Some Statistics

Data Type
- Categorical
  - fixed set of possible value
  - not continuous
- Numerical
  - from measurement
  - different numerical values have different meanings
  - number is not necessarily numerical variable
  - number can be categorical

Tidy data:
- 1.Each variable must have its own column
- 2.Each variable must have its own row
- 3. Each value must have its own cell

Plot

1 Categorical Data
- x-axis: a category
- # level = # bar
- y-axis: count, frequency
- count larger -> bar higher
- library(ggplot2)
  ggplot(data = mtcars, aes(x = am)) + geom_bar()
  *x must be a catecorical variable

1 Numerical Data
- Histogram
  - Symmetric/Left-skewed/Right-skewed
  - Uniform/Unimodal/Bimodal/Multimodal
  - Abnormal points
  - ggplot(data = mtcars, aes(x = hp)) + geom_histogram()
- Boxplot
  - Median:
    50% value is higher
    50% value is lower
  - 1st Quartile:
    25% is lower
    median of lower half
  - 3rd Quartile:
    25% is higher
    median of higher half
  - IQR: 3rd Quartile - 1st Quartile
  - Whisker: no longer than 1.5IQR
  - Symmetric/Left-skewed/Right-skewed
  - Light tail/Heavy tail, Small variation/Large variation
  - library(ggplot2)
    ggplot(data = mtcars, aes(x ="hp", y = hp)) + geom_boxplot()

Data vs. Data
- 1. Number against Number
  (* y-axis against x-axis)
  - scatterplot
    - Correlated/Uncorrelated
      - Confounding effect:
        $A \sim C$
        $B \sim C$
        $A !\sim C$
    - Strong, Moderate, Weak
    - Positive, Negative
    - Linear relationship
      - A curve cannot represent linear relationship
      - Use Transformation(eg. log transformation)
        convert non-linear relationship to linear relationship
    - Outlier
    - Cluster
  - linear regression
- 2. Number against category
  - Boxplot
    - compare level ~ distribution of dependent variable
    - ggplot(data = mtcars, aes(x = factor(am), y = hp)) +
      geom_boxplot()
  - Linear regression with indicator variable

Theories of Histogram
- bin = $[X_0 + m*h, X_0 + (m+1)h]$
  (m is index, range from -1 to anywhere you need)
- h y = $1/n$ # {Xi in bin}
  (h y is area)
  - y = height of bin
  - n = sample size
  - h = width of bin
  - binwidth wider -> bin height lower
  - more #{Xi in bin} -> bin height higher
- $y = 1/hn$ #{Xi in bin}
  => $f(x\_bar) = 1/hn$#{Xi in bin}
  (histogram estimater)

Estimator of Density
- $hy = 1/n$ # {Xi in bin}
  => $2 hy = 1/n$ #{Xi in X-h, X+h}
  - $f(x\_bar) = 1/hn \ 1/2$ #{Xi in X-h, X+h}
  - weight functon
    (to count # {Xi in X-h, X+h})
    - w(x) = 1/2 if |x| < 1
    - w(x) = 0 if |x| >= 1
  - naive estimator of density:
    f(x_bar) = 1/hn * sum(w(Xi - X)/h))

Inference

Population
- Parameters
  - population mean: u
  - population proportion: p

Sample
- Statistics
  - sample mean: x_bar
  - sample proportion: p_bar

Core
- Use Sample Statistics to predict Population Parameter

Central limit theorem
- get sample > 1000 times, calculate every sample
  mean/proportion
- Show Empirical distribution features
- Sampling distribution:
  (histogram: distribution of sample mean/proportion)
  - unimodal symmetric
  - mid point is population parameters(mean/proportion)
  - sample size larger
    -> sampling distribution more centred around mode

Hypothesis testing
- 1. Hypothesis
  - Null Hypothesis: nothing/random:
    (=)
    - one proportion test: p_1 = 0.5
    - two proportion test: p_male = p_female
    - two mean test: u_male = u_female
  - Alternative Hypothesis: complementary set of Null hypothesis:
    (!= or > or <)
    - one proportion test: p_2 != 0.5
    - two proportion test: p_male != p_female
    - two mean test: u_male != u_female
- 2. Test statistics
  - sample proportion
- 3. Simulation
  - get Sampling distribution according to null hypothesis
- 4. Caluclate P value
  - Def: given null hypothesis, the probability of having our
    observation or more extreme observations
- 5. Conclusion
  - P value < 0.05, reject null hypothesis, accept alternative
    hypothesis
  - P value > 0.05, there is no evidence to reject null hypothesis, so
    we accept it

Confidence Interval
- In statistics, a confidence interval (CI) is a type of interval
  estimate(of a population parameter) that is computed from the
  observed data
  (common: 0.9, 0.95, 0.99)
- 1. R code:
  0.025 quartile
  0.975 quartile
  to get 95% CI
- 2. Count:
- Confidence level increase -> CI width increase
- We are 95% percent confident that the true population mean is
  within (10, 100).

Classification
- features
  => Analysis
  =>Classifications
  =>Validation
- Separation of training and testing
- Decision Tree