THE OHIO STATE UNIVERSITY

Connecting Math and Statistics with GAISE

Bradford R. Findell, PhD

Columbus Reimagine Conference

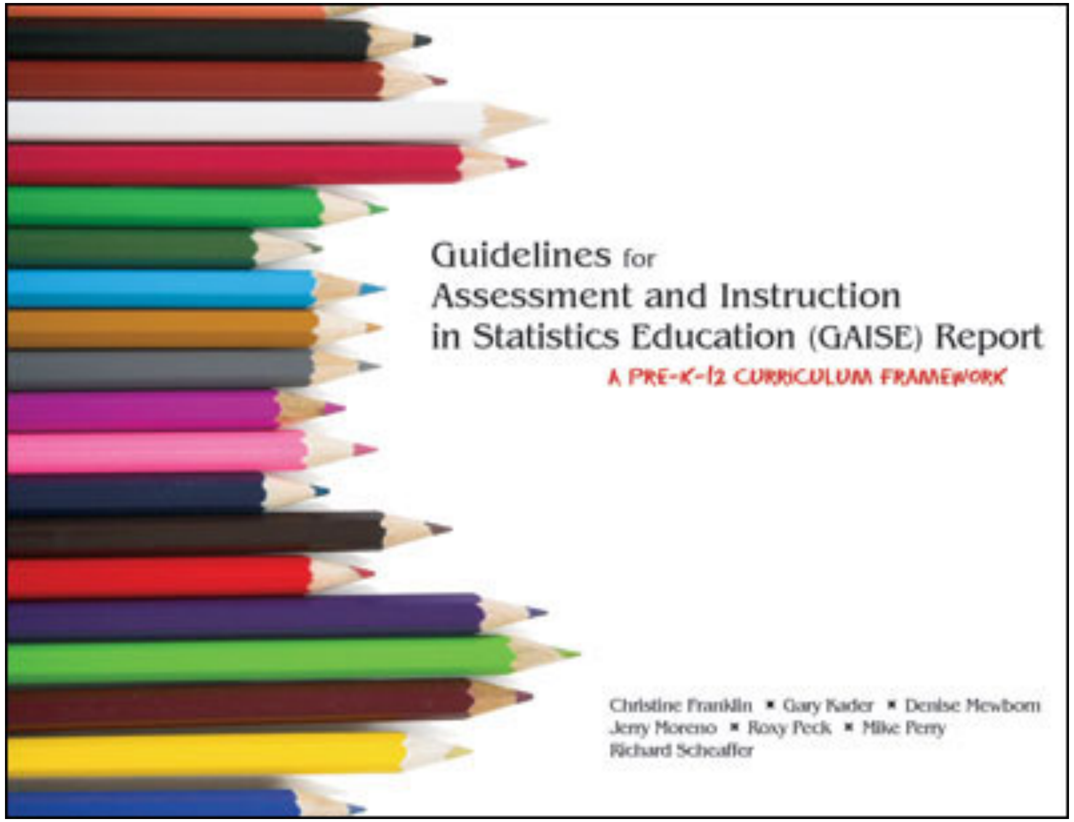
June 14, 2017

Slides will be available at
http://bradfindell.com

The GAISE Model

- Statistical problem solving is a four-step process:
  1. Formulate questions
  2. Collect data
  3. Analyze data
  4. Interpret results

- from the American Statistical Association (ASA, 2005)
  - Full report



Why does this matter?

2017 revision of Ohio's Learning Standards for Mathematics

GAISE Model in Grade 6

Original Standards	Revised Standards
Develop understanding of statistical variability. **6.SP.1** Recognize a statistical question as one that anticipates variability in the data related to the question and accounts for it in the answers. *For example, "How old am I?" is not a statistical question, but "How old are the students in my school?" is a statistical question because one anticipates variability in students' ages.*	**Develop understanding of statistical problem solving.** **6.SP.1** Develop statistical reasoning by using the GAISE model: **a.** Formulate Questions: Recognize and formulate a statistical question as one that anticipates variability and can be answered with quantitative data. *For example, "How old am I?" is not a statistical question, but "How old are the students in my school?" is a statistical question because of the variability in students' ages.* (GAISE Model, step 1) **b.** Collect Data: Design and use a plan to collect appropriate data to answer a statistical question. (GAISE Model, step 2) **c.** Analyze Data: Select appropriate graphical methods and numerical measures to analyze data by displaying variability within a group, comparing individual to individual, and comparing individual to group. (GAISE Model, step 3) **d.** Interpret Results: Draw logical conclusions from the data based on the original question. (GAISE Model, step 4)

Reiterate GAISE Model in Grade 7

Original Standards	Revised Standards
Use random sampling to draw inferences about a population. **7.SP.2** Use data from a random sample to draw inferences about a population with an unknown characteristic of interest. Generate multiple samples (or simulated samples) of the same size to gauge the variation in estimates or predictions. *For example, estimate the mean word length in a book by randomly sampling words from the book; predict the winner of a school election based on randomly sampled survey data. Gauge how far off the estimate or prediction might be.*	**Broaden understanding of statistical problem solving.** **7.SP.2** Broaden statistical reasoning by using the GAISE model. **a.** Formulate Questions: Recognize and formulate a statistical question as one that anticipates variability and can be answered with quantitative data. *For example, "How do the heights of seventh graders compare to the heights of eighth graders?"* (GAISE Model, step 1) **b.** Collect Data: Design and use a plan to collect appropriate data to answer a statistical question. (GAISE Model, step 2) **c.** Analyze Data: Select appropriate graphical methods and numerical measures to analyze data by displaying variability within a group, comparing individual to individual, and comparing individual to group. (GAISE Model, step 3) **d.** Interpret Results: Draw logical conclusions and make generalizations from the data based on the original question. (GAISE Model, step 4)

Remove MAD from Grade 6

Original Standards	Revised Standards
Summarize and describe distributions. **6.SP.5** Summarize numerical data sets in relation to their context, such as by: **a.** Reporting the number of observations. **b.** Describing the nature of the attribute under investigation, including how it was measured and its units of measurement. **c.** Giving quantitative measures of center (median and/or mean) and variability (interquartile range and/or mean absolute deviation), as well as describing any overall pattern and any striking deviations from the overall pattern with reference to the context in which the data were gathered. **d.** Relating the choice of measures of center and variability to the shape of the data distribution and the context in which the data were gathered.	**Summarize and describe distributions.** **6.SP.5** Summarize numerical data sets in relation to their context. **a.** Report the number of observations. **b.** Describe the nature of the attribute under investigation, including how it was measured and its units of measurement. **c.** Find the quantitative measures of center (median and/or mean) for a numerical data set and recognize that this value summarizes the data set with a single number. Interpret mean as an equal or fair share. Find measures of variability (range and interquartile range) as well as informally describe the shape and the presence of clusters, gaps, peaks, and outliers in a distribution. **d.** Choose the measures of center and variability, based on the shape of the data distribution and the context in which the data were gathered.

Include MAD in Grade 7

Original Standards	Revised Standards
Draw informal comparative inferences about two populations. **7.SP.3** Informally assess the degree of visual overlap of two numerical data distributions with similar variabilities, measuring the difference between the centers by expressing it as a multiple of a measure of variability. *For example, the mean height of players on the basketball team is 10 cm greater than the mean height of players on the soccer team, about twice the variability (mean absolute deviation) on either team; on a dot plot, the separation between the two distributions of heights is noticeable.*	**Summarize and describe distributions representing one population and draw informal comparisons between two populations.** **7.SP.3** Describe and analyze distributions. **a.** Summarize quantitative data sets in relation to their context by using mean absolute deviation (MAD), interpreting mean as a balance point. **b.** Informally assess the degree of visual overlap of two numerical data distributions with roughly equal variabilities, measuring the difference between the centers by expressing it as a multiple of a measure of variability. *For example, the mean height of players on the basketball team is 10 cm greater than the mean height of players on the soccer team, about twice the variability (mean absolute deviation) on either team; on a dot plot (line plot), the separation between the two distributions of heights is noticeable.*

Move Sampling and Inference to HS

Original Standards	Revised Standards
Use random sampling to draw inferences about a population. **7.SP.1** Understand that statistics can be used to gain information about a population by examining a sample of the population; generalizations about a population from a sample are valid only if the sample is representative of that population. Understand that random sampling tends to produce representative samples and support valid inferences.	**Use sampling to draw conclusions about a population.** **7.SP. 1** Understand that statistics can be used to gain information about a population by examining a sample of the population. **a.** Differentiate between a sample and a population. **b.** Understand that conclusions and generalizations about a population are valid only if the sample is representative of that population. Develop an informal understanding of bias.
Draw informal comparative inferences about two populations. **7.SP.4** Use measures of center and measures of variability for numerical data from random samples to draw informal comparative inferences about two populations. *For example, decide whether the words in a chapter of a seventh-grade science book are generally longer than the words in a chapter of a fourth-grade science book.*	Deleted Standard

Reiterate the GAISE Model in Grade 8

Original Standards	Revised Standards
Investigate patterns of association in bivariate data. **8.SP.1** Construct and interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association.	**Investigate patterns of association in bivariate data.** **8.SP.1** Construct and interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantities. Describe patterns such as clustering; outliers; positive, negative, or no association; and linear association and nonlinear association. (GAISE Model, steps 3 and 4)
Investigate patterns of association in bivariate data. **8.SP.2** Know that straight lines are widely used to model relationships between two quantitative variables. For scatter plots that suggest a linear association, informally fit a straight line, and informally assess the model fit by judging the closeness of the data points to the line.	**Investigate patterns of association in bivariate data.** **8.SP.2** Understand that straight lines are widely used to model relationships between two quantitative variables. For scatter plots that suggest a linear association, informally fit a straight line, and informally assess the model fit by judging the closeness of the data points to the line. (GAISE Model, steps 3 and 4)
Investigate patterns of association in bivariate data. **8.SP.3** Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept. *For example, in a linear model for a biology experiment, interpret a slope of 1.5 cm/hr as meaning that an additional hour of sunlight each day is associated with an additional 1.5 cm in mature plant height.*	**Investigate patterns of association in bivariate data.** **8.SP.3** Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept. *For example, in a linear model for a biology experiment, interpret a slope of 1.5 cm/hr as meaning that an additional hour of sunlight each day is associated with an additional 1.5 cm in mature plant height.* (GAISE Model, steps 3 and 4)

Reiterate the GAISE Model in High School

Original Standards	Revised Standards
Summarize, represent, and interpret data on a single count or measurement variable. **S.ID.1** Represent data with plots on the real number line (dot plots, histograms, and box plots).*	**Summarize, represent, and interpret data on a single count or measurement variable.** **S.ID.1** Represent data with plots on the real number line (dot plots, histograms, and box plots) in the context of real-world applications using the GAISE model.*
Summarize, represent, and interpret data on a single count or measurement variable. **S.ID.2** Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets.*	**Summarize, represent, and interpret data on a single count or measurement variable.** **S.ID.2** In the context of real-world applications by using the GAISE model, use statistics appropriate to the shape of the data distribution to compare center (median and mean) and spread (mean absolute deviation, interquartile range, and standard deviation) of two or more different data sets. *
Summarize, represent, and interpret data on a single count or measurement variable. **S.ID.3** Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).*	**Summarize, represent, and interpret data on a single count or measurement variable.** **S.ID.3** In the context of real-world applications by using the GAISE model, interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers). *

Let’s Do Some Statistics!

Literary Computing

- “The authorship of several literary works is often a topic for debate. Were some of the works attributed to William Shakespeare actually written by Francis Bacon or Christopher Marlowe? Which of the anonymously published Federalist Papers were written by Alexander Hamilton, which by James Madison, which by John Jay? Who were the authors of the writings contained in the Bible? The field of “literary computing” began to find ways of numerically analyzing authors’ works, looking at variables such as sentence length and rates of occurrence of specific words.” (Rossman and Chance, 2003)

Gettysburg Address

- Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.
- Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.
- But, in a larger sense, we can not dedicate—we can not consecrate—we can not hallow—this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us—that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion—that we here highly resolve that these dead shall not have died in vain—that this nation, under God, shall have a new birth of freedom—and that government of the people, by the people, for the people, shall not perish from the earth.

What Questions Can We Ask?

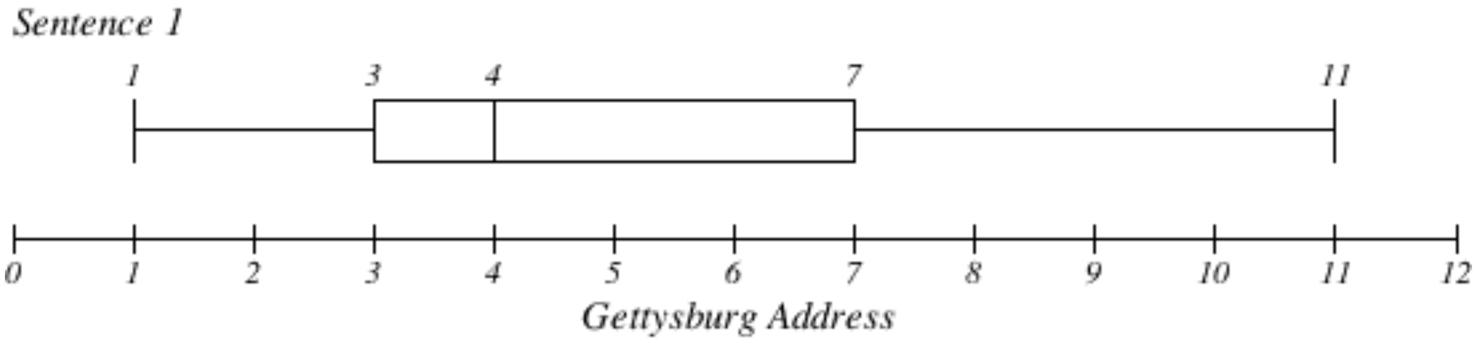
- Average number of syllables?
- Average number letters?
- Commonness of adjectives? Describing people positively?
- Readability?
- Usage of “uncommon” or ”old fashioned” words…
- Distribution of parts of speech

The First Sentence

- Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.
- Word lengths:
  - 4, 5, 3, 5, 5, 3, 3, 7, 7, 5, 2, 4, 9, 1, 3, 6, 9, 2, 7, 3, 9, 2, 3, 11, 4, 3, 3, 3, 7, 5
- Sorted:
  - 1, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 6, 7, 7, 7, 7, 9, 9, 9, 11
- Min: 1; First quartile: 3; Median: 4; Third quartile: 7; Max: 11
  - 1, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 6, 7, 7, 7, 7, 9, 9, 9, 11

A Box Plot

- Min: 1; First quartile: 3; Median: 4; Third quartile: 7; Max: 11
  - 1, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 6, 7, 7, 7, 7, 7, 9, 9, 9, 11



Progression of Content

- In grade 6, use box plots and the five-number summary
- In grade 7, compute mean and mean absolute deviation (MAD)
- In high school, compute mean and standard deviation
- In high school, use samples to estimate mean and standard deviation of a population
- Tools for creating box plots:
  - http://www.imathas.com/stattools/boxplot.html
- Sampling from Gettysburg Address:
  - http://www.rossmanchance.com/applets/OneSample.html?population=gettysburg

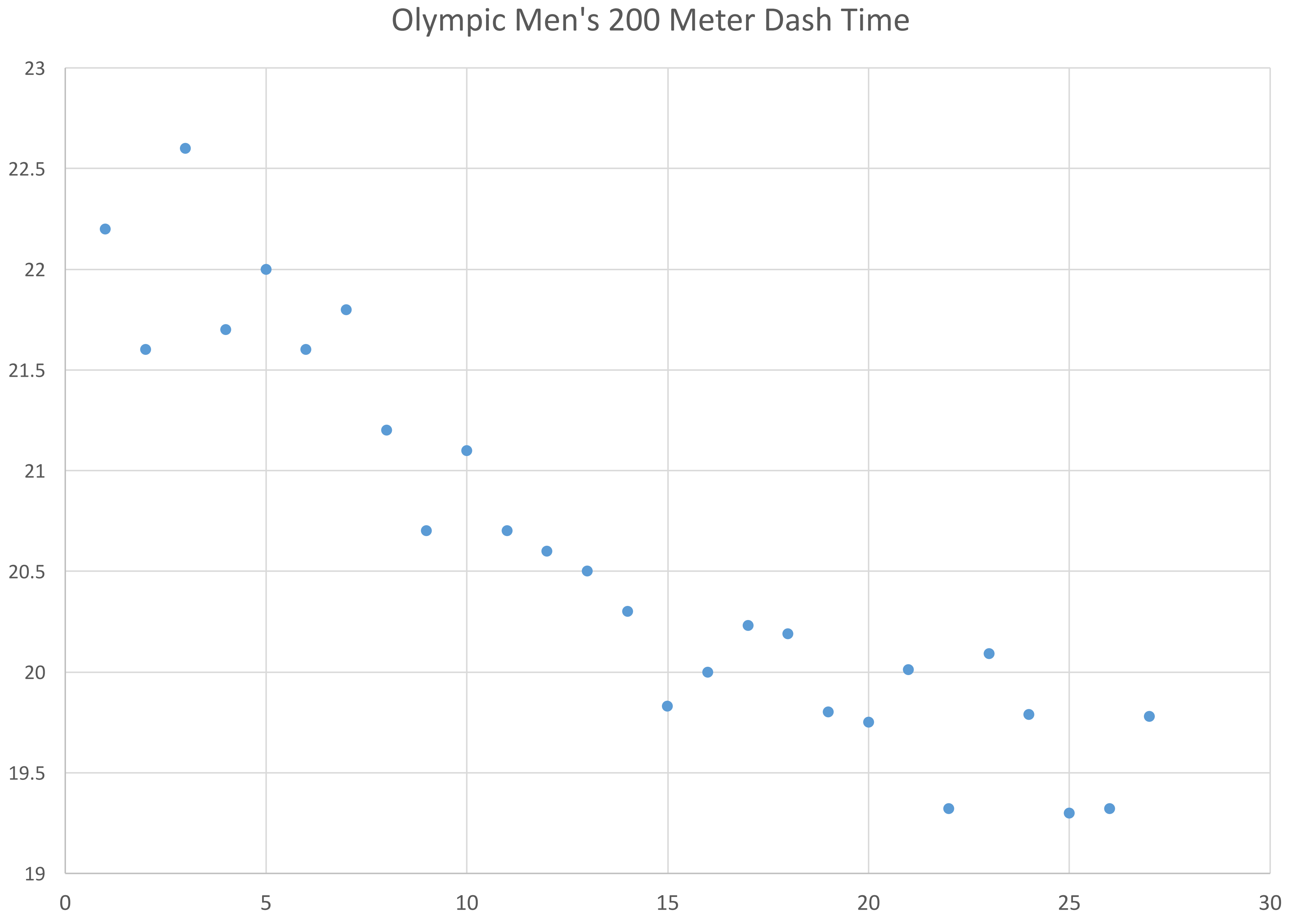
Olympics: Men’s 200m Dash

Year	Winner	Time
1900	Walter Tewksbury, United States	22.2
1904	Archie Hahn, United States	21.6
1908	Robert Kerr, Canada	22.6
1912	Ralph Craig, United States	21.7
1920	Allan Woodring, United States	22.0
1924	Jackson Scholz, United States	21.6
1928	Percy Williams, Canada	21.8
1932	Eddie Tolan, United States	21.2
1936	Jesse Owens, United States	20.7
1948	Mel Patton, United States	21.1
1952	Andrew Stanfield, United States	20.7
1956	Bobby Morrow, United States	20.6
1960	Livio Berruti, Italy	20.5
1964	Henry Carr, United States	20.3

Year	Winner	Time
1968	Tommie Smith, United States	19.83
1972	Valeri Borzov, USSR	20
1976	Donald Quarrie, Jamaica	20.23
1980	Pietro Mennea, Italy	20.19
1984	Carl Lewis, United States	19.8
1988	Joe DeLoach, United States	19.75
1992	Mike Marsh, United States	20.01
1996	Michael Johnson, United States	19.32
2000	Konstantinos Kenteris, Greece	20.09
2004	Shawn Crawford, United States	19.79
2008	Usain Bolt, Jamaica	19.3
2012	Usain Bolt, Jamaica	19.32
2016	Usain Bolt, Jamaica	19.78

What Questions Can We Ask?

- How strong is the correlation between year and time?
- What about altitude?
- What about wind aided?
- How fast was second place?
- Technology for timing? How accurate is the data?
- Is a linear model a good fit?
- Use model to estimate intermediate or future times?
- When does the model break down?


Olympic Men's 200 Meter Dash Time
23
22.5
22
21.5
21
20.5
20
19.5
19
0
5
10
15
20
25
30

Progression of Content

- Grade 8
  - Scatter plot and visually estimate line of fit
  - Interpret slope and intercepts
- High school: Modeling
  - Regression to determine line of best fit
  - More sophisticated interpretation
  - Correlation
  - Correlation vs. causation
  - Nonlinear models