



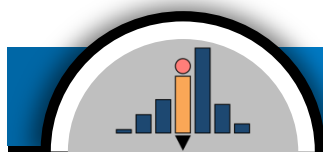
Big Data Technology Projects for the Triola Statistics Series

Mario F. Triola

Contents

I. Body Temperatures Project	1
II. Old Faithful Geyser Project	3
III. Powerball Challenge	5

Visit www.TriolaStats.com for additional
Projects, Challenges and Textbook Supplement



Triola – Body Temperatures Project

One of the six recommendations included among the *Guidelines for Assessment and Instruction in Statistics Education* is to foster **active learning** in the classroom. The author and many notable leaders in statistics education enthusiastically encourage active involvement in introductory statistics courses, and these projects provide suggestions for such involvement.

This *Body Temperatures* project covers multiple statistical concepts and can be ongoing throughout the statistics course or is suitable as an end-of-semester project.

Oral Report: Students typically have some reluctance to speak in public, so a brief oral report can be very helpful in building confidence. The oral report is an activity that can better prepare students for future professional activities. For large classes or online classes, posters or PowerPoint presentations could be used instead of in-class presentations.

Group projects are more effective because they help develop the interpersonal skills that are so necessary in today's working environment. Groups of three, four, or five students work well. The professor should select groups with consideration given to relevant factors such as past class performance and attendance.

Body Temperatures Project

This is a relatively large-scale project that uses multiple data sets and statistical tools. Project data sets can be downloaded from [TriolaStats.com](https://www.triolastats.com) and are included in [Statdisk.com](https://www.statdisk.com) (see “Data Sets” in top menu).

Data Sets (click here to download)

“Body Temperatures Civil War Union Army Veterans 1860-1940”¹

“Body Temperatures NHANES I 1975”²

“Body Temperatures Maryland 1992”³ (Data Set 5 in *Elementary Statistics* 14th Edition)

“Body Temperatures Stanford 2023”⁴

Project Introduction

Carl A. Wunderlich was a physician in Germany. He collected body temperatures from about 25,000 patients in Leipzig and, in 1871, he published results in which he identified 98.6°F (or 37.0°) as the mean human body temperature, with a range of 97.2°F to 99.5°F. For decades, the mean was believed to be 98.6°F, but more recent studies have suggested different values.

¹ Study of Union Army Veterans of the Civil War (UAVCW) data from “Decreasing human body temperature in the United States since the Industrial Revolution,” by Myroslava Protsiv, Catherine Ley, Joanna Lankester, Trevor Hastie, Julie Parsonnet. PMID: 31908267

² National Health and Nutrition Examination Survey conducted by the National Center for Health Statistics.

³ The data were provided by the authors of this article published in 1992: “A critical appraisal of 98.6 °F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich” by Mackowiak PA, Wasserman SS, Levine MM. *JAMA*. 1992;268(12):1578-1580. doi:10.1001/jama.1992.03490120092034

⁴ The data accompanied this article: “Defining Usual Oral Temperature Ranges in Outpatients Using an Unsupervised Learning Algorithm” by Catherine Ley, PhD; Frederik Heath, MS; Trevor Hastie, PhD; Zijun Gao, PhD; Myroslava Protsiv, MS; Julie Parsonnet, MD. *JAMA Intern Med*. 2023;183(10):1128-1135. doi:10.1001/jamainternmed.2023.4291

Suggested Topics

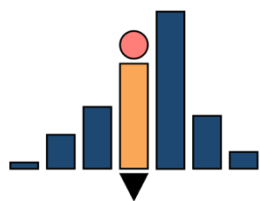
This project involves analysis of human body temperatures. Suggested topics are given below.

1. Is the mean body temperature equal to 98.6°F, as is commonly believed? If not, what single value is the actual mean body temperature? What is a 95% confidence interval estimate of the actual mean body temperature?
2. What is the level of measurement of body temperatures (ratio, interval, ordinal, nominal)? Are the temperatures discrete data or continuous data? Are the temperatures quantitative data or categorical data? If a data set is used to calculate a mean temperature, is the result a statistic or a parameter?
3. Are human body temperatures normally distributed?
4. Are human body temperatures changing over time? If so, how are they changing?
5. Is there a *statistically significant* difference between the mean body temperature of males and the mean body temperature of females? Does that difference have *practical significance*?
6. Which of these factors affect body temperatures: age, height, weight, time of day, month, gender?
7. Identify a regression equation that can be used to predict the body temperature of someone with a given gender, age, height, weight, time of day of the measurement, and month of the measurement. Use that equation to predict the body temperature of a female of age 21 with a height of 64 inches, a weight of 120 pounds, and a temperature measured at 10:00 AM during the month of November. Use the regression equation to estimate your own body temperature and determine whether the result is reasonably accurate.

Project Deliverables

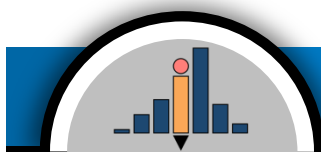
For this project, a brief written report should be submitted, and it should include the following components:

- List of data collected along with a description of how the data were obtained.
- Description of the method of analysis.
- Relevant graphs and/or statistics, including displays from technology.
- Statement of conclusions.
- Reasons why the results might not be correct, along with a description of ways in which the study could be improved. (It may not be practical for students to use a sound sampling method. Practical considerations may require that they use a convenience sample, and stress that this is perfectly acceptable as long as they explain how their sampling method is flawed and how it could have been improved with sufficient resources of time, money, and effort.)



Triola Stats

Visit www.TriolaStats.com for additional
Projects, Challenges and Textbook Supplements



Triola – Old Faithful Geyser Project

One of the six recommendations included among the *Guidelines for Assessment and Instruction in Statistics Education* is to foster **active learning** in the classroom. The author and many notable leaders in statistics education enthusiastically encourage active involvement in introductory statistics courses, and these projects provide suggestions for such involvement.

This *Old Faithful Geyser* project covers multiple statistical concepts and can be ongoing throughout the statistics course or is suitable as an end-of-semester project.

Oral Report: Students typically have some reluctance to speak in public, so a brief oral report can be very helpful in building confidence. The oral report is an activity that can better prepare students for future professional activities. For large classes or online classes, posters or PowerPoint presentations could be used instead of in-class presentations.

Group projects are more effective because they help develop the interpersonal skills that are so necessary in today's working environment. Groups of three, four, or five students work well. The professor should select groups with consideration given to relevant factors such as past class performance and attendance.

Old Faithful Geyser Project

This is a relatively large-scale project that uses multiple data sets and statistical tools. Project data sets can be downloaded from [TriolaStats.com](https://www.triolastats.com) and are included in [Statdisk.com](https://www.statdisk.com) (see “Data Sets” in top menu).

Data Sets ([click here to download](#))

Old Faithful 1990

Old Faithful 1995 (Data Set 26 in *Elementary Statistics* 14th Edition)

Old Faithful 2008

Old Faithful 2015

Old Faithful 2020

Project Introduction

The Old Faithful Geyser, located in Yellowstone National Park, is one of the most notable geysers in the world. This project involves measurements of the geyser from different time periods.

Project Topics

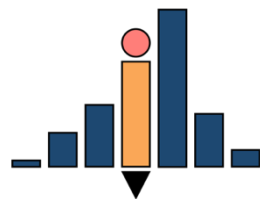
1. Consider the two variables of duration times and the time intervals after an eruption. What is the level of measurement of each of those two variables (ratio, interval, ordinal, nominal)? Are data from each of two variables discrete data or continuous data? Are data from each of the two variables quantitative data or categorical data? If data from one of the variables is used to calculate a mean, is the result a statistic or a parameter?
2. Generate histograms of the distribution of duration times for each of the five time periods. Use the histograms to characterize the nature of the distribution for each of the five time periods (normal, uniform, etc). Does it appear that the distribution of duration times is changing over time?

3. Generate histograms of the distribution of time intervals after an eruption for each of the five time periods. Use the histograms to characterize the nature of the distribution for each of the five time periods (normal, uniform, etc). Does it appear that the distribution of time intervals after an eruption is changing over time?
4. Find the mean of the duration times for each of the five time periods. Does it appear that the mean is changing over time?
5. Find the mean of the time intervals after an eruption for each of the five time periods. Does it appear that the mean is changing over time?
6. Is there a correlation between the duration times and the times after an eruption?
7. Find an equation that could be used to predict the time interval after an eruption given the duration time of the eruption.
8. Is there a correlation between the duration time and the height of a geyser?

Project Deliverables

For this project, a brief written report should be submitted, and it should include the following components:

- List of data collected along with a description of how the data were obtained.
- Description of the method of analysis.
- Relevant graphs and/or statistics, including displays from technology.
- Statement of conclusions.
- Reasons why the results might not be correct, along with a description of ways in which the study could be improved. (It may not be practical for students to use a sound sampling method. Practical considerations may require that they use a convenience sample, and stress that this is perfectly acceptable as long as they explain how their sampling method is flawed and how it could have been improved with sufficient resources of time, money, and effort.)



Triola Stats

Visit www.TriolaStats.com for additional
Projects, Challenges and Textbook Supplements



Triola – Powerball Challenge

Powerball Challenge

Powerball numbers from 846 lottery drawings are included in the “Powerball Numbers” data set. Using this data set, work individually or in groups to answer the Challenge Questions listed below.

Data Set (click here to download)

“Powerball Numbers”

This data set is also included in [Statdisk.com](https://www.statdisk.com) (see “Data Sets” in the Statdisk top menu).

Relevant Chapters

The Powerball Challenge questions relate to statistical concepts covered in Chapter 4 “Probability” and Chapter 11 “Goodness-of-fit and Contingency Tables”.

Challenge Questions

1. Construct a histogram of all of the selected numbers, excluding the red Powerball numbers. What appears to be the distribution of those numbers?
2. Construct a histogram of the red Powerball numbers. What appears to be the distribution of those numbers?
3. Consider the five Powerball numbers that are selected, not including the red Powerball number. Find the probability that when someone randomly selects those five numbers (without replacement), they select the same five numbers that are drawn in the next lottery.
4. Consider the red Powerball number. Find the probability that when someone randomly selects the red Powerball number, they select the same number that is drawn in the next lottery.
5. Find the probability of winning the Powerball lottery by selecting the winning five numbers and the red Powerball number.
6. The Powerball lottery is designed to be random in the sense that all selected numbers are equally likely. Use a goodness-of-fit test to test the claim that the selected numbers (excluding the red Powerball numbers) are equally likely. Does it appear that the selected numbers are randomly selected?
7. Use a goodness-of-fit test to test the claim that the selected red Powerball numbers are equally likely. Does it appear that red Powerball numbers are equally likely?



Visit www.TriolaStats.com for additional
Projects, Challenges and Textbook Supplements