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 at 12:45pm.
4. I will press the record button now.

Unit 2: Data Summary and Visualization II

2. Numerical variables II

Stat 140 - 02

Mount Holyoke College

Dr. Shan Shan

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

Outline

1. Today: Summary statistics

2. Main ideas
    1. Measures of Central Tendency
    2. Measures of Dispersion
    3. Boxplot and 5-Number Summaries

3. Summary

Given a visualization of the distribution of a numerical variable, one needs to find ways to summarize the information that facilitate understanding and insight. Two standard tools for this are:

1. Measures of Central Tendency (mean, median)
2. Measures of Dispersion (standard deviation, range, IQR)

1. Today: Summary statistics

2. Main ideas
    1. Measures of Central Tendency
    2. Measures of Dispersion
    3. Boxplot and 5-Number Summaries

3. Summary

1. Today: Summary statistics

2. Main ideas
    1. Measures of Central Tendency
    2. Measures of Dispersion
    3. Boxplot and 5-Number Summaries

3. Summary

Suppose we observe n numbers, $x_1, \ldots, x_n$.

There are two commonly used statistics used to summarize the **center** of the distribution of these values:

- The **mean** is the average of these values (add them up and divide by n). We use $\bar{x}$ to denote the mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + \cdots + x_n}{n}$$

- The **median** is the middle value when you arrange them in order. (If the sample size n is even, you take the average of the middle two values)

Suppose one observes the following data:

$$1, 0, 2, -2, 1, -2, 5, -1$$

The mean is $\bar{X} = \frac{1}{8}(1 + 0 + 2 - 2 + 1 - 2 + 5 - 1) = 0.5$.

Let's order the data,

$$-2, -2, -1, 0, 1, 1, 2, 5$$

The median is the average of 0 and 1, or 0.5.

Poll question

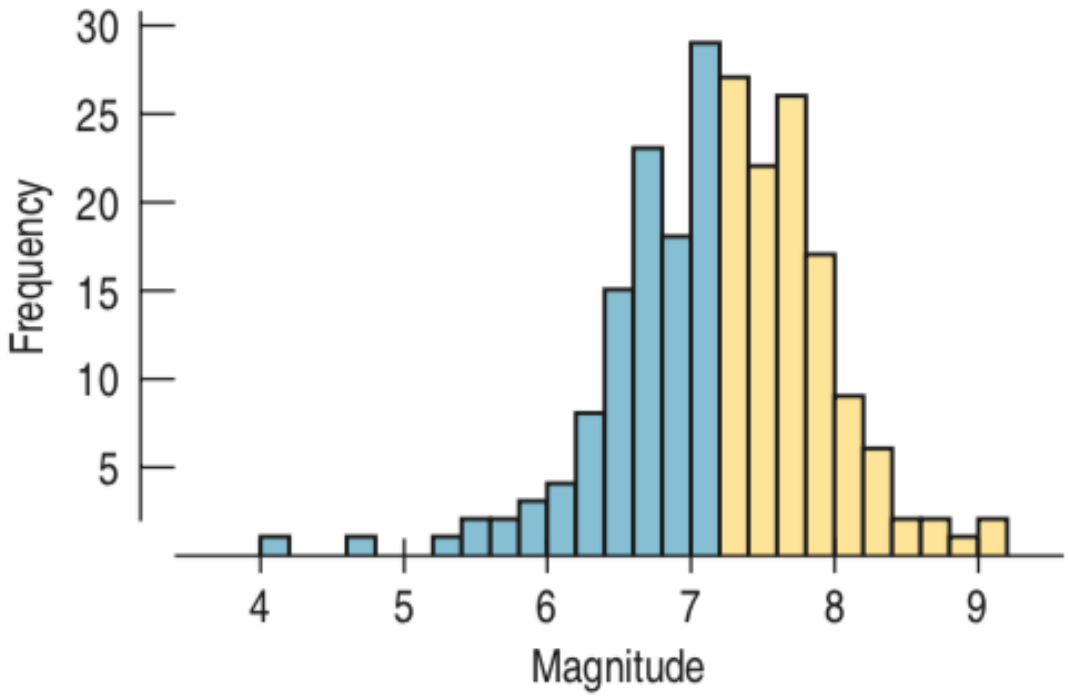
How do the mean and median of the following two datasets compare?
Dataset 1: 30, 50, 70, 90
Dataset 2: 30, 50, 70, 1000

a. $\bar{x}_1 = \bar{x}_2, \text{median}_1 = \text{median}_2$
b. $\bar{x}_1 < \bar{x}_2, \text{median}_1 = \text{median}_2$
c. $\bar{x}_1 < \bar{x}_2, \text{median}_1 < \text{median}_2$
d. $\bar{x}_1 > \bar{x}_2, \text{median}_1 < \text{median}_2$
e. $\bar{x}_1 > \bar{x}_2, \text{median}_1 = \text{median}_2$

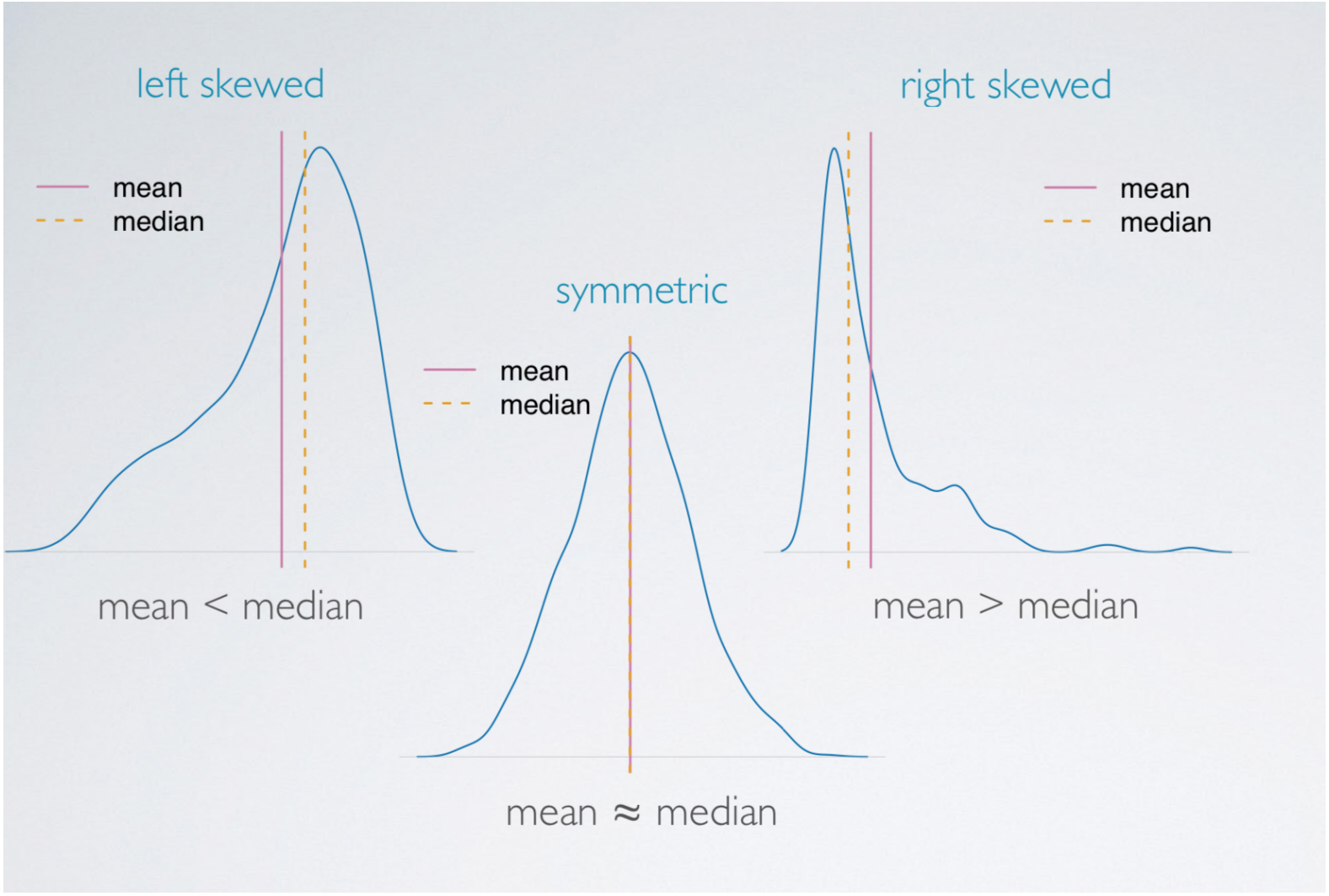
The mean can be pulled in misleading directions if there are outliers. A single large or small datum will have a large influence on the mean, but not on the median.

An outlier is an incorrect or unrepresentative observation that is very different from the others in the sample.

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



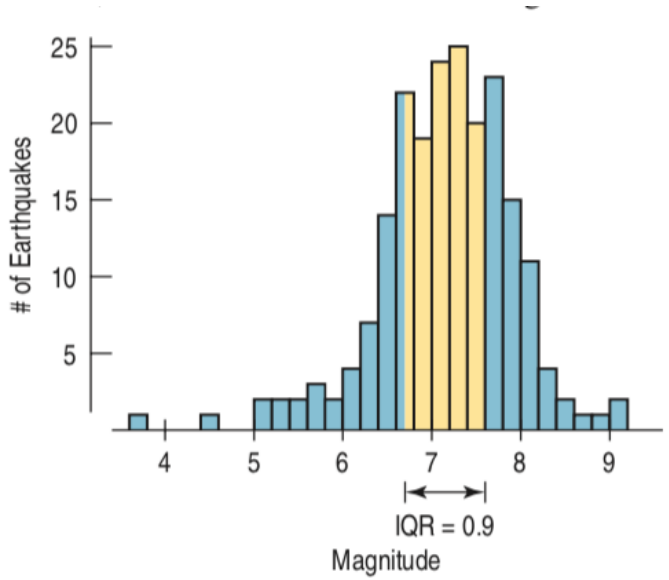
The mean can be pulled towards the tail if the distribution is skewed.



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.

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.

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

2. The **variance** is (almost) the average squared difference of each observation from the mean.

$$\text{Variance} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

3. The **standard deviation** is the square root of the variance. Intuitively, you can think of it as the average distance of the data points from the mean (although technically, that's not exactly right).

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

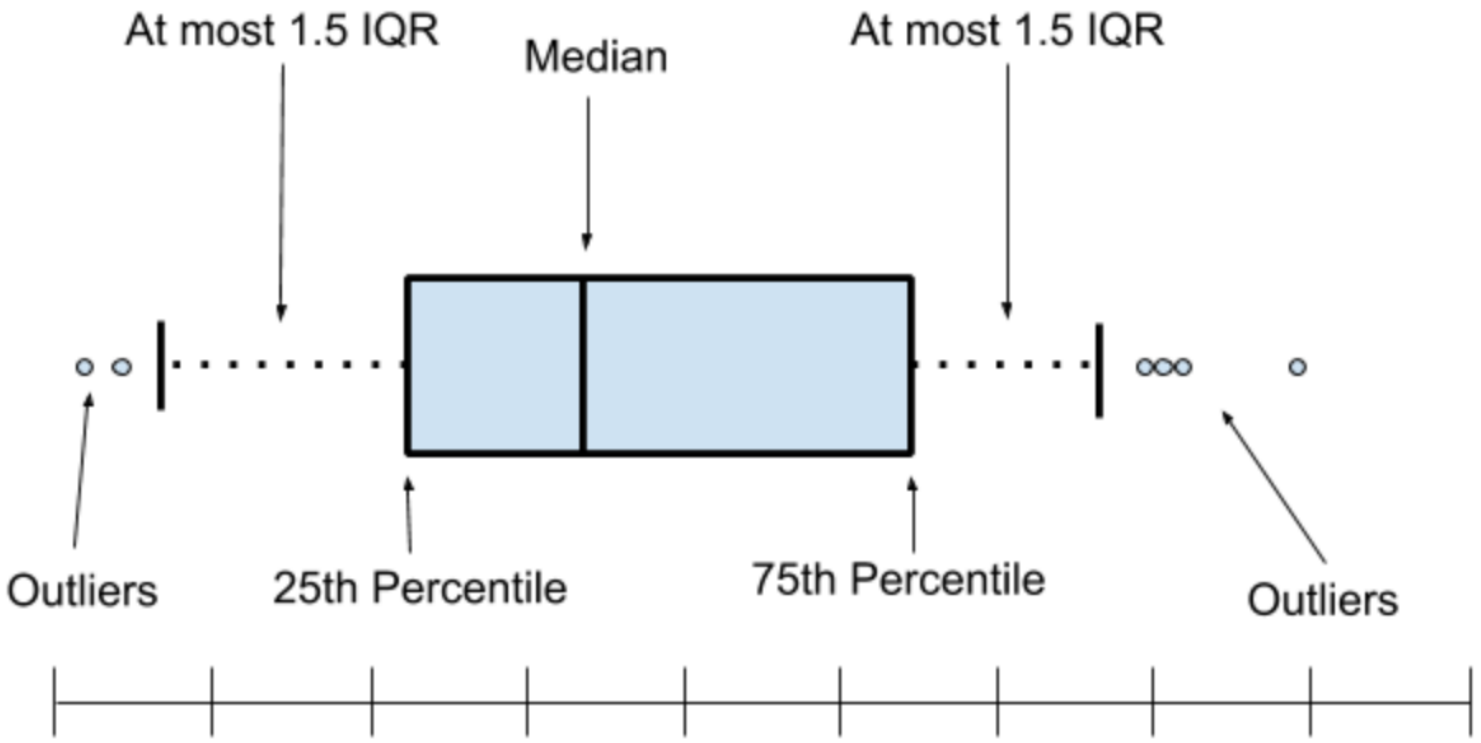
Outline

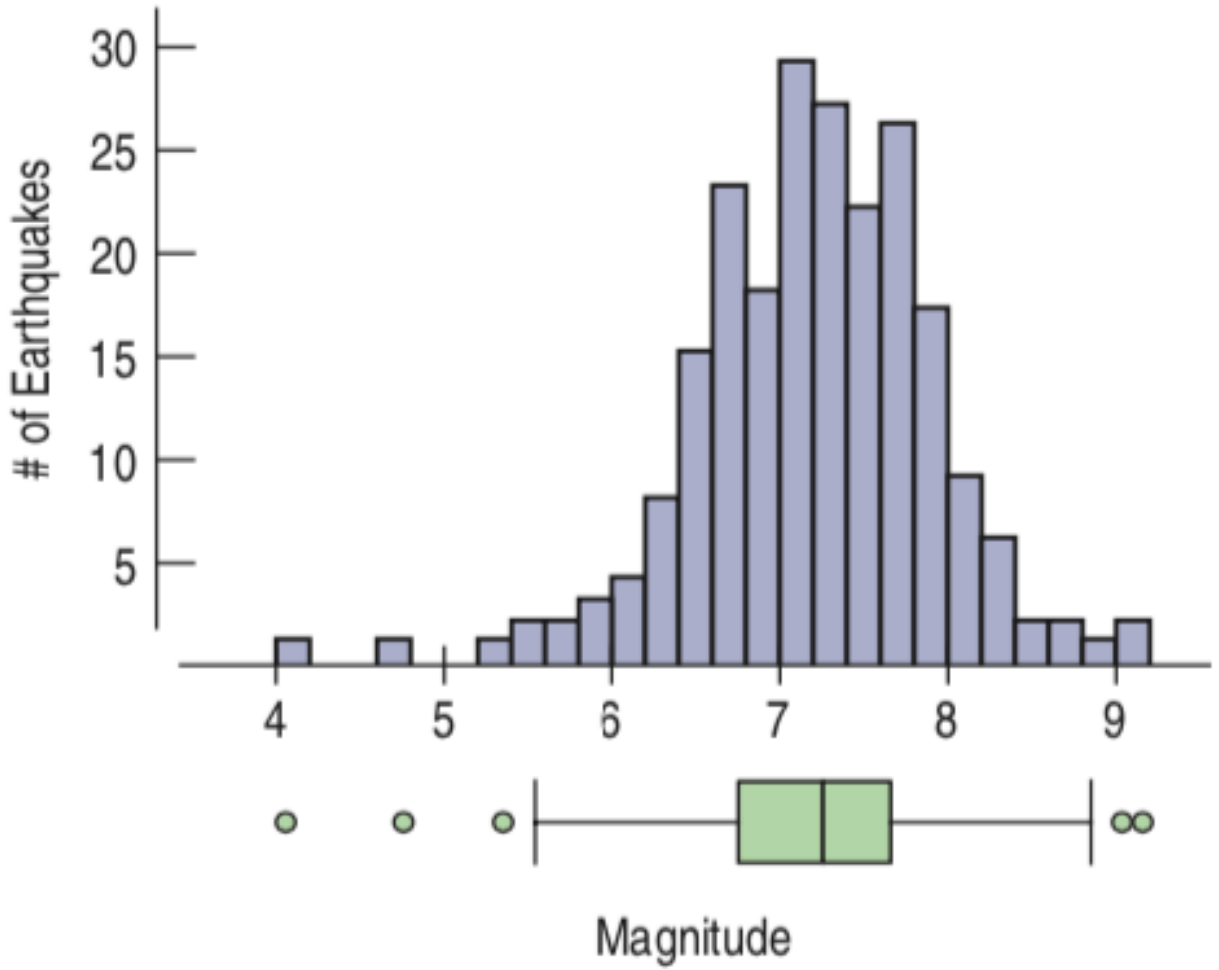
1. Today: Summary statistics

2. Main ideas
    1. Measures of Central Tendency
    2. Measures of Dispersion
    3. Boxplot and 5-Number Summaries

3. Summary

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

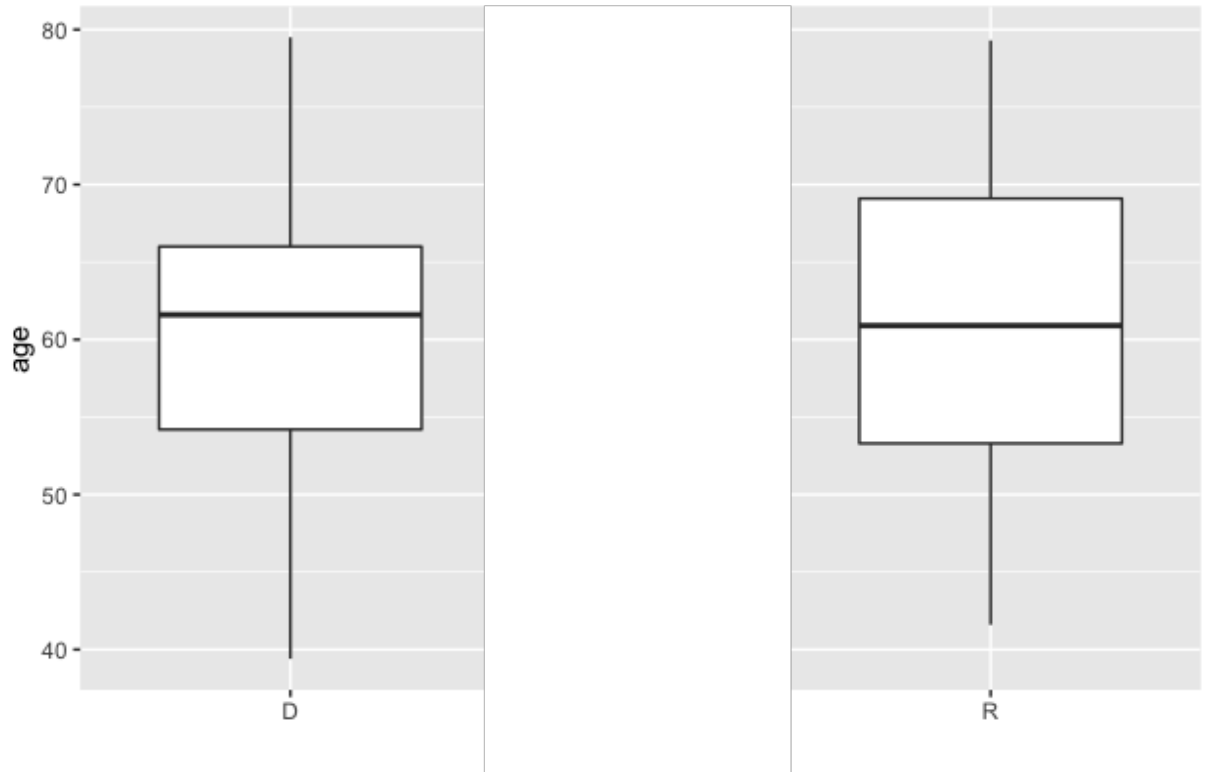

At most 1.5 IQR
Median
At most 1.5 IQR
Outliers
25th Percentile
75th Percentile
Outliers
X Axis
Shows data range and labels
the values you are graphing.


# of Earthquakes
30
25
20
15
10
5
4
5
6
7
8
9
Magnitude

Poll question

Which party had the highest median age?

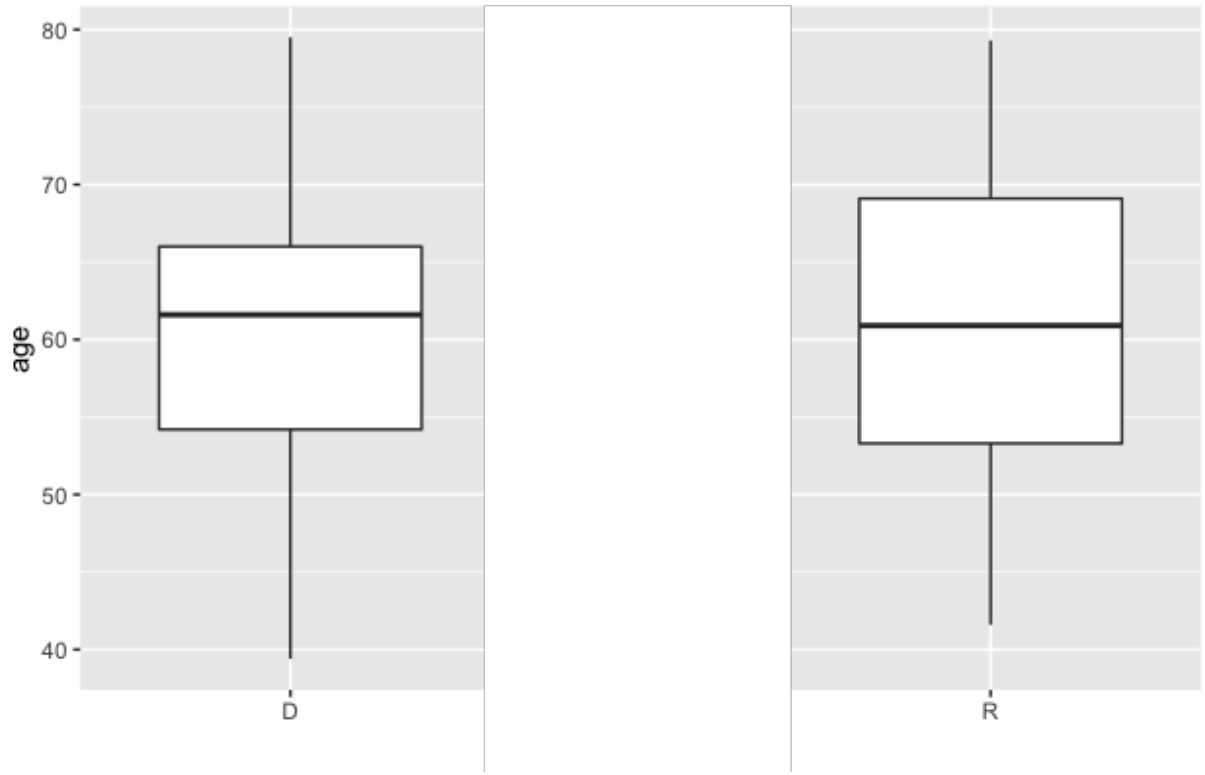
a. Democrat
b. Republican



Poll question

The youngest member of the senate belonged to which party?

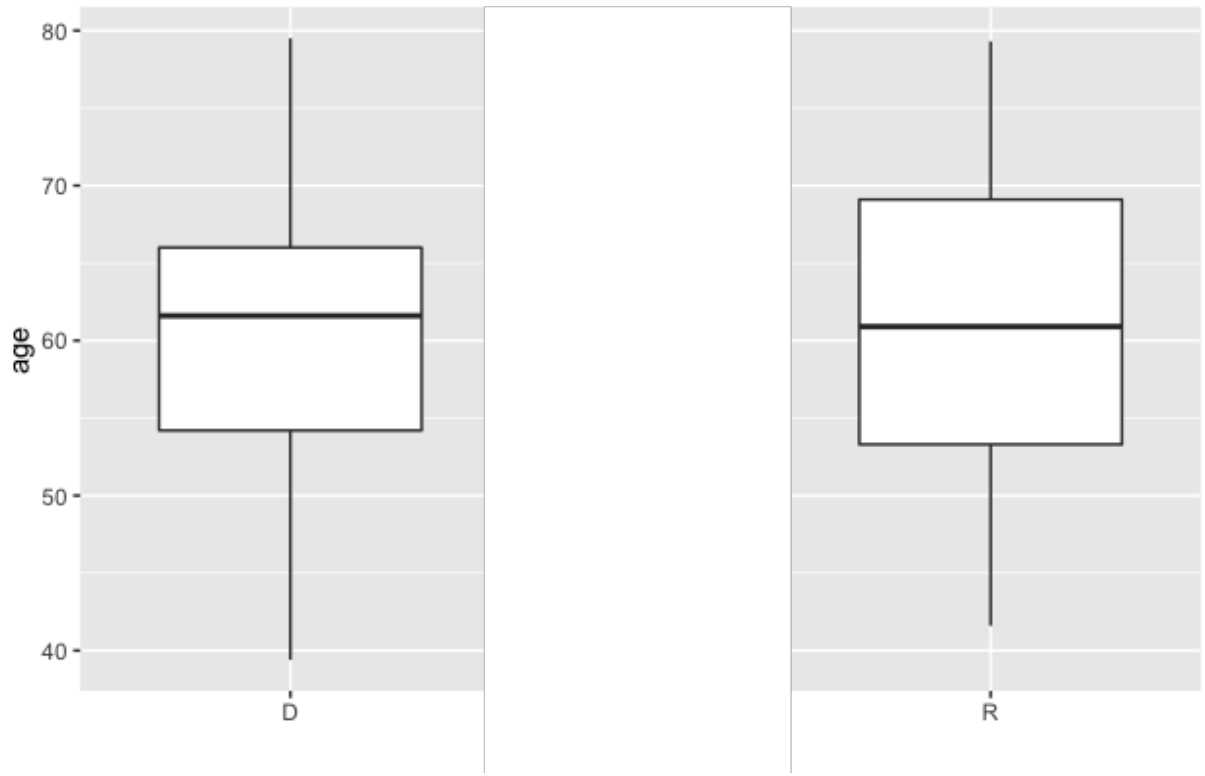
a. Democrat
b. Republican



Poll question

75% of Republican senators were younger than what age?

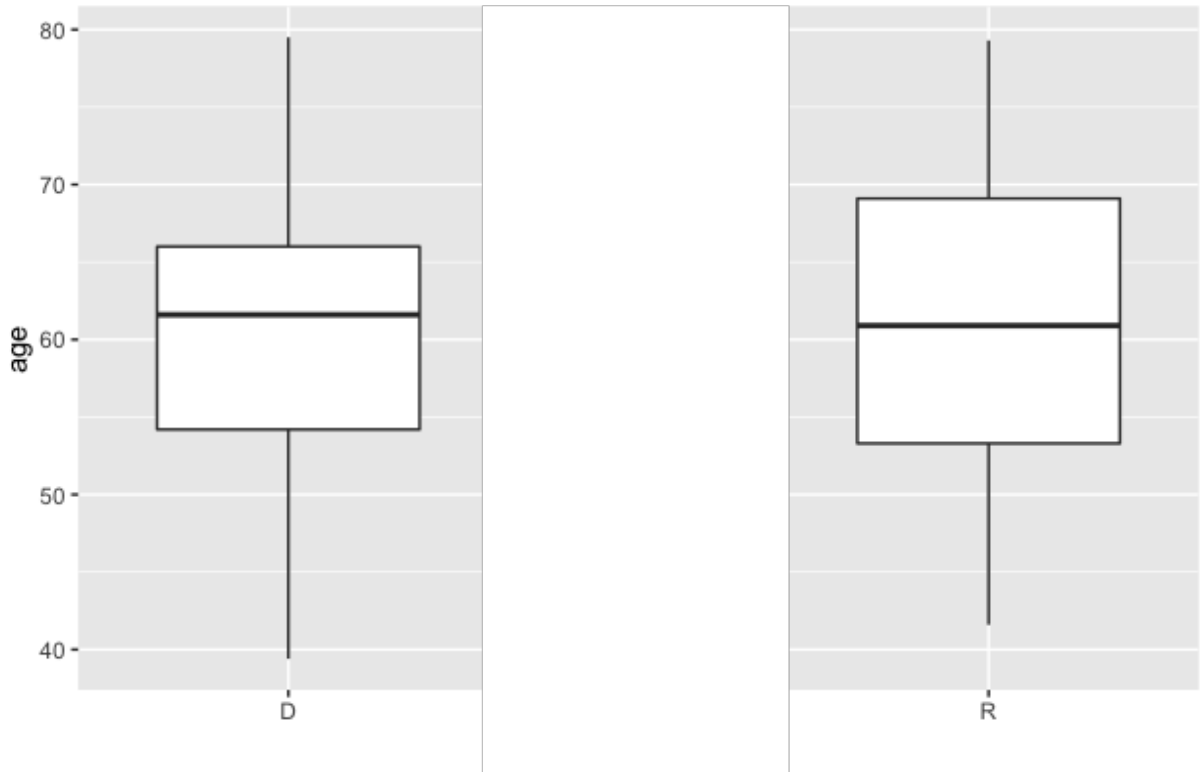
a. 55
b. 70



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



Tutorial exercise: For the rest of the class

Finish 06 Exercise

See link in zoom chat or on daily schedule

Goal:

(1) Practice computing mean/mdian by hand

(2) Practice computing summary statitics by R

Submit your .pdf file on Moodle.

Outline

1. Today: Summary statistics

2. Main ideas
    1. Measures of Central Tendency
    2. Measures of Dispersion
    3. Boxplot and 5-Number Summaries

3. Summary

1. Measures of Central Tendency
2. Measures of Dispersion
3. Boxplot and 5-Number Summaries

Message:
Mean, Variance, and Standard deviation are sensitive to outliers and skewness. They should only be used when a distribution looks "nice" (unimodal, symmetric, no outliers). Otherwise, use median and IQR to summarize center and spread.