Pictures to visualize the distribution of data:

- Numerical data: dot plot, histogram, density plot
- Categorical data: bar chart, pie chart

Features of a distribution:

Feature	**Measure**	
	Numerical data	*Categorical data*
Shape	Symmetric, Skewed	N/A
Center	Mode, Mean, Median	Mode
Spread	Standard deviation, Interquartile range	Number of categories

Section 3.4
Comparing Measures of Center

- In a symmetric distribution, the mean and the median are approximately the same.
- In a right-skewed distribution, the mean tends to be greater than the median.
- In a left-skewed distribution, the mean tends to be less than the median.

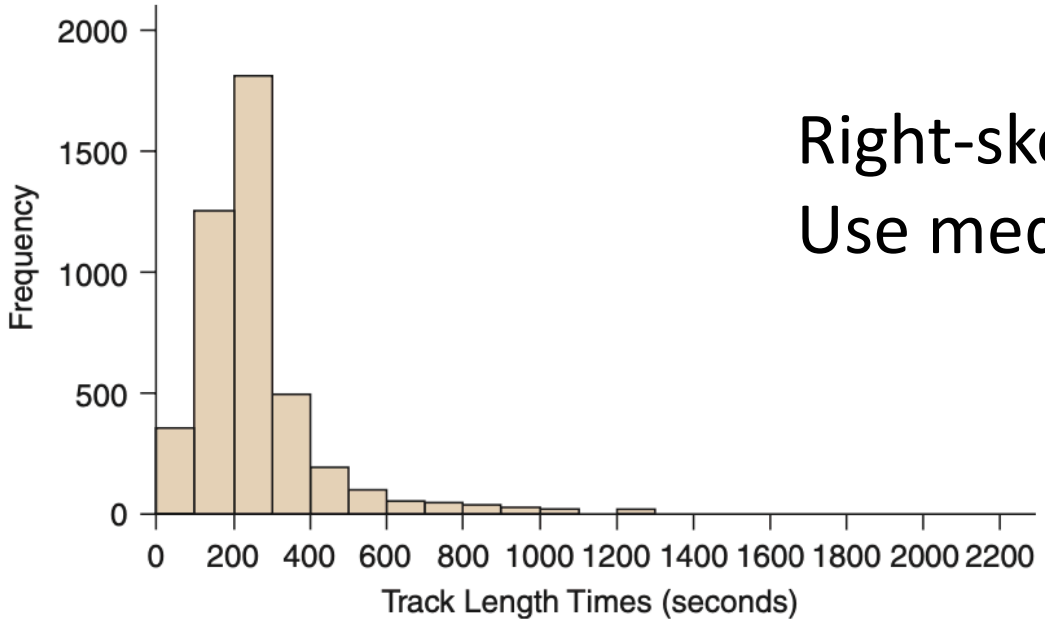
Mean vs Median

Shape	Measure for Center	Measure for Spread
Skewed	Median	IQR
Symmetric	Mean or Median	STD or IQR

- Median is more resistant to outliers. Outliers can appear even in symmetric distribution.
- Mean is sensitive to all values of the data. Use it you need a measure for center that is sensitive to all data values.
- Mean is mathematically more useful because you can compute other mathematical quantities through the mean, such as variance.

How to choose a suitable measure?

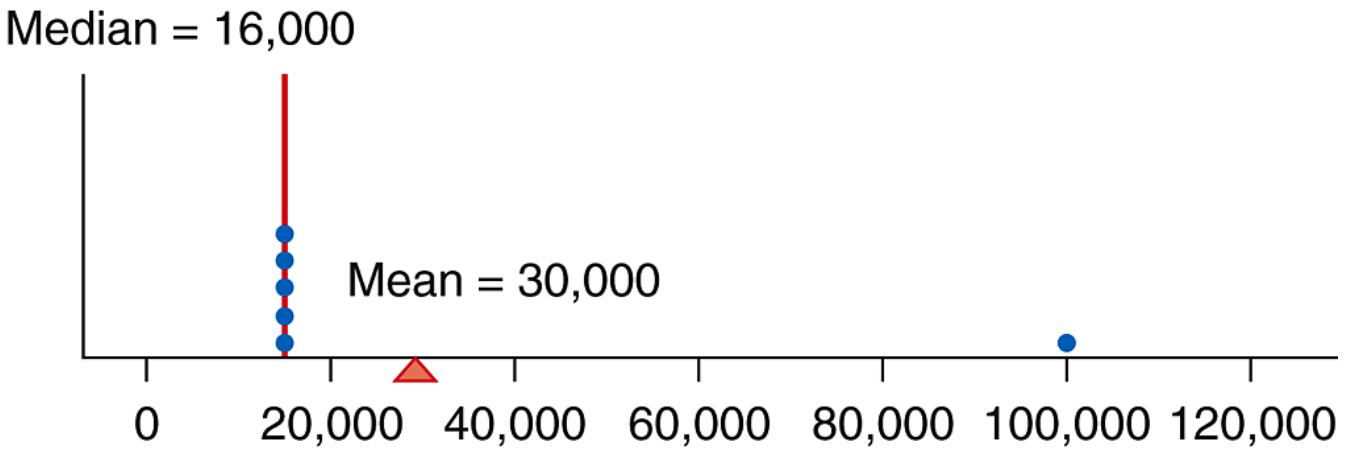
A person created a data set of the music tracks in his digital library. He wants to describe the distribution of song lengths.



Right-skewed!
Use median and IQR

How to choose a suitable measure?

A fast-food restaurant has five employees. Each employee's annual income is about $16,000 per year. The owner, on the other hand, makes $100,000 per year.



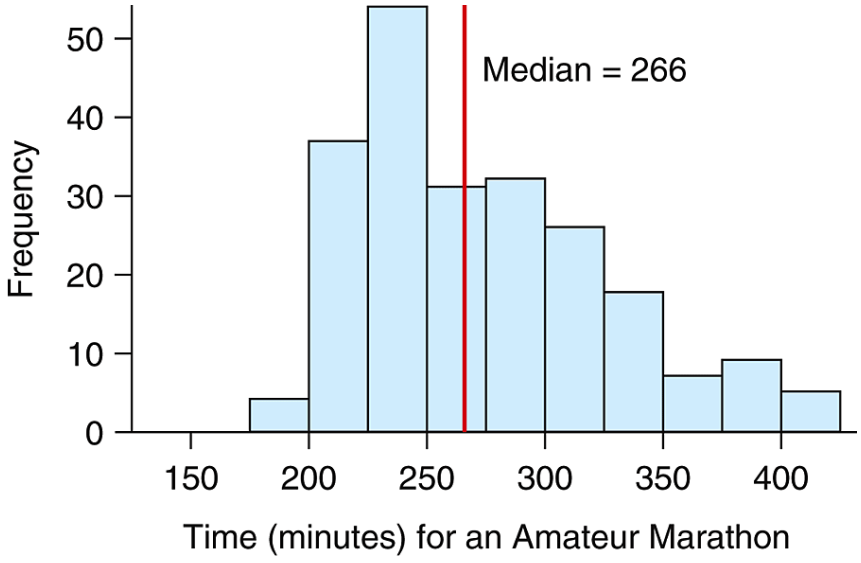
Comparing Different Distributions

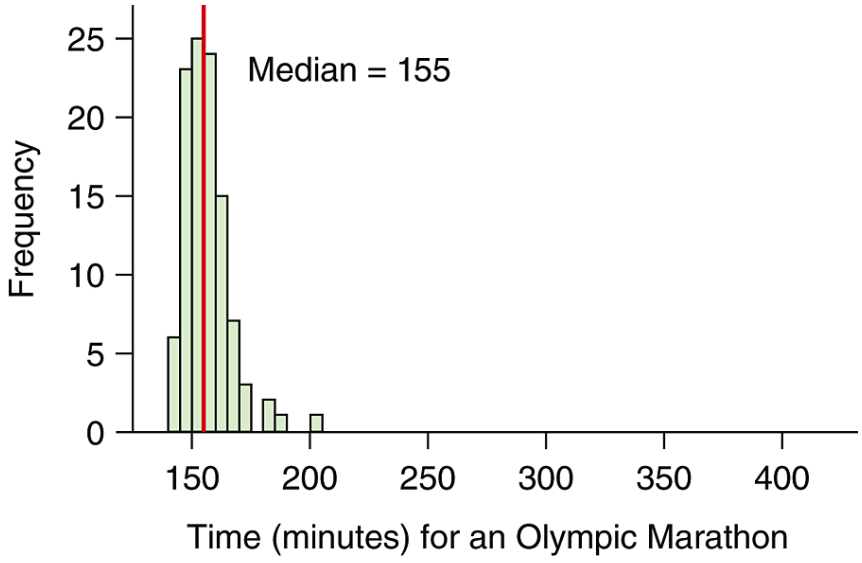
When comparing two distributions:

- Always use the same measures of center and spread for both distributions. Otherwise, the comparison is not valid.
- If one of the distributions is skewed, use Median and IQR to compare both distributions!

Example

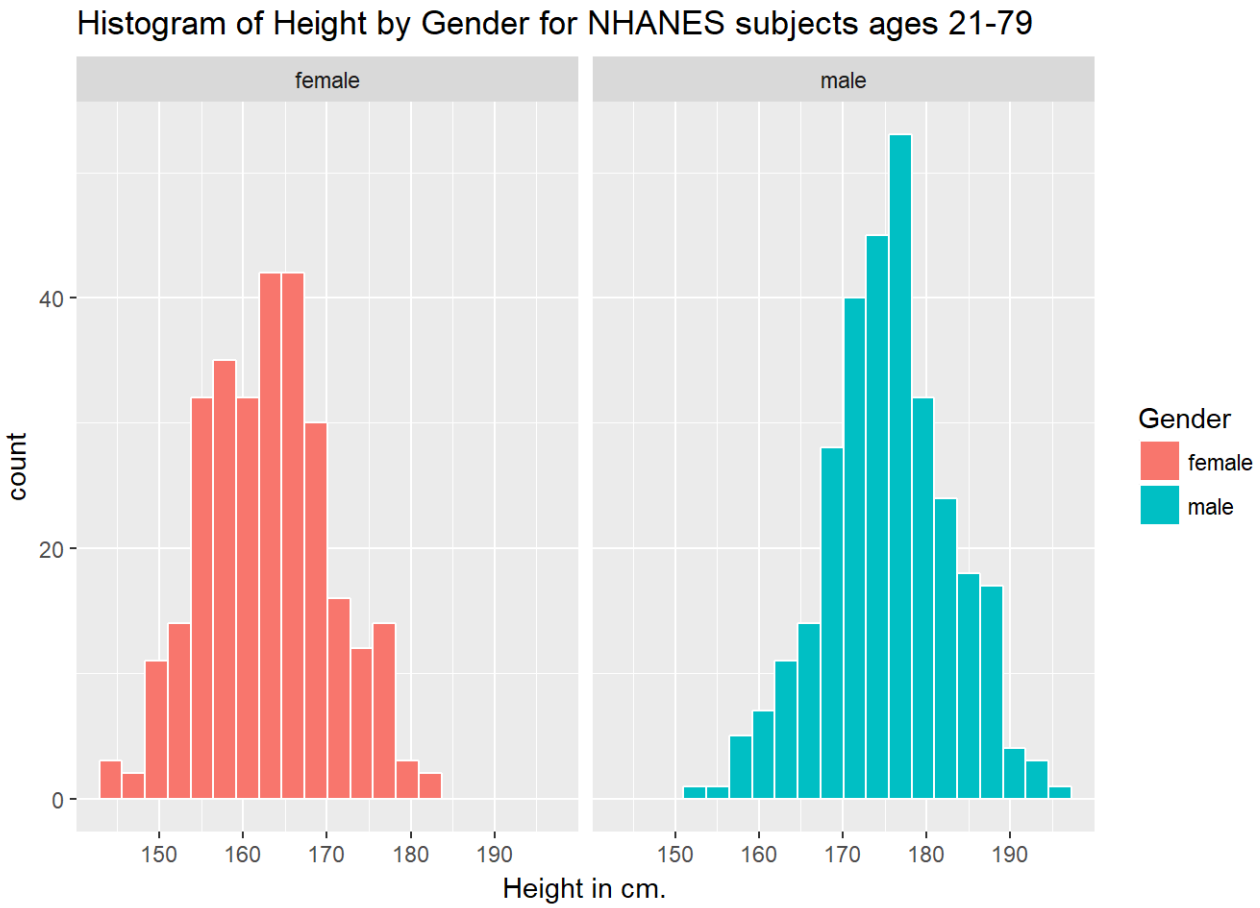
Comparing the distributions of running times for amateur and Olympic marathon runners is below.





Example

Compare the height distributions of female and male.



Section 3.5
Use boxplot to display summaries

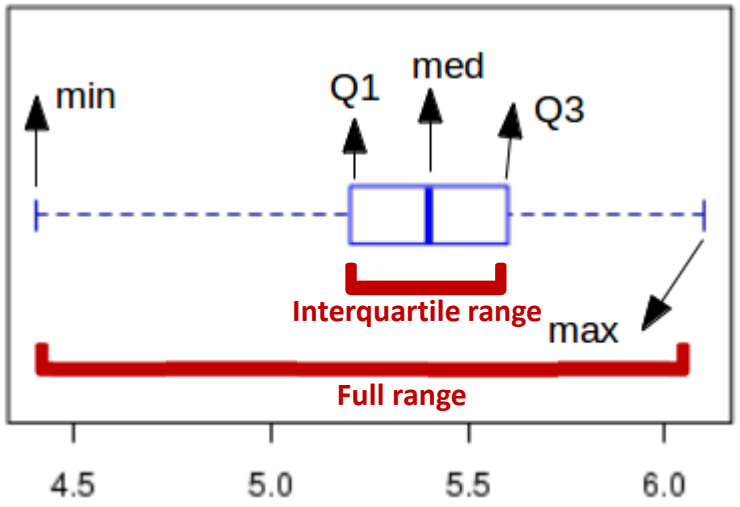
- In a symmetric distribution, the mean and the median are approximately the same.
- In a right-skewed distribution, the mean tends to be greater than the median.
- In a left-skewed distribution, the mean tends to be less than the median.

Creating Box Plots

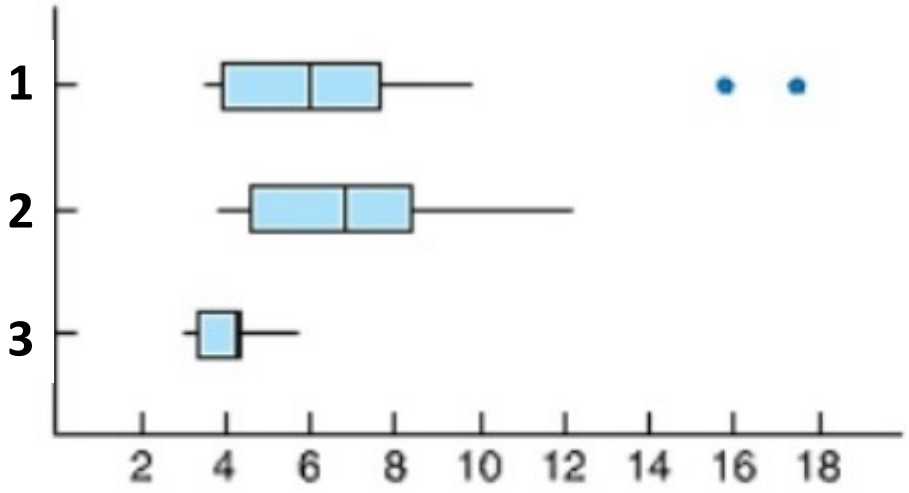
- The 5-number summary:

  Minimum, Q1, Median, Q3, Maximum

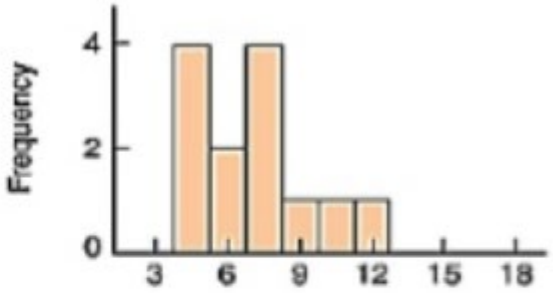
- Use the 5-number summary and a number line to create a boxplot.



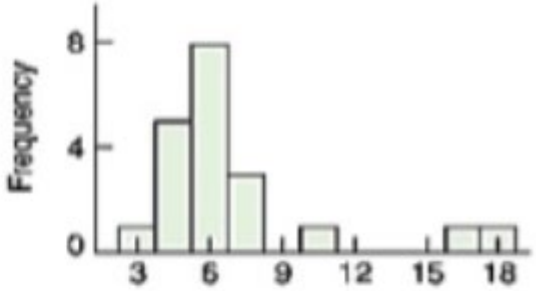
Match each histogram with its boxplot.



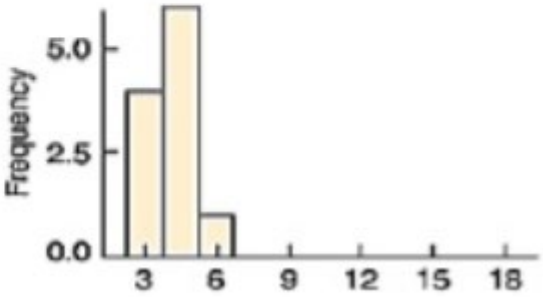
X



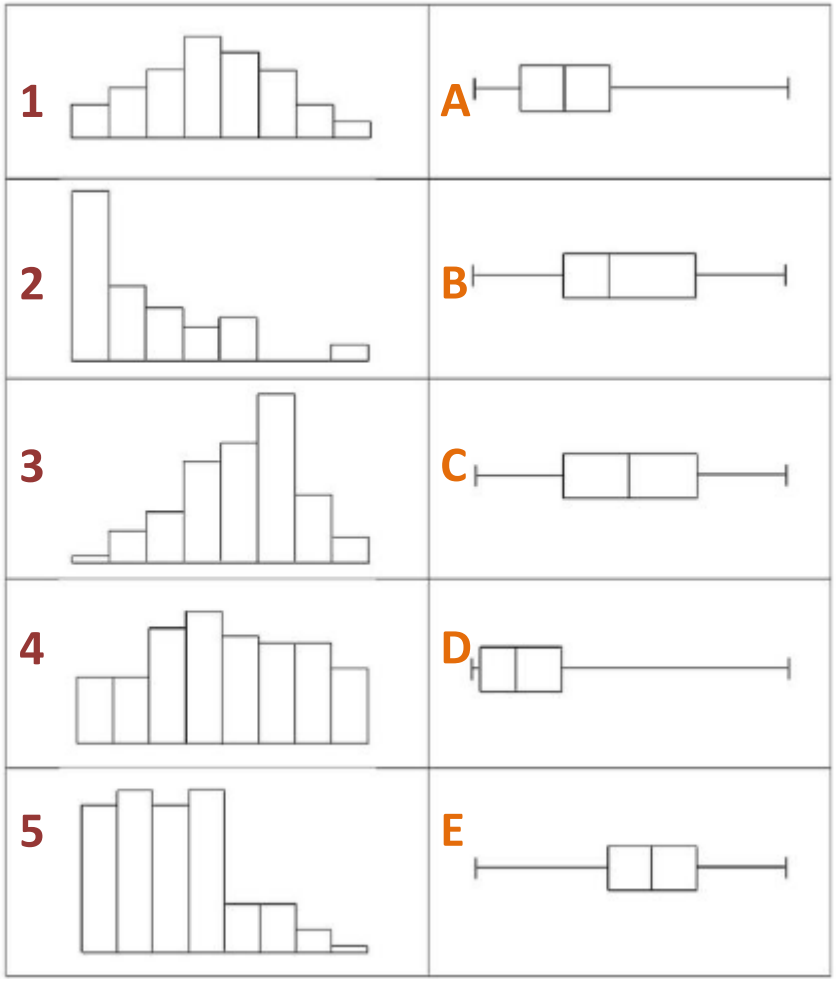
Y



Z



Match each histogram to the correct boxplot



End of 3.5

Finding Outliers

Outliers - Extreme data values

General Rule for finding outliers:

- Find the **fences** (“cutoffs”) for usual data values:

$$\text{Lower fence} = Q1 - 1.5(IQR)$$

$$\text{Upper fence} = Q3 + 1.5(IQR)$$

- Values more extreme than the fences are outliers (values less than lower fence or greater than upper fence).

Finding Outliers: Example

The first and third quartiles in the distribution of daily high temperatures in San Francisco are $59°\text{F}$ and $70°\text{F}$, respectively. Using these values, what temperatures would be considered outliers in San Francisco?

- Fences: $\text{Lower fence} = 59 - 1.5(70 - 59) = 42.5°\text{F}$

  $\text{Upper fence} = 70 + 1.5(70 - 59) = 86.5°\text{F}$

- Outliers: Any temperature below $42.5°\text{F}$ or above $86.5°\text{F}$ would be considered an outlier.