

Paper SD-159

Using SAS® to Create an Effect Size Resampling Distribution for a Statistical Test

Peter Wludyka, University of North Florida

ABSTRACT

One starts with data to perform a statistical test of a hypothesis. An effect size is associated with a particular test/sample and this effect size can be used to decide whether there is a clinically/operationally significant effect. Since the data (the sample) is usually all a researcher knows factually regarding the phenomenon under study, one can imagine that by sampling (resampling) with replacement from that original data that additional information about the hypothesis and phenomenon/study can be acquired. One way to acquire such information is to repeatedly resample for the original data set (using, for example, PROC SURVEYSELECT) and at each iteration (replication of the data set) perform the statistical test of interest and calculate the corresponding effect size. At the end of this stage one has R effect sizes (R is typically greater than 1,000), one for each performance of the statistical test. This effect size distribution can be presented in a histogram. Uses for this distribution and its relation to the p-value resampling distribution which was presented at SESUG 2014 will be explored.

INTRODUCTION

Typical of resampling (bootstrap) methods the basic idea is that the sample represents the population from which the sample arose; that is, one treats the sample as if it were a population. Given that a statistical test is performed on the sample there is an (estimated) effect size associated with that test. One might wonder how this effect size might change were a “different sample” selected from the same population. The surrogate for this “different sample” is a sample drawn with replacement from the initial sample. Repeated resampling and performance of the statistical test on each of these resamples gives rise to the effect size resampling distribution. The basic idea is described in the schematic (Figure 1) which shows the logical structure of the problem. A SAS implementation is somewhat different since rather than “looping” one creates a data set containing the “loop” information

The research question as well as the sampling methodology and nature of the measurements collected determine (influence) the choice of statistical test. The effect size resampling distribution describes tests results that might be associated with repeating the study with the same population.

This paper builds on ideas from a paper in SESUG 2014 “Using SAS to Create the p-value Resampling Distribution for a Statistical Test” and in this paper we will look at both distributions to see in particular the effects of changing sample sizes.

In this paper the general method for creating the effect size resampling distribution using SAS will be presented. Two examples will be presented which illustrate how one might create the effect size resampling distribution. For methods associated with both retrospective and prospective power estimates based on the p-value resampling distribution see the 2014 paper (Wludyka and Smotherman).

THE EFFECT SIZE AND P-VALUE RESAMPLING DISTRIBUTIONS

The following will be illustrated through the examples:

- Step1: Perform statistical test and calculate the (estimated) effect size and the p-value for reference.
- Step2: Resample from the original sample R times for each sample size of interest
- Step3: Calculate the effect size for each resample (replicate within sample size); do the same for the p-value if that is of interest.
- Step 4: Summarize and organize the effect sizes (p-values if wanted). Typically a histogram tells the story along with indications of key percentiles.

EXAMPLE 1: ONE SAMPLE T-TEST

We illustrate with a simple one sample problem with hypotheses:

$$H0: \mu_Y = 140 \quad (1)$$

$$H1: \mu_Y \neq 140.$$

Step1. The analysis is performed using the SAS program which follows in which the sample is a pseudo random sample of size $n=10$ from $Y \sim N(143, 7^2)$, a normal population with mean 143 and standard deviation 7. The actual (true but never known) effect size is $\frac{|\mu - \mu_0|}{\sigma} = \frac{143 - 140}{7} = 0.429$. The actual sample is in Output 1; the summary statistics and output from PROC TTEST are in Output 2. The estimated effect size is

$$d = \frac{|\bar{Y} - \mu_0|}{S} = \frac{143.26 - 140}{5.02} = 0.6482$$

which is a medium effect size (Cohen, 1992). Based on the p -value = 0.0706 ($t = 2.05$) one fails to reject the null hypothesis at the 5% level of significance. In decision theoretic terms the story is over and one concludes that there is not enough evidence to conclude the population mean is not 140. This test assumes (approximate) normality and under that assumption the power of the test is 44.9% (See Output 2, which describes the scenario for the power analysis from PROC POWER). This assessment of retrospective power is based on the notion that were the true population mean and standard deviation equal to the sample values (Output 2) then the power of a one sample t -test of hypothesis (1) is 44.9% assuming normality.

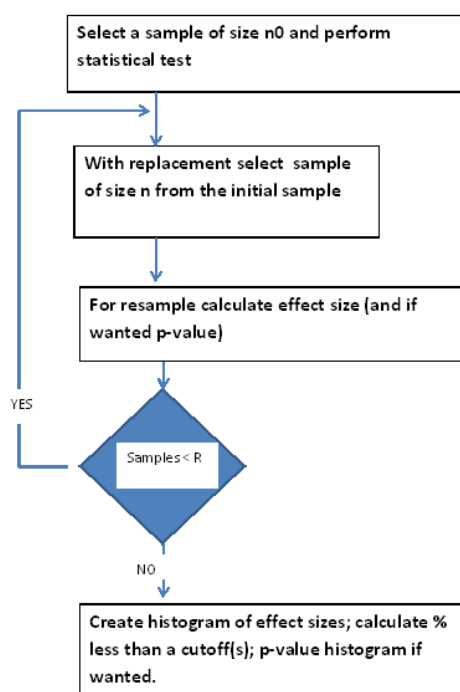


Figure 1. The Process of Finding the resampling distribution.

Obs	1	2	3	4	5	6	7	8	9	10
y	149.1	136.5	150.1	139.2	140.1	138.4	145.4	140.9	143.3	149.6

Output 1: From PROC Print Showing original Data Set.

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	143.3	5.0195	1.5873	136.5	150.1
Mean	95% CL Mean	Std Dev	95% CL Std Dev		
143.3	139.7	146.8	5.0195	3.4526	9.1637
DF	t Value	Pr > t			
9	2.05	0.0706			

Output 2: From PROC TTEST, Summary Statistics for the Original Sample in Output 1.

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Mean	3.2538
Standard Deviation	5.0195
Total Sample Size	10
Number of Sides	2
Null Mean	0
Alpha	0.05
Computed Power	0.449

Output 3: From a PROC POWER, Power of t-Test Based on Observed Statistics in Output2.

Step 2. Generate R resamples from the data in Output 1 for selected sample sizes. In this example there are R = 5000 resamples for each of n = 10, 15, and 20. See STEP 2 in the SAS program in which

```
proc surveyselect data=Ydata out=outbootBn10;
```

is used to generate 5000 resamples of size 10 and put them in the data set outbootBn10. In the code for that PROC `method= urs` indicates that sampling with replacement is to take place. The three PROC SURVEYSELECTs generate resamples for each of the three sample sizes.

Step 3: Calculate the estimated effect size for each of the samples (and p-value if wanted).

Step 3a: Use PROC MEANS to find the mean and standard deviation of each sample and put those in the data set ysummary, which in this case is indexed by ss (sample size) and replicate. This is sufficient for the simple example of a one sample t-test; however, in a more complex implementation other statistics will need to be calculated.

Step 3b: Use a DATA step to calculate the effect sizes (and p-values). The details in this step depend on the specifics of the test. For the one sample t-test the effect size is $|(\text{mean} - \text{hypothesized mean})|/(\text{standard deviation})$ and the p-value is based on the t-distribution (see SAS program for details).

Step 4: Summarize the effect size distribution. PROC CAPABILITY was used to create Figures 2, 3, and 4. COMPHIST creates the three stacked histograms seen in Figures 3 and 4.

Interpretation of the Effect Size Resampling Distribution

The inset in Figure 2 reports that there were R = 5000 resamples. The two vertical lines (technically lower and upper spec limits in the context of PROC CAPABILITY) are used to mark off medium and large effect sizes. LSL:Medium Effect corresponds to a medium effect size based on d. So, were the researcher to sample repeatedly from a

population similar to the one in Output 1 about 28.28% of the time the researcher would report an effect size less than a medium effect size of 0.5 (and 71.72% of the time an effect size greater than the medium effect cutoff). About 32.40% of the samples would show an effect size classified as large effect. Note that

$$\left| \frac{\bar{Y} - \mu_0}{S} \right| > 0.5$$

implies

$$|\bar{Y} - \mu_0| > 0.5S$$

That is, the sample mean differs from the hypothesized mean by more than $\frac{1}{2}$ an observed standard deviation for a medium effect size. For assessing clinical/operational significance the effect size is more meaningful than the p-value which is strongly influenced by the sample size. Note that in the one sample t-test the t statistic $|t| = d\sqrt{n}$. That is, the t statistic is proportional to the effect size with proportionality factor root n.

The role of sample size in the effect size resampling distribution can be seen in Figure 3. Note that the mean effect size over the 5000 resamples is almost identical for the three sample sizes. This mean is slightly higher than $d = 0.649$, the effect size from the original sample. Why?

As the sample size increases the probability of reporting an effect size of medium or more increases (71.72%, 76.22%, 80.76% for $n = 10, 15, 20$ respectively). Since the effect size for the initial sample is greater than 0.5, this makes sense and basically says that were one to take a sample of size $n = 20$ from a population like the one from which the initial sample were taken, then about 81% of the time one could report a medium or better effect size (and a large effect size 27.5% of the time).

Figure 3 can be used to see the effect of sample size on the likelihood of rejecting the null hypothesis. At the 5% level of significance the power based on the p-value resampling distribution is $(1 - .58)100\% = 42\%$, which is the percentage of p-values less than 0.05. For $n = 15$ the power increases to 68.4%; and for $n = 20$ the power increases to 85.1%. One can create a power curve based on these values. The inset for Figure 3 also contains power estimates for $\alpha = 0.1$. See Wludyka and Smotherman (2014) for details. Typically one assesses statistical significance first, so examining the p-value resampling distribution shows one likely scenarios when sampling from a population like the one represented in Output 1. In our example the observed p-value was 0.0706. Were this a preliminary study, one might examine the p-value resampling distribution to establish a sample size for a confirmatory study.

The resampling distributions for effect size and p-values differ from analytic studies such as power studies in as far as no theoretical probability distribution is assumed. The conclusions are of the form: these are distributions of outcomes when sampling from a population such as the one originally sampled. These conclusions do not require that the specific statistical test chosen is appropriate, only that it be used with subsequent samples. In particular, for this example one need not assume normality.

Generalizations of the one sample test

- Since a pre post one group study design can be reduced to a single continuous variable by differencing the previous example can be used to create an effect size (and p-value) resampling distribution for a paired test.
- Any univariate statistical test can be the subject of the analysis so long as there is a well defined effect size (or p-value).

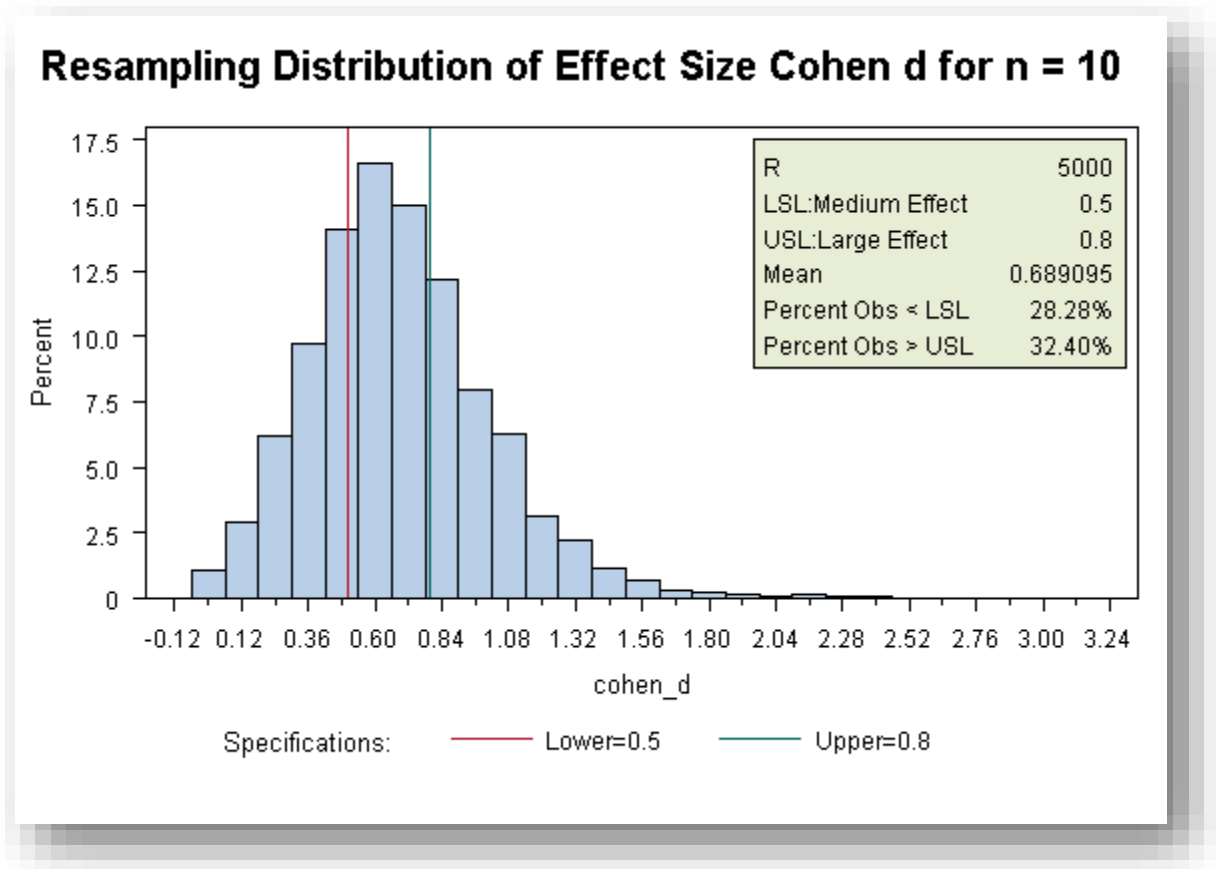


FIGURE 2. Effect Size Resampling Distribution for n = 10 for Data Set in Output 1.

Resampling Distribution of Effect Size Cohen d for n = 10, 15, 20

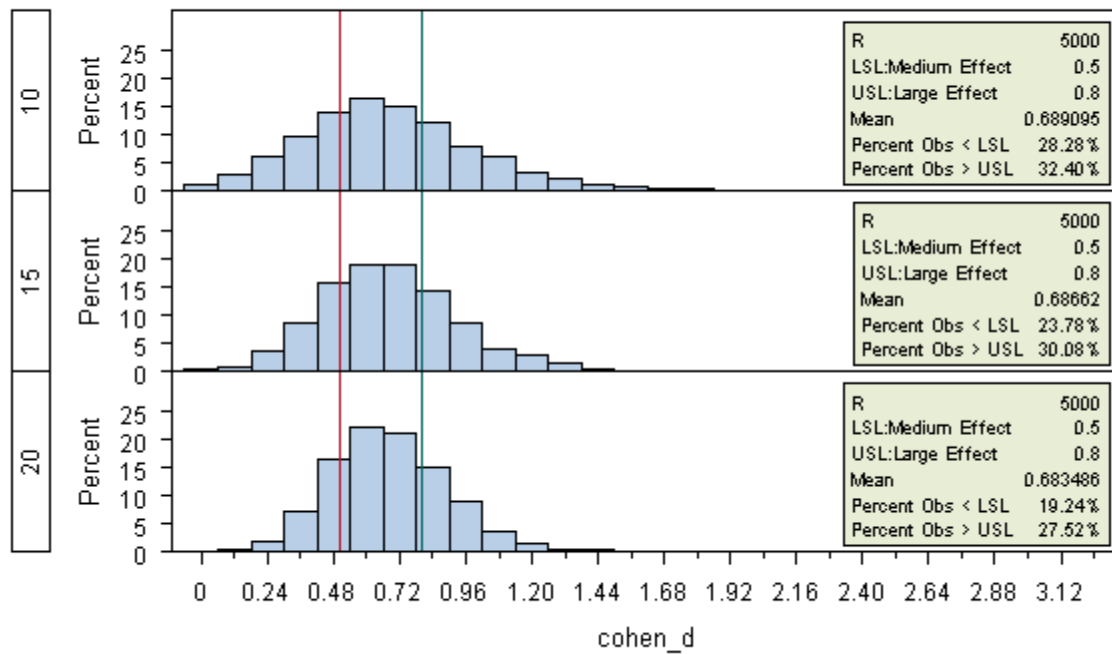


FIGURE 3. Effect Size Resampling Distribution for n = 10, 15, and 20 for Data Set in Output 1.

Resampling Distribution of p-value for n = 10, 15, 20

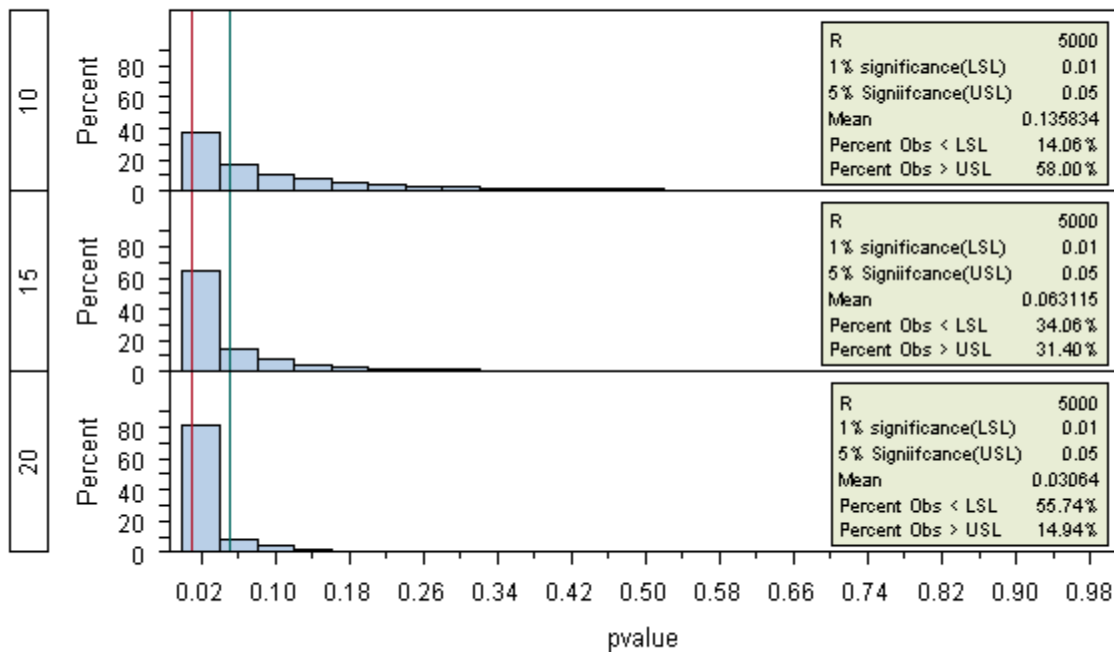


FIGURE 4. P-Value Resampling Distribution for n = 10, 15, and 20 for Data Set in Output 1.

SAS program 1 (for Example 1) is listed below.

```

***Generate the original sample 1A
normal
mean =
sigma = ;
DATA Ydata;
    seed = 12345679;
    mu = 143;
    sigma = 7;
    H0mu = 140;
    n = 10;
    do i = 1 to n;
        call rannor(seed, norm01);
        y = mu + sigma*norm01;
    output;
    end; run;
ods rtf;
proc print data = Ydata;var i y;
***perform t test;

```

```

proc ttest data = Ydata h0 = 140;
/*1*/
    title 't test on Mean';
    var y; run;
***calculate power based on observed mean and standard deviation;
proc power ;
/*2*/
    title 'power for t test based on normality with mean and SD of sample';
        onesamplemeans
            mean      = 3.2538
            ntotal    = 10
            stddev    = 5.0195
    power = .; run;
ods rtf close;
*****
STEP 2: Create subsamples of
sampsiz =
R = rep =
note: in this example 5000 for n = 10, 15, 20 respectively
*****/;
proc surveyselect data=Ydata out=outbootBn10
    seed=30459585
    method= urs
    sampsiz =10
    outhits
    rep=5000;
    title 'select bootstrap samples'; run;
proc surveyselect data=Ydata out=outbootBn15
    seed=30459587
    method= urs
    sampsiz =15
    outhits
    rep=5000;
proc surveyselect data=Ydata out=outbootBn20
    seed=30459589
    method= urs
    sampsiz =20
    outhits
    rep=5000;
    title 'select bootstrap samples'; run;
DATA outbootBn10; set outbootBn10;
ss = 10;
DATA outbootBn15; set outbootBn15;
ss = 15;
DATA outbootBn20; set outbootBn20;
ss = 20;
DATA outbootB; set outbootBn10 outbootBn15 outbootBn20;
/****STEP 3: calculate effect size (and p-value) for each sample)
proc sort data = outbootB; by ss replicate;
/** get statistics needed for effect size and pvalue for each sample */
proc means data = outbootB noprint;
    var y; by ss replicate;
    output out = ysummary n=ny mean = meany stddev = stddevy var =
vary;run;
/** calculate effect size and pvalue for each sample*/
DATA ysummary; set ysummary;
    df = ny-1;

```

```

ObsDiff = meany - 140;
SEM = stddevy/(ny**0.5);
tstat= obsdiff/SEM;
pvalue =(1-probt(abs(tstat),df))*2;
cohen_d = abs(ObsDiff)/stddevy;

run;

/*
multiple resampling distribution histograms
*****/

proc capability data = ysummary /*noprint*/;
spec lsl=0.50 usl = 0.80;
var Cohen d;
comphistogram cohen_d / class = (ss)nrows =3;
inset n='R' lsl='LSL:Medium Effect' usl = 'USL:Large Effect' mean lslpct
uslpct / cfill = ywh pos=ne;
title1 'Resampling Distribution of Effect Size Cohen d for n = 10, 15, 20';
run;
proc capability data = ysummary /*noprint*/;
spec lsl=0.010 usl = 0.05;
var pvalue;
comphistogram pvalue / class = (ss)nrows =3;
inset n='R' lsl='1% significance (LSL)' usl = '5% Significance (USL)' mean
lslpct uslpct / cfill = ywh pos=ne;
title1 'Resampling Distribution of p-value for n = 10, 15, 20';
run;
/**** Creating a single effect size histogram*/
DATA ysummary; set ysummary;
if ss ^ = 10 then delete;run;
proc capability data = ysummary /*noprint*/;
spec lsl=0.50 usl = 0.80;
var Cohen_d;
histogram cohen_d /*/ class = (ss)nrows =1*/;
inset n='R' lsl='LSL:Medium Effect' usl = 'USL:Large Effect' mean lslpct
uslpct / cfill = ywh pos=ne;
title1 'Resampling Distribution of Effect Size Cohen d for n = 10';
run;
quit;

```

Example 2: Two independent samples t-test

This test arises in experimental and observational studies in which two groups are being compared with respect to their population means. WLOG one can think of a treatment and a control group with means μ_T and μ_C . The null hypothesis is $\mu_T = \mu_C$. In this example one draws samples from the treatment group and from the control group. Since they are independent any pairing of samples from the treatments and the controls can be used to perform an independent samples t-test. The samples can be drawn using PROC SURVEYSELECT and then merged. In SAS one creates a data set in which each row in the data set corresponds to a “pairing” of a treatment resample and a controls resample and contains summary statistics $\bar{Y}_T, S_T, n_T, \bar{Y}_C, S_C, n_C$. For the equal variances version of the two independent samples t-test the effect size is

$$\frac{\bar{Y}_T - \bar{Y}_C}{S_p}$$

in which S_p is the pooled estimate of the common variance. The p-value resampling distribution can be created from the same summary data set.

CONCLUSION

The resampling distribution of effect sizes can be easily constructed for a statistical test or procedure using bootstrap methods. Similarly one can construct the p-value resampling distribution. These two distributions offer useful information regarding an outcome variable from a study. The p-value distribution is a useful nonparametric method for assessing retrospective (and prospective power for future studies with differing sample sizes). The effect size distribution can be used to assess clinical/operational significance for future studies. It also puts the observed effect size for the study in context in as far as the effect size resampling distribution allows one to assess the likelihood of observing an effect size of a certain magnitude.

REFERENCES

- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.
- Wludyka, P., and Smotherman, C. (2014). Using SAS to Create a p-value Resampling Distribution for a Statistical Test, SESUG 2014 (SD 91), Myrtle Beach, SC.

CONTACT INFORMATION

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

Name: Peter Wludyka
Enterprise: University of North Florida / Department of Mathematics and Statistics
Address:
City, State ZIP:
E-mail: pwludyka@unf.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.