Week 2: Data Summary and Visualization

3. Numerical variables I

Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

Outline

1. Numerical variables

2. Today: Visualizing the distribution of a numerical variable

3. Main ideas
   1. Histogram
   2. Density plot
   3. Center and Spread
   4. Boxplot and 5-Number Summaries

4. Summary

So far: Categorical variables

- 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
  - Shape: skewness, modality, outliers
  - Summary: center and spread
- Relationship between two numerical variables (next week)
  - Association and correlation
- Plotting in R and exploratory data analysis

Outline

1. Numerical variables

2. Today: Visualizing the distribution of a numerical variable

3. Main ideas
    1. Histogram
    2. Density plot
    3. Center and Spread
    4. Boxplot and 5-Number Summaries

4. Summary

Outline

1. Numerical variables

2. Today: Visualizing the distribution of a numerical variable

3. Main ideas
    1. Histogram
    2. Density plot
    3. Center and Spread
    4. Boxplot and 5-Number Summaries

4. Summary

1. Numerical variables

2. Today: Visualizing the distribution of a numerical variable

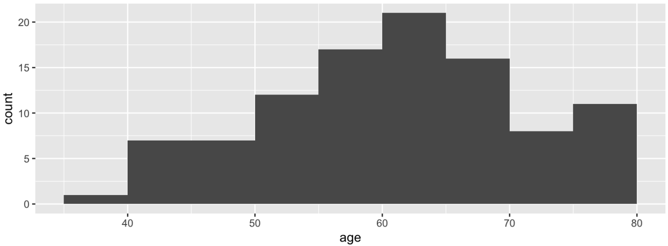
3. Main ideas
    1. Histogram
    2. Density plot
    3. Center and Spread
    4. Boxplot and 5-Number Summaries

4. Summary

Histogram

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

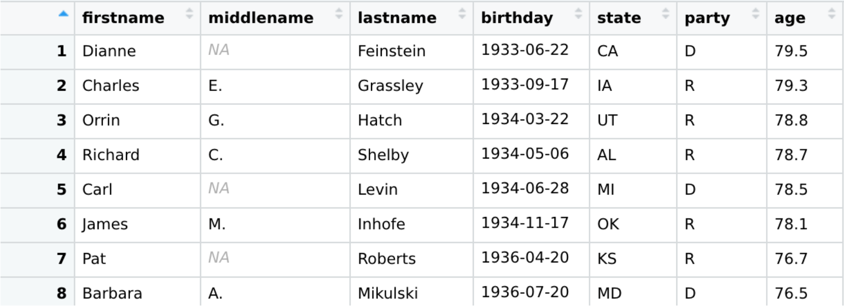
The x axis, representing the numerical variable, is divided into bins of equal width, and the height of each bar represents the number of units in that bin.



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

A word on the 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

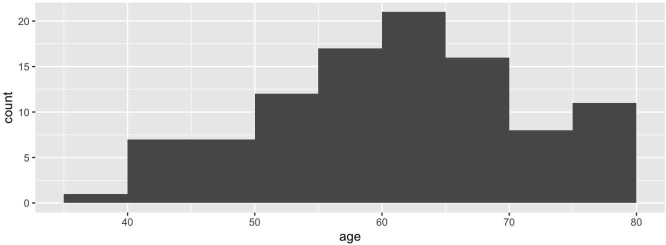
The ‘senate 113‘ data set is from the *fivethirtyeight* package.

Poll question

What are the observational units in this data set?

a. US Senate
b. All senators in the 113th US Senate
c. Age of the senators

Warm up



Poll question

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

a 1

b 7

c 14

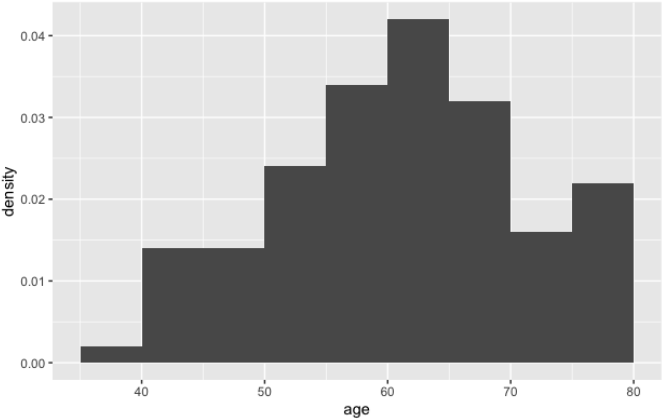
Density histogram

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).

Think harder

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?

Outline

1. Numerical variables

2. Today: Visualizing the distribution of a numerical variable

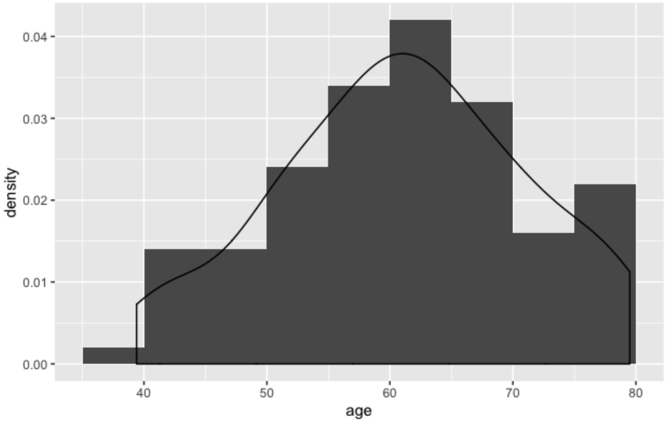
3. Main ideas
    1. Histogram
    2. Density plot
    3. Center and Spread
    4. Boxplot and 5-Number Summaries

4. Summary

Density plot

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?

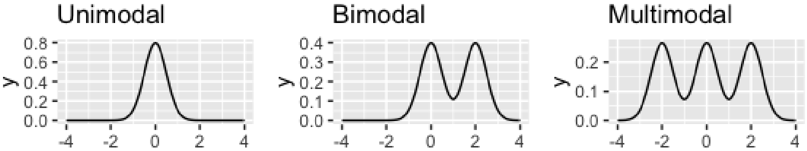
Describing the shapes

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

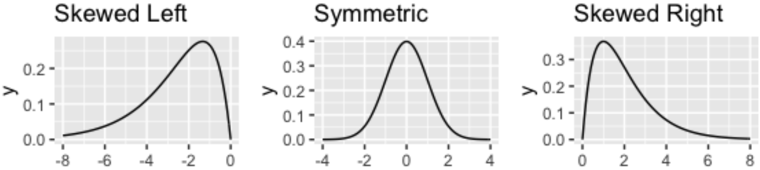
1. mode
2. skewness
3. gaps or outliers

(1) mode

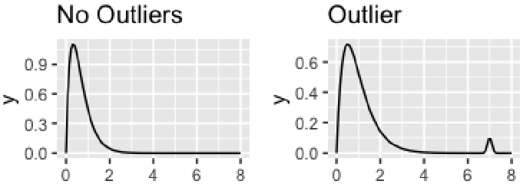
A mode is a local peak in the distribution.



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.



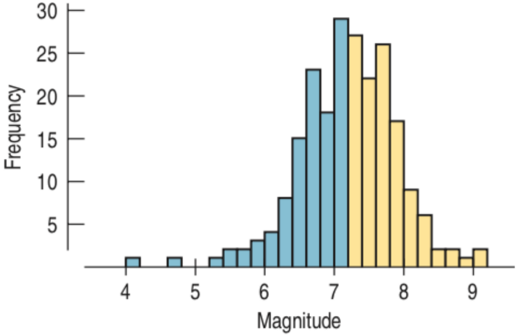
1. Numerical variables

2. Today: Visualizing the distribution of a numerical variable

3. Main ideas
    1. Histogram
    2. Density plot
    3. Center and Spread
    4. Boxplot and 5-Number Summaries

4. Summary

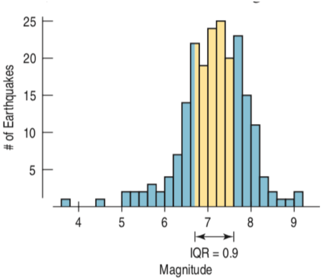
Median is the center of a histogram. Half of the data are less than the median and half are greater than the median.



There are three common measures of the **spread** of a distribution (how "wide" is it?):

1. The **inter-quartile range** (IQR):

$$\text{IQR} = \text{Q3 - Q1} = \text{75th percentile - 25th percentile}$$



The IQR is the width of an interval covering the middle half of the data.

Terminology

1. The 75th percentile (the number such that 75% of the data are less than that value, and 25% are greater than that value). Also called the third quartile.
2. The 25th percentile (the number such that 25% of the data are less than that value, and 75% are greater than that value). Also called the first quartile.

Outline

1. Numerical variables

2. Today: Visualizing the distribution of a numerical variable

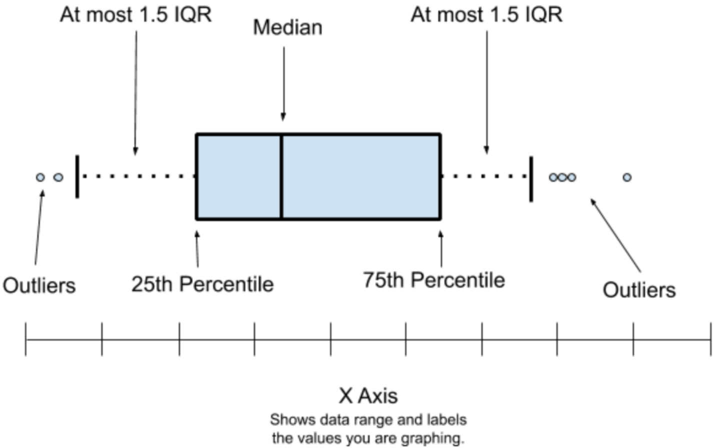
3. Main ideas
    1. Histogram
    2. Density plot
    3. Center and Spread
    4. Boxplot and 5-Number Summaries

4. Summary

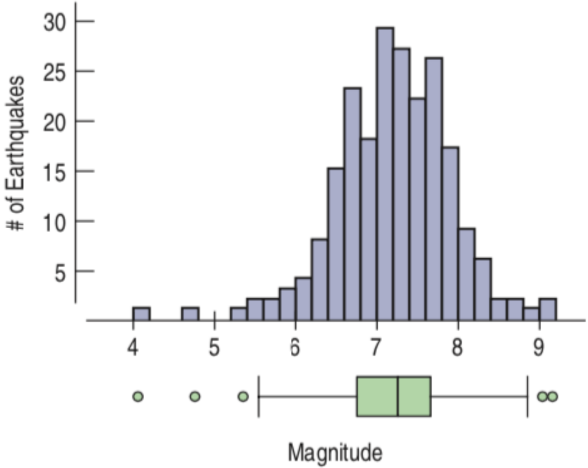
5-Number Summaries

1. The maximum: the largest value in the data set
2. The 75th percentile (the number such that 75% of the data are less than that value, and 25% are greater than that value). Also called the third quartile.
3. The median (the number such that half of the data are less than that value and half are greater than that value)
4. The 25th percentile (the number such that 25% of the data are less than that value, and 75% are greater than that value). Also called the first quartile.
5. The minimum: the smallest value in the data set

Boxplot


At most 1.5 IQR
Median
At most 1.5 IQR
Outliers
25th Percentile
75th Percentile
Outliers
X Axis
Shows data range and labels
the values you are graphing.

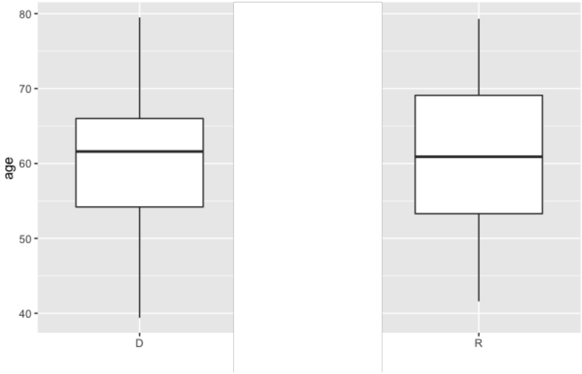
Boxplot



Poll question

Which party had the highest median age?

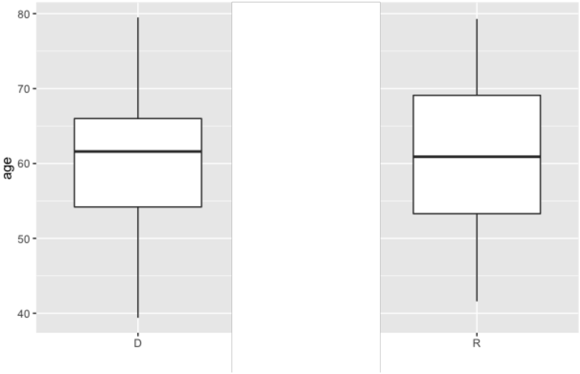
a Democrat
b Republican



Poll question

The youngest member of the senate belonged to which party?

a. Democrat
b. Republican

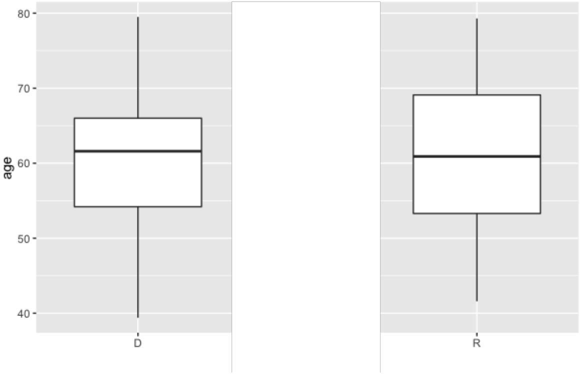


Poll question

75% of Republican senators were younger than what age?

ⓐ 55

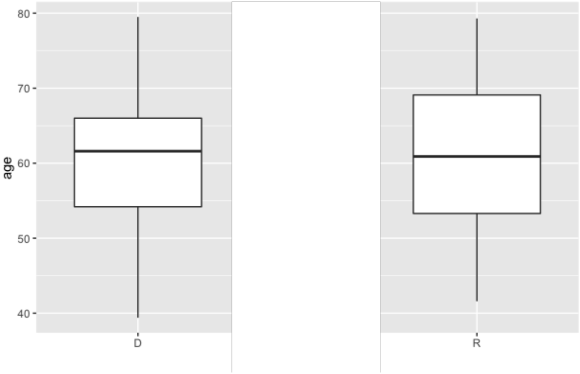
ⓑ 70



Poll question

How wide of an interval would you need to cover the ages of the middle half of Democratic senators?

- a 10
- b 40



Outline

1. Numerical variables

2. Today: Visualizing the distribution of a numerical variable

3. Main ideas
   1. Histogram
   2. Density plot
   3. Center and Spread
   4. Boxplot and 5-Number Summaries

4. Summary

Summary of main ideas

1. Histogram
2. Density plot
3. Center and Spread
4. Boxplot and 5-Number Summaries