

1. Would you help me? If yes, unmute yourself, and let me know.
 - ① Be my co-host, and admit people from the waiting room.
 - ② If there is a question in the chat, please remind me.
 - ③ If you notice somebody raises their hand during lecture, please remind me.
2. Type your questions, if any, in the chat now.
3. I will set an alarm for 1hr 15min at 12:45pm.
4. I will press the record button now.

Unit 2: Data Summary and Visualization II

1. Numerical variables I

Stat 140 - 02

Mount Holyoke College

Dr. Shan Shan

Slides posted at <http://sshanshans.github.io/stat140>

1. Announcement
2. This week: Numerical variables
3. Today: Visualizing the distribution of a numerical variable
4. Main ideas
 1. Histogram
 2. Density plot
5. Summary





- ▶ Gianna Cai
- ▶ Alex Moreno
- ▶ Grace Wason
- ▶ Jingyi Wu

Office hours and links are posted on Moodle.

Go to Moodle

Go to Media Gallery (Side bar on the left)

You may need to wait for a few minutes for all the videos to show up

 question @19   

stop following 25 views

Actions ▾

Recorded Video

How do we access the recorded video on Moodle? I clicked on the link [Zoom Class Sessions \(Live & Recorded\)](#)External tool, but there has no video to be watched. Only an upcoming Zoom link is provided.

other

~ An instructor (Shan Shan) thinks this is a good question ~

The second evaluative assignment (10% of course grade) will be announced on Friday. This is a small group project. Please form a team of 4 people before class on Friday, and post your team information on Piazza.

Each team will nominate a group captain.

Sep 03, Thur 3:15 PM

MHC MATH/STAT DEPT
Virtual Tea Program
Presents Session 02



①

Evaluating contributions of
point-of-care data to inform
forecasts of influenza

Minh Tam Hoang

②

On Parking Sequences

Ruozhen Gong
Jingyi Wu

1. Announcement
2. This week: Numerical variables
3. Today: Visualizing the distribution of a numerical variable
4. Main ideas
 1. Histogram
 2. Density plot
5. Summary

- ▶ Distribution of one categorical variable
 - Bar plot
- ▶ Relationship between two categorical variables
 - Conditional distribution
- ▶ Intro to R, Rstudio and Rmarkdown

- ▶ Distribution of one numerical variable (Day 1,2)
 - Shape: skewness, modality, outliers
 - Summary: center and spread
- ▶ Relationship between two numerical variables (Day 3)
 - Association and correlation
- ▶ Plotting in R and exploratory data analysis (Day 4,5)

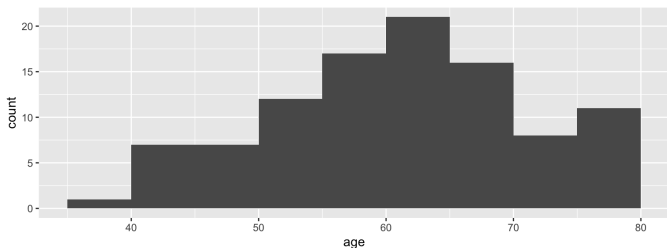
1. Announcement
2. This week: Numerical variables
3. Today: Visualizing the distribution of a numerical variable
4. Main ideas
 1. Histogram
 2. Density plot
5. Summary

1. Announcement
2. This week: Numerical variables
3. Today: Visualizing the distribution of a numerical variable
4. Main ideas
 1. Histogram
 2. Density plot
5. Summary

1. Announcement
2. This week: Numerical variables
3. Today: Visualizing the distribution of a numerical variable
4. Main ideas
 1. Histogram
 2. Density plot
5. Summary

Histograms are a common type of plot for displaying the distribution a numerical variable.

The x axis, representing the numerical variable, is divided into bins of equal width, and the height of each bar represents the number of units in that bin.



Histogram of the age variable in the 'senate 113' dataset.

In this class, we will be using a data set called 'senate 113' with information about the senators in the 113th US Senate (this is the senate that came into session in January 2013).

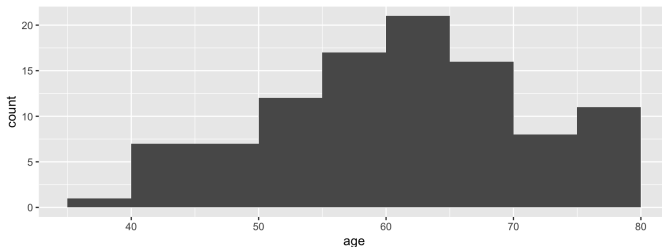
	▲ firstname ▾	middlename ▾	lastname ▾	birthday ▾	state ▾	party ▾	age ▾
1	Dianne	NA	Feinstein	1933-06-22	CA	D	79.5
2	Charles	E.	Grassley	1933-09-17	IA	R	79.3
3	Orrin	G.	Hatch	1934-03-22	UT	R	78.8
4	Richard	C.	Shelby	1934-05-06	AL	R	78.7
5	Carl	NA	Levin	1934-06-28	MI	D	78.5
6	James	M.	Inhofe	1934-11-17	OK	R	78.1
7	Pat	NA	Roberts	1936-04-20	KS	R	76.7
8	Barbara	A.	Mikulski	1936-07-20	MD	D	76.5

The 'senate 113' data set is from the *fivethirtyeight* package.

Poll question

What is each observational unit in this data set?

- a US Senate
- b All senators in the 113th US Senate
- c A senator in the 113th US Senate



Poll question

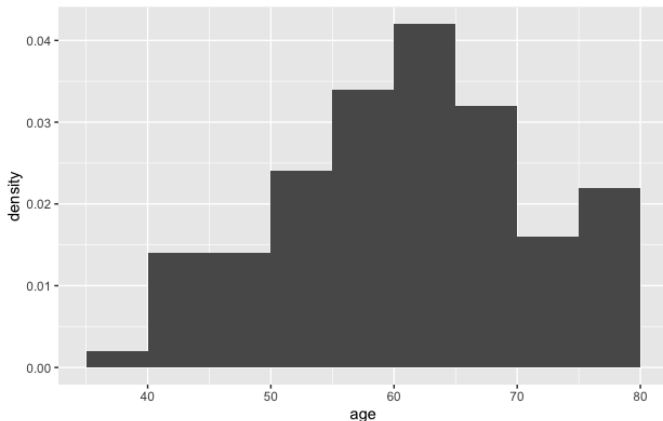
Based on this histogram, how many senators were aged between 40 and 50?

- a 1
- b 7
- c 14

You can also plot a **density histogram** where the vertical axis is density.

- ▶ Count: The **height** of each bar is the number of observational units in that bin.
- ▶ Density: The **area** of each bar is the proportion of observational units in that bin. (The height is whatever it needs to be to make the area work out correctly).

The following density histogram describes the distribution of the age variable in the senate 113 data set

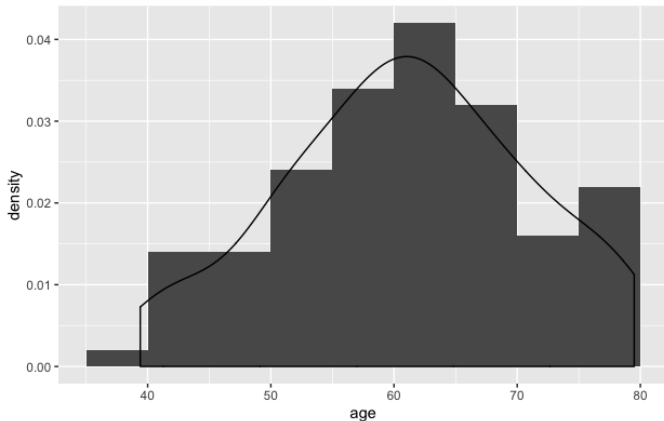


What's the sum of the area of all bars in the density histogram?

1. Announcement
2. This week: Numerical variables
3. Today: Visualizing the distribution of a numerical variable
4. Main ideas
 1. Histogram
 2. Density plot
5. Summary

A density plot is basically a smoothed density histogram.

Below is the density plot of the 'senate 113' dataset.

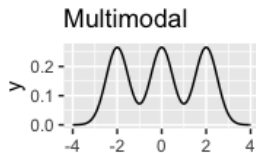
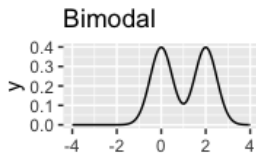
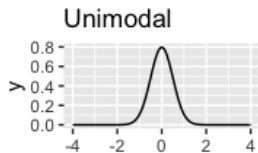


Challenge yourself: what's the area under the density plot?

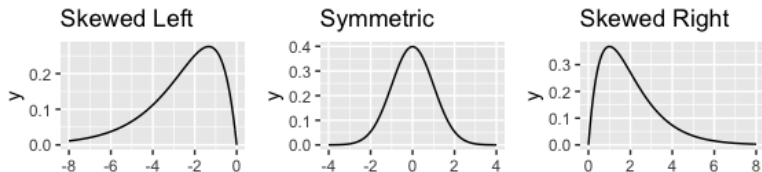
When I look at a histogram or density plot, I'm evaluating three characteristics of the plot:

1. mode
2. skewness
3. gaps or outliers

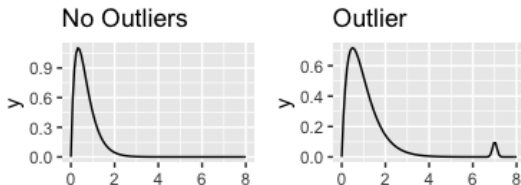
A mode is a local peak in the distribution.



If a distribution is skewed, it's **skewed towards the tail**.



An outlier is a data value that “doesn’t fit” with the rest of the data.



1. Announcement
2. This week: Numerical variables
3. Today: Visualizing the distribution of a numerical variable
4. Main ideas
 1. Histogram
 2. Density plot
5. Summary

1. Histogram
2. Density plot