Chapter 260

Tests for One ROC Curve

Introduction

Receiver operating characteristic (ROC) curves are used to assess the accuracy of a diagnostic test. The technique is used when you have a criterion variable which will be used to make a yes or no decision based on the value of this variable. The area under the ROC curve (AUC) is a popular summary index of an ROC curve.

This module computes power and sample size when a new diagnostic test is compared to an existing (gold) standard. Two approaches are available: the approach of Hanley and McNeil (1982) is used when the criterion variable is continuous and the approach of Obuchowski and McClish (1997) is used when the criterion variable is a discrete rating scale.

Technical Details

In the following, we suppose that we have two groups of patients, those with a condition of interest (the positive group) and those without it (the negative group). This classification may be known from extensive diagnosis or based on the value of another diagnostic test. The diagnostic test of interest is performed on each patient and the resulting test value is recorded. At each specified cutoff value of the criterion variable, the true positive rate (TPR) and the false positive rate (FPR) are calculated. A plot of the TPR versus the FPR allows you study the consequences of using various cutoff values. This plot is called the *ROC curve*.

It should be noted that TPR is similar to the statistical power of the diagnostic test at a particular cutoff value of the criterion variable. Similarly, FPR is an estimate of the probability that the diagnostic test results in a type I (alpha) error. Thus, the ROC curve may be interpreted as a plot of the diagnostic test's power versus it's significance level at various possible criterion cutoff values.

Users of ROC curves have developed special names for TPR and FPR. They call TPR the *sensitivity* of the test and 1 - FPR the *specificity* of the test. Statisticians will be more familiar with using the word *power* instead of sensitivity and the phrase '1 - *alpha*' instead of specificity.

An ROC curve may be summarized by the area under it (AUC). This area has an additional interpretation. Suppose that a rater is asked to study two subjects, one that is actually disease positive and one that is disease negative. The AUC is equal to the probability that the rater will give the disease positive subject a higher score than the disease negative subject. That is, the AUC is the probability that the rater will correctly order the two subjects as to which is more likely to have the disease.

Several methods of computing the AUC have been proposed. One method uses the trapezoidal rule to calculate the AUC directly. Another method, called the *binormal model*, computes the area by fitting two normal distributions to the data.

The Binormal Model

Let *X* denote the distribution of the criterion variable for negative (normal) patients and *Y* denote the distribution of the criterion variable for positive (diseased) patients. It is assumed that

$$X \sim N(\mu_-, \sigma_-^2)$$

and

$$Y \sim N(\mu_+, \sigma_+^2)$$

For a particular cutoff value of the criterion variable, *c*, the true positive rate is given by

$$\begin{aligned} TPR(c) &= \mathrm{P}(Y > c) \\ &= 1 - \Phi\left(\frac{c - \mu_+}{\sigma_+}\right) \\ &= \Phi\left(\frac{\mu_+ - c}{\sigma_+}\right) \end{aligned}$$

where $\Phi(z)$ is the cumulative normal distribution.

Similarly, the false positive rate is given by

$$\begin{aligned} FPR(c) &= \mathrm{P}(X > c) \\ &= 1 - \Phi\left(\frac{c - \mu_-}{\sigma_-}\right) \\ &= \Phi\left(\frac{\mu_- - c}{\sigma_-}\right) \end{aligned}$$

The ROC curve is thus the curve traced out by the functions

$$[FPR(c), TPR(c)] = \left[\Phi\left(\frac{\mu_- - c}{\sigma_-}\right), \Phi\left(\frac{\mu_+ - c}{\sigma_+}\right)\right]$$

The area under the ROC curve, AUC, is defined as

$$\begin{aligned} \theta &= \int_{-\infty}^{\infty} TPR(c) FPR'(c) dc \\ &= \int_{-\infty}^{\infty} \Phi\left(\frac{\mu_+ - c}{\sigma_+}\right) \phi\left(\frac{\mu_- - c}{\sigma_-}\right)\left(-\frac{1}{\sigma_-}\right) dc \\ &= \int_{-\infty}^{\infty} \Phi(A + Bv)\phi(v) dv \\ &= \Phi\left(\frac{A}{\sqrt{1 + B^2}}\right) \end{aligned}$$

where

$$c = \mu_- - v\sigma_-$$

$$A = \frac{|\mu_+ - \mu_-|}{\sigma_+}$$

$$B = \frac{\sigma_-}{\sigma_+}$$

Maximum likelihood estimates of *A* and *B* can be computed and used to compute AUC. The variances and covariance of these MLE's can be estimated from Fisher's information matrix.

Define $\Delta = \theta_1 - \theta_0$ to be the difference in the accuracies (AUC's) between the new and standard tests. The test statistic for this test is

$$Z = \frac{\hat{\Delta}}{\sqrt{V(\theta_0)}} = \frac{\hat{\theta}_1 - \theta_0}{\sqrt{V(\theta_0)}}$$

where $V(\theta_0)$ is the variance of $\hat{\theta}$ under the null hypothesis of equality. The above test statistic gives the following formulae for computing sample size or power

$$N_+ = \frac{\left(z_\alpha\sqrt{V(\theta_0)} + z_\beta\sqrt{V(\theta_1)}\right)^2}{(\theta_1 - \theta_0)^2}$$

$$\beta = \Phi\left(\frac{|\theta_1 - \theta_0|\sqrt{N_+} - z_\alpha\sqrt{V(\theta_0)}}{\sqrt{V(\theta_1)}}\right)$$

Discrete (Rating) Data

For a criterion variable yielding a discrete rating and with $R = N_-/N_+$, Obuchowski (1998) recommends

$$V(\theta) = f^2\left(1 + \frac{B^2}{R} + \frac{A^2}{2}\right) + g^2B^2\left(\frac{1+R}{2R}\right)$$

where

$$f = \frac{E_1}{\sqrt{2\pi(1+B^2)}}$$

$$g = -\frac{ABE_1}{\sqrt{2\pi(1+B^2)^3}}$$

$$E_1 = \exp\left(-\frac{A^2}{2+2B^2}\right)$$

The value of *A* can be found as

$$A = \Phi^{-1}(\theta)\sqrt{1 + B^2}$$

For the most conservative results, Obuchowski (1998) recommends setting $B = 1$, so that

$$A = \Phi^{-1}(\theta)\sqrt{2}$$

Continuous Data

For a criterion variable yielding a continuous result and with $R = N_-/N_+$, Obuchowski (1998) suggests that the following formula of Hanley and McNeil (1983) is more appropriate

$$V(\theta) = \frac{\theta}{R(2-\theta)} + \frac{2\theta^2}{1+\theta} - \theta^2\left(\frac{1+R}{R}\right)$$

Example 1 – Calculating Power

An investigator wants to study the accuracy of a diagnostic test which yields measurements on a rating scale from 1 to 5. Historically, such tests have had an AUC of 0.80. The investigator wants to investigate three alternative AUC values: 0.825, 0.850, and 0.900. A two-sided test is planned with a significance level of 0.05. Since no other information is available, B is set to 1.0. The investigator would like to achieve a power of 90% in the study. Patients without the disease under study are about twice as frequent as patients with the disease. The investigator wants to see results for a sample size of up to 6000 patients.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab	
Solve For	**Power**
Alternative Hypothesis	**Two-Sided Test**
Alpha	**0.05**
Group Allocation	**Enter N+ and R, where N- = R × N+**
N+	**20 50 100 250 500 1000 2000**
R	**2**
AUC0 (Area Under Curve|H0)	**0.80**
AUC1 (Area Under Curve|H1)	**0.825 0.85 0.90**
Type of Data	**Discrete (Ratings)**
B (SD Ratio = SD-/SD+)	**1.0**
Lower FPR	**0.00**
Upper FPR	**1.00**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Reports

Numeric Results

Solve For:	Power
Alternative Hypothesis:	Two-Sided (H1: AUC ≠ AUC0)
Type of Data:	Discrete (Ratings)
B (SD Ratio):	1
FPR Limits:	Lower = 0, Upper = 1

	Sample Size					Area Under the ROC Curve						
				R (N- / N+)		Actual			Adjusted for FPR Limits			
Power	N+	N-	N	Target	Actual	H0 AUC0	H1 AUC1	Difference Diff	H0 AUC0'	H1 AUC1'	Difference Diff'	Alpha
0.04806	20	40	60	2	2	0.8	0.825	0.025	0.8	0.825	0.025	0.05
0.07393	50	100	150	2	2	0.8	0.825	0.025	0.8	0.825	0.025	0.05
0.11455	100	200	300	2	2	0.8	0.825	0.025	0.8	0.825	0.025	0.05
0.23648	250	500	750	2	2	0.8	0.825	0.025	0.8	0.825	0.025	0.05
0.43207	500	1000	1500	2	2	0.8	0.825	0.025	0.8	0.825	0.025	0.05
0.72636	1000	2000	3000	2	2	0.8	0.825	0.025	0.8	0.825	0.025	0.05
0.95496	2000	4000	6000	2	2	0.8	0.825	0.025	0.8	0.825	0.025	0.05
0.08703	20	40	60	2	2	0.8	0.850	0.050	0.8	0.850	0.050	0.05
0.18340	50	100	150	2	2	0.8	0.850	0.050	0.8	0.850	0.050	0.05
0.34913	100	200	300	2	2	0.8	0.850	0.050	0.8	0.850	0.050	0.05
0.73689	250	500	750	2	2	0.8	0.850	0.050	0.8	0.850	0.050	0.05
0.96287	500	1000	1500	2	2	0.8	0.850	0.050	0.8	0.850	0.050	0.05
0.99968	1000	2000	3000	2	2	0.8	0.850	0.050	0.8	0.850	0.050	0.05
1.00000	2000	4000	6000	2	2	0.8	0.850	0.050	0.8	0.850	0.050	0.05
0.24895	20	40	60	2	2	0.8	0.900	0.100	0.8	0.900	0.100	0.05
0.65634	50	100	150	2	2	0.8	0.900	0.100	0.8	0.900	0.100	0.05
0.94738	100	200	300	2	2	0.8	0.900	0.100	0.8	0.900	0.100	0.05
0.99997	250	500	750	2	2	0.8	0.900	0.100	0.8	0.900	0.100	0.05
1.00000	500	1000	1500	2	2	0.8	0.900	0.100	0.8	0.900	0.100	0.05
1.00000	1000	2000	3000	2	2	0.8	0.900	0.100	0.8	0.900	0.100	0.05
1.00000	2000	4000	6000	2	2	0.8	0.900	0.100	0.8	0.900	0.100	0.05

Power	The probability of rejecting a false null hypothesis when the alternative hypothesis is true.
N+ and N-	The number of items sampled from each population.
N	The total sample size. N = (N+) + (N-).
Target R	The desired ratio (or ratios) of R entered in the procedure. R is the ratio of N- to N+, so that N- = R × N+.
Actual R	The value for R obtained in this scenario. Because N+ and N- are discrete, this value is sometimes slightly different than the target R.
AUC0 and AUC1	The actual areas under the ROC curve for the null and alternative hypotheses, respectively.
Diff	The difference to be detected. Diff = AUC1 - AUC0.
AUC0' and AUC1'	The adjusted (rescaled to account for FPR limits) areas under the ROC curve for the null and alternative hypotheses, respectively.
Diff'	The adjusted (rescaled) difference to be detected. Diff = AUC1' - AUC0'.
Alpha	The probability of rejecting a true null hypothesis.
B (SD Ratio)	The ratio of the standard deviations of the negative and positive groups. B = SD- / SD+.
FPR Limits	The lower and upper bounds on the false positive rates.

Summary Statements

A two-group (positive/negative or with condition/without) design with discrete (rating scale) response data will be used to test the area under the ROC curve against the null value 0.8. The comparison will be made using a two-sided Z-test with a Type I error rate (α) of 0.05. The area under the curve will be computed between the X-axis (false positive rate) values of 0 and 1. The ratio of the standard deviation of the responses in the negative group to the standard deviation of the responses in the positive group (SD- / SD+) is assumed to be 1. To detect an area under the curve of 0.825 with sample sizes of 20 for the positive (with condition) group and 40 for the negative (without condition) group, the power is 0.04806.

Dropout-Inflated Sample Size

	Sample Size			Dropout-Inflated Enrollment Sample Size			Expected Number of Dropouts		
Dropout Rate	N+	N-	N	N+'	N-'	N'	D+	D-	D
20%	20	40	60	25	50	75	5	10	15
20%	50	100	150	63	125	188	13	25	38
20%	100	200	300	125	250	375	25	50	75
20%	250	500	750	313	625	938	63	125	188
20%	500	1000	1500	625	1250	1875	125	250	375
20%	1000	2000	3000	1250	2500	3750	250	500	750
20%	2000	4000	6000	2500	5000	7500	500	1000	1500

Dropout Rate	The percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e., will be treated as "missing"). Abbreviated as DR.
N+, N-, and N	The evaluable sample sizes at which power is computed (as entered by the user). If N+ and N- subjects are evaluated out of the N+' and N-' subjects that are enrolled in the study, the design will achieve the stated power.
N+', N-', and N'	The number of subjects that should be enrolled in the study in order to obtain N+, N-, and N evaluable subjects, based on the assumed dropout rate. N+' and N-' are calculated by inflating N+ and N- using the formulas N+' = N+ / (1 - DR) and N-' = N- / (1 - DR), with N+' and N-' always rounded up. (See Julious, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. (2018) pages 32-33.)
D+, D-, and D	The expected number of dropouts. D+ = N+' - N+, D- = N-' - N-, and D = D+ + D-.

Dropout Summary Statements

Anticipating a 20% dropout rate, 25 subjects should be enrolled in Group 1, and 50 in Group 2, to obtain final group sample sizes of 20 and 40, respectively.

References

Hanley, J. A. and McNeil, B. J. 1983. 'A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases.' Radiology, 148, 839-843. September, 1983.

Obuchowski, N. and McClish, D. 1997. 'Sample Size Determination for Diagnostic Accuracy Studies Involving Binormal ROC Curve Indices.' Statistics in Medicine, 16, pages 1529-1542.

This report shows the power for each of the sample sizes. Most of the definitions are standard. However, a special explanation must be given for AUC and AUC'.

AUC

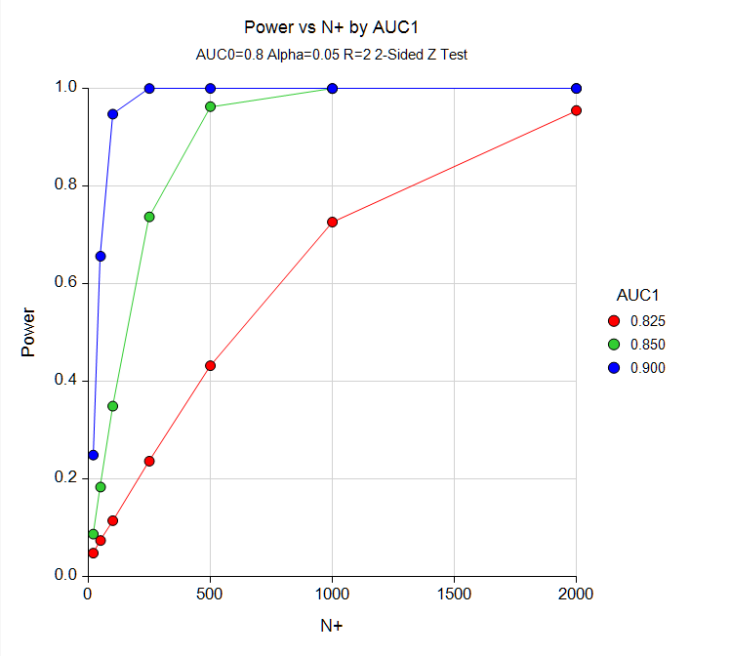
This is the actual area under the curve. This value will equal the adjusted area when the FPR range is set from 0.0 to 1.0. Otherwise, these values will be different.

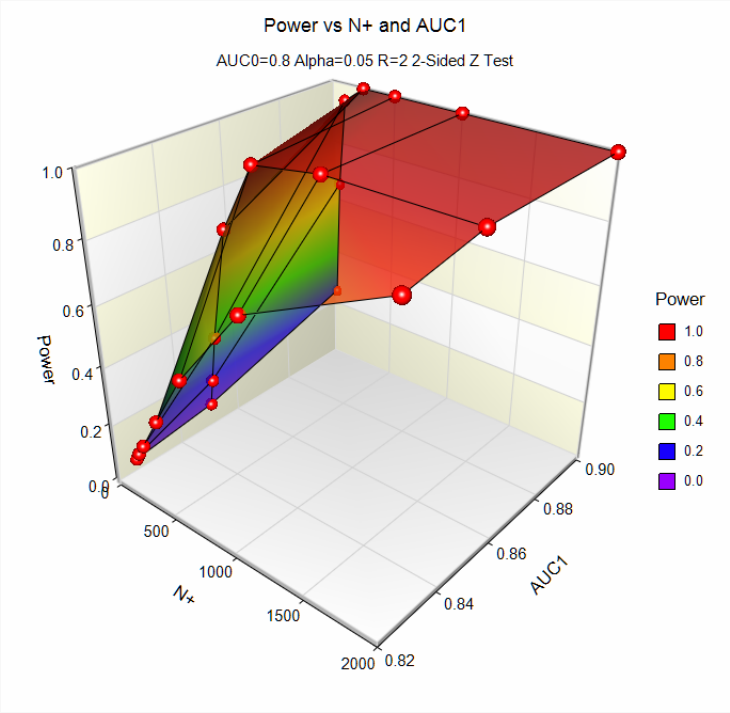
AUC'

This adjusted value is only useful when the FPR range is not 0.0 to 1.0. An adjustment is applied so that the minimum area is 0.5 and the maximum area is 1.0. This will yield the values of AUC that resulted when the FPR range was 0.0 to 1.0.

Plots Section

Plots





These plots show the power versus the sample size for the three values of AUC1.

Example 2 – Calculating Sample Size

Continuing on with Example1, the investigator wants to know the exact sample size needed for each of the three values of AUC2. The investigator wants to look at the Numeric Report.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For	**Sample Size**
Alternative Hypothesis	**Two-Sided Test**
Power	**0.90**
Alpha	**0.05**
Group Allocation	**Enter R = N-/N+, solve for N+ and N-**
R	**2**
AUC0 (Area Under Curve|H0)	**0.80**
AUC1 (Area Under Curve|H1)	**0.825 0.85 0.90**
Type of Data	**Discrete (Ratings)**
B (SD Ratio = SD-/SD+)	**1.0**
Lower FPR	**0.00**
Upper FPR	**1.00**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Solve For:	Sample Size
Alternative Hypothesis:	Two-Sided (H1: AUC ≠ AUC0)
Type of Data:	Discrete (Ratings)
B (SD Ratio):	1
FPR Limits:	Lower = 0, Upper = 1

Power		Sample Size					Area Under the ROC Curve						
					R (N- / N+)		Actual			Adjusted for FPR Limits			
Target	Actual	N+	N-	N	Target	Actual	H0 AUC0	H1 AUC1	Difference Diff	H0 AUC0'	H1 AUC1'	Difference Diff'	Alpha
0.9	0.90010	1582	3164	4746	2	2	0.8	0.825	0.025	0.8	0.825	0.025	0.05
0.9	0.90069	381	762	1143	2	2	0.8	0.850	0.050	0.8	0.850	0.050	0.05
0.9	0.90243	85	170	255	2	2	0.8	0.900	0.100	0.8	0.900	0.100	0.05

This report shows the sample size needed to achieve 90% power for each value of AUC1.

Example 3 – Partial Area under Curve

Continuing on with Example 2, the investigator knows that FPR values between 0.0 and 0.20 are the only values of interest. Hence, he wants to investigate the sample size needed when the FPR range is confined to this range.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 3** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For	**Sample Size**
Alternative Hypothesis	**Two-Sided Test**
Power	**0.90**
Alpha	**0.05**
Group Allocation	**Enter R = N-/N+, solve for N+ and N-**
R	**2**
AUC0 (Area Under Curve|H0)	**0.80**
AUC1 (Area Under Curve|H1)	**0.825 0.85 0.90**
Type of Data	**Discrete (Ratings)**
B (SD Ratio = SD-/SD+)	**1.0**
Lower FPR	**0.00**
Upper FPR	**0.20**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Solve For:	Sample Size
Alternative Hypothesis:	Two-Sided (H1: AUC ≠ AUC0)
Type of Data:	Discrete (Ratings)
B (SD Ratio):	1
FPR Limits:	Lower = 0, Upper = 0.2

		Sample Size					Area Under the ROC Curve						
							Actual			Adjusted for FPR Limits			
Power					R (N- / N+)								
Target	Actual	N+	N-	N	Target	Actual	H0 AUC0	H1 AUC1	Difference Diff	H0 AUC0'	H1 AUC1'	Difference Diff'	Alpha
0.9	0.90009	2663	5326	7989	2	2	0.128	0.137	0.009	0.8	0.825	0.025	0.05
0.9	0.90018	645	1290	1935	2	2	0.128	0.146	0.018	0.8	0.850	0.050	0.05
0.9	0.90133	144	288	432	2	2	0.128	0.164	0.036	0.8	0.900	0.100	0.05

Note that the necessary sample size has almost doubled.

Example 4 – Validation using Obuchowski and McClish (1997)

The formulas used in this module were given in Obuchowski and McClish (1997). On page 1538, they provide an example which will be duplicated here. The study investigated the accuracy of MRI for detecting abnormalities in patients with symptomatic knees. In order to do this, they wanted to know the sample size that would be needed to construct a 95% confidence interval so that the length of the confidence interval is no more than 0.10.

The measure of diagnostic accuracy is the AUC from an FPR of 0.0 to an FPR of 1.0. The allocation ratio is 1.5. B = 1.0. The value of A is found to be 1.2. This translates to an AUC0 of 0.7995. The value of AUC1 = AUC0 + 0.10 / 2, where 0.10 is the maximum length of the confidence interval. A two-tailed confidence interval is envisioned in which alpha is 0.05. In order to find the sample size of a confidence interval, the power is set to 50%. In their article, they found N+ = 161 and N- = 242.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 4** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab	
Solve For	**Sample Size**
Alternative Hypothesis	**Two-Sided Test**
Power	**0.50**
Alpha	**0.05**
Group Allocation	**Enter R = N-/N+, solve for N+ and N-**
R	**1.5**
AUC0 (Area Under Curve|H0)	**0.7995**
AUC1 (Area Under Curve|H1)	**0.8495**
Type of Data	**Discrete (Ratings)**
B (SD Ratio = SD-/SD+)	**1.0**
Lower FPR	**0.00**
Upper FPR	**1.00**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Solve For:	Sample Size
Alternative Hypothesis:	Two-Sided (H1: AUC ≠ AUC0)
Type of Data:	Discrete (Ratings)
B (SD Ratio):	1
FPR Limits:	Lower = 0, Upper = 1

Power		Sample Size					Area Under the ROC Curve						
					R (N- / N+)		Actual			Adjusted for FPR Limits			
Target	Actual	N+	N-	N	Target	Actual	H0 AUC0	H1 AUC1	Difference Diff	H0 AUC0'	H1 AUC1'	Difference Diff'	Alpha
0.5	0.50262	162	243	405	1.5	1.5	0.7995	0.8495	0.05	0.7995	0.8495	0.05	0.05

Note that the sample sizes of 162 and 243 are within one of the results of Obuchowski. The difference occurs because their values of 161 and 242 produce a power that is slightly less than 0.5, so **PASS** increased the sample size slightly.