Statistical Significance Tests to assess validity of AI models

Given multiple AI models, how do we compare which model will perform better than others on unseen data? The key to the answer lies in statistical significance tests. Statistical tests help us make conclusions about the population using information about samples and help us rule out anomalous or chance results.

Every statistical test begins with a null hypothesis, which is a specific statement about a population parameter made for the purposes of argument. The alternative hypothesis includes all other feasible values for the population parameter besides the value stated in the null hypothesis. When comparing two AI models, the null hypothesis is that the median performance level of Model 1 is equal to that of Model 2 while the alternative hypothesis is that the performance of Model 1 is not equal to Model 2, i.e. the median performance level of Model 1 is greater than Model 2 or vice versa.

To compare if one model is better than the other, we consider some metric of performance. One of the most common metrics to measure performance of the AI models that classify an input into a binary output is Area under the Receiving Operator Characteristics curve (AUROC).

There are two types of statistical tests -

1. **Parametric tests** - These tests assume that the population data follows a normal distribution. Parametric tests are in general more powerful (require a smaller sample size) than nonparametric tests **(Chin, 2008)**. However, this also makes the tests more prone to Type I error. Some popular parametric tests are as follows:

    - One-Sample t/Z-Test
    - Unpaired 2 Sample t/Z Test
    - Paired 2 Sample Z/t-Test
    - ANOVA

2. **Non-parametric tests** - No assumptions about the distribution of population data is made. As these tests, in general, require a larger sample size than parametric tests to reject the null hypothesis, the tests are more prone to Type II error. A few examples of non-parametric tests are:

    - One-Sample Sign Test
    - One-Sample Wilcoxon Signed Rank Test
    - Mann-Whitney U Test

In the case of hypothesis testing for population proportion, the Z-statistic (and Z-test) is always used as when sample sizes are large, the approximate normality assumptions hold for both the sample proportion and the test statistic. **(Zou, et al. 2003)**

In case of hypothesis testing for population mean, the Z-statistic (and Z-test) is used when the population is normally distributed and the population standard deviation is known or when the sample size is greater than 30. The t-statistic (and t-test) is used when there are unknown population variances but the populations are normally distributed. If such assumptions about the normal distribution of the population cannot be made, we may try nonparametric methods. **(Park, Hun Myoung 2009)**

Let us look at each of these tests in more detail.

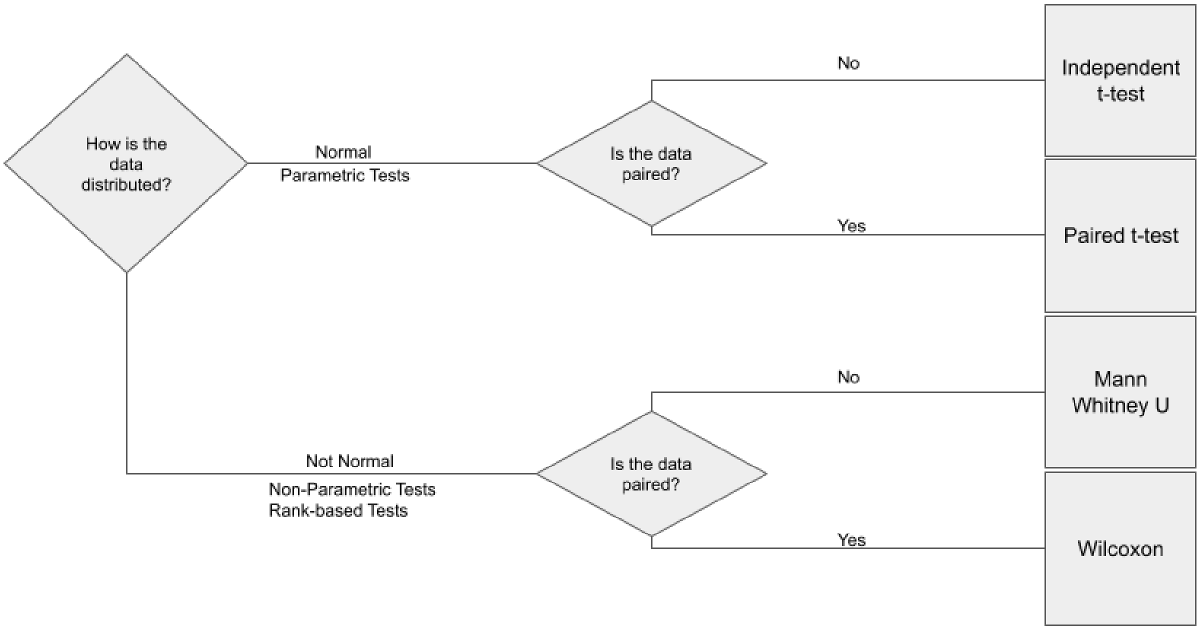
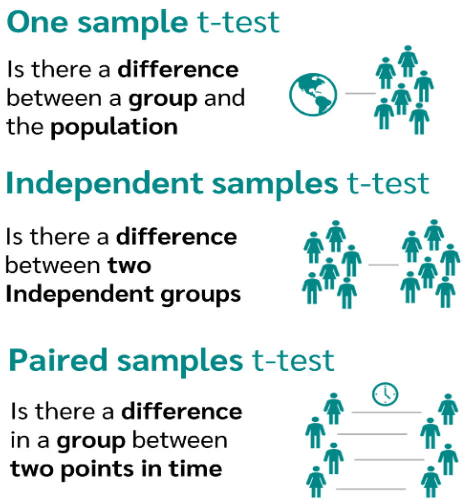


Figure 1. A simple algorithm to determine which statistical test must be performed

Types Of Parametric Tests



One sample t-test

Is there a **difference** between a **group** and the **population**

Independent samples t-test

Is there a **difference** between **two Independent groups**

Paired samples t-test

Is there a **difference** in a **group** between **two points in time**

Figure 2. Types of t-tests. One sample t-test to compare sample mean with population mean proposed in null hypothesis. Independent samples t-test to compare means of two populations. Paired sample t-test to identify effects of a treatment on a single sample. Source : Datatab.net

- **One Sample t/Z-Test:-** It is a test to compare the sample mean or proportion with the population mean or proportion proposed in the null hypothesis. **(Whitlock & Schluter, 1989 Analysis of Biological data, 2nd edition)**

- **Unpaired two-Sample t/Z Test:-** It is a test to compare two different population means or population proportions. Independent samples from two populations are taken and sample means or proportions are calculated.

- **Matched pair Z/t-Test:-** This test is used to compare the effects of a treatment on a single population.
  A random sample is taken from a population and a parameter is measured for each member of the sample before and after treatment . The mean of differences before and after treatment is calculated. The null hypothesis is that the treatment has no effect on the population. If the population standard deviation of difference in means is known, we use the Z-test, otherwise we use the t-test.

- **ANOVA:-** The Analysis of variance is a generalisation of the unpaired two-sample t-test for means. It provides a statistical test of whether two or more population means are equal. The null hypothesis of ANOVA is that the population means are the same for all treatments/populations. Rejecting the null hypothesis in ANOVA is evidence that the mean of at least one group is different from the others. The key insight of ANOVA is that we can estimate how much variation among group means ought to be present from sampling error alone if the null hypothesis is true.

An F-statistic is calculated that tells the ratio between group mean square and error mean square. The group mean square is proportional to the observed amount of variation among the group sample means. The error mean square estimates the variance among subjects that belong to the same group. If the null hypothesis is true, and the means do not differ, the group and error mean squares on average will be similar and F should be close to 1. If the null hypothesis is false, the real differences among group means should inflate the group mean square and F is expected to exceed 1. Those are the only two possibilities. Using an F-distribution, we can carry out a formal test to accept or reject the null hypothesis. **(Whitlock & Schluter, 1989 Analysis of Biological data, 2nd edition)**

Types Of Non-Parametric Tests

- **One-Sample Sign Test:-** This test is used to test whether the median value for a single data set is equal to a hypothesized value. Each value in the sample above the hypothesized median is given a positive sign and each value below the hypothesized median in the sample is given a negative sign. The One-sample sign test compares the number of negative signs with the number of positive signs. The null hypothesis is that the number of positive signs is equal to the number of negative signs.

- **Wilcoxon Signed Rank Test:-** Analogue to the matched pair t-test, this test is used to compare effects of a treatment on a single population, if the differences between the matched pairs are non-normally distributed.
  To perform this test, the matched pair differences are signed and ranked. The test statistic is calculated by summing the similarly signed ranks and taking the smaller sum of the two.

  Depending on the test statistic, we may or may not get significant evidence for our hypothesis.

- **Mann-Whitney U Test:-** To compare 2 different population medians. Independent samples from 2 populations are taken and observations of both samples are ranked. Sum of ranks of each sample is calculated and the test statistic is the smallest sum.

  Depending on the test statistic, we may or may not get significant evidence for the hypothesis.

Let’s look at a problem we often encounter at Deeptek.AI. Consider an AI model to identify the presence of lung nodules in Chest X rays (CXRs) in a CXR scan. We need to carry out a statistical significance test of whether the radiologists aided with the AI model can better identify nodules in a CXR scan. Suppose we gave ten radiologists a sample of 4500 CXR scans collected from multiple centers. They were made to identify nodules in each sample with and without AI aid. Aided and unaided ROC curves were constructed, based on the results of each of the samples. The AUROC values were calculated. Results are shown below.

Unaided	Aided
0.7653	0.7883
0.7614	0.8027
0.7627	0.7990
0.7562	0.7917
0.7572	0.8043
0.7536	0.7922
0.7626	0.8007
0.7608	0.7894
0.7567	0.7919
0.7505	0.7966

Table 1. An example list of 10 pairs of AUROC values based on radiologists detecting nodules in chest X-rays without the AI aid (unaided) and with aid (aided).

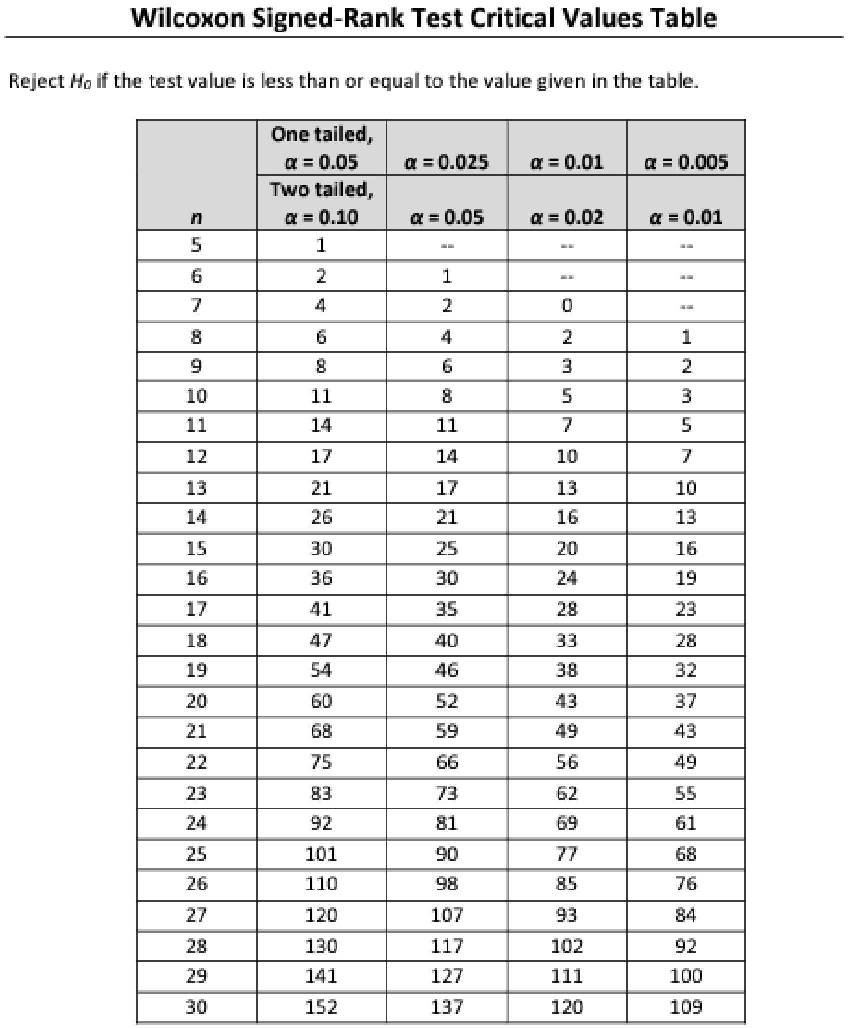
This data is sample input for a statistical significance test to be carried out to see whether radiologists aided with the model consistently perform better than when they were unaided. As we have paired data and the differences between each matched pair may not be normally distributed, we perform the Wilcoxon Signed Rank test on the data to assess whether the AUROC are consistently higher for radiologists aided with the AI than those unaided.

Unaided	Aided	Difference (Aided-Unaided)	Absolute Difference	Rank	Signed Rank
0.7653	0.7883	0.0230	0.0230	1	1

0.7614	0.8027	0.0413	0.0413	8	8
0.7627	0.7990	0.0363	0.0363	5	5
0.7562	0.7917	0.0356	0.0356	4	4
0.7572	0.8043	0.0471	0.0471	9	9
0.7536	0.7922	0.0386	0.0386	7	7
0.7626	0.8007	0.0381	0.0381	6	6
0.7608	0.7894	0.0286	0.0286	2	2
0.7567	0.7919	0.0353	0.0353	3	3
0.7505	0.7966	0.0462	0.0462	10	10

Table 2. The differences between unaided and aided AUROC values in table 1 are calculated, signed and ranked.

Test statistic is 0 as sum of negative signed ranks is 0.



Wilcoxon Signed-Rank Test Critical Values Table

Reject H_0 if the test value is less than or equal to the value given in the table.

	One tailed, $\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
n	Two tailed, $\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.02$	$\alpha = 0.01$
5	1	--	--	--
6	2	1	--	--
7	4	2	0	--
8	6	4	2	1
9	8	6	3	2
10	11	8	5	3
11	14	11	7	5
12	17	14	10	7
13	21	17	13	10
14	26	21	16	13
15	30	25	20	16
16	36	30	24	19
17	41	35	28	23
18	47	40	33	28
19	54	46	38	32
20	60	52	43	37
21	68	59	49	43
22	75	66	56	49
23	83	73	62	55
24	92	81	69	61
25	101	90	77	68
26	110	98	85	76
27	120	107	93	84
28	130	117	102	92
29	141	127	111	100
30	152	137	120	109

Table 3. Critical values of Wilcoxon Signed-Rank Test. Source : Openpress of The University of Saskatchewan, Canada

As we can see, we have evidence at the 0.005 significance level that the AUROC are consistently higher for radiologists aided with the AI than those unaided, since our test statistic (0) is less than critical value when n=10.

References

T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," in *Neural Computation*, vol. 10, no. 7, pp. 1895-1923, 1 Oct. 1998, doi: 10.1162/089976698300017197.

Richard Chin, Bruce Y. Lee, in Principles and Practice of Clinical Trial Medicine, 2008

Comparing Group Means: T-tests and One-way ANOVA Using Stata, SAS, R, and SPSS (Park, Hun Myoung 2009)

Zou, K. H., Fielding, J. R., Silverman, S. G., & Tempany, C. M. C. (2003). Hypothesis Testing I: Proportions. Radiology

Whitlock & Schluter, 1989 Analysis of Biological data, 2nd edition

OPENPRESS.USASK.CA

Datatab.net