Significance Testing in Natural and Biological Systems: A Review

Subhash Kak
Oklahoma State University, Stillwater

Abstract. It is generally accepted that the reproducibility crisis in the fields of psychology, biomedical science, social science, cognitive science, and neuroscience is in part due to the misuse of the Null Hypothesis Significance Testing (NHST), the default method of analysis for research in these fields. We review the shortcomings in the use of NHST and then go beyond these to consider additional issues. Current significance testing methods assume that the data is representative of stable characteristics associated with the population, which may not be true as in the case of seasonal variation in immune response and strength of pathogens. Indeed, many natural systems are time-varying and some are scale-free, which requires design of new methods for such cases. We also consider the problem from the perspective of information efficiency and since three-way logic is superior to two-way logic, we argue that adding a third hypothesis may be beneficial in certain applications.

Introduction

The basis of human judgment is rational thinking within the framework of chosen beliefs, and it is clear that the judgment can be no better than the beliefs. One might think that data-based decisions, as in statistical reasoning or machine intelligence, do not suffer from shortcomings related to beliefs, but, in reality, that is not so. Seemingly neutral data-driven methods that are routinely used in many fields can lead to false judgments if used inappropriately.

The dominant statistical method in biomedical, social science and psychological research is the null hypothesis significance testing (NHST), which scholars believe is partly responsible for the replication crisis of natural systems as in social science, psychology, cognitive neuroscience, and biomedical science [1][2][3], but it continues to be the default method [4][5][6].

The Open Science Collaboration (2015) replicated 100 landmark studies in the field of psychology, and in this set less than half yielded results sufficiently similar to the claims. Though 97% of the original studies produced statistically significant results, only 36% of

the replication studies did so [7]. Efforts to replicate research in other fields such as behavioral economics [8], medicine [9], genetics [10] and neuroscience [11] likewise produce poor results. A survey published in *Nature* in 2016 reported that researchers had been unable to reproduce over 70% of the findings of other scientists [12]. Just a few years' prior, researchers at the biotechnology firm Amgen reported only 11% of the pre-clinical cancer studies could be replicated [13].

It has been shown that it is surprisingly simple to increase the probability of obtaining positive results for false hypotheses, using research practices that are considered conventional. One can be easily led to confirmation of the hypothesis and “[i]n many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not.” [14]

In 2016, the American Statistical Association issued a statement of principles regarding the misuses of and misinterpretations of NHST [15], stressing that “p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone. Neither do they measure the size of an effect or the importance of a result, or provide a good measure of evidence regarding a model or hypothesis.”

But the problem is more complicated than just the use of NHST, which assumes population properties remain unchanging. In reality, diseases and physiological processes have periodicities [15], the immune function has seasonal changes [16][17], and pathogens also peak at different times of the year [18]. Furthermore, many biological and natural processes are scale-invariant [19][20][21][22] and so they do not follow the normal distribution that is implicitly assumed in many hypothesis tests which requires the use of maximum-likelihood methods and other tests [23]. This article reviews these issues that should be helpful in formulating new tests for hypothesis verification in systems with varying characteristics. It reviews the mathematical basis of the result that three-way logic is superior to two-way logic [24][25][26] and, by extension, testing three hypotheses is better than testing two, which provides the rationale for adding a third hypothesis to the experiment, which may be beneficial in certain natural systems applications quite like it is done for many engineering applications.

Background

There are complex social reasons why popular significance tests continue to be used in spite of their well-known shortcomings. Amongst these reasons is that the investigators want quick confirmation of their hypothesis using a method that is widely used, which is

consistent with the wish of the sponsors to monetize the claimed innovation as soon as possible.

Marcia Angell, who was for two decades the editor of *The New England Journal of Medicine*, said [27]: “It is simply no longer possible to believe much of the clinical research that is published, or to rely on the judgment of trusted physicians or authoritative medical guidelines.” She believes that the drug companies are mainly responsible for this situation: “Over the past two decades the pharmaceutical industry has moved very far from its original high purpose of discovering and producing useful new drugs. Now primarily a marketing machine to sell drugs of dubious benefit, this industry uses its wealth and power to co-opt every institution that might stand in its way, including the US Congress, the FDA, academic medical centers, and the medical profession itself.” [28]

More recently, Richard Horton, editor of *The Lancet*, wrote that “The case against [medical] science is straightforward: much of the scientific literature, perhaps half, may simply be untrue. Afflicted by studies with small sample sizes, tiny effects, invalid exploratory analyses, and flagrant conflicts of interest, together with an obsession for pursuing fashionable trends of dubious importance, science has taken a turn towards darkness.” [29]

It is generally agreed that the biggest reason behind irreproducible research is bad design and inherent biases. In a well-known paper, John Ioannidis claimed that most published results are false [30], and they merely present the investigator’s biases in a dressed-up manner. The way significance testing is used, it is hard to separate the good findings from the bad, which has implications for the usefulness of clinical research. “[E]xamination of the meta-analyses appearing in *Psychological Bulletin* from 1978 to 2006 shows that most employ a statistically inappropriate model for meta-analysis (the fixed effects model) and that 90% do not correct for the biasing effects of measurement error.”[31]

To deal with inherent biases, of which some researchers may not even be consciously aware, there is need for new kind of design that approaches the hypothesis from different perspectives. It has been suggested that there should be a clear identification of the underlying uncertainty model associated with the scientific study [32][33][34] [35] and some have even argued that NHST should be abandoned [36][37].

Data and hypothesis testing

A statistical hypothesis is a conjecture concerning the unknown probability distribution associated with the observed data X . The significance test is used to check the tenability

of the hypotheses. Generally, the method requires reducing the data to a single numerical statistic T whose marginal probability distribution is closely connected to a main question of interest.

Before the start of the experiment, two hypotheses are defined:

1. Null Hypothesis, H_0 : there is no significant effect
2. Alternative Hypothesis, H_a : there is some significant effect

The data is used to calculate a p-value that is then compared with the critical value α , which is set before starting the experiment. If the p-value is less than the critical value, then the effect is deemed significant and the null hypothesis is rejected. If p-value is more than the critical value, it is concluded that there is no significant effect and the null hypothesis is not rejected.

There are four possible outcomes, with two representing correct decisions and two representing errors.

	Accept the Null	Reject the Null
Null hypothesis true	Correct	Type I Error <i>False positive</i>
Null hypothesis false	Type II Error <i>False negative</i>	Correct

The likelihood that a test will be able to detect a property in the data depends on the strength of that property in the population. In a research project, the investigator does not know this strength, for the estimation of this value may itself be one of the purposes of the study. Instead, the investigator must choose the size of the sample so as to be able to detect the property in the sample. In other words, the size of the sample is often correlated with the presence of absence of the effect.

One can reduce the risk of committing a type I error by using a lower value for α . However, by doing so one will increase the probability of false negatives (a type II error). The probability of making a type II error is β , and this is related to the power of the

statistical test (power = $1 - \beta$). One can decrease the risk of a type II error by ensuring the test has enough power by ensuring that the sample size is large enough.

The working of the test is clear if the probabilities associated with the population are known (as would be the case in an engineered system, but it is very unlikely in a biological or social system). Thus in a hypothesis about the mean of a population, one can choose a sample size n and find the sample mean and calculate the sample standard deviation if the population standard deviation is unknown. The population can be normal distributed with known or unknown variance in which cases n can be small, or it may not be normal with known or unknown variance in which case n must be large (typically $n \geq 30$). Clearly, one needs a larger n if the distribution is unknown.

One speaks of three hypotheses.

- (i) $H_0 : \mu = \mu_0; H_a : \mu \neq \mu_0.$
- (ii) $H_0 : \mu = \mu_0; H_a : \mu > \mu_0.$
- (iii) $H_0 : \mu = \mu_0; H_a : \mu < \mu_0.$

If the population is normal and σ is known, one computes $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$; otherwise $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$, that represent the normalized statistic associated with the samples. If σ is not known, then one uses the sample standard deviation and this becomes a source of error if the sample size is small.

Figure 1. p-value is the area under the curve to the right of Z_c in the standard normal distribution

Suppose $Z_c = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ is the computed value of the test statistic. We use the following table to describe what is being computed:

Table. Significance testing with critical value α

H_a	$H_a : \mu \neq \mu_0$	$H_a : \mu > \mu_0$	$H_a : \mu < \mu_0$
p-value	$2 P(Z > Z_c)$	$P(Z > Z_c)$	$P(Z < Z_c)$
Decision	Reject H_0 if p-value $< \alpha$		

As example, let there be a sample size of $n=16$ to determine if the sample is representative of the larger population (H_0) which has the mean of 100. Formally, $H_0: \mu=100$, which means that we must consider the p-value to be $2 P(Z > |Z_c|)$, that is, one integrates both sides of the tail of the distribution. The larger population is known to have a standard deviation of 16. The sample has a mean of 108 and we use $\alpha = 0.05$ (which is a popular choice in many social science experiments):

$$Z_c = \frac{108 - 100}{16/\sqrt{16}} = 2$$

The p-value corresponding to $Z_c=2$ is the area under the normal distribution to the right of $Z=2$ (Figure 1). This probability of $P(Z_c \geq 2)$ may be read off from a table and it equals 0.023. Since 2×0.023 is less than $\alpha = 0.05$, the null hypothesis may be rejected. This means that the sample is not representative of the larger population.

On the other hand, if the sample size was 8, $Z_c = \frac{108-100}{16/\sqrt{8}} = 1.414$. The p-value corresponding to this is 0.078, two times which is larger than 0.05, and so the null hypothesis will be accepted this time. The likelihood of a mean of 108 in a smaller sample of 8 is more likely (within the significance level of $\alpha=0.05$) than in the larger sample of 16 for in the smaller sample a value far from the mean can change the sample mean much more easily than in a larger sample. In case, the hypothesis is defined in terms of just greater than or less than the statistic, only one side of the tail of the standard normal distribution will have to be integrated.

This example shows how the choice of the specific members of the test sample can lead to vastly varying results. In an experiment where the investigator is looking for a new

effect as represented by the experimental finding, there will be a tendency to privilege readings with the property over those that do not satisfy it. In other words, the hypothesis may be considered confirmed by the investigator's bias in favor of it and by casting out readings that go against the hypothesis.

If the statistics are not normal, then one needs a larger sample. For unknown distributions, one can use tight bounds such as the Hoeffding's inequality or variations thereof that provides an upper bound on the probability that the sum of bounded independent random variables deviates from its expected value by more than a certain amount [38]. There will also be more complex situations where the elements switch characteristics [39], or where the underlying phenomenon is fractal or fractal-like [40]. Such cases will require investigation within the context of the application.

The power law

If the statistics are according to a power law, which is encountered often in natural systems, then also one needs to define the hypotheses in a different way with the corresponding significance test.

Experiments have shown (e.g. [23][25]) that many phenomena follow the following power law approximately for large values of x :

$$p(x) = ax^{-\gamma} \quad (1)$$

where γ is a parameter whose value is typically in the range $2 < \gamma < 3$, and a is a constant that is needed to normalize the distribution. The main characteristic of this distribution is its heavy-tailed nature but the value for small x can have considerable variation with measurable effects (Figure 2). The distribution could be discrete or continuous.

For an experiment, the hypothesis could be whether the sample belongs to the green ($\gamma = 2.7$), the blue ($\gamma = 2.9$), or the orange ($\gamma = 2.1$) populations. In the case of a mixed population where all these samples are present in known or unknown proportions, the null hypothesis could, for example, be: Is the sample predominantly green? In this case, of course, the term "predominant" will have to be suitably defined.

For the case, where there is a x_{min} associated with the distribution, the power law may be written as below after finding out the value of a to normalize the distribution:

$$p(x) = \begin{cases} \frac{\gamma-1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\gamma}, & x \geq x_{min} \\ 0, & x < x_{min} \end{cases} \quad (2)$$

Its moment $E(X^m)$ is easily found to be:

Figure 2. Three power law distributions (not normalized) with different γ

$$E(X^m) = \frac{\gamma-1}{\gamma-1-m} x_{min}^m \quad (3)$$

It is clear that all moments for which $m \geq \gamma - 1$ diverge.

Specifically,

$$E(X) = \frac{\gamma-1}{\gamma-2} x_{min}; E(X^2) = \frac{\gamma-1}{\gamma-3} x_{min}^2 \quad (4)$$

$$VAR(X) = \frac{(\gamma-1)x_{min}^2}{(\gamma-2)^2 (\gamma-3)} \quad (5)$$

This power law has a well-defined mean only if $\gamma > 2$, and it has a finite variance only if $\gamma > 3$. This means that one cannot apply traditional statistics based on standard deviation.

Another specific distribution with the exponent γ nearly close to 1 is the Zipf's law for discrete variable k , where N is the number of elements, and k is the rank:.

$$f(k; \gamma, N) = \frac{1/k^\gamma}{\sum_1^N 1/n^\gamma} \quad (6)$$

In hypotheses related to power law one could, in principle, determine modes of behavior correlated with the exponent. For reliable estimate of the exponent of the power law distribution, maximum likelihood methods are used.

Multiple hypotheses

There is a deeper problem with hypothesis testing that requires attention. This is the question related to the nature of the population on which the null and the alternative hypotheses are being framed.

Binary hypotheses seem intuitive and natural: a person has a specific disease or does not have it. Populations are described in binary fashion in terms of gender or as adults versus children, although this can be enlarged to three or more classes. On the other hand, age is a continuous variable that can be mapped into hypotheses that are non-binary, where one might be looking for effects that are age-specific.

If one were speaking of dealing with a disease, in the general case, one can divide a sample of subjects into at least three classes:

1. Those who respond to the drug versus those who don't
2. Those who respond to placebo versus those who don't
3. Those who get well on their own versus those who don't

These classes may be seen as arising out of the mind-body connection [41][42] and the unknown aspects of the working of the immune system that lead to the placebo and nocebo effects [43][44][45].

Let us label the chosen hypotheses appropriate for the population under study as 1, 2, 3, ... d . If the hypotheses are well defined, then from considerations related to maximization of entropy, their individual probabilities should be equal to $1/d$. Therefore, the information associated with each hypothesis is $\ln d$. This argument has been generalized from the problem of dimensions where it has recently been used with surprising results [46][47][48].

Clearly, this information increases as d increases. But this increase must be balanced against the cost of the use of the larger count of the hypotheses. Information efficiency per hypothesis is:

$$E(d) = \frac{\ln d}{d} \quad (7)$$

Its maximum value is obtained by taking the derivative of $E(d)$ and equating that to zero. This yields $d_{optimum} = e = 2.71828\dots$. In the consideration of hypotheses (as against dimensions of space where noninteger dimensions are mathematically possible) we can only count in integers. In other words:

Theorem. The optimal number of hypotheses based on information efficiency considerations is 3.

Table 1 gives the value of $E(d)$ in bits (where the measure is $\frac{\log_2 d}{d}$) for d ranging from 2 to 5.

Table 1. Efficiency of hypotheses classes ranging from 2 to 5

d	2	3	4	5
$E(d)$ bits	0.500	0.528	0.500	0.465

Figure 3. Efficiency of hypothesis classes for $d= 2, 3, 4, 5$

Use of three classes, rather than two, improves the information efficiency by a value of 5.6 percent.

It must be stressed that since the result is based on information, it can only be probabilistically true. What that means is that not all experiments will be better off with three hypotheses, and it is quite possible that a specific experiment will do better with two. This means that considerations of design will have to go into deciding what is the best course of action, unless mathematical criteria can be determined that lead to this judgment.

As an aside, it is significant that a ternary classification of patients as well as processes is used in at least one medicine system (Ayurveda) [49][50][51]. A genomic correlation with the ternary classification has been demonstrated for that system [52]. The three-hypotheses approach may be seen to work in different ways. It could be yes/no/maybe or yes/no/indeterminate for a pathology or in terms of some other trichotomy.

Obviously, multiple hypotheses can present challenges of design. One will have to examine the analytical implications of modifying the regime of two testing classes to three. Various questions about this third class may be asked. What are the conditions under which the use of the third class is justified? The third class may make it easy to triangulate the study so that biases are minimized.

As an example, if seasonality of pathogens is a factor in an investigation, then in addition to groups of subjects who get the drug under study, and those that get the placebo, there could also be a third group that get neither the drug nor the placebo. This will make the seasonal relationship an important parameter related to the investigation.

In some three-hypothesis test problems, the null hypothesis may be regarding the medial result. In textile engineering, the three-hypothesis test problem could be to decide whether the difference of strength amongst two yarns is zero (the null hypothesis), positive or negative [53]. Although the null hypothesis is what one seeks here, the two alternatives provide information on the risks and costs. Sequential probability tests, based on Bayesian optimality and generalized likelihood ratios have been devised [54][55][56] that may be useful in certain applications.

Conclusions

This paper is a review of issues with significance testing for natural systems, such as those in biological and social science applications. In experiments where previously unknown effects the distribution related to the effect is unknown, the investigator's confirmation bias can easily affect null hypothesis significance testing and, therefore, its use can lead to erroneous results.

Most methods assume that the data is representative of stable characteristics associated with the population, but this may not be true as in the case of seasonal variation of pathogens or the immune response. Issues with significance testing for power law distributions were described. The paper also looked at the problem of hypothesis testing from the perspective of information efficiency and argued that at least for some problems the use of a third hypothesis may be called for since three-way logic is superior to two-way logic. How these three hypotheses may be defined would depend on the nature of the problem and would require further theoretical and experimental investigations.

References

1. Begley, C. G., and Ellis, L. M. Raise standards for preclinical cancer research. *Nature* 483, 531–533 (2012)
2. Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., et al. Estimating the reproducibility of psychological science. *Science* 349, 943. (2015)
3. Szucs, D., Ioannidis, J. P. A. When null hypothesis significance testing is unsuitable for research: a reassessment. *Frontiers in Human Neuroscience* 11, 390 (2017).
4. Bakan, D. The test of significance in psychological research. *Psychol. Bull.* 66, 423–437 (1966)
5. Edwards, A. W. F. (1972). *Likelihood: An Account of the Statistical Concept of Likelihood and Its Application to Scientific Inference*. Cambridge, UK: Cambridge University Press.
6. Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* 46, 806–834.
7. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 349 (2015)
8. Camerer, C. F., et al. Evaluating replicability of laboratory experiments in economics. *Science* 351, 1433–1436 (2016)
9. Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* 10, 712–712 (2011)
10. Munafò, M. R. Reliability and replicability of genetic association studies. *Addiction* 104, 1439–1440 (2009)

11. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376 (2013)
12. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 452–454 (26 May 2016)
13. Baker, M. and Dolgin, E. Cancer reproducibility project releases first results. *Nature* 541, 269–270 (19 January 2017)
14. Simmonss, J.P., Nelson, L.D., Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* (2011)
15. Wasserstein, R.L., Lazar, N.A. The ASA's statement on P-values: Context, process, and purpose. *Am Stat.* 70, 129–33 (2016)
16. Dopico, X., Evangelou, M., Ferreira, R. et al. Widespread seasonal gene expression reveals annual differences in human immunity and physiology. *Nat Commun* 6, 7000 (2015)
17. Nelson R.J., Demas, G.E. Seasonal changes in immune function. *Q Rev Biol.* 71, 511-48 (1996)
18. Martinez, M.E. The calendar of epidemics: Seasonal cycles of infectious diseases. *PLoS Pathog* 14(11) (2018)
19. Dowell, S.F. Seasonal variation in host susceptibility and cycles of certain infectious diseases. *Emerging Infection Diseases* 7, 369-374 (2001)
20. Moscoso del Prado Martín F. Hypothesis testing on the fractal structure of behavioral sequences: The Bayesian assessment of scaling methodology. *Psychol Methods*, 18(4), 514-34 (2013)
21. Yu, S. Kak, S., Social network dynamics: an attention economics perspective. In *Social Networks: A Framework of Computational Intelligence*, Witold Pedrycz and Shyi-Ming Chen (editors). Springer Verlag, pp. 225-258 (2014)
22. Kak, S. Power series models of self-similarity in social networks. *Information Sciences*, 376, 31-38 (2017)
23. Clauset, A., Shalizi, C.R., Newman, M.E.J., Power-law distributions in empirical data. *SIAM Review* 51, 661-703 (2009)
24. Kak, S. On ternary coding and three-valued logic. arXiv:1807.06419 (2018)
25. Kak, S. The base-e representation of numbers and the power law. *Circuits Syst. Signal Process.* 40, 490-500 (2021).
26. Kak, S. The intrinsic dimensionality of data. *Circuits Syst. Signal Process.* 40, 2599-2607 (2021); <https://doi.org/10.1007/s00034-020-01583-8>
27. Angell M. Drug Companies & Doctors: A Story of Corruption. *The New York Review of Books* magazine. [Last accessed August 5, 2015]. Available from: <http://www.nybooks.com/articles/archives/2009/jan/15/drug-companies-doctors-a-story-of-corruption/>
28. Angell, M. The truth about the drug companies. *The New York Review*. (2004)
29. Horton R. Offline: What is medicine's 5 sigma? (2015); www.thelancet.com. Available from: <http://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736%2815%2960696-1.pdf>.

30. Ionnides, J.P.A. Why most published research findings are false. *PLOS Medicine*. (2005).
31. Schmidt, F. Detecting and correcting the lies that the data tell. *Perspectives on Psychological Science* 5, 233-242 (2010)
32. Munafo, M.R. and Smith, G.D. Robust research needs many lines of evidence, *Nature* 553, 399-401 (2018)
33. Evans, S.R. Fundamentals of clinical trial design. *J Exp Stroke Transl Med*. 3(1): 19–27 (2010)
34. Smith, R.J. The continuing misuse of null hypothesis significance testing in biological anthropology. *American J. of Physical Anthropology* 166, 236 – 245 (2018)
35. McNair, L. Basket clinical trial designs. *Clinical Researcher* 34 (2020)
36. Bickel, D.R. Null hypothesis significance testing interpreted and calibrated by estimating probabilities of sign errors. *The American Statistician* 75, 104-112 (2021)
37. McShane, B.B. et al. Abandon statistical significance. *The American Statistician* 73 (2019)
38. Kak, S. Threshold detection error bounds. *Journal of the Inst. of Electronics and Telecomm. Engineers* 30, 29-35 (1984)
39. Kak, S. Reasoning in systems with elements that randomly switch characteristics. *arXiv* (2017). <https://arxiv.org/abs/1712.04909>
40. Kak, S. Fractals with optimal information dimension. *Circuits Syst. Signal Process*. 40 (2021)
41. Stetka, B. Important link between the brain and the immune system found. *Scientific American*, (2015).
42. Hohwy J. *The Predictive Mind*. Oxford University Press (2013).
43. Ongaro G, Kaptchuk T.J. Symptom perception, placebo effects, and the Bayesian brain. *Pain* 2019;160:1-4 (2019)
44. Büchel C., Geuter S., Sprenger C., Eippert F. Placebo analgesia: a predictive coding perspective. *Neuron* 81:1223-39 (2014).
45. Kaptchuk T.J. Open-Label Placebo: Reflections on a Research Agenda. *Perspect Biol Med* 61:311-34. (2018)
46. Kak, S. Information theory and dimensionality of space. *Scientific Reports* 10, 20733 (2020)
47. Kak, S. Asymptotic freedom in noninteger spaces. *Scientific Reports* 11, 1–5 (2021). <https://www.nature.com/articles/s41598-021-83002-9>
48. Kak, S. Information theory of evolutionary stages in noninteger dimensional spaces. *TechRxiv* (2021); https://www.techrxiv.org/articles/preprint/Information_Theory_of_Evolutionary_Stages_in_Noninteger_Dimensional_Spaces/13775005
49. Rastogi, S., Building bridges between Ayurveda and modern science. *Int. J. Ayurveda Res* 1, 41-46 (2010)
50. Shirolkar, A., Chakraborty, S., Mandal, M., Dabur, R. Plasma metabolomics reveal the correlation of metabolic pathways and Prakritis of humans. *Journal of Ayurveda and Integrative Medicine* 9, 113-122 (2018)
51. Rioux, J. A complex, nonlinear dynamic systems perspective on Ayurveda and Ayurvedic research. *Journal of alternative and complementary medicine* 18, 709-18 (2012)

52. Govindaraj, P., Nizamuddin, S., Sharath, A. et al. Genome-wide analysis correlates Ayurveda Prakriti. *Sci Rep* 5, 15786 (2015).
53. Anderson, S.L. A simple method of comparing the breaking loads of two yarns. *Textile Institute* 45, 472–479 (1954)
54. Payton, M.E. and Young, L.J. A sequential procedure for deciding among three hypotheses. *Sequential Analysis* 13, 277–300 (1994)
55. Draglia, V.P., Tartakovsky, A.G., and Veeravalli, V.V. Multihypothesis sequential probability ratio tests. I. Asymptotic optimality. *IEEE Transactions on Information Theory* 45, 2448–2461 (1999)
56. Dragalin, V.P., Tartakovsky, A.G., and Veeravalli, V.V. Multihypothesis sequential probability ratio tests. II. Accurate Asymptotic Expansions for the Expected Sample Size. *IEEE Transactions on Information Theory* 46, 1366–13831 (2000)