21 INDIANS WINS IN A ROW
(AL RECORD)

STATS DURING STREAK

104 RUN DIFFERENTIAL

41 HOME RUNS

7 SHUTOUTS

1.79 TEAM ERA

NEXT UP:
1916 GIANTS
(26 STRAIGHT WINS)
A handful of statistics that illustrate their dominance during this historic run:

- 105 – Run differential - they've outscored their opponents 142-37.
- 41 – Home runs. Francisco Lindor leads with nine.
- 7 – Shutouts.
- 1.58 – Team ERA.
- 9 – Innings trailed, out of 199 innings.
- 19 – Wins from the starting rotation. Trevor Bauer and Corey Kluber each have four.
- 200-37 – Strikeout-walk ratio.

September 13, 2017
<table>
<thead>
<tr>
<th>University</th>
<th>City, State</th>
<th>#1 in National Universities</th>
<th>#2 in National Universities</th>
<th>#3 in National Universities (tie)</th>
<th>#4 in National Universities</th>
<th>#5 in National Universities (tie)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Princeton University</td>
<td>Princeton, NJ</td>
<td>$47,140</td>
<td>5,400</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harvard University</td>
<td>Cambridge, MA</td>
<td>$48,949</td>
<td>6,710</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University of Chicago</td>
<td>Chicago, IL</td>
<td>$54,825</td>
<td>5,941</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yale University</td>
<td>New Haven, CT</td>
<td>$51,400</td>
<td>5,472</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Columbia University</td>
<td>New York, NY</td>
<td>$57,208</td>
<td>6,113</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Princeton University</td>
<td>Massachusetts Institute of Technology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>---------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Princeton, NJ</td>
<td>Cambridge, MA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$47,140</td>
<td>$49,892</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#1 in National Universities</td>
<td>#5 in National Universities (tie)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Harvard University</th>
<th>Stanford University</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cambridge, MA</td>
<td>Stanford, CA</td>
</tr>
<tr>
<td>$48,949</td>
<td>$49,617</td>
</tr>
<tr>
<td>#2 in National Universities</td>
<td>#5 in National Universities (tie)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>University of Chicago</th>
<th>University of Pennsylvania</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chicago, IL</td>
<td>Philadelphia, PA</td>
</tr>
<tr>
<td>$54,825</td>
<td>$53,534</td>
</tr>
<tr>
<td>#3 in National Universities (tie)</td>
<td>#8 in National Universities</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Yale University</th>
<th>Duke University</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Haven, CT</td>
<td>Durham, NC</td>
</tr>
<tr>
<td>$51,400</td>
<td>$53,744</td>
</tr>
<tr>
<td>#3 in National Universities (tie)</td>
<td>#9 in National Universities</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Columbia University</th>
<th>California Institute of Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York, NY</td>
<td>Pasadena, CA</td>
</tr>
<tr>
<td>$57,208</td>
<td>$49,908</td>
</tr>
<tr>
<td>#5 in National Universities (tie)</td>
<td>#10 in National Universities</td>
</tr>
<tr>
<td>University</td>
<td>City, State</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Princeton University</td>
<td>Princeton, NJ</td>
</tr>
<tr>
<td>Harvard University</td>
<td>Cambridge, MA</td>
</tr>
<tr>
<td>University of Chicago</td>
<td>Chicago, IL</td>
</tr>
<tr>
<td>Yale University</td>
<td>New Haven, CT</td>
</tr>
<tr>
<td>Columbia University</td>
<td>New York, NY</td>
</tr>
<tr>
<td>Dartmouth College</td>
<td>Hanover, NH</td>
</tr>
<tr>
<td>Northwestern University</td>
<td>Evanston, IL</td>
</tr>
<tr>
<td>Cornell University</td>
<td>Ithaca, NY</td>
</tr>
</tbody>
</table>
Measures of Central Tendency

Mean, Median, and Mode
Population study

- The population under study is the 65 students in our class.
- We asked you how many languages you speak (besides English).
- Here were your responses:

```
0 1 0 0 1 1 2 1 4 1 2 1 2 2 1 2 0 1 1 2 1 0 2 0 1 0 0 1
1 2 0 1 1 1 1 0 2 1 1 1 0 1 1 2 0 1 2 0 1 0 1 1 1 1 1 1 0 2 0 1
```

- Each response is a measurement, or an outcome
- In this lecture, we will discuss ways to summarize data from a sample, using as an example these made-up responses
Frequency distribution

- A count of the number of times each outcome occurs is called a frequency distribution.
- This information can be conveyed in a table or a plot.
- A plot of a frequency distribution of numerical data is called a histogram.

<table>
<thead>
<tr>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>
The **mean** of a frequency distribution

- Recall our dataset

  0 1 0 0 1 1 0 0 1 1 2 1 4 1 2 1 2 2 1 2 0 1 1 2 1 0 2 0 1 0 0 1
  1 2 0 1 1 1 1 0 2 1 1 1 0 1 1 2 0 1 2 0 1 0 1 1 1 1 1 1 0 2 0 1

- The **mean** (sum/sample size) of these numbers is 0.95
- Note that no student reports knowing 0.95 languages
- Still, this number is one way of summarizing the frequency distribution:

  “On average, the students in this class know 0.95 languages (beyond English).”
Another way to calculate the mean

- Use the frequency distribution
- Calculate the sum of the data points by first multiplying the outcome and frequency columns together, and then adding up the results

  \[ \text{sum} = (0 \times 18) + (1 \times 34) + (2 \times 12) + (4 \times 1) = 0 + 34 + 24 + 4 = 62 \]

- The mean is then the sum divided by the sample size (65)

  \[ \text{mean} = \frac{\text{sum}}{\text{sample size}} = \frac{62}{65} \approx 0.95 \]
Yet another way to calculate the mean

- We could also calculate the mean by multiplying the outcome column by relative frequencies (frequency/sample size), and adding up the results

\[
\text{mean} = (0 \times 0.28) + (1 \times 0.52) + (2 \times 0.18) + (3 \times 0) + (4 \times 0.02) \approx 0.95
\]

- The mean of a sample is the weighted average of the distinct outcomes
- The weights are the relative frequencies with which those outcomes occur
But the mean is not perfect
(No descriptive statistic ever is)

Example:

- A ten person first-year seminar has 9 first-year students who are 18 years old, and one who is 45
- The mean student age is 20.7
- Yet almost no one in the class is even 20!
- The mean is not an ideal way to summarize data in this case
The **median** of a frequency distribution

- The median is the “middle” of a frequency distribution

- Assume a sample: 12, 5, 3, 4, 5
- Sorting these data gives: 3, 4, 5, 5, 12
- The median value of this sample is 5

- If the sample size is even: e.g., 2, 3, 4, 5, 5, 12
- The median is the mean of the middle two numbers
- \((4 + 5)/2 = 4.5\)

- Q: What is the median age in the aforementioned first-year seminar?
The median is the “middle” point of a distribution

- Consider another sample: 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4
- The median is the outcome that divides the distribution of outcomes in half
- 50% of the outcomes are at or below the median, and 50% are at or above it
- The area in a histogram to the right of the median equals the area to its left

Blue Area = Red Area

\[6 \times 1 + 8 \times 0.5 = 8 \times 0.5 + 3 \times 1 + 3 \times 1\]

10 = 10
The mean is the “balance” point of a distribution

- You can think of the mean as the distribution’s center of gravity
- The sum total of the distances of all the points less than the mean equals the sum total of the distances of all the points greater

\[(2.15-1) \times 6 + (2.15-2) \times 8 = (3-2.15) \times 3 + (4-2.15) \times 3\]
\[(1.15) \times 6 + (.15) \times 8 = (.85) \times 3 + (1.85) \times 3\]
\[8.1 = 8.1\]
iClicker Q: The mean vs. the median

If a student scores in the top half (upper 50%) of the class on an exam, did the student score above the mean?

A: Yes, the student scored above the mean

B: No, the student scored below the mean

C: Not sure; there isn’t enough information to answer this question

D: Not sure, because I don’t understand the mean and median well enough yet to even hazard a guess (but I’m interested in learning this stuff!)
C (or D): Not Enough Information

- The student is not necessarily above (or below) the mean
- The **mean** is the “balance” point of a distribution
- The **median** is the “middle” point of a distribution
- The order of these two points can vary with the distribution
Symmetric (i.e., non-skewed) distributions

- In a symmetric distribution, the left mirrors the right
- So the mean (the balance point) equals the median (the middle)
- E.g., the mean and the median of the frequency distribution for the sample 1, 2, 2, 3 are both 2
Asymmetric (i.e., skewed) distributions

- In an asymmetric distribution, the left does not mirror the right
- So the mean and the median are not necessarily equal
- E.g., the mean and the median of the frequency distribution for the sample 1, 2, 2, 10 are not equal

<table>
<thead>
<tr>
<th>outcome</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
</tr>
</tbody>
</table>

median = 2  mean = 3.75
Mean vs. Median

- The gold sample has an outlier: 10
- So its histogram is skewed right
- Likewise, its balance point (the mean) falls to the right of the middle point (the median)
- Ultimately, only 25% of the outcomes are above the mean

<table>
<thead>
<tr>
<th>outcome</th>
<th>blue_frequency</th>
<th>gold_frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
\text{mean} = \text{median} = 2 \\
\text{median} = 2 \quad \text{mean} = 3.75
\]
iClicker Q: The mean vs. the median

If a student scores in the top half (upper 50%) of the class on an exam, did the student score above the mean?

A: Yes, the student scored above the mean

B: No, the student scored below the mean

C: Not sure; there isn’t enough information to answer this question

D: Not sure, because I don’t understand the mean and median well enough yet to even hazard a guess (but I’m interested in learning this stuff!)
C: Not Enough Information

- A student who scores in the upper 50% might have scored 10, but they might have scored 2.
- It’s impossible to tell if they scored above the mean or not!

<table>
<thead>
<tr>
<th>outcome</th>
<th>blue_frequency</th>
<th>gold_frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

mean = median = 2
median = 2  mean = 3.75
A real-world example of a skewed distribution

- Delay times are in minutes
- A negative observation means the flight left earlier than scheduled
- The histogram is right skewed
- It has a long right-hand tail
- The mean is being pulled away from the median in the direction of the tail, so we expect the mean to be greater than the median:

\[
\text{mean} = 9.135913 \quad \text{median} = -2
\]
Another example of a skewed real-world distribution: Public school funding
Mean vs. median, rule-of-thumb
Bush tax cuts

● GOP reported 92 million Americans would get a tax cut, averaging $1,083
● But actually, the median was less than $100
● Incredibly rich outliers received much larger tax cuts and skewed the mean
Bernie’s Sanders Talking Points

- Average contribution is $27
- Distribution must have been right-skewed
- So isn’t the median smaller still?
- I’m guessing it wasn’t
Rule of Thumb

- For skewed distributions, the mean lies in the direction of the skew (towards the longer tail) relative to the median
- But this rule of thumb is not always true!
Mode

- The mode is the most popular outcome; the one that occurs most often.
- The mode is useful for summarizing categorical data. In the distribution of letters in English text, the mode is e.
- In plurality voting, the winning candidate is the mode; the one with the most votes.
- The mode need not be unique; some distributions are bi- or multi-modal.
Measures of Central Tendency

- Mean: numerical, preferably symmetric data
- Median: ordered, possibly skewed data
- Mode: any data, but especially non-numeric data
  - e.g., hair color, food items, movies, etc.