Statistical Data Set Comparison for Continuous, Dependent Data

Timothy C. Smith

David Taylor Model Basin (NSWC/CD)

ABSTRACT

Classical statistical methods exist for the comparison of the mean or variance of two sets of independent data. One of the prime requirements is that the samples be independent. This requirement is problematic for continuous random data, such as ship motion. In this case, samples are strongly dependent to samples near them but are independent only of samples far from them. The nearness or farness is determined by the autocorrelation function. This paper discusses an approach similar to classical discrete statistical comparisons but applicable to continuous, dependent data. Through use of the autocorrelation function the method determines an equivalent independent number of samples without drastically sub-sampling the data set. The method will be demonstrated using simulated roll data of the Office of Naval Research Topside Series tumblehome hull form.

KEYWORDS

Hypothesis testing; Continuous Data

INTRODUCTION

Many fields, from engineering to medicine to manufacturing, require the statistical comparison of data sets. The common question asked is if two data sets are statistically from the same population. Classical statistical methods have been developed since the 1800's to compare means and variances. These methods involve the comparison of the data sets to theoretical probability distributions to determine if the difference in data sets is within a specified statistical significance. The Student's t-test for mean values and F-test for variances are commonly used.

In order to apply these classical statistical methods, a number of assumptions must be made about the data samples. The most important are that the data are discrete samples and the samples are independent of each other. Each sample is a separate entity and is not influenced by other samples. This is often the case in material sampling, drug trials, and manufacturing. The fill level of one bottle on an assembly line does not depend on the fill level of other bottles on the assembly line.

However, not all data are discrete and independent. This is the case with ship motion data, such as roll. A roll time history is a sampling of a continuous process – roll does not change in discrete steps, but rather as a smooth continuum. Additionally, there is a memory effect due to waves radiated from the body which still continue to influence the body. This means that present motion is dependent on previous motions. The amount and duration of influence is directly calculated from the autocorrelation function.

The autocorrelation function can be used to sub-sample continuous data to generate independent data samples. The data samples are randomly selected within data windows that are separated by the time of it takes the autocorrelation function to go to zero. Depending on the process, for example parametric rolling which has a very long autocorrelation decay time, sub-sampling can greatly reduce the amount of data or require a very large amount of data. This is both inefficient and expensive.

This paper discusses an approach to apply classical statistical methods to continuous, dependent data using the autocorrelation function to "fine" or penalize the data for violation of the assumptions of discrete, independent statistics. This approach is further described in Priestley (1981). It is applicable to any continuous, dependent data set including full- and model-scale measurements, and simulation of ship motions.

To verify the applicability, we apply this approach to the comparison of model test and simulation data.

THEORY

The applicable classical statistical hypothesis tests are: the N-test, the Student's t-tests and F-tests (Bendat and Piersol 1966). The N-test uses the normal distribution to compare mean values. It is appropriate for large numbers of degrees of freedom.

The Student's t-test compares mean values using the t-distribution, which is appropriate for relatively small degrees of freedom, less than 30. The t-distribution is a normal distribution divided by a chi squared distribution. The t-distribution is defined by a variance and number of degrees of freedom. For large degrees of freedom the Student's t-test approaches the N-test.

The F-test uses the F-distribution which is the ratio of two chi squared distributions and is defined by the degrees of freedom for each chi squared distribution.

To compare ship motions, we are most interested in the variance, as the means are usually close to zero and the important part is the variation about that mean. We will use it to compare the variance of an ensemble of realizations, i.e., the variance of a test condition. The variance of the variance is also required. The calculation of the ensemble variance and variance of variance are detailed in Priestley (1981). Additionally, we will compare the data sets if all the realizations in an ensemble are concatenated together into a single long realization.

Briefly restating Priestley (1981), for an ensemble of N records, the autocorrelation function of each record time history is determined and cut at the point it begins to increase due to numerical noise. The cutting procedure uses the envelope of the autocorrelation function to determine the beginning of numerical noise and provide a smooth transition to zero. The autocorrelation function was transformed to a smooth spectrum as a check on the cutting process. The ensemble autocorrelation function is the weighted average of the autocorrelation function for each record in the ensemble. The weighting function, W_i, is the number of samples in a record divided by the total number of samples in the ensemble. The variance of the variance, VV_{irec}, are calculated for each record using Eqn. 1.

$$VV_{irec} = \frac{2}{N_{irec}} \sum_{i=-(N_{irec}-1)}^{N_{irec}-1} \left(1 - \frac{|i|}{N_{irec}}\right)\left(R_{|i|}\right)^2 \qquad (1)$$

where N_{irec} is the number of samples in a record and $R_{|i|}$ is the autocorrelation function for *each record or ensemble*.

The ensemble variance (*MVa*) and variance of the variance (*VVa*) are calculated using Eqn.

$$MVa = \sum_{i=0}^{N_{rec}-1} \left[Mv_i + (Mm_i)^2\right](W_i) - Mma^2 \qquad (2)$$

$$VVa = \sum_{i=0}^{N_{rec}-1} VV_i (W_i)^2 \qquad (3)$$

where Mv_i is the variance of each record, Mm_i is the mean value for each record, and *Mma* is the ensemble mean.

In this application of Student's t-test, the ensemble variance takes the place of the mean and the ensemble variance of variance takes the place of the variance. The number of realizations in the ensemble becomes the degrees of freedom.

Using the F-test to compare variances becomes more problematic. In the case of continuous data, where there can be hundreds and thousands of data samples, some equivalent number of independent data samples is required. If the number of data samples is used as the degrees of freedom, the F-test degenerates to comparing delta functions. While possessing a high degree of confidence, the result is not very helpful. Using sub-sampling to reduce the number of degrees of freedom results in very large variances. This is also an unhelpful result. Hence, there is the need for an equivalent number of independent degrees of freedom, which need not be an integer.

The equivalent number of independent degrees of freedom is the number of degrees of freedom for a chi squared distribution that has the same confidence interval as the data. The confidence interval of a given confidence, e.g., 95%, contains 95% of the data. For a normal distribution and 95% confidence, 95% of the data are within 1.96 standard deviations from the mean. Using the same concept, but with variance, the confidence limits for 95% are: $MVa \pm 1.96\sqrt{VVa}$ or $MVa \pm 1.69\sqrt{VVa}$ for 90% confidence. These values have already been penalized for dependent data and are equivalent independent data values.

The implementation determines, for a given number of degrees of freedom, the difference between the data confidence interval upper and lower bounds and the chi squared distribution. The equivalent degree of freedom is the number that minimizes the difference between the data upper and lower confidence interval bounds and a chi squared distribution, as seen in Fig. 1.

Numerically, this was implemented as a two step process. First, the maximum difference of the upper or lower bound for comparison was calculated for a range of degrees of freedom. Then the minimum value was found. This process was done for both data sets to be compared as the F-test requires the degrees of freedom from both sets of data.

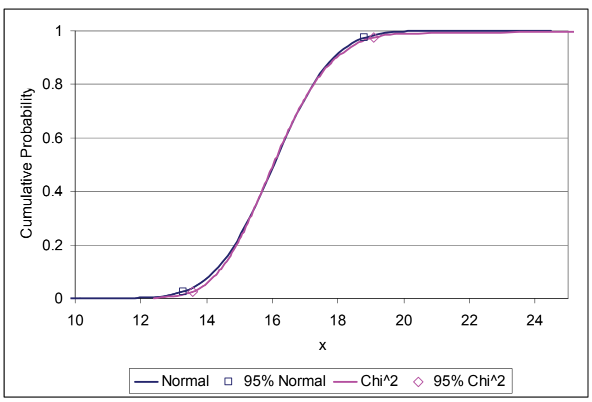

Fig. 1. Comparison of normal distribution and the equivalent chi squared distribution from matching the confidence interval.

SHIP

Simulations were made for the Office of Naval Research Topside series tumblehome hull form (Bishop 2005). This test series used the same underwater geometry with three different topsides – flared, wall-sided, and tumblehome. Experimental measurements of roll decay and regular transfer functions were made; hence, the lack of irregular seas model data for comparison. Still the hull form is useful for simulation. At this GM, the peak of the righting arm curve is 20 degrees and GZ is negative at 40.6 degrees. The principal dimensions are listed in Table 1.

The simulation realizations were made for long-crested and short-crested Sea State 8 seas. The significant wave height was 11.5m with a 13.5 sec average period for both seaways. The spectral shape was Bretschneider. The short-crested seas used five heading components with $\cos^{2n}$ spreading over ±45 degrees. Six realizations were made for each seaway. The ship speed was 16 knots and the relative wave heading was 220 degrees (40 degrees off starboard bow).

Table 1. Principal dimensions for full-scale model ONR tumblehome as simulated.

Parameter	Units	
Length between perpendiculars	m	154.00
Beam	m	18.802
Draft fwd	m	5.50
Draft aft	m	5.50
Displacement	tonnes SW	8,487.8
GM	m	1.114
Bilge keel span	m	1.250

SIMULATION PROGRAM

The ship motion simulation program used for comparison was FREDYN. FREDYN is a time domain, quasi-nonlinear, dynamic stability model that can simulate the motions of a free-running ship in a seaway under control of an autopilot. The architecture of FREDYN is based on the de Kat and Paulling model (1989) which in essence adds together the relevant force contributions in the equations of motion. The theoretical model consists of a "blended" non-linear strip theory approach, where linear radiation and diffraction potential flow, and non-linear Froude-Krylov are added to maneuvering and viscous drag forces. The radiation and diffraction forces are calculated over the calm water wetted surface. The Froude-Krylov forces have a hydrostatic component and a hydrodynamic component from the incident wave and are integrated over the instantaneous wetted surface. The maneuvering and viscous drag forces include: hull resistance; viscous damping, e.g., roll damping; rudder; skeg; propeller; and wind forces. The wave field is modeled as a summation of sine waves.

COMPARISON

A Student's t and F-test were done for the same test condition – bow seas, 16 knots, and Sea State 8. Six FREDYN realizations were made for each seaway. Each realization had the same wave spectrum but a different wave time history. The portion of the realization where the wave force ramped up was excluded from analysis. The total time of all the realizations equaled 34 minutes of full-scale time.

Using Priestley (1981), the wave elevation and roll angle ensemble variance and variance of the variance are given in Table 2 for the two seaways. Table 3 presents the variance and equivalent independent degrees of freedom for concatenated data. As a check, the ensemble and concatenated variances and variance of variances are very close to each other.

Table 2. Wave and roll statistics by record and ensemble.

	Long-crested		Short-crested	
Record	Wave Variance (m^2)	Roll Variance (deg^2)	Wave Variance (m^2)	Roll Variance (deg^2)
1	9.635	27.517	7.413	16.203
2	7.896	17.952	6.775	17.275
3	7.945	24.956	6.924	11.720
4	8.862	18.767	7.494	16.223
5	7.916	22.634	7.426	17.668
6	7.008	34.555	5.442	16.996
Avg	8.210	24.397	6.912	16.014
Ensemble				
Var	8.210	24.532	6.913	16.056
Var of Var	0.3534	6.0389	0.2622	1.9809
# records	6	6	6	6
Equiv Indep DoF	Wave	Roll	Wave	Roll
1	93	47	75	50
2	64	23	64	56
3	65	40	66	28
4	80	25	77	50
5	65	34	76	58
6	52	71	43	54
Avg DoF	69.8	40.0	66.8	49.3

Also, the sum of the individual record equivalent independent degrees of freedom is comparable to the concatenated equivalent independent degrees of freedom. With these data, we now proceed with the Student's t and F-tests taking our null hypothesis that the variances are equal for 95% confidence (α=0.05).

Without looking at the results, we can expect to accept the waves as the same significant wave heights that were specified, but not the roll as the spread waves will excite the ship from different headings. Or we can expect to reject both as spreading could affect both the waves and response. From simply looking at the results, we expect the latter due to the large differences between them.

Table 3. Wave and roll statistics for concatenation of all records.

	Long-crested		Short-crested	
	Wave Variance (m^2)	Roll Variance (deg^2)	Wave Variance (m^2)	Roll Variance (deg^2)
Var	8.205	24.517	6.908	16.046
Var of Var	0.427	6.896	0.287	1.980
Equiv Indep DoF	321	180	338	266
Time (sec)	2041	2041	2041	2041

Student's t-test

As mentioned earlier, we are substituting the variance for the mean and the variance of variance for the variance in the standard formulation. The number of degrees of freedom equals the number of records in the ensemble. Table 4 has the Student's t-test parameters calculated assuming unequal variances and unequal samples (degrees of freedom). Assuming equal number of samples is a special case of the unequal samples assumption and no error is introduced by using equal number of samples. And the test statistic is:

$$\frac{|x_1 - x_2|}{\sqrt{\frac{S_1}{n_1 - 1} + \frac{S_2}{n_2 - 1}}} \qquad (4)$$

Where x_i, S_i, and n_i are the mean, variance, and number of degrees of freedom, respectively, for data set i in the standard formulation. We are substituting variance for the mean and variance of variance for the variance.

For 95% confidence, the critical t value is 2.228 for a two-tailed test. If the test statistic is greater than the critical value, the null hypothesis is rejected and the variances cannot be considered equal. Thus, neither the waves nor roll can be considered to have the same variance.

The Student's t-test was applied to the concatenated data. Those results are in Table 5. In this case, the large number of degrees of freedom results in a normal distribution as seen by the critical t value equaling nearly 1.961, the normal distribution value for 95% confidence. Again, both the waves and roll cannot be considered to have the same variance.

Table 4 Student's t-test results for ensemble data.

	Wave Elevation		Roll Angle	
	Long	Short	Long	Short
Mean (Variance)	8.210	6.913	24.53	16.06
Variance (VoV)	0.353	0.262	6.039	1.981
Number of DoF	6	6	6	6
Pooled Variance	0.308		4.010	
alpha	0.050		0.050	
Number of DoF	10.0		10.0	
T-Critical	2.228		2.228	
Test Statistic	3.699		6.693	
	Reject		Reject	

Table 5. Student's t-test results for concatenated data.

	Wave Elevation		Roll Angle	
	Long	Short	Long	Short
Mean (Variance)	8.205	6.908	24.52	16.05
Variance(VoV)	0.427	0.287	6.896	1.980
Number of DoF	321.0	338.0	180.0	266.0
Pooled Variance	0.355		3.962	
alpha	0.050		0.050	
Number of DoF	657.0		444.0	
T-Crit	1.976		1.982	
Test Statistic	27.74		39.50	
	Reject		Reject	

F-test

The F-test statistic is the ratio of the two variances. As general practice, the ratio is formed to be greater than one, though this is not necessary. Again, if the test statistic is greater than the critical value, the null hypothesis is rejected and the variances cannot be considered equal. If the ratio is formed as less than one, then the probability is taken as one minus the probability, and now the test statistic needs to be greater than the critical value to *accept* the null hypothesis.

The F-test results for the ensemble data are in Table 6 and the concatenated data are in Table 7. The ensemble average number of equivalent independent degrees of freedom is used for the ensemble data. Using ensemble data, both the waves and roll angle can be accepted as from the same data set. This is opposite the Student's t-test results. Using concatenated data, this case accepted the waves as having the same variance and rejected roll. This is the first of our expected results.

Table 6. F-test results for ensemble data.

	Wave Elevation		Roll Angle	
	Long	Short	Long	Short
Ensemble Variance	8.210	6.913	24.53	16.06
Avg Equiv Indep DoF	69.8	49.3	40.0	49.3
alpha	0.05		0.05	
F-Crit	1.536		1.639	
Test Statistic	1.188		1.528	
	Accept		Accept	

Comparing the maximum and minimum records in each ensemble using the F-test also produced mixed results. In this case, we know the records come from the same data set and we should accept them as so. However, long-crested roll was rejected with a test statistic of 1.93 vice a critical value of 1.85. This reflects the large variance of variance seen in the long-crested roll ensemble. This is a function of non-linearity in the system increasing the uncertainty.

Table7. F-test results for concatenated data.

	Wave Elevation		Roll Angle	
	Long	Short	Long	Short
Concatenated Variance	8.205	6.908	24.52	16.05
Equiv Indep DoF	125	90	1071	462
alpha	0.05		0.05	
F-Crit	1.375		1.141	
Test Statistic	1.188		1.528	
	Accept		Reject	

CONCLUSION

The statistical comparison of data sets is a common problem. Statistical formulations since the 1800's assume discrete independent data samples. As a result, the typical statistical comparisons such as Student's t- and F-test are not strictly applicable to continuous, dependent data, e.g., ship roll motion.

This paper discussed an approach using the autocorrelation function to "fine" dependent data to calculate the variance of variance. The paper also demonstrated a confidence interval matching technique to determine the equivalent independent values for a time history of dependent data. Interestingly, the equivalent number of independent values was more than simply dividing the total time by the time for the autocorrelation to go to zero. This implies more information was retained and greater accuracy attained than simply sub-sampling the time history. Also, both the ensemble and concatenated data approach produced comparable variance, variance of the variance, and equivalent independent degrees of freedom.

This approach was applied to both the Student's t- and F-tests to compare long-crested to short-crested simulation data. Both hypothesis test results met an expected outcome, albeit different ones. The Student's t-test is somewhat preferred as it produced consistent results with both ensemble and concatenated data. The results do show the danger of blindly following the hypothesis testing results.

ACKNOWLEDGEMENT

The author is most grateful to Dr. Vadim Belenky for helpful discussions concerning this work.

REFERENCES

Bendat, J. S. and Piersol, A. G., (1966) Measurement and Analysis of Random Data, John Wiley & Sons

Bishop, R. C., Belknap, W. F., Turner, C., Simon, B. S., and Kim, J. H., (2005) Parametric Investigation on the Influence of GM, Roll Damping, and Above-Water Form on the Roll Response of Model 5613, NSWCCD-TR-50-2005/27.

de Kat, J. O. and Paulling, J. R., (1989) "The Simulation of Ship Motions and Capsizing in Severe Seas," SNAME Trans, , vol. 97.

Priestley, M. B. (1981) Spectral Analysis and Time Series Vol. 1 Univariate Series, Academic Press