## SAMPLE EXAM QUESTIONS – 23 SEPT – ANSWERS FROM PETER

**Question 1: Short questions**

Mark each statement as **True or False**. Justify your answers and, if false, provide the correct answer.

1. **The following code takes the gapminder data, and produces a scatter plot of life expectancy (lifeExp) vs GDP (gdpPercapita) where all points are coloured blue.** True or False? Justify your answers and, if false, provide the correct answer.

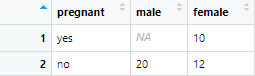




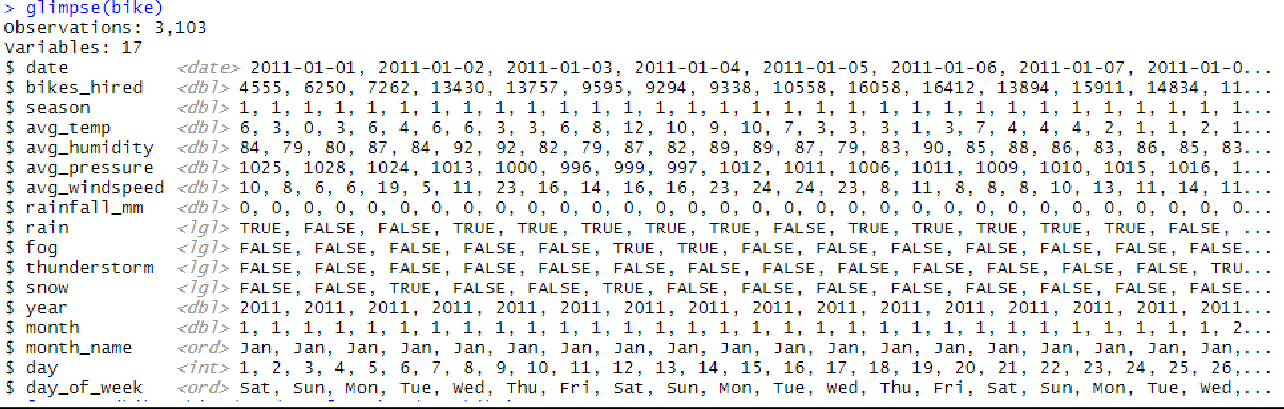


1. **The following dataframe is in tidy format.** True or False? Justify your answers and, if false, provide the code to make the dataframe tidy.





#### The dataframe *bike* contains data on the number of bikes rented out in London. You can glimpse its structure below





**To create a boxplot of *bikes\_hired* on a month-by-month basis, we use**

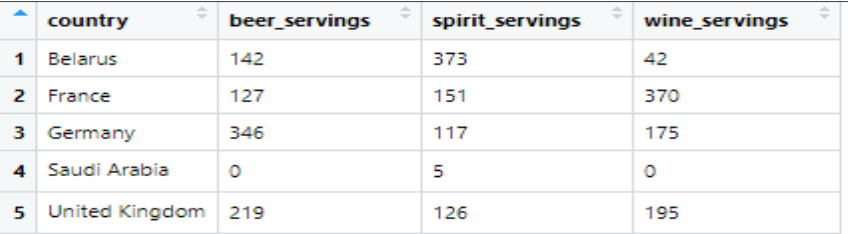
**ggplot**(data = bike, mapping = **aes**(x = month, y = bikes\_hired)) +

**geom\_boxplot**()

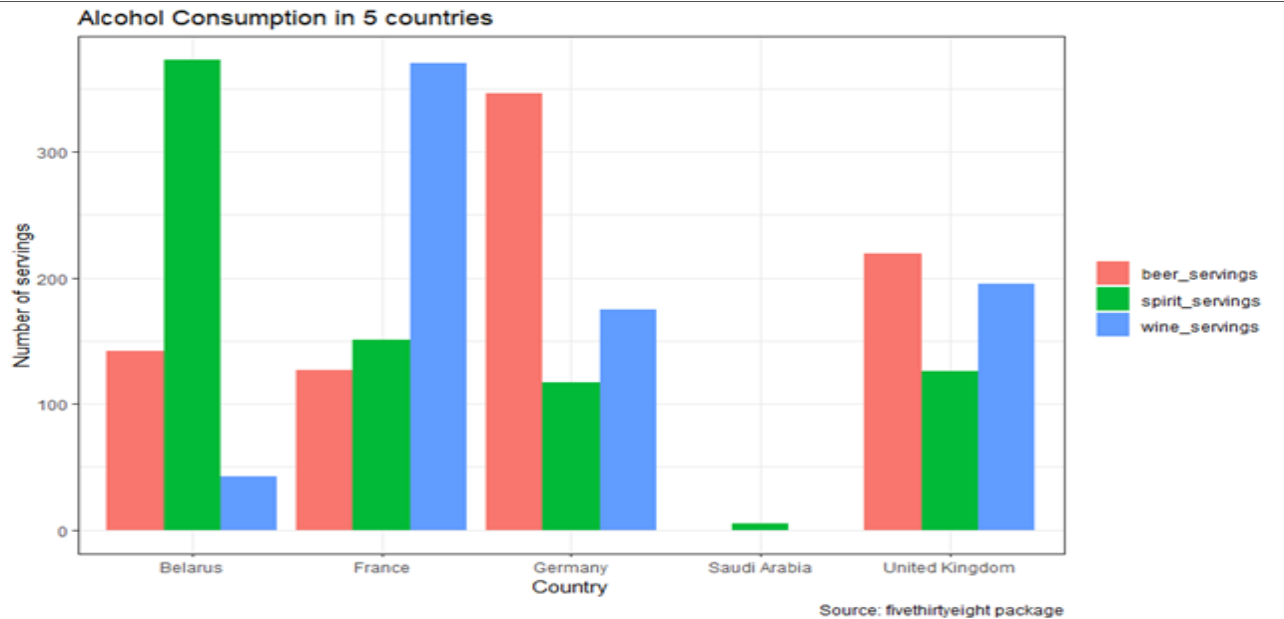
|  |  |
| --- | --- |
|  | |
|  |  |

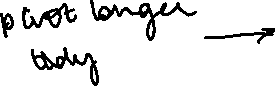
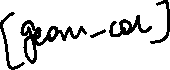


#### The *fivethirtyeight* package, has a dataset *drinks* with data on the number of servings to help us identify where people drink the most beer, wine and spirits. From this dataset se selected a few countries and the resulting dataset, *small\_drinks,* is shown below



**Using tidyverse packages and functions, how you would you create this plot?**





**Question 2**

In determining automobile mileage ratings, R, we rely on the relationship between the distance travelled and the amount of fuel consumed by a vehicle that ascertains the automobiles’ fuel efficiency. The measure used for this purpose is expressed in “miles-per-gallon”, mpg. It was found that the mpg in the city for a certain model is normally distributed, with a mean of 22.5mpg and a standard deviation of 1.5mpg.

1. **You buy a car of this model to drive it mostly in the city. What is the probability that its mpg in the city is more than 24mpg?**



R is Normally distributed N(22.5; 1.5)

CLT: we want standardized Normal variable

P[R > 24mpg] = P[(R-22.5)/1.5 > (24-22.5)/1.5] = P[Z > 1] = (1-0.68)/2 approx 16%.

From tables: 15.87%

* 1. **Probability of X falling between 2 numbers and , is given by:**



**. What is the probability that car’s mpg in the city is between 21.5mpg and 23 mpg?**

P[21.5 < R < 23] = P[(21.5-22.5)/1.5 < Z < (23-22.5)/1.5] = P[-0.67 < Z < 0.33] = 1 – 37.07% - 25.14% = 37.79%

* 1. W**hat is the probability that its mpg in the city equals to the average, i.e., exactly 22.5mpg?**



P[R = 22.5] = ?

Probability is 0. Why? The R is continuous random variable.

1. **Find the mileage rating that the upper 5% of the cars of this model achieve.**



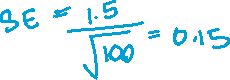
P[R > x ] = 5%

P[Z > (x-22.5)/1.5] = 5%

(x – 22.5)/1.5 = 1.645, solve equation for x.

Answer: x = 24.9675mpg.

1. **Suppose that the car manufacturer of this model, samples 100 cars from its assembly line and tests them for mileage ratings. What is the probability that the sample mean will be greater than 21mpg? The standard deviation of this sample of 100 cars is 1.5 mpg.**



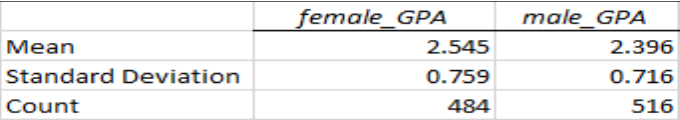
Standard error (SE) = 1.5/sqrt(100)

Sample mean is W.

Then P[W> 21] = P[(W-22.5)/0.15 > (21-22.5)/0.15] = P(T stats > -10) = 1.

# Question 3

We want to study whether there is any difference in male and female first year GPAs at US colleges and universities. We collected a sample of 1000 students and the summary statistics are given below:





#### State what is the population, the sample, the parameter you want to infer, and the available sample statistic



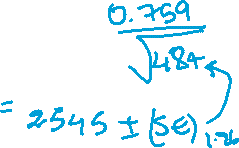
The population represents all first year students at US col and uni.

The sample is 1000 students and their GPAs (484 females, and 516 males).

Parameter is the population difference of female – male GPAs, \mu\_fem - \mu\_male

Sample statistics = 0.149.

1. **Construct two 95% confidence intervals; one for the mean female GPA and one for the mean male GPA. Do you have to make any assumptions?**



SE\_female = 0.759/sqrt(484) = 0.0345

SE\_male = 0.716/sqrt(516) = 0.032

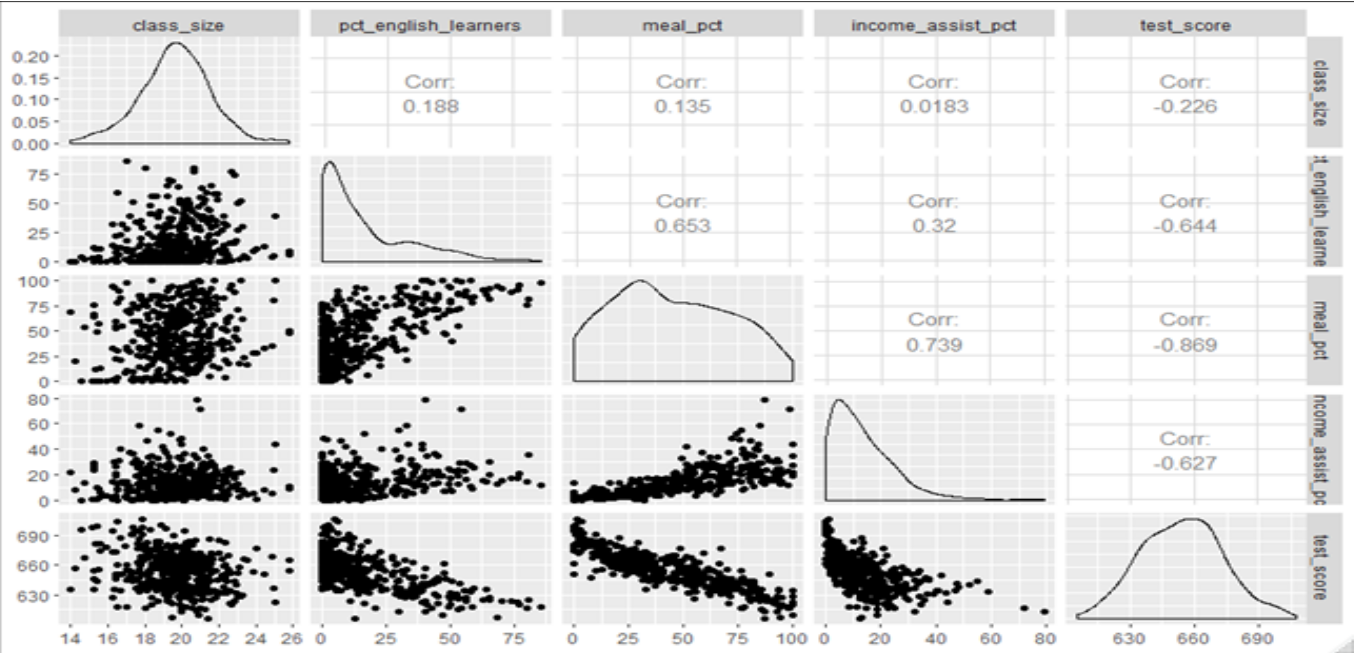
CI\_female: \mu\_female +- 1.96\*SE\_female = [2.477; 2.612]

CI\_male: \mu\_male +-1/96\*SE\_male = [2.334; 2.458]

Confidence intervals do not overlap. So, I would REJECT THE H0! No assumptions are needed. Since there is sufficient sample size n>>30; the sample mean follows a NORMAL DISTRIBUTION. (CLT applies).

1. **Based on this sample, test whether or not the mean difference of GPAs for first year students is the same or not. Use a 5% significance level. Conduct a hypothesis testing, state the null and the alternative, calculate a t-statistic for the difference, and finally state what you decide/infer.**



**Question 4**



H0: \diff (\mu\_female - \mu\_male) = 0 vs

H1: \diff (\mu\_\female - \mu\_male) doesn’t equal 0.

SE for the difference: sqrt (0.759^2/484+0.716^2/516) ~ = 0.047

So, T stats will be then T = (Evidence – H0 Claim)/SE = (0.149-0)/0.047 = 3.19.

With confidence level of 95% we can reject the null hypothesis since T-stats is greater (in absolute terms) than T critical (approx. 2).

There is a significant difference in mean of GPAs between men and female in the first year students at US colleges.

(T stats > T critical, the p-value should be smaller than 5%. Conclusion: Reject the H0!)

We wanted to examine the effect small class sizes have on standardised test scores. We collected data from 420 elementary school districts in California and the following table shows a scatterplot- correlation matrix of all available variables

**Table 3.1 Scatter plot - correlation matrix of available variables**

**The variables are as follows:**

### **class:** average class size in school district

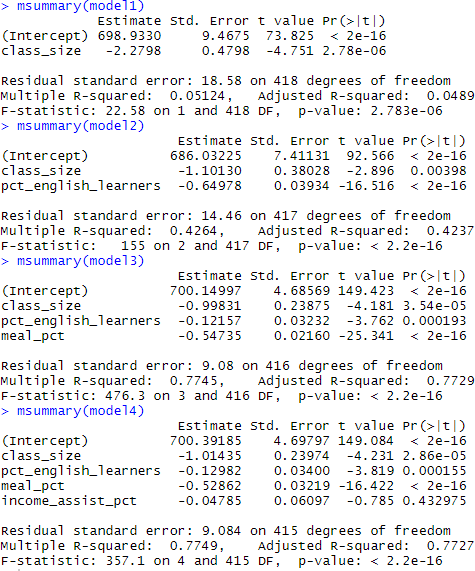
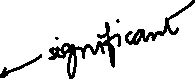
* + **pct\_enlish\_learners:** percentage of students in the district for whom English is not their native language
  + **meal\_pct:** percentage of students in the district receiving free school meals
  + **income\_assist\_pct:** percentage of students whose families were in an income support programme

### **test\_score:** the average standardised score in the school district

In addition, in our quest to understand what explains variability in ***test\_score***, we have run four regression models, 1 through 4, the summary results of which are shown in table 3.2 below.



**Table 4.2 Four regression models**



#### Looking at model 1, is class size a significant predictor of *test score*? What proportion of the overall variability in *test\_score* does *class\_*size explain



Model 1: T-stats for class\_size = -2.2798/SE = -4.751, (R output, p-value is extremely small <5%)

The variable the class\_size is significant predictor of the test\_score.

R2 is about 5%, so the class\_size explain only 5% of overall variability in the test\_score.

1. **Consider model2. Are both explanatory variables significant? What is the proportion of variability in**

***test score* that is explained by model 2? What is the effect of class size and why has it changed?**

Model 2:

Class\_size T stats = -1.101/0.38 = -2.896

PCT\_english\_learners T stats = -0.6497/0.0393 = -16….

Both explanatory variables are significant, T STATS (in abs terms) > 2. And, we are now able to explain more than 42% of overall variance.

--

%English\_learners brings new and more information; evidence is the correlation matrix with positive correlation between the class\_size and %English\_learners corr(X1,X2)=0.188!

Correlation between the test\_score and %English\_learners corr(Y, X2) is also significantly large and negative -0.64

The effect of the class\_size reduces from -2.28 to -1.10!

1. **Consider models 3 and 4. Which one do you choose and why? Given your choice, predict the test score a school district with class size = 22, pct\_english\_learners = 25, meal\_pct = 60, and income\_assist\_pct = 10 is likely to get and give an approximate 95% prediction interval.**

In Model 4, for variabke income\_assit\_pct T stats = -0.785 and -2 <-0.785<2, this variable is not significant. There is a hight negative correlation between the test\_score and the meal\_pct; secondly, there is also high positive correlation between two explanatory variables, the meal\_pct and the income\_assit 0.739 which may/could result in the co-linearity issues in MODEL 4!

So, Model 3 is better from inference point of view. All explanatory variables are significant! The ADJ R2 can explain 77.29+% of overall variance (the highest).

In R: lm.predict()

MODEL 3;

TEST\_SCORE\_ESTIMATOR = 700.14 – 0.99\*22- 0.12\*25-0.54\*60 =

(point-wise prediction)

95% of your prediction: TEST\_SCORE\_ESTIMATOR +- 2(T Critical)\*9.08 (Residual Standard Error) = [624.15, 660.5] (Alex)

1. **Looking again at your best model, the teachers’ union claims that reducing class size by five (5) students, will improve test scores by at least 25 points. Do you agree or disagree with this claim?**

**We are looking at MODEL 3!**

**The coefficient of the class\_size is -0.998, so the effect of the adding five students will be approximately 5\*0.998 ~ 5.**

**Our 95% CI will change. How? 5+-2\*9.08 = [-13, 23], and you can easily see that 25 points is outside of our updated 95% CI - > that means we disagree the claim of the teacher’s UNION!**

*Different approach:*

*Fitted new value for score: TEST\_SCORE\_ESTIMATOR\_SMALLER\_CLASS = 700.14 – 0.99\*17- 0.12\*25-0.54\*60 = 647.91*

*Fitted new value for score 95% CI: [630, 666] difference from (iii) in CI [+6, +6]*

Different approach

Firstly start with CI for the coefficient of the class\_size from Model 3: -0.998 +-2\*0.2389…

To compare more linear regressions you use ANOVA()

**Table 1: The standard Normal distribution P(Z>z)**



X

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Z = (X - μ)/σ** | **0.00** | **0.01** | **0.02** | **0.03** | **0.04** | **0.05** | **0.06** | **0.07** | **0.08** | **0.09** |
| **0.00** | 0.5000 | 0.4960 | 0.4920 | 0.4880 | 0.4840 | 0.4801 | 0.4761 | 0.4721 | 0.4681 | 0.4641 |
| **0.10** | 0.4602 | 0.4562 | 0.4522 | 0.4483 | 0.4443 | 0.4404 | 0.4364 | 0.4325 | 0.4286 | 0.4247 |
| **0.20** | 0.4207 | 0.4168 | 0.4129 | 0.4090 | 0.4052 | 0.4013 | 0.3974 | 0.3936 | 0.3897 | 0.3859 |
| **0.30** | 0.3821 | 0.3783 | 0.3745 | 0.3707 | 0.3669 | 0.3632 | 0.3594 | 0.3557 | 0.3520 | 0.3483 |
| **0.40** | 0.3446 | 0.3409 | 0.3372 | 0.3336 | 0.3300 | 0.3264 | 0.3228 | 0.3192 | 0.3156 | 0.3121 |
| **0.50** | 0.3085 | 0.3050 | 0.3015 | 0.2981 | 0.2946 | 0.2912 | 0.2877 | 0.2843 | 0.2810 | 0.2776 |
| **0.60** | 0.2743 | 0.2709 | 0.2676 | 0.2643 | 0.2611 | 0.2578 | 0.2546 | 0.2514 | 0.2483 | 0.2451 |
| **0.70** | 0.2420 | 0.2389 | 0.2358 | 0.2327 | 0.2296 | 0.2266 | 0.2236 | 0.2206 | 0.2177 | 0.2148 |
| **0.80** | 0.2119 | 0.2090 | 0.2061 | 0.2033 | 0.2005 | 0.1977 | 0.1949 | 0.1922 | 0.1894 | 0.1867 |
| **0.90** | 0.1841 | 0.1814 | 0.1788 | 0.1762 | 0.1736 | 0.1711 | 0.1685 | 0.1660 | 0.1635 | 0.1611 |
| **1.00** | 0.1587 | 0.1562 | 0.1539 | 0.1515 | 0.1492 | 0.1469 | 0.1446 | 0.1423 | 0.1401 | 0.1379 |
| **1.10** | 0.1357 | 0.1335 | 0.1314 | 0.1292 | 0.1271 | 0.1251 | 0.1230 | 0.1210 | 0.1190 | 0.1170 |
| **1.20** | 0.1151 | 0.1131 | 0.1112 | 0.1093 | 0.1075 | 0.1056 | 0.1038 | 0.1020 | 0.1003 | 0.0985 |
| **1.30** | 0.0968 | 0.0951 | 0.0934 | 0.0918 | 0.0901 | 0.0885 | 0.0869 | 0.0853 | 0.0838 | 0.0823 |
| **1.40** | 0.0808 | 0.0793 | 0.0778 | 0.0764 | 0.0749 | 0.0735 | 0.0721 | 0.0708 | 0.0694 | 0.0681 |
| **1.50** | 0.0668 | 0.0655 | 0.0643 | 0.0630 | 0.0618 | 0.0606 | 0.0594 | 0.0582 | 0.0571 | 0.0559 |
| **1.60** | 0.0548 | 0.0537 | 0.0526 | 0.0516 | 0.0505 | 0.0495 | 0.0485 | 0.0475 | 0.0465 | 0.0455 |
| **1.70** | 0.0446 | 0.0436 | 0.0427 | 0.0418 | 0.0409 | 0.0401 | 0.0392 | 0.0384 | 0.0375 | 0.0367 |
| **1.80** | 0.0359 | 0.0351 | 0.0344 | 0.0336 | 0.0329 | 0.0322 | 0.0314 | 0.0307 | 0.0301 | 0.0294 |
| **1.90** | 0.0287 | 0.0281 | 0.0274 | 0.0268 | 0.0262 | 0.0256 | 0.0250 | 0.0244 | 0.0239 | 0.0233 |
| **2.00** | 0.0228 | 0.0222 | 0.0217 | 0.0212 | 0.0207 | 0.0202 | 0.0197 | 0.0192 | 0.0188 | 0.0183 |
| **2.10** | 0.0179 | 0.0174 | 0.0170 | 0.0166 | 0.0162 | 0.0158 | 0.0154 | 0.0150 | 0.0146 | 0.0143 |
| **2.20** | 0.0139 | 0.0136 | 0.0132 | 0.0129 | 0.0125 | 0.0122 | 0.0119 | 0.0116 | 0.0113 | 0.0110 |
| **2.30** | 0.0107 | 0.0104 | 0.0102 | 0.0099 | 0.0096 | 0.0094 | 0.0091 | 0.0089 | 0.0087 | 0.0084 |
| **2.40** | 0.0082 | 0.0080 | 0.0078 | 0.0075 | 0.0073 | 0.0071 | 0.0069 | 0.0068 | 0.0066 | 0.0064 |
| **2.50** | 0.0062 | 0.0060 | 0.0059 | 0.0057 | 0.0055 | 0.0054 | 0.0052 | 0.0051 | 0.0049 | 0.0048 |
| **2.60** | 0.0047 | 0.0045 | 0.0044 | 0.0043 | 0.0041 | 0.0040 | 0.0039 | 0.0038 | 0.0037 | 0.0036 |
| **2.70** | 0.0035 | 0.0034 | 0.0033 | 0.0032 | 0.0031 | 0.0030 | 0.0029 | 0.0028 | 0.0027 | 0.0026 |
| **2.80** | 0.0026 | 0.0025 | 0.0024 | 0.0023 | 0.0023 | 0.0022 | 0.0021 | 0.0021 | 0.0020 | 0.0019 |
| **2.90** | 0.0019 | 0.0018 | 0.0018 | 0.0017 | 0.0016 | 0.0016 | 0.0015 | 0.0015 | 0.0014 | 0.0014 |
| **3.00** | 0.0013 | 0.0013 | 0.0013 | 0.0012 | 0.0012 | 0.0011 | 0.0011 | 0.0011 | 0.0010 | 0.0010 |

**Table 2: The standard Normal distribution P(Z<z)**

