In general, you don't need to memorise the R code, as we don't have any questions in the final exam asking you to fill in the code. However, you need to understand R outputs to answer many questions. For example, means, standard deviations, and quantiles used in hypothesis tests. Note that this outline of contents may not cover every aspect of the final exam, but it should provide you with a skeleton of key topics and concepts that you must review.

==========

Weeks 1 and 2

- Frequency table, barplot
- Histogram -- it's important to know the difference between density scale histograms and frequency scale histograms, and be able to calculate the area for each class interval.
- Boxplot -- it's important to understand how and what numerical summaries are used to build the boxplot. Being able to determine the skewness of a data set from the boxplot (and histogram as well) is very useful.
- Numerical summaries -- order-based (median, quantiles, and interquartile range) and average-based (mean and SD). There are two versions of SDs. The definition of these is assumed knowledge and their formula will not be provided in the final.
- Be familiar with the concept of outliers and know how to detect them using graphical summaries.

Week 3

- Normal distribution
    - Properties of normal distribution -- quantiles, percentiles, equivalence between two normals under linear transform, standardisation and z-score, 68%-95%-99.7% Rule ...
    - A very important application is to use the normal distribution to approximate a data set, and then work out the proportion of certain events. For example, the proportion of heights in a population falls into a certain interval.

  You need to know everything regarding the normal distributions
- Correlation coefficient
    - Understand the definition (average of the products of Z scores)
    - Its value is always between -1 and 1
    - Being able to interpret linear association using the values of correlation coefficient (and vice versa)
    - Properties: it is symmetric and invariant to scaling and shift
    - Can be sensitive to outliers
    - Nonlinear association cannot be determined (thinking about the Anscombes Quartet)

Non-examinable topic(s):

- You don't need to memorise the density function of the normal distribution.

Week 4

- Regression Line
    - Can work out the regression using five numerical summaries (means, sample SDs, and the correlation coefficient)
    - Coefficient of determination, which tells the proportion of the variation (compared to the baseline prediction) in the dependent variable that can be explained by the linear model
- Residual plot
    - If the linear fit is appropriate for the data, it should show no pattern (random points around 0)
    - By checking the patterns of the residuals, the residual plot is a diagnostic plot to check the appropriateness of a linear model
    - Homogeneity of the spread (homoscedasticity or heteroscedasticity)
- Probability
    - Basic definitions and conditional probability
    - Addition and multiplication rule
    - Chance simulation (sample with/without replacement)

Non-examinable topic(s):

- Why the Coefficient of determination is the same as r^2.
- The derivation of the optimality of linear regression model

Week 5

- Box models
    - The expected value and the standard error of random draws
    - The expected value and the standard error of summation of random draws
    - The expected value and the standard error of summation of sample sums
        - E(S) = n x mean; and SE(S) – SD x sqrt(n)
    - The expected value and the standard error of summation of sample means
        - E(\bar X) = mean; and SE(S) – SD / sqrt(n)

Non-examinable topic(s):

- We don’t expect you to derive the expected value and the standard error of summation of random draws, but other parts are expected.

Week 6

- Central limit theorem
    - Know how the sample sum and sample average will follow the normal curve in the limit
    - Know how to apply the standard units of the sample sum and sample average (using the expected values and the SE), so their standard unit (z-score) follows the standard normal in the limit.
- Unknown proportion
    - Prediction interval – an interval for the sample average defined by known population parameters (mean and SD of the box)
    - Consistency between data and population proportion defined by the prediction interval
    - Confidence interval – an interval for estimating unknown population proportion using a given sample
    - Confidence intervals change from sample to sample, only a certain percentage of them (e.g., 95%) specified by the confidence level covers the fixed population proportion
    - We use R to work out Wilson's confidence level

Non-examinable topic(s):

- We don't expect you to derive Wilson's confidence level by hand (reading R output is sufficient).

Week 7

- Confidence interval for unknown mean
    - Know how to work out the confidence interval for unknown mean, with known population SD
- Z test for proportion
    - Z-statistic, under the null hypothesis, we can set up the expected value and SE of the sample proportion, and hence the Z-statistic.
    - The distribution of the test statistic is defined by the null hypothesis.
    - Level of significance / false alarm rate – even the null hypothesis is true, we can still falsely reject at this rate. This also define how we can interpret the evidence in the data.
    - Know how to set up the alternative hypothesis based on the context and how to get the P-value depending on the alternative.
    - We can make equivalent decisions based on the confidence interval.

Week 8

- Z test for mean
    - The five steps of hypothesis test.
    - Applying confidence intervals to make equivalent decisions (only covered for the two-sided case).

Week 9

- One-sample T-test (for unknown population SD)
    - Know when to apply Student's t-distribution and its degrees of freedom
    - Assumptions (near normality in particular) involved and how to check them
    - P-value for the T-test
    - Critical region of rejection
    - Confidence intervals for estimating the population mean with unknown population SD
    - Equivalence between the decisions using P-value and those based on confidence intervals
- One-sample T-test using bootstrap simulation
    - Using for loop
    - Bootstrap simulation
    - Simulation-based P-value
    - Simulation-based (equal tail) confidence intervals – no longer have symmetry

Non-examinable topic(s):

- The derivation of Sample SD using the box model

Week 10

- Two-sample T-test
    - The classical two-sample T-test
        - Assumptions, pooled estimated of the common SD, estimated SE for the sample mean difference, degrees of freedoms
    - The Welch test
        - Assumptions, estimated SE for the sample mean difference
    - Two-sample Z test – a much simpler case than T test
    - Bootstrap simulation
    - Confidence intervals for the population mean difference

Non-examinable topic(s):

- We didn't discuss how to work out the degrees of freedom for the Welch test (we use R for this)
- We expect you to know how to apply the pooled estimate formula for the common SD, but not its full derivation.

Week 11

- Chi-squared test for goodness of fit
    - Test statistic, degrees of freedom of the Chi-squared distribution (k-1), P-value (one-sided test)
    - Checking assumptions – all expected frequencies >= 5
    - Using simulation – we know the exact box model under the null hypothesis

- Chi-squared test for independence
    - Estimate the expected frequencies using the independence assumption (under H0)
    - The degrees of freedom for Chi-squared distribution

Non-examinable topic(s):

- We didn't discuss how to carry out bootstrap simulation for the test for independence (using R for this).

Week 12

- Inference for simple linear regression
    - Know the degrees of freedom in estimating the SD of the regression error
        - (n - (p+1))
    - Know how to get the confidence interval of the regression coefficient for given standard errors
    - Know how to calculate the observed test statistic and P-value for given standard errors
    - Can interpret the null and alternative hypotheses
- Multiple linear regression
    - Can interpret the fitted model
    - Know the degrees of freedom in estimating the SD of the regression error
        - (n - (p+1))
    - Know how to get the confidence interval of the regression coefficient for given standard errors
    - Know how to calculate the observed test statistic and P-value for given standard errors
    - Can interpret the null and alternative hypotheses
- Can interpret the fitted models (with/without log-transform)

Non-examinable topic(s):

- We don't expect you to derive the SE for regression coefficients. Only knowing how to read summary(lm(...)) is sufficient

Week 13

- F-test – can interpret the hypotheses of F-test and make decisions based on the P-value
- Know the difference between R-squared (coefficient of determination) and Adjusted R-squared
- Model selection – can perform forward and backward model selection using F-test
- Logistic regression – can interpret the model in terms of odds

Non-examinable topic(s):

- Anything not mentioned above