NONPARAMETRICS

WHEN TO USE:

- Non-parametrics do not require knowledge about the underlying distribution. Samples do not have to come from normal distribution.
- Lots of the tests can be used on categorical data.
- The calculations can be easier than other methods.

DEFINITIONS:

- Robust: insensitive to deviations/outliers
- Breakdown Point: a measure of robustness, the contaminated fraction that ruins how a test works
- Rank: a number assigned to a datapoint according to its order in a sorted list

TESTS:

Sign Test:

Converts data points into +/- to see which has more Then see if it is significant that +/- has more.

- Data must be randomly selected
- Does not have to come from a particular distribution

Does The Median Of The Variables HAVE THE VALUE d_0?

$H_0 = d = d_0$, $H_1 = d \neq d_0$, Random Variables: $X_1, X_2, \ldots X_i$

$$S(d_0) = \sum_{i=1}^{n} \operatorname{sgn}(X_i - d_0)$$

→ Can use z-statistic/binomial probabilities to see if significant

$$z = \frac{(x + 0.5) - n/2}{\sqrt{n}/2}$$

x - # of the less freq. sign

n - total number

Wilcoxon Signed Ranks

Used to test if the population of matched-pair differences has a median other than 0. So, if the two samples are different in median.

H_0 : the matched pairs have differences w/ a median of 0.

H_1 : the matched pairs have differences w/ a non-zero median

Example: compare rainfall in January vs. July

Location	January	July	diff.	Abs.	rank
Houston	0	5	5	5	3
Dallas	2	1	-1	1	1
Austion	1	3	2	2	2

Steps: Find the differences in the paired samples
↳ rank them

Sum the ranks of the negative diff. $T_- = 1$

Sum the ranks of the positive diff. $T_+ = 5$

W_{stat} = smaller of the Ts, so $W_{stat} = 1$

Then use signed ranked table w/ W_{stat} (have to pick $W_{crit.}$)

Or for larger samples: $$Z = \frac{W_{stat} - \frac{n(n+1)}{4}}{\sqrt{n(n+1)(2n+1)/24}}$$

Wilcoxon Rank Sum

This test uses ranks of sample data from two independent population to test if they have different medians. If the medians are equal the ranks should be even.

H_0 : Samples come from populations w/ equal medians

H_1 : Samples come from pop. w/ different medians

Steps: Asign ranks to all of the data points

Sample 1	rank	Sample 2	rank
3	1	12	8
5	3	6	4
11	7	7	5.5
4	2	7	5.5

Sum the ranks of the samples

$T_1 = 13, \quad T_2 = 23$

Calculate test statistic

$$z = \frac{T_1 - \mu_1}{\sigma_1} \qquad \mu_1 = \frac{n_1(n_1+n_2+1)}{2} \qquad \sigma_1 = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$$

either sample can be used as sample 1

↓

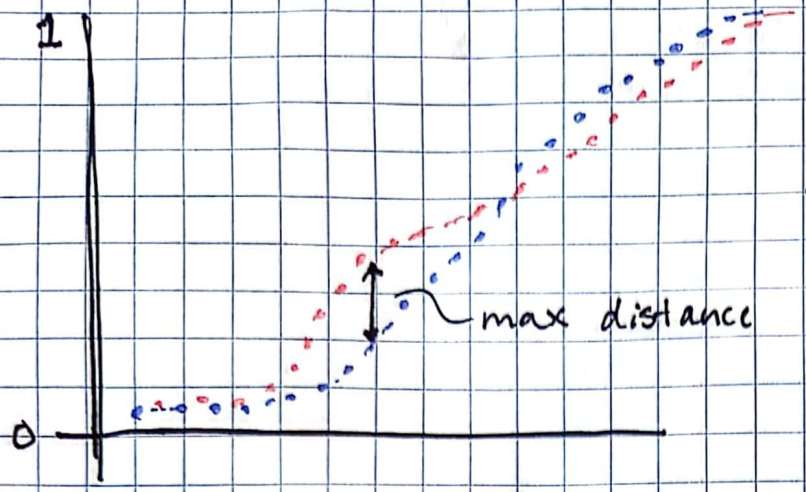
If more than 2 independent samples use Kruskal - Wallis Test

KS - Test (Kolmogorov - Smirnov)

Used to compare the shape of distributions. Can compare shape between two samples or the shape of one sample to a known dist. (like normal, exponential etc.)

H_0 : two samples are drawn from the same distribution

H_1 : the samples are drawn from different dist.



We want to know if these two distributions are the same or not.

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \leq x]$$

e.d.f, heights from 0 to 1 steps of $1/n$

For 1 sample test:

$$M_{KS} = \sqrt{n}\ \max_x |\hat{F}_n(x) - F_0(x)|$$

F_0 is a pre-specified function

measured the max diff. between e.d.f & model. The larger M_{KS} is the more confident you can be in the rejection of the null.

For 2 sample test:

$$M_{KS} = \max_x |\hat{F}_m(x) - \hat{G}_n(x)|$$

$\hat{F}_m$ & $\hat{G}_m$ are the edfs of two different samples

The maximum vertical distance between the functions in measured.

For More 1-sample tests | 2 sample dist. tests:

Cramer-von-Mises Statistic

tests to see if the e.d.f of the data and a model are different. (like the KS test).

$$T_{CVM} = n\int_{-\infty}^{\infty} [\hat{F}_n(x) - F_0(x)]^2\, dF_0(x) = \frac{1}{12n} + \sum_{i=1}^{n}\left(\frac{2i-1}{2n} - F(X_i)\right)$$

sum of the squared differences

CvM better describes global & local differences so it can be better than the KS test.

ANDERSON-DARLING TEST (AD)

CvM / KS don't assess differences in the high/low x values well so AD, weights the CvM stat.

$$A^2_{AD,n} = n \sum_{i=1}^{n} \frac{[i/n - F_0(x)]^2}{F_0(x_i)(1-F_0(x_i))}$$

↑ still has the same goal as KS & CvM, just a different way of calculating the statistic

Kendall Tau's Correlation Test

Used to determine if there is a relationship between two variables.

Steps:

- Rank each variable seperately
- For each mass object determine which lumin. rank corresponds to it.
- Count the # of concordant & discordant points
- Calculate statistic:

$$\tau = \frac{C-D}{C+D} \qquad z = \frac{3\tau\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}$$

Example: mass & Luminosity of star

Mass	Lumin.	C	D
1	2	3	1
2	1	3	0
3	4	1	1
4	3	1	0
5	6		
		Total: 8	2

Spearman-ρ Test

Similar to Kendall-τ

Step

- Give each object a ranking for each variable
- Find the difference between the two ranks
- Use ρ stat.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

- Significance level: $t = \frac{\rho \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$

Example.

Mass	Lumin.	d
1	2	-1
2	1	1
3	4	-1
4	3	1
5	6	-1
6	5	1

χ^2 Test

Used on categorical data to see if attribute is more common among a specific type. This is used on contingency tables.

Example:

	Class I	Class II	Total
Yes	a	b	n_1
No	c	d	n_2
Total	c_1	c_2	n

So, is a property more common in Class II?

$$T_{\chi^2} = \sqrt{\frac{n}{n_1 n_2}} \frac{ad-bc}{\sqrt{c_1 c_2}} + 0.5$$

$$\longrightarrow T_{\chi^2, r \times c} = \sum_{i=1}^{c} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

O - observed

$E - n_i c_j / n$

to generalize to rxc matrix