Powerful statistical tests for ordered data

Jürgen Köfinger* and Gerhard Hummer†

Ordered data abound, yet statistical tests tend to ignore this order or do not make full use of it. We present non-parametric tests which probe the sign order of the residuals comprehensively. Compared to Pearson's χ^2 test and commonly used sign-based tests, the h and (χ^2, h) tests have superior statistical power over orders of magnitudes of the numbers of data points.

Extracting information from one-dimensionally ordered data by model fitting is a fundamental and ubiquitous task in all sciences. Order in data is induced by time and frequency, space and wavelength, experimental conditions like concentration, and countless other order parameters. Good fits aim to extract the maximum amount of useful information from the data. Statistical tests ensure the goodness of fits.

Widely used statistical tests, however, are blind to order in the data and thus discard valuable information. For example, Pearson's[1] χ^2 depends only on the absolute values of the N residuals $r_i = f_i - d_i$ between model f_i and data d_i. Yet, practitioners do not take a reasonable χ^2 and the resulting high score in a goodness-of-fit test at face value. They carefully inspect the residuals visually and search for recognizable patterns. Systematic deviations or correlations indicate that one or more of the underlying assumptions are wrong.

The human intuition needed to judge the randomness of the residuals is quantified in sign-based tests. The runs test of Wald and Wolfowitz[2] uses the number of runs r as a test statistic. A run is defined by all consecutive residuals of equal sign $s_i = \mathrm{sign}(r_i)$. Schilling's test[3] uses the length $l_{\max}$ of the longest run as a test statistic and is applied in the CorMap[4] method, for example. However, the runs test and the longest run test both probe only a small part of the information contained in the ordered signs and, therefore, have lower power than Pearson's χ^2 (Supplementary Fig. 1).

Here we introduce test statistics using the full distribution of run lengths and thus assessing sign order comprehensively. We collect the lengths of runs of positive and negative signs in a single histogram h (Fig. 1 and Supplementary Fig. 2) and calculate its probability $p(h)$ analytically (Methods). Probabilities of common sign-based test statistics can be obtained by marginalization. The number of runs r normalizes h and $l_{\max}$ is the length of the longest run counted in h. The h statistic is independent of the error model, the error estimates, and the magnitudes of the residuals, and consequently of the χ^2 statistic.

For accurate errors, combinations of sign-order information and χ^2 give even more powerful tests. The weighted runs test of Beaujean and Caldwell uses the most poorly χ^2-weighted run as a statistic[5]. Here we combine the full run length histogram h with Pearson's χ^2 statistic into the comprehensive (χ^2, h) statistic, with a probability that factorizes into $p(\chi^2, h) = p(\chi^2)p(h)$ (Methods). Note that we distinguish probabilities and probability densities by the argument of $p(\cdot)$.

As in the Fisher-Irwin test[7,8], we use the probability $p(h)$ as a measure for typicality of an observed run length distribution. Typical distributions h close to the expected distribution[6] $\propto 2^{-l}$ of the run length l have high probability. Atypical or extreme distributions h with, for example, multiple long runs will have low probabilities (Fig. 1). We introduce the cumulative distribution function $\mathrm{cdf}(\mathcal{I})$ of the Shannon information[9] $\mathcal{I} = -\ln p(h)$ to calculate P-values, $P(\mathcal{I}) = 1 - \mathrm{cdf}(\mathcal{I})$ (Supplementary Fig. 3).

To combine the h and χ^2 statistics, we extend the probability-based definition of typicality to probability densities $p(x)$ of random continuous variables x. Following Jaynes[10], we define the Shannon information as $\mathcal{I}(x) = \ln m(x)/p(x)$, where $m(x)$ is the invariant measure. With this definition, both Shannon information and P-values are invariant under variable transformations. For the dimensionless χ^2 statistic, $m(\chi^2) = \mathrm{const}$.

The Shannon information distribution (SID) for the h statistic can be calculated by exact enumeration of the partitions of the integer N or estimated from randomly generated sign configurations[5]). For the (χ^2, h) statistic, we multiply $p(h)$ with exact probabilities $p(\chi^2)$. Shannon information values are additive and compiled in cumulative distribution functions (Methods and Supplementary Fig. 3).

Shifted gamma distribution with three parameters accurately describe the SIDs (Supplementary Note 1 and Supplementary Figs. 4-6). Using B-splines[11] to represent the dependence of the gamma distribution parameters on the number of data points N, we can calculate P-values for all of our test statistics from $N \approx 50$ to $N = 100000$ accurately and efficiently (Supplementary Figs. 6 and 7). The gamma distribution form of the SID is a general feature (Supplementary Note 1). We note here that it also

* Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Max-von-Laue-Straße 3, 60438 Frankfurt am Main, Germany; Author to whom correspondence should be addressed. Electronicmail: juergen.koefinger@biophys.mpg.de

† Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Max-von-Laue-Straße 3, 60438 Frankfurt am Main, Germany; Institute for Biophysics, Goethe University, 60438 Frankfurt am Main, Germany

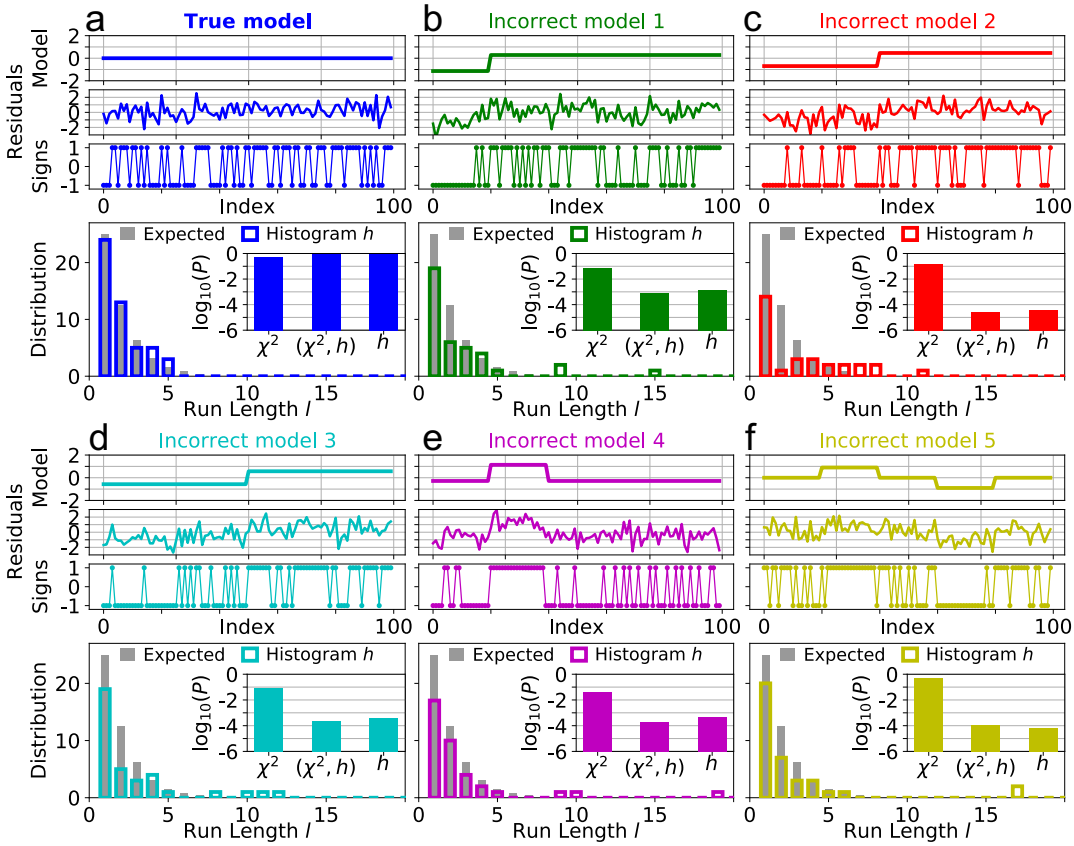


FIG. 1. **The histogram h of the run lengths of the signs of the residuals, and its combination with Pearson's χ^2, (χ^2, h), sensitively detect systematic deviations not detected by χ^2.** (a-f) Results for six different models (top panels) with $N = 100$ data points and added Gaussian noise with zero mean and unit standard deviation. (Top to bottom) Models, residuals, signs of the residuals, run length distribution, and bar plot of the P-values for the χ^2, (χ^2, h), and h tests as inset. For the true model (a), the run length multiplicities fluctuate about their expected values[6] $2^{-l-1}N/(1-2^{-N/2})$ (gray bars in bottom panels) and all P-values are large. For the incorrect models 1-5 (b-f), weight in the run length histograms is shifted to larger lengths. The resulting P-values for the h and (χ^2, h) tests are orders of magnitude lower compared to the χ^2 test.

captures $\mathrm{cdf}(-\ln P)$ defining the statistical power.

The h test detects systematic deviations masked by common tests. We generated six realizations of N=100 ordered data points according to the true model (Fig. 1a) with added uncorrelated Gaussian noise of zero mean and unit variance. The χ^2, (χ^2, h), and h tests all give a high P-value to the true model. However, for the incorrect models 1-5, the (χ^2, h) and h tests consistently outperform the χ^2 test, with P-values that are at least one order of magnitude smaller (Fig. 1b-f). h dominates in the (χ^2, h) test, with similarly low P-values for h and (χ^2, h) tests. Note that the combined (χ^2, h) test usually has a lower P-value than the individual tests for incorrect models (Fig. 2).

The h and (χ^2, h) tests outperform the χ^2 test over orders of magnitudes in the data size N (Fig. 2). We added noise with different standard errors to the five incorrect models (Fig. 1b-f) and evaluated the statistical power of these tests. Our tests consistently outperform the χ^2 test, as exemplified by the power correlations for $N = 500, 5000$, and 50000 at significance level $\alpha = 0.01$ (Fig. 2a-c). Also, the h test has superior statistical power compared to the runs test[2] (Supplementary Fig. 8) and the longest-run test[3] (Supplementary Fig. 9).

To systematically compare the data-size dependence of the statistical power, we assessed the β-risk of our tests evaluated for noise levels at which the χ^2 test loses its power. The β-risk of accepting an incorrect model, given by one minus the statistical power, is close to zero for noise levels where the test is powerful. For given significance level α, model m, and number of data points N, we determine the value of the standard deviation $\tilde{\sigma}$ such that the β-risk of the χ^2 test has a fixed value of 4α, i.e., $\beta(N;\chi^2) \equiv 1 - \mathrm{pow}(N, \chi^2, \tilde{\sigma}, m, \alpha) = 4\alpha$, where $\mathrm{pow}(\cdot)$ is the statistical power (Methods). We then evaluate the β-risk of a test T for $\tilde{\sigma}$ as $\beta(N;T) \equiv 1 - \mathrm{pow}(N, T, \tilde{\sigma}, m, \alpha)$. The β-risk ratio $\beta(N;T)/\beta(N;\chi^2) = \beta(N;T)/(4\alpha)$ is thus the factor by which the β-risk changes if we use

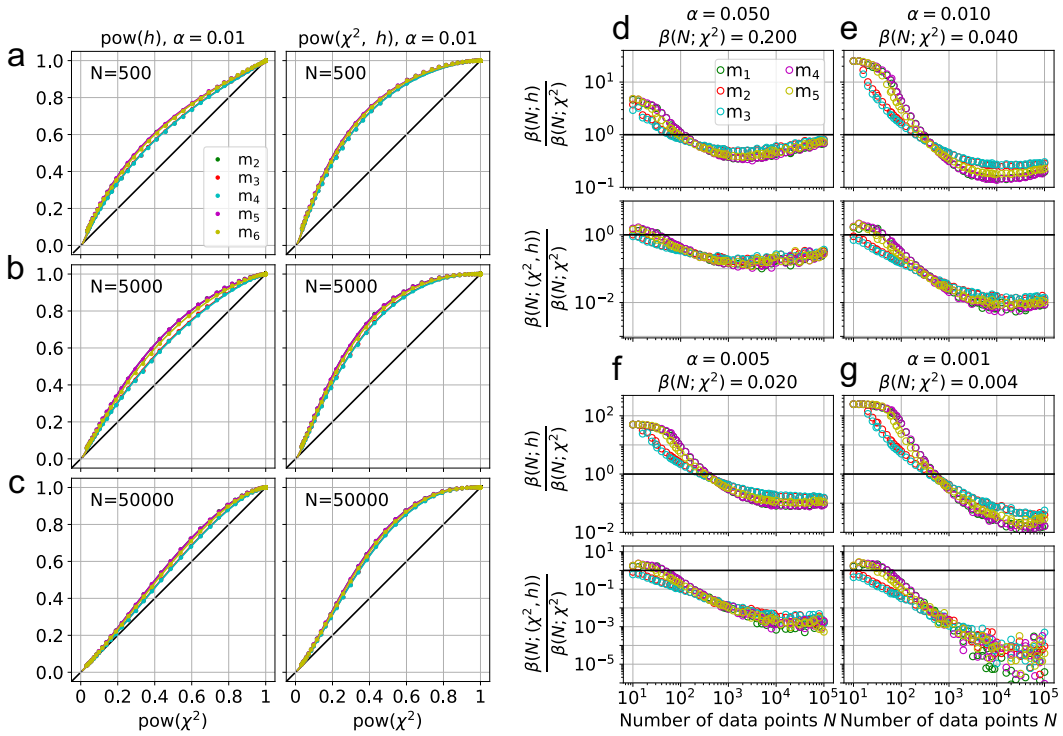


FIG. 2. **The h and (χ^2, h) tests have superior statistical power compared to the χ^2 test over orders of magnitude in the number of data points.** (a-c) Power correlation for the h statistic (left) and (χ^2, h) statistic (right) with the power of the χ^2 statistics for (a) $N = 500$, (b) $N = 5000$, (c) $N = 50000$ data points for the alternative models 1-5 shown in Fig.1f-j and a significance level $\alpha = 0.01$. Noise levels increase from right to left and top to bottom. The lines in the correlation plot are fits based on an empirical equation describing the dependence of the statistical power on the noise (Methods). (d-g) Ratio of the β-risks of the h test (top) and the (χ^2, h) test (bottom) to the β-risk of the χ^2 test as function of the number of data points. Noise levels were set by fixing the β-risk of the χ^2 test at $\beta(N; \chi^2) = 4\alpha$ for significance levels $\alpha = 0.05$ (d), 0.01 (e), 0.005 (f), 0.001 (g). Below the black horizontal lines, the β-risk is lower than that of the χ^2 test.

test T instead of the χ^2 test for noise levels where the χ^2 test has a β-risk of 4α. β-risk ratios larger/smaller than one correspond to the test T being less/more powerful than the χ^2 test.

Although the h statistic relies on the signs of the residuals only, it performs significantly better than the χ^2 statistic over a large range of data sizes (Fig. 2d-g, top panels). With increasing data size, the β-risk ratios decrease and drop to values as low as $\sim$0.4, $\sim$0.2, $\sim$0.1, $\sim$0.02 for $\alpha = 0.05, 0.01, 0.005$, and 0.001, respectively. The locations of the minima and the point of equal power, where the β-risk ratio is one, move to larger sizes for larger α-values. Uncertainties in the standard errors, which we assume here to be known exactly, will further degrade the performance of the χ^2 test but not of the h test.

The (χ^2, h) statistic has lower β-risk than the χ^2 statistic already for a few tens of data points (Fig. 2d-g, bottom panels). The data-size dependence has a shape similar to that for h but shifted to even smaller β-risk ratios. The ratios drop to values of $\sim$0.1, $\sim$0.01, $\sim 10^{-3}$, and $\sim 10^{-4}$ for $\alpha = 0.05, 0.01, 0.005$, and 0.001, respectively.

Asymmetric sign probabilities can be conveniently handled using the $h^{\pm}$ and $(\chi^2, h^{\pm})$ statistics (Methods, Supplementary Note 2, and Supplementary Fig. 10). Asymmetric sign probabilities arise, for instance, for Poisson noise with low count numbers. The $h^{\pm} = (h^+, h^-)$ statistic consists of separate histograms h^+ and h^- for runs of positive and negative signs, respectively. For symmetric probabilities, the statistical powers of $h^{\pm}$-based tests and h-based tests are comparable (Supplementary Figs. 12-15). Using $h^{\pm}$, we avoid the extra step of marginalizing $p(h^{\pm})$ to obtain $p(h)$ for asymmetric sign probabilities. When testing if two samples have been drawn from the same distribution following Wald and Wolfowitz[2], the superior h or $h^{\pm}$ tests should replace the runs test. $p(h)$ and $p(h^{\pm})$ are properly normalized and can be readily used in Bayesian inference and machine learning to avoid over- and underfitting.

We confirmed our findings and illustrate the sizeable benefit of using our tests for 353 models fitted to small-angle scattering X-ray data in SASBDB[12] (Supplementary Note 3 and Supplementary Fig. 11). About one quarter of all good fits w.r.t. χ^2 at $\alpha = 0.01$, are poor fits w.r.t. h and thus flagged for re-examination (Supplementary Fig. 12). The CorMap[4] method identifies 10%

as poor fits, indicating that it would benefit significantly from replacing the longest run test with the h test.

It is straightforward to extend the h and $h^{\pm}$ statistics to ordered data subject to correlated noise. Interpreting the signs of the residuals as spins, the uncorrelated and correlated cases correspond to one-dimensional Ising models without and with spin-spin couplings. Constant nearest-neighbor couplings correspond to the simplest case of exponentially decaying correlations. Partition functions and run length distributions h and $h^{\pm}$ can then readily be calculated using the machinery of statistical mechanics (Methods). Proceeding as before in the case of uncorrelated noise, one can then calculate P-values as measures of surprise.

An open source Python 3 implementation is available free of charge at https://github.com/bio-phys/hplusminus.

ACKNOWLEDGMENTS

We thank Drs. Ruth Pfeiffer, Mitchell H. Gail, and Roberto Covino for helpful discussions. This work was supported by the Max Planck Society.

[1] Pearson, K. On the Criterion that a given System of Deviations from the Probable in the Case of Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random sampling. *Philos. Mag.* **50**, 157–175 (1900).

[2] Wald, A. & Wolfowitz, J. On a test whether two samples are from the same population. *Ann. Math. Stat.* **11**, 147–162 (1940).

[3] Schilling, M. F. The longest run of heads. *Coll. Math. J.* **21**, 196–207 (1990).

[4] Franke, D., Jeffries, C. M. & Svergun, D. I. Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra. *Nat. Meth.* **12**, 419+ (2015).

[5] Beaujean, F. & Caldwell, A. A test statistic for weighted runs. *J. Stat. Plan. Infer.* **141**, 3437 – 3446 (2011).

[6] Denisov, S. I. & Hänggi, P. Domain statistics in a finite ising chain. *Phys. Rev. E* **71**, 046137.

[7] Fisher, R. A. The logic of inductive inference. *J. R. Stat. Soc.* **98**, 39–82 (1935).

[8] Irwin, J. O. Tests of significance for differences between percentages based on small numbers. *Metron* **12**, 83–94 (1935).

[9] Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

[10] Jaynes, E. T. *Statistical Physics*, chap. Information Theory and Statistical Mechanics, p. 181 (K. Ford (ed.), Benjamin, New York, 1963).

[11] Prautzsch, H., Boehm, W. & Paluszny, M. *Bézier and B-Spline Techniques*. Mathematics and Visualization (Springer, Berlin and Heidelberg, 2002).

[12] Kikhney, A. G., Borges, C. R., Molodenskiy, D. S., Jeffries, C. M. & Svergun, D. I. SASBDB: Towards an automatically curated and validated repository for biological scattering data. *Protein Sci.* **29**, 66–75 (2020).

METHODS

The h statistic. The histogram of run lengths for a given configuration is given by the vector of integers $h \equiv (h_1, h_2, \ldots, h_N)$, where $0 \leq h_i \leq N$ is the number of runs of length i in a sign configuration. We collect the lengths of runs of positive signs and the lengths of runs of negative signs in the single histogram h. Note that $\sum_{i=1}^{N} h_i = n$, where n is the number of runs, and that $\sum_{i=1}^{N} i h_i = N$, where N is the number of data points. This means that h is a partition of the integer N.

We assume that a configuration of N signs is generated by randomly drawing positive and negative signs with equal probability. Then, the conditional probability to observe n runs is given by

$$p(n|N) = \frac{1}{2^{N-1}} \binom{N-1}{n-1} \tag{1}$$

The binomial coefficient $\binom{N-1}{n-1}$ is the multiplicity of dividing N sites in n segments and 2^{N-1} is the normalization constant. Here and in the following we use the convention that the binomial coefficient $\binom{N}{n} = 0$ for $n < 0$ and $n > N$.

We calculate the number of possibilities to arrange n runs. If all runs were of different length then there would be $n!$ possibilities. However, we have h_i runs of length i such that we have to correct $n!$ by dividing out $h_i!$ for all i, i.e., the multiplicity of h is given by the multinomial coefficient $n!/\prod_{i=1}^{N} h_i!$. The normalization constant is given by the total number of configurations with n segments, i.e., by $\binom{N-1}{n-1}$. Consequently, the probability to observe a run length distribution h for a given number of data points N and number of runs n is given by

$$p(h|n, N) = \frac{1}{\binom{N-1}{n-1}} \frac{n!}{\prod_{i=1}^{N} h_i!} \tag{2}$$

Thus, we obtain for the joint probability $p(h) \equiv p(h, n|N) = p(h|n, N)p(n|N)$ of observing a run length distribution h with n runs of finite length

$$p(h) = \frac{1}{2^{N-1}} \frac{n!}{\prod_i h_i!} \tag{3}$$

In the Supplementary Note 2, we present a corresponding expression for $p(h)$ generalized to asymmetric probabilities for the signs.

The $h^{\pm}$ statistic. The $h^{\pm} = (h^+, h^-)$ statistic is given by the pair of run length histograms of runs with positive signs, $h^+ = (h_1^+, \ldots, h_N^+)$, and runs with negative signs, $h^- = (h_1^-, \ldots, h_N^-)$. Note that $\sum_{i=1}^{N} h_i^S = n^S$ and $\sum_{i=1}^{N} i h_i^S = N^S$, where n^S is the number of runs with sign $S = +, -$ and N^S is the number of positive $(S = +)$ and negative $(S = -)$ signs.

We consider run length distributions for positive and negative signs separately. Of n runs, n^+ runs have sign $s = +1$ and n^- runs have sign $s = -1$, i.e., $n = n^+ + n^-$. Positive and negative runs alternate such that $|n^+ - n^-| = 0$ if n is even and $|n^+ - n^-| = 1$ if n is odd. The conditional probability to observe n^+ runs of sign $+1$, and consequently $n^- = n - n^+$ runs of sign -1, is given by

$$p(n^+|n \text{ even}, N) = \delta\left(n^+, \frac{n}{2}\right) \tag{4}$$

for even n and by

$$p(n^+|n \text{ odd}, N) = \frac{1}{2}\left[\delta\left(n^+, \frac{n+1}{2}\right) + \delta\left(n^+, \frac{n-1}{2}\right)\right] \tag{5}$$

for odd n. Here, $\delta(x, y) = 1$ if $x = y$ and zero otherwise. In summary,

$$p(n^+|n, N) = \begin{cases} p(n^+|n \text{ even}, N) \text{ for } n \text{ even} \\ p(n^+|n \text{ odd}, N) \text{ for } n \text{ odd} \end{cases} \tag{6}$$

The conditional probability to observe $N^+ = N - N^-$ positive signs is determined by the product of the multiplicities of distributing N^+ positive signs on n^+ runs and N^- negative signs on n^- runs, i.e.,

$$p(N^+|n, n^+, N) = \frac{1}{Z} \binom{N^+ - 1}{n^+ - 1} \binom{N^- - 1}{n^- - 1}$$

for $n^+ > 0$ and $n^- > 0$ and by $p(N^+|n, n^+, N) = 0$ else. The normalization constant Z is given by

$$Z = \sum_{N^+ = n^+}^{N - n^-} \binom{N^+ - 1}{n^+ - 1} \binom{N^- - 1}{n^- - 1} = \left[\binom{N - n^+ - 1}{n^- - 1} \, {}_2F_1(n^+, n^+ + n^- - N; 1 + n^+ - N; 1)\right]^{-1} \tag{7}$$

where ${}_2F_1$ is the ordinary hypergeometric function.

The conditional probability to observe a run length histogram $h^S \equiv (h_1^S, h_2^S, \ldots, h_N^S)$, where $S = +, -$, is

$$p(h^S|n^S, N^S, N) = \binom{n^S}{h_1 \ldots h_{N^S}} \left[\binom{N^S - 1}{n^S - 1}\right]^{-1} \tag{8}$$

for $n^S > 0$ and by $p(h^S|n^S, N^S, N) = 1$ otherwise.

Consequently, the probability to observe the two run length histograms h^+ and h^- is given by

$$p(h^{\pm}) = \times p(h^+|n^+, N^+, N) p(h^-|n - n^+, N - N^+, N) \times p(N^+|n, n^+, N) p(n^+|n, N) p(n|N) \tag{9}$$

In the Supplementary Note 2, we present a corresponding expression for $p(h^{\pm})$ generalized to asymmetric probabilities for the signs.

Combination with χ^2 statistic. Pearson's χ^2 test is based on the probability density of χ^2 given by

$$p(\chi^2|k) = \frac{\left(\chi^2\right)^{k/2-1} e^{-\chi^2/2}}{2^{k/2} \Gamma\left(\frac{k}{2}\right)} \tag{10}$$

where k is the number of degrees of freedom. For $k \geq 2$, $p(\chi^2|k)$ has a single peak at $k-2$. We introduce $p(\chi^2) = p(\chi^2|N)$.

P-value calculation using the Shannon information distribution. We assume that our discrete statistic of state i is called x_i and has a probability p_i. For a given value x_k we now calculate the probability that a state x_j randomly sampled from the discrete distribution $\{x_k, p_k\}$ has a probability $p_j \leq p_k$, or equally that $\mathcal{I}_j \geq \mathcal{I}_k$ where $\mathcal{I}_k = -\ln p_k$ is the Shannon information[9]. The probability distribution of the Shannon information is then given by counting all states that have the same probability, i.e.,

$$p(\mathcal{I}_k) = p_k \sum_i \delta[\mathcal{I}_k - \mathcal{I}_i] \tag{11}$$

where the sum extends over all states and where $\delta[y] = 1$ if $y = 0$ and zero else. Thus, the P-value is determined by the cumulative distribution function of the Shannon information,

$$P(\mathcal{I}_k) = \mathrm{ccdf}(\mathcal{I}_k) = 1 - \mathrm{cdf}(\mathcal{I}_k) = \sum_i \Theta[\mathcal{I}_k - \mathcal{I}_i] p_i \tag{12}$$

where $\Theta[y] = 1$ if $y \geq 0$ and zero otherwise. We introduced $\mathrm{ccdf}(\mathcal{I}_k)$ for the complementary cumulative distribution function.

For example, for the probability distribution given by eq 9, we obtain

$$\mathrm{ccdf}(\mathcal{I}) = \sum_{h^\pm} p\left(h^\pm\right) \Theta\left[-\ln p\left(h^\pm | N\right) - \mathcal{I}\right] \tag{13}$$

where we sum over all run length distributions $h^\pm$ and where $\mathcal{I}$ is the Shannon information of the sample. Using that $p\left(h^\pm\right) = \frac{1}{2^N} \sum_{s_1=\pm 1} \cdots \sum_{s_N=\pm 1} \delta(h^\pm(\{s_i\}) - h^\pm)$, where we sum over all 2^N sign configurations $\{s_i\}$, the above equation can be rewritten as

$$\mathrm{ccdf}(\mathcal{I}) = \frac{1}{2^N} \sum_{s_1=\pm 1} \cdots \sum_{s_N=\pm 1} \Theta\left[-\ln p\left(h^\pm(\{s_i\})\right) - \mathcal{I}\right] \tag{14}$$

where $\Theta[y] = 1$ if $y \geq 0$ and zero else.

For continuous random variables x and constant invariant measure $m(x)$, we obtain

$$p(\mathcal{I}) = \int p(x) \delta[\mathcal{I}(x) - \mathcal{I}] \mathrm{d}x \tag{15}$$

where $\delta(\cdot)$ is Dirac's δ-function. We obtain that

$$p(\mathcal{I}) = \sum_i \frac{e^{-\mathcal{I}(x_i)}}{|\mathcal{I}'(x_i)|} \tag{16}$$

where x_i are the solutions to $\mathcal{I}(x) = -\ln p(x)$ and $\mathcal{I}'(x)$ is the first derivative of $\mathcal{I}(x)$. The P-value is given by

$$P(\mathcal{I}) = \mathrm{ccdf}(\mathcal{I}) = 1 - \mathrm{cdf}(\mathcal{I}) = \int p(x) \Theta[\mathcal{I}(x) - \mathcal{I}] \mathrm{d}x \tag{17}$$

In practice, we generate M sign configurations randomly, calculate the probabilities according to eqs 3 or 9, and rank them from smallest to largest, i.e., $(\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_M)$ with $\mathcal{I}_i \leq \mathcal{I}_{i+1}$. We then calculate the P-value for the Shannon information $\mathcal{I}$ of a sign configuration by counting all sampled values of the Shannon information $\mathcal{I}_i \geq \mathcal{I}$ and dividing by the number of samples M, i.e.,

$$P(\mathcal{I}) = \frac{k}{M} \tag{18}$$

with k given such that $\mathcal{I}_k \geq \mathcal{I}$ and $\mathcal{I}_{k+1} < \mathcal{I}$. Note that by calculating the Shannon information instead of probabilities we avoid numerical problems arising when evaluating eqs 3 and 9 directly.

For a continuous random variable and a constant invariant measure $m(x)$, we use the probability density to rank samples according to typicality and to calculate the P-value as for discrete random variables.

The SIDs of various probability distributions closely follow gamma distributions (Supplementary Note 1). The SIDs of the Gaussian distribution and the exponential distribution are given exactly by shifted gamma distributions. For the truncated Gaussian distribution we obtain a truncated and shifted gamma distribution. The SID of the χ^2 distribution can be accurately approximated by a shifted gamma distribution and this approximation becomes more accurate for increasing number of degrees of freedom (Supplementary Fig. 6)

Statistical power. The cumulative P-value distribution function defines the statistical power of a test statistic for a given model m, i.e.,

$$\mathrm{pow}(N, T, \sigma, m, \alpha) = \int_0^\alpha p(P|N, T, \sigma, m) \mathrm{d}P \tag{19}$$

where P denotes the P-value and $p(P|N, T, \sigma, m)$ is the distribution function of P-values generated by all noise realizations with standard deviation σ. The statistical power quantifies how strongly the P-value distribution function is peaked at $P = \alpha$. If m is the true model then the P-value distribution is flat and $\mathrm{pow}(N, T, \sigma, m, \alpha) = \alpha$.

Empirically, we find that the statistical power as a function of the noise is well described by

$$\mathrm{pow}(\sigma; \alpha, k_i, b_i) = \alpha^{\exp\left[-\frac{k_i}{(\sigma - b_i)^2}\right]} \tag{20}$$

for $\sigma \geq b_i$, where k_i and b_i are positive constants for a given test statistic T_i, model, and α value. For $\sigma \leq b_i$, we set $\mathrm{pow}(\sigma; \alpha, k_i, b_i) = 1$.

From eq 20, we can express σ and obtain for a given value of the statistical power $\overline{\mathrm{pow}}$

$$\sigma(\overline{\mathrm{pow}}) = \sqrt{\frac{k_i}{\ln \frac{\ln \alpha}{\ln \overline{\mathrm{pow}}}}} + b_i \tag{21}$$

We can use this expression to describe the correlation of the statistical power of a test T_j with a test T_i by

inserting eq 21 into eq 20

$$\ln \mathrm{pow}(\sigma(\overline{\mathrm{pow}}); \alpha, k_j, b_j) = \exp\left[-\frac{k_j}{(\sigma(\overline{\mathrm{pow}}) - b_j)^2}\right] \ln \alpha \quad (22)$$

Correlated noise. We now introduce sign correlations by biasing configurations represented by histograms. The histograms h and the pair of histograms $h^{\pm} = (h^+, h^-)$ are coarse-grained representations of the sign configuration. h and $h^{\pm}$ are complete descriptions of the configuration space in the sense that each sign configuration can be assigned to exactly one histogram h and exactly one pair of histograms $h^{\pm}$. A physically appealing way to introduce correlations is to interpret the signs as spins of a one-dimensional Ising model. We define energy potentials $E(h)$ and $E(h^{\pm})$ that depend on h and $h^{\pm}$, respectively. The probability to observe the histogram h then becomes

$$p(h|E) \propto p(h) \exp[-E(h)] \quad (23)$$

and analogously for $h^{\pm}$.

For nearest neighbor coupling J we have $E = -J\sum_{i=1}^{N-1} s_i s_{i+1} = -J(N - 2n + 1)$ with the number of runs given by $n = \sum_{i=1}^{N} h_i$. For this simplest correlated model, the Boltzmann distribution is given by $\exp[-E]/\cosh^{N-1}(J)$ and the sign-sign correlations decay exponentially in the thermodynamic limit, $\langle s_i s_j \rangle \propto \exp[-|i - j|/\xi]$ with $\xi = 1/\ln\tanh(J)$. We can also add a field coupling to the spins, which then represents asymmetric error distributions. With these newly defined probabilities, we can proceed as in the case for uncorrelated noise and calculate P-values as measures for typicality.

Calculation details. We define five models for the difference between the true model (Fig. 1a) and alternative models (Fig. 1b-f). For the true model, this difference is zero. The five alternative models $m_1, \ldots, m_5$ have been generated using step functions and they have been least-square fitted to the true model. Using a scale parameter $0 < a < 1$, we define steps at indices round(aN) for each data size N. Model 1, 2, and 3 have a single step at $a = 0.2, 0.4$, and 0.5, respectively. Model 4 has two steps at $a = 0.2$ and $a = 0.4$. Model 5 has 4 steps at $a = 0.2, 0.4, 0.6$, and 0.8. We have chosen the plateau values, rounded to the second digit after the comma, of the models in Fig. 1b-f as $(-0.76, 0.19)$ for model 1, $(-0.47, 0.31)$ for model 2, $(-0.38, 0.38)$ for model 3, $(-0.19, 0.76, -0.19)$ for model 4, $(0, -0.60, 0, 0.60, 0)$ for model 5.

We then resolved these models with equally spaced data points. We used two sets of numbers of data points. To highlight results for individual values of N, we chose the numbers of data points as $N = i \times 10^j$, where $i = 1, 2, \ldots 9$ and $j = 1, \ldots 4$, and $N = 10^5$. To study the size dependence of the statistical power and the β-risk we use an approximately logarithmic scale for N, i.e., $N = \mathrm{round}(10^{1+0.1i})$ with $i = 0, 1, \ldots, 40$ such that $N = 10, 13, 16, \ldots, 79433, 100000$.

To the models we added uncorrelated Gaussian noise $\mathcal{N}(0, \sigma)$ with zero mean and standard deviation σ. To simplify the scan of the noise level for the different models and data sizes, we chose the standard deviation σ by setting the variance to $\sigma^2 = s^2\sqrt{N}/\sum_{i=1}^{N} f_i^2$. f_i are the values of the model and s is a scaling factor. We set $s = s'/1.8$, where s' takes on 181 equidistant values with spacing 0.05 in the closed interval $[1, 10]$. With these definitions we make sure that we approximately cover the same range of the statistical power for the χ^2 test for all N.

For the P-value calculation, we have to calculate the distribution of the test statistics for the true model. To do so, we added normally distributed noise to the models, evaluated for all noise realization the statistics and their probabilities, and recorded them in lists. We evaluated the χ^2, h, and $h^{\pm}$ statistic for 10^6 samples. For the (χ^2, h) and $(\chi^2, h^{\pm})$ statistic, we take advantage of the independence of the χ^2 statistic and the h and $h^{\pm}$ statistic. We generated 5000 χ^2-values and evaluated their log-probabilities. We also generated 20000 sign configurations and evaluated h and $h^{\pm}$, and the respective log-probabilities for those. We then summed up the Shannon information values corresponding to the products $p(\chi^2)p(h)$ and $p(\chi^2)p(h^{\pm})$, sorted the list, and sampled only every 100th value to reduce the size. We then formed cumulative histograms of the Shannon information for each statistic (Supplementary Fig. 3 for an example with $N = 100$).

We perform least-square fits of the cumulative gamma distribution function to each of the cumulative Shannon information distribution functions and extracted the shape parameter α, the inverse scale parameter β, and the shift or location parameter $\mathcal{I}_o$. Note that we use a different font for the gamma distribution parameters α and β to distinguish them from the significance level α and the β-risk. We fitted smoothing B-splines[11] to these parameters as functions of the logarithm of the number of data points N (Supplementary Fig. 6).

To calculate the statistical power for the considered models, we generated for each model and σ-value 100000 samples. For comparison, we evaluated the P-values for these samples using the numerically estimated cumulative SID for the true model and using its gamma distribution approximation. We collected the P-values in cumulative distribution functions. We obtained the statistical power by evaluating the linearly interpolating function of the cumulative P-value distribution function for the given significance level α.

We calculated the β-risk ratio for the P-values which we obtained using the numerically estimated cumulative SID and its gamma approximation. These estimates agree excellently with each other (Supplementary Fig. 16). In Fig. 2, we show results from a power-law fit to the β-risk ratio calculated using the gamma approximation. We fitted a function $c\beta^d$ with fit parameters

c and d to the β-risk ratio as a function of the β-risk for the χ^2 test in the interval $\beta \in [10^{-3}, 10^{-1}]$. The β-risk ratios values calculated for the gamma approximation and calculated from the power-law approximation of the β-risk ratio agree accurately with each other (Supplementary Figs. 15 and 17). Importantly, we can reliably extrapolate to extremely low values of the β-risk ratio. For example, for the (χ^2, h) test and a significance level $\alpha = 0.001$, the extrapolated β-risk ratio reaches $\sim 10^{-5}$ for $N \gtrsim 10^4$. That is, the β-risk of the (χ^2, h) test is $\sim 10^{-5} \times 4\alpha = 4 \times 10^{-8}$. To estimate the β-risk ratio without extrapolation, we would have to sample significantly more than 10^9 noise realizations.

We used the Z-score approximation of the runs test of Wald and Wolfowitz and calculated two-tailed P-values for $Z = (r - \mu_\mathrm{r})/\sigma_\mathrm{r}$ where r is the number of runs and where $\mu_\mathrm{r} = 1 + (2N^+N^-)/N$ and $\sigma_\mathrm{r}^2 = (2N^+N^-(2N^+N^- - N))/(N^2(N - 1))$.

We calculated the cumulative probabilities $\mathrm{cdf}(l_\mathrm{max})$ for the longest run l_max following Ref. 3. We then calculate a two-tailed P-value by $P(l_\mathrm{max}) = 2\mathrm{cdf}(l_\mathrm{max})$ if $\mathrm{cdf}(l_\mathrm{max}) < 0.5$ and $P(l_\mathrm{max}) = 2[1 - \mathrm{cdf}(l_\mathrm{max})]$ else.

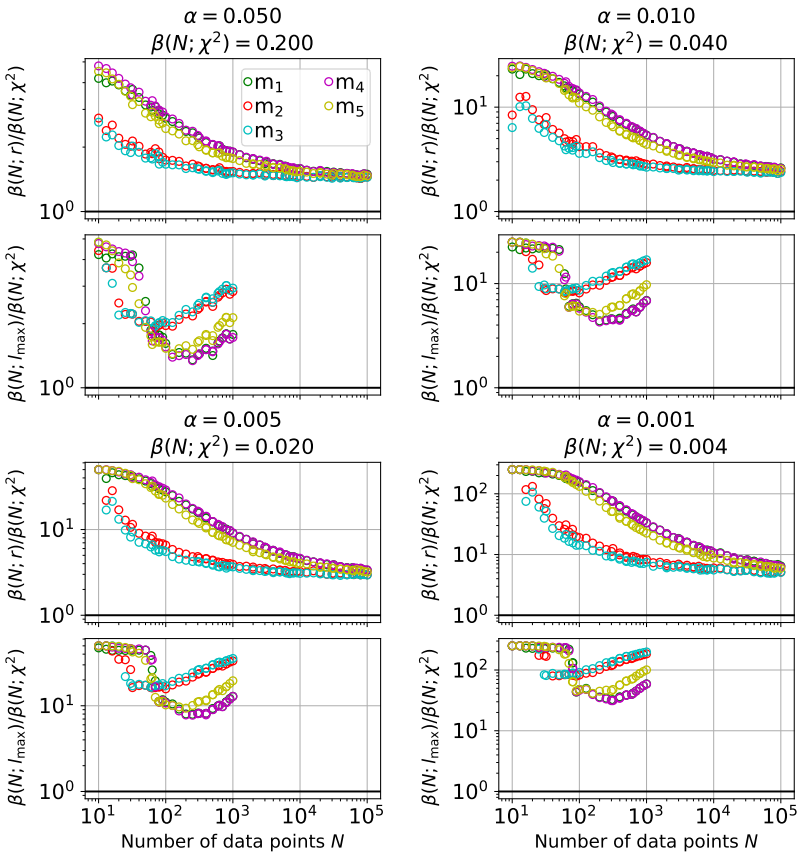
For all considered statistics, system sizes, and noise levels, and for given α-value, we fitted eq 20 to the statistical power as a function of σ using the fit paramters k_i an b_i. We use these fitted values of the parameters to reproduce the correlations of the statistical powers of two test statistics using eq 22 as shown in Fig. 2a-c.

Supplementary information: Powerful statistical tests for ordered data

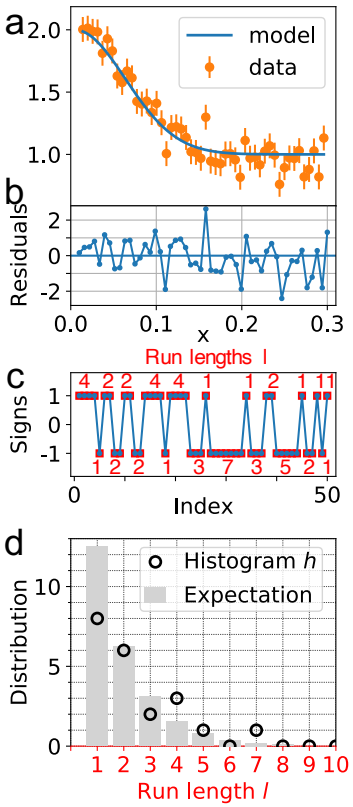
Jürgen Köfinger[1,*] and Gerhard Hummer[1,2]

[1] *Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Max-von-Laue-Straße 3, 60438 Frankfurt am Main, Germany*
[2] *Institute for Biophysics, Goethe University, 60438 Frankfurt am Main, Germany*

[*] Author to whom correspondence should be addressed. Electronic mail: juergen.koefinger@biophys.mpg.de

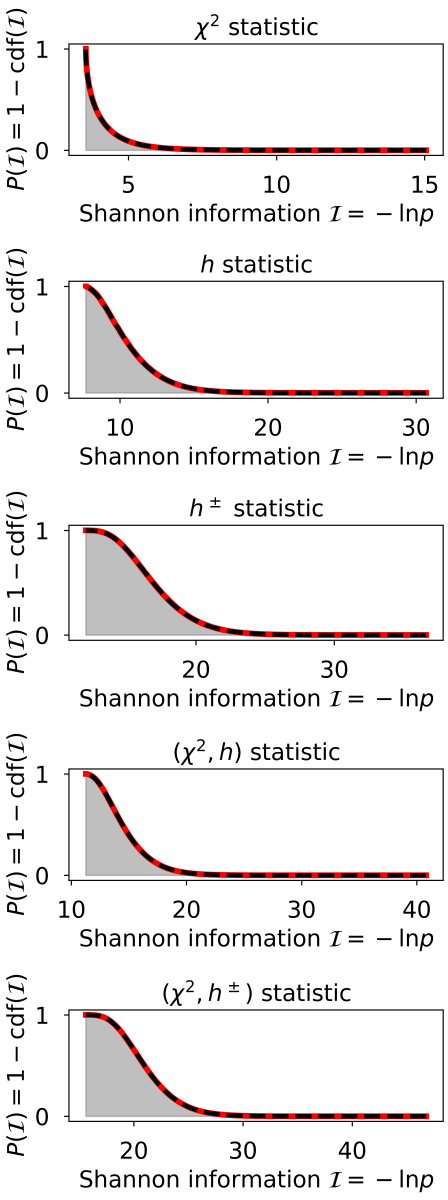


Supplementary Figure 1. **The runs test r of Wald and Wolfowitz[1] and the longest-run test $l_{\max}$ by Schilling[2] perform worse than the χ^2 test for all data sizes.** For the models shown in Fig. 1b-f of the main text (colors), we show the β-risk relative to the β-risk of the χ^2 statistic as a function of the number of data points N. We fixed the β-risk of the χ^2 test as $\beta(N;\chi^2) = 4\alpha$ for the significance level $\alpha = 0.05, 0.01, 0.005, 0.001$ and evaluated the β-risk for the r and $l_{\max}$ test statistics for the corresponding noise levels. The black horizontal lines indicate a β-risk ratio of one where both tests have equal power. In each of the four subplots, the top panel shows results for the runs test of Wald and Wolfowitz (r) and the bottom panel shows results for Schilling's longest-run test ($l_{\max}$). Note that we evaluated the longest run test of Schilling for data sizes up to $N = 1000$ only because the recursive calculation of the probability[2] fails for larger sizes.



Supplementary Figure 2. **Illustration of run-length distributions underlying the h test statistic.** (a) To assess the agreement of a model (blue line) with the data (orange symbols with errorbars indicating the standard error of the mean), we (b) plot the residuals scaled by the inverse standard error of the mean. (c) The signs of the residuals (blue disks) form runs of consecutive sign values (red bars). The red numbers next to the runs indicate their lengths. (d) The multiplicities of the run lengths l (circles) fluctuate about their expected values[3] $2^{-l-1}N/(1-2^{-N/2})$ (gray bars).

SUPPLEMENTARY NOTE 1: PROBABILITY DISTRIBUTION OF THE SHANNON INFORMATION



Supplementary Figure 3. **We use cumulative distribution functions $\mathrm{cdf}(\mathcal{I})$ of the Shannon information $\mathcal{I}$ to calculate P-values, $P(\mathcal{I}) = 1 - \mathrm{cdf}(\mathcal{I})$.** The Shannon information is given by $\mathcal{I} = -\ln p$, where p is the probability (density) of the test statistic. Top to bottom, we show results for the χ^2, h, $h^\pm$, (χ^2, h) and $(\chi^2, h^\pm)$ statistics for $N = 100$ data points. We randomly generated sign configurations and collected the corresponding values of the Shannon information $\mathcal{I}$ in complementary cumulative distributions functions $1-\mathrm{cdf}(\mathcal{I})$ (red solid lines). These functions closely follow complementary cumulative shifted gamma distributions, which have been least-square fitted to the numerical data (black dashed line). For the values of the fit parameters see Supplementary Fig. 6.

We can calculate the Shannon information distribution $p(\mathcal{I})$ of a probability distribution $p(x)$ of a continuous scalar variable x directly via

$$p(\mathcal{I}) = \sum_i \frac{p^2(x_i)}{|p'(x_i)|} = \sum_i \frac{e^{-\mathcal{I}(x_i)}}{|\mathcal{I}'(x_i)|} \tag{1}$$

where $p'(x)$ and $\mathcal{I}'(x)$ are the first derivative of the probability density $p(x)$ and the Shannon information $\mathcal{I}(x)$ with respect to x, respectively, and where the x_i are solutions to

$$\mathcal{I} = -\ln p(x) \tag{2}$$

The Shannon information distribution follows a gamma distribution in many important cases. In the following four examples, $p(\mathcal{I})$ is either exactly given by, proportional to, or approximated by a shifted gamma distribution,

$$\gamma(\mathcal{I} - \mathcal{I}_o, \alpha, \beta) = \frac{\beta^{-\alpha}(\mathcal{I} - \mathcal{I}_o)^{\alpha-1} e^{-\beta(\mathcal{I}-\mathcal{I}_o)}}{\Gamma(\alpha)} \tag{3}$$

α is the shape parameter and β the inverse scale parameter. The shift parameter $\mathcal{I}_o$ is determined by the largest value of the probability density, $p_{\max}$, as $\mathcal{I}_o = -\ln p_{\max}$.

Exponential distribution. For the exponential distribution

$$p(x) = k e^{-kx} \tag{4}$$

we obtain

$$p(\mathcal{I}) = \frac{e^{-\mathcal{I}}}{k} = e^{-(\mathcal{I}+\ln k)} = \gamma(\mathcal{I} - \mathcal{I}_o; 1, 1) \tag{5}$$

where $\mathcal{I}_o = -\ln k$ and $\mathcal{I}_o \le \mathcal{I} \le -\infty$.

Gaussian distribution. For the Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6}$$

we obtain

$$p(\mathcal{I}) = \gamma\left(\mathcal{I} - \mathcal{I}_o; \frac{1}{2}, 1\right) \tag{7}$$

where $\mathcal{I}_o = \ln\sqrt{2\pi\sigma^2}$.

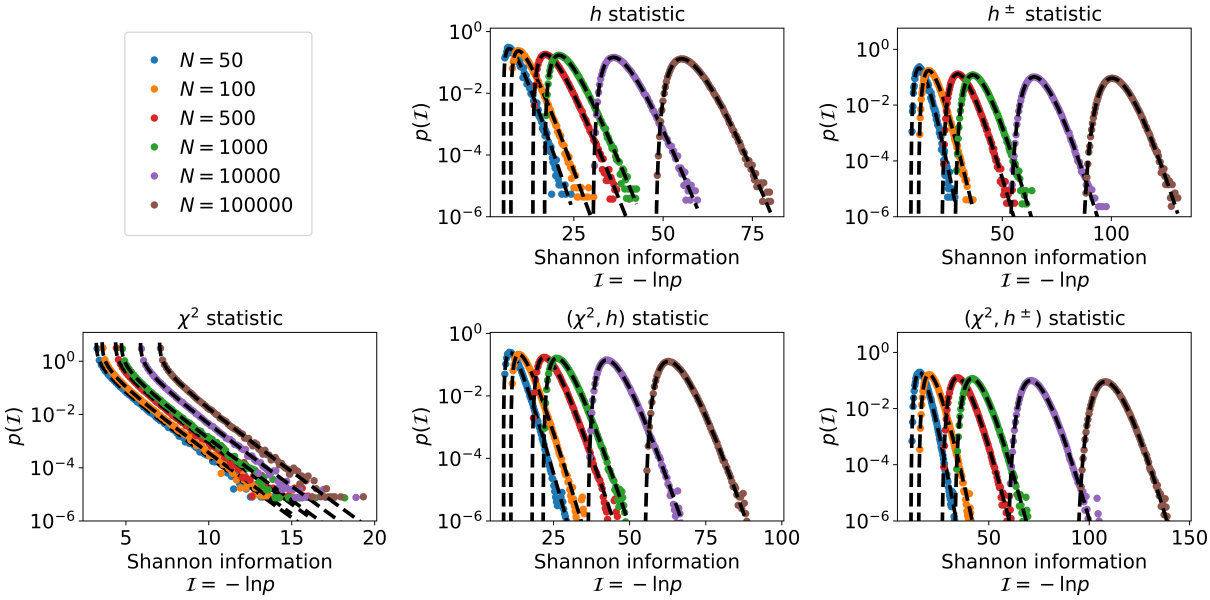
Truncated Gaussian distribution. For the truncated Gaussian distribution given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}\,\mathrm{erf}\left(\frac{w}{\sqrt{2\sigma^2}}\right)} e^{-\frac{x^2}{2\sigma^2}} \tag{8}$$

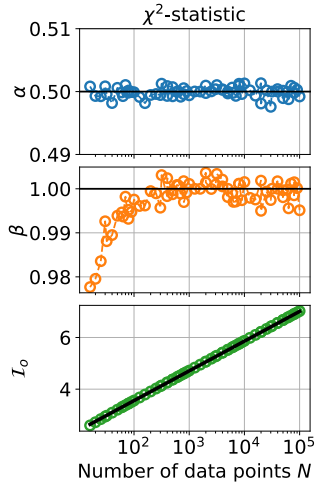
for $-w < x < w$ and zero otherwise, we obtain that the probability distribution of the Shannon information is proportional to a shifted gamma distribution, i.e.,

$$p(\mathcal{I}) = \mathrm{erf}^{-1}\left(\frac{w}{\sqrt{2\sigma^2}}\right) \gamma\left(\mathcal{I} - \mathcal{I}_o; \frac{1}{2}, 1\right) \tag{9}$$

where $\mathcal{I}_o = \ln\sqrt{2\pi\sigma^2}/\mathrm{erf}\left(\frac{w}{\sqrt{2\sigma^2}}\right)$ and $\mathcal{I}_o \le \mathcal{I} \le -\ln p(w)$.



Supplementary Figure 4. **The Shannon information distribution functions closely follow gamma distributions.** We plot on a semi-log scale numerical results for the Shannon information distributions functions $p(\mathcal{I})$ for the χ^2, h, $h^{\pm}$, (χ^2, h), and $(\chi^2, h^{\pm})$ statistics for $N = 50$ (blue), $N = 100$ (orange), $N = 500$ (red), $N = 1000$ (green), $N = 10000$ (magenta), $N = 100000$ (brown) data points. Least-square fits of the gamma distribution (black dashed lines) agree excellently with the numerical results. For the values of the fit parameters see Supplementary Figs. 5 and 6.



Supplementary Figure 5. **The Shannon information distribution function of the χ^2 statistic can be summarized by only three parameters for given N.** The shifted gamma distributions $\gamma(\mathcal{I} - \mathcal{I}_o; \alpha, \beta)$ (eq 3) have been least-square fitted to the numerically determined Shannon information distributions $p(\mathcal{I})$ using $\mathcal{I}_o$, α, and β as fit parameters. We show as black lines the parameters for the approximate expression given by eq 12 with $\alpha = 1/2$, $\beta = 1$, and $\mathcal{I}_o = -\ln p(N-2|N)$, where $p(N-2|N)$ is the χ^2 distribution given by eq 10 of the Methods evaluated at its maximum.

Multivariate normal distribution. The multivariate normal distribution in n dimensions is given by

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2} \tag{10}$$

Its maximum defines $\mathcal{I}_o = \frac{n}{2}\ln(2\pi)$. Using spherical coordinates in n dimensions and that the surface of an n-dimensional sphere is given by $S = 2\pi^{\frac{n}{2}} R^{n-1}/\Gamma\left(\frac{n}{2}\right)$, with $R^2 = \sum_{i=1}^{n} x_i^2$, we obtain

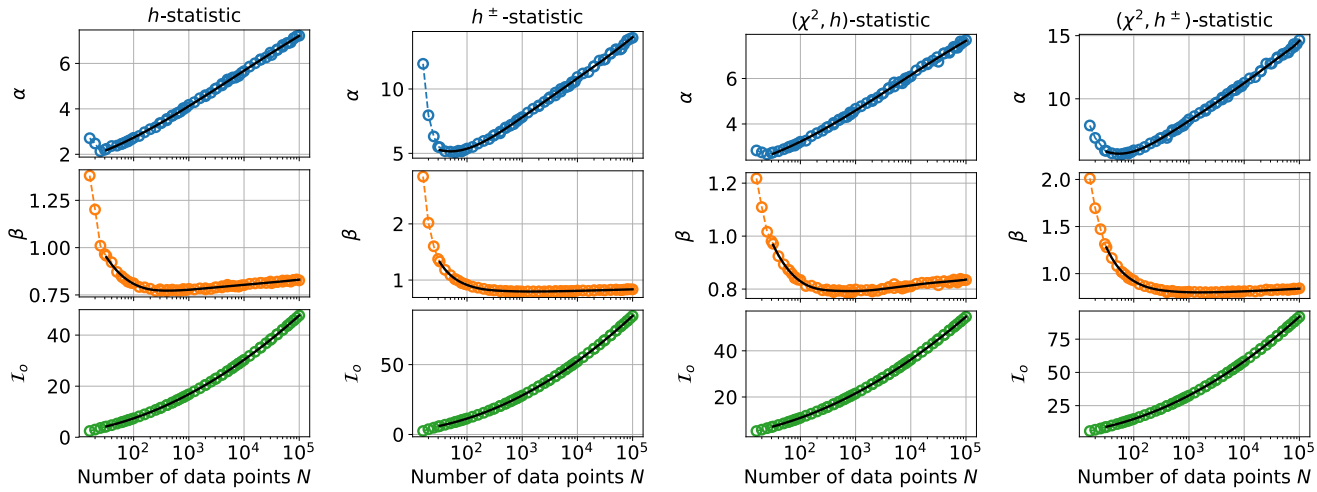
$$p(\mathcal{I}) = \gamma\left(\mathcal{I} - \mathcal{I}_o; \frac{n}{2}, 1\right) \tag{11}$$

χ^2-distribution. As we show next, the Shannon information distribution function of the χ^2-distribution $p(\chi^2|N)$, defined in eq 10 of the Methods, closely follows a gamma distribution even for just a few degrees of freedom. That is,

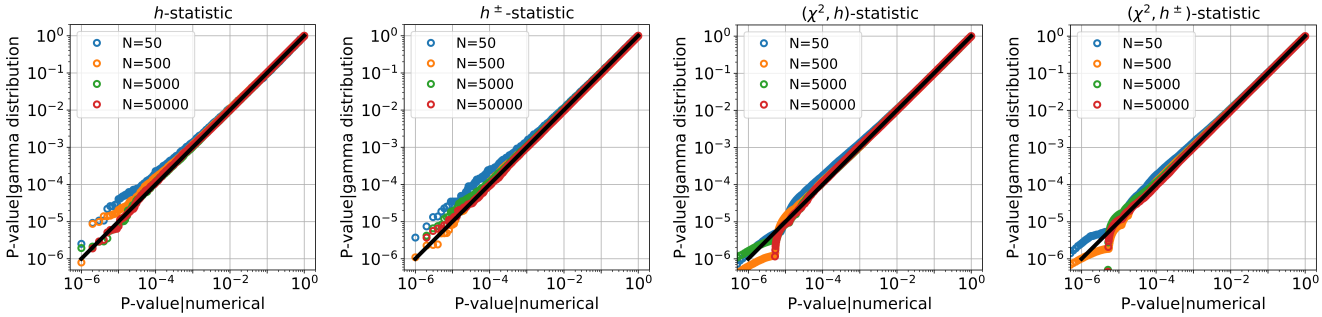
$$p(\mathcal{I}) \approx \gamma\left(\mathcal{I} - \mathcal{I}_o; \frac{1}{2}, 1\right) \tag{12}$$

with $\mathcal{I}_o = -\ln p(N-2; N)$ and $\mathcal{I}_o < \mathcal{I} < \infty$. The values of the shape parameter α and the inverse scale parameter β are identical to the values for a Gaussian distribution. The reason is that the χ^2 distribution approaches the Gaussian distribution for increasing number of degrees of freedom. Consequently, this approximation becomes more accurate for larger N.

In the following we derive the gamma distribution approximation of the Shannon information distribution for



Supplementary Figure 6. **The Shannon information distribution function of a statistic is summarized by only three parameters for given N.** The shifted gamma distributions $\gamma(\mathcal{I}-\mathcal{I}_o,\alpha,\beta)$ (eq 3) have been least-square fitted to the numerically determined Shannon information distributions $p(\mathcal{I})$ using the location parameter $\mathcal{I}_o$, the shape parameter α, and the inverse scale parameter β as fit parameters. Results for the h, $h^{\pm}$, (χ^2,h), and $(\chi^2,h^{\pm})$ statistics (left to right) show a smooth dependence on the data size for all statistics. The black lines show smoothing B-spline fits to the parameters as a function of $\log_{10}(N)$ in the range from $N=30$ to $N=100000$. With this B-spline representation, we can accurately and efficiently calculate P-values from $N\approx 50$ to $N=100000$ data points.



Supplementary Figure 7. **Validation of the P-value calculation using Shannon information distributions represented by gamma distributions.** We use smoothing B-spline representations for the parameters of the gamma distribution α, β, $\mathcal{I}_o$ as functions of the numbers of data points N (vertical axis; see Supplementary Fig. 6) and compare it to our numerical results generated by sampling random sign configurations (horizontal axis). Results for $N=50,500,5000$, and 50000 confirm that the B-spline representation is accurate and that we can use it to accurately calculate P-values from $N\approx 50$ to $N=100000$.

the χ^2-distribution to highlight the underlying approximations. The derivative of the Shannon information is given by

$$\mathcal{I}'(x)=\frac{2-k}{2}\frac{1}{x}+\frac{1}{2}=\frac{1}{2}\left[1-\frac{k-2}{x}\right] \tag{13}$$

The two solutions of $\mathcal{I}=-\ln p(x)$ are given by

$$x_n=(2-k)W_n\left[A(\mathcal{I})\right] \tag{14}$$

where we introduced

$$A(\mathcal{I})=\frac{\left[e^{-\mathcal{I}}2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)\right]^{\frac{2}{k-2}}}{2-k} \tag{15}$$

and where $n=0,-1$ indicate the two branches $W_n(\cdot)$ of the Lambert W function. This function, also called the product logarithm, is defined as the inverse function of $f(x)=xe^x$. We thus obtain for $x_{-1}>k-2$

$$|\mathcal{I}'(x_{-1})|=\frac{1}{2}\left[1-\frac{k-2}{x_{-1}}\right] \tag{16}$$

and for $x_0<k-2$

$$|\mathcal{I}'(x_0)|=\frac{1}{2}\left[\frac{k-2}{x_0}-1\right] \tag{17}$$

such that the Shannon information distribution can be

written as

$$p(\mathcal{I}) = 2e^{-\mathcal{I}} \left[\frac{1}{\frac{k-2}{x_0} - 1} - \frac{1}{\frac{k-2}{x_{-1}} - 1} \right] \tag{18}$$

Inserting, eq 14 into eq 18, we obtain

$$p(\mathcal{I}) = 2e^{-\mathcal{I}} \left[\frac{W_{-1}\left[A(\mathcal{I})\right]}{W_{-1}\left[A(\mathcal{I})\right] + 1} - \frac{W_0\left[A(\mathcal{I})\right]}{W_0\left[A(\mathcal{I})\right] + 1} \right] \tag{19}$$

Using that the derivative of the Lambert W function is given by

$$W_n'(x) = \frac{W_n(x)}{W_n(x) + 1} \tag{20}$$

we obtain

$$p(\mathcal{I}) = 2e^{-\mathcal{I}} A(\mathcal{I}) \left[W_{-1}'\left[A(\mathcal{I})\right] - W_0'\left[A(\mathcal{I})\right] \right] \tag{21}$$

Differentiating the Taylor series of $W_{-1}(y) - W_0(y)$ at $y = -1/e$, we obtain

$$W_{-1}'(y) - W_0'(y) = -\frac{\sqrt{2e}}{\sqrt{\frac{1}{e} + y}} - \frac{11e^{3/2}\sqrt{\frac{1}{e} + y}}{6\sqrt{2}} - \frac{769e^{5/2}\left(\frac{1}{e} + y\right)^{3/2}}{432\sqrt{2}} + \dots \tag{22}$$

We multiply this expression by y and use only the first term, which is given by

$$-\frac{\sqrt{2e}y}{\sqrt{\frac{1}{e} + y}} = \frac{\sqrt{\frac{2}{e}}}{\sqrt{\frac{1}{e} + y}} - \sqrt{2e}\sqrt{\frac{1}{e} + y} \tag{23}$$

For $y \approx -1/e$, we can neglect the second term and we obtain

$$y\left[W_{-1}'(y) - W_0'(y)\right] \approx \sqrt{\frac{2}{e}} \left(y + \frac{1}{e}\right)^{-\frac{1}{2}} \tag{24}$$

Next, we evaluate this expression for $y = A(\mathcal{I})$. The Taylor series expansion of $\left[e^{-\mathcal{I}}\right]^{\frac{2}{k-2}}$ at $\mathcal{I}_o$ is given by

$$\left[e^{-\mathcal{I}}\right]^{\frac{2}{k-2}} = \left(e^{-\mathcal{I}_o}\right)^{\frac{2}{-2+k}} - \frac{2\left(e^{-\mathcal{I}_o}\right)^{\frac{2}{-2+k}}(\mathcal{I} - \mathcal{I}_o)}{-2+k} + \dots \tag{25}$$

We use the first two terms of this series and express $A(\mathcal{I})$ as

$$\begin{aligned} A(\mathcal{I}) &= \frac{\left[e^{-\mathcal{I}} 2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)\right]^{\frac{2}{k-2}}}{2-k} \\ &\approx \left[\left(e^{-\mathcal{I}_o}\right)^{\frac{2}{-2+k}} - \frac{2\left(e^{-\mathcal{I}_o}\right)^{\frac{2}{-2+k}}(\mathcal{I} - \mathcal{I}_o)}{-2+k}\right] \frac{\left[2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)\right]^{\frac{2}{k-2}}}{2-k} \\ &= \left[1 - \frac{2(\mathcal{I} - \mathcal{I}_o)}{-2+k}\right] \frac{\left[2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right) e^{-\mathcal{I}_o}\right]^{\frac{2}{k-2}}}{2-k} \end{aligned} \tag{26}$$

For the second factor, we obtain

$$\frac{\left[2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right) e^{-\mathcal{I}_o}\right]^{\frac{2}{k-2}}}{2-k} = -\frac{1}{e} \tag{27}$$

such that

$$A(\mathcal{I}) \approx -\frac{1}{e}\left[1 - \frac{2(\mathcal{I} - \mathcal{I}_o)}{-2+k}\right] \tag{28}$$

Thus, using this expression for $y = A(\mathcal{I})$, we obtain

$$y\left[W_{-1}'(y) - W_0'(y)\right] \approx \sqrt{\frac{1}{e}} \left(\frac{\mathcal{I} - \mathcal{I}_o}{-2+k}\right)^{-\frac{1}{2}} \tag{29}$$

In this approximation, $p(\mathcal{I})$ is proportional to $(\mathcal{I} - \mathcal{I}_o)^{-\frac{1}{2}} e^{-\mathcal{I}}$. The normalization constant is given by

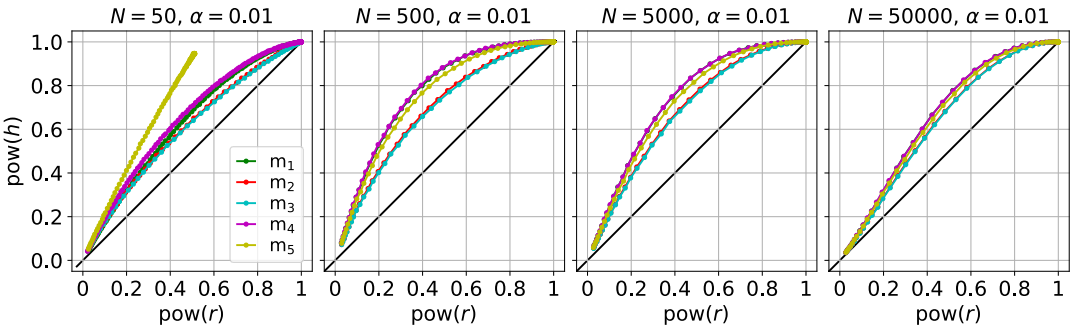
$$\int_{\mathcal{I}_o}^{\infty} (\mathcal{I} - \mathcal{I}_o)^{-\frac{1}{2}} e^{-\mathcal{I}} \mathrm{d}\mathcal{I} = e^{-\mathcal{I}_o}\sqrt{\pi} = e^{-\mathcal{I}_o}\Gamma\left(\frac{1}{2}\right) \tag{30}$$

such that we obtain

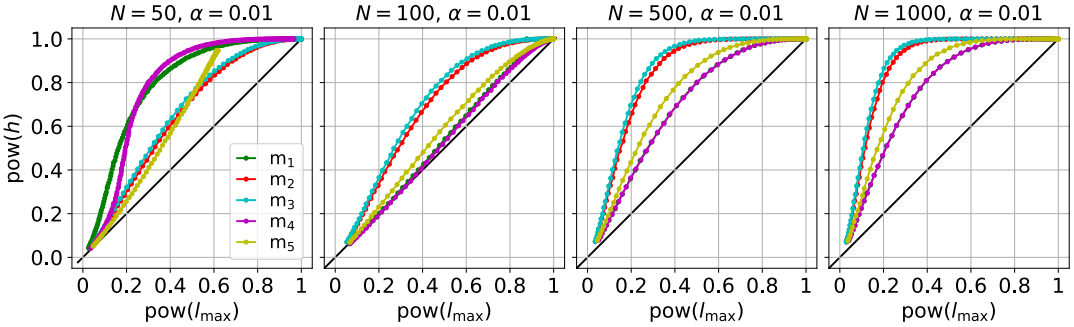
$$p(\mathcal{I}) \approx \frac{(\mathcal{I} - \mathcal{I}_o)^{-\frac{1}{2}} \exp(-(\mathcal{I} - \mathcal{I}_o))}{\Gamma\left(\frac{1}{2}\right)} \tag{31}$$

equal to eq 12.

COMPARISON OF THE h TEST TO OTHER SIGN-BASED TESTS



Supplementary Figure 8. **The h test has larger statistical power than the runs test of Wald and Wolfowitz.** Power correlation of the h test (vertical axis) with the runs test of Wald and Wolfowitz (horizontal axis, r is the number of runs) for the models (colors) shown in Fig. 1b-f of the main text and $N = 50, 500, 5000$, and 50000 data points. The significance level is $\alpha = 0.01$. Note that for model 5 (yellow) and $N = 50$ the power of the h test reaches $\sim$0.95 and the power of the runs test reaches $\sim$0.5 for the smallest errors.



Supplementary Figure 9. **The h test has larger statistical power than Schilling's longest run test.** Power correlation of the h test (vertical axis) with the longest run statistic, $l_{\max}$, (horizontal axis) for the models (colors) shown in Fig. 1b-f of the main text and $N = 50, 100, 500$ and 1000 data points. The significance level is $\alpha = 0.01$. Note that for model 5 (yellow) and $N = 50$ the power of the h test reaches $\sim$0.8 and the power of the longest run test reaches $\sim$0.6 for the smallest errors.

SUPPLEMENTARY NOTE 2: GENERALIZATION TO ASYMMETRIC SIGN PROBABILITIES

In the derivation of eq 9 of the Methods, we assumed that positive and negative signs are equally likely. Here, we assume signs have probabilities p_+ to be positive and $p_- = 1 - p_+$ to be negative. For these asymmetric probabilities, we derive the probability of observing distributions of run length of positive and negative signs corresponding to eq 9 of the Methods. We write this probability as

$$\begin{aligned} &p(h^{\pm}|p_+) = \\ &\times p(h^+|n^+, N^+, N) p(h^-|n - n^+, N - N^+, N) \\ &\times p(n^+, n^-|N^+, N) p(N^+|N, p_+) \end{aligned} \tag{32}$$

where $p(h^+|n^+, N^+, N)$ and $p(h^-|n - n^+, N - N^+, N)$ are given by eq 8. In the following, we derive expressions for $p(n^+, n^-|N)$ and $p(N^+|N)$.

The probability that we find N^+ positive signs is given by the binomial distribution as

$$p(N^+|N, p_+) = \binom{N}{N^+} p_+^{N^+} p_-^{N-N^+} . \tag{33}$$

Next we calculate the joint probability $p(n^+, n^-|N^+, N)$ to find $n_\pm$ runs with signs ± 1. To take into account that in an alternating sequence of positive and negative runs, n^+ and n^- cannot deviate from each other by more than one, we introduce $\Delta = n^+ - n^- = -1, 0, 1$. For the two ordered states with all signs positive ($N^+ = N$) or all signs negative ($N^+ = 0$), the probabilities are equal one,

$$p(n^+ = 1, \Delta = 1|N^+ = N, N) = 1 \tag{34}$$

and

$$p(n^+ = 0, \Delta = -1|N^+ = 0, N) = 1 \tag{35}$$

and zero otherwise. For $N^+ \neq 0, N$, the conditional probability is given by

$$p(n^+, \Delta|N^+, N) = \frac{1 + \delta(\Delta, 0)}{Z} \binom{N^+ - 1}{n^+ - 1} \binom{N^- - 1}{n^+ - \Delta - 1} \tag{36}$$

For $\Delta = 0$, $1 + \delta(\Delta, 0) = 2$ which accounts for the two possible arrangements of runs for $n^+ = n^-$, starting the sequence of runs either with a positive or a negative run. The normalization constant Z is given by the double sum

$$Z = \sum_{n^+ = n^+_{\min}}^{n^+_{\max}} \sum_{\Delta = \Delta_{\min}(n^+)}^{\Delta_{\max}(n^+)} [1 + \delta(\Delta, 0)] \binom{N^+ - 1}{n^+ - 1} \binom{N^- - 1}{n^+ - \Delta - 1} \tag{37}$$

where the lower limit for the number of positive runs n^+ is given by

$$n^+_{\min} = \begin{cases} 0 \text{ for } N^+ = 0, \text{ and} \\ 1 \text{ otherwise.} \end{cases} \tag{38}$$

and the upper limit by

$$n^+_{\max} = \min(N^+, N^- + 1) . \tag{39}$$

The lower limit for Δ, and therefore the upper limit for the number of negative runs $n^- = n^+ + \Delta$, is given by

$$\Delta_{\min} = \begin{cases} 1 \text{ for } n^+ = N^- + 1, \\ 0 \text{ for } n^+ = N^-, \text{ and} \\ -1 \text{ otherwise,} \end{cases} \tag{40}$$

and the upper limit by

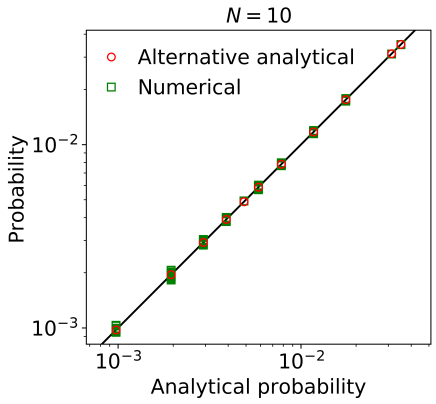
$$\Delta_{\max} = \begin{cases} -1 \text{ for } n^+ = 0, \\ 0 \text{ for } n^+ = 1, \text{ and} \\ 1 \text{ otherwise.} \end{cases} \tag{41}$$

We show numerically that for $p_+ = p_- = 1/2$, eq 32 gives the same results as eq 9 of the Methods. For $N = 10$, we generate 10^6 sign configurations, for which we evaluate $p(h^{\pm})$ given by eq 9 of the Methods and $p(h^{\pm}|p_+ = 1/2)$ given by eq 32 presented here. Additionally, we determine the probabilities of the pairs of distributions (h^+, h^-) by counting. To do so, we convert (h^+, h^-) into unique strings. We find that both expressions agree well with each other and with the numerically determined probabilities (see Supplementary Fig. 10).

We can generalize $p(h)$ given by eq 3 of the Methods to asymmetric sign probabilities by summing eq 32 over all histograms h^+ and $h^- = h - h^+$, i.e.,

$$p(h|p_+) = \sum_{h^+} p((h^+, h - h^+)|p_+) \tag{42}$$

where we use that $h^{\pm} = (h^+, h - h^+)$. The sum over all histograms for the positive signs, $\sum_{h^+}$, has to be performed under the constraints $\sum_{i=1}^N h_i^S = n^S$ and $\sum_{i=1}^N i h_i^S = N^S$ for $S = +, -$ and $h_i = h_i^+ + h_i^-$.

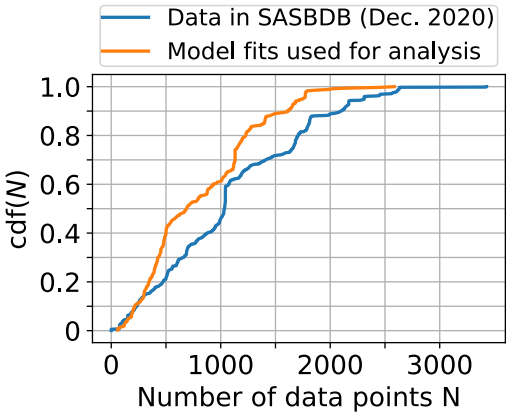


Supplementary Figure 10. **Validation of eq 9 of the Methods (bottom axis) and the alternative expression given by eq 32 (red) using numerical data for N = 10 data points.** We calculated the run length histograms h^+ and h^- for 10^6 randomly generated sign configurations. Numerically determined probability values for (h^+, h^-) (green) agree well with the analytical expressions. Note that different run length histograms can have the same multiplicities and thus the same probabilities. For the example considered here we count the 179 unique histograms $h^\pm$ expected from the theory of integer partitions and the 11 unique nonzero probability values given by sums of powers of 1/2. The smallest probability value is given by $1/2^{10} \approx 0.98 \times 10^{-3}$.

SUPPLEMENTARY NOTE 3: APPLICATION TO SASBDB

We downloaded all 1776 available data sets from the SASBDB[4] (Dec. 2020), 1221 of which contained at least one model. The total number of available models was 1781. 1510 of the corresponding model files contained four columns, i.e., the values of the momentum transfer q, the experimental intensity, the errors, and the model itself. From these, we removed stretches at the smallest q-values where the measured intensity has been set to a constant value due to the extrapolation of the model intensity to $q = 0$. For the 1510 models, we recalculated the values of χ^2 and compared them to the corresponding values in the database. In all cases we assumed that the number of degrees of freedom is given by the number of data points. We found that for 1265 models belonging to 1069 experimental scattering intensities, the recalculated χ^2 value was within $\pm 1\%$ of the database value. We discarded all other models.

Many of the remaining models were extremely poor fits, which we discarded. Not all models is SASBDB are meant to be accurate. Instead, they are used to illustrate that a given model does not fit the data at all or to interpret the intensity using simple geometrical bodies. Thus, we analyze all 353 models for which the P-values for both the χ^2 and h test were above 10^{-6}.



Supplementary Figure 11. **The cumulative distribution function of the data sizes.** We downloaded all 1221 data sets from SASBDB (blue). We selected all 353 model fits with P-values larger than 10^{-6} for both the χ^2 and h tests for further analysis (orange).

Our results for the synthetic data in the main text show that the h- and $h^{\pm}$-based tests are powerful given the data size distributions of SASBDB (Supplementary Fig. 11). 80% of the data sets have more than 500 data points and the median is at $N \approx 1000$. For the 353 models we used for the following analysis, 60% of the models have more than 500 data points. The median of the size distribution is at $N \approx 700$. That is, the size distribution for the full SASBDB is shifted to larger values where our sign-based tests become even more powerful.

We calculated the P-values for all models for the χ^2, h, $h^{\pm}$, (χ^2, h), and $(\chi^2, h^{\pm})$ statistics using the gamma distribution approximation of the Shannon information distribution. Additionally, we compare with the CorMap test[5], for which we take the P-values directly from the data base.

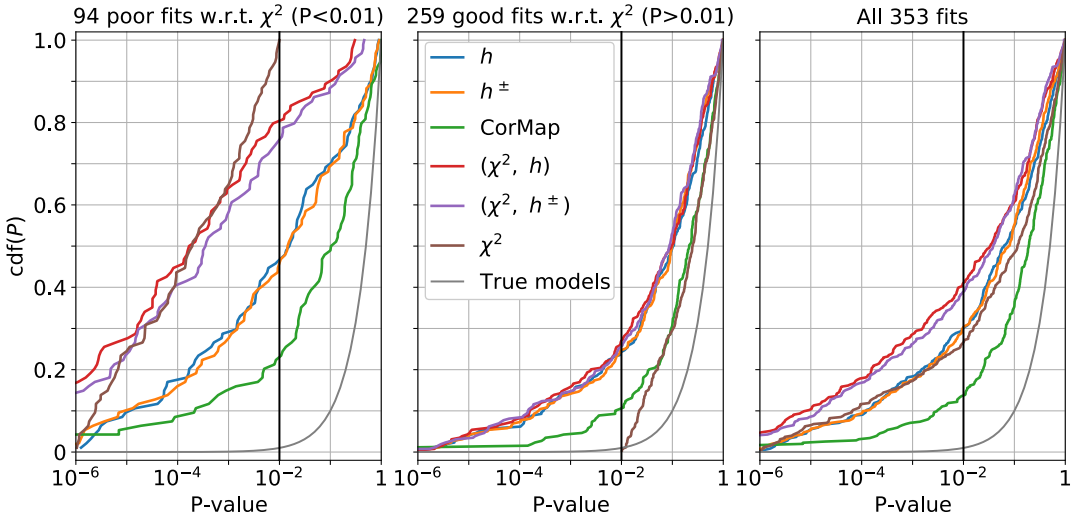
We analyzed the statistical power of the various tests for three different sets of fitted models: (1) Poor fits w.r.t. χ^2 for which $P(\chi^2) < 0.01$, (2) good fits w.r.t. χ^2 for which $P(\chi^2) > 0.01$, and (3) poor and good fits together (Supplementary Fig. 12). For all sets we find that the CDFs of the P-values of the h and $h^{\pm}$ statistics are highly correlated. The CDFs of (χ^2, h) and $(\chi^2, h^{\pm})$ statistics show a similarly strong correlation. Given a test statistic, we refer to fits with $P < 0.01$ as 'poor fits' and fits with $P > 0.01$ as 'good fits' in the following.

For the poor fits w.r.t. χ^2, we find that $\sim$20% are good fits w.r.t. (χ^2, h) and $(\chi^2, h^{\pm})$ (Supplementary Fig. 12, left panel). These fits have poor but not terribly poor χ^2-values and do not show systematic deviations or correlations according to the sign-based tests. Likely reasons are overfitting or that the models are good, but the error estimates are a bit off. About 55% of the poor fits w.r.t. χ^2 are good fits w.r.t. h and $h^{\pm}$. That is, the corresponding signs of the residuals pass this test for randomness. In contrast, $\sim$80% of the poor fits pass the CorMap test for randomness. That is, the CorMap test, which applies the longest-run test[2], is less powerful.

For the good fits w.r.t. χ^2, the CDF curves of the (χ^2, h), $(\chi^2, h^{\pm})$, h and $h^{\pm}$ statistics lie on top of each other (Supplementary Fig. 12, center panel). That is, for good χ^2 values the P-values of the combined tests (χ^2, h) and $(\chi^2, h^{\pm})$ are determined by h and $h^{\pm}$, respectively. Our tests identify $\sim$25% of the good fits w.r.t. χ^2 as poor fits. That is, the corresponding residuals show correlations or systematic deviations. The CorMap test identifies only $\sim$10% of the good fits w.r.t. χ^2 as poor fits.

Considering all good and poor fits together, we find that the h and $h^{\pm}$ test perform as well as the χ^2 test (Supplementary Fig. 12, right panel). At $P = 0.01$, the sign-based tests identify about $\sim$30% as poor fits, the χ^2 test $\sim$27%. In contrast, the (χ^2, h) and $(\chi^2, h^{\pm})$ identify $\sim$40% as poor fits. In comparison, the CorMap test identifies $\sim$15% as poor fits. Thus, more than 10% of all the models do not pass the (χ^2, h) and $(\chi^2, h^{\pm})$ tests even though they pass the χ^2 test.

In summary, our results for fitted models in SASBDB confirm the superior statistical power of our sign-based tests presented in the main text. Importantly, these results show the sizable benefit of using our tests to ensure model quality.



Supplementary Figure 12. **The cumulative distribution function of the P-values of fitted models in SASBDB evaluated for various tests (colors).** The gray line represents the CDF expected for true models, $\mathrm{cdf}(P) = P$, as a reference. We discarded extremely poor models for which $P < 10^{-6}$ for both the χ^2 and h test in this analysis. The black vertical line indicates the value $P = 10^{-2}$, which we use to distinguish poor (left of the line) from good fits (right of the line). We show CDFs for 94 poor fits w.r.t. χ^2 with $P(\chi^2) < 10^{-2}$ (left panel), for 259 good fits w.r.t. χ^2 with $P(\chi^2) > 10^{-2}$ (center panel) and all 353 poor and good fits fits together (right panel).

[1] Wald, A. & Wolfowitz, J. On a test whether two samples are from the same population. *Ann. Math. Stat.* **11**, 147–162 (1940).

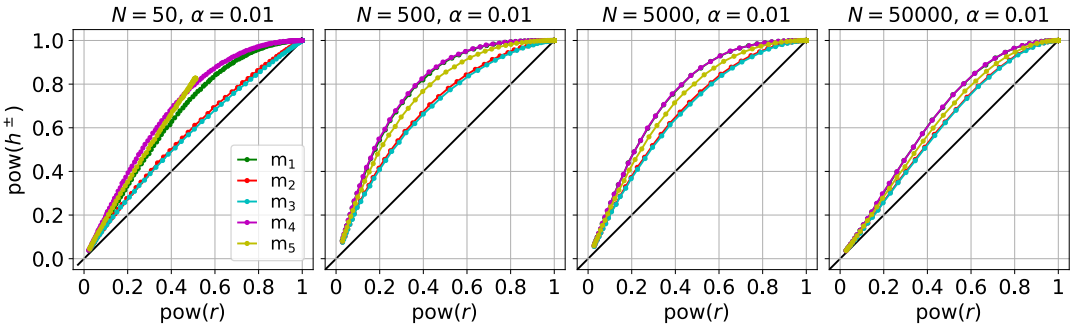
[2] Schilling, M. F. The longest run of heads. *Coll. Math. J.* **21**, 196–207 (1990).

[3] Denisov, S. I. & Hänggi, P. Domain statistics in a finite ising chain. *Phys. Rev. E* **71**, 046137.

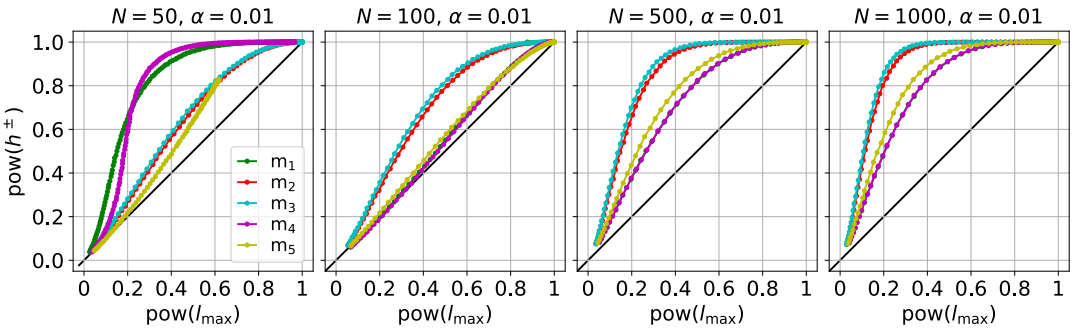
[4] Kikhney, A. G., Borges, C. R., Molodenskiy, D. S., Jeffries, C. M. & Svergun, D. I. SASBDB: Towards an automatically curated and validated repository for biological scattering data. *Protein Sci.* **29**, 66–75 (2020).

[5] Franke, D., Jeffries, C. M. & Svergun, D. I. Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra. *Nat. Meth.* **12**, 419+ (2015).

COMPARISON OF THE $h^{\pm}$ TEST TO OTHER SIGN-BASED TESTS

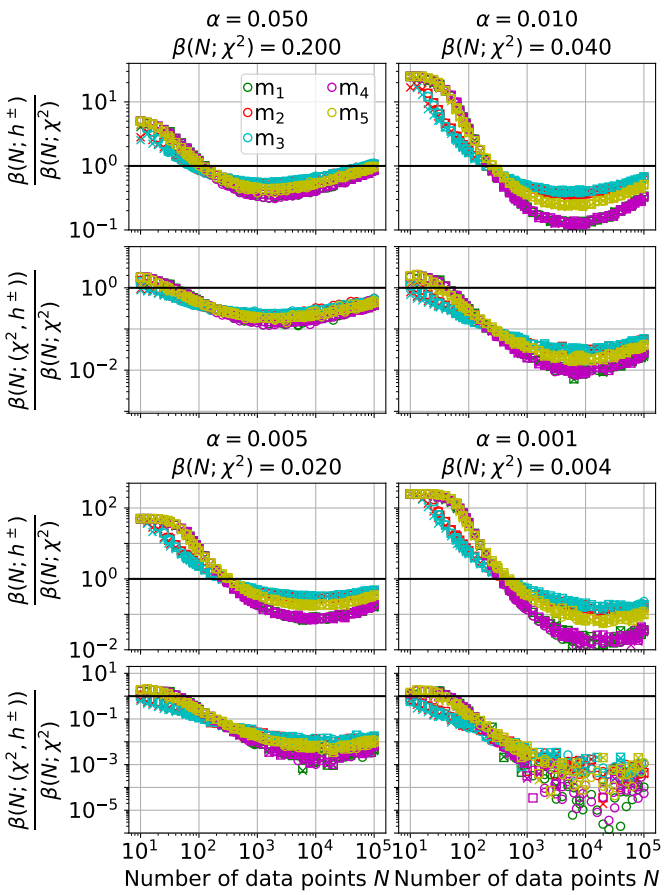


Supplementary Figure 13. **The $h^{\pm}$ test has larger statistical power than the runs test of Wald and Wolfowitz.** Power correlation of the $h^{\pm}$ test (vertical axis) with the runs test of Wald and Wolfowitz (horizontal axis, r is the number of runs) for the models (colors) shown in Fig. 1b-f of the main text and $N = 50, 500, 5000$, and 50000 data points. The significance level is $\alpha = 0.01$. Note that for model 5 (yellow) and $N = 50$ the power of the h test reaches ∼0.8 and the power of the runs test reaches ∼0.5 for the smallest errors.

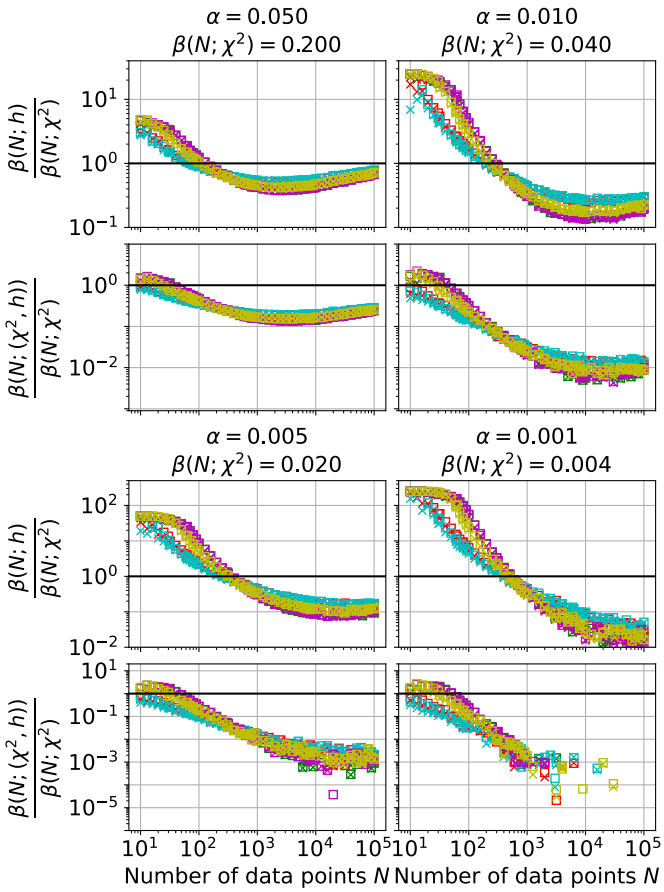


Supplementary Figure 14. **The $h^{\pm}$ test has larger statistical power than Schilling's longest run test.** Power correlation of the $h^{\pm}$ test (vertical axis) with the longest run statistic, $l_{\max}$, (horizontal axis) for the models (colors) shown in Fig. 1b-f of the main text and $N = 50, 100, 500$, and 1000 data points. The significance level is $\alpha = 0.01$. Note that for model 5 (yellow) and $N = 50$ the power of the h test reaches ∼0.8 and the power of the longest run test reaches ∼0.5 for the smallest errors.

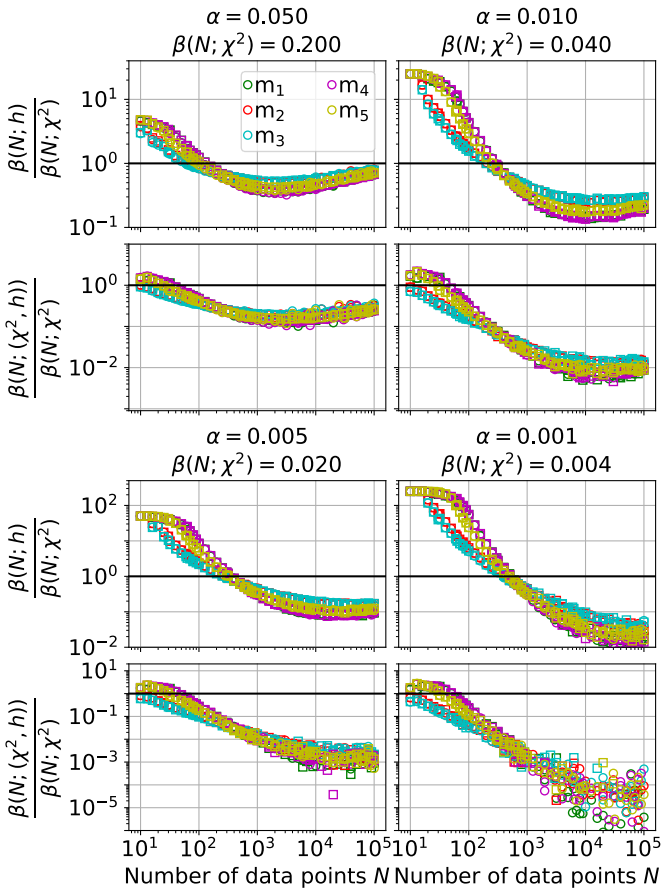
β-RISK RATIO CALCULATION AND GAMMA APPROXIMATION OF THE SHANNON INFORMATION DISTRIBUTION



Supplementary Figure 15. **β-risk ratios for the $h^{\pm}$ statistic and the $(\chi^2, h^{\pm})$ statistic.** For the models shown in Fig. 1b-f of the main text, we show the β-risk of the $h^{\pm}$ and $(\chi^2, h^{\pm})$ tests relative to the β-risk of the χ^2 statistic as a function of the number of data points for the significance levels $\alpha = 0.05, 0.01, 0.005, 0.001$. Results from the numerically determined SIDs (crosses), the gamma distribution approximation (squares), and the power-law fits of the β-risk ratio calculated using the gamma approximation (circles) agree excellently with each other. For the (χ^2, h) test at $\alpha = 0.001$ and $N \gtrsim 1000$ values of the β-risk ratio determined from the numerically determined SIDs and from their gamma distributions approximations are missing due to numerical limitations. The power-law fits can be used for extrapolation in this regime.



Supplementary Figure 16. **β-risk ratios calculated from the numerically determined SIDs (crosses) and from their gamma distribution approximations (squares) agree excellently with each other.** We show results for the models shown in Fig. 1b-f of the main text (colors). Note that due to numerical limitations, β-risk ratio values are missing for significance levels $\alpha = 0.001$ and data sizes $N \gtrsim 1000$ for the (χ^2, h) test.



Supplementary Figure 17. **β-risk ratios calculated using the gamma distribution approximation of the SID (squares) agree excellently with the results from power-law fits to the β-risk ratios (circles).** We show results for the models shown in Fig. 1b-f of the main text (colors). Note that due to numerical limitations, β-risk ratio values calculated using the gamma approximation are missing for significance levels $\alpha = 0.001$ and data sizes $N \gtrsim 1000$ for the (χ^2, h) test. The power-law fits can be used for extrapolation in this regime.