Sequential Sampling

Summary

The **Sequential Sampling** procedure implements various Sequential Probability Ratio Tests (SPRTs). Unlike statistical tests which have a fixed sample size, the number of samples required by sequential tests is not predetermined. Instead, after each sample is taken, one of 3 decisions is made:

1. Stop the test and reject the null hypothesis.
2. Stop the test and accept the null hypothesis.
3. Continue sampling.

I n many cases, the SPRT will come to a decision with fewer samples than would have been required for a fixed size test.

Sample StatFolio: *wald1.sgp*

Sample Data:

The file *wald1.sgd* contains a set of observations presented in the classic book by Wald (1947). The data are shown in the following table:

Sample	*Measurement*
1	151
2	144
3	121
4	137
5	138
6	136
7	155
8	160
9	144
10	145
11	130
12	120
13	104
14	140
15	125
16	106
17	145
18	123
19	138
20	108

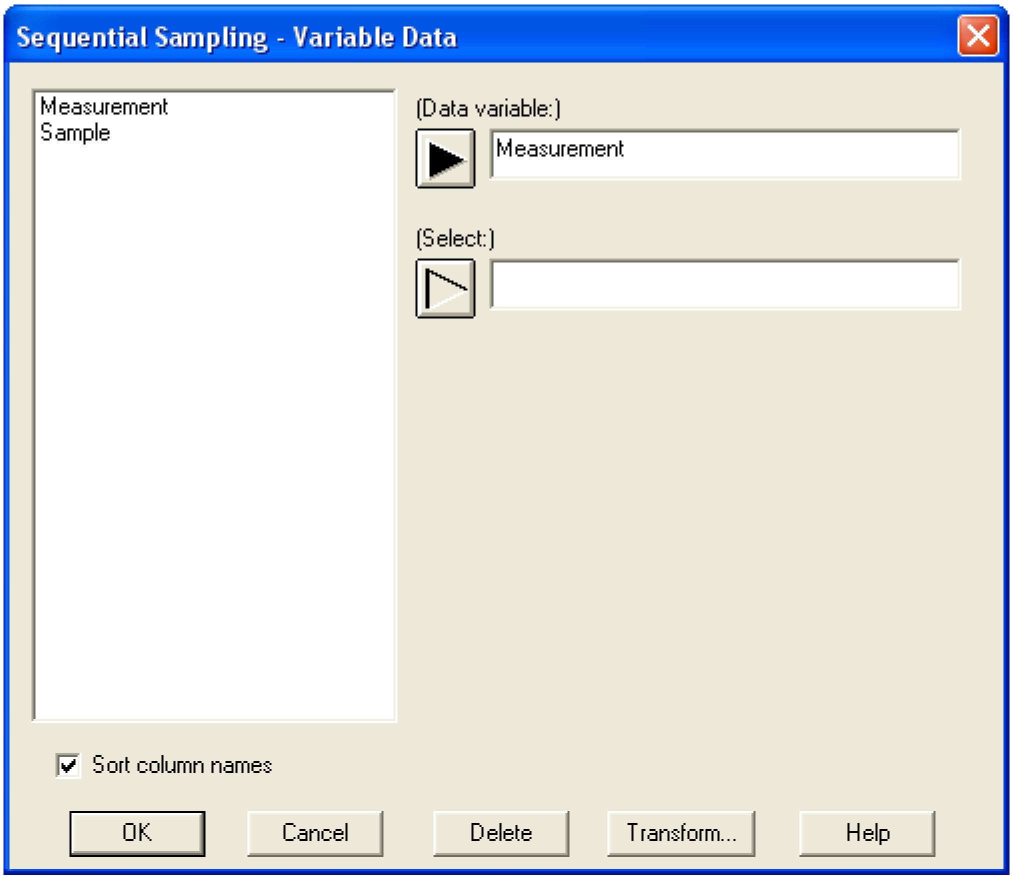
It is assumed that the data are random samples from a normal distribution with known standard deviation σ = 25. It is desired to test the following hypotheses concerning the mean of the distribution:

Null hypothesis: $\mu_0 = 135$

Alternative hypothesis: $\mu_1 = 150$

Data Input

The data input dialog box requests the name of the column containing the measurements:

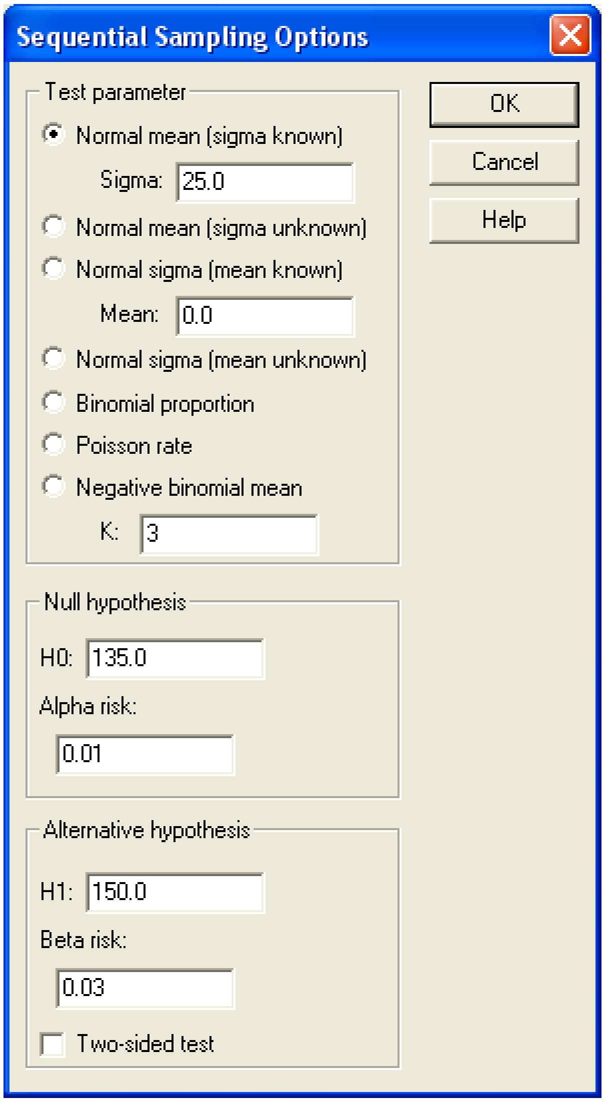


- **Data variable**: an optional column containing the data that has been collected so far. If no data has yet been collected, this field may be left blank.

- **Select**: subset selection.

Analysis Options

After the data has been specified, the *Analysis Options* dialog box is displayed in order to specify the test to be conducted:



- **Test parameter**: the parameter whose value is specified by the null and alternative hypotheses. Options are:

    - *Normal mean (sigma known):* the mean μ of a normal distribution, assuming that the value of the standard deviation σ is known. The value of sigma must be specified.

    - *Normal mean (sigma unknown):* the mean μ of a normal distribution, assuming that the value of the standard deviation σ is not known. The value of sigma will be estimated from the data.

    - *Normal sigma (mean known):* the standard deviation σ of a normal distribution, assuming that the value of the mean μ is known. The value of the mean must be specified.

    - *Normal sigma (mean unknown):* the standard deviation σ of a normal distribution, assuming that the value of the mean μ is not known. The value of the mean will be estimated from the data.

  - *Binomial proportion p*: the probability of an event for the binomial distribution. If this option is selected, all data values must be either 0 or 1.

  - *Poisson rate λ*: the rate parameter or mean of the Poisson distribution. If this option is selected, all data values must be non-negative integers.

  - *Negative binomial mean*: the mean of the negative binomial distribution. If this option is selected, all data values must be non-negative integers. This distribution is often used in place of the Poisson distribution for overdispersed counts (counts for which the variance is greater than the mean). The value of the parameter *k*, (an integer greater than or equal to 1), must be specified.

- **Null hypothesis H0**: the value of the parameter specified by the null hypothesis.

- **Null hypothesis alpha risk**: the α-risk of the test. This is the probability that the test will end with an incorrect rejection of the null hypothesis.

- **Alternative hypothesis H1**: the value of the parameter specified by the alternative hypothesis.

- **Alternative hypothesis beta risk**: the β-risk of the test. This is the probability that the test will end with an incorrect acceptance of the null hypothesis.

- **Two-sided test**: whether the test is two-sided or one-sided (available when testing the normal mean only). If a two-sided test is selected, the alternative hypothesis is assumed to consist of two values equally spaced around the null hypothesis.

Analysis Summary

The *Analysis Summary* displays sample statistics for the data (if any), a summary of the hypothesis test, and the status of the test:

Sequential Analysis (One Sample)
Data variable: Measurement

Count	20
Average	133.5
Median	137.5
Standard deviation	16.0049
Minimum	104.0
Maximum	160.0
Stnd. skewness	-0.768226
Stnd. Kurtosis	-0.527452

Hypothesis Test

Null hypothesis	mean = 135.0	alpha risk = 0.01
Alternative hypothesis	mean = 150.0	beta risk = 0.03

Known sigma = 25.0

Decision: reject null hypothesis at sample 20

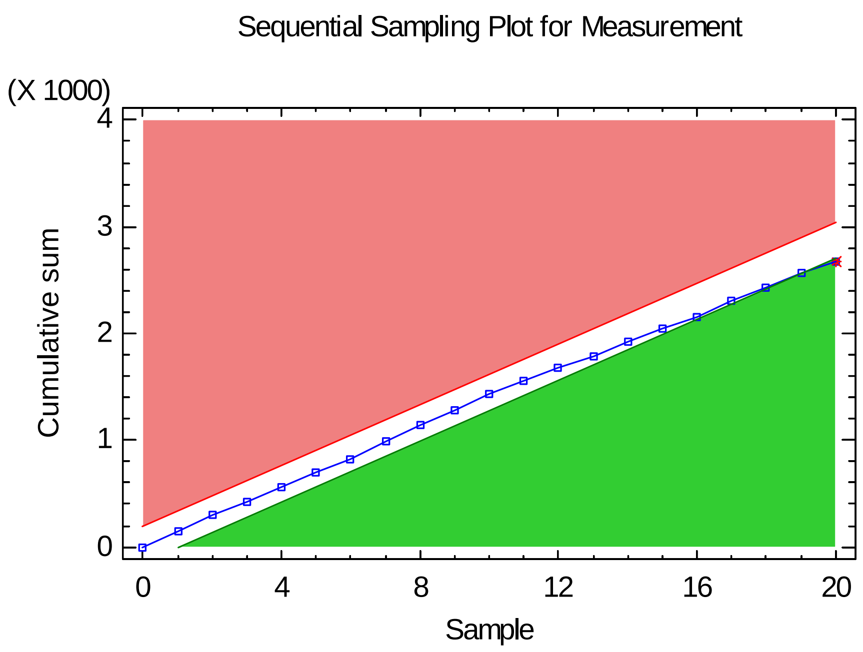
After each data value, a decision is made to accept or reject the null hypothesis if there is sufficient information to stop the test. Otherwise, the program will indicate that continued sampling (more data) is required. Typically, observations are collected and added to the data column one at a time until enough data have been collected to stop the test.

Cumulative Plot

The *Cumulative Plot* illustrates the decision regions appropriate for a sequential sampling plan. For the current example, it plots the cumulative sum of the data values after *m* samples have been taken. Let X_i be the observed data value for the i-th sample. Then the cumulative sum is defined by

$$T_m = \sum_{i=1}^{m} X_i \tag{1}$$

In the plot below, these sums are plotted versus sample number:



Also included on the plot are three regions:

1. A rejection region, colored red. If the cumulative sum enters the rejection region, sampling is stopped and the null hypothesis is rejected.

2. An acceptance region, colored green. If the cumulative sum enters the acceptance region, sampling is stopped and the null hypothesis is accepted.

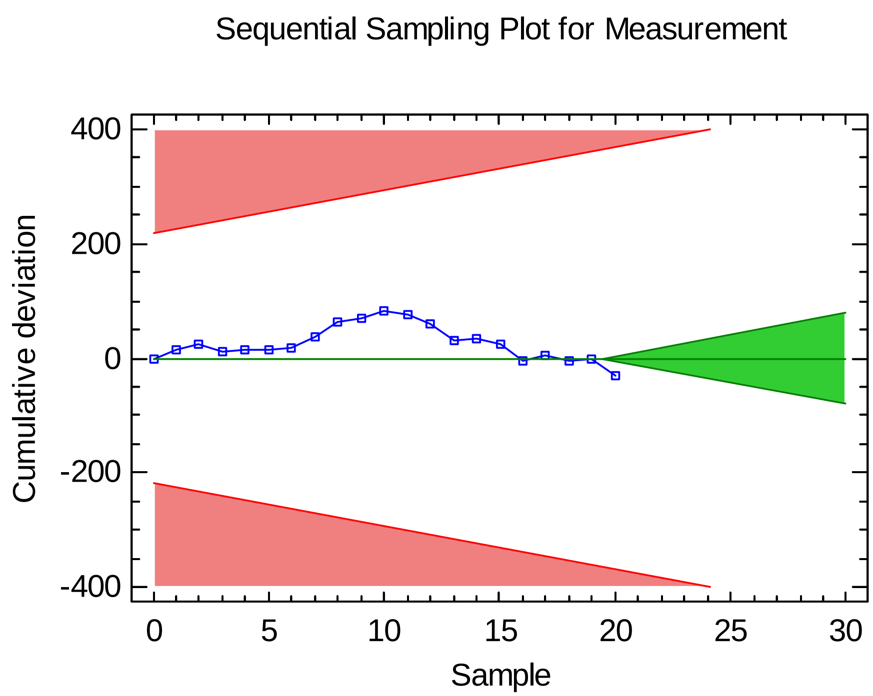
3. A white region. As long as the cumulative sum remains in this region, additional samples must be collected.

The point at which the statistic first enters the acceptance or rejection region, at which further data collection would stop, is indicated by a red asterisk. For the example data, sampling would continue until m = 20 samples have been collected, at which time the null hypothesis would be accepted.

If a two-sided test of were selected, the plot would show instead the cumulative sum of deviations from the null hypothesis, defined by

$$T_m = \sum_{i=1}^{m} \left(X_i - \mu_0 \right) \tag{2}$$

The rejection region then consist of two sections, one for alternative hypotheses greater than the null and another for alternative hypotheses less than the null. A typical example is shown below:



In the plot above, sampling would continue beyond m = 20, since all of the points are within the white zone.

If the value of sigma was not known, the SPRT performed would be a sequential t-test. In such a test, the quantity plotted is the log likelihood ratio defined by

$$T_m = \ln\left(\frac{f\left(t \mid \nu, \delta\right)}{f\left(t \mid \nu, 0\right)} \right) \tag{3}$$

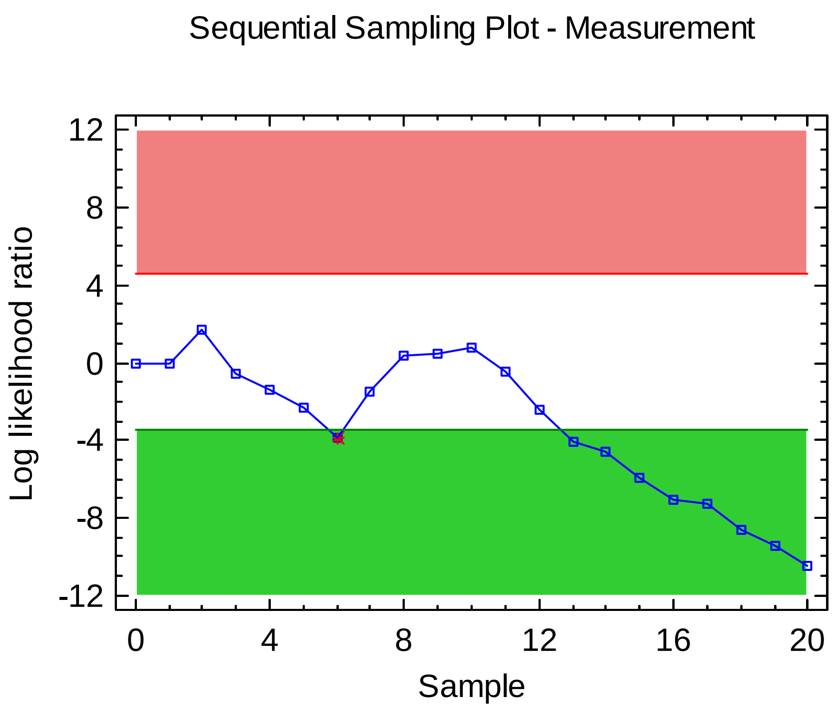
where t is the standard t-statistic calculated from the sample mean and sample standard deviation

$$t = \frac{\left(\bar{x} - \mu_0\right)}{s / \sqrt{m}} \tag{4}$$

and $f(t|\nu,\delta)$ is the probability density function for the non-central t-distribution with ν degrees of freedom and non-centrality parameter

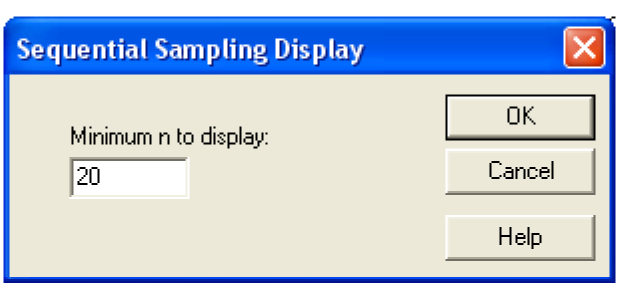
$$\delta = \frac{(\mu_1 - \mu_0)}{s/\sqrt{n}} \tag{5}$$

The decision limits are then two parallel lines as shown below:



In the plot above, the null hypothesis would have been accepted after 6 samples were collected, since the log likelihood ratio enters the green acceptance zone at that point.

Pane Options



Minimum n to display: The plot can be extended beyond the end of the data by specifying a larger sample size.

Decision Numbers

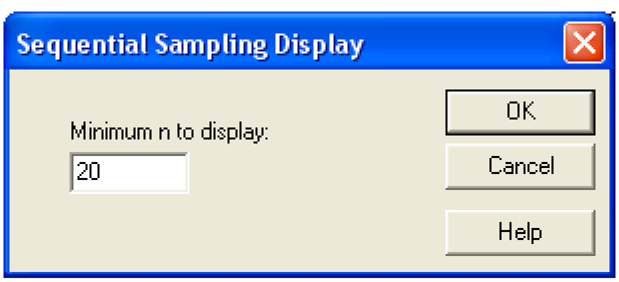
This table displays the value of the plotted statistic and corresponding acceptance and rejection values after each sample is collected:

Decision Numbers

		Acceptance	*Rejection*
Sample	*Cumulative sum*	*Number*	*Number*
1	151.0	-3.18782	333.113
2	295.0	139.312	475.613
3	416.0	281.812	618.113
4	553.0	424.312	760.613
5	691.0	566.812	903.113
6	827.0	709.312	1045.61
7	982.0	851.812	1188.11
8	1142.0	994.312	1330.61
9	1286.0	1136.81	1473.11
10	1431.0	1279.31	1615.61
11	1561.0	1421.81	1758.11
12	1681.0	1564.31	1900.61
13	1785.0	1706.81	2043.11
14	1925.0	1849.31	2185.61
15	2050.0	1991.81	2328.11
16	2156.0	2134.31	2470.61
17	2301.0	2276.81	2613.11
18	2424.0	2419.31	2755.61
19	2562.0	2561.81	2898.11
20	2670.0	2704.31	3040.61

If the data leads to acceptance or rejection of the null hypothesis, the value of the statistic at which that occurs is shown in red.

Pane Options



Minimum n to display: The table can be extended beyond the end of the data by specifying a larger sample size.

Test Performance

The *Test Performance* table displays the properties of the sequential test:

Test Performance

	Mean	*Prob(accept null)*	*Average sample number*
Null hyp.	135.0	0.99	18.9766
	136.5	0.975793	22.9245
	138.0	0.943178	28.1286
	139.5	0.874201	34.4606
	141.0	0.748428	40.7227
	142.5	0.5	
	144.0	0.371918	43.6911
	145.5	0.215877	39.3378
	147.0	0.11574	33.7087
	148.5	0.0595041	28.4336
Alt. hyp.	150.0	0.03	24.0699

Sample size for fixed test n = 50

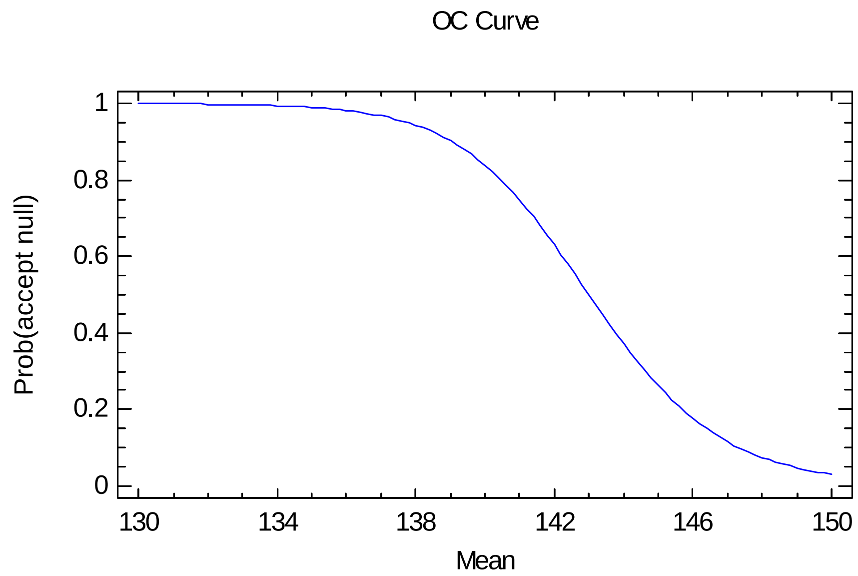
The table has rows for various values of the parameter being tested. Two important quantities are displayed:

1. *Prob(accept null)* – This is the probability that the test will lead to acceptance of the null hypothesis. The alpha and beta risk set this probability when the mean exactly equals the null and alternative hypotheses, respectively. At values between the null and the alternative, the probability will vary.
2. *Average sample number* – The average number of samples that must be taken before the test either accepts or rejects the null hypothesis.

The table also shows the sample size required by a test with predetermined sample size. In most cases, the average sample size for the sequential test will be smaller than for the fixed size test.

O. C. Curve

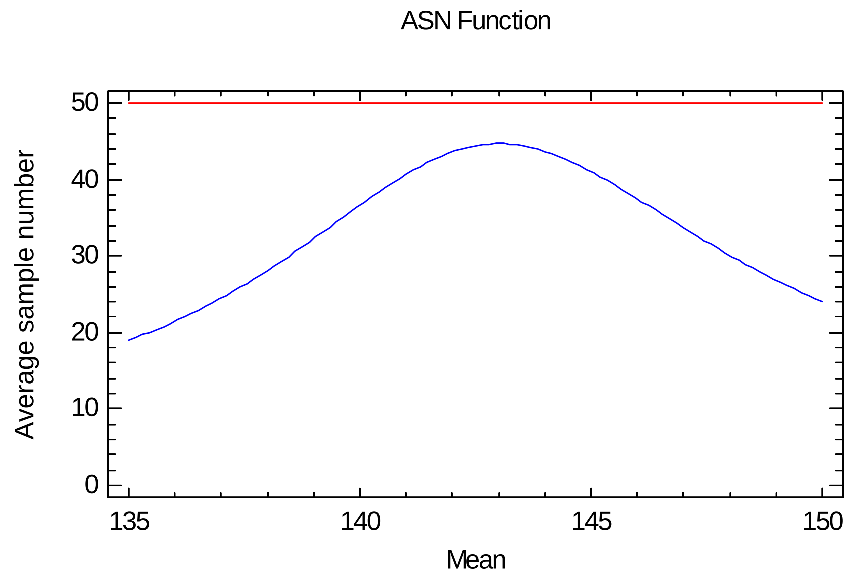
The *O. C. Curve* or *Operating Characteristic Curve* plots the probability that the null hypothesis will be accepted versus various values of the parameter being tested:



This curve passes through the points (μ_0, α) and (μ_1, β).

ASN Function

The *ASN Function* or *Average Sample Number Function* plots the average sample size required before the null hypothesis is either accepted or rejected as a function of the true value of the parameter being tested:



A horizontal line is also displayed at the sample size required for a fixed size test.

Formulas

Test of Normal Mean (One-Sided, Sigma Known)

Test statistic: $T_m = \sum_{i=1}^{m} X_i$ (6)

Acceptance number: $T_m = \frac{\sigma^2}{\mu_1 - \mu_0} \ln\left(\frac{\beta}{1-\alpha}\right) + \left(\frac{\mu_1 + \mu_0}{2}\right)m$ (7)

Rejection number: $T_m = \frac{\sigma^2}{\mu_1 - \mu_0} \ln\left(\frac{1-\beta}{\alpha}\right) + \left(\frac{\mu_1 + \mu_0}{2}\right)m$ (8)

Test of Normal Mean (Two-Sided, Sigma Known)

Test statistic: $T_m = \sum_{i=1}^{m} \left(X_i - \mu_0\right)$ (9)

Acceptance number: $T_m = \frac{\sigma^2}{\delta} \ln\left(\frac{\beta}{1-\alpha/2}\right) + \left(\frac{\delta}{2}\right)m$ (10)

Rejection number: $T_m = \frac{\sigma^2}{\delta} \ln\left(\frac{1-\beta}{\alpha/2}\right) + \left(\frac{\delta}{2}\right)m$ (11)

where

$$\delta = \pm\left|\mu_1 - \mu_0\right| \quad (12)$$

Test of Normal Mean (One-Sided, Sigma Unknown)

Test statistic: $T_m = \ln\left(\frac{f\left(t \mid \nu, \delta\right)}{f\left(t \mid \nu, 0\right)}\right)$ (13)

Acceptance number: $T_m = \ln\left(\frac{\beta}{1-\alpha}\right)$ (14)

Rejection number: $T_m = \ln\left(\frac{1-\beta}{\alpha}\right)$ (15)

Test of Normal Mean (Two-Sided, Sigma Unknown)

Test statistic: $T_m = \ln\left(\frac{f(t \mid \nu, \delta)}{f(t \mid \nu, 0)}\right)$ (16)

Acceptance number: $T_m = \ln\left(\frac{\beta}{1-\alpha/2}\right)$ (17)

Rejection number: $T_m = \ln\left(\frac{1-\beta}{\alpha/2}\right)$ (18)

Test of Normal Sigma (Mean Known)

Test statistic: $T_m = \sum_{i=1}^{m} X_i$ (19)

Acceptance number: $T_m = \frac{2\ln\left(\frac{\beta}{1-\alpha}\right)}{\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right)} + \frac{\ln\left(\frac{\sigma_1^2}{\sigma_0^2}\right)}{\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right)} m$ (20)

Rejection number: $T_m = \frac{2\ln\left(\frac{1-\beta}{\alpha}\right)}{\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right)} + \frac{\ln\left(\frac{\sigma_1^2}{\sigma_0^2}\right)}{\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right)} m$ (21)

Test of Normal Sigma (Mean Unknown)

This case uses the same statistic as the test which assumes that the mean is known. The acceptance and rejection numbers after *m* samples have been collected correspond to the acceptance and rejection numbers after *m-1* samples when the mean is known.

Test of Binomial Proportion

Test statistic: $T_m = \sum_{i=1}^{m} X_i$ (22)

Acceptance number: $T_m = \frac{\ln\left(\frac{\beta}{1-\alpha}\right)}{\ln\left(\frac{p_1}{p_0}\right) - \ln\left(\frac{1-p_1}{1-p_0}\right)} + \left(\frac{\ln\left(\frac{1-p_0}{1-p_1}\right)}{\ln\left(\frac{p_1}{p_0}\right) - \ln\left(\frac{1-p_1}{1-p_0}\right)}\right)m$ (23)

Rejection number: $T_m = \frac{\ln\left(\frac{1-\beta}{\alpha}\right)}{\ln\left(\frac{p_1}{p_0}\right) - \ln\left(\frac{1-p_1}{1-p_0}\right)} + \left(\frac{\ln\left(\frac{1-p_0}{1-p_1}\right)}{\ln\left(\frac{p_1}{p_0}\right) - \ln\left(\frac{1-p_1}{1-p_0}\right)}\right)m$ (24)

Test of Poisson Rate

Test statistic: $T_m = \sum_{i=1}^{m} X_i$ (25)

Acceptance number: $T_m = \frac{\ln\left(\frac{\beta}{1-\alpha}\right)}{\ln\left(\frac{\lambda_1}{\lambda_0}\right)} + \left(\frac{\lambda_1 - \lambda_0}{\ln\left(\frac{\lambda_1}{\lambda_0}\right)}\right)m$ (26)

Rejection number: $T_m = \frac{\ln\left(\frac{1-\beta}{\alpha}\right)}{\ln\left(\frac{\lambda_1}{\lambda_0}\right)} + \left(\frac{\lambda_1 - \lambda_0}{\ln\left(\frac{\lambda_1}{\lambda_0}\right)}\right)m$ (27)

Test of Negative Binomial Mean

Test statistic: $T_m = \sum_{i=1}^{m} X_i$ (28)

Acceptance number: $T_m = \frac{\ln\left(\frac{\beta}{1-\alpha}\right)}{\ln\left(\frac{\mu_1(\mu_0 + k)}{\mu_0(\mu_1 + k)}\right)} + k\left(\frac{\ln\left(\frac{\mu_1 + k}{\mu_0 + k}\right)}{\ln\left(\frac{\mu_1(\mu_0 + k)}{\mu_0(\mu_1 + k)}\right)}\right)m$ (29)

Rejection number: $$T_m = \frac{\ln\left(\frac{1-\beta}{\alpha}\right)}{\ln\left(\frac{\mu_1(\mu_0+k)}{\mu_0(\mu_1+k)}\right)} + k\left(\frac{\ln\left(\frac{\mu_1+k}{\mu_0+k}\right)}{\ln\left(\frac{\mu_1(\mu_0+k)}{\mu_0(\mu_1+k)}\right)}\right)m \qquad (30)$$

<u>Operating Characteristic Curve</u>

$$L(\theta) = \frac{\left(\frac{1-\beta}{\alpha}\right)^{h(\theta)} - 1}{\left(\frac{1-\beta}{\alpha}\right)^{h(\theta)} - \left(\frac{\beta}{1-\alpha}\right)^{h(\theta)}} \qquad (31)$$

where $h(\theta)$ is a quantity that depends on the value of the unknown parameter θ (see Wald, 1947).

<u>Average Sample Number</u>

$$E_\theta(n) = \frac{L(\theta)\ln\left(\frac{\beta}{1-\alpha}\right) + [1-L(\theta)]\ln\left(\frac{1-\beta}{\alpha}\right)}{E_\theta\left(\frac{f(x,\theta_1)}{f(x,\theta_0)}\right)} \qquad (32)$$

where $f(x,\theta)$ is the probability density or mass function for the data variable.

