3k running assignment

Prachi Tiwari, id=23102057

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 is that the mean 3k running time for a particular population is 10.2 minutes, whereas the alternative hypothesis is that the mean 3k running time for the population is different from 10.2 minutes, using 0.05 being the significance threshold for the test.

**5 marks**

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

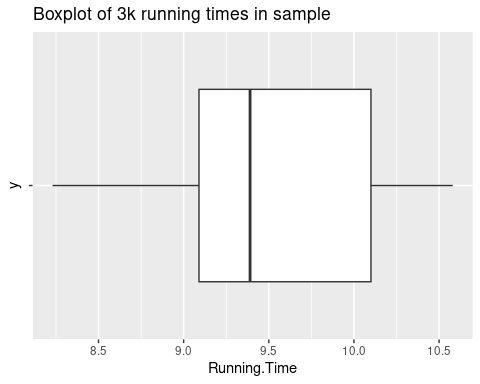
1. What statistical test is appropriate to test the Task 1 hypotheses? State the assumptions needed for the test to be valid, and how they could be justified in this study. (1-2 sentences).

ANSWER: To test the Task1 hypotheses, we can use one-sample t-test. Assumption needed for valid test is data must be normally distributed and data is a random sample from population. It can be justified by assessing the distribution of 3k running times and verifying the equality of variances.

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

ANSWER: Below code output shows partially right skewed boxplot without any outliers, means not equally distributed which contradicts assumption of t-test that populations being compared have approximately normal distributions.

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 yielded a t-value of -3.9871 and a p-value of 0.00119, both of which are less than the accepted significance threshold of 0.05. As a result, we reject the null hypothesis that 3k running time for population is 10.2 minutes in this case.

1. The percent confidence interval range calculated is between 9.061532 and 9.854718 minutes, indicating that we can be 95% certain that the mean 3k running time for the population will fall within this range. Furthermore, the interval excludes the hypothesised mean running duration of 10.2 minutes, which supports the alternative hypothesis.

**5 marks**

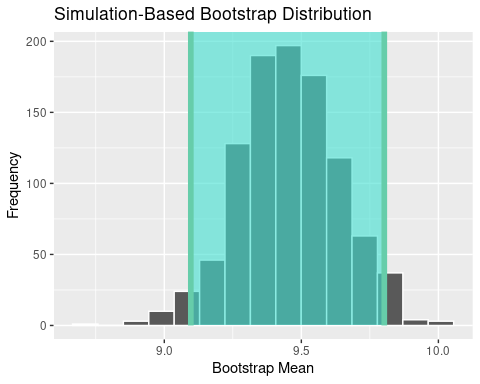
#### Task 5: A bootstrap 95% confidence interval for the mean 3k time

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 95% 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

(1-2 sentences)

ANSWER: The 95% confidence interval generated using the t-distribution approach implies data normality and is best suited for small sample sizes. The 95% confidence interval obtained from bootstrap resampling, on the other hand, makes no assumptions about normality or sample size. When the intervals are close together, it indicates that the t-distribution assumptions are appropriate for the sample data. If the intervals differ greatly, it indicates that the data may not follow the t-distribution assumptions, making the bootstrap interval more reliable.

**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 adjusted R-code generates an output with a tolerance range of 7.64 to 11.28 minutes. As a result, we can be 95% positive that the quickest 10% of our research sample will run 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**

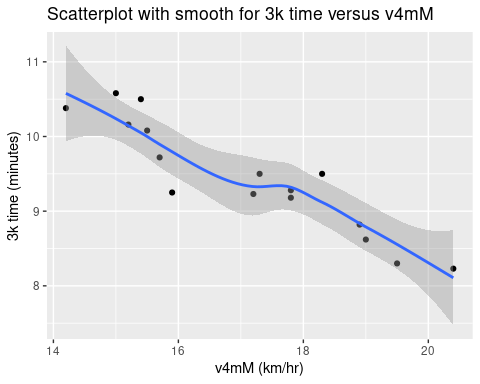
## Linear Regression

#### 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 scatterplot curve exhibiting a linear regression model between 3k time and v4mM suggests a possible link between these two variables. It follows the data points in a falling pattern, implying a credible linear relationship between running duration and v4mM.

running %>% ggplot(aes(x = v4mM, y = Running.Time)) + geom\_point() + geom\_smooth() + labs(title = "Scatterplot with smooth for 3k time versus v4mM", x = "v4mM (km/hr)", y = "3k time (minutes)")



**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: The scatterplot and correlation coefficient analysis show that there is a strong negative association between ‘Running.Time’ and ‘v4mM,’ with a correlation coefficient of -0.925857. The scatterplot’s declining trend emphasises this significant negative connection, demonstrating that as velocity (v4mM) increases, so does running time. This means that runners go faster at greater lactate concentrations.

**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

**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?

(2-3 sentences)

ANSWER: The equation of the line of best fit for the simple linear regression of ‘Running.Time’ on v4mM is: Running.Time = (slope) \* v4mM + (intercept) The slope reflects how much ‘Running.Time’ changes for a unit change in ‘v4mM,’ suggesting that when ‘Running.Time’ lowers, running velocity (‘v4mM’) increases, which is consistent with a negative slope. The intercept, although theoretically the anticipated ‘Running.Time’ when ‘v4mM’ is zero, lacks a meaningful interpretation in this context because running velocity cannot be 0, signalling the runner is in an idle state.

**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?

(1-2 sentences)

ANSWER: The measured running time of 8.82 minutes differs from the expected running time of ‘v4mM’ = 18.9 km/hr. This discrepancy emerges because the projected running time offers an average estimate for persons running at that pace over the whole population, but the observed running time refers to a specific individual’s record. The residual value is -0.18 when computed as the difference between the actual running time (8.82 minutes) and the anticipated running time (9.00 minutes) using the linear regression model. This negative residual suggests that the model overestimated the time required to run this scenario.

lm1<-lm(Running.Time ~ v4mM, data = running)  
newdata <-data.frame(v4mM = 18.9)  
predict(lm1, newdata = newdata)

## 1   
## 8.775336

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**

#### 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)

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

1. Do you have any concern about this prediction, and if so why? (1 sentence)

ANSWER: There is a substantial worry about the prediction when utilising a running speed value for ‘v4mM’ of 2.6 km/hr, which is much lower than the initial value of 18.9 km/hr utilised during linear regression model training. This slower running pace falls outside of the dataset’s coverage range, increasing the likelihood of generating erroneous and misleading findings.

**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.

ANSWER: The hypothesis test and confidence interval analysis support the alternative hypothesis that the population’s mean 3k running time is different from 10.2 minutes. The estimated sample mean of 9.458125 falls below 10.2 minutes at a 95% confidence level, inside the interval range of 9.06 to 9.85, indicating quicker running times for athletes. Concerning the tolerance interval, it is anticipated that 90% of the population’s 3k times will lie within the range of 7.635421 to 11.28083 minutes with 95% confidence. This implies that the majority of athletes run faster than the estimated mean running time.

**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: A substantial and statistically significant correlation coefficient is found when a linear model is established between 3k running time and running speed at a blood lactate concentration of 4 mmol/litre (‘v4mM’). This conclusion implies that increasing running speed is connected with a reduction in running time. The linear regression model, which was used to estimate the mean 3k running time based on ‘v4mM’ running speed, demonstrates ‘v4mM’’s efficacy as a robust predictor in the linear regression model. When dealing with lower running speeds, such as 2.6 km/hr, care is advised because it falls outside the dataset range.

**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**.