

answer for ex3.1

Jiaqi Wang

2025-09-25

Question1

```
ultra_altered <- read.csv(here::here("data","ultrarunning_altered.csv"))
pb100k_dec <- ultra_altered$pb100k_dec
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##   filter, lag
```

```
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

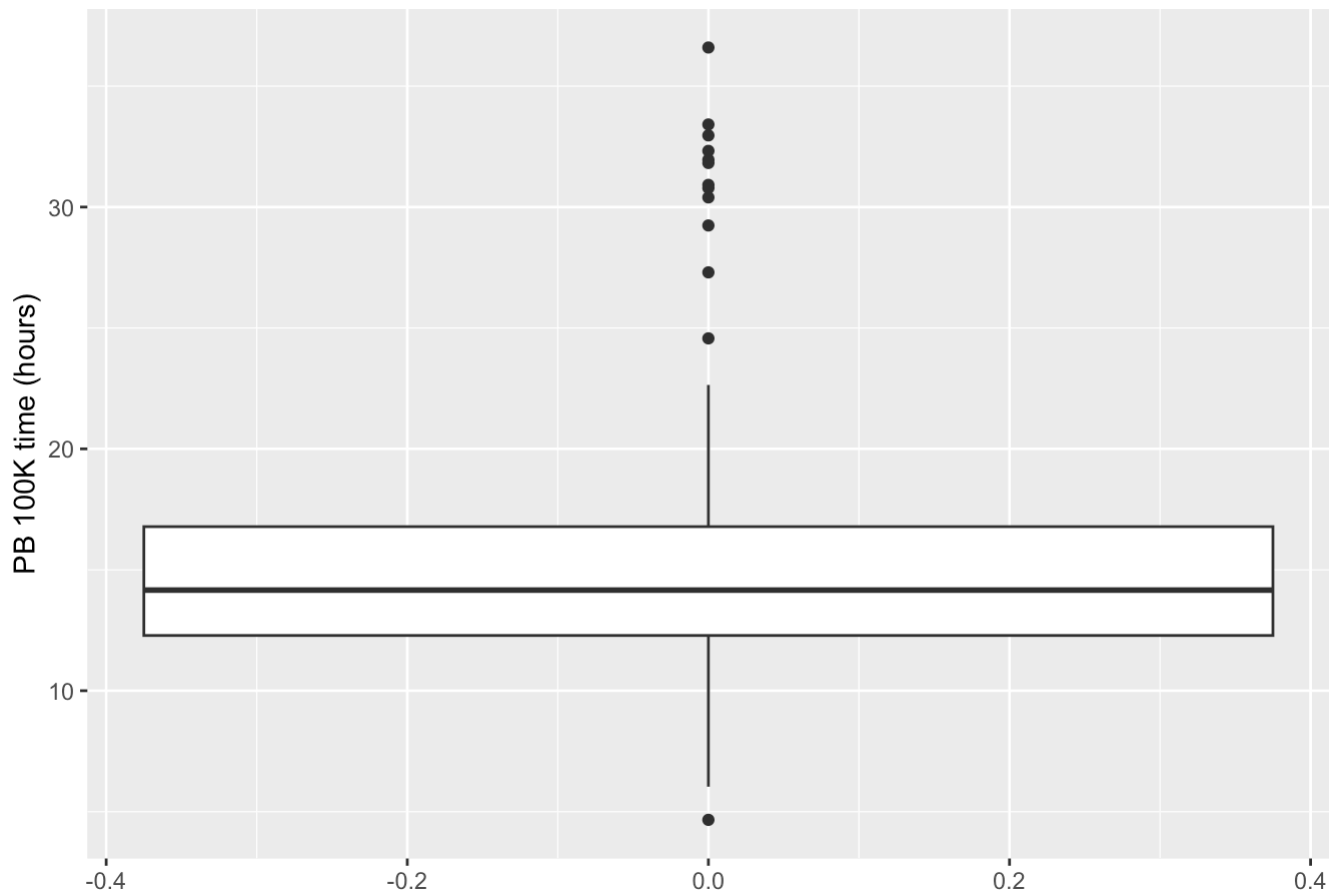
```
## — Attaching core tidyverse packages — tidyverse 2.0.0 —
## ✓ forcats   1.0.0   ✓ readr       2.1.5
## ✓ ggplot2   3.5.2   ✓ stringr    1.5.1
## ✓ lubridate 1.9.4   ✓ tibble     3.3.0
## ✓ purrr     1.1.0   ✓ tidyr      1.3.1
```

```
## — Conflicts — tidyverse_conflicts() —
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()     masks stats::lag()
## ⓘ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(moments)

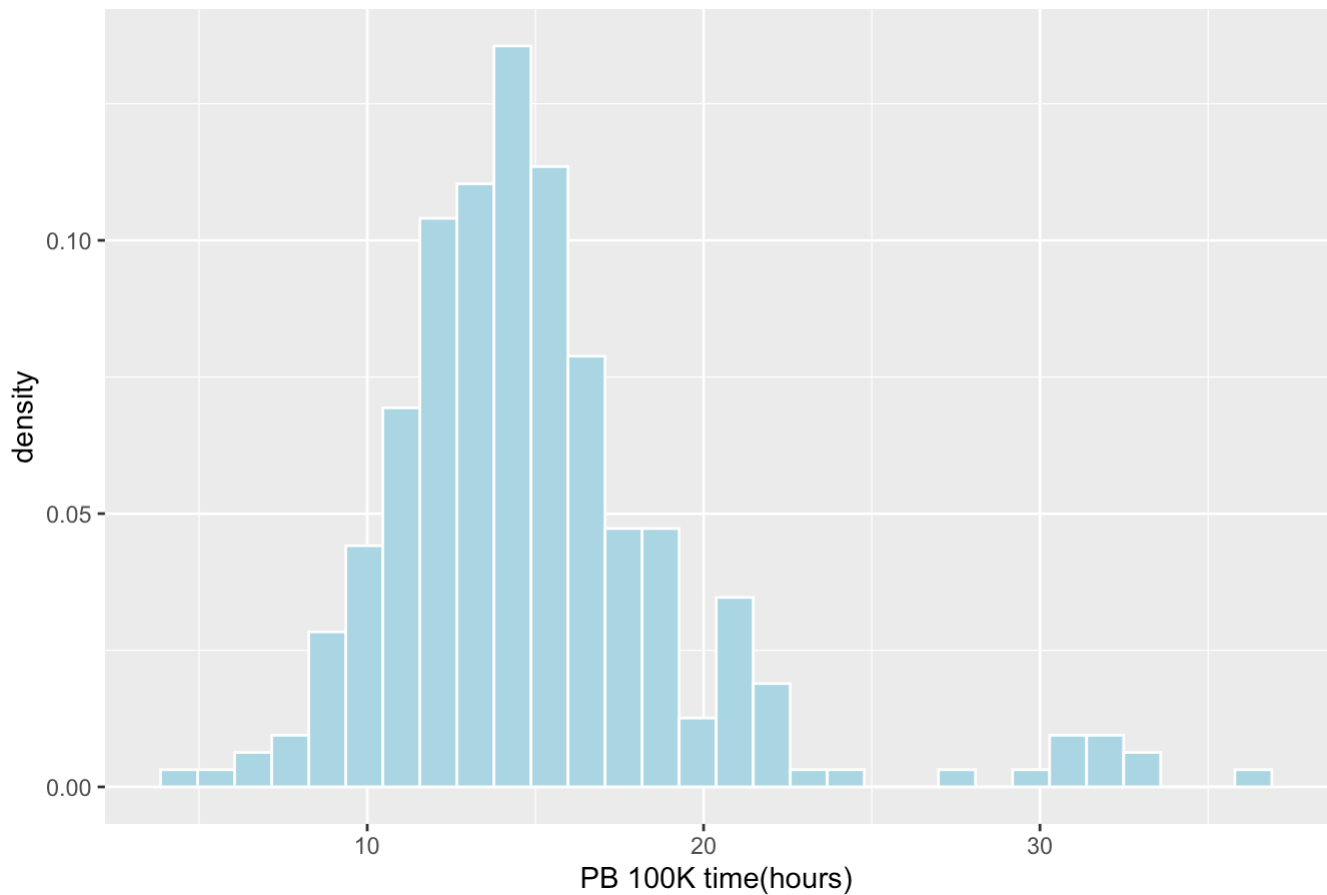
ggplot(ultra_altered, aes(y = pb100k_dec)) +
  geom_boxplot() +
  labs(y = "PB 100K time (hours)", title = "Boxplot of pb100k_dec" )
```

Boxplot of pb100k_dec



```
ggplot(ultra_altered, aes(x = pb100k_dec)) +  
  geom_histogram(aes(y = after_stat(density)),  
    bins = 30, fill = "lightblue", color = "white")+  
  labs(x = "PB 100K time(hours)", title = "histogram of pb100k_dec")
```

histogram of pb100k_dec



```
y <- ultra_altered |> drop_na(pb100k_dec) |> pull(pb100k_dec)
skew_val <- skewness(y)
skew_val
```

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

```
mean_val <- mean(y)
sd_val <- sd(y)
med_val <- median(y)
iqr_val <- IQR(y)
mean_val; sd_val; med_val; iqr_val
```

```
## [1] 15.04504
```

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

```
## [1] 14.16044
```

```
## [1] 4.499292
```

```
mean_val - med_val
```

```
## [1] 0.8846023
```

```
quantile(y, probs = c(0.25, 0.5, 0.75))
```

```
##          25%          50%          75%  
## 12.28475 14.16044 16.78404
```

#The mean finishing time (15.05 hours) is higher than the median (14.16 hours). The mean of 15.05 hours falls within the IQR (12.0–16.5 hours), at about 68% of the distance from Q1 to Q3, i.e., closer to Q3. This matches the boxplot, where the mean appears in the upper half of the box.”

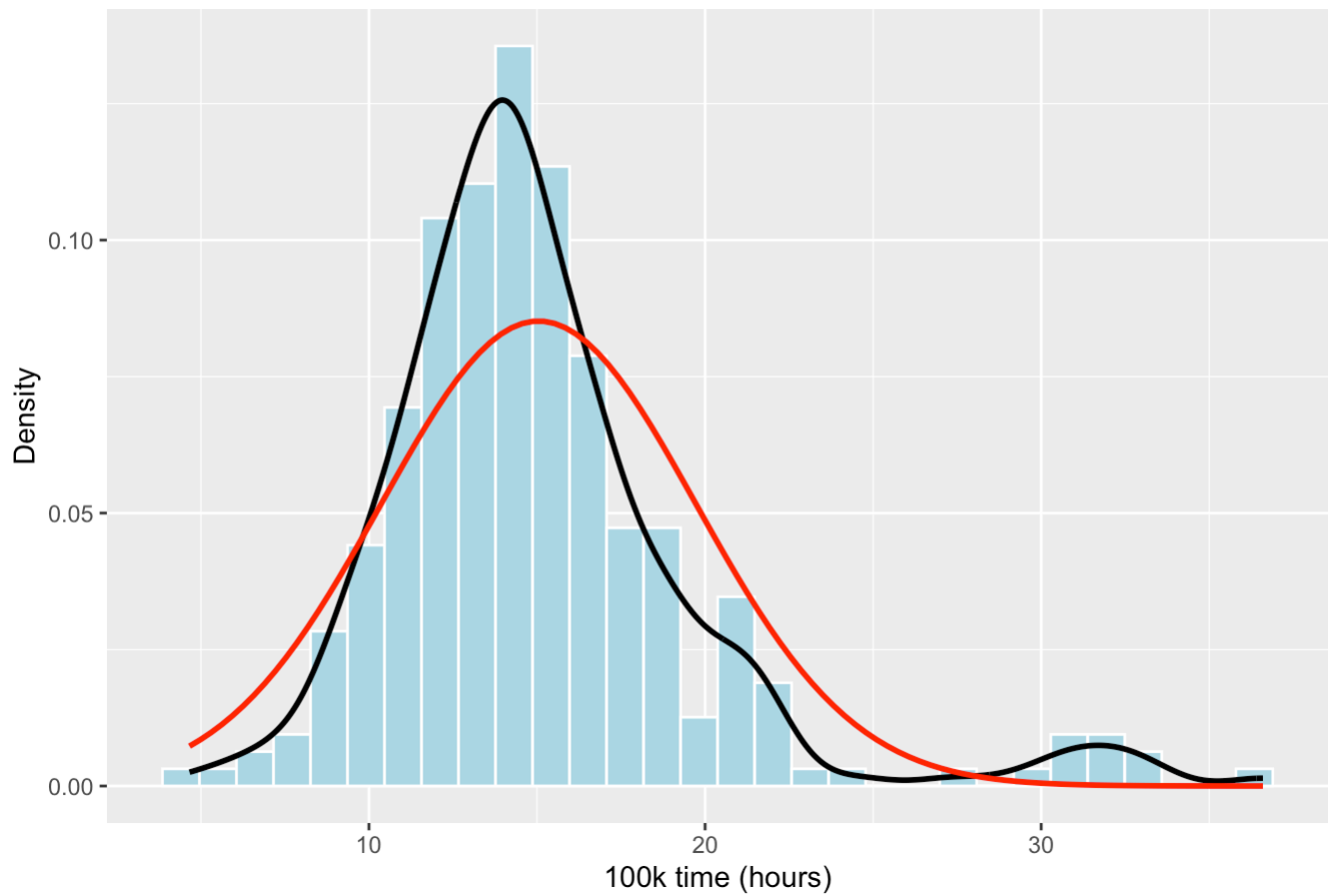
Question 2

```
prop_above_mean <- mean(y > mean_val)  
prop_above_mean
```

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

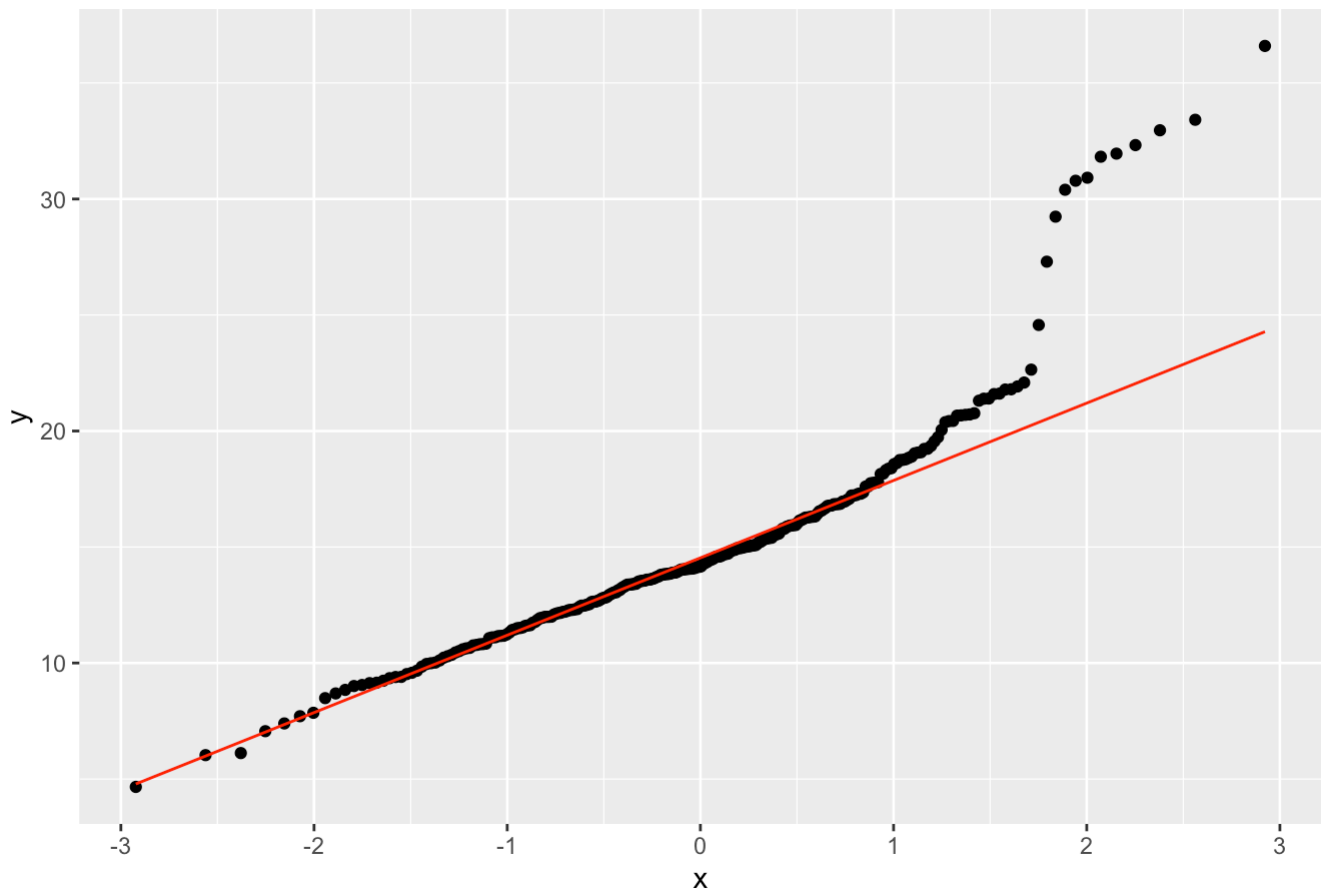
```
library(ggplot2)  
  
ggplot(data.frame(y), aes(x = y)) +  
  geom_histogram(aes(y = after_stat(density)), bins = 30,  
                 fill = "lightblue", color = "white") +  
  geom_density(linewidth = 1) +  
  stat_function(fun = dnorm, args = list(mean = mean_val, sd = sd_val),  
               linewidth = 1, color = "red") +  
  labs(x = "100k time (hours)", y = "Density",  
       title = "Histogram + Kernel Density + Normal Curve")
```

Histogram + Kernel Density + Normal Curve



```
# Q-Q  
ggplot(data.frame(sample = y), aes(sample = sample)) +  
  stat_qq() + stat_qq_line(color = "red") +  
  labs(title = "Normal Q-Q Plot")
```

Normal Q-Q Plot



```
ttest_ultra = t.test(y,mu = 14)
ttest_ultra$statistic
```

```
##          t
## 3.786211
```

```
ttest_ultra$parameter
```

```
## df
## 287
```

```
ttest_ultra$p.value
```

```
## [1] 0.0001862835
```

```
ttest_ultra$conf.int
```

```
## [1] 14.50178 15.58831
## attr(,"conf.level")
## [1] 0.95
```

```
ttest_ultra$estimate
```

```
## mean of x
## 15.04504
```

#We conducted a one-sample t-test to compare the mean ultramarathon finishing time against 14 hours ($H_0: \mu = 14$). The sample mean was 15.05 hours. The test yielded $t(287) = 3.79$, $p < 0.001$, with a 95% confidence interval for the mean of [14.50, 15.59]. Since 14 does not fall within this interval, we reject the null hypothesis and conclude that the true mean finishing time is significantly greater than 14 hours.

Question 4

```
successes <- sum(ultra_altered$pb100k_dec > 14)

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
```

```
wilcox.test(ultra_altered$pb100k_dec,
            mu = 14,
            conf.int = TRUE,
            correct = FALSE)
```

```
##
## Wilcoxon signed rank test
##
## data: ultra_altered$pb100k_dec
## V = 24277, p-value = 0.01419
## alternative hypothesis: true location is not equal to 14
## 95 percent confidence interval:
## 14.10189 14.95838
## sample estimates:
## (pseudo)median
## 14.5147
```

#1. Sign test: The binomial sign test yielded a p-value of about 0.175, which is not statistically significant. This suggests insufficient evidence to reject the null hypothesis that the population median is 14 hours.

#2. Wilcoxon signed-rank test: The Wilcoxon test gave $V = 24277$, $p = 0.014$, with a 95% confidence interval of [14.10, 14.96] and an estimated median (Hodges–Lehmann) of 14.51 hours. Since 14 is not contained in this interval and the p-value is below 0.05, we reject the null hypothesis, indicating that the true median differs significantly from 14 hours.

#3. Comparison and interpretation: The two tests gave different p-values because they use different information: the sign test only considers whether observations are above or below 14, while the Wilcoxon test also incorporates the magnitude of the differences. When the distribution is approximately symmetric, the Wilcoxon test is more powerful. In our data, which are somewhat right-skewed, the tests diverged. If the research question is about the typical central tendency (median) in a skewed distribution, the Wilcoxon test is generally preferable, while the sign test remains the most robust under extreme asymmetry.

Question 5

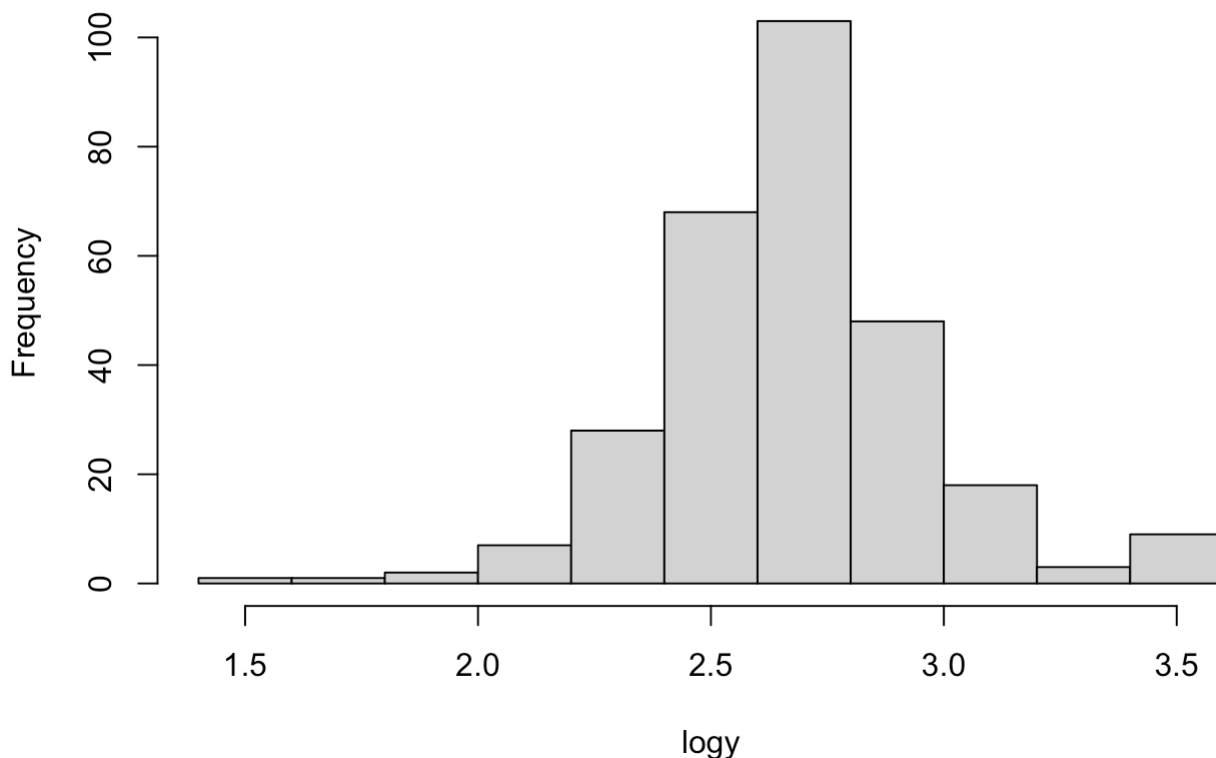
```
logy <- log(y)
log_fit <- t.test(logy)
geo_point <- exp(log_fit$estimate)
geo_ci    <- exp(log_fit$conf.int)
geo_point; geo_ci
```

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

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

```
hist(logy)
```

Histogram of logy



#After log-transforming ultramarathon finishing times, the histogram and Q-Q plot indicate that the distribution is much closer to normal compared to the raw scale. A t-test for the mean of the log-transformed times gave a 95% confidence interval on the log scale of [2.635, 2.702]. Exponentiating the point estimate and interval yields a geometric mean finishing time of 14.42 hours, with a 95% confidence interval of [13.95, 14.91] hours. Importantly, this back-transformed estimate and interval refer to the geometric mean (which, under a log-normal model, equals the median on the original scale), rather than the arithmetic mean.