3k running assignment

Sahil Chitnis, id=23100637

22 November, 2023

## Inference and linear regression for 3 km Running Times based on laboratory testing

### Study Description:

Sixteen well-trained male middle and long distance runners performed independently a 3 km time trial and a number of running tests in the laboratory.

One of the tests was running velocity at a blood lactate concentration of 4 mmol/L (v4mM), measured in km/hr. For background knowledge to understand something about the study, it requires considerable physical effort to achieve a blood lactate concentration of 4 mmol/L in well-trained athletes (if you want to know more about blood lactate and athletic performance here is a reference <https://www.trainingpeaks.com/blog/what-is-lactate-and-lactate-threshold/>).

Other tests were running velocity at Lactate Threshold (vTlac), and at VO2 max (V02Max) both measured in km/hr. In addition, a measure of running economy, oxygen uptake (measured in ml/kg/min) at running speeds of 14.5km/hr and 16.1 km/hr were measured and the variable names are Rel.14.5 and Rel.16.1.

All the laboratory testing took place on a motorised treadmill, and distance running performance was determined by 3 km time trials on an indoor 200m track, 3k time in minutes was recorded as Running.Time.

Based on data on other populations, the sports scientist doing the study wants to know what evidence the sample provides that the mean 3k running time in this population is different to 10 minutes 12 seconds (10.2 minutes), and also wants to estimate a plausible interval for the population mean 3k running time. The sports scientist also wants to estimate the 3k time below which the fastest 10% of the population times lie. Finally, the sports scientist wants to investigate if it is possible to predict 3k running time from running velocity at a blood lactate concentration of 4 mmol/L.

### Aims:

To investigate the following:

1. Is the population mean 3k running time different to 10 minutes 12 seconds (ie as decimal 10.2 minutes)? What is a plausible range in which the population mean 3k time is likely to lie? And what time are we confident that the top 10% of the population 3k times will be faster than?
2. Can we use linear regression to predict 3 km running time in minutes (Running.Time) from running speed at lactate 4 mmol/L in km per hour (v4mM)?

It is decided in advance that the acceptable significance level for hypothesis testing is = 0.05.

# load required libraries  
library(tidyverse)  
library(infer)  
library(tolerance)

### Read the data and see a few rows

running = read.csv("3krunning.csv", header = TRUE)  
head(running)

## Running.Time v4mM vTlac Rel.14.5 Rel.16.1 VO2Max  
## 1 8.23 20.4 19.5 47.1 52.4 23.4  
## 2 8.30 19.5 18.2 48.1 60.0 23.5  
## 3 8.62 19.0 17.3 50.3 56.8 22.0  
## 4 8.82 18.9 17.8 51.8 56.1 23.0  
## 5 9.18 17.8 16.5 48.7 54.1 21.5  
## 6 9.23 17.2 15.6 50.5 59.6 20.5

### Summary Statistics

Shown below are summary statistics for some of the columns in the dataset.

running %>% summarise(  
 count = n(),  
 meantime = mean(Running.Time),  
 sdtime = sd(Running.Time),  
 meanv4mM = mean(v4mM),  
 sdv4mM = sd(v4mM),  
 meanRel14.5 = mean(Rel.14.5),  
 sdRel14.5 = sd(Rel.14.5),  
   
 )

## count meantime sdtime meanv4mM sdv4mM meanRel14.5 sdRel14.5  
## 1 16 9.458125 0.744269 17.06875 1.848141 51.59375 3.289877

## Hypothesis test and confidence interval for mean 3k time

#### Task 1: Write the null and alternative hypotheses and significance level of the test (text, 1 sentence)

i.e. the hypothesis test for Aim (1), first sentence.

Answer: The null hypothesis (H0) for this test is that the population mean 3k running time is equal to 10.2 minutes, and the alternative hypothesis (H1) is that the population mean 3k running time is not equal to 10.2 minutes.This will be valid for significance level of 0.05. **5 marks**

#### Task 2): Appropriate statistical test and assumptions

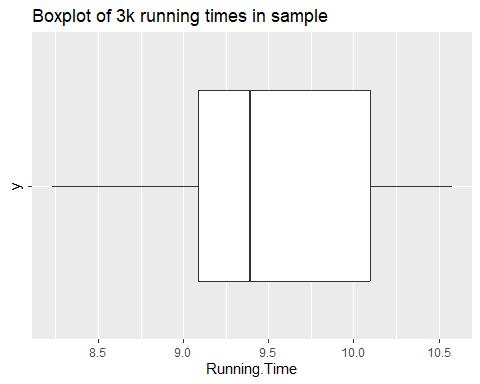
1. What statistical test is appropriate to test the Task 1 hypotheses? State the assumptions needed for the test to be valid, and how they could be justified in this study. (1-2 sentences).

Answer: As we are comparing the mean to a known value i.e 10.2 and we also don’t know the standard deviation a one sample t-test would be a appropriate to test the hypotheses.Here I am assuming that the data is normally distributed and the observations are independent.

1. Comment on the output of the R code below in relation to an assumption for the test (1 sentence)

Answer: The boxplot output for the 3k running times describes a fairly symmetric distribution of the data without the presence of any outliers which supports the normality assumption for validating the One-sample test considered for the hypothesis.

running %>% ggplot(aes(x =Running.Time, y = "" )) + geom\_boxplot() + labs(title = "Boxplot of 3k running times in sample")



**10 marks**

#### Task 3: Carry out the hypothesis test and estimate a 95% confidence interval for the true mean 3k time

The code below does a two-sided t test for the null hypothesis that the true mean Rel.14.5 is equal to 50. Adapt the code below to carry out the hypothesis test from Task 1.

running %>% dplyr::select(Running.Time) %>% t.test(mu = 10.2)

##   
## One Sample t-test  
##   
## data: .  
## t = -3.9871, df = 15, p-value = 0.00119  
## alternative hypothesis: true mean is not equal to 10.2  
## 95 percent confidence interval:  
## 9.061532 9.854718  
## sample estimates:  
## mean of x   
## 9.458125

**5 marks**

#### Task 4: Interpret (a) the results of the hypothesis test (b) the 95% confidence interval from Task 3

Note: you need to interpret the results after you have adapted the code in Task 3. You will not get marks for this question if you intepret the hypothesis test for the Rel.14.5 variable. (1-2 sentences)

Answer:- a) The hypothesis test results indicate that the p-value is 0.00119 and the t-statistic is -3.9871 with 15 degrees of freedom. We reject the null hypothesis because the p-value is smaller than the significance level of 0.05. b) This indicates that there is a 95% confidence interval around the true population mean 3k running time. Further evidence that the population mean 3k running time is significantly different from 10.2 minutes can be found in the fact that the interval does not contain the value 10.2 minutes.

**5 marks**

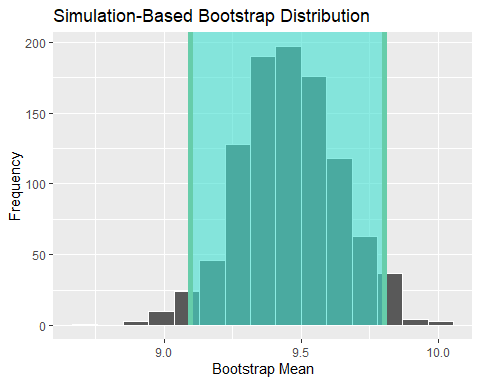
#### Task 5: A bootstrap 95% confidence interval for the mean 3k time

The code below estimates a 90% confidence interval for the mean 3k time using the bootstrap method. Change the number of bootstrap replications to 1000, and change the confidence level to 95%.

set.seed(123456789) # this is for replicable simulations - do not change this  
boot <- running %>%   
 specify(response = Running.Time) %>%  
 generate(reps = 1000, type = 'bootstrap') %>%  
 calculate(stat = 'mean')  
  
### Produce the 90% bootstrap confidence interval  
### using the generated bootstrap distribution of the sample statistic  
get\_ci(boot, level = 0.95)

## # A tibble: 1 × 2  
## lower\_ci upper\_ci  
## <dbl> <dbl>  
## 1 9.10 9.80

### plot the histogram of bootstrap statistics  
### and shading the bootstrap confidence interval.   
boot %>%   
 visualize(method = "simulation") +  
 shade\_confidence\_interval(get\_ci(boot))+  
 xlab("Bootstrap Mean")+  
 ylab("Frequency")



**5 marks**

#### Task 6: Compare the 95% confidence interval using the t distribution with the bootstrap 95% CI

Answer:- Comparing values of the upper and lower bounds of the 95% confidence interval to the histogram of the bootstrap mean shows consistent results. The lower confidence interval is 9.096812 and the upper confidence interval is 9.802625 and this correlates to the selected interval on the histogram. According to the t-test the values within 95% confidence interval the range are (9.061532, 9.854718). This shows a larger interval than the bootstrap, but still similar.

**5 marks**

#### Task 7: A tolerance interval for the range in which we are 95% confident 90% of the population of 3k times will lie.

1. The code below estimates a tolerance interval for the range in which we are 95% confident 95% of the population of 3k times will lie. Adapt the code to estimate a tolerance interval for the range in which we are 95% confident 90% of the population of 3k times will lie.
2. The men’s world record for the 3k is 7 mins 20.67 seconds by Daniel Komen of Kenya in 1996. Using your adapted tolerance interval, state the running time you are 95% confident that the fastest 10% of our study population will achieve. (1 sentence)

Answer: The output of the modified R-code has a tolerance range of 7.64 to 11.28 minutes. Consequently, we have a 95% confidence level that the top 10% of the study population will complete the course in 7.64 minutes or less.

normtol.int(running$Running.Time, alpha = 0.05, P = 0.90, side = 2)

## alpha P x.bar 2-sided.lower 2-sided.upper  
## 1 0.05 0.9 9.458125 7.635421 11.28083

**5 marks**

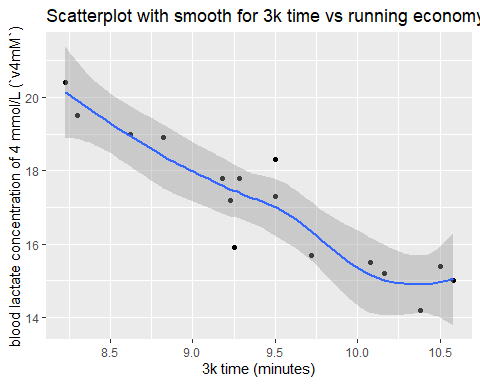
## Linear Regression

#### Task 8: Labelled scatterplot with smooth of Running.Time versus v4mM

1. The code below makes a scatterplot with smooth of Running.Time versus Rel.14.5. Adapt the code to make a scatterplot with smooth of Running.Time versus v4mM. Remember to update the title, axis labels, and units.
2. What does the smoother suggest regarding the suitability of a simple linear regression model for the relationship between Running.Time and v4mM? (1 sentence)

Answer:- The smoother shows a nonlinear relationship between running.time and rel.14.5.

running %>% ggplot(aes(x = Running.Time, y = v4mM)) + geom\_point() + geom\_smooth() + labs(title = "Scatterplot with smooth for 3k time vs running economy at 14 kph", x = "3k time (minutes)", y = "blood lactate concentration of 4 mmol/L (`v4mM`)")



**5 marks**

#### Task 9: Correlation coefficient between v4mM and Running.Time

1. The code below calculates the sample correlation coefficient between Running.Time and Rel.14.5. Change to code to calculate the sample correlation coefficient between Running.Time and v4mM.

running %>% dplyr::select (Running.Time, v4mM) %>% cor()

## Running.Time v4mM  
## Running.Time 1.000000 -0.925857  
## v4mM -0.925857 1.000000

1. Interpret the relationship between Running.Time and v4mM, referring to the relevant scatterplot and correlation coefficient.

Answer:- This value is close to -1, indicating a strong negative linear relationship between these two variables.

**10 marks**

#### Task 10: Line of best fit for the linear relationship between 3 km running time and v4mM in the sample.

1. The code below calculates the simple linear regression for Running.Time versus Rel.14.5. Change the code to produce the simple linear regression for Running.Time versus v4mM.

lm1<-lm(Running.Time ~ v4mM, data = running)  
lm1

##   
## Call:  
## lm(formula = Running.Time ~ v4mM, data = running)  
##   
## Coefficients:  
## (Intercept) v4mM   
## 15.8223 -0.3729

observed\_range <- range(running$v4mM)  
  
# Display the observed range  
print(observed\_range)

## [1] 14.2 20.4

**5 marks**

#### Task 11: Write down the equation of the line of best fit from for the simple linear regression of Running.Time on v4mM and provide an interpretation of the slope and intercept. Does the intercept have a physically meaningful interpretation?

Answer:- Running Time= 15.8223 + (-0.3729) \* V4mM. The slope of the line is -0.3729. This means that for each additional km/hr increase in the running speed at a blood lactate concentration of 4 mmol/L (v4mM), we expect the 3 km running time (Running.Time) to decrease by 0.3729 minutes, on average, assuming all other factors are held constant.The intercept of the line is 15.8223. This is the expected value of Running.Time when v4mM is 0.

**10 marks**

#### Task 12: Predict the running time (i.e. Running.Time) when running speed at blood lactate concentration 4 mmol/litre (i.e. v4mM) is 18.9 km per hour.

The code below predicts Running.Time when Rel.14.5 = 48 ml/kg/min from the linear regression model of Running.Time on Rel.14.5. It also shows the first few rows of the dataset.

1. Adapt the code to predict Running.Time when v4mM=18.9 km/hr from a linear regression of Running.Time on v4mM.
2. Why is the predicted running time at v4mM=18.9 different to 8.82, the observed running time when v4mM is 18.9 km/hr? - see observation 4 in the dataset. What is the value of the residual for this observation?

Answer:- Since the linear regression model relies on the relationship between Running, the expected running time at v4mM=18.9 may differ from the observed running time of 8.82.Time and v4mM across all observations in the dataset, not just one.

The residual for this observation is the difference between the observed and predicted running time ie -0.18.

(1-2 sentences)

lm1<-lm(Running.Time ~ v4mM, data = running)  
newdata <-data.frame(v4mM = 8.82)  
predict(lm1, newdata = newdata)

## 1   
## 12.5337

head(running)

## Running.Time v4mM vTlac Rel.14.5 Rel.16.1 VO2Max  
## 1 8.23 20.4 19.5 47.1 52.4 23.4  
## 2 8.30 19.5 18.2 48.1 60.0 23.5  
## 3 8.62 19.0 17.3 50.3 56.8 22.0  
## 4 8.82 18.9 17.8 51.8 56.1 23.0  
## 5 9.18 17.8 16.5 48.7 54.1 21.5  
## 6 9.23 17.2 15.6 50.5 59.6 20.5

v4mM\_observation\_4 <- 18.9  
  
# Predicted running time for observation 4  
predicted\_running\_time\_observation\_4 <- predict(lm1, newdata = data.frame(v4mM = v4mM\_observation\_4))  
  
# Display the predicted running time  
print(predicted\_running\_time\_observation\_4)

## 1   
## 8.775336

#calculating residual  
residual <- 8.82 - predict(lm1, newdata = newdata)  
residual

## 1   
## -3.713704

**10 marks**

#### Task 13: Predict the running time (i.e. Running.Time) when v4mM is 2.6 km per hour. Explain if you have any concern related to this prediction.

1. Write code in the chunk below to predict Running.Time when v4mM= 2.6 km/hr. (Hint: copy and adapt the relevant code from the chunk above)
2. Do you have any concern about this prediction, and if so why? (1 sentence)

Answer:- The V4mM value of 2.6 is outside of the v4mM dataset range. We can make a guess as to what the corresponding running time might be when the value of v4mM is 2.6, but this is not a real value or meaningful interpretation of the dataset.Due to this reasoning the predidtion for v4mM 2.6 is unrelaiable.

lm1<-lm(Running.Time ~ v4mM, data = running)  
newdata <-data.frame(v4mM = 2.6)  
predict(lm1, newdata = newdata)

## 1   
## 14.85286

head(running)

## Running.Time v4mM vTlac Rel.14.5 Rel.16.1 VO2Max  
## 1 8.23 20.4 19.5 47.1 52.4 23.4  
## 2 8.30 19.5 18.2 48.1 60.0 23.5  
## 3 8.62 19.0 17.3 50.3 56.8 22.0  
## 4 8.82 18.9 17.8 51.8 56.1 23.0  
## 5 9.18 17.8 16.5 48.7 54.1 21.5  
## 6 9.23 17.2 15.6 50.5 59.6 20.5

**10 marks**

## Overall Conclusion and presentation

#### Task 14: State your overall conclusions in relation to the two main aims of the study (3-4 sentences for each)

Aim (1) From the hypothesis test, confidence intervals for the true mean, and tolerance interval. We came to the conclusion that the alternative hypothesis is not true, based on the results of the hypothesis test, which is 10.2.

Answer:- The p value is extremely small, as we can see after running the aforementioned code, so we accept the alternative hypothesis, which states that the true mean is not equal to 10.2 minutes. The projected mean of 10.2 did not fall into the range of values within the one sample t-test 95% confidence interval, which supported the rejection of the null hypothesis. The range of values was 9.061532, 9.854718. The values inside the t-test range were confirmed to be accurate, and the alternative hypothesis was still accepted, as indicated by the bootstrap interval of 9.096812 to 9.802625. Understanding the bootstrapping intervals and t-test allows us to draw the following conclusions:

90% of the population mean has running times below 9.8 minutes with 95% confidence.

**5 marks**

Aim (2) From fitting a linear model for the relationship between 3k running time and the running speed at blood lactate concentration 4 mmol/litre.

Answer:- Given the higher values of v4mM, we can assume shorter running times with a negative correlation coefficient of 0.926. In other words, runners who have higher levels of v4mM run faster than those who have lower levels.

**5 marks**

#### Task 15: Knit your .Rmd file to Word and submit the resulting Word document (alternatively knit to html, save the html as a pdf and submit the pdf). Make sure you have edited the .Rmd to include your own name and student ID.

If you submit an .Rmd file you will **lose 20 marks**, and if you submit a file missing your name and/or ID you will **lose 10 marks**.