BIOSTAT 702: Exercise 3.1

One Sample Inference for a Continous Outcome

August 18, 2025

Contents

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 interepret 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.

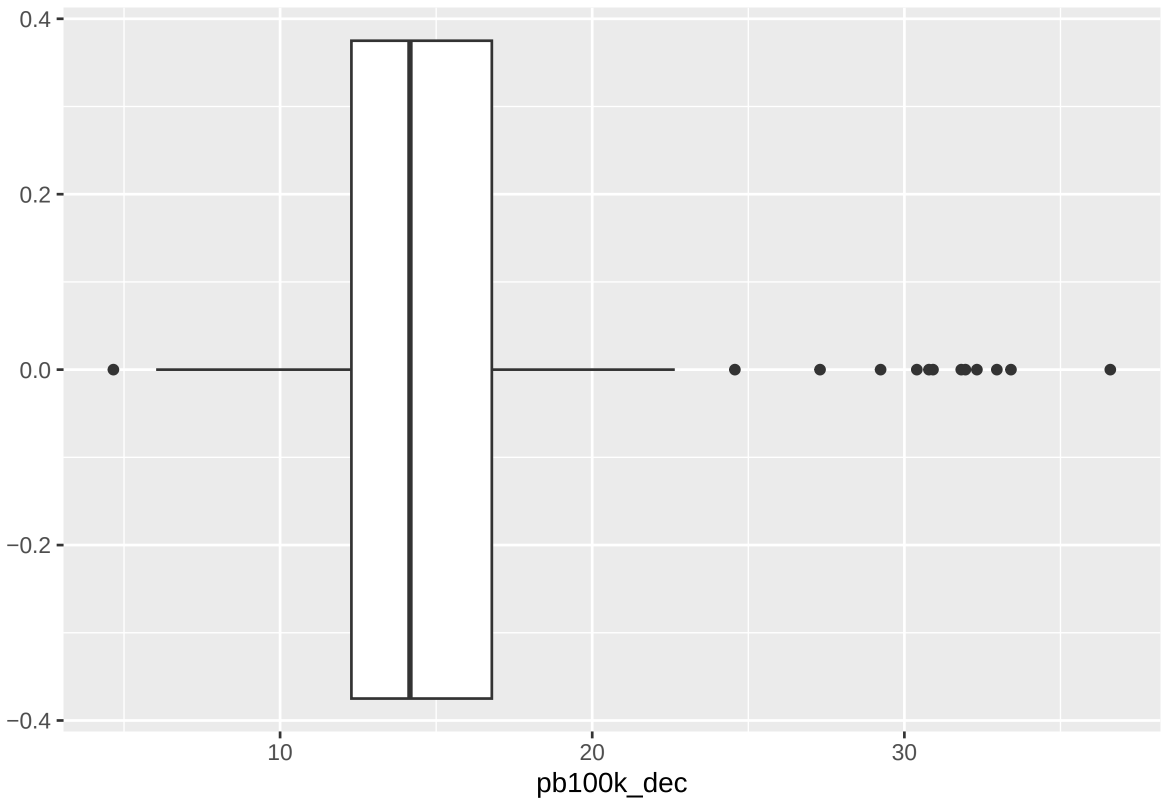
1. the altered ultrarunning dataset

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.

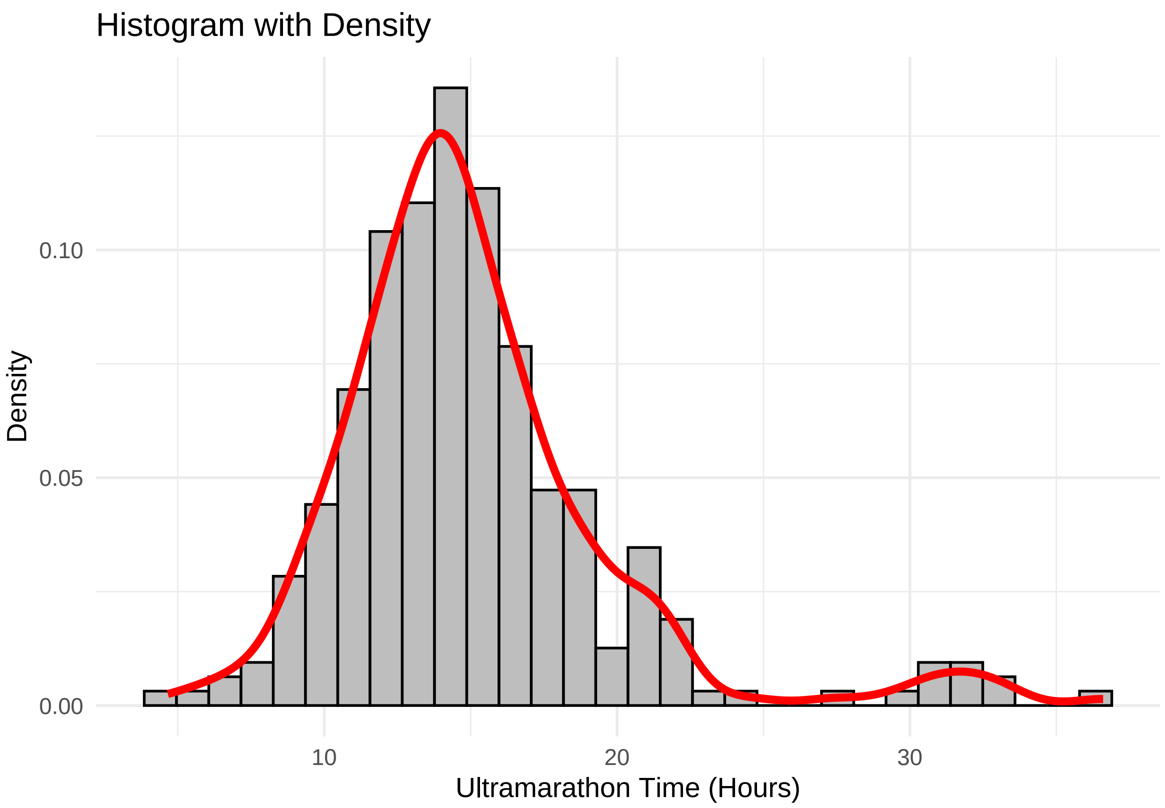
```r
ultra_altered <- read.csv(here::here("Datasets/ultrarunning/ultrarunning_altered.csv"))
ggplot(data = ultra_altered, aes(pb100k_dec)) +
  geom_boxplot()
```



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.

```r
ggplot(ultra_altered, aes(x = pb100k_dec)) +
  geom_histogram(aes(y = ..density..),
                 fill = "grey", color = "black") +
  geom_density(color = "red", size = 1.5) +
  labs(title = "Histogram with Density",
       x = "Ultramarathon Time (Hours)",
       y = "Density") +
  theme_minimal()
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



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?

   The skewness is 1.67, which is a substantial right skew, consistent with what we see in the histogram.

```
moments::skewness(ultra_altered$pb100k_dec)
```

```
## [1] 1.66582
```

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.

```
summary(ultra_altered$pb100k_dec)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.66   12.28   14.16   15.05   16.78   36.60
```

```
sd(ultra_altered$pb100k_dec)
```

```
## [1] 4.684084
```

The mean is 15.05, which is between the 50th (median) and 75th percentiles of 14.16 and 16.78. The interquartile range is 4.6 (Q3-Q1) and the standard deviation is 4.68.

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.

```
mean(ultra_altered$pb100k_dec > mean(ultra_altered$pb100k_dec))
```

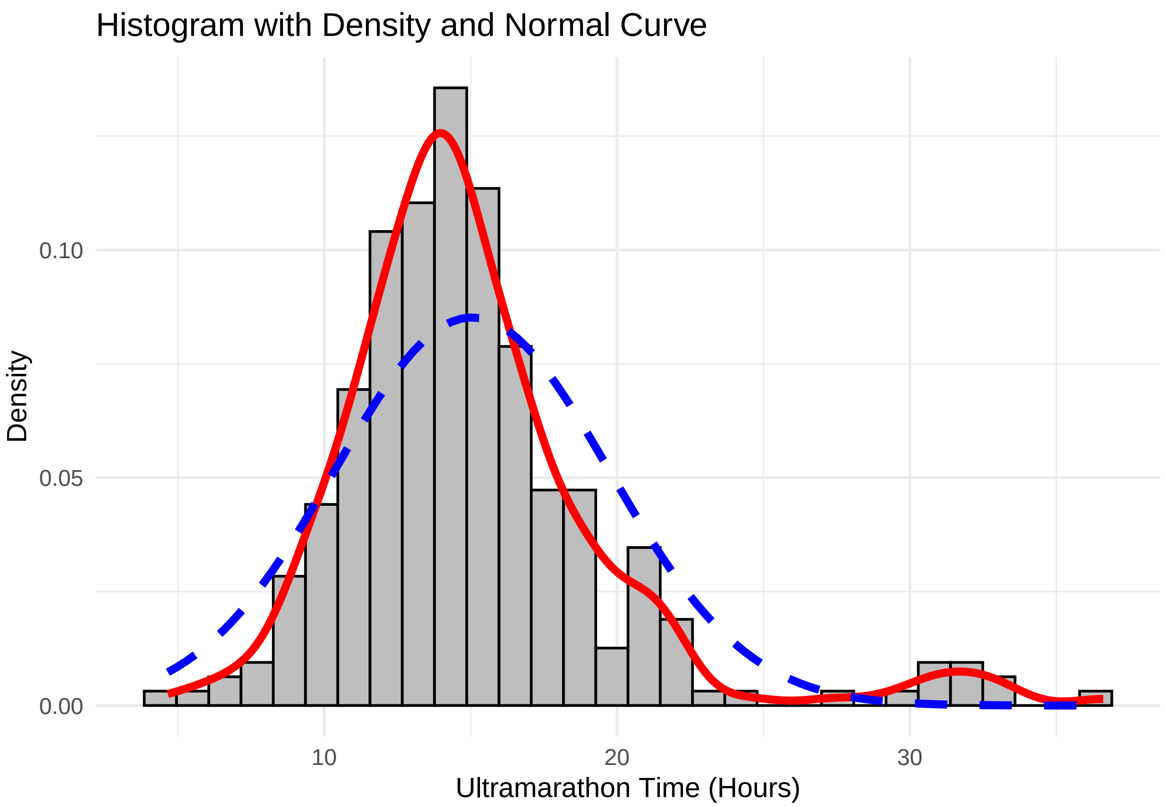
```
## [1] 0.3993056
```

about 40% of values are above the mean.

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.

```
# Create a data frame for the normal density curve
normal_density <- data.frame(
  x = seq(min(ultra_altered$pb100k_dec),
          max(ultra_altered$pb100k_dec), length.out = 100)) %>%
  mutate(y = dnorm(x, mean = mean(ultra_altered$pb100k_dec),
                   sd = sd(ultra_altered$pb100k_dec)))

# Plot
ggplot(ultra_altered, aes(x = pb100k_dec)) +
  geom_histogram(aes(y = ..density..),
                 fill = "grey", color = "black") +
  geom_density(color = "red", size = 1.5) +
  geom_line(data = normal_density, aes(x = x, y = y),
            color = "blue", size = 1.5, linetype = "dashed") +
  labs(title = "Histogram with Density and Normal Curve",
       x = "Ultramarathon Time (Hours)",
       y = "Density") +
  theme_minimal()
```

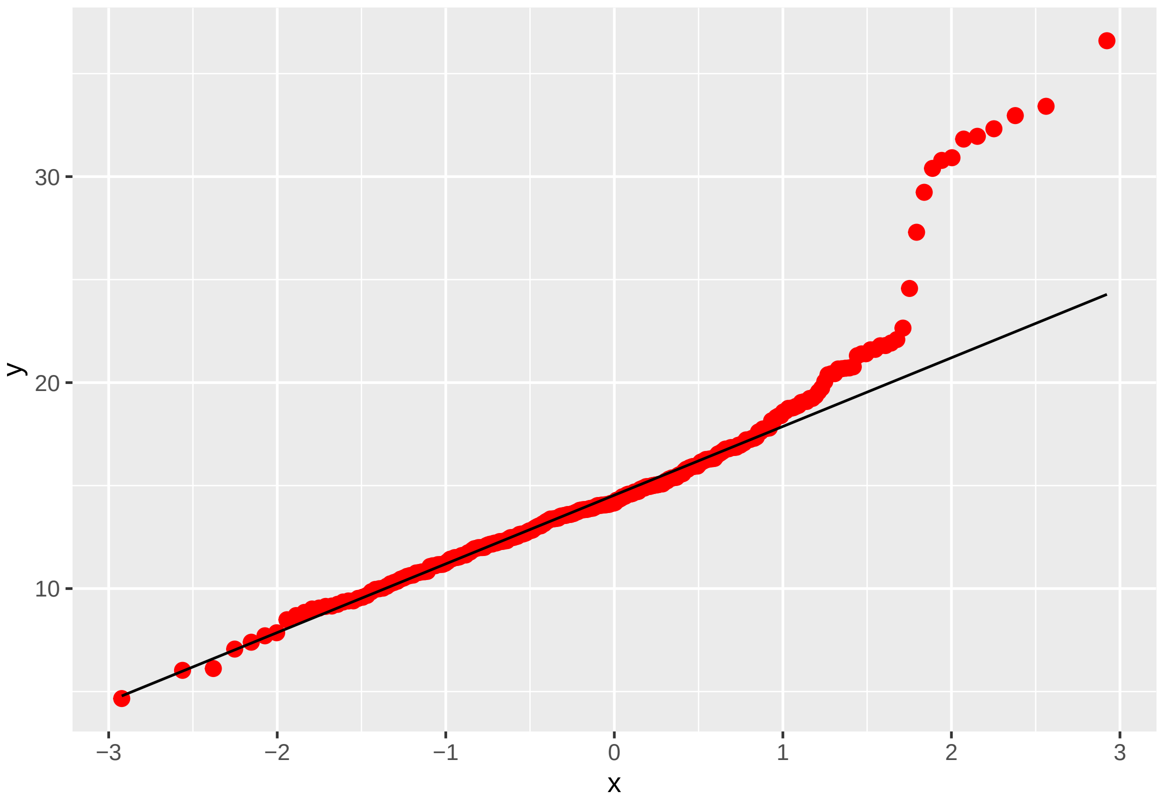
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



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.

   In a normal Q-Q plot the dots should fall on the reference line if the data are drawn from a normal distribution. In this example the dots are clearly not on the line, suggesting the data are not normal.

```r
ggplot(ultra_altered, aes(sample = pb100k_dec)) +
  stat_qq(size = 2.5, color = 'red') +
  stat_qq_line()
```



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.

The test is for the mean. In this case we reject the null hypothesis that the mean ultrarunning time is equal to 14.

```r
t.test(ultra_altered$pb100k_dec, mu=14)
```

```
##
##  One Sample t-test
##
## data:  ultra_altered$pb100k_dec
## t = 3.7862, df = 287, p-value = 0.0001863
## alternative hypothesis: true mean is not equal to 14
## 95 percent confidence interval:
##  14.50178 15.58831
## sample estimates:
## mean of x
##  15.04504
```

Question 4

1. Perform a sign test using this null hypothesis value. Is this testing the mean or the median?

   This is a test of the median.

```
# number of successes is number greater than mu_0
successes <- sum(ultra_altered$pb100k_dec > 14)

# trials is the total number in the sample
trials <- nrow(ultra_altered)

binom.test(successes,trials, 0.5)
```

```
##
##  Exact binomial test
##
## data:  successes and trials
## number of successes = 156, number of trials = 288, p-value = 0.1752
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4822100 0.6002573
## sample estimates:
## probability of success
##              0.5416667
```

2. Perform a signed rank test using this null hypothesis value. Is this testing the mean or the median?

   This is also a test of the median.

```
wilcox.test(ultra_altered$pb100k_dec,
            mu = 14)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  ultra_altered$pb100k_dec
## V = 24277, p-value = 0.01421
## alternative hypothesis: true location is not equal to 14
```

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?

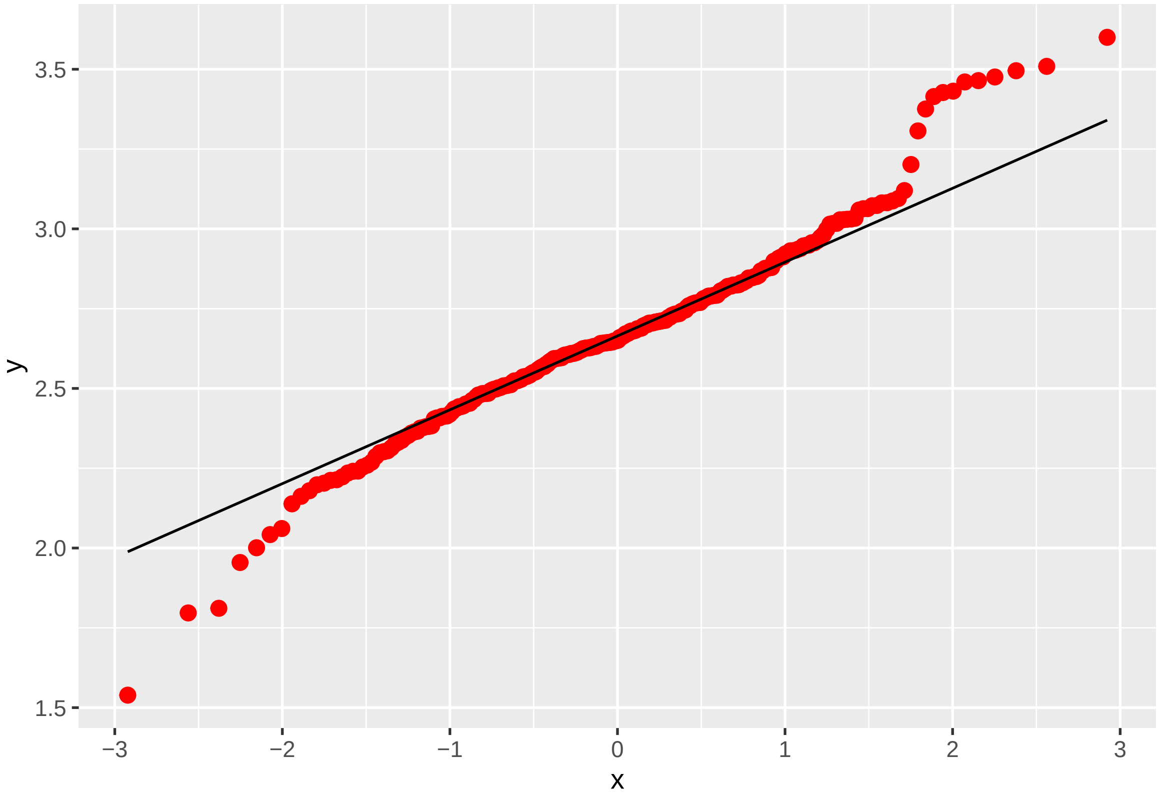
   The p-values from the t-test, sign test, and signed-rank test differ because each test makes different assumptions and targets different aspects of the data. The p-value for the t-test is very sensitive to outliers, since it directly tests the mean, which as we saw earlier, is quite different from 14. The signed-rank test uses the magnitude and direction of the values centered on 14 for the ranks. While this is more robust to outliers than the t-test, it still assumes symmetry about the null value. In our sample, the positive outliers dominate the ranks due to the asymmetry so this may not be a reliable test of the median due to this assumption violation. The sign test requires less assumptions and is therefore more reliable, but has lower power. Ultimately, if we do want to test the median to lessen the effect of the outliers, the sign test is most appropriate and may have adequate power given our large sample size. We may suggest to the investigator that a follow-up study would strengthen this result. If we do want to test the mean and are interested in the effect of the outliers, the t-test may still be useful, as long as it is understood that the result reflects the influence of the outliers.

Question 5

Sometimes when we have a distribution that is right-skewed, a log-normal distribution might be more fitting that 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.

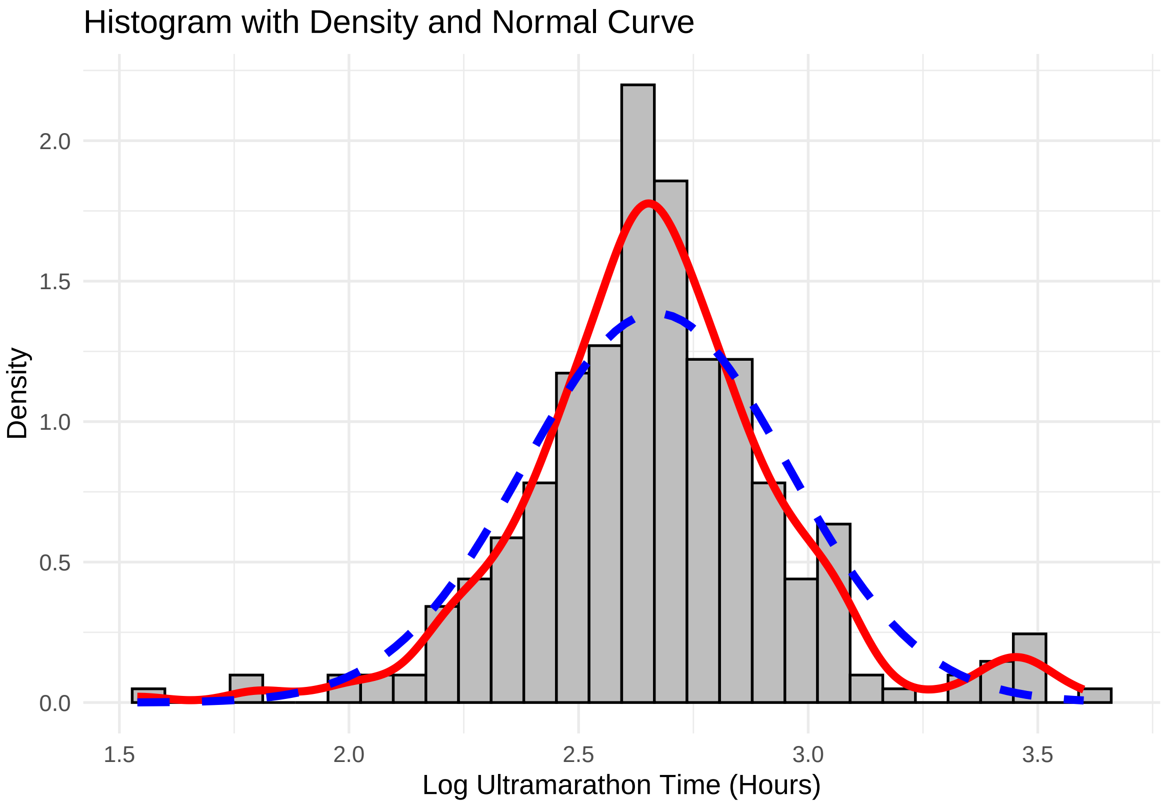
```r
# Natural log of the ultramarathon times
ultra_altered = ultra_altered %>% mutate(log_pb100k_dec = log(pb100k_dec))

# Check for normality using a Q-Q plot
ggplot(ultra_altered, aes(sample = log_pb100k_dec)) +
  stat_qq(size = 2.5, color = 'red') +
  stat_qq_line()
```



```r
# Create a data frame for the normal density curve
normal_density <- data.frame(
  x = seq(min(ultra_altered$log_pb100k_dec),
          max(ultra_altered$log_pb100k_dec), length.out = 100)) %>%
  mutate(y = dnorm(x, mean = mean(ultra_altered$log_pb100k_dec),
                   sd = sd(ultra_altered$log_pb100k_dec)))
```

```r
# Plot
ggplot(ultra_altered, aes(x = log_pb100k_dec)) +
  geom_histogram(aes(y = ..density..),
                 fill = "grey", color = "black") +
  geom_density(color = "red", size = 1.5) +
  geom_line(data = normal_density, aes(x = x, y = y),
            color = "blue", size = 1.5, linetype = "dashed") +
  labs(title = "Histogram with Density and Normal Curve",
       x = "Log Ultramarathon Time (Hours)",
       y = "Density") +
  theme_minimal()
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



```r
# run t-test
ttest_log = t.test(ultra_altered$log_pb100k_dec)

# Exponentiate
exp(ttest_log$estimate)
```

```
## mean of x
##   14.4213
```

```r
exp(ttest_log$conf.int)
```

```
## [1] 13.94777 14.91090
## attr(,"conf.level")
## [1] 0.95
```

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?

   What we actually have an estimate of is the geometric mean (the exponentiated mean of the log ultramarathon times). This will be less than the mean of the untransformed data. The geometric mean is a better representation of the "middle" of a skewed distribution than the arithmetic mean.

$$\exp\left(\frac{1}{n}\sum_i \log(x_i)\right) = \exp\left(\frac{1}{n}\log(\prod_i x_i)\right) = \exp\log\left(\prod_i x_i\right)^{1/n}$$