1. (a) Cape Town, South Africa, had a serious water shortage this year. One possible reason for this is below-average recent rainfall. Here is the annual rainfall (in mm) at Cape Town Airport since 2010:

Year	2010	2011	2012	2013	2014	2015	2016	2017	$\bar{x}$	s
Rainfall	368	375	408	679	513	327	221	153	380.5	163.7

You can use the following output from Matlab to answer the questions below.

$\texttt{tinv}(0.9, 6) = 1.440,$	$\texttt{tinv}(0.9, 7) = 1.415,$	$\texttt{tinv}(0.9, 8) = 1.397,$
$\texttt{tinv}(0.95, 6) = 1.943,$	$\texttt{tinv}(0.95, 7) = 1.895,$	$\texttt{tinv}(0.95, 8) = 1.860,$
$\texttt{tinv}(0.975, 6) = 2.447,$	$\texttt{tinv}(0.975, 7) = 2.365,$	$\texttt{tinv}(0.975, 8) = 2.306,$
$\texttt{tinv}(0.98, 6) = 2.612,$	$\texttt{tinv}(0.98, 7) = 2.517,$	$\texttt{tinv}(0.98, 8) = 2.449,$
$\texttt{tinv}(0.99, 6) = 3.143,$	$\texttt{tinv}(0.99, 7) = 2.998,$	$\texttt{tinv}(0.99, 8) = 2.897.$

i. [**3 marks**] Use these data to find a 95% confidence interval for the average recent Cape Town rainfall.

ii. [**3 marks**] What assumptions did you need to make to answer part i)? If is possible to check these assumptions, explain how.

iii. [**4 marks**] Prior to 2010, the average annual rainfall at Cape Town Airport was 515 mm per year.
Use a hypothesis test to determine if there is evidence that total rainfall in Cape Town has reduced in recent years.

iv. [**1 mark**] Let's say we are interested in which city gets more rain – Cape Town or Durban (a city located near Cape Town). We obtain rainfall data for 2010-17 from Durban.
What procedure would you use to test for a difference in average rainfall between Cape Town and Durban?

(b) Let X be a random variable whose probability density function is:

$$f(x) = \begin{cases} \frac{1}{\mu} e^{-x/\mu} & \text{if } x \geq 0 \\ 0 & x < 0 \end{cases}$$

where $\mu > 0$ is a parameter. It can be shown that $E(X) = \mu$.

i. [**2 marks**] Use integration to show that $\mathbb{E}(X^2) = 2\mu^2$.

ii. [**1 mark**] Hence show that it is not generally true that

$$\mathbb{E}\left[g(X)\right] = g\left[\mathbb{E}(X)\right]$$

iii. Consider a random sample of size 2 from X, which we will write as (X_1, X_2).

A. [**1 mark**] Write down the joint density function of X_1 and X_2.

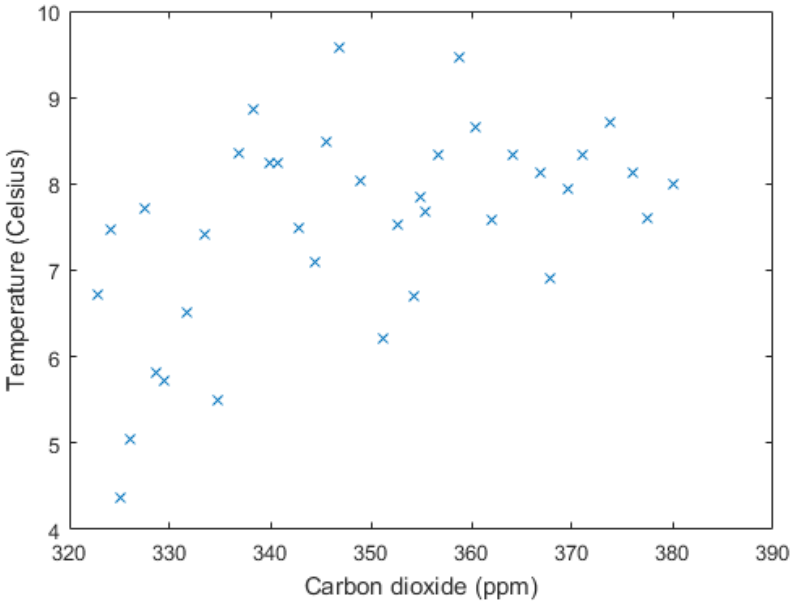
B. [**3 marks**] Hence find the probability density function of the sum $Y = X_1 + X_2$.

C. [**2 marks**] Assume we take a large random sample of size n from X, which we will write as $(X_1, X_2, \ldots, X_n)$. Let $\bar{X}$ be its sample mean,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

What is the approximate distribution of $\bar{X}$?

2. Mauna Loa Observatory is a remote high altitude atmospheric research facility that has been taking measurements for over 50 years, well-known for its documentation of recent climate trends. Below we analyse average carbon dioxide concentration (`x`, parts per million, ppm) and average temperature (`y`, in degrees Celsius) in December each year from 1970 to 2006.




```
Linear regression model:
    y ~ 1 + x

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   --------    -------    -------    --------

    (Intercept)      -4.581     3.5384    -1.2947     0.20392
    x              0.034691    0.01012     3.4278    0.001573


Number of observations: 37, Error degrees of freedom: 35
Root Mean Squared Error: 1.03
R-squared: 0.251,  Adjusted R-Squared 0.23
F-statistic vs. constant model: 11.7, p-value = 0.00157
```

The following summary statistics were also obtained for `x`:

$$\sum_{i=1}^{n} x_i = 12,921 \qquad \sum_{i=1}^{n} (x_i - \bar{x})^2 = 10,397$$

You can use the following output from Matlab to answer the questions below.

$\texttt{tinv}(0.9, 35) = 1.306,$	$\texttt{tinv}(0.9, 36) = 1.306,$	$\texttt{tinv}(0.9, 37) = 1.305,$
$\texttt{tinv}(0.95, 35) = 1.690,$	$\texttt{tinv}(0.95, 36) = 1.688,$	$\texttt{tinv}(0.95, 37) = 1.687,$
$\texttt{tinv}(0.975, 35) = 2.030,$	$\texttt{tinv}(0.975, 36) = 2.028,$	$\texttt{tinv}(0.975, 37) = 2.026,$
$\texttt{tinv}(0.98, 35) = 2.133,$	$\texttt{tinv}(0.98, 36) = 2.131,$	$\texttt{tinv}(0.98, 37) = 2.129,$
$\texttt{tinv}(0.99, 35) = 2.438,$	$\texttt{tinv}(0.99, 36) = 2.435,$	$\texttt{tinv}(0.99, 37) = 2.431.$

(a) [**1 mark**] How strong is the relationship between carbon dioxide and temperature? Use an appropriate numerical measure in your answer.

(b) i. [**2 marks**] Find a 95% confidence interval for the expected increase in December temperature if carbon dioxide concentration increases by one part per million.

ii. [**1 mark**] Hence find a 95% confidence interval for the expected increase in December temperature if carbon dioxide concentration increases by 20ppm (as is expected to happen over the next decade).

(c) Many climate scientists advocate trying to stabilise atmospheric carbon dioxide concentration at 450ppm.

i. [**2 marks**] What is the predicted average December temperature at Mauna Loa Observatory when carbon dioxide concentration is 450ppm?

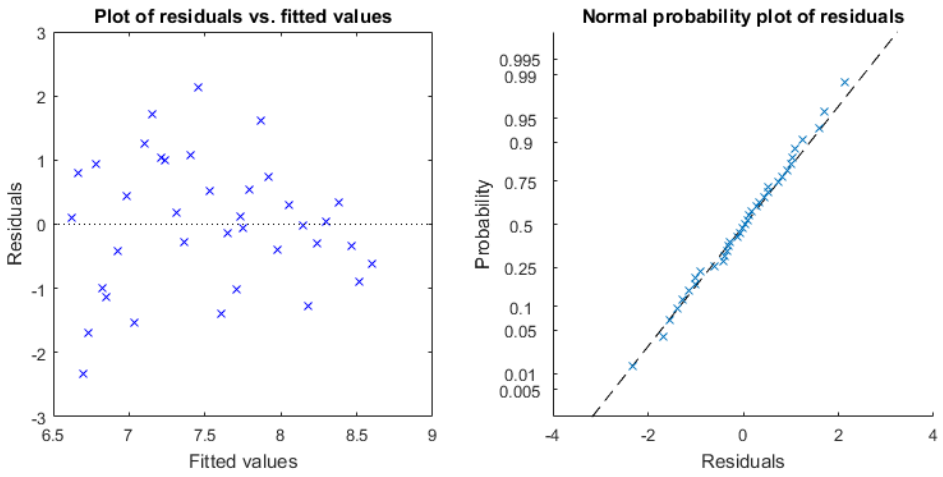
ii. [**2 marks**] Construct a 98% confidence interval for an observed December value for temperature, when carbon dioxide concentration is 450ppm.

iii. [**1 mark**] Why should we be cautious about this prediction?

(d) To answer part b), several assumptions were made about the data.

i. [**2 marks**] What are these assumptions?

ii. [**2 marks**] Explain what each of the below plots tells us about whether or not these assumptions are satisfied.



iii. [**1 mark**] Consider now the assumptions for part c). While the same assumptions were made as in part b), some of them take on greater importance in part c). Briefly explain.

(e) [**3 marks**] If two random variables X and Y are independent, then

$$\mathbb{E}\left[g(X)h(Y)\right] = \mathbb{E}\left[g(X)\right]\,\mathbb{E}\left[h(Y)\right]$$

Use this result to show that a non-zero correlation between carbon dioxide and temperature implies that the two variables are dependent.

(f) [**1 mark**] Does the regression output suggest that there is evidence that carbon dioxide and temperature are correlated? In your answer, refer to the relevant parts of the output.

(g) [**1 mark**] Is this an observational study or an experiment?

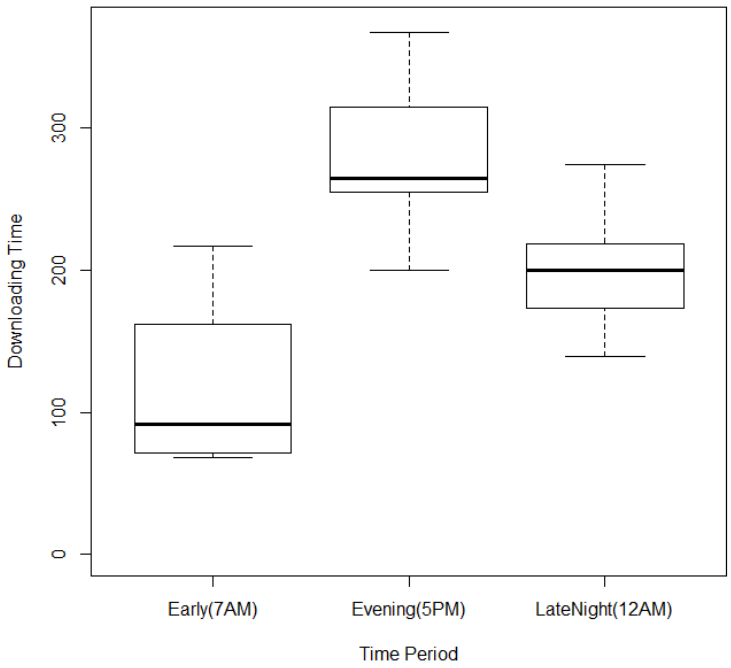
(h) [**1 mark**] Sceptics argue that increasing global temperatures may not be due to increasing atmospheric carbon dioxide concentration, but may instead be due to sunspot activity or other reasons.

Use ideas from the course to briefly explain whether or not the Mauna Loa data disproves sceptics' claims.

3. To see how much of a difference time of day makes on the speed at which he could download files, a college sophomore performed an experiment. He placed a file on a remote server and then proceeded to download it at three different time periods of the day. He downloaded the file 24 times in all, 8 times in each time period. The downloading times (in seconds) are summarised in the table below:

Early (7a.m.)	Evening (5p.m.)	Late night (12a.m.)
68	299	216
138	367	175
75	331	274
186	257	171
68	260	187
217	269	213
93	252	221
90	200	139
$\bar{x}_1 = 116.88$	$\bar{x}_2 = 279.38$	$\bar{x}_3 = 199.50$
$s_1 = 57.41$	$s_2 = 51.77$	$s_3 = 40.88$

Comparative boxplots are given in the figure below.



You can use the following output from Matlab to answer the questions below.

$\texttt{tinv}(0.9, 2) = 1.886$,	$\texttt{tinv}(0.9, 21) = 1.323$,	$\texttt{tinv}(0.9, 23) = 1.320$,
$\texttt{tinv}(0.95, 2) = 2.920$,	$\texttt{tinv}(0.95, 21) = 1.721$,	$\texttt{tinv}(0.95, 23) = 1.714$,
$\texttt{tinv}(0.975, 2) = 4.303$,	$\texttt{tinv}(0.975, 21) = 2.080$,	$\texttt{tinv}(0.975, 23) = 2.069$,
$\texttt{tcdf}(3.2729, 2) = 0.959$,	$\texttt{tcdf}(3.2729, 21) = 0.9982$,	$\texttt{tcdf}(3.2729, 23) = 0.9983$,
$\texttt{finv}(0.95, 2, 21) = 3.467$,	$\texttt{finv}(0.95, 2, 23) = 3.422$,	$\texttt{finv}(0.95, 3, 24) = 3.009$,
$\texttt{finv}(0.975, 2, 21) = 4.420$,	$\texttt{finv}(0.975, 2, 23) = 4.349$,	$\texttt{finv}(0.975, 3, 24) = 3.721$,
$\texttt{fcdf}(20.72, 2, 21) = 1$,	$\texttt{fcdf}(20.72, 2, 23) = 1$,	$\texttt{fcdf}(20.72, 3, 24) = 1$.

(a) [**2 marks**] What do the boxplots tell you about the downloading speed at different time periods of day? Comment on the shape, range and location.

(b) [**3 marks**] List three assumptions that need to be valid for an Analysis of Variance (ANOVA) to test whether there is a difference in average downloading time among the three time periods. Which of these three can be checked by considering the accompanying summary statistics? Explain whether these verifiable assumption(s) are supported.

Assume from now on that these assumptions are valid.

(c) [**3 marks**] An ANOVA table was partially constructed to summarise the data:

Source	df	SS	MS	F
Treatment	**(1)**	105635	**(4)**	**(5)**
Error	21	**(3)**	2549	
Total	**(2)**	159166		

Copy the ANOVA table in your answer booklet. Complete the table by determining the missing values (1)–(5), and stating how you computed the missing entries.

(d) [**4 marks**] Using a significance level of $\alpha = 0.05$, carry out the ANOVA F-test to determine whether there is a difference in average downloading time among the three time periods.

(*You can use the numerical values found in the above ANOVA table; however, you are required to write the detail of the test: null and alternative hypotheses; observed value of the test statistic; rejection criterion or the p-value (you may use bounds for the p-value, specify the degrees of freedom if applicable); conclusion in plain language.*)

(e) [**4 marks**] From the previous results, construct a 90% two-sided confidence interval on the difference between the "true" downloading times in the evening and late at night, that is, $\mu_2 - \mu_3$.

(f) [**4 marks**] Using the Bonferroni adjustment, carry out a t-test comparing the "true" average downloading time in the early morning and late at night. Does this allow you to come to the same conclusion as the ANOVA F-test in d), at overall level $\alpha = 0.05$? Explain.

(*You can use the numerical values found in the above ANOVA table; however, you are required to write the detail of the test: null and alternative hypotheses; observed value of the test statistic and p-value (specify the degrees of freedom if applicable).*)

4. You can use the following output from Matlab to answer the questions below.

```
>> norminv([0.9 0.925 0.95 0.975 0.98 0.99])

ans =

    1.2816    1.4395    1.6449    1.9600    2.0537    2.3263

>> normcdf([0.2 1/sqrt(5) 1 2 2/sqrt(5) 3])

ans =

    0.5793    0.6726    0.8413    0.9772    0.8145    0.9987
```

(a) The lengths of similar components produced by a company are approximated by a normal distribution model with a mean of 5 cm and a standard deviation of 0.02 cm.

i. [**2 marks**] If a component is chosen at random, what is the probability that the length of this component is between 4.98 and 5.02 cm?

ii. [**2 marks**] What is the probability that among 5 randomly selected components, 3 will be between 4.98 and 5.02 cm?

(b) i. [**3 marks**] At an electronics plant, in a sample of 100 new workers, 49% met the production quota.
Construct a 95% confidence interval for the true proportion π of new workers who will meet the production quota.

ii. [**3 marks**] What assumptions did you make when constructing this confidence interval? Where possible, comment on whether these are reasonable.

iii. [**3 marks**] Another sample of new workers is to be collected, to improve the estimate of the proportion of new workers who meet the production quota.
In order for the estimate to have an error less than 0.02 with confidence level 0.95, how many new workers must be included in the new sample, regardless of the true value of π?

iv. [**5 marks**] A new training program is introduced, and it is claimed that more than 90% of new workers who attend the training program will meet the production quota. But in a sample of 100 new workers who complete the training program, only 84% meet the production quota.

Does the sample provide enough evidence to indicate invalidate the claim that at least 90% of new workers will meet the production quota? Carry out a suitable hypothesis test at 5% level of significance.

(Write the detail of the test: null and alternative hypotheses, rejection criterion, observed value of the test statistic, p-value, conclusion in plain language.)

v. [**2 marks**] Assume that 70% of new employees do the training program. Given the above information, what is your best estimate of the overall proportion of new employees meeting the production quota?