Statistical Testing of RNGs

For a sequence of numbers to be considered a sequence of randomly acquired numbers, it must have two basic statistical properties:

- Uniformly distributed values
- Independence from previous values drawn

A sequence fails because patterning is present in the sequence with some degree of statistical certainty. This may manifest itself by:

- Clustering or bias occurs at some point (ie., statistically, the sequence is not uniformly distributed)
  - This includes sequences of tuples drawn from the sequence (if truly random, there should be n-space uniformity)
- Standard statistical measures are too high or too low
  - Mean value
  - Standard deviation
- Statistically significant patterns occur (indicating dependencies)
  - Alternating patterns
  - Distribution patterns (correlation)
  - Other patterning

It is always possible that some underlying patterns in a sequence will go undetected.

NIST and RNGs

In recognition of the importance of random number generators, particularly for areas such as cryptography, the NIST (http://csrc.nist.gov/rng/) has established guidelines for random number generation and testing with three primary goals:

1. to develop a battery of statistical tests to detect non-randomness in binary sequences constructed using random number generators and pseudo-random number generators utilized in cryptographic applications

2. to produce documentation and software implementation of these tests

3. to provide guidance in the use and application of these tests

Typical Tests for Randomness as Reported in 1979

(Bright, H. and R. Enison. *Computing Surveys* 11(4). 1979.
Quasi-Random Number Sequences from a Long-Period TLP Generator
with Remarks on Application to Cryptography)

1. *Equidistribution or Frequency Test.* Count the number of times a member of the sequence falls into a given interval. The number should be approximately the same for intervals of the same length if the sequence is uniform.

2. *Serial Test.* This is a higher dimensional version of the equidistribution test. Count the number of times a *k*-tuple of members of the sequence falls into a given *k*-dimensional cell. If this test is passed, the sequence is said to be *k*-distributed.
   Other tests of *k*-distributivity are:

   1. Poker test: Consider groups of *k* members of the sequence and count the number of distinct values represented in each group (e.g., a hand of 5 cards falls into one of a number of groups, such as one pair, two pairs, three of a kind, etc).

   2. Maximum (minimum) of *k:* Plot the distribution of the function max (min) [of a *k*-tuple].

   3. Sum of *k*

3. *Gap Test.* Plot the distribution of gaps in the sequence of various lengths (typically, a gap is the distance between an item and its next recurrence)

4. *Runs Test.* Plot the distribution of the runs up (monotonic increasing) and runs down (monotonic decreasing), or the distribution of the runs above the mean and those below the mean, or the distribution of the run lengths.

5. *Coupon Collector's Test.* Choose a suitably small integer *d* and divide the universe into *d* intervals. Then each member of the

sequence falls into one such interval. Plot the distribution of runs of various lengths required to have all *d* intervals represented. The idea is that you are trying to collect all of the "coupons".

6. *Permutation Test.* Study the order relations between the members of the sequence in groups of *k.* Each of the *k!* possible orders should occur about equally often. If the universe is large, the probability of equality is small; otherwise, equal members may be disregarded.

7. *Serial Correlation Test.* Compute the correlation coefficient between consecutive members of the sequence. This gives the serial correlation for lag 1. Similarly, one may get the serial correlation for lag *k* by computing the correlation coefficient between x_i and x_{i+k}. This is to show that the members of the sequence are independent. Recall: the correlation coefficient references the method of least squares for fitting a line to a set of points. The closer the correlation coefficient is to 1, the more confidence in using the regression line as a predictor. Hence, the an RNG should exhibit small serial correlation values.

Generally, the RNG is assumed to be producing probabilities; ie., numbers in the interval [0,1].

Knuth, *Seminumerical Algorithms*, (Addison-Wesley, 1997 – 3rd edition) is an excellent source for a discussion on testing random number generators.

An often cited battery of tests for random number generators is Marsaglia's 1995 Diehard collection
(http://stat.fsu.edu/pub/diehard/)
Both source and object files are provided, but at least some of the source is "an f2c conversion of patched and jumbled Fortran code," which may limit its usefulness.

Background for Hypothesis Testing

Given n observations, X_1, X_2, … , X_n (assumed to be from independent, identically distributed random variables), with (unknown) population mean μ and variance σ^2. The *sample mean*

$$\overline{X}(n) = \frac{\sum_{i=1}^{n} X_i}{n}$$

provides an (unbiased) estimator for μ (meaning that if we perform a large number of independent trials, each producing a value for $\overline{X}(n)$, the average of the $\overline{X}(n)$'s will be a close approximation to μ).

Likewise the *sample variance*:

$$S^2(n) = \frac{\sum_{i=1}^{n} \left[X_i - \overline{X}(n)\right]^2}{n-1}$$

is an estimator for σ^2 (the population variance is calculated by dividing by n; division by n-1 is used in this case because it can be proved that the sample variance is an unbiased estimator for σ^2 when dealing with independent, identically distributed random variables)

Basically, two random variables are independent if there is no relationship between them; ie., the values assumed by one random variable tell us nothing about how the values of the other one distribute and vice-versa.

Of course, using $\overline{X}(n)$ to estimate μ has little validity without knowing more information to provide an assessment of how closely $\overline{X}(n)$ approximates μ.

For the mean, it is easy to show that

- $E(cX) = cE(X)$
- $E(X_1 + X_2) = E(X_1) + E(X_2)$

For the variance,

- $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$, but
- $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$ only if the X_i's are independent

Evidently, we must examine $\mathrm{Var}(\overline{X}(n))$ to see how good an approximation to μ we can expect. Consider that

$$\mathrm{Var}(\overline{X}(n)) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)$$ ***if the X_i's are independent***.

If so, then $\mathrm{Var}(\overline{X}(n)) = \sigma^2/n$

In other words, when the X_i's are independent, it is quite clear that with larger values of n, the variance shrinks and averaging the sample means provides a good estimate for μ.

Does this matter? Law and Kelton report that simulation output data are almost always correlated; ie., the set of means acquired from multiple simulation runs lack independence. In this case, using the sample variance as an estimator may significantly underestimate the population variance, in turn leading to erroneous estimates of μ.

Covariance and Correlation as Measures of Independence Between Two Random Variables

For two random variables X_1 and X_2, the ***covariance*** is given by

$$\mathrm{Cov}(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E(X_1X_2) - \mu_1\mu_2$$

$$\boxed{E[(X_1 - \mu_1)(X_2 - \mu_2)] = E(X_1X_2) - \mu_1\underset{\mu_2}{\cancel{E(X_2)}} - \mu_2\underset{\mu_1}{\cancel{E(X_1)}} + \mu_1\mu_2 = E(X_1X_2) - \mu_1\mu_2}$$

If X_1 and X_2 are independent, then $E(X_1X_2) = E(X_1)E(X_2)$ (this can be determined from the fact that for independent events x and y, $P(xy) = P(x)P(y)$). Hence, when X_1 and X_2 are independent, $\mathrm{Cov}(X_1, X_2) = 0$.

The relationship $\sigma^2 = E[(X-\mu)^2] = E(X^2) - \mu^2$ is not hard to show (we did it earlier), so in particular, $\mathrm{Cov}(X_1, X_1) = E(X_1^2) - \mu_1^2 = \sigma_1^2$ (a boundary case, since nothing could be less independent than a random variable compared with itself).

When $\mathrm{Cov}(X_1, X_2) = 0$, the two random variables are said to be ***uncorrelated***. If X_1 and X_2 are independent, we noted $\mathrm{Cov}(X_1, X_2) = 0$, which indicates there is some relationship between covariance and independence, but it turns out that this is not a necessary and sufficient condition (so, for those interested in statistical theory, criteria under which this is a necessary and sufficient condition for independence provides a natural object of study).

When $\mathrm{Cov}(X_1, X_2) > 0$, the two random variables are said to be ***positively correlated*** (the counterpart being ***negatively correlated***). Intuitively, positive correlation typically occurs when both $X_1 > \mu_1$ and $X_2 > \mu_2$ or $X_1 < \mu_1$ and $X_2 < \mu_2$ (with the obvious counterpart for negative). In particular, when negative correlation is present, you expect that when one sample mean is on the high side, the other will be on the low side.

To normalize the statistic, the ***correlation*** ρ_{12} (also known as the *correlation coefficient*) is defined by

$$\rho_{12} = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_1 \sigma_2}$$

Here we now have

$$-1 \leq \rho_{12} \leq 1$$

This follows from the fact that $E(X_1^2)E(X_2^2) \geq [E(X_1)E(X_2)]^2$
(*Schwarz's inequality*)

The closer ρ is to 1 or -1 , the greater the dependence relationship between X_1 and X_2.

Simulation models typically employ random variables for input so the model outputs are also random. A ***stochastic process*** is one for which the model outcomes are ordered over time, so in general, the collection of random variables representing one of the outputs of a simulation model over multiple runs, $X_1, X_2, \ldots, X_n$ is a stochastic process.

Reaching conclusions about the model from the implicit stochastic process may require making assumptions about the nature of the stochastic process that may not be true, but are necessary for the application of statistical analysis. For example, an assumption that the theoretical values of μ and σ are the same across $X_1, X_2, \ldots, X_n$. Similarly, the assumption that the covariances $\mathrm{Cov}(X_i. X_{i+t})$ are the same for all t tends to hold only after a "warmup" period for the model to stabilize (the term "*covariance stationary*" is used to describe this situation).

Confidence Intervals

When using data from multiple simulation runs, the objective is to make inferences based on the data provided by the runs. *Point estimates* (eg., a mean value) are one form of inference. *Interval estimates* (eg., a confidence interval) are another form of inference.

When collecting sample means X_i

$$\overline{X}(n) = \frac{\sum_{i=1}^{n} X_i}{n}$$

we expect balance in the sense that E($\overline{X}(n)$) = μ; ie., the collected set of sample means balance around the population mean μ. This is the intuitive basis underlying the ***Central Limit Theorem***, namely, that the sampled means distribute (approximately) normally around μ for large enough sample size n. Moreover, for large n, the mean and variance of the distribution of sample means are given (approximately) by μ and σ^2/n.

Note that the Central Limit Theorem assumes that the random variables X_1, X_2, …, X_n corresponding to the sample means are independent and identically distributed. Since they derive from the same pdfs, they will in general be identically distributed, but will lack independence as noted by Law and Kelton; ie., we are likely to underestimate σ^2 when working with the outcomes of simulation experiments with the objective of estimating μ. The importance is more in the context of explaining why for smaller n, there may be a great deal of variance among the simulation outcomes.

σ^2 is not a known, but if independence of the random variables can be assumed, it can be replaced by the sample standard deviation $S^2(n)$ as discussed earlier.

If we normalize by switching to the sequence of random variables Z_n given by

$$Z_n = \frac{\overline{X}(n) - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

then $E(Z_n) \approx 0$ and $Var(Z_n) = E(Z_n^2) \approx 1$; ie., the distribution approximates the standard normal distribution (keep in mind that we are assuming that the sample means $X_1, X_2, \ldots, X_n$ are random variables which distribute normally as supported by the central limit theorem).

Substituting the sample variance in Z_n for σ^2 we get the values

$$t_n = \frac{\overline{X}(n) - \mu}{\sqrt{\frac{S^2(n)}{n}}}$$

which can be shown to approximate a distribution referred to in the literature as ***Student's t distribution*** (so-called because it was published under the pen name Student by W. S. Gossett in 1908).

Assuming independence throughout, the t distribution is given by

$$\frac{W}{\sqrt{V/n}}$$

where W is given by the standard normal distribution and V is given by the chi-squared distribution with n degrees of freedom; ie., the gamma distribution with control values $\alpha = n/2$ and $\beta = 2$. The calculations are usually given in a table based on "degrees of freedom" $n \geq 2$ and "confidence" values γ.

We won’t get into the details regarding how the t distribution was originally arrived at, but technically, the t distribution with n degrees of freedom is given by the pdf

$$f(x) = \frac{\Gamma\left[\frac{n+1}{2}\right]}{\sqrt{\pi n}\cdot\Gamma\left[\frac{n}{2}\right]\cdot\left(1+\frac{x^2}{n}\right)^{(n+1)/2}}$$

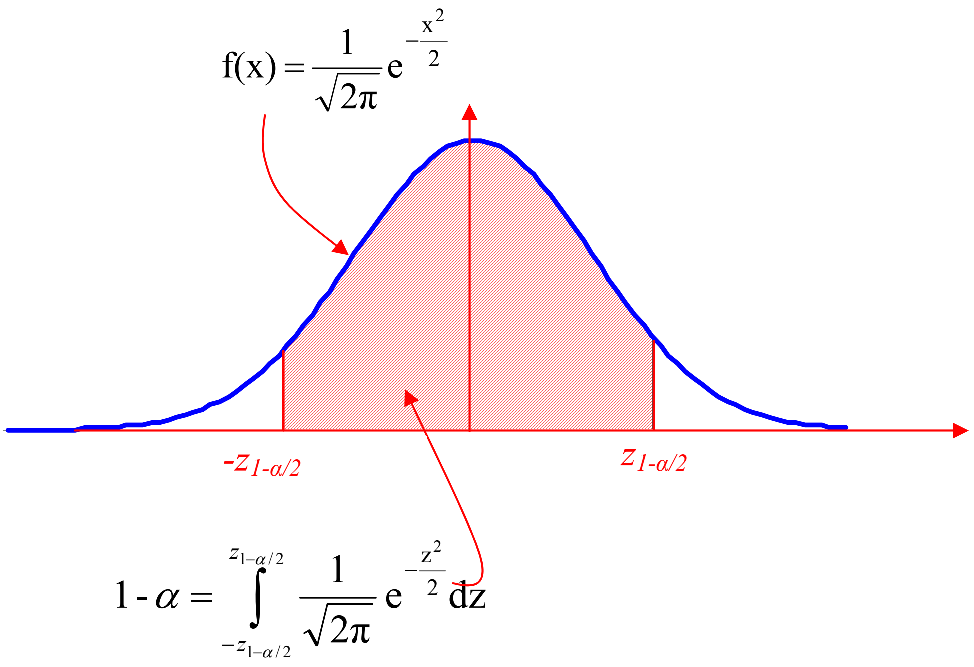
and as might be expected closely approximates the normal distribution, particularly for large n. It’s key virtue is that it provides an estimate for what constitutes “large enough n”.

In our case, the t_n values

$$t_n = \frac{\overline{X}(n)-\mu}{\sqrt{\frac{S^2(n)}{n}}}$$

approximate a t distribution with n-1 degrees of freedom (from the fact that the sample variance is over n-1 intervals).

For our sequence of random variables $X_1, X_2, \ldots, X_n$ we expect with probability approximately 1- α (shaded area below) that the values will fall in the interval $[-z_{1-\alpha/2}, z_{1-\alpha/2}]$, where the notation $z_{1-\alpha/2}$ designates the symmetric points on the axis yielding area 1- α.



$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$1-\alpha = \int_{-z_{1-\alpha/2}}^{z_{1-\alpha/2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

$[-z_{1-\alpha/2}, z_{1-\alpha/2}]$ is called the ***confidence interval*** for 100(1-α) percent confidence. For a simulation outcome, either the confidence interval contains the mean or it does not. If the means are normally distributed, are independent, and one conducts a sufficiently large number of experiments, then the proportion of these whose values fall within the confidence interval should be approximately (1-α).

Recall:

$$Z_n = \frac{\overline{X}(n) - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

$$t_n = \frac{\overline{X}(n) - \mu}{\sqrt{\frac{S^2(n)}{n}}}$$

With our data, we could then look at the interval

$$\overline{X}(n) \pm z_{1-\alpha/2}\sqrt{\frac{S^2(n)}{n}}$$

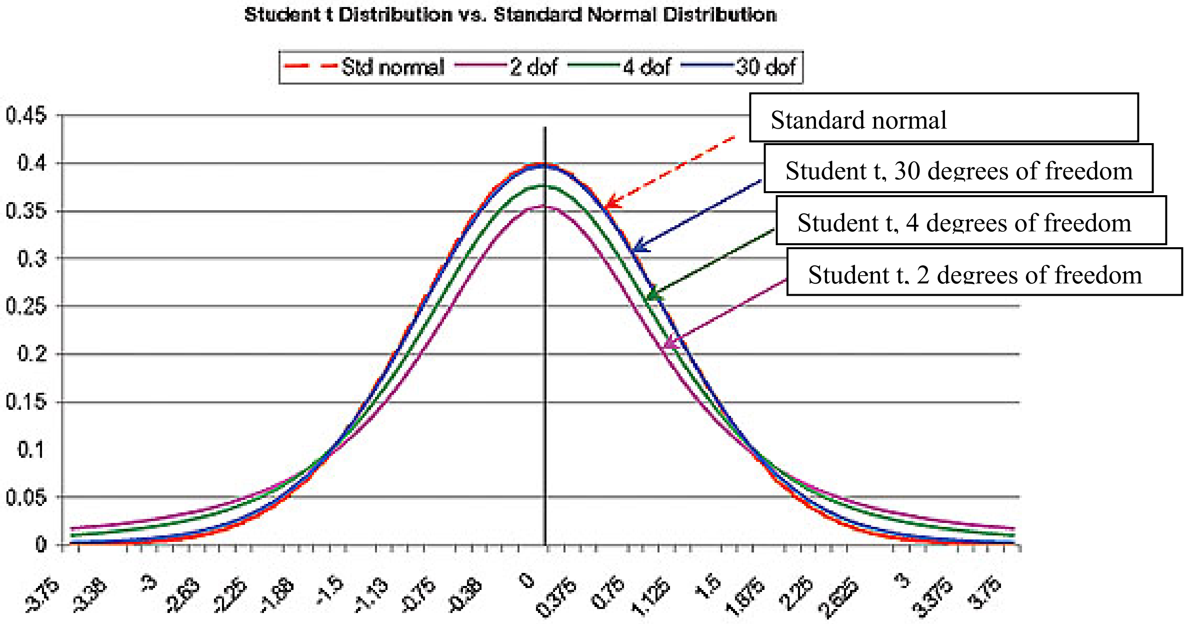
The trouble is that this sheds no light on “sufficiently large”.

Since the sequence t_1, t_2, … t_n approximates a t distribution with n-1 degrees of freedom, we have a natural alternative for getting confidence intervals that do take into account the number of outcomes. Hence, for any finite n, we can use the interval

$$\overline{X}(n) \pm t_{n-1,1-\alpha/2}\sqrt{\frac{S^2(n)}{n}}$$

where $t_{n-1,1-\alpha/2}$ is the upper critical point with n - 1 degrees of freedom for a probability (area across the interval) of 1- α.

This is a measure of how precisely we know μ for the experiment.



Remarks:

- the t distribution has longer tails than the normal distribution; ie., $t_{n-1,1-\alpha/2} > z_{1-\alpha/2}$, so it is a more conservative testing mechanism.
- if we increase n by a factor of 4 to 4n, then the control number for the confidence interval

  $$t_{n-1,1-\alpha/2}\sqrt{\frac{S^2(n)}{n}}$$

  is decreased by a factor of approximately 2.
- as $n\to\infty$, the values $t_{n-1,1-\alpha/2} \to z_{1-\alpha/2}$
- the approximate cumulative point differences (absolute values) and the cumulative $\Delta x = 0.125$ area differences between the Student t and standard normal distributions for various degrees of freedom ranges as follows:

Degrees of freedom	Point difference	Area difference
2	6.12%	6.23%
4	4.09%	1.88%
8	2.27%	0.50%
16	1.19%	0.14%
32	0.68%	0.05%
64	0.46%	0.02%

Hypothesis Testing

Law and Kelton report an interesting experiment that motivates hypothesis testing:

- For each of normal, exponential, chi square, lognormal, and hyperexponential distributions
- For each of n=5, 10, 20 and 40 as sample sizes
- Generate the sample and compute its average
- Do 500 times and keep track of whether or not the average is in the confidence interval given by $t_{n-1,1-\alpha/2}$

In other words, for an exponential distribution of known mean μ and for n=10, 10 observations are generated, and the average computed. It is either in the confidence interval given by $t_{n-1,1-\alpha/2}$ or it is not.

Tabulation of the results for a 90 percent confidence interval:

Distribution	**Skewness**	**n=5**	**n=10**	**n=20**	**n=40**
Normal	0.00	0.910	0.902	0.898	0.900
Exponential	2.00	0.854	0.878	0.870	0.890
Chi square (1 df)	2.83	0.810	0.830	0.848	0.890
Lognormal	6.18	0.758	0.768	0.842	0.852
Hyperexponential	6.43	0.584	0.586	0.682	0.774

As predicted by the Central Limit Theorem, for each distribution the coverage gets closer to 0.90 as n increases. Skewness evidently affects the size of the sample required. The skewness is a measure of symmetry, equal to 0 for a symmetric distribution. It is given by the ratio

$$\frac{E\left[(X-\mu)^3\right]}{\sigma^3} = \frac{\int_{-\infty}^{\infty}(x-\mu)^3 f(x)dx}{\sigma^3}$$

(contrast this with the self correlation measure $E[(X-\mu)^2]/\sigma^2 = 1$)

In our scenario, we have have a sequence of random variables $X_1, X_2, \ldots, X_n$ corresponding to the sample means, and assume the criteria of the Central Limit Theorem are met, and so these are (approximately) normally distributed.

We wish to test the "*null hypothesis*" H_0 that $\mu = \mu_0$, where μ_0 is a fixed hypothesized value for μ. Our expectation is that if $\left|\overline{X}(n) - \mu_0\right|$ is large too frequently, then H_0 is not likely to be true.

If H_0 is true, we now know that the t_n statistic has a t distribution with n-1 degrees of freedom. Therefore, our hypothesis test for $\mu = \mu_0$ is arguably best handled using the t statistic and is as follows:

if $|t_n| > t_{n-1,1-\alpha/2}$ then "reject" H_0

This is the same as saying the test value falls outside of the confidence interval. The probability that the statistic falls into the rejection region when H_0 is true is call the *level* of the test, typically chosen equal to 0.05 or 0.10. We expect to reject in no more than 5% or 10% of cases if H_0 is true.

Two types of errors are possible when performing a hypothesis test

1. Type I error: H_0 is rejected even though true
2. Type II error: H_0 is accepted even though false

The probability α is under the experimenter's control and bounds the probability of a Type I error. In contrast, for a fixed level α, the probability of a Type II error depends on what is actually true and in general has no bound under experimenter control.

If n is increased, then the Type II error count is reduced, but the bounding issue remains. For this reason, the terminology employed in practice is "reject" and "fail to reject", since when we fail to reject, we don't know with any definable certainty that H_0 is true (ie., the test may not be powerful enough to make the distinction).

In general practice, the t statistic can be used if the distribution of the sample values is of the "mounded" variety.

Example:

Suppose the assumed mean for a system is 3000 and we prepare a simulation model for the system. For 10 runs, we have 9 degrees of freedom. Suppose that we obtain the means

μ_1=3005, μ_2=2996, μ_3=2928, μ_4=3003, μ_5=2932,
μ_6=2937, μ_7=2961, μ_8=3015, μ_9=2945, μ_{10}=3001

The Null hypothesis H_0 is that the mean for the simulation model is μ_0=3000; ie.,

$$\left|\overline{X}(n) - 3000\right| = 0$$

From the Student t table, if we want 95% confidence, we want interval end points of $-t_{9,0.975}$ and $t_{9,0.975}$ (2.5% on each end). The tabulated values are then in the $t_{0.025}$ column at degrees of freedom ν=9; ie., 2.26. For 99% confidence, the critical value is 3.25.

The computed values are $\overline{X}(10)$ = 2972.3 and S (10) = 34.84266

The computed t statistic is

$$t = \frac{\overline{X}(10) - \mu_0}{\sqrt{\dfrac{S^2(10)}{10}}} = -2.51402$$

$$\boxed{S^2(n) = \frac{\sum_{i=1}^{n}\left[X_i - \overline{X}(n)\right]^2}{n-1}}$$

falls outside of the confidence interval. Hence, we can reject the Null hypothesis with 95% confidence (meaning in this case 95% confidence that the mean is < 3000), but we cannot reject it with 99% confidence.

More to the point, we can construct a 95% confidence interval around the mean as

$$\overline{X}(10) \pm 2.26 \cdot \sqrt{\frac{S^2(10)}{10}} = 2972.3 \pm 2.26\frac{34.84266}{\sqrt{10}} = 2972.3 \pm 24.9$$

For the 99% case, the interval widens to ±35.8, which includes 3000. Note that if n is increased, the confidence interval shortens.

The t statistic is employed where data has mound characteristics. The test statistic utilized should be one that can arguably represent the sampled data. In testing a sequence of numbers for randomness, uniformity implies the test statistic must compare the sampled values to the uniform distribution, for example.

A Specific Test for Randomness: The Runs Test

A sequence of values can be uniformly distributed but still exhibit characteristics of dependence.

Consider the sequence of integers
6, 1, 6, 1, 4, 6, 7, 6, 10, 7, 7, 3, 3, 1, 8, 2, 5, 9, 8, 5, 10, 4, 4, 4, 6, 3, 9, 10, 8, 7
Flag each number with a + if it is followed by a larger number and with a – if it is followed by a smaller number.
- + - + + + - + - + - + - + - + + - - + - + + + - + + - -

Each string of +'s and –'s forms a run, so our sequence of 30 numbers has 21 runs. There can be too few runs (e.g., the numbers simply ascend) or too many runs (e.g. alternating runs of length 1).

By using combinatorial counting methods (where every combination of +'s and –'s is equally probable, in our case 2^{30} of them), it can be shown that the mean number of runs μ_r for sequences of length N is given by

$$\mu_r = \frac{2N-1}{3} \text{ with variance } \sigma_r^2 = \frac{16N-29}{90}$$

We know that since we are dealing with means, we can use the Student t distribution for the test statistic.

For N=30, $\mu_r = 19.67$ and $\sigma_r = 3.877$. The Null hypothesis is that the sample mean is the same as μ_r; ie., with some degree of confidence, no pattern indicating dependence is present.

For our run count N=30, a=21, the t statistic is

$$t = \frac{a - 19.67}{3.877/\sqrt{30}} = 1.879$$

With 29 degrees of freedom, the $t_{0.025}$ column value is 2.04 and for 99% confidence is 2.76. Hence, we fail to reject the Null hypothesis with 95% confidence, indicating that the test does not call independence into question by a significant amount (5%); ie., independence is not rejected based on this test. Note: because σ_r is known, the z statistic could have been used; namely, $(a - \mu_r)/\sigma_r = 0.34$. The 99% value is 2.33. The 64% value is 0.36 (in which case the test does not call independence into question at least 36% of the time), demonstrating that the t-test is more conservative.