Paper SD05

Using the SAS® System to Construct *n*-Values Plots

Peter Wludyka, University of North Florida, Jacksonville, FL; Amy Cox, University of North Florida, Jacksonville, FL

ABSTRACT

Often a statistical test is performed in which the observed outcome favors the alternative but the evidence in favor of the alternative is not statistically significant. Suppose that, for example, a researcher tests the hypothesis that the mean is 10 versus the alternative is less than 10 using a *z*-test. Suppose further that based on a sample of size 8, the p-value is 0.078 (hence the null hypothesis is not rejected with level of significance 0.05). Given the observed mean (X-bar) how large would the sample have to have been in order for the hypothesis to be rejected? An *n*-values plot can be used to answer this question. This plot has vertical axis *n* (sample sizes) and horizontal axis alpha (level of significance). Points forming an *n*-values line containing combinations of (alpha, *n*) with the minimum n required to rejected at alpha are plotted. A horizontal line at n =8 in this example would be plotted (the *n*-values line will intersect the *n* = 8 line at alpha = 0.078. Using the plot one can "see" that a sample of size 11 would have rejected at level of significance 0.05. SAS macros for generating plots for various commonly used tests will be presented.

INTRODUCTION

Several macros will be presented that can be used to produce n-values plots. The following test will be considered: one-sample z-test and t-test for a mean; one sample z-test for a proportion; ANOVA F-test for two or more means; and 2 by 2 table tests.

ONE SAMPLE *Z*-TEST AND PROTOTYPE

Consider the standard *z*-test for the mean. The one sided hypothesis is

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_A &: \mu < \mu_0 \end{aligned} \qquad (1)$$

where μ_0 is the hypothesized mean. Suppose for example that one wishes to test the hypothesis that the average diameter of a widget is 10mm versus the alternative that the average diameter is less than 10 mm. Then the hypothesis becomes

$$\begin{aligned} H_0 &: \mu = 10 \\ H_A &: \mu < 10 \end{aligned} \qquad (2)$$

The *z*-test is appropriate whenever the parent population is normal with known variance σ^2. Suppose that in this example the standard deviation is known to be 3mm. Suppose that a random sample (X_i distributed $NID(\mu, \sigma^2)$) of size $n = 8$ is selected from the population of widgets and the sample mean is 8.5mm (that is, $\overline{X} = 8.5$). The z-statistic is

$$z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma} = \frac{(8.5 - 10)\sqrt{8}}{3} = -1.414 \qquad (3)$$

and the p-value is 0.07865. Note that the investigator will not reject the null hypothesis at level of significance $\alpha = 0.05$ since the *p*-value is not less than alpha (or equivalently, z is not less than $-z_\alpha = -z_{.05} = -1.645$). This interpretation of the sample mean of 8.5 as being insufficiently small to reject (2) is dependent on the sample size. Had the same result obtained with a larger sample size the decision may have been different. How large would *n* have to be in order for the observed mean to be sufficiently small to lead to rejection of (2)? Simple algebraic manipulation of (3) produces

$$\left(\frac{\sigma z_\alpha}{\overline{X} - \mu_0} \right)^2 = n = 10.82 \qquad (4)$$

Then, a sample of size 11 would have been sufficient to reject (2) with a sample mean of 8.5mm. A more interesting result is produced by an n-vales plot (see Figure 1) in which the pairs (α, n) derived from equation (4) are plotted. The *n*-values line identifies the minimum sample size required (treating *n* as continuous) to reject the null hypothesis at level of significance alpha. Note that the line relating alpha and the sample size is downward sloping, indicating the well known fact that rejecting the null hypothesis at a low alpha requires a larger sample size than rejecting with a higher alpha.

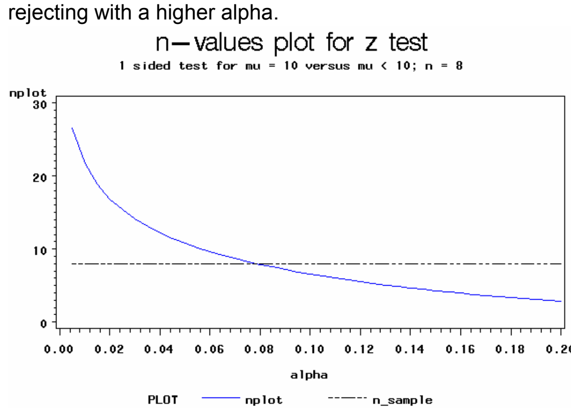

Figure 1: N-Values Plot For One Sided Z-Test Example

The macro that produces Figure 1 also produces the output below, which for reference outputs the test statistic (-1.41421) and the test p-value (0.07865). The variable n_size identifies the sample size to reject the null hypothesis at a specified level of significance. For example, rejection at 0.025 would require a sample of size 16.

n-values plot for z test
1 sided test for mu = 10 versus mu < 10; n = 8
05:42 Wednesday,

Obs	alpha	nplot	n_size	t_alpha	teststat	pvalue
1	0.240	1.9955	2	0.74594	-1.41421	0.078650
2	0.245	1.9061	2	0.72842	-1.41421	0.078650
3	0.250	1.8197	2	0.71114	-1.41421	0.078650
4	0.195	2.9558	3	0.91627	-1.41421	0.078650
5	0.200	2.8333	3	0.89603	-1.41421	0.078650
6	0.205	2.7152	3	0.87616	-1.41421	0.078650
7	0.210	2.6013	3	0.85664	-1.41421	0.078650
8	0.215	2.4913	3	0.83745	-1.41421	0.078650
9	0.220	2.3851	3	0.81858	-1.41421	0.078650
10	0.225	2.2826	3	0.80000	-1.41421	0.078650
11	0.230	2.1836	3	0.78172	-1.41421	0.078650
12	0.235	2.0879	3	0.76370	-1.41421	0.078650
13	0.160	3.9558	4	1.07029	-1.41421	0.078650
37	0.075	8.2890	9	1.61659	-1.41421	0.078650
38	0.060	9.6693	10	1.77021	-1.41421	0.078650
39	0.065	9.1700	10	1.71532	-1.41421	0.078650
40	0.050	10.8222	11	1.89458	-1.41421	0.078650
41	0.055	10.2169	11	1.82966	-1.41421	0.078650
42	0.045	11.4975	12	1.96615	-1.41421	0.078650
43	0.040	12.2596	13	2.04601	-1.41421	0.078650
44	0.035	13.1321	14	2.13645	-1.41421	0.078650
45	0.030	14.1495	15	2.24088	-1.41421	0.078650
46	0.025	15.3658	16	2.36462	-1.41421	0.078650
47	0.020	16.8715	17	2.51675	-1.41421	0.078650
48	0.015	18.8372	19	2.71457	-1.41421	0.078650

Interpretation of the n-values plot

Classical statistical decision theory suggests choosing alpha prior to performing the analysis. Typically today researchers employ *p*-values. When the observed outcome of an experiment favors the alternative, but does not indicate statistical significance, the n-values plot expresses the outcome of a mind experiment in which one supposes that new data consistent with what has been observed can be generated. The restriction on the mind experiment is that the statistic on which the test is based remains invariant. That is, in this example, the sample mean is still 8.5mm but the interpretation of this observed value is conditioned on the sample size.

In contrasts with power studies (curves) the *n*-values plot is based on the null-distribution (whereas the latter is based on various non-null distributions).

The n-values plot quantifies statements such as: "The results were nearly significant" by supplying a sample size at which significance would, ceteris paribus, lead to rejection. This approach is consistent with current analyses in which models are ranked by p-values.

Two sided z-test

Many times, statistical tests are conducted to see if the observed mean is different (either greater than or less than) the hypothesized mean. The two-sided hypothesis for this case is

$$H_0 : \mu = \mu_0 \qquad (5)$$
$$H_A : \mu \neq \mu_0$$

In this case, the z-statistic is still calculated using the formula (3). However, the p-value changes since this z-statistic is no longer being compared to z_α but to $z_{\alpha/2}$. For example, suppose that in the same study of average widget diameter, the experimenter wishes to test the hypothesis that the average diameter is 10mm versus the alternative that the average diameter is not 10mm, the hypothesis becomes

$$H_0 : \mu = 10 \qquad (6)$$
$$H_A : \mu \neq 10$$

The z-statistics remains –1.414 as calculated is (3). However, the p-value now changed to .1573. Again, the null hypothesis will not be rejected at the α=.05 level of significance since the p-value is not less than alpha (or the z statistics is not less than $-z_{\alpha/2} = -1.96$). To determine how large an n would be required to reject the null hypothesis in a two tailed test, equation (4) changes to

$$\left(\frac{\sigma z_{\alpha/2}}{\bar{X} - \mu_0}\right)^2 = n = 15.36 \qquad (7)$$

Thus showing that a sample of size 16 would be sufficiently large to reject the null hypothesis for this two-tailed test. Note that the only difference between equation (4) and equation (7) is the z-value used in the computation. Because $z_{\alpha/2}$ is always smaller than z_α, the n-value required for the two-tailed test is always larger than that required for the one tailed test. A plot of the n-values shows how large of an n is required for the test to be significant at a given alpha.

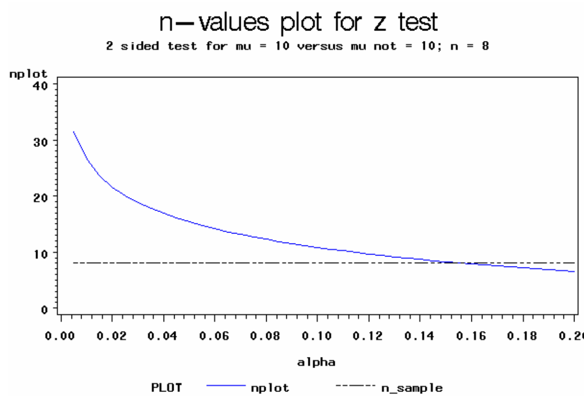

Figure 2: N-Values Plot For Two Sided Z-Test Example

ONE SAMPLE *t*-TEST

In situations where the population is normal, but the population variance (σ^2) is unknown, a z-test no longer an appropriate test for the equality of the mean to some target value. However, a t-test based on s^2 (the sample variance) is appropriate. The t-test statistic depends on the number degrees of freedom (n-1) as well as the level of significance.

One sided t-test

Returning to the previous example, suppose the population variance of widgets is unknown. The experimenter selects a sample of n=8 widgets. He calculates the sample mean and standard deviation of these 8 widgets to be 8.5 and 3, respectively. A t-test is now appropriate. Suppose the hypotheses are the same as in (2). The *t* statistic is now calculated using the formula

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{(\bar{X} - \mu_0)\sqrt{n}}{s} = \frac{(8.5 - 10)\sqrt{8}}{3} = -1.414 \qquad (8)$$

The test statistic is (coincidentally) the same as in the z-test; however, the p-value now becomes .1001. At the α=.05 level of significance, there is not enough evidence to support the alternative hypothesis. Equivalently, $t = -1.414$ is not less than $t_{.05,7} = -1.895$. To determine how large a sample would have been necessary for a sample mean of 8.5 to be significant, equation (4) becomes

$$\left(\frac{s t_\alpha}{\bar{X} - \mu_0}\right)^2 = n = 14.364 \qquad (9)$$

So a sample of size 15 would be sufficient for the findings to be significant at the .05 level (provided the same sample mean and variance were unchanged). An n-values plot of the sample size versus the significance level shows the sample size required for a hypothesis to be significant at a given alpha.

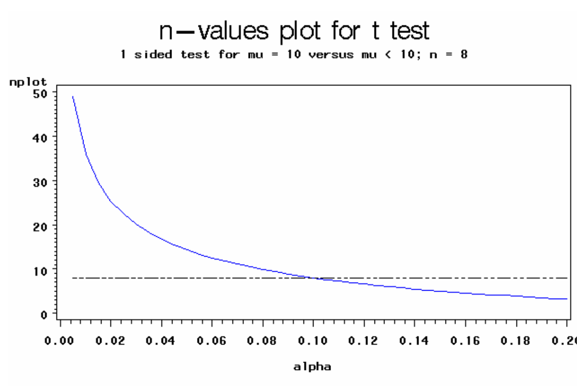

Figure 3: N-Values Plot For One Sided T-Test Example

More detail can be found in the tabular output, which also

n-values plot for t test
1 sided test for mu = 10 versus mu < 10; n = 8
05:42 Wednesday,

Obs	alpha	nplot	n_size	t_alpha	teststat	pvalue
1	0.210	2.9353	3	0.85664	-1.41421	0.10010
2	0.215	2.8053	3	0.83745	-1.41421	0.10010
3	0.220	2.6803	3	0.81858	-1.41421	0.10010
4	0.225	2.5600	3	0.80000	-1.41421	0.10010
5	0.230	2.4443	3	0.78172	-1.41421	0.10010
6	0.235	2.3329	3	0.76370	-1.41421	0.10010
7	0.240	2.2257	3	0.74594	-1.41421	0.10010
8	0.245	2.1224	3	0.72842	-1.41421	0.10010
9	0.250	2.0229	3	0.71114	-1.41421	0.10010
10	0.180	3.8371	4	0.97942	-1.41421	0.10010
11	0.185	3.6706	4	0.95794	-1.41421	0.10010
12	0.190	3.5111	4	0.93690	-1.41421	0.10010
13	0.195	3.3582	4	0.91627	-1.41421	0.10010
14	0.200	3.2115	4	0.89603	-1.41421	0.10010
15	0.205	3.0706	4	0.87616	-1.41421	0.10010
16	0.155	4.7909	5	1.09440	-1.41421	0.10010
17	0.160	4.5821	5	1.07029	-1.41421	0.10010
35	0.085	9.3570	10	1.52946	-1.41421	0.10010
36	0.075	10.4535	11	1.61659	-1.41421	0.10010
37	0.065	11.7693	12	1.71532	-1.41421	0.10010
38	0.070	11.0795	12	1.66430	-1.41421	0.10010
39	0.060	12.5345	13	1.77021	-1.41421	0.10010
40	0.055	13.3906	14	1.82966	-1.41421	0.10010
41	0.050	14.3577	15	1.89458	-1.41421	0.10010
42	0.045	15.4630	16	1.96615	-1.41421	0.10010
43	0.040	16.7446	17	2.04601	-1.41421	0.10010

shows that a sample of 15 would be sufficient to reject the hypothesis that the population mean is 10mm.

Two sided t-test

Similarly to the z-test, for the two sided alternative equation (9)

$$\left(\frac{s t_{\alpha/2}}{\bar{X} - \mu_0} \right)^2 = n \qquad (10)$$

This formula yields an n of 22.37 for our example, meaning that for the observed sample mean of 8.5 to be statistically different than the hypothesized mean of 10, a sample of 23 would be required as seen in the n-values plot.

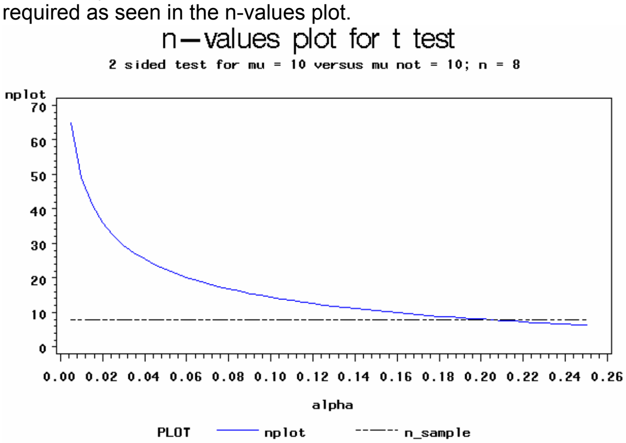

Figure 4: N-Values Plot For Two Sided T-Test Example

From the graph (Figure 4) the sample size required to reject appears to be around 22 or 23. The tabular output reveals that 23 widgets would be required.

n-values plot for t test
2 sided test for mu = 10 versus mu not = 10; n = 8
05:42 Wednesday,

Obs	alpha	nplot	n_size	t_alpha	teststat	pvalue
1	0.230	6.9152	7	1.31484	-1.41421	0.20020
2	0.235	6.7526	7	1.29928	-1.41421	0.20020
3	0.240	6.5948	7	1.28401	-1.41421	0.20020
4	0.245	6.4416	7	1.26902	-1.41421	0.20020
5	0.250	6.2929	7	1.25428	-1.41421	0.20020
6	0.205	7.8104	8	1.39736	-1.41421	0.20020
7	0.210	7.6195	8	1.38017	-1.41421	0.20020
8	0.215	7.4348	8	1.36334	-1.41421	0.20020
9	0.220	7.2560	8	1.34685	-1.41421	0.20020
10	0.225	7.0829	8	1.33069	-1.41421	0.20020
11	0.180	8.8730	9	1.48938	-1.41421	0.20020
12	0.185	8.6446	9	1.47009	-1.41421	0.20020
34	0.080	16.7446	17	2.04601	-1.41421	0.20020
35	0.085	16.0789	17	2.00492	-1.41421	0.20020
36	0.075	17.4679	18	2.08973	-1.41421	0.20020
37	0.070	18.2577	19	2.13645	-1.41421	0.20020
38	0.065	19.1256	20	2.18664	-1.41421	0.20020
39	0.060	20.0862	21	2.24088	-1.41421	0.20020
40	0.055	21.1580	22	2.29989	-1.41421	0.20020
41	0.050	22.3658	23	2.36462	-1.41421	0.20020
42	0.045	23.7430	24	2.43634	-1.41421	0.20020
43	0.040	25.3362	26	2.51675	-1.41421	0.20020
44	0.035	27.2128	28	2.60830	-1.41421	0.20020
45	0.030	29.4756	30	2.71457	-1.41421	0.20020

Paired T-Test

For the paired t-test use the differences and this reduces to the previous one sample cases.

SAS MACRO FOR ONE SAMPLE TESTS ON THE MEAN

The actual macro statement to produce Figure 4 is:

```
%nvalues1(n = 8, x_bar = 8.5, sigmaX = 3,
       nullmean = 10, sided =2,testtype =2);
```

The macro identifies the sample size, the sample mean, the known (or estimated) standard deviation, the hypothesized mean, a coded sided-variable that identifies the test as less than, equal to, or greater than, and a coded testtype variable that specifies the test as a z-test or t-test.

```
%macro nvalues1(
       n = ,
       x_bar =,
       sigmaX =,
       nullmean = ,
       sided =,  /* 1 = <, 2 not = , 3 = > */
       testtype = /* 1 = z-test, 2 = t-test */)
;
data plotdat;
If &sided = 1 and &x_bar > &nullmean
       then put '****sample mean does not favor
       alternative: plot invalid************';
If &sided = 3 and &x_bar < &nullmean
       then put '*****sample mean does not favor
       alternative: plot invalid************';
do alpha = .005 to .20 by .005;

%local sidenum;
%local sign;


%if &sided=1 or &sided=3 %then %let sidenum=1;
%if &sided=2 %then %let sidenum=2;

%if &sided=1 %then %let sign=<;
%if &sided=2 %then %let sign= not =;
%if &sided=3 %then %let sign=>;

%if &testtype=1 %then %let testsym = z;
%if &testtype=2 %then %let testsym = t;

teststat = (&x_bar-&nullmean)*sqrt(&n)/&sigmax;
if &testtype=1 and &sided = 1
       then pvalue = probnorm(teststat);
if &testtype=1 and &sided = 3
       then pvalue = 1-probnorm(teststat);
if &testtype=1 and &sided = 2
```

```
        then pvalue = 2*min(probnorm(teststat),1-
        probnorm(teststat));

if &testtype=2 and &sided = 1
        then pvalue = probt(teststat,&n-1);
if &testtype=2 and &sided = 3
        then pvalue = 1-probt(teststat,&n-1);
if &testtype=2 and &sided = 2
        then pvalue = 2*min(probt(teststat,&n-
        1),1-probt(teststat,&n-1));
n_sample = &n;
dft = n_sample-1;

if &sided = 2 then z_alpha = probit(alpha/2);
        else z_alpha = probit(alpha);
        if &sided = 2 then t_alpha = tinv(1-
        alpha/2,dft);
        else t_alpha = tinv(1-alpha,dft);
If &testtype = 1 then nplot =
((&sigmaX*z_alpha)/(&x_bar-&nullmean))**2;
        else if &testtype = 2
                then nplot =
((&sigmaX*t_alpha)/(&x_bar-&nullmean))**2;
        else nplot = 0;
n_size = floor(nplot)+1;
output;
end;
proc gplot;
        plot nplot*alpha = 1 n_sample*alpha =2 /
overlay legend;
        symbol1 c=BLUE,i=join, l=1, v=none;
         symbol2 c=BLACK, i=join, l=14, v=none;
title "n-values plot for &testsym test";
title2 "&sidenum sided test for mu = &nullmean
versus mu &sign &nullmean; n = &n";
run;

proc print data = plotdat;
        var alpha nplot n_size t_alpha teststat
        pvalue;
title " n-values plot for &testsym test";
title2 "&sidenum sided test for mu = &nullmean
versus mu &sign &nullmean; n = &n";

run;
%mend nvalues1;
```

ONE SAMPLE TEST FOR A PROPORTION

When the parameter under study is a proportion an exact test based on the binomial distribution or a large sample z-test can be performed to test whether the proportion is equal to some target value (p). We will present only the latter. The decision is based on the observed proportion ($\hat{p}$). In this case, if performing a one sided test, the hypotheses in (1) change to

$$H_0 : p = p_0$$
$$H_A : p < p_0 \qquad (11)$$

And if performing a two sided test, the hypotheses in (5) becomes

$$H_0 : p = p_0$$
$$H_A : p \neq p_0 \qquad (12)$$

To determine how large of a sample must be used for the findings to show a significant difference, the formula for a one sided test (4) changes to

$$\left(\frac{\sqrt{pq}\,z_\alpha}{\hat{p}-p}\right)^2 = n \qquad (13)$$

because the for variance of the sample proportion is proportional to pq, where *p* is the population proportion and *q* = *1-p*.

If the test is a two-sided test, then formula (13) becomes

$$\left(\frac{\sqrt{pq}\,z_{\alpha/2}}{\hat{p}-p}\right)^2 = n \qquad (14)$$

Note, that again the only difference between the formulas for a one sided and two sided test, (13) and (14), is the z-value (tail area)

For example, suppose that when studying the widgets, our experimenter desires to test the hypothesis that more then 20% of widgets have diameters less than 10mm. He takes a sample of 30 widgets, and finds that 9 are less than 10mm in diameter. The hypotheses are

$$H_0 : p = .20$$
$$H_A : p > .20$$

The z-value associated with this test is computed by

$$z = \frac{\hat{p}-p}{\sqrt{pq}/\sqrt{n}} = \frac{(\hat{p}-p)\sqrt{n}}{\sqrt{pq}} = \frac{(\frac{9}{30}-.20)\sqrt{30}}{\sqrt{(.20)(.80)}} = 1.369$$

At the .05 level of significance, the result is not significant, since the associated p-value of .08545 is not less than .05. To determine how large an *n* would have been necessary to obtain a significant result, we use formula (13).

$$\left(\frac{z_\alpha\sqrt{pq}}{\bar{X}-\mu_0}\right)^2 = \left(\frac{1.645\sqrt{.20(.80)}}{\frac{9}{30}-.20}\right)^2 = 43.29$$

This (see also the tabular output below) shows that a sample of 44 would be necessary for the observed proportion to lead to rejection of the null hypothesis. The n-values plot (see Figure 5) shows how large of a sample would be necessary to obtain significance at a given alpha value.

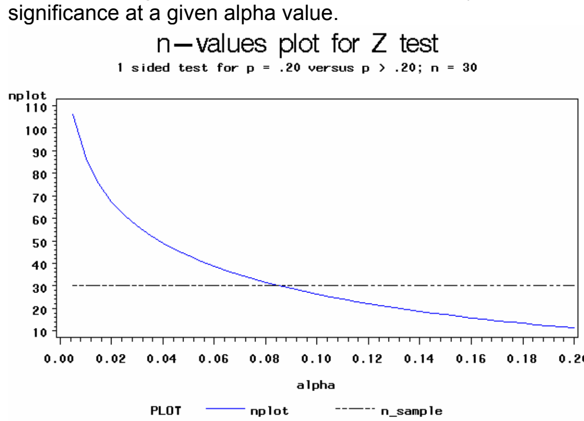

Figure 5: n-Values Plot for Proportion z-Test Example

The non-graphical output can be used to focus in on details:

```
                 n-values plot for z test
 1 sided test for mu = .20 versus p > .20; n = 30
                                                  05:42
```

Obs	alpha	nplot	n_size	z_alpha
1	0.005	106.158	107	-2.57583
2	0.010	86.590	87	-2.32635
3	0.015	75.349	76	-2.17009
4	0.020	67.486	68	-2.05375
5	0.025	61.463	62	-1.95996
6	0.030	56.598	57	-1.88079
7	0.035	52.528	53	-1.81191
8	0.040	49.038	50	-1.75069
9	0.045	45.990	46	-1.69540
10	0.050	43.289	44	-1.64485
11	0.055	40.868	41	-1.59819
12	0.060	38.677	39	-1.55477
13	0.065	36.680	37	-1.51410
14	0.070	34.847	35	-1.47579
15	0.075	33.156	34	-1.43953
16	0.080	31.588	32	-1.40507
17	0.085	30.127	31	-1.37220

Macro for test on a Single Proportion

This macro statement produces the above output:

```
%nvaluesp1(n = 30, p_hat = .30,
          nullprop = .20, sided =3 );
The macro code follows:
%macro nvaluesp1(
     n = ,
     p_hat =,/*sample proportion*/
     nullprop = , /*hypothesized proportion*/
     sided =,  /* 1 = <, 2 not = , 3 = > */
     ) ;
data plotdat;
If &sided = 1 and &p_hat > &nullprop
     then put '****sample proportion does
          not favor alternative: plot
          invalid***********';
If &sided = 3 and &p_hat < &nullprop
     then put '*****sample proportion does
          not favor alternative: plot
          invalid***********';
do alpha = .005 to .20 by .005;

     n_sample = &n;

     p= &nullprop;
     q = 1-p;
     variance=p*q;
     stdev=sqrt(variance);
     %local sidenum;
     %local sign;

     %if &sided=1 or &sided=3
          %then %let sidenum=1;
     %if &sided=2
          %then %let sidenum=2;

     %if &sided=1
          %then %let sign=<;
     %if &sided=2
          %then %let sign= not =;
     %if &sided=3
          %then %let sign=>;
if &sided = 2 then z_alpha = probit(alpha/2);
          else z_alpha = probit(alpha);
nplot = ((stdev*z_alpha)/(&p_hat-p))**2;

n_size = floor(nplot)+1;
output;
end;
proc gplot;
          plot nplot*alpha = 1 n_sample*alpha =2 /
overlay legend;
          symbol1 c=BLUE,i=join, l=1, v=none;
           symbol2 c=BLACK, i=join, l=14, v=none;
title "n-values plot for Z test";
title2 "&sidenum sided test for p = &nullprop
versus p &sign &nullprop; n = &n";
run;

proc print data = plotdat;
          var alpha nplot n_size z_alpha;
title " n-values plot for z test";
title2 "&sidenum sided test for mu = &nullprop
versus p &sign &nullprop; n = &n";
run;
    %mend nvaluesp1;
```

TESTS ON SEVERAL MEANS

Suppose that one is testing the hypothesis that the means of three populations are equal, and that the test is to be based on three independent samples of size n = 5 (for each treatment). The data below illustrates the idea.

treat 1	treat 2	treat 3
5	11	5
7	12	6
4	6	6
6	5	5
9	10	8

An Anova F-test yields

Table 1: ANOVA F-Table (sample size 5)

SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
treat 1	5	31	6.2	3.7		
treat 2	5	44	8.8	9.7		
treat 3	5	30	6	1.5		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	24.4	2	12.2000	2.4564	0.1276	3.8853
Within Groups	59.6	12	4.9667			
Total	84	14				

The p-value is too large to reject the null hypothesis at alpha = 0.05. Suppose that the sample were twice as large and in fact consisted of a replicate of the original sample.

Treat 1	treat 2	treat 3
5	11	5
7	12	6
4	6	6
6	5	5
9	10	8
5	11	5

7	12	6
4	6	6
6	5	5
9	10	8

The ANOVA F-Test on this data set produces

Table 2: ANOVA F-Table (sample size 10

SUMMARY

Groups	Count	Sum	Average	Variance
treat 1	10	62	6.2	3.2889
treat 2	10	88	8.8	8.6222
treat 3	10	60	6	1.3333

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	48.8	2	24.4000	5.5268	0.0097	3.3541
Within Groups	119.2	27	4.4148			
Total	168	29				

Of course, the averages are the same. The p-value is less than 0.05, and hence, this mind experiment suggests that a sample of size 10 (instead of 5) would lead to rejection of the null hypothesis. An n-values plot can be generated based on this reasoning. Identify this as the Replicate Sample Approach (SRA). A simple mathematical relationship connects the two F statistics

$$F_{tn} = F_n\left(\frac{tn-1}{n-1}\right) \qquad (15)$$

$$F_{10} = F_5\left(\frac{2(5)-1}{5-1}\right) = 2.4564\left(\frac{9}{4}\right) = 5.5268$$

where the subscript on the F statistic refers to the sample size and *t* is the multiple by which the sample size is altered.

SAS ANOVA-F Macro Output

The SAS macro statement

```
%nvalanova(n = 5, k=3,F = 2.4564);
```

produces the graph in Figure 6 and the output below.

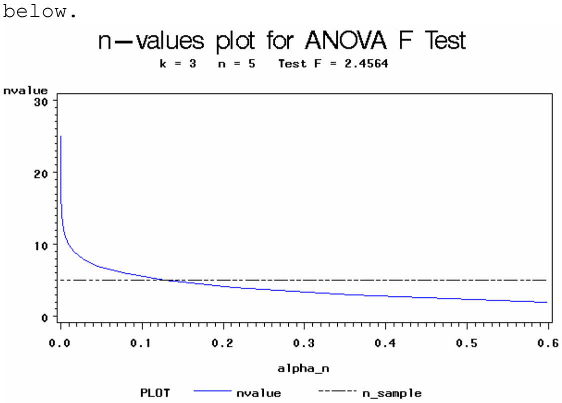

Figure 6: N-Values Plot For ANOVA F-test Example

Observe that the graph indicates that a sample of size about 7 would be sufficient. The tabular output indicates that for n=7 the corresponding p-value is 0.04557, which is significant at alpha = 0.05. Note that for n =10 the F-statistic is 5.5269.

```
    n-values plot for ANOVA F Test
 k = 3    n = 5    Test F = 2.456376

 Obs     nvalue        F        alpha_n

   1       25       14.7384     0.00000
   2       24       14.1243     0.00001
   3       23       13.5102     0.00001
   4       22       12.8961     0.00002
   5       21       12.2820     0.00003
   6       20       11.6679     0.00006
   7       19       11.0538     0.00009
   8       18       10.4397     0.00016
   9       17        9.8256     0.00026
  10       16        9.2115     0.00044
  11       15        8.5974     0.00074
  12       14        7.9833     0.00124
  13       13        7.3692     0.00208
  14       12        6.7551     0.00347
  15       11        6.1410     0.00581
  16       10        5.5269     0.00973
  17        9        4.9128     0.01628
  18        8        4.2987     0.02724
  19        7        3.6846     0.04557
  20        6        3.0705     0.07625
  21        5        2.4564     0.12758
  22        4        1.8423     0.21347
  23        3        1.2282     0.35719
  24        2        0.6141     0.59765
```

ANOVA-F n-Values Macro

The macro statement that produced the n-values plot seen in Figure 6 is

```
%nvalanova(n = 5, k=3,F = 2.4564);
```

This macro requires the sample size used in the study, the number of treatment levels, and the F-statistic computed for the test.

```
/****************************
n-values plot for multiple mean comparisons
using ANOVA
*****************************/

%macro nvalanova(
        n = ,
        k = ,
        F =
        ) ;
data plotdat;
numdf = &k-1;

mult = 5;
nmult = mult*&n;
do nvalue = 2 to nmult;
        n_sample = &n;
        denomdf = (nvalue-1)*&k;
        F = &F*(nvalue - 1)/(n_sample-1);
        alpha_n =1- probf(f,numdf,denomdf);
output;
end;
proc sort data=plotdat; by alpha_n;
proc print; var nvalue F alpha_n;
title "n-values plot for ANOVA F Test";
title2 "k = &k   n = &n   Test F = &F  ";
proc gplot;
        plot nvalue*alpha_n =1 n_sample*alpha_n =
2 / overlay legend;
        symbol1 c=BLUE,i=join, l=1,
                v=none;
```

```
        symbol2 c=BLACK, i=join, l=14, v=none;
title "n-values plot for ANOVA F Test";
title2 "k = &k   n = &n   Test F = &F  ";
run;
   %mend nvalanova;
```

TABLE TESTS

Tables arise in the analysis of categorical data. Only 2 by 2 tables will be considered in this paper. The ideas extend naturally to 2 by S and other tables. Consider a test for the effectiveness of a DRUG with the following results.

	Fav	UnFav	SUMS	%Fav
Test	23	37	60	38.33%
Placebo	16	48	64	25.00%
SUMS	39	85	124	

The design is one in which subjects are randomized between placebo and test groups. One method of analysis is to calculate a Chi-squared statistic. We will look at the statistic Q (the one identified as the Mantel-Haenszel Chi-Square in SAS output from PROC FREQ). See for example Stokes, Davis, and Koch (2000) for a complete discussion). Denoting the cell counts by n_{ij}, the row sums by r_i and the column sums by c_j, the statistic Q is

$$Q = \frac{(n_{11} - m_{11})^2}{v_{11}} \quad (16)$$

where the expected cell count

$$m_{ij} = \frac{r_i c_j}{n} \quad (17)$$

and the variance is

$$v_{ij} = \frac{r_1 r_2 c_1 c_2}{n^2 (n-1)} \quad (18)$$

Note that

$$n = r_1 + r_2 = c_1 + c_2$$

is the (total) sample size, and the row sums represent the number of test and placebo subjects respectively. Given the study results the hypothesis that the response rates are different for the two groups cannot be rejected at the 5% level since Q = 2.533 (p-value = .1115). Note that the observed rates for favorable response are 38.33% for the treatment group and 25% for the placebo group.

The question is: were these rates to hold for a larger sample, how large would that sample have to be in order to arrive at a significant result? The notion of larger sample needs to be pinned down. One notion would be to increase on the number of subjects in the test group. Another would be to increase the total number of subjects. First, we will examine increasing the test subjects only. Letting the proportion responding favorably remain constant, and changing r_1 to r_1^* the cell counts become

$$n_{11}^* = \left(\frac{n_{11}}{r_1}\right) r_1^* \qquad n_{12}^* = \left(\frac{n_{12}}{r_1}\right) r_1^* \quad (19)$$

Hence,

$$c_1^* = n_{11}^* + n_{21} \qquad c_2^* = n_{12}^* + n_{22} \quad (20)$$

Then the calculation of Q follows from:

$$m_{11}^* = \frac{r_1^* c_1^*}{n^*}$$

$$v_{11}^* = \frac{r_1^* r_2 c_1^* c_2^*}{(n^*)^2 (n^* - 1)} \quad (21)$$

$$Q^* = \frac{(n_{11}^* - m_{11}^*)^2}{v_{11}^*}$$

The p-value (alpha) is given by

$$\text{alpha\_n} = P(x > Q^* \mid x \sim \chi^2(1)) \quad (22)$$

The n-values plot can take two forms, depending on whether the vertical axis is the sample size for the test group or the total sample size; that is, the pairs:

$$(\text{alpha\_n}, r_1^*)$$

or

$$(\text{alpha\_n}, n^*)$$

are plotted. The circumstance in which both the test and placebo groups increase in size can be handled by allocating each increase in n between the two groups according to the proportions arising in the study(in this example, 60/124 for the test group). Alternatively any proportion of interest can be used. Note that when both test and placebo groups are increased all of the cell counts are altered (proportionately). The Pearson Chi-squared is

$$Q_p = \left(\frac{n}{n-1} Q\right) \quad (23)$$

so an n-values plot for this test statistics is easy to produce.

Tables SAS Macro Output

The macro statement that produced the n-values plot seen in figure 10 is

```
%nvaltables(n11=23,n12=37,n21=16,n22=48,
       increase=2,stat=1,prop=0);
```

In this macro statement, the four table values are identified as the n_{ij}. The increase variable indicates whether to increase the n-value of the test group only or both groups. The stat variable indicates whether the Q or Q_p statistic should be used as the test statistic. Lastly, the proportion of n to be allocated to the test and placebo groups is needed (it will default to the proportion previously in the study if set to zero).

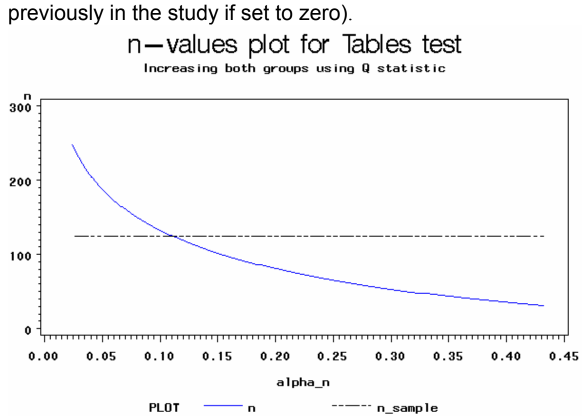

Figure 7: N-Values Plot for Table Test

From the plot (see Figure 7) it is apparent that a substantial increase in the sample size would be needed for the observed difference in the proportion to be significant at the 5% level. The non-graphical output clarifies the analysis. Note that about 188 subjects would have to be enrolled in the study to reject the no effect hypothesis.

n-values plot for Tables test
Increasing both groups using Q statistic
05:42 Wednesday,

Obs	n	n11_new	n12_new	n21_new	n22_new	alpha_n
49	90.52	16.79	27.01	11.68	35.04	0.17454
50	91.76	17.02	27.38	11.84	35.52	0.17158
51	93.00	17.25	27.75	12.00	36.00	0.16869
52	94.24	17.48	28.12	12.16	36.48	0.16584
53	95.48	17.71	28.49	12.32	36.96	0.16306
54	96.72	17.94	28.86	12.48	37.44	0.16032
55	97.96	18.17	29.23	12.64	37.92	0.15764
56	99.20	18.40	29.60	12.80	38.40	0.15501
57	100.44	18.63	29.97	12.96	38.88	0.15243
58	101.68	18.86	30.34	13.12	39.36	0.14989
59	102.92	19.09	30.71	13.28	39.84	0.14741
60	104.16	19.32	31.08	13.44	40.32	0.14497
61	105.40	19.55	31.45	13.60	40.80	0.14258
75	122.76	22.77	36.63	15.84	47.52	0.11331
76	124.00	23.00	37.00	16.00	48.00	0.11149
77	125.24	23.23	37.37	16.16	48.48	0.10970
78	126.48	23.46	37.74	16.32	48.96	0.10795
79	127.72	23.69	38.11	16.48	49.44	0.10622
80	128.96	23.92	38.48	16.64	49.92	0.10452
81	130.20	24.15	38.85	16.80	50.40	0.10286
124	183.52	34.04	54.76	23.68	71.04	0.052534
125	184.76	34.27	55.13	23.84	71.52	0.051738
126	186.00	34.50	55.50	24.00	72.00	0.050955
127	187.24	34.73	55.87	24.16	72.48	0.050184
128	188.48	34.96	56.24	24.32	72.96	0.049426
129	189.72	35.19	56.61	24.48	73.44	0.048680
130	190.96	35.42	56.98	24.64	73.92	0.047945
131	192.20	35.65	57.35	24.80	74.40	0.047223
132	193.44	35.88	57.72	24.96	74.88	0.046512

It is interesting to compare this approach to one in which only the test group sample size is increased. The macro becomes:

```
%nvaltables(n11=23,n12=37,n21=16,n22=48,increase
=1,stat=1,prop=0);

producing the output below.
```

Obs	n	n11_new	n12_new	n21_new	n22_new	alpha_n
23	120.4	21.62	34.78	16	48	0.11675
24	121.6	22.08	35.52	16	48	0.11493
25	122.8	22.54	36.26	16	48	0.11318
26	124.0	23.00	37.00	16	48	0.11149
27	125.2	23.46	37.74	16	48	0.10986
28	126.4	23.92	38.48	16	48	0.10830
157	281.2	83.26	133.94	16	48	0.050215
158	282.4	83.72	134.68	16	48	0.050088
159	283.6	84.18	135.42	16	48	0.049964
160	284.8	84.64	136.16	16	48	0.049840
161	286.0	85.10	136.90	16	48	0.049718
162	287.2	85.56	137.64	16	48	0.049597
163	288.4	86.02	138.38	16	48	0.049478
164	289.6	86.48	139.12	16	48	0.049360
165	290.8	86.94	139.86	16	48	0.049243

In this case the sample size needed is about 283. This is an increase in 159 test subjects. On the other hand, increasing both the test and placebo groups from a test-group n of 60 to a test-group n of 91 and the placebo-group from 64 to 97, for an crease of 65 or so, is associated with a significant result..

Tables SAS Macro

```
/*********************************
n-values plot for tables test
**********************************/

%macro nvaltables(
        n11 = ,
        n12 = ,
        n21 = ,
        n22 = ,
        increase = , /* 1=test group , 2=both
                        groups */
        stat = , /* 1=Q , 2=QP */
        prop = /*proportion of addition subjects
                added = 0 to use study prop. or
                if entered 1 for increase     */
        ) ;

data plotdat;
n_sample = &n11 + &n12+ &n21 + &n22;
        c1 = &n11 + &n21;
        c2 = &n12 + &n22;
        r1 = &n11 + &n12;
        r2 = &n21 + &n22;
%local inc_name;
%local stat_name;

%if &increase=1 %then %let inc_name=test group
only;
%if &increase=2 %then %let inc_name=both groups;
%if &stat=1 %then %let stat_name = Q;
%if &stat=2 %then %let stat_name=Qp;

prop2=&prop;
if &prop =0 then prop2=.5;

do  npct_inc = -50 to 300 by 2;

        if &increase=1 then
                n11_new=&n11*(1+npct_inc/100);
        if &increase=1 then
                n12_new=&n12*(1+npct_inc/100);
        if &increase=1 then n21_new=&n21;
        if &increase=1 then n22_new=&n22;
        if &increase=1 then
                r1_new=n11_new+n12_new;
        if &increase=1 then r2_new=r2;

        if &increase=2 then n11_new =
                &n11*(1+npct_inc/100)*prop2;
        if &increase=2 then n12_new =
                &n12*(1+npct_inc/100)*prop2;
        if &increase=2 then
                n21_new=&n21*(1+npct_inc/100)*(1-
        prop2);
        if &increase=2 then n22_new
                =&n22*(1+npct_inc/100)*(1-prop2);
        if &increase=2 then
                r1_new=n11_new+n12_new;
        if &increase=2 then
                r2_new=n21_new+n22_new;

        c1_new = n11_new + n21_new;
        c2_new = n12_new + n22_new;
        n=c1_new+c2_new;

        m11 = (r1_new*c1_new)/n;
        v11 =(r1_new*r2_new*c1_new*c2_new)
                /(n**2*(n-1));
        Q=(n11_new-m11)**2/v11;

        if &stat=1 then alpha_n=1-probchi(Q,1);
        if &stat=2 then alpha_n=1-probchi(n/(n-
                1)*Q,1);
output;
end;
proc print; var n n11_new n12_new n21_new
        n22_new alpha_n;
title "n-values plot for Tables test";
```

```
title2 "Increasing &inc_name using &stat_name
statistic";

proc gplot;
       plot n*alpha_n = 1 n_sample*alpha_n =2 /
overlay legend;
       symbol1 c=BLUE,i=join, l=1, v=none;
        symbol2 c=BLACK, i=join, l=14, v=none;
title "n-values plot for Tables test";
title2 "Increasing &inc_name using &stat_name
statistic";
run;
run;
   %mend nvaltables;
```

OTHER TESTS

The methods offered in this paper can be extended to an assortment of tests.

CONCLUSION

The n-values plots in this paper can be useful in data analysis. The most natural area of application is in circumstances in which the test statistics does not clear the hurdle of statistical significance but does at least suggest the existence of some effect of difference. The plot offers insight into the sample size required, ceteris paribus, to reach statistical significance. Along with other analyses, including, retrospective and other power studies, n-values plots can be used to plan additional studies and sampling.

REFERENCES

Categorical Data Analysis Using the SAS System, 2nd Edition, Stokes, M., Davis, C, and Koch, G.2000, SAS Institute Inc, Cary, N.C.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.
Contact the authors at:

Peter Wludyka
Associate Professor of Statistics
Director,
Center for Research and Consulting in Statistics
University of North Florida
Jacksonville, Florida
Work Phone: 904-620-1048
Fax: 904-620-2818
Email: pwludyka@unf.edu
Web: www.unf.edu/coas/math-stat/~pwludyka

Amy Cox
University of North Florida
Jacksonville, FL
Work Phone: 904-620-2819
Email: coxa0003@unf.edu