

# BIOSTAT 702: Exercise 3.1

## One Sample Inference for a Continuous Outcome

Fall 2025

### Contents

<b>Learning Objectives</b>	<b>1</b>
<b>How to Do This Exercise</b>	<b>1</b>
<b>Grading Rubric</b>	<b>2</b>
<b>Resources</b>	<b>2</b>
<b>Question 1</b>	<b>2</b>
<b>Question 2</b>	<b>2</b>
<b>Question 3</b>	<b>3</b>
<b>Question 4</b>	<b>3</b>
<b>Question 5</b>	<b>3</b>

### Learning Objectives

1. Execute commonly used 1-sample tests for a continuous outcome, including the t-test, the sign test, and the signed rank test
2. Use Q-Q plots to evaluate continuous data distributions
3. Know how to determine when parametric and non-parametric tests are appropriate
4. Understand how to use and interpret basic data transformations

### How to Do This Exercise

We recommend that you read this entire document prior to answering any of the questions. If anything is unclear please ask for help from the instructors or TAs before getting started. You are also allowed to ask for help from the instructors or TAs while you are working on the assignment. You may collaborate with your classmates on this assignment—in fact, we encourage this—and use any technology resources available

to you, including Internet searches, generative AI tools, etc. However, if you collaborate with others on this assignment please be aware that *you must submit answers to the questions written in your own words. This means that you should not quote phrases from other sources, including AI tools, even with proper attribution.* Although quoting with proper attribution is good scholarly practice, it will be considered failure to follow the instructions for this assignment and you will be asked to revise and resubmit your answer. In this eventuality, points may be deducted in accordance with the grading rubric for this assignment as described below. Finally, you do not need to cite sources that you used to answer the questions for this assignment.

## Grading Rubric

The assignment is worth 20 points (4 points per question). The points for each question are awarded as follows: 3 points for answering all parts of the question and following directions, and 1 point for a correct answer. Partial credit may be awarded at the instructor's discretion.

## Resources

The following resources on Canvas / the internet will be helpful for answering the questions for this exercise.

1. the altered ultrarunning dataset

## Question 1

The dataset `ultrarunning_altered.csv` contains the same variables and research goals as the ultrarunning dataset we have encountered before, except now, some of the ultramarathon times have been altered slightly for learning purposes.

1. Create a box plot of the outcome variable, `pb100k_dec`. You should see a few observations flagged as high outliers. The investigators would especially focus on these as targets of future emotional intelligence interventions.
2. Create a histogram, and overlay a kernel density plot. This is one way to generate a smoothed version of the histogram. You should find that the density is highest for short-medium ultramarathon times and then quickly trails off.
3. One of the reasons to visualize a continuous outcome variable is to help choose whether or not to use parametric procedures that are based on the normal distribution. For example, a 1-sample t-test comparing mean ultramarathon times to a value of 14 hours (i.e., under the null hypothesis) (1) explicitly assumes normality; and (2) implicitly assumes that the mean is the appropriate parameter to study. As discussed, normal distributions are completely characterized by their mean and standard deviation, and the mean and median are identical. Normal distributions also have a skewness of 0. What is the skewness of the ultramarathon times? Is this value consistent with the histogram?

## Question 2

1. Let's explore the representativeness of the mean in another way. Calculate the mean and standard deviation (i.e., a standard descriptive presentation when data are normally distributed). Calculate the median and interquartile range (i.e., a standard descriptive presentation when the data are not normally distributed). How different are the mean and median? You should find that the mean falls near the top of the interquartile range, which is also apparent on the box plot.

2. What percentage of observations are above the mean? If the distribution is symmetric, this should be 50%, except for the impact of sampling variability.
3. Append a normal density to your histogram, with the mean and standard deviation equivalent to those in the sample. By eye, it should not provide a great fit to the data.
4. Create a Q-Q plot comparing the distribution of ultramarathon times to normality. In plain English, explain what a Q-Q plot does and what you found.

### Question 3

The investigator hypothesizes that the “average” ultramarathon time of their target population is 14 hours. It isn’t clear whether they are referring to the median or the mean. Perform a t-test using this null hypothesis value. Is this testing the mean or the median? Write an interpretation of the result.

### Question 4

1. Perform a sign test using this null hypothesis value. Is this testing the mean or the median?
2. Perform a signed rank test using this null hypothesis value. Is this testing the mean or the median?
3. The above p-values for the t-test, sign test, and signed-rank test are not consistent with one another. Why not? Which p-value should you use and why?

### Question 5

Sometimes when we have a distribution that is right-skewed, a log-normal distribution might be more fitting than a normal distribution. Therefore, it is common to run hypothesis tests for the mean of the log of the variable of interest, if that transformation better fits the normality assumptions than the untransformed variable.

1. Take the log of ultramarathon times (i.e., to the base  $e$ ). Check for normality. Create a 95% confidence interval for the mean of the log-ultramarathon times. Exponentiate the point estimates and the upper and lower endpoints.
2. There’s one final technical detail to consider. Do the estimate and confidence interval you just created pertain to the mean, the median, or something else?