Descriptive Statistics

1. Mean.
    1.a. Obtain the mean *weight* of all chicken from the *ChickWeight* data set. Then obtain the mean weight by chicken
    1.b. Obtain the mean of a 3x4 matrix
    1.c. Create a list out of the *co2* data set, where elements represent years. Obtain mean CO2 concentration.

2. Quantiles. For the *weight* data in the *ChickWeight* data set
    2.1. Generate 7 quantiles of equal probability
    2.2. Generate percentiles
    2.3. Generate 10 quantiles of unequal probability

3. Variation. From the *DNase* data set
    3.1. Obtain the MAD of density for different levels of concentration
    3.2. Obtain the variance and standard deviation of density
    3.3. Obtain the range of density by run

4. Co-variation: Using Edgar Anderson's *iris* data
    4.a) Determine the co-variance between sepal length and width
    4.b) Evaluate the correlation between sepal length and petal length

Probability Distributions

1. Generate sample data [1000], PDF, CDF and quantiles for the following
    - 1.a) Standard normal distribution
    - 1.b) Binomial distribution: Number of trials = 50, Prob of success = 0.3
    - 1.c) Chi-squared distribution: Degrees of freedom = 10
    - 1.d) Exponential distribution: Rate = 1.3
    - 1.e) Uniform distribution: Min = 10, Max = 15
    - 1.f) T-Distribution: Degrees of freedom = 20

Hypothesis Testing

1. One sample T-test. For the eruptions data in the Old Faithful geyser in Yellowstone [dataset *faithful*], choose a value (μ) for the population mean "close to" the sample mean. Null hypothesis - H_0: Population mean = μ
    1.a) Perform a two-sided T-test. Alternate hypothesis - H_a: Population mean $\neq \mu$
    1.b) Perform a one sided test. Alternate hypothesis - H_a: Population mean $> \mu$
    1.c) Perform a one sided test. Alternate hypothesis - H_a: Population mean $< \mu$

2. One sample T-test. Repeat 1) for a value of the population mean that is much lesser than the sample mean

3. One sample T-test. Repeat 1) for a value of the population mean that is much greater than the sample mean

4. Two sample T-test. Using the Michelson speed of light dataset *morley*, test that the average speed of light in either case is the same. Null hypothesis: H_0: Differences in Population mean = 0
    4.a) Perform a two-sided T-test. Alternate hypothesis - H_a: Pop mean delta $\neq 0$
    4.b) Perform a one sided test. Alternate hypothesis - H_a: Pop mean delta > 0
    4.c) Perform a one sided test. Alternate hypothesis - H_a: Pop mean delta < 0

5. Two sample T-test. Repeat 4) for a value of population mean delta that is greater than zero

6. Two sample T-test. Repeat 4) for a value of population mean delta that is lesser than zero

7. Perform a KS-Test on Edgar Anderson's *iris* data on the hypothesis H_0: data sets (by species) vary significantly

8. Perform an F test on the data Edgar Anderson's *iris* data on the hypothesis H_0: $\sigma 1/\sigma 2 = 1$

9. Perform an F test on the data Edgar Anderson's *iris* data on the hypothesis H_0: $\sigma 1/\sigma 2$ = {value greater than 1}

10. Perform an F test on the data Edgar Anderson's *iris* data on the hypothesis H_0: $\sigma 1/\sigma 2$ = {value lesser than 1}

Linear Models

1. Formula objects: Test symbol usage
    1.a) Symbols *,-,^,:,/ and function *I*()
    1.b) Create different formulas for linear models: one/two/three variables, transformation, Classification analysis, polynomial regression, and nested classification

2. Creating Models: Using the Motor Trend car data - *mtcars* - in package *datasets*, create a linear model of *hp* as a function of *disp*, *mpg* and *wt*.
    2.a) Obtain summary level info about the model
    2.b) Obtain model coefficients, residuals and fitted values
    2.c) Perform an ANOVA on the model
    2.d) Obtain the qr decomposition of the model

3. Creating Models: Using the Motor Trend car data, now tweak the model and repeat 2a) - d)
    3.a) Perform a transformation
    3.b) Modify the terms. Ex., use the (arithmetic) inverse of *mpg*.
    3.c) Include interactions between terms

4. Creating Models: Using Edgar Anderson's *iris* data, create of Classification model of Petal length as a function of Species, Sepal Length and Sepal Width
    4.a) Obtain summary level info about the model
    4.b) Obtain model coefficients, residuals and fitted values
    4.c) Perform an ANOVA on the model
    4.d) Obtain the qr decomposition of the model

5. Updating Models:
    5.a) Simulate the addition of a term to the model from 2)
    5.b) Simulate the drop of a term from the model from 2)
    5.c) Make a permanent change to the model from 2)

Generalized Linear Models

1. Use the Bayshore Medical data on Low Birth Weights and update the GLM from the prior video - add terms, take terms out, treat some terms as factors etc. For each iteration,
    1.a) Obtain summary level info about the model
    1.b) Obtain model coefficients, residuals and fitted values
    1.c) Perform an ANOVA on the model

Non linear regression

1. Using the *kirby2* data (file = kirby2.csv), perform a non-linear regression of y with respect to x
    - 1.a) Obtain summary level info about the model
    - 1.b) Obtain model coefficients, residuals and fitted values
    - 1.c) Obtain the variance-co-variance matrix and calculate model predictions and standard errors

    Model: $y = (b_1 + b_2x + b_3x^2)/(1 + b_4x + b_5x^2)$

    Starting values: $b_1 = 2$, $b_2 = -0.1$, $b_3 = 0.003$, $b_4 = -0.001$ and $b_5 = 0.00001$

Tree models

1. Use the *PlantGrowth* dataset and generate a tree model of weight by group
    1.a) Obtain summary level information about the tree model
    1.b) Plot the tree model along with text

2. Change the minimum number of observations at a node for a split and re-generate the tree model. Repeat steps 1a), 1b)

3. Change the minimum improvement to fit for a split to be considered and re-generate the tree model. Repeat steps 1a), 1b)

4. Change the minimum number of observations at a leaf and re-generate the tree model. Repeat steps 1a), 1b)