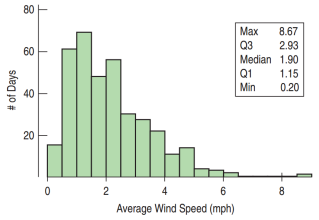
Lesson 09: In what ways, can we compare distributions and look for trends and patterns in centers and spreads?

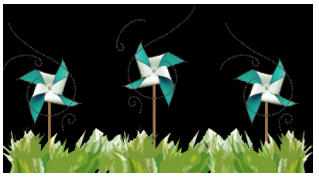
Q: Describe the distribution.



A histogram of daily Average Wind Speed *for 1989.*

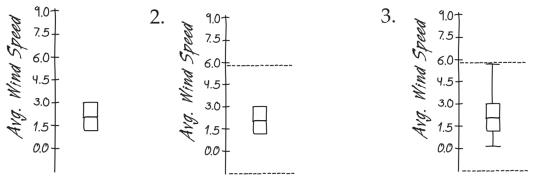
Let's start with the "big picture." Here's a histogram and 5-number summary of the `Average Wind Speed` for every day in 1989. Because of the skewness, we'll report the median and IQR. We can see that the distribution of `Average Wind Speed` is unimodal and skewed to the right. Median daily wind speed is about 1.90 mph, and on half of the days, the average wind speed is between 1.15 and 2.93 mph. We also see a rather windy 8.67-mph day. Was that unusually windy or just the windiest day of the year? To answer that, we'll need to work with the summaries a bit more.

WHO	Days during 1989
WHAT	Average daily wind speed (mph), Average barometric pressure (mb), Average daily temperature (deg Celsius)
WHEN	1989
WHERE	Hopkins Forest, in Western Massachusetts
WHY	Long-term observations to study ecology and climate



Box plots and 5 number summaries

1. 2. 3.



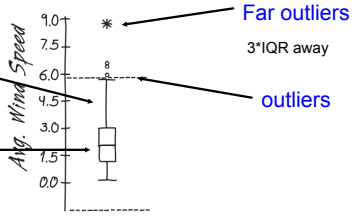
What was our standard to call a data value an outlier?

longer tail indicating skewness

Median is roughly centered

middle half of the data is symmetric

4.

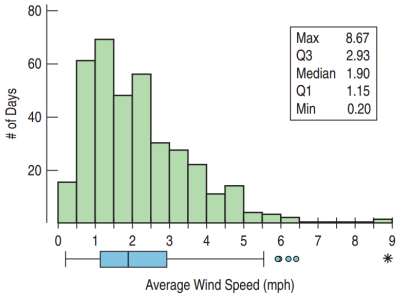


Far outliers

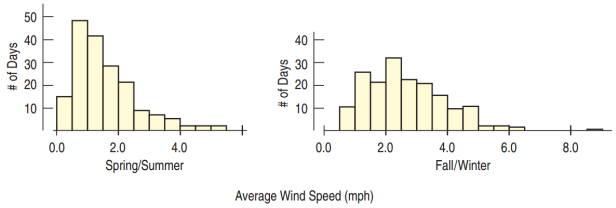
3*IQR away

outliers

By turning the boxplot and putting it on the same scale as the histogram, we can compare both displays of the daily wind speeds and see how each represents the distribution.



Comparing groups with histograms



Is it windier in the winter or the summer?

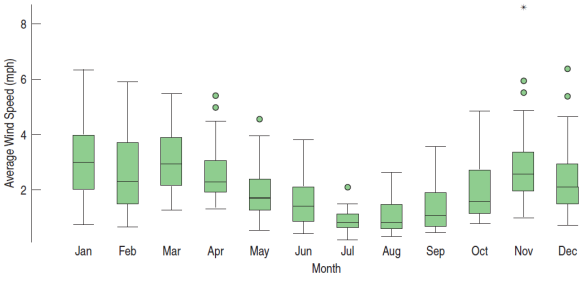
Summaries for *Average Wind Speed* by Season

Group	Mean	StdDev	Median	IQR
Fall/Winter	2.71	1.36	2.47	1.87
Spring/Summer	1.56	1.01	1.34	1.32

Center and Spread

Center and Spread

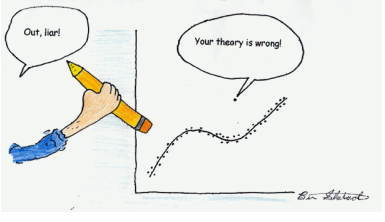
Are some months windier than others?



Why not histograms or stem and leaf plots?

Boxplots of the average daily wind speed for each month show seasonal patterns in both the centers and spreads.

When we looked at a boxplot of wind speeds for the entire year, there were only 5 outliers. Now, when we group the days by `Month`, the boxplots display more days as outliers and call out one in November as a far outlier. The boxplots show different outliers than before because some days that seemed ordinary when placed against the entire year's data looked like outliers for the month that they're in. That windy day in July certainly wouldn't stand out in November or December, but for July, it was remarkable.



Don't treat them like liars!

What should we do with outliers?

Try to understand them in the context of data

- Look at the gap between the case and the rest of the data.
- May be an error...correct it if you can.
- Or may be a correct but extraordinary value...worth investigating.
- You may have to report summaries with and without the outlier, so the reader may judge by himself or herself the validity of the data.
  - > If there are any clear outliers and you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be quite revealing.
- Note: The median and IQR are not likely to be affected by the outliers.

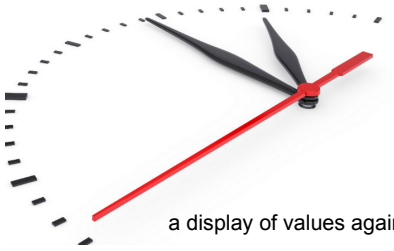
What should you NOT do with outliers?

- We should not ignore outliers
- if you decide to exclude the outlier, you must justify your decision.

WIND, SNOW, COLD GIVE N.E. A TASTE OF WINTER

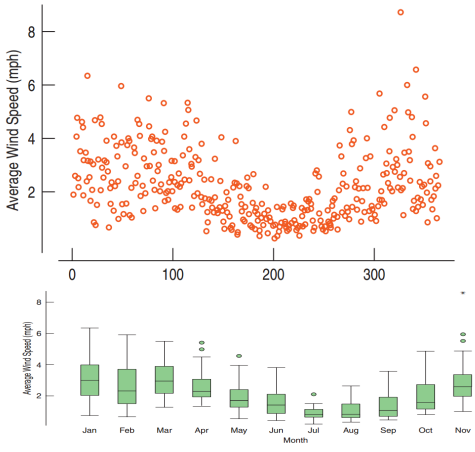
Published on November 22, 1989
Author: Andrew Dabilis, Globe Staff

An intense storm roared like the Montreal Express through New England yesterday, bringing frigid winds of up to 55 m.p.h., 2 feet of snow in some parts of Vermont and a preview of winter after weeks of mild weather. Residents throughout the region awoke yesterday to an icy vortex that lifted an airplane off the runway in Newark and made driving dangerous in New England because of rapidly shifting winds that seemed to come from all directions.



Timeplot

a display of values against time



A timeplot of Average Wind Speed *shows the overall pattern and changes in variation.*

Looking into the future

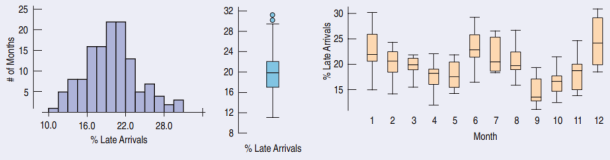
Using Hopkins forest data, you can safely say next June will be like the last.

But we certainly wouldn't predict another storm based on singular November day.

Don't assume a trend will continue.

Just checking

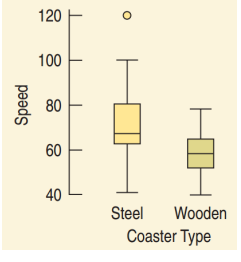
The Bureau of Transportation Statistics of the U.S. Department of Transportation collects and publishes statistics on airline travel (www.transtats.bts.gov). Here are three displays of the % of flights arriving late each month from 1995 through 2005:



1. Describe what the histogram says about late arrivals.
2. What does the boxplot of late arrivals suggest that you can't see in the histogram?
3. Describe the patterns shown in the boxplots by month. At what time of year are flights least likely to be late? Can you suggest reasons for this pattern?

Roller coasters are a thrill ride in many amusement parks worldwide. And thrill seekers want a coaster that goes fast. There are two main types of roller coasters: those with wooden tracks and those with steel tracks. Do they typically run at different speeds? Here are boxplots:

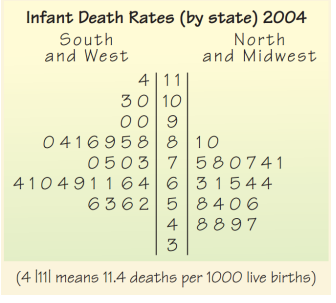
Question: Compare the speeds of wood and steel roller coasters.



Answer

Overall, wooden-track roller coasters are slower than steel-track coasters. In fact, the fastest half of the steel coasters are faster than three quarters of the wooden coasters. Although the IQRs of the two groups are similar, the range of speeds among steel coasters is larger than the range for wooden coasters. The distribution of speeds of wooden coasters appears to be roughly symmetric, but the speeds of the steel coasters are skewed to the right, and there is a high outlier at 120 mph. We should look into why that steel coaster is so fast.

In 2004 the infant death rate in the United States was 6.8 deaths per 1000 live births. The Kaiser Family Foundation collected data from all 50 states and the District of Columbia, allowing us to look at different regions of the country. Since there are only 51 data values, a back-to-back stem-andleaf plot is an effective display. Here's one comparing infant death rates in the Northeast and Midwest to those in the South and West. In this display the stems run down the middle of the plot, with the leaves for the two regions to the left or right. Be careful when you read the values on the left:4|11| means a rate of 11.4 deaths per 1000 live birth for one of the southern or western states.



How do infant death rates compare for these regions?

In general, infant death rates were generally higher for states in the South and West than in the Northeast and Midwest. The distribution for the northeastern and midwestern states is roughly uniform, varying from a low of 4.8 to a high of 8.1 deaths per 1000 live births. Ten southern and western states had higher infant death rates than any in the Northeast or Midwest, with one state over 11. Rates varied more widely in the South and West, where the distribution is skewed to the right and possibly bimodal. We should investigate further to see which states represent the cluster of high death rates.