Stats in R

- **Check if data is normally distributed or not**
- **Based on that you can know whether to apply parametric methods or not**
- **Check the homoscedasticity of the data**

Step 1: Test normality of data using **Shapiro Wilk Test, Anderson Darling Test**

Step 2: If data is normally distributed, you can use the **T test**(if you do not know the stdev of the data and n > 30)**/Z test(**if we know the stddev of the data and n>30**)**, otherwise Mann Whitney U test.
The **Mann Whitney U test** can be used for examining the difference in median of two independent populations.
Z test is used as a proportions test for big parametric populations(n>30)
2 sample T test is called the Welch T test.(numerical outcome variable Y, categorical explanatory variable.
Foe example: in the cars dataset in R, am - transmission of car(0-manual, 1-automatic), wt-weight
# outcome variable = mtcars$wt, always on the left side
# explanatory varible = mtcars$am, on the right side

Step 3: Analysis of Variance(ANOVA) is used for normally distributed and independent data sets. We can use it to compare the means of different groups(**more than 2**).
If the data is not normally distributed the equivalent test(to ANOVA) would be the **Kruskal Wallis test** for non parametric data(non normally distributed data).

Step 4: **Levene Test** for **homoscedasticity**, used to compare the variances of different groups
Levene test needs the car package.

Step 5: **Chi squared test** of independence, parametric(works on normally distributed data)
It can be used to test the independence of two categorical variables.
For non parametric data(non normally distributed data) the equivalent to the Chi Squared test of independance is the **Fishers exact test**

Step 6: Checking **correlation**
Pearson correlation coefficient(default) -
2 numeric variables,
linear coefficient,
parametric data(normal distribution)

Spearman rank correlation coefficient -
2 numeric variables,

Monotonic coefficient,
non-parametric data(non normal distribution)

Kendall rank correlation coefficient -
x-y concordance of pairs,
non-parametric data(non normal distribution)

T-Test

A t-test is used for testing the mean of one population against a standard or comparing the means of two populations if you **do not know the populations' standard deviation** and when you have a **limited sample (n < 30)**. If you know the populations' standard deviation, you may use a z-test. **A z-score tells you how many standard deviations from the mean your result is.** You can use your knowledge of normal distributions (like the 68 95 and 99.7 rule) or the z-table to determine what percentage of the population will fall below or above your result.
Example:Measuring the average diameter of shafts from a certain machine when you have a small sample.

Z-Test

A z-test is used for testing the mean of a population versus a standard, or comparing the means of two populations, with **large (n ≥ 30) samples** whether you **know the population standard deviation** or not. It is also used for testing the proportion of some characteristic versus a standard proportion, or comparing the proportions of two populations.
Example:Comparing the average engineering salaries of men versus women.
Example: Comparing the fraction defectives from 2 production lines.

F-Test

An F-test is used to compare **2 populations' variances**. The samples can be any size. It is the basis of ANOVA.
Example: Comparing the variability of bolt diameters from two machines.

Difference between T test and Z test:

1. Z-test is a statistical hypothesis test that follows a normal distribution while T-test follows a Student's T-distribution.
2. A T-test is appropriate when you are handling small samples (n < 30) while a Z-test is appropriate when you are handling moderate to large samples (n > 30).
3. T-test is more adaptable than Z-test since Z-test will often require certain conditions to be reliable. Additionally, T-test has many methods that will suit any need.
4. T-tests are more commonly used than Z-tests.
5. Z-tests are preferred than T-tests when standard deviations are known.

Outlier Detection

- 3 sigma edit rule(t = 3)
    - [mean - t * SD, mean + t * SD] (3 standard deviations on either side)
    - We choose t = 3 as for normal distributions, the probability of observing data outside the 3 std dev's is 0.3%
    - **Drawback**: Mean and std dev are sensitive to outliers data
- Boxplot method:
    - [Q1 - C * IQD, Q3 - C * IQD]
    - IQD = Q3 - Q1
    - C = 1.5
- **Univariate** Outlier detection
    - Dixon Test
    - Grubbs Test
- **Multivariate** Outlier detection
    - Mvoutlier R package - sign1, sign2, pcout
    - Sign1 - based on the covariance metrics

Linear Modelling

- Linear model
    - Normally distributed error terms
    - Constant variance
    - Suitable for problems where the outcome variable is continuous
    - Y = a + bx + e
- Generalized linear model
    - Non normally distributed error terms
    - Non constant variance
    - Suitable for problems where the outcome variabel is a count(poisson distribution), proportion between 0,1(binomial distribution), survival analysis
    - Log y = a + bx + e
- Polynomial regression

    - Used when there is a non linear relationship between the outcome variable and the predictor
    - Very sensitive to outliers
    - It does not assume that the variables are monotonic and constant across the range of a variable(like in linear regression)
    - Smoothers:
        - Simple smoothers - SMA(Simple moving average), EMA(Exponential moving average), used in stock trading.
        - LOESS: locally weight the observations. It is a locally weighted least squared regression. The data is averaged out but more weight is put on the local data at any given point. More emphasis will fall on points in the middle and not at the start and end
        - Smothers vs roughness trade off
    - Splines
        - Fitting several different sub regression lines each of which will fit the local portion of the data. The slope changes based on the local set of data.
        - Penalized splines - advanced splines that allow for automatic data driven selection of smoothness. It tries to minimize the least squared error. It is called BROKEN STICK REGRESSION

Linear Regression

- Assumes features are independent of each other
- Requires homoscedasticity of variances between features(homogeneity of variance)