THE UNIVERSITY OF SYDNEY

Room Number _____

Seat Number _____

Student Number |__|__|__|__|__|__|__|__|__|

ANONYMOUSLY MARKED

(Please do not write your name on this exam paper)

CONFIDENTIAL EXAM PAPER

This paper is not to be removed from the exam venue

Mathematics and Statistics

EXAMINATION

Semester 1 - Final, 2023

MATH1005-1 Statistical Thinking with Data (Paper)

EXAM WRITING TIME: 1.5 hours

READING TIME: 10 minutes

EXAM CONDITIONS:
1. Closed book: no reference materials/resources are permitted

MATERIALS PERMITTED IN THE EXAM VENUE:
(No electronic aids are permitted e.g. laptops, phones)

Calculator - handheld

MATERIALS TO BE SUPPLIED TO STUDENTS:
Answer sheet: Gradescope MCQ

INSTRUCTIONS TO STUDENTS:
This examination has two sections: Multiple Choice and Extended Answer.

The Multiple Choice Section is worth 50% of the total examination.

The Extended Answer Section is worth 50% of the total examination.

Answers to the Multiple Choice questions must be entered on the Multiple Choice Answer Sheet.

Please tick the box to confirm that your examination paper is complete. ☐

For Examiner Use Only

Q	Mark
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	

Total _____

SOLUTIONS

Spacer (Multiple Choice Section)

Multiple Choice Section

This section contains 20 multiple choice questions. Select the correct option.

- Each question has exactly one correct answer.
- Each question is of equal value.
- There are no penalties for wrong answers.
- Space for working is provided, but working will not be marked.

Answers must be entered on the Multiple Choice Answer Sheet.

Question 1 (Mean and SD) Points: 1.5

In a dataset of size 8, the mean is 7 and and standard deviation is 4. We add 4 to each observation in the dataset. The new mean and SD are respectively

(A) 7 and 4

(B) 11 and 8

(C) 7 and 8

→ (D) 11 and 4

Question 2 (Mean) Points: 1.5

In a survey of 100 people, the average number of hours of exercise per week was found to be 3.5. If an additional 50 people who exercise an average of 5 hours per week are added to the survey, what is the new average number of hours of exercise per week?

(A) 3.8

→ (B) 4.0

(C) 4.2

(D) 4.4

Question 3 (Probability) Points: 1.5

A deck with 52 cards is shuffled and two cards are dealt without replacement. There are four suits: diamonds, clubs, hearts and spades, with 13 cards of each. What is the chance that both cards are diamonds? Round to 2 d.pl.

(A) 0.02

(B) 0.04

→ (C) 0.06

(D) 0.08

SOLUTIONS

turn to page 3

Question 4 (Quantiles) Points: 1.5

Select the correct statement from below.

(A) If a histogram is skewed to the left, then the median is smaller than the mean.

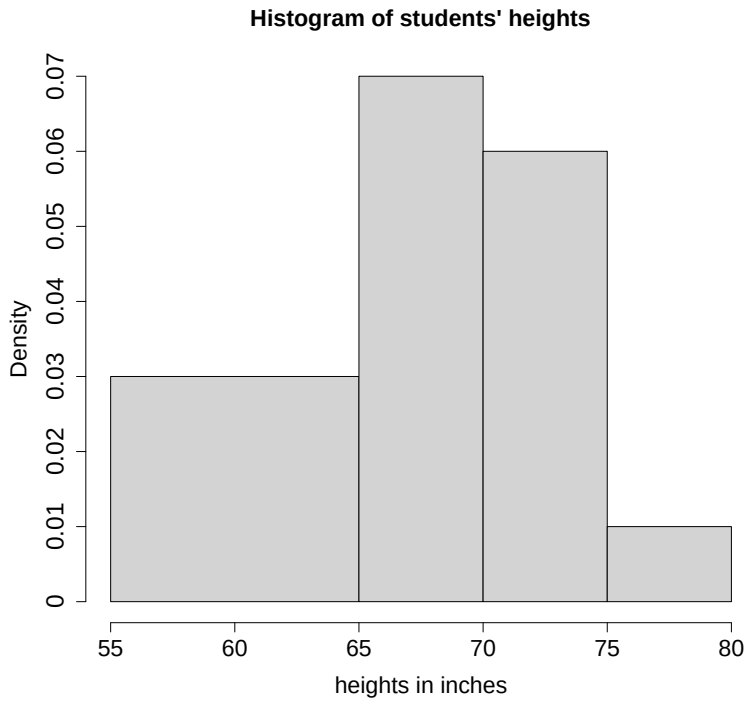
(B) If two lists of numbers have the same mean and same SD, then they should have the same 1st quartile.

→ (C) Let A and B be the 80th percentiles of two different normal curves. Both curves have mean 0, and the standard deviations are 1 and 5, respectively. Then A is smaller than B.

(D) Median is more sensitive to extreme values than mean.

Question 5 (Histogram) Points: 1.5

The histogram shows the distribution of students' heights in a class. Suppose we decide to merge the bins for students of height 55-65 and 65-70 into one bin and redraw the histogram. The height of the new bin 55-70 is:



→ (A) 0.043

(B) 0.055

(C) 0.05

(D) 0.053

Question 6 (Normal distribution) Points: 1.5

The daily sales of a typical sales person at a local clothing store can be assumed to follow the normal curve, with a mean of \$720 and a standard deviation of \$50. Daniel is a salesman in the store. His current sales is at the 25th percentile. If he would like to get to the 75th percentile, how much more does he have to sell? Select the line of R code that will give the correct answer.

SOLUTIONS

turn to page 4

Question 6 (continued)

(A) `qnorm(0.25, 720, 50)-qnorm(0.75, 720, 50)`

(B) `pnorm(0.25, 720, 50)-pnorm(0.75, 720, 50)`

(C) `pnorm(0.75, 720, 50)-pnorm(0.25, 720, 50)`

→ (D) `qnorm(0.75, 720, 50)-qnorm(0.25, 720, 50)`

Question 7 (Normal distribution) Points: 1.5

The weight of the box is normally distributed with a mean of 10 pounds and a variance of 4 pounds^2. The box is going to be shipped to a country where weights are commonly reported in kilograms. To convert from pounds to kilograms, one must use the formula kg = lb/2.205, where kg is the weight in kilograms and lb is the weight in pounds. What is the probability that the weight of the box in kilograms is greater than 4 kilograms, which is the weight limit for packages being shipped to that country? (Round to 2 d.pl.) You may find the following R output useful:

```
> pnorm(0.295)
[1] 0.6160031
> pnorm(0.59)
[1] 0.7224047
> pnorm(0.89)
[1] 0.8132671
```

→ (A) 0.72

(B) 0.62

(C) 0.38

(D) 0.81

Question 8 (Correlation) Points: 1.5

Select the correct statement from below.

(A) If the correlation coefficient between two variables X and Y is $r = 0.95$, it means 95% of the points in the scatterplot lie on a straight line.

(B) If $r = 0$ for two variables X and Y, it means there is no relationship between them.

→ (C) In a University class, the correlation between the quantitative two variables, *students' age* and *students' birth year*, is -1.

(D) More than one of the other listed options is correct.

Question 9 (Correlation) Points: 1.5

Consider the following two tables with pairs of values for x and y,

x	y
2	4
4	1
4	5
8	3

x	y
1	9
2	3
2	11
4	$*$

What is the value of $*$ in the second table so that the two tables have the same correlation coefficient

SOLUTIONS

turn to page 5

Question 9 (continued)

between x and y?

(A) 3

→ (B) 7

(C) 8

(D) 12

Question 10 (Causation) Points: 1.5

In one large introductory class, it was noted that students who attended tutoring after the midterm experienced a larger increase in their final scores than students who did not. Is this convincing evidence that tutoring helps?

(A) We can conclude that tutoring helps based on the data.

(B) We can conclude that tutoring helps if the class is large enough.

→ (C) We cannot conclude that tutoring helps because this is an observational study.

(D) We cannot conclude that tutoring helps because the observed margin of increase in the final scores might not be large enough.

Question 11 (what-is-SE) Points: 1.5

A random sample of size $n = 16$ is to be taken with replacement from a box containing $N = 50$ tickets, each bearing a number. The list of numbers has SD $\sigma = 3$. The standard error of the **sum** of the numbers drawn is

(A) $\frac{12\sqrt{34}}{7}$

(B) $\frac{3\sqrt{34}}{28}$

→ (C) 12

(D) $\frac{3}{4}$

Question 12 (SE-larger-smaller) Points: 1.5

A random sample of size $n > 1$ is taken from a box of tickets, each bearing a number. The standard deviation of the numbers on the tickets is σ. The standard error of the **average** of numbers drawn is

(A) equal to σ

(B) equal to σ/n

→ (C) smaller than σ

(D) larger than σ

Question 13 (Box-predict-sum) Points: 1.5

One hundred draws will be made at random with replacement from one of the following boxes: Box 1 has tickets 3 and 7; Box 2 has tickets 4 and 6. You are asked to guess what the sum will be, and you win \$1 if you are right to within 10. Which Box maximises your chances, and (to within ±2%) what is the chance

Question 13 (continued)

you win?

(A) Box 2 and 96%

→ (B) Box 2 and 69%

(C) Box 1 and 69%

(D) Box 1 and 96%

Question 14 (CLT) Points: 1.5

A random sample of size 400 is taken from a large population with mean $\mu = 12$ and standard deviation $\sigma = 2.5$. The probability that the sample average is between 11.875 and 12.25 is closest to

→ (A) 81%

(B) 67%

(C) 95%

(D) 47.5%

Question 15 (which-hyp) Points: 1.5

In a "taste test" challenge, 100 randomly sampled students try two different drinks, Brand A and Brand B (in a double-blind arrangement) to see which drink they prefer. If p denotes the proportion of the student population who prefer Brand A, which alternative hypothesis should be used to examine the question: "Is one of the brands more popular than the other?"?

→ (A) $p \neq 0.5$

(B) $p = 0.5$

(C) $p > 0.5$

(D) $p < 0.5$

Question 16 (z-or-t-stat) Points: 1.5

A random sample of size 100 is taken from a large population. The sample mean is 26.4 while the sample standard deviation is 7.0. The value of a t-statistic for testing the null hypothesis that the population mean equals 25 is given by

→ (A) 2

(B) 3.77

(C) 2.8

(D) 5.28

Question 17 (chi-sq-dice) Points: 1.5

A six-sided die, with faces numbered 1, 2, 3, 4, 5, 6, is rolled 30 times. The numbers rolled each time are stored in the R vector `rolls`. Using the R output below, determine the value of the χ^2 test statistic for testing that each face is equally likely.

SOLUTIONS

turn to page 7

Question 17 (continued)

```
> rolls
 [1] 6 4 1 2 1 3 2 4 2 5 1 3 4 4 5 6 1 5 4 3 5 4 1 2 6 5 6 6 4 5
> table(rolls)
rolls
1 2 3 4 5 6
5 4 3 7 6 5
> sum((table(rolls)-5)^2)
[1] 10
```

→ (A) 2

(B) 10

(C) 2.4

(D) 12

Question 18 (chi-sq-indep) Points: 1.5

Consider the table below which classifies a random sample of 100 people according to gender and handedness.

	Left-handed	Right-handed	Total
Female	4	44	48
Not female	7	45	52
Total	11	89	100

For testing the hypothesis that gender and handedness are independent, what is the correct number of degrees of freedom?

→ (A) 1

(B) 2

(C) 3

(D) 4

Question 19 (pooled-sd-est) Points: 1.5

A two-sample t-test is to be performed. One sample of size 15 has a standard deviation of 5.61. The other sample of size 32 has a standard devation of 3.78. Both samples are assumed to be randomly drawn from normal populations with a common SD σ. For the purposes of the t-test, what value (rounded to 2 decimal places) is used to estimate σ? **Note:** it is possible to deduce the correct answer without using a calculator.

(A) 3.78

→ (B) 4.43

(C) 4.70

(D) 5.61

SOLUTIONS

turn to page 8

Question 20 (t-based-CI) Points: 1.5

A random sample of size $n = 20$ is taken from a normal population with mean μ and SD σ both unknown. Consider the following R output, where the sample values are stored in `x`:

```
> mean(x)
[1] 14.83756
> sd(x)
[1] 2.975263
> qt(c(0.90, 0.95, 0.975, 0.98, 0.99, 0.995), df = 18)
[1] 1.330391 1.734064 2.100922 2.213703 2.552380 2.878440
> qt(c(0.90, 0.95, 0.975, 0.98, 0.99, 0.995), df = 19)
[1] 1.327728 1.729133 2.093024 2.204701 2.539483 2.860935
> qt(c(0.90, 0.95, 0.975, 0.98, 0.99, 0.995), df = 20)
[1] 1.325341 1.724718 2.085963 2.196658 2.527977 2.845340
> qt(c(0.90, 0.95, 0.975, 0.98, 0.99, 0.995), df = 21)
[1] 1.323188 1.720743 2.079614 2.189427 2.517648 2.831360
> qt(c(0.90, 0.95, 0.975, 0.98, 0.99, 0.995), df = 22)
[1] 1.321237 1.717144 2.073873 2.182893 2.508325 2.818756
```

A 95% confidence interval for μ is given by $14.8 \pm c\,\widehat{\text{SE}}$ where $(c, \widehat{\text{SE}})$ are given by

→ (A) $(2.09, 0.67)$

(B) $(1.73, 0.67)$

(C) $(2.09, 2.98)$

(D) $(1.33, 0.67)$

Spacer (Short Answers Section)

Extended Answer Questions

Some marks will be awarded for working. Show the working in the space provided below each question.

Question 21 (Extended Question 1) Points: 10

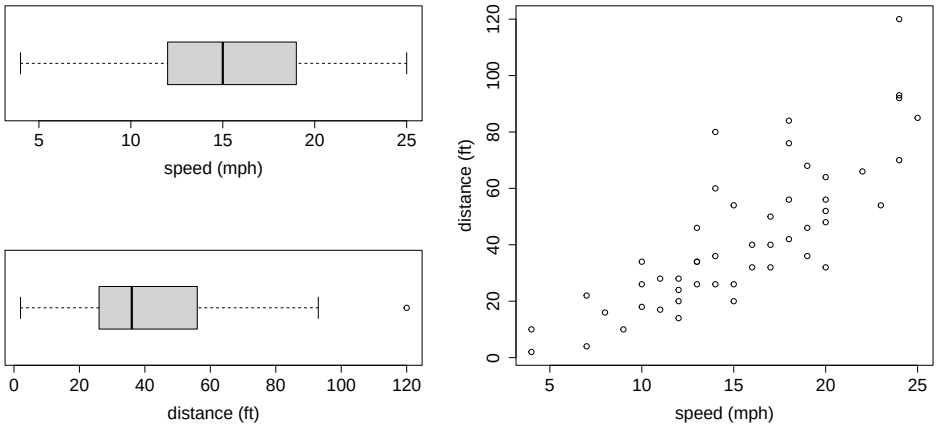
The dataset `cars` contains observations of 50 cars, with measurements of their speed in miles per hour (mph) and their corresponding stopping distance in feet (ft). Look over the R output and then answer the following questions.

SOLUTIONS

turn to page 9

Question 21 (continued)

```
> head(cars)
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
> summary(cars)
     speed           dist       
 Min.   : 4.0   Min.   :  2.00  
 1st Qu.:12.0   1st Qu.: 26.00  
 Median :15.0   Median : 36.00  
 Mean   :15.4   Mean   : 42.98  
 3rd Qu.:19.0   3rd Qu.: 56.00  
 Max.   :25.0   Max.   :120.00  
```



(a) What is the IQR of speed?

(b) Assume that speed follows a normal distribution. Which expression below, a, b or c, calculates an estimated SD of the normal curve? Explain briefly.

```
a=(19-15.4)/qnorm(0.75)
b=qnorm(0.75)/(19-15.4)
c=pnorm(19-15.4)/0.75
```

(c) Comment on the shape of the box plot for distance. Explain briefly why this shape is reasonable given the context of the study.

(d) Using the following output, write down an expression predicting the stopping distance of a car travelling at 21 mph. You do not need to evaluate the expression.

```
> dist=cars$dist
> speed=cars$speed
> L=lm(dist~speed)
> round(L$coeff, 3)
(Intercept)       speed 
    -17.579       3.932 
```

SOLUTIONS

turn to page 10

Question 21 (continued)

(e) Would we be able to predict the stopping distance of a car travelling at 35 mph using this regression? Explain.

(f) A previous study plotted the average stopping distance against the average speed of cars at each of 25 testing sites. These data had a clear positive correlation, with $r = 0.95$. The study concluded that driving at a faster speed leads to a longer stopping distance. Identify at least two problems with this argument.

Solution:

(a) 19-12=7

(b) a. Standardising, $(19 - 15.4)/\text{SD}$ equals qnorm(0.75). Rearranging gives the expression.

(c) Right skewed with one outlier. Some cars may have had difficulty stopping; distance is nonnegative; any other reasonable comment about the context.

(d) $3.932 \times 21 - 17.579$

(e) No. 35 is very far away from the bulk of the data, we cannot extrapolate.

(f) First, correlation does not imply causation. Possible confounders include tyres, weather, etc. Second, the correlation between averaged variables at a testing site level overstates the dependence at an individual sample level.

Solution: Solution

SOLUTIONS

turn to page 11

Question 22 (Extended Question 2) Points: 10

An investigator is interested in how many people make more than \$95,000 a year in a company, and he draws a random sample of 100 employees without replacement from a total of 10,000 employees. He counts 20 people earning more than \$95,000 in his sample. The investigator read on a website that 15% of people in the company make more than \$95,000.

(a) The R code below defines a box model, followed by simulating 10000 replicates of the experiment done by the investigator. Due to the large number of employees, the investigator decides to ignore the difference between sampling with and without replacement. Fill in the three blanks as indicated by "??". What does this code calculate in the end?

```
draw=function(){
    return(sum(sample(c(rep(1,??), rep(0,??)), size=??, replace=T))>=20)
}
mean(replicate(10000, draw()))
```

Solution: Since sampling is done with replacement, the first two ?? may be replaced by any pair of numbers whose ratio is $\frac{15}{85}$, e.g. `(15, 85)` or `(150, 850)` or even `(3, 17)`.

The third ?? should be the sample size, $n = 100$.

In the end, the code calculates the fraction of times out of 10000 replicates that we observe 20 or more 1s, i.e. 20 or more people earning more than \$95,000 in the sample.

(b) Under the box model, write down the expected value of the count in the investigators experiment.

Solution: In general, the sum of the draws has expected value $n\mu$ where n is the sample size and μ is the mean of the box. The proportion of 1s in the box is $\mu = p = \frac{15}{100} = \frac{3}{20} = 0.15$. Thus the expected value of the sum is $n\mu = 100 \times 0.15 = 15$.

(c) Explain why, under the box model, the standard error of the count in the investigator's experiment is 3.57 (to 2 decimal places).

Solution: The SD of the box is $\sigma = \sqrt{p(1-p)}$ and so the standard error of the count (sum) is thus $\sigma\sqrt{n} = \sqrt{np(1-p)} = \sqrt{100 * 0.15 * 0.85} \approx 3.57$. In general, the SE of the sum of the draws is $\sqrt{n}\sigma$ $n = 100$ is the sample size and σ is the SD of the box. There is a convenient computing formula for the SD:

$$\text{SD} = \sqrt{\text{MSQ} - (\text{AVG})^2}$$

where MSQ is the "mean square" of the box and AVG is the average. When the box only contains 0s and 1s, the mean and mean square are both equal to p, the proportion of 1s in the box, so that

$$\sigma = \sqrt{p - p^2} = \sqrt{p(1-p)}\,.$$

Here $p = 0.15$ so we get a standard error of $\sqrt{100 \times 0.15 \times 0.85} \approx 3.57$.

(d) Provide a normal approximation to the chance of getting a count of 20 or more under this box model. The R output below may be useful for this.

SOLUTIONS

turn to page 12

Question 22 (continued)

```
> z = (60:75)/50
> cbind(z, pnorm(z, lower.tail=FALSE))
         z
 [1,] 1.20 0.11506967
 [2,] 1.22 0.11123244
 [3,] 1.24 0.10748770
 [4,] 1.26 0.10383468
 [5,] 1.28 0.10027257
 [6,] 1.30 0.09680048
 [7,] 1.32 0.09341751
 [8,] 1.34 0.09012267
 [9,] 1.36 0.08691496
[10,] 1.38 0.08379332
[11,] 1.40 0.08075666
[12,] 1.42 0.07780384
[13,] 1.44 0.07493370
[14,] 1.46 0.07214504
[15,] 1.48 0.06943662
[16,] 1.50 0.06680720
```

Solution: As a z-score, the count 20 corresponds to $(20-15)/3.57 \approx 1.4$. Using the R output the area under the (standard) normal curve to the right of 1.4 is (approx.) 8.1%.

(e) The exact chance of 20 or more is given by the R command below:

```
> 1-pbinom(19, 100, 0.15)
[1] 0.1065443
```

Does this cast any doubt on the claims of the website? Explain.

Solution: We can use the probability above to perform an hypothesis test of the null hypothesis $H_0: p = 0.15$. The precise alternative hypothesis that should be used depends on what "deviations" from H_0 the investigator was interested in *before they saw the data*. It is tempting to consider a one-sided alternative $H_1: p > 0.15$ although this would only be valid if the investigator was interested in such an alternative *before they saw the data*. If there is any doubt, the more "conservative" two-sided alternative should be used. This would result in a P-value equal to $2 \times 0.1065 \approx 0.213$. In either case, this is a large p-value and thus does not provide any evidence against the null hypothesis. One would then conclude that the data are consistent with the company's claim that 15% of people make more than \$95,000.

Solution: Solution

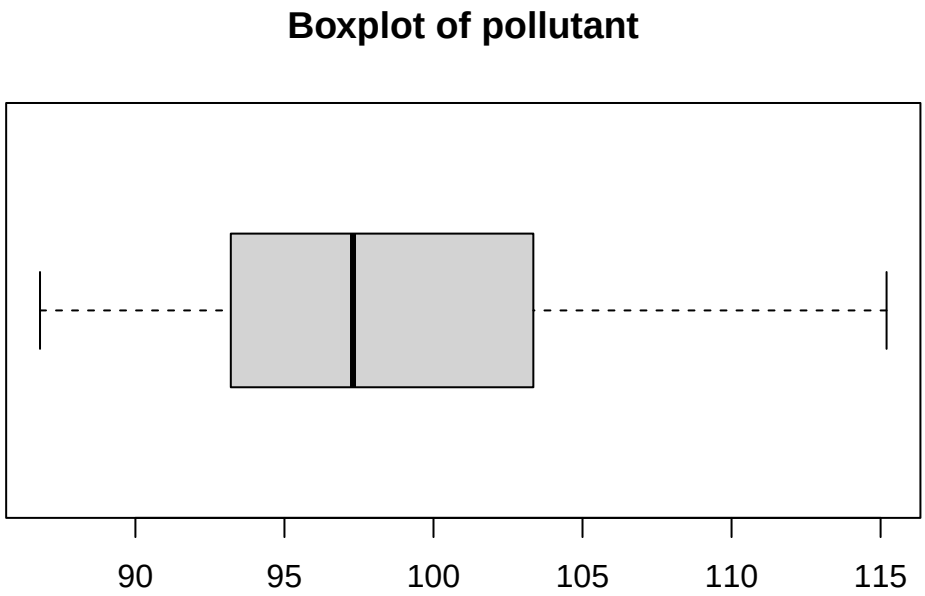
SOLUTIONS

turn to page 13

Question 23 (Extended Question 3) Points: 10

A manufacturing plant which emits a certain waste product into a nearby waterway has claimed to have implemented measures to reduce the average daily amount of pollution being created, which was previously measured to be 100 ppm (parts per million) of a certain pollutant. To back up their claim, measurements of pollutant levels on 36 consecutive days were taken, with a view to performing a t-test. Consider the R output below and the questions following it:

```
> pollutant
 [1] 115.2 109.7  90.6 104.2 106.5  92.9  93.3 107.2  97.7  94.1
[11]  93.1  97.6  86.8 101.2 102.9 100.1  98.8  90.3 103.8  98.0
[21]  95.5 108.0  93.5  86.8  92.4 102.6 114.1  97.0  90.8 110.9
[31] 102.1  87.3  96.8  93.7  94.6  95.7
> boxplot(pollutant, horizontal=TRUE, main="Boxplot of pollutant")
```

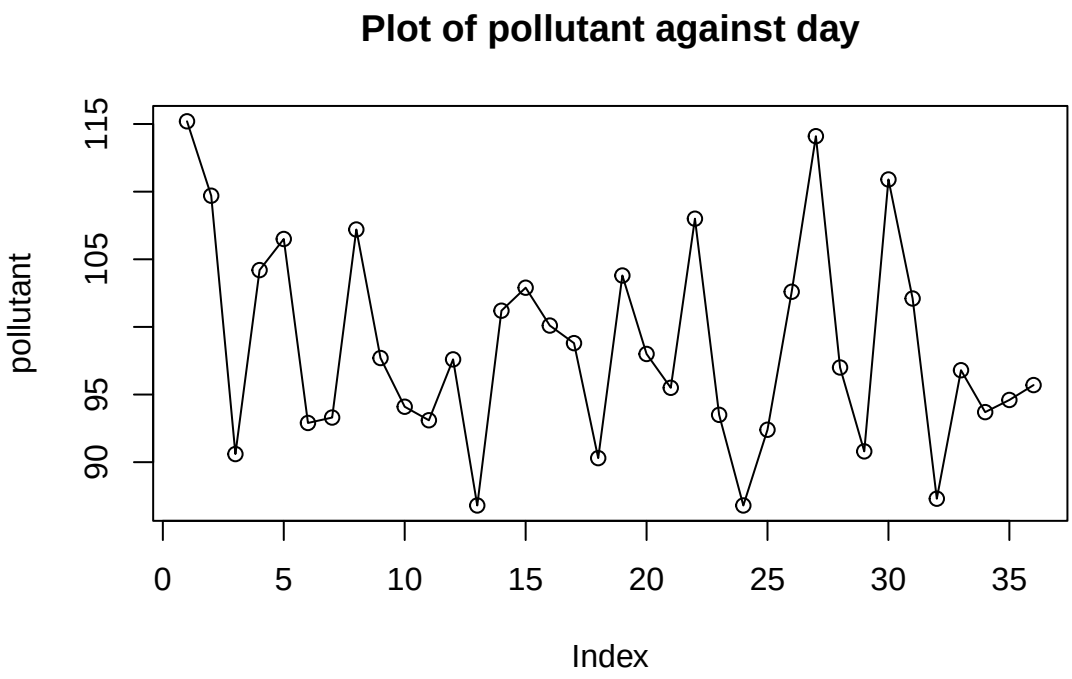



```
> plot(pollutant, main="Plot of pollutant against day")
> lines(pollutant)
```

SOLUTIONS

turn to page 14

Question 23 (continued)




```
> mean(pollutant)
[1] 98.49444
> sd(pollutant)
[1] 7.502874
> qt(c(0.8, 0.9, 0.95, 0.975,  0.99, 0.995), df = 34)
[1] 0.8523212 1.3069516 1.6909243 2.0322445 2.4411496 2.7283944
> qt(c(0.8, 0.9, 0.95, 0.975,  0.99, 0.995), df = 35)
[1] 0.8520119 1.3062118 1.6895725 2.0301079 2.4377225 2.7238056
> qt(c(0.8, 0.9, 0.95, 0.975,  0.99, 0.995), df = 36)
[1] 0.851720 1.305514 1.688298 2.028094 2.434494 2.719485
> qt(c(0.8, 0.9, 0.95, 0.975,  0.99, 0.995), df = 37)
[1] 0.851444 1.304854 1.687094 2.026192 2.431447 2.715409
> qt(c(0.8, 0.9, 0.95, 0.975,  0.99, 0.995), df = 38)
[1] 0.8511828 1.3042302 1.6859545 2.0243942 2.4285676 2.7115576
```

(a) What assumption underlying the t-test is the boxplot designed to assess? What do you conclude?

Solution: The t-test assumes the data is like a random sample from a normal population. The boxplot is designed to assess the normality assumption. Since it is reasonably symmetric with no outliers, the normality assumption seems reasonable.

(b) What assumption underlying the t-test is the plot against day designed to assess? What do you conclude?

Solution: The t-test assumes the data is like a random sample from a normal population. The plot against day is designed to see if there might be dependence across days (which is not desirable and would contradict the independence implicit in a random sample). There is no strong indication of serial dependence so there is nothing here to contradict the assumption of a random sample.

(c) Write down appropriate null and alternative hypotheses in terms of the parameter μ, representing the daily average amount of pollutant created *after* the measures have been implemented.

Question 23 (continued)

Solution: $H_0 : \mu = 100$ against $H_1 : \mu < 100$.

(d) Calculate the estimated standard error of the sample mean.

Solution:

```
> est.se = sd(pollutant)/sqrt(36)
> est.se
[1] 1.250479
```

(e) Use the R output to explain why the p-value is between 10% and 20%.

Solution: The observed value of the t-statistic is $(98.49 - 100)/1.25 \approx -1.2$. The correct number of degrees of freedom is 35 ($n - 1$ where $n = 36$ is the sample size). The 80th percentile of t_{35} is 0.85, the 90th percentile is 1.3. Since 1.2 is between these, by symmetry the observed value -1.2 is between the 10th and 20th percentiles (in the lower tail). Thus since the test is one-sided the p-value is between 10% and 20%.

(f) Write a sentence giving your overall conclusion.

Solution: **Short version:** The p-value is large, thus the data is consistent with the null hypothesis of no reduction.

In more detail: While the average pollutant output over the 36 days is slightly below 100, the apparent reduction is not *significantly* less than 100. So there has *either* been no "true" reduction, *or* there has been a small "true" reduction but more data is needed to detect it.

SOLUTIONS

This is the end of the examination paper