Non-parametric statistical inference

Non-parametric hypothesis testing

- There are situations for which is not possible to apply parametric statistics:
  - Data have a ranking but no clear numerical interpretation, such as user preferences;
  - You are interest on a parameter of the population for which the distribution is unknown (medians, variances, percentiles, etc.).
- Pos: Non-parametric methods make fewer assumptions than parametric, they are distribution free
- Cons: In cases where a parametric test would be appropriate, non-parametric tests have less power.

Test normality assumption

- Normality of the data is required for t-test. You may need to test whether your data follows a normal distribution.
- Graphical methods for testing normality include the histogram and qq-plot.
- Non-parametric tests such as Kolmogorov-Smirnov and Shapiro-Walk can also be used.

Mann-Withney test

- This test assumes only an ordinal level of measurement, since it is based on the ranking of scores. It tests whether the two samples come from the same population (equivalent to the two-sample unpaired t-test).

1. Rank the set of n_1 and n_2 scores from lowest (rank 1) to highest.

2. Let R_i be the sum of ranks of the i-th sample

3. Test statistics is $U = \min\left(n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1,\ n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2\right)$

- The values for U are tabulated for $n<20$. It approximates the normal distribution for larger sizes.

Wilcoxon signed-ranks test

- This test assumes an interval level of measurement. It tests whether the two paired samples come from the same population (equivalent to the two-sample matched-pair t-test).

1. Compute the differences between each pair and rank them.

2. Each rank is given the sign of the difference it corresponds to.

3. Sum the positive ranks and sum the negative ranks. The test statistic is the smallest sum.

- The values are tabulated for n<20. It approximates the normal distribution for larger sizes.

(Two sample) Kolmogorov-Smirnov test

- KS test compares the two empirical distribution functions. The test statistic is

$$D = sup_x \left| F_{1,n_1}(x) - F_{2,n_2}(x) \right|$$

where F_{1,n_1} and F_{2,n_2} are the empirical distribution functions and sup is the supremum function. The values of the test statistic are tabulated.