin widths allows us to make different interpretations of the data.
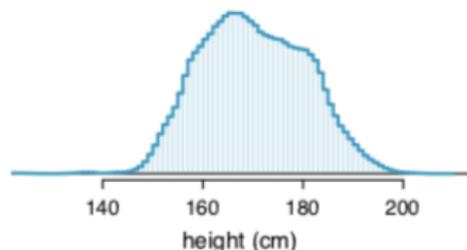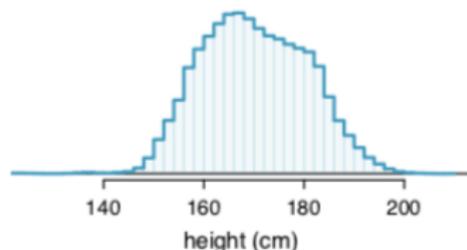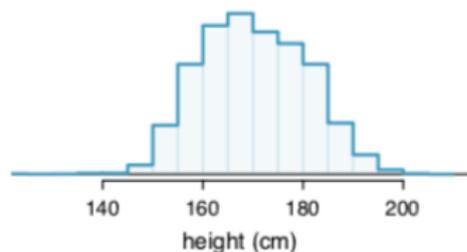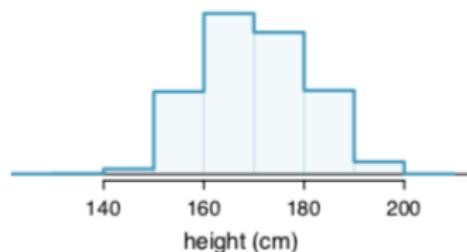
- By decreasing bin widths substantially, we "smooth out" the bumps in the histogram.

Example
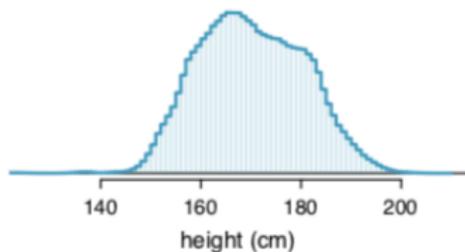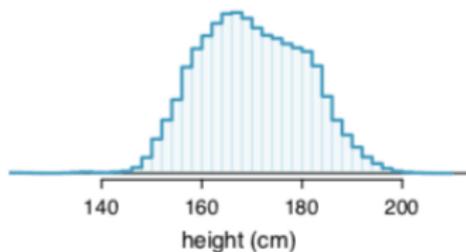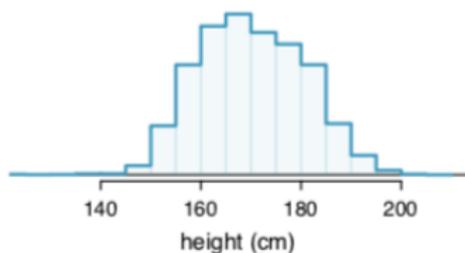
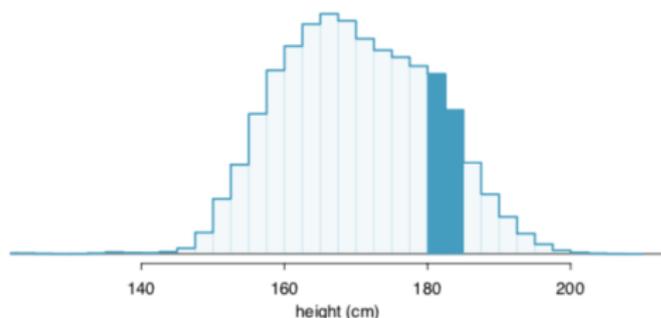What proportion of the sample is between 180 cm and 185 cm tall?

Example

To find the proportion of the sample between 180 and 185 cm,

- Add up the heights of the bins in the range 180 cm and 185 and divide by the sample size.
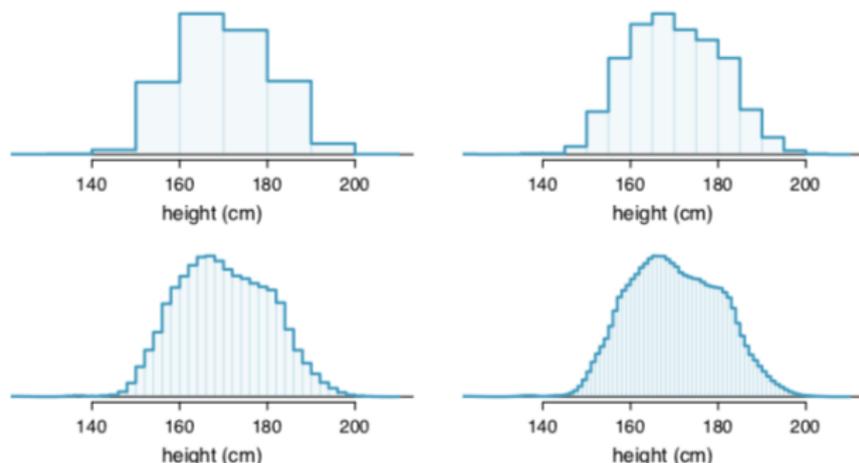
Example

- This can be done with the two shaded bins. These have counts of 195,307 and 156,239 people:

$$\frac{195307 + 156239}{3000000} = 0.1172$$

- This fraction is the same as the proportion of the histogram's area that falls in the range 180 to 185 cm.

Histograms to Continuous Distributions

- Let's examine the transition from the boxy hollow histogram (top left)) to the much smoother one (bottom right).
- The last plot has so many bins that the histogram is starting to resemble a smooth curve.
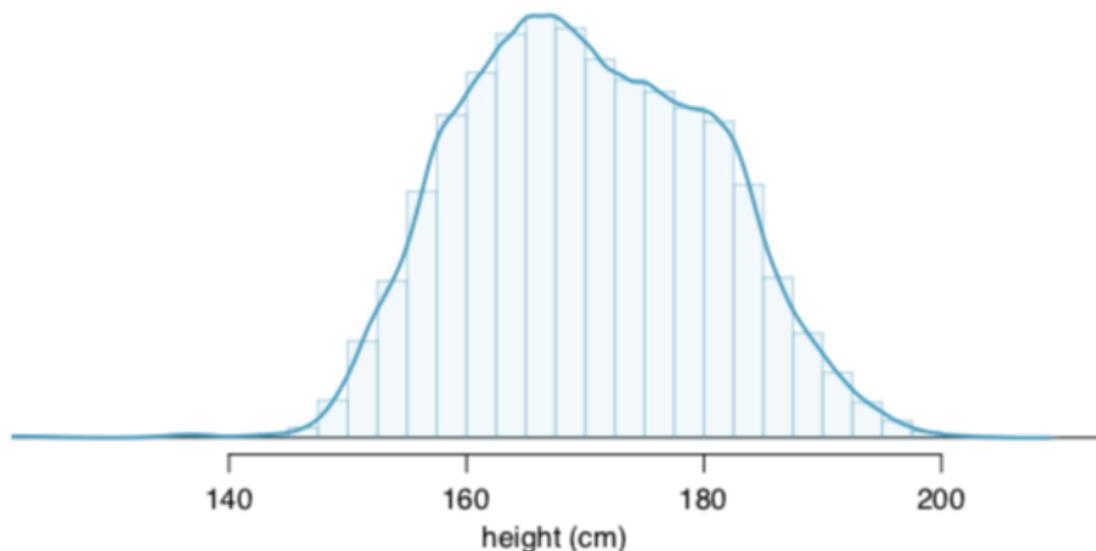
Histograms to Continuous Distributions
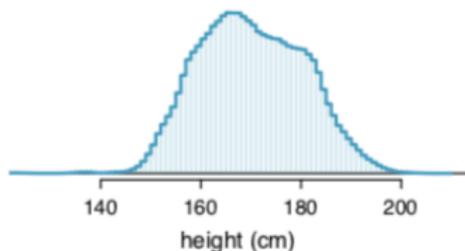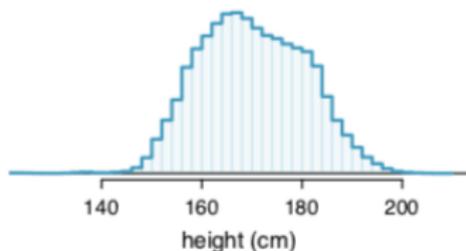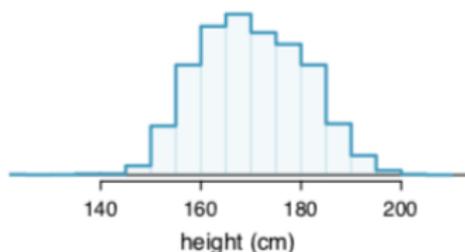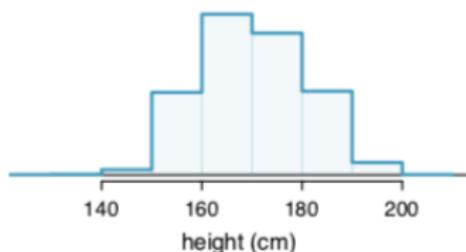
Population height as a *continuous* numerical variable might best be explained by a curve that represents the outline of extremely slim bins.

Histograms to Continuous Distributions

- This smooth curve represents a **probability density function**.
  - There are also called a **density** or **distribution**.
- A density has a special property: the total area under the density's curve is 1.

Histograms to Continuous Distributions

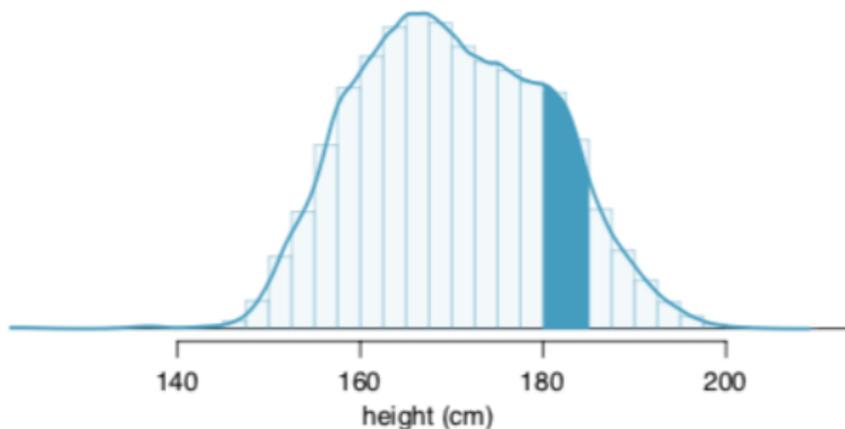Here, such a curve is shown overlaid on a histogram of the sample.

Probabilities from Continuous Distributions

We computed the proportion of individuals with heights 180 to 185 cm as

$$\frac{\text{number of people between 180 and 185cm}}{\text{total sample size}},$$

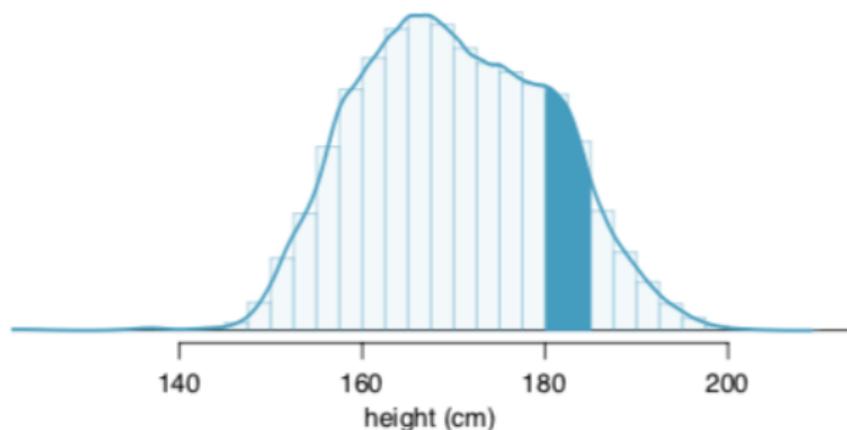the fraction of the histogram's area in this region.

Probabilities from Continuous Distributions

We can also use the area in the shaded region under the curve to find a probability (using calculus or with the help of a computer):

$$P(\text{height between 180 and 185}) = \text{area between 180 and 185} = 0.1157$$

Probabilities from Continuous Distributions

The probability that a randomly selected person is between 180 and 185 cm is 0.1157.

This is very close to the estimate from the previous example, when we used the histogram bins instead of the curve.

Rossman and Chance Applets

www.rossmanchance.com/applets/

This website contains a collection of applets that may be helpful in conceptualizing some of the topics from class.

Random Number Generator

www.rossmanchance.com/applets/RandomGen/GenRandom01.htm

- Generate random numbers with or without replacement.
- Suppose we gave everyone in the class a number.
- Then we can use the random number generator to choose random samples.

www.rossmanchance.com/applets/Dotplot.html

- Generate a random sample.
- See histograms, dotplots, and boxplots of your random sample.
  - Change bin widths for histograms.
- Check the mean, standard deviation, median, and IQR for your sample.
- You may also use your own data!

Monty Hall

www.rossmanchance.com/applets/MontyHall/Monty04.html

This applet simulates the Monty Hall problem we discussed yesterday.

Random Babies

www.rossmanchance.com/applets/randomBabies/RandomBabies.html

- This shows some number of babies being randomly assigned to families.
- It tracks the average number of correct matches.
- This and the Monty Hall applet are also demonstrations of the Law of Large Numbers.

Midterm

This concludes the material for the midterm.

Tomorrow we are moving on to Chapter 4!

Questions?

Are there any questions you'd like answered before Monday?