Notes on robust statistics

January 6, 2025

Introduction

Robust statistics refers to methods designed to provide reliable results even when some assumptions about the data are violated, such as (i) the presence of outliers and / or (ii) wrong assumption on type of distribution.[1] Robust statistics is therefore particularly useful in analysing real-world data, which next to never meet the assumptions of traditional statistical methods. They are essential for reliable data analysis. The text below is a summary of methods for robust estimation of location, scale, linear regression, correlation, and statistical tests.

How to measure quality of statistical methods

When deciding on which statistical method to use, we need to understand their differences and be able to quantify them. In robust statistics, we are interested in two properties: robustness and efficiency.

Robustness

Robustness refers to the ability of a statistical method to provide consistent results even when the data are corrupted by the presence of outliers. The measures includes *breakdown point*, *influence function*, which summarised in *gross error sensitivity*.

Breakdown Point: is equal to the smallest fraction of data points that can be replaced with outliers before the estimator gives an arbitrarily large result. A higher breakdown point indicates a more robust method. For example, the median has a breakdown point of 50%, meaning it can handle up to 50% of the data being outliers without giving an extreme result.

Influence Function: measures sensitivity of an estimator to small changes in the data at a given point. If the influence function is bounded, then the estimator is robust and it would not be affected by small perturbations or outliers. Influence function is defined as

$$\mathrm{IF}(x|p,\eta) = \lim_{\epsilon \to 0} \frac{\eta((1-\epsilon)p + \epsilon\delta_x) - \eta(p)}{\epsilon},$$

where p is the assumed probability density function of data, η is the estimator, and x is a point in which the influence function is evaluated.

Gross error sensitivity: is a maximum over the influence function over the support of the measuresments. It summarizes influence function into a single number. If smaller than infinity, the the estimator is robust to some degree.

$$\mathrm{GES}(p,\eta) = \sup_x |\mathrm{IF}(x)|$$

Efficiency

Efficiency of the method relates variance of the estimate with the number of data points.[2] The robust methods typically sacrifices efficiency for the robustness. If one method is less efficient than the other, then it requires more data points to achieve same variance of the estimate.

[1] Usually most people assume normal distribution, many times without knowing it.

[2] The estimate is random variable, because it is a function of random variables (the data points are realizations of these random variables). The distribution of the estimate depends on the type of the estimator and the distribution function of observed random variables (data points).

When judging the efficiency, it usually related to some well known well established estimator (mean for location, variance for spread) on data points following some known distributions (Normal distribution). The most common measure is **Asymptotic Relative Efficiency** (ARE), which is computed as a fraction of variances of two estimators $\hat{\eta}_1$, $\hat{\eta}_2$, with number of data-points going to infinity. ARE helps you to estimate, how many more data points do you need to get an estimate of the same quality (variance). ARE is defined as

$$\mathrm{ARE}(\hat{\eta}_1, \hat{\eta}_2, p) = \frac{V_2}{V_1},$$

where $\frac{V_1}{n}$, $\frac{V_2}{n}$ are variances of estimators $\hat{\eta}_1$, $\hat{\eta}_2$ of a parameter μ of probability distribution p. Be aware though that computing variances of estimates is generally difficult and might not be analytical.

Estimators of location

Mean: Mean is optimal if the data are sampled from the normal distribution, but it has zero break point. This means that even one outlier can render the estimate useless. While extremely popular, it should be used with caution, as there robust alternatives with only slightly worse efficiency.

For n data points $\{x_i\}_{i=1}^n$, mean is computed as $\frac{1}{n}\sum_i x_i$.

Median: The median is the middle value of a set of sorted points (in case of the cardinality of set is even, it is the average of two closest to the middle.) The median is optimal estimator of location of Laplace distribution. It is extremely robust with 50% breaking point, but it is relatively innefficient, since its asymptotic relative efficiency (ARE) with respect to the normal distribution is 0.637.

For n data points $\{x_i\}_{i=1}^n$, median is the solution of the following optimization problem

$$\mathrm{med}\left(\{x_i\}_{i=1}^n\right) = \arg\min_t \left| \sum_i \mathrm{I}(x_i \geq t) - \sum_i \mathrm{I}(x_i \leq t) \right|,$$

where $\mathrm{I}(\cdot)$ is an indicator function being one if its argument is true and zero otherwise.

Trimmed Mean: Computes mean from a set of measurements, where q% highest and q% lowest values are removed. The rationale behind is that errors in measurements have highest influence when they have extreme values. By removing tails, these extreme values are removed. It is very likely that removing tails remove some normal samples as well, which decreases the efficiency of the estimator. The breakdown point is controlled by q and it is equal to q. The ARE with respect to the normal distribution and for $q = 10\%$ is 0.943.

Winsorized Mean: Computes mean from a set of measurements, where q% highest and lowest values are replaced by closest valid values. The method is based derived from M-estimators explained below, where the rationale becomes clear. Winsorized mean can be useful for datasets with skewed distributions. The breakdown point is controlled by q and it is equal to q.

Hodges-Lehman: For n data points $\{x_i\}_{i=1}^n$, the Hodges-Lehmann estimator is calculated by:

1. Forming all possible pairs $\{(x_i, x_j)\}_{i=2,j=1}^{n,i}$.
2. Calculating the average of each pair.
3. Taking the median of these averages.

The formula is therefore $\text{med}\left(\left\{\frac{1}{2}(x_i - x_j)\right\}_{i=2,j=1}^{n,i}\right)$. It is very robust with 0.29 breakdown and very efficient with ARE = 0.955 against Normal distribution. While not very widely known, it is an excellent choice.

Estimators of scale

Sample standard deviation (STD) is the most common method to estimate scale. For n data points $\{x_i\}_{i=1}^n$, the unbiassed version is computed as

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

and the biassed formula version as

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The biassed version is sometimes preferred, as it has lower variance. The method is fragile, as the breakdown point is 0, but it is optimal for Normal distribution.

IQR One of the most common measures is the Interquartile Range (IQR), which is the difference between the 75th and 25th percentiles of the data. IQR is robust, but it is not very efficient. The IQR is only indirectly related to variance. To obtain estimate of the variance, it has to be divided by $2\sqrt{2}\text{erf}^{-1}(\frac{1}{2}) \approx 1.349$. The breakdown point is 25%.

Median absolute deviation (MAD) estimates the scale of the data by taking the median of the absolute differences between each data point and the median of the data. It is robust and efficient, with a breakdown point of 50%. ARE of MAD with respect to the the sample standard deviation is 37%. For n data points $\{x_i\}_{i=1}^n$, MAD is computed as

$$\text{MAD} = \text{med}\{|x_i - \text{med}\{x_i\}|\}.$$

S_n Robust and efficient estimator of scale was proposed in [3]. For n data points $\{x_i\}_{i=1}^n$, the formula is

$$S_n = \text{med}_i\{\text{med}_j|x_i - x_j|\}.$$

To use it to estimate standard deviation of normal distribution, the value has to be multiplied by 1.1926, i.e. $\hat{\sigma} = 1.1926 \cdot S_n$. The method is robust with breakdown point 29% and efficient with ARE to normal distribution being 0.86.

Q_n The second estimator of scale proposed in [3] uses formula

$$Q_n = \{|x_i - x_j||i < j\}_{q_{25}}.$$

To use it to estimate standard deviation of normal distribution, the value has to be multiplied by 1.0483. This method is more robust than S_n with breakdown point 50%, but it is slightly less efficient with ARE to normal distribution being 0.82.

M-estimators

Normal distribution — mean

Why mean is so popular? It turns out that that mean is an optimal (minimum variance and unbiassed) estimator for the normal distribution. This is because it is solution to maximum likelihood estimator. Assuming n data points $\{x_i\}_{i=1}^n$, the derivation is as follows. The estimate of μ is the point at which the likelihood function for data points attains maximum, i.e.

$$\arg\max_{\mu} \mathcal{L} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}.$$

To remove the product, we can find the maximum of the logarithm of $\mathcal{L}$ instead. Optimizing the logarithm does not change the solution, since the argument is positive and the logarithm is a monotonous function. With that the optimization problem simplifies to

$$\arg\max_{\mu} \log \mathcal{L} = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \log \sqrt{2\pi}\sigma$$

To find an extrema, we take the derivative with respect to μ, set it to zero, and after algebric simplification, we obtain a formula for mean

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Laplace distribution — median

Let's now assume the data points to be drawn from Laplace distribution, i.e $x_i \sim \text{Laplace}(\mu, \sigma)$ with probability density function $p(x|\mu, \sigma) = \frac{1}{2\sigma} e^{-\frac{1}{\sigma}|x-\mu|}$. Following the same steps as above, we start by formulating the likelihood function as

$$\arg\max_{\mu} \mathcal{L} = \prod_{i} \frac{1}{2\sigma} e^{-\frac{1}{\sigma}|x_i-\mu|}.$$

Again, we maximise its logarithm

$$\arg\max_{\mu} \log L = -\frac{1}{\sigma} \sum_{i} |x_i - \mu| - \log 2\sigma.$$

We derive the estimator by setting the gradient to zero and do little bit of algebraic manipulations.

$$0 = \sum_{i} \text{sgn}(x_i - \mu) = \sum_{i} \text{I}[x_i > \mu] - \sum_{i} \text{I}[x_i < \mu]$$

and we have obtained the median.

Generalization

Let's now assume a family of exponential probability distributions with density functions defined as $\frac{1}{Z(\sigma)} e^{-\rho(\frac{x-\mu}{\sigma})}$. The properties of the distribution are determined by a function ρ, which can be any function with the integral $\int_{-\infty}^{\infty} e^{-\rho(\frac{x}{\sigma})} dx = Z(\sigma)$ being finite, i.e. $Z(\sigma) =< +\infty$. $Z(\sigma)$ is called a partition constant and normalizes the probability density function such that its integral over the range is one. The distribution is parametrized by the location and scale parameters, as we usually assume ρ to be symmetric.

We can derive a maximum likelihood estimator by repeating the above steps for a general function σ. The likelihood function is defined as

$$\arg\max_{\mu} \mathcal{L} = \prod_i \frac{1}{Z(\sigma)} e^{-\rho(\frac{x_i-\mu}{\sigma})}.$$

We take its logarithm

$$\arg\max_{\mu} \log \mathcal{L} = -\sum_i \rho(\frac{x_i-\mu}{\sigma}) - \log(\sigma),$$

compute gradient, and find point that is equal to zero.

$$0 = \sum_i \rho'(\frac{x_i-\mu}{\sigma}).$$

Now realize, that the robustness robustness of the solution is determined by the properties of ρ'. If $\rho'(x)$ is bounded, then the solution is robust, and the other way around if $\rho'(x)$ is unbounded, the solution is not robust. In fact, there is a very close connection of $\rho'(x)$ to the influence function introduced above to charaterize robustnes of detectors. This property is used to design a family of functions ρ, such that their maximum likelihood estimates yield a robust estimator of μ.

We list three most widely known functions.

Huber loss A (not called Huber) [2], defined as

$$\rho = \begin{cases} \frac{x^2}{2} & |x| < a \\ \frac{a^2}{2} & |x| \geq a \end{cases}$$

with derivative

$$\rho' = \begin{cases} x & |x| < a \\ 0 & |x| \geq a \end{cases}$$

which yields an estimator $\mu = \frac{1}{n_{<a}} \sum_{i||x_i|<a} x_i$, which corresponds to *trimming estimator* of location introduced above

Huber loss B (called Huber) [2], defined as

$$\rho = \begin{cases} \frac{x^2}{2} & |x| < a \\ a|x| - \frac{a^2}{2} & |x| \geq a \end{cases}$$

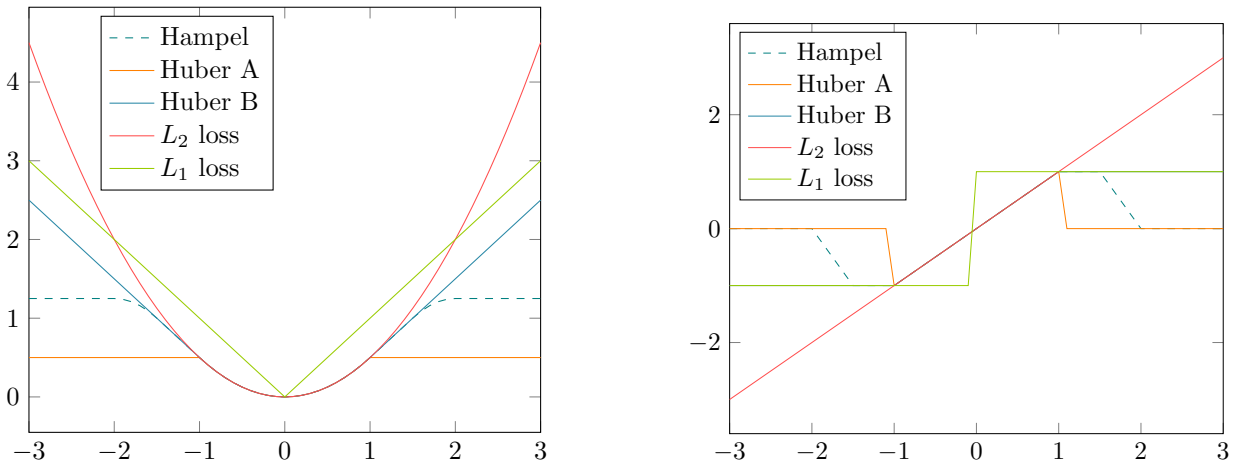
with derivative

$$\rho' = \begin{cases} x & |x| < a \\ a \cdot \mathrm{sgn}(x) & |x| \geq a \end{cases}$$

which yields an estimator $\mu = \frac{1}{n}\left[\sum_{i|abs(x_i)<a} x_i + n_{>a} \cdot a\right]$, which corresponds to *Windsorizing estimator* of location introdused above. Huber loss B is very popular in machine learning, because it is convex and therefore leads to optimization problems with more stable convergence.

Hampel loss [1] defined as

$$\rho = \begin{cases} \frac{x^2}{2} & 0 \leq x < a \\ ax - \frac{a^2}{2} & a \leq x < b \\ \frac{a(x-c)^2}{2(b-c)} + \frac{1}{2}a(b+c-a) & b \leq x < c \\ \frac{1}{2}a(b+c-a) & c \leq x \end{cases}$$



(a) ρ (b) ρ'

Figure 1: Functions ρ used in M-estimators of location (left) together with their derivatives ρ'.

with derivative

$$\rho' = \begin{cases} x & 0 \le x < a \\ a & a \le x < b \\ \frac{a(x-c)}{b-c} & b \le x < c \\ 0 & c \le x \end{cases}$$

Hampel loss is complicated and rarely used in practice due to the large number of parameters. The shape of its influence function was used during design of robust estimator of scale S_n and Q_n.

Robust linear regression

The math behind M-estimators can be used to design robust linear regression. The standard linear regression is based on the assumption that the residuals are normally distributed. The robust alternative is to use M-estimators with a robust ρ function.

To derive the robust linear regression, we assume a linear model of dependant variable y on independant variables $x \in \mathbb{R}^d$ subjected to aditive noise P_ϵ i.e.

$$y = x^{\mathrm{T}}\beta + \epsilon, \epsilon \sim P_\epsilon$$

The standard linear regression is based on the assumption that the residuals are normally distributed, i.e. $P_\epsilon = \mathcal{N}(0, \sigma^2)$. Assuming a set of data points $\{(x_i, y_i\}_{i=1}^n$, we can use maximum likelihood estimate method to estimate parameters as above to estimate parameters β of the model. Therefore, the maximum likelihood estimate of β is the solution of the following optimization problem

$$\hat{\beta} = \arg\min_\beta \sum_i (x_i^{\mathrm{T}}\beta - y_i)^2.$$

Assuming different distribution of noise, e.g. Laplace $\epsilon \sim \mathrm{Laplace}(0, \sigma)$, following the same steps we obtain recover **median absolute regression**

$$\hat{\beta} = \arg\min_\beta \sum_i |x_i^{\mathrm{T}}\beta - y_i|.$$

Replacing the ρ function with Huber B function we recover **least trimmed squares (LTS)**

$$\hat{\beta} = \arg\min_\beta \sum_i (x_i^{\mathrm{T}}\beta - y_i)^2_{(q)},$$

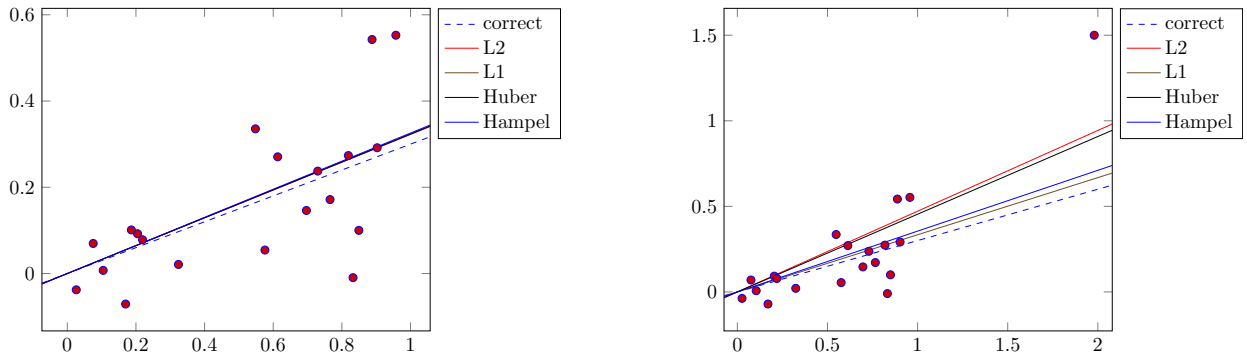


Figure 2: An example of robust regression on a simple 1d problem $y = 0.1x + \mathcal{N}(0, 0.1)$. The left figure is without outliers, the right figure is with outliers.

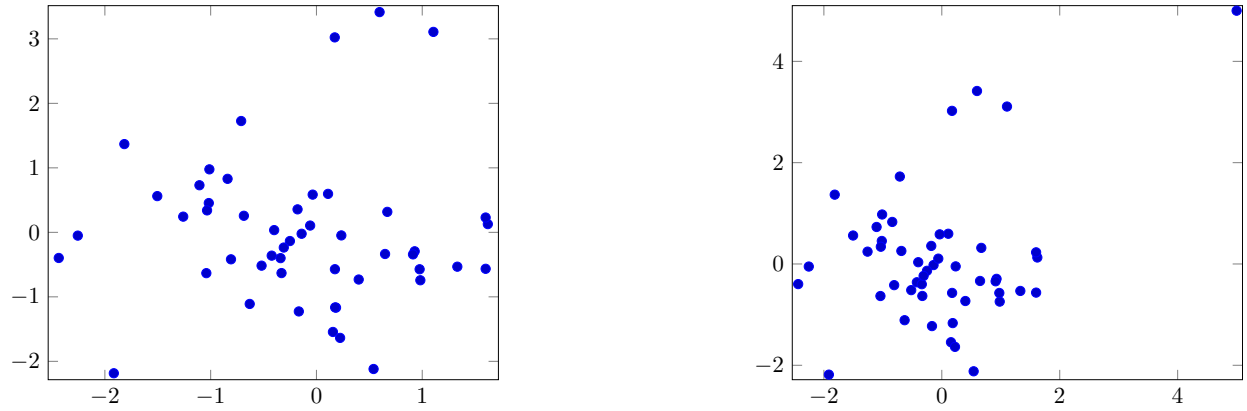


Figure 3: Figures show samples drawn from normal distribution with diagonal covariance matrix (i.e. x_i and y_i are independent). The left Figure shows samples without any outlier, for which Pearson correlation coefficient is $\rho = -0.0008$ suggesting the data are independent. The right figure shows (almost) the same samples with one outlier, for which Pearson correlation coefficient is $\rho = -0.3245$ suggesting the data are dependent, which is wrong.

where $\cdot_{(q)}$ represents an order statistics (quantile).

An example of robust 1d regression is showin in Figure 2.

Robust estimate of correlation

Correlation in statistics referes to any type of relationship between two random variables. In practice, the term "correlation" almost always implies linear correlation measured by *Pearson's correlation.*

Let's assume that we have two sets of data points $\{(x_i, y_i)\}_{i=1}^n$. The **Pearson's correlation** is defined as

$$\rho_{X,Y} = \frac{\frac{1}{n}\sum_i[(x_i - \bar{x})(y_i - \bar{y})]}{\sigma_x \sigma_y},$$

where $\bar{x}$, $\bar{y}$ denotes a sample mean, σ_x and σ_y are estimates of standard deviation of $\{(x_i)\}_{i=1}^n$ and $\{(y_i)\}_{i=1}^n$ respectively. It should be now obvious that the Pearson's correlation is very fragile with breakdown point of zero, i.e. even a single bad data point can arbitrarily largely influence the resulting estimate (see Figure 3 (right).)

The most common robust version of correlation is **Spearman's correlation**. The idea behind is to use rank of data points instead of their values. For example if we have data points

$$\{(0.1, 0.5), (0.05, 0.7), (0.4, 0.1), (0.2, 100)\},$$

their ranks will be

$$\{(2, 2), (1, 3), (4, 1), (3, 4)\}.$$

Spearman's correlation uses rank to compute the correlation similarly to Pearson's, i.e.

$$r_s = \frac{\frac{1}{n}\sum_i [(r_i^x - \bar{r}_x)(r_i^y - \bar{r}_y)]}{\sigma_{r_x}\sigma_{r_y}},$$

where r_i^x denotes a rank of x_i in $\{x_i\}_{i=1}^n$, $\bar{r}_x$ denotes a sample mean of ranks, and σ_{r_x} standard deviation of ranks (and similarly for y). The Spearman's correlation is robust with breakdown point of 50%. For the example with an outlier in Figure 3 (right), the Spearman's correlation is $r_s = -0.0008$ suggesting the data are independent, which is correct.

Kendalls'τ correlation goes further and replaces all quantities with their relative order. The formula uses definition of *concordant* and *disconcordant* pairs.

- *Concordant* pairs are pairs of observations where orders of (x_i, x_j) and (y_i, y_j) are the same, i.e. $x_i < x_j$ and $y_i < y_j$ or $x_i > x_j$ and $y_i > y_j$;
- *disconcordant* pairs are pairs of observations where orders of (x_i, x_j) and (y_i, y_j) do not agree, i.e. $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$.

Denoting number of concordant pairs as n_c and disconcordant pairs as n_d, the Kendalls'τ is defined as

$$r_k = \frac{1}{\binom{n}{2}}(n_c - n_d).$$

Kendall's τ has interesting properties. If the variables x_i and y_i in $\{(x_i, y_i)\}_{i=1}^n$ are independent, then the coefficient r_k has known distribution, i.e. $r_k \sim \mathcal{N}\left(0, \frac{2(2n+5)}{9n(n-1)}\right)$, which can be used in statistical testing of independence.

The above definition of Kendalls'τ is for continuous data. For discrete data, the definition is slightly different, as we have to account for ties.

Robust statistical tests

Robust statistical tests are robust variants of statistical tests. These tests should be used when we do not know the distribution of the data points (recall popular t-tests assumes Normal distribution) or data might contain outliers.

Sign test assumes data points to to come in pairs $\{(x_i, y_i)\}_{i=1}^n$ (to be matched in the language of statistics). *Sign test* tests if differences between pairs of observations are consistent.

The test statistic is computed as follows:

1. discard samples for which $|y_i - x_i| = 0$;
2. test statistic

$$W = \sum_{i=1}^{n} I(y_i > x_i)$$

Under null hypothesis W follows binomial distribution $\mathrm{Bi}(N, 0.5)$.

Wilcoxon-signed rank test assumes data points to come in pairs $\{(x_i, y_i)\}_{i=1}^n$ (to be matched in the language of statistics). The test tests if location of distribution generating $\{x_i\}_{i=1}^n$ is the same as that generating $\{y_i\}_{i=1}^n$. The idea is to test, if difference between pairs follows a symmetric distribution around zero.

The test statistic W is computed as follows:

1. calculate $\{|y_i - x_i|\}_{i=1}^n$ and discard those with $|y_i - x_i| = 0$ (for simplicity of the exposition, we assume there are not such points);
2. rank remaining samples according to $|y_i - x_i|$ (we denote the rank as r_i);
3. test statistic
$$W = \sum_{i=1}^{n} [\mathrm{sgn}(y_i - x_i) \cdot r_i]$$

Under null hypothesis, where we assume that locations are equal, the test statistic W has known distribution with zero mean and variance $\sigma_w^2 = \frac{n(n+1)(2n+1)}{6}$. For small n critical values of W are tabulated, for large n it approximated by Normal distribution due to central limit theorem.

Wilcoxon-signed rank test should be go-to test, because even for the case where both $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ follows normal distribution and paired t-test is more appropriate, the Wilcoxon-signed rank test has efficiency ARE = 0.95, which is very good. In other words for its robustness, we pay very little in terms of efficiency.

Mann-Whitney U-test is non-paired version of the *Wilcoxon-signed rank test*. The test tests if two populations $\{x_i\}_{i=1}^{n_x}$ and $\{y_i\}_{i=1}^{n_y}$ have the same location.

The test statistic is computed as follows:

1. Calculate ranks of all samples together $\{x_i\}_{i=1}^{n_x} \cup \{y_i\}_{i=1}^{n_y}$.
2. Sum ranks of samples from the first population, R_x.
3. Sum ranks of samples from the second population, R_y.
4. Calculate $U_x = R_x - \frac{n_x(n_1+1)}{2}$ and $U_y = R_y - \frac{n_y(n_y+1)}{2}$.
5. $U = \min\{U_x, U_y\}$

For small n_x and n_y, the critical values of test statistic U for null hypothesis are tabulated. For large n_x and n_y we use again approximation by Normal distribution $\mathcal{N}\left(\frac{n_x n_y}{2}, \frac{n_x n_y (n_x+n_y+1)}{12}\right)$.

Bibliography

[1] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

[2] Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:492–518, 1964.

[3] Peter J Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993.