**UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG**

**SCHOOL OF PHYSIOLOGY**

**PHSL 4005: BSc (Honours)**

**FINAL EXAMINATION – 31 August 2016**

**EXAM 2** (open book): STATISTICS

Duration: 4 hours

------------------------------------------------------------------------------------------------------------

**INSTRUCTIONS TO CANDIDATES**

For the duration of this examination you will not be confined to the examination venue, but you must stay in the School of Physiology. It is an *open-book* examination, and you may use resources on the internet, textbooks, or [course notes](https://painblogr.org/biostatistics) (<http://painblogr.org/biostatistics/>) to assist you. However, you may not converse (via any medium) with anyone, or solicit help through internet resources such as Stack Exchange. **Cell phones are not allowed during examinations/tests. Calculators may be used.**

An electronic copy of this document, with active links can be located at the following URL: <https://dl.dropboxusercontent.com/u/11805474/painblogr/biostats/exam-data-2016/exam-script.docx>

**YOU MAY NOT DISCUSS THE QUESTIONS OR ANSWERS WITH ANYONE.**

Reproducibility (30)

Question 1: (20)

Question 2: (20)

Question 3: (20)

Question 4: (10)

Total Marks for the Exam = 100

**Examination requirements**

For questions 1 to 3 you will be required to:

1. Download and import datasets from the URLs provided;
2. Tidy the datasets (if required);
3. Generate summary statistics and exploratory figures/tables to characterise the data;
4. Undertake hypothesis testing. In each case, you must clearly state the:
   * Hypothesis being tested;
   * Type of data being analysed;
   * Statistical test(s) chosen to assess the hypothesis/hypotheses;
   * The assumptions of the test(s);
   * Core information describing the output of the statistical test(s);
   * Conclusions you can draw from the analysis.
5. Use your analysis to answer the posed question (where appropriate)

Data provided in questions 2 to 4 of this examination are adapted from real data, the sources of which are available on request (after the examination).

**Assessment of reproducibility**

(**30 marks** will be allocated to this aspect of the examination)

You may use any statistical or graphing software to complete these tasks, but irrespective of the platform used, you **are required** to upload a single RMarkdown document (configured to produce an html output) to a repository (repo) called *'exam-2016'* in your GitHub account. You **must** ensure that all the inputs (e.g., data files) and outputs (e.g., html, figures) produced by the analysis script have been pushed to the repo. The files in the repository **must have at least one commit**. In the case of more than one commit, **only the final commit will be assessed for marks**. A link to the GitHub repository must be emailed to [Peter Kamerman](mailto:peter.kamerman@wits.ac.za) ([peter.kamerman@wits.ac.za](mailto:peter.kamerman@wits.ac.za))

The purpose of this exercise is to assess your ability to undertake reproducible statistical analysis. As such, your repo will be assessed based on whether the repo can be cloned and your analysis reproduced by running the RMarkdown script. The following mark allocation will be used to assess your repo:

* A functional link to a repo entitled, 'exam-2016' was emailed **[5 marks]**;
* The presence in the repo of a single RMarkdown script, with associated input files and output files, all with at least one commit **[5 marks]**;
* Reproducibility of the analyses (i.e., can the outputs of the analysis script be reproduced by cloning the repo and running the script locally) **[10 marks]**;
* Clarity of the script (i.e., are the analyses well commented, such that someone else can follow the data cleaning, graphing, analysis steps followed) **[10 marks]**

*Note: If you do not use R/RStudio to analyse the data, mark allocation will be based on the reproducibility of the methods used relative to that which is achievable when using R/RStudio.*

## Question 1

*Access the data at:* [*https://dl.dropboxusercontent.com/u/11805474/painblogr/biostats/exam-data-2016/question1.csv*](https://dl.dropboxusercontent.com/u/11805474/painblogr/biostats/exam-data-2016/question1.csv)

The *question1.csv* dataset is based on an experiment that measured mean oral body temperature and heart rate at rest in male and female students. The dataset includes 130 observations of the two variables. The data coding is as follows:

* **Column 1:** Body temperature (degrees Celsius).
* **Column 2:** Heart rate in males (beats per minute).
* **Column 3:** Heart rate in females (beats per minute).

What statistical inferences can you make from these data?

[20]

## Question 2

*Access the data at:* [*https://dl.dropboxusercontent.com/u/11805474/painblogr/biostats/exam-data-2016/question2.csv*](https://dl.dropboxusercontent.com/u/11805474/painblogr/biostats/exam-data-2016/question2.csv)

The *question2.csv* dataset includes observations of 151 intoxicated test subjects walking a three metre line. Variables include: the direction to which subjects first stumbled, the side of the line on which they ended up, and a record of subjects' dominant hand. The data coding is as follows:

* **Column 1:** Subject identity number
* **Column 2:** Subject sex (1 = male, 0 = female)
* **Column 3:** Dominant hand (1 = right-handed, 0 = left-handed)
* **Column 4:** Direction of first stumble (1 = to the right, 0 = to the left)
* **Column 5:** Final position relative to the line at the end of the 3m walk (1 = to the right of the line, 0 = to the left of the line)

Using these data, what inferences can you make about handedness and ataxic walking in intoxicated individuals?

[20]

## Question 3

*Access the data at:* [*https://dl.dropboxusercontent.com/u/11805474/painblogr/biostats/exam-data-2016/question3.csv*](https://dl.dropboxusercontent.com/u/11805474/painblogr/biostats/exam-data-2016/question3.csv)

The data in *question.3.csv* were collected by an amateur runner, using a GPS enabled sports watch, when she ran her standard 6.4km training route. The overweight runner was tracking her performance over the course of 19 runs, each spaced two days apart. Variables reported here include run number, running time, and estimated calories consumed.

* **Column 1:** The run number (1-19).
* **Column 2:** Running time (s).
* **Column 3:** Calories consumed (cal).

Can the runner use her running time over the 6.4km route to predict the number of calories she consumed during the run? If so, predict the number of calories consumed if she reduced her time to complete the course to 30 minutes.

[20]

## Question 4

For the previous questions you were required to provide exploratory plots of the data. Exploratory plots are 'quick and dirty' with no tidying of the plots required beyond making them intelligible to the investigators. However, these plots are not normally publication-quality.

Below I have provided you with a script that you can ‘copy-paste’ into RStudio to generate a fake data set that models the relationship between students' project marks (expressed as percent) for a BSc Honours in Physiology course and their final marks (expressed as percent) over a ten-year period. Run the script, and then plot a black and white, publication-quality figure depicting the relationship between project mark and final mark, and save the image in PDF format.

Provide an explanation for the axis scale you chose.

foo <- rnorm(10000, mean = 60, sd = 3) # final mark  
bar <- rnorm(10000, mean = 68, sd = 5) # project mark  
baz <- rep(seq(from = 1997, to = 2006), each = 1) # years  
  
year <- sample(baz, 150, replace = TRUE,  
 prob = c(0.05, 0.05, 0.08, 0.08,   
 0.1, 0.13, 0.14, 0.13, 0.12, 0.12))  
project\_mark <- sample(bar, 150, replace = TRUE)  
final\_mark <- sample(foo, 150, replace = TRUE)  
  
plot\_data <- data\_frame(year = year,  
 project\_mark = round(project\_mark, 1),  
 final\_mark = round(final\_mark, 1)) %>%  
 arrange(year)

[10]