Model name	Hierarchical Model	Output type	Common settings / metrics	Common Frequentist test (Parametric)	Common Frequentist test (Non-Parametric)	Assumptions	Additional info
Binary model	Bernoulli distribution with Beta prior	binary output	correct vs incorrect predictions	Binomial test	bootstrap / permutation	2	
Binomial model	Binomial distribution with Beta prior	binomial output	Exact match, Accuracy, Recall, UAS (sentencelevel), LAS (sentencelevel)*	Binomial test	bootstrap / permutation	2,3,6	
Metric model	T-Student distribution with muliple priors	metric observations	Exact match, Accuracy, Recall, UAS (sentencelevel), LAS (sentencelevel), running time. energy usage, L2 error	t-test	bootstrap / permutation	1,2,4	In this model (unlike frequentist t-test) outliers don't need to be discarded manually to realize the strict normality assumption. The observations are assumed to follow t-distirbution with unknown normality parameter.
Count model	Negative Binomial distribution with Normal prior	counts	The count of certain patterns an algorithm could find in a big pool, in a fixed amount of time. Notice that you can't convert this into a ratio form, since there is no well-defined denominator. Ex: measuring how many of questions could be answered correctly (from an infinite pool of questions) by a particular QA systems, in a limited minute (the system is allowed to skip the questions too)		bootstrap / permutation	2,5	
Ordinal model	Normal distribution with parameterized tresholds	ordinals	Collection of objects/labels arranged in a certain ordering, not necessarily with a metric distance between them; for example sentiment labels (https://www.aclweb.org/anthology/S16-1001.pdf), product review categories, grammaticality of sentences		bootstrap / permutation	2	
2	TBA (not implemented yet)	Contengency table/confusion matrix			McNemar's test, bootstrap / permutation	2	
Generalized metric model	TBA (not implemented yet)	metric observations			bootstrap / permutation	2	

Assumption 1: assuming that the observations are distributed as a t-student (a normal distribution with potentially longer tales).

Assumption 2: the observations from each group are assumed to be i.i.d, conditioned on the inherent characterstics of two systems

Assumption 3: Could be used when the total number of instances (the denominators) is known.

Assumption 4: Could be used when (1) granularity (denomenator) is high enough to assume the variable is continuous Or (2) when the variable is inherently continous

Assumption 5: the observations expected to be follow a Negative-Binomail / Poisson distribution.

Assumption 6: the observations are expected to follow a binomial-distribution.

Assumption 7: the observations are expected to be normally-distributed.