**Brief Overview of Nonparametric methods**

**Reference: Engineering Biostatistics (by Branislav Vidakovic)**

Book and Matlab codes available at: <http://statbook.gatech.edu/index.html>

1. **Plotting Methods**

* **Probability-Probability (P-P) Plot**
  + **Compares empirical CDF to proposed theoretical CDF**
  + **Requires location and scale of theoretical distribution to be defined**
  + **Not sensitive to extreme outliers**
  + **Easier to see differences in middle of distribution**
* **Quantile-Quantile (Q-Q) Plot**
  + **More robust and popular than P-P plots**
  + **2 types:**
    - **Empirical distribution vs theoretical distribution**
    - **Empirical distribution vs empirical distribution**
    - **Can be misleading and may look approximately linear in the central part for many candidate distributions (be careful)**

****

Figure 01 – Comparison of PP (left) and QQ (right) plots – QQ is sensitive to obs from tails

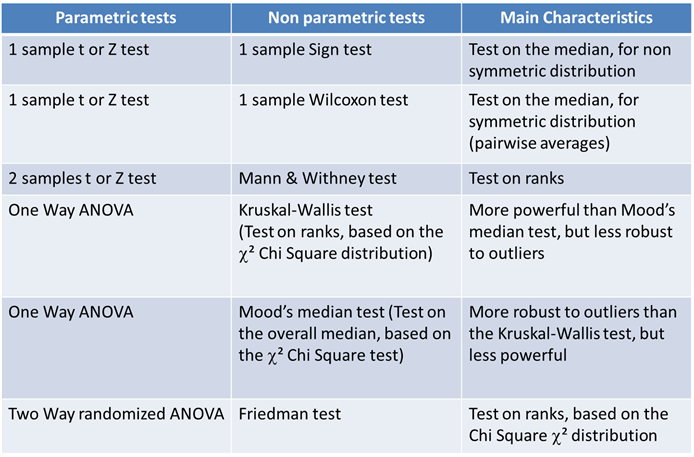
From Engineering Biostatistics (by Branislav Vidakovic), p.850.

1. **Goodness-of-fitness tests (chapter 17 of the book)**

* **Pearson’s Chi-square**
  + **Used to test if empirical distribution is equal to theoretical distribution**
    - **Test can be decisive only if the model does not fit the data well.**
    - **Good for discrete data and special distributions that cannot be fit in a Kolmogorov-type statistical test.**
    - **Degrees of freedom = number of intervals – 1 or, if k parameters need to be specified, Degrees of freedom = number of intervals – k - 1**
  + **Rules for selecting size of bins and interval:** 
    - **For continuous distributions, take intervals approximately equal in prob.**
    - **All np > 1 and at least 80% of np’s > 5s where np = expected frequencies**
      * **Use Yates correction if expected frequencies <5.**
  + **Shortcomings:** 
    - **Depends on n:**
    - **Large n 🡪 overpowered test**
    - **Small n 🡪 Lacks power**
    - **Arbitrary interval size for continuous distribution**
    - **Not efficient for testing known continuous distributions**
* **Kolmogorov-Smirnov Test**
  + **Distribution free method**
  + **No sample size requirements or observed frequency limitations**
  + **Used for continuous distributions**
  + **Compares the maximum distance between two CDFs**
  + **One sample test (“Kolmogorov test”):** 
    - **Compares empirical and theoretical CDF**
  + **Two sample test (“Smirnov test”):**
    - **Compares two samples to see if their distributions are same**
* **Cramer-von Mises, Watson’s & Rosenblatt’s tests**
  + **Integrates differences between two CDFs**
  + **One Sample test (CvM, Watson’s):**
    - **Pretty much anything that CvM is used on, Watson can also be used on with slight modification**
  + **Two samples test (Rosenblatt’s):** 
    - **Finds distance between empirical CDF’s using L2 norm and then find empirical joint distribution**
    - **Better than KS test at detecting tail differences; worse than KS test when detecting differences in middle of distribution**
* **Tests for Departures from Normality**
  + **Jarque-Bera**
  + **Shapiro-Wilk**

1. **Distribution free Methods (chapter 18 of the book)**

* **Sign Test**
  + **Very low power compared to Wilcoxon Signed Rank Test**
  + **Doesn’t care for magnitude against mean or median**
  + **Works for any distribution**
* **Wilcoxon Signed-Rank Test (WSiRT)**
  + **Signs and ranks are considered: it’s the difference between ranks of positive differences and ranks of negative differences**
  + **Can be used to test hypothesis of location**
    - **Assumes symmetry of distribution**
* **Wilcoxon Sum-Rank Test (WSuRT)**
  + **Used in place of two sample t-test when populations are independent but possibly not normal**
  + **Null hypothesis is that two populations are equal**
  + **Assumes shapes of distributions are similar**
* **Kruskal-Wallis Test**
  + **Compares three or more samples**
  + **Null: all populations have identical distribution functions**
  + **Assumes independence among population and assumes group population distributions to be similar in shape.**
  + **Allows pairwise comparisons**
* **Friedman’s Test**
  + **For repeated measurements**
  + **Used when normality is in question or when variances possibly vary from population to population**
  + **Allows pairwise comparisons**
* **Parametric vs Nonparametric counterpart (Summary Table)**

****

1. **Contingency Tables (chapter 12 of the book)**

* **Examples of 2-way and 3-way tables in Excel File (Contingency Table.xlsx)**
* **Two-way tables:**
  + **Chi-square and Likelihood ratio statistics: Used to measure the deviation from independence of the factors by comparing *nij* (observed) and *eij* (expected) over all cells**
  + **Chi-square is not normalized and depends on sample size; distribution depends on number of columns and rows**
  + **Assumptions for both:** 
    - **No empty cells**
    - **Observed counts should not be less than 5 for at least 80% of the cells**
    - **Only total sample size is defined in advance without restriction on marginal counts**
      * **If row or column sums are present in advance, it tests for the homogeneity of one factor with respect to other**
      * **If both marginals are fixed in advance test becomes hypergeometric distribution, use Fisher’s exact test.**
      * **Fisher exact test is based on the null hypothesis that two factors, each with two factor levels, are independent, conditional on fixing marginal frequencies for *both* factors**
  + **Mc Nemar’s Test:** 
    - **Determines if two population proportions resulting from marginal sums are different**
    - **Can test for symmetry and test for marginal homogeneity; for a 2x2 table, these metrics are the same**
    - **Two proportions are NOT independent**
    - **The marginal totals, unlike Fisher Exact Test, are not fixed**
    - **How two-sample paired *t*-test is for normally distributed responses, the McNemar test is used for paired binary responses.**
  + **Cochran’s Q Test**
    - **Equivalent to McNemar’s when pairs are organized as 2x2**
    - **Sensitive to deviations in probabilities**
    - **Often applied to large N problems**
* **Three-way tables:**
  + **Complete Independence: R (rows), C (columns) and P (pages) are independent**
    - **Use Chi-square or Likelihood Ratio statistic**
  + **Joint Independence: test if each single factor is independent of other two**
    - **If rejected, could have many options:**
    - **All three (R, C, P) are dependent**
    - **Two are dependent but both are independent of third**
    - **Use Chi-square**
  + **Conditional Independence: factors may be related via joint relation, but when accounting for the relation, are independent of each other**
    - **Use Chi-square**
  + **Stuart Maxwell Test:**
    - **Use Stuart Maxwell test to test for equality of marginal distributions**
    - **It’s a 3x3 version of McNemar’s test**
* **Stratified tables:**
  + **For 2 or 3 tables when there’s a variable that stratifies data**
  + **Multiple tables make inference more precise by controlling for an influential variable**
  + **Mantel Haenszel’s Test:** 
    - * **Test conditional independence of two factors**
      * **Measure degree of conditional association (risk factors)**
      * **Meta-Analysis of conditional odds ratios**