University of Pretoria
Department of Industrial and Systems Engineering

Industrial Analysis
Bedryfsanalise

BAN 313

Internal examiner:
Interne eksaminator: Prof. Johan W. Joubert

External examiner:
Eksterne eksaminator: Mr. Werner W. van Niekerk

Answer all questions on *clickUP*.

Beantwoord al die vrae op clickUP

Complete all **9** questions for **20** marks
*Beantwoord al **9** vrae vir **20** punte*

Total time: 90 minutes
Totale tyd: 90 minute

This is strictly an individual assessment. You are welcome to access any documented material, but no communication with (any) other individual(s) via any mode or means.

Please take note of the last question, which requires that you upload the R/RMarkdown file that you used to complete your calculations. This must be a ***single file***, so ensure that you plan and set up your R session accordingly.

The internal examiner is available during the course of the test on +27 82 338 0565.

Census data

The following questions deal with the given sample as taken from the 2011 Census *Public Use Micro Sample* (PUMS). You are given the metadata (`metadata.pdf`), the record layout (`recordLayout.xls`), the municipal locations (`municipalities.csv`), and the compressed (GZIP) files of persons (`persons.txt.gz`) and households (`households.txt.gz`). For all the questions, merge the `persons.txt.gz` and `households.txt.gz` and only use those complete records for which you have data entries at both individual and household level.

2 1. For how many persons (individuals) do you have complete records? That is, where there are records for the household and its individual members. _____

2 2. What proportion of individual persons born in Mpumalanga make use of a communal refuse dump? Give your answer as a fraction and round your answer to 4 decimal places. For example, if you believe the answer is 12.34%, given your answer as `0.1234` and make sure to use a decimal point (not comma). _____

4 3. Since we only gave you a sample of the census data, give the 98% confidence interval for your estimate in question 2. Similar to the previous question, round your answer for each interval limit to four digits.
Lower limit: _____; upper limit: _____.

2 4. In calculating the confidence limits in 3, what is the *standard error* you used? Similar to the previous questions, round your answer to four digits. _____

2 5. In a press release by the national Department of Environmental Affairs, the government claims that 58% of persons in South Africa receive waste collection services by their local authority at least once a week. Which of the following statements do you agree with?

   A. Their estimate was conservative (too low).
   B. Their estimate was correct.
   C. Their estimate was ambitious (too high).
   D. None of the above. The conditions are not met to allow for inference.

2 6. Motivate your answer in question 5 statistically using a 90% confidence level.

Emissions

The given compressed file, `pems.txt.zip`, contains the emissions field test data as recorded with the Centre of Transport Development's Portable Emissions Measurement System (PEMS). These field tests were conducted using the Road-Rail Vehicle (RRV), a medium-heavy Isuzu FTR850 truck. The linked *YouTube* video shows the PEMS equipment mounted on the back of the truck. Don't waste too much time watching it now; you can watch it after the test. It is just provided for context. The equipment is connected to the vehicle's exhaust and a variety of sensors and onboard vehicle diagnostics are recorded. The goal of the data is to assist in understanding the variation in emissions concentrations under real driving conditions in South Africa.

Multiple trips' data are included in this set and each row represents a single measurement record captured by the PEMS equipment. The first column contains the date and time of the record. With the exception of the first column, all others are numeric values. The data set is not curated (cleaned) so may contain missing values.

The second and third columns indicate the unique trip number, and the driver conducting the trip, respectively. Column 4 indicates the load on the vehicle. Columns 5–8 provide the geospatial positioning system (GPS) readings in the form of the latitude, longitude, altitude and (derived) speed. The next three columns, 9–11, provide the ambient readings from the weather probe. The next 15 columns are variables related to the vehicle diagnostics, and the remainder of the columns deal with the emissions. Answer the following questions using the given data set.

2 7. The variable `co2_mass` gives the instantaneous quantity (in grams, g) of CO_2 (carbon dioxide) emitted for that second. What is the total quantity of CO_2, in *kilograms* (kg), emitted over the course of trip 3? Round your answer to 2 decimal places. For example, if you believe the answer is 1.234kg, give your answer as `1.23` and remember to use a decimal point (not comma). _____

2 8. The variable `humidity` is the ambient humidity (given in % relative humidity). What would be your best estimate for the distribution of the humidity readings during trip 8?

A. Normal distribution with $\mu = 17.22$ and $\sigma = 1.87$.

B. Uniform distribution with min $= 14.1$ and max $= 24$.

C. Exponential distribution with $\lambda = 1/17.22 = 0.058$.

D. Poisson distribution with $\lambda = 17.22$.

E. None of the above.

2 9. Submit your supporting code (R script, RMarkdown document, or compressed folder) as a **single file**, using your student number (with the prefix 'u') as the filename. For example, `u01234567.R`, `u01234567.Rmd`, or `u01234567.zip`.

⋆⋆⋆ **end of paper** ⋆⋆⋆
einde van vraestel

Formulas

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad var = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$\Pr(A^c) = 1 - \Pr(A) \qquad \Pr(\text{A and B}) = \Pr(\text{A}) \times \Pr(\text{B})$$

$$\Pr(\text{A or B}) = \Pr(\text{A}) + \Pr(\text{B}) - \Pr(\text{A and B}) \qquad \Pr(A|B) = \frac{\Pr(\text{A and B})}{\Pr(\text{B})}$$

$$z = \frac{x - \mu}{\sigma} \qquad x = \mu + z\sigma$$

$$Q_1 - 1.5 \times IQR, \quad Q_3 + 1.5 \times IQR$$

$$\hat{p} \pm z_{score} \times SE_{\hat{p}} \qquad SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad z = \frac{\hat{p} - p_0}{SE_0} \qquad SE_0 = \sqrt{\frac{p_0(1-p_0)}{n}}$$

$$\bar{x} \pm t_{score} \times SE_{\bar{x}} \qquad t = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} \qquad SE_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad df = n - 1$$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{score} \times SE \qquad SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE_0} \qquad SE_0 = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{score} \times SE \qquad t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE} \qquad SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad df = \min(n_1 - 1, n_2 - 1)$$

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \qquad \text{expected} = \frac{\text{row} \times \text{column}}{\text{total}} \qquad df = (r-1) \times (c-1)$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad \hat{\beta}_1 = r\left(\frac{s_y}{s_x}\right) \qquad \text{residual} = y - \hat{y} \qquad s = \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}}$$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \qquad t = \frac{\hat{\beta}_1 - 0}{SE_{\hat{\beta}_1}} \qquad \hat{\beta}_1 \pm t_{score} \times SE_{\hat{\beta}_1} \qquad df = n - 2$$