Statistical Computing

Ma Data Science — Prof. Tim Downie

Practice Exam – January 2020

Surname, forenames: Matriculation nr.:

Allowed material: Formula booklet, two sheets (four sides) of A4-Paper with *hand written* notes, calculator, pen, pencil ruler and blank paper.

Only use pencil for diagrams, all other writing should be done with non-erasable pen. Correction fluid (Tipp-Ex etc.) is not allowed. **Write your name on each page**. Clearly label each exam question number/parts on your exam script. Please leave a few lines between your answers to each question for marking purposes.

The duration is 90 minutes. The marks from your R-Test will be added to your marks from this Exam. To pass the course you need 50 marks in total.

Important: Many marks are given for your working. Make sure you hand in all relevant working and calculations.

Do not write in this section:

Question	1	2	3	4	Test	Total*	Grade
Marks							
Maximum	x	x	x	x	20	100	

NB: the marks per question are only allocated for the actual Exams

* part-marks will be rounded up to the nearest integer

Provisional grading scheme

1.0	95	–		2.0	80	–	84	3.0	65	–	69	4.0	50	–	55
1.3	90	–	94	2.3	75	–	79	3.3	60	–	64				
1.7	85	–	89	2.7	70	–	74	3.7	55	–	59	5.0	0	–	49

Question 1 Descriptive Statistics and Simulation **x Marks**

(a) Three samples are given below. You are told that in each case the arithmetic mean to two significant figures is one of the five following values: 3.65, 7.99, 14.18, 19.08, 20.40
Write in the corresponding mean value in the column provided.

									Mean
Sample 1	10.73	10.86	12.18	12.67	15.35	17.02	17.11	17.55	
Sample 2	2.01	3.65	4.20	4.93	5.53	10.55	14.33	18.72	
Sample 3	16.79	17.44	18.42	19.26	19.68	20.18	20.36	20.52	

(b) Three further samples are given below. You are told that in each case the standard deviation is one of the following five values -0.15, 0.21, 2.85, 24.01, 105.23. Write in the corresponding standard deviation and variance in the columns provided.

									Std. Dev.	Variance
Sample 4	27.44	29.36	32.35	48.39	67.66	71.62	79.16	87.27		
Sample 5	1.93	1.97	3.70	4.01	4.26	7.16	8.41	9.35		
Sample 6	0.13	0.34	0.55	0.57	0.58	0.70	0.71	0.78		

(c) Calculate the median, lower quartile and upper quartile for the following sample of 20 values.

6.75	15.83	17.23	18.73	20.71	22.31	24.71	24.88	29.41	30.39
30.99	31.46	32.66	33.09	36.73	37.27	42.32	43.86	45.74	50.99

(d) A random sample $x_1, \ldots, x_{10}$ has the descriptive statistics given in the following table. The x values are transformed using the formula $y_i = 100 + 5x_i$. Write in the corresponding descriptive statistics for the y values.

	x	y
Mean	23.0	
Median	22.5	
Variance	17.64	
Standard deviation	4.2	

(e) The following random sample from a normal distribution was generated by the given R Code.

```
> set.seed(101)
> rnorm(5)
[1] -0.3260365  0.5524619 -0.6749438  0.2143595  0.3107692
```

(i) What are the two parameters for this normal distribution.

(ii) Give an arithmetic formula, which converts the above sample into a sample from a $N(10, 4)$ distribution.

(iii) Give an R-command to simulate 5 random numbers from a $N(10, 4)$ distribution.

(iv) Explain the purpose of the command `set.seed(101)` in the above code.

Question 2 R Programming **x Marks**

Answer the following questions relating to this R code.

```
my.func<-function(x){
  n<-length(x)
  xx<-sort(x)
  if(n %% 2 == 0) return(sum(xx[n/2+0:1])/2)
  xx[(n+1)/2]
}
> y<-c(2,3.5,5,4.2,10.1)
> my.func(y)
[1] 4.2
```

(a) What is the "`%%`" operator in the fourth line?

(b) Explain what property of `x` is being tested in the logical condition `(n %% 2 == 0)`.

(c) Why does the command `return` only need to be specified once, and why is it not necessary to specify `if(n %% 2 == 1)` on the last line of the function?

(d) What is the usual name given the statistic that `my.func` represents?

(e) Why in the thrid line are the data sorted using the command `xx<-sort(x)`?

Question 3 Confidence Interval **x Marks**

An experiment was carried out investigating the strength of steel straps used for loading pallets. From a sample of 28 steel straps the following statistics were obtained (the units are in Newtons).

Sample mean: $1042N$ Sample standard deviation $51N$

(a) Obtain a 95% confidence interval for the population mean μ. The relevant t-distribution quantile is $t = 2.05$.

(b) What is the R-command to obtain the t-distribution quantile in part (a)?

(c) Give an exact interpretation of the confidence interval in the context of this experiment.

(d) How does the width of the confidence interval change, when the confidence level is decreased?

(e) How does the width of the confidence interval change, when the sample size is increased?

(f) The producer of the steel straps would like to use these data to test whether the population mean is 1050 Newtons.

(i) State the null and alternative hypotheses for this test.

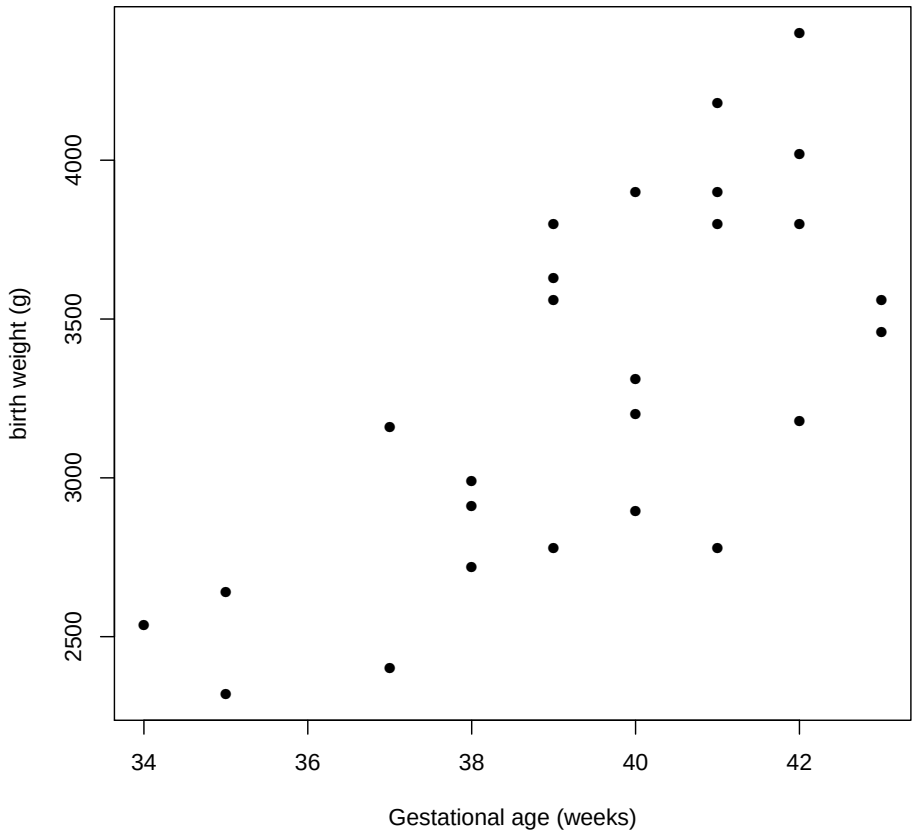
(ii) What is possible to conclude about this hypothesis test from your confidence interval, without needing to do any further calculations? Include a brief explanation of your conclusion.

Question 4 Regression **x Marks**

Here are some data for 26 babies born in a London hospital in a particular week. The babies are all boys of the same race. The data are their birth weights in grammes (y) and gestational age (length of pregnancy) to the nearest week (x).

Gestational Age	42	41	39	40	40	40	39	39	41	42
Birth Weight	3180	2780	3630	3900	3310	2896	2780	3800	3900	4020
Gestational Age	41	43	42	41	38	37	38	43	35	37
Birth Weight	4180	3460	4400	3800	2990	3160	2720	3560	2640	2400
Gestational Age	35	38	40	42	39	34				
Birth Weight	2320	2910	3200	3800	3560	2538				

Here is a scatter plot of the data.



The following calculations are provided:
$\overline{x}$=39.46, $\overline{y}$=3301.3, s_x=2.45, s_y=578.3 and the covariance of x and y is s_{xy}=1000.8

A regression line of the form $y = a + bx$ is to be fitted.

(a) Calculate the regression coefficients a and b.

(b) Give an interpretation of the regression coefficients a and b, in terms of gestational age and birth weight.

(c) Calculate the fitted value for the first observation and the last observation.

(d) Use your answers from part (c) to draw the regression line on the scatter plot.

(e) Calculate the residual for the first observation.