SESSION 8

SIMULATIONS AND STATISTICS

Moderator—Barbara Basden, California State University, Fresno

Anatomy of a DATASIM simulation: The Doob and Gross horn-honking study

DRAKE R. BRADLEY
Bates College, Lewiston, Maine

The program DATASIM is used to simulate the classic "horn-honking" study by Doob and Gross (1968). In a 2×2 field experiment, Doob and Gross investigated the effects of status of frustrator—a low- or high-status car blocking an intersection—on the latency to honk among male and female drivers. The present paper illustrates how to extract the values of simulation parameters from the published study, how to initialize the simulation in DATASIM, and how to generate and analyze the simulated data. Certain complications arise because the latency data collected by Doob and Gross were nonnormally distributed, cell variances were heterogeneous, and sample sizes were unequal. DATASIM is able to incorporate these features in the simulation, and several methods for assessing the quality of the simulation are illustrated. In addition, sampling experiments are reported, which were performed to assess the joint and individual effects of nonnormality and heterogeneity on the Type I and Type II error rates of the F test. The paper concludes with some practical suggestions regarding how researchers can evaluate, and adjust for, the effects of such violations.

In 1968, Doob and Gross published a study in which they investigated the effects of status on inhibiting aggression. The experimental paradigm was simple: The experimenters would drive a low- or high-status car to an intersection, wait for the light to change to green, and then just remain stationary. They were interested in measuring the horn-honking behavior of the driver in the car behind them, which they did by counting the total number of honks made during the 12-sec cycle of the light, and by ascertaining the latency to the first honk. (If the driver never honked, he or she was assigned the maximum possible latency of 12 sec.) The experimenters recorded the sex of the driver, and subsequently analyzed the latency data by conducting a 2×2 (status $\times$ sex) analysis of variance. They also analyzed the data by conduct-

ing a 2×3 chi-square test to compare the number of drivers who never honked, honked once, or honked twice (or more) in the low- and in the high-status conditions. The results were clear and compelling: Drivers honked less frequently, and took longer to initiate the first honk, when they were blocked by a high-status as opposed to a low-status car. Doob and Gross interpreted this as support for the hypothesis that status can inhibit aggression. An additional finding was that males honked sooner than females, which Doob and Gross attributed to the higher levels of aggression typically shown by males.

For instructors who teach experimental psychology, research methods, or statistics, the Doob and Gross (1968) paper is an excellent journal article to have students read and discuss. It is short and relatively easy to understand, it describes a "real-life" phenomenon that most people have experienced first-hand, and it illustrates both the advantages and the pitfalls of conducting an experiment in a naturalistic setting. Also, a hypothesis is clearly stated and derived from previous research, variables are operationally defined, sources of confounding are identified, and so on. Finally, the study raises a number of interesting design and data analysis issues. These include alternative methods for analyzing the data (2×2 ANOVA vs. 2×3 chi-square), issues concerning the assumptions of

The computer simulations reported in this paper were conducted on a Macintosh IIfx supplied by Apple Computer and the Consortium of Liberal Arts Colleges. The author is indebted to Apple and to the CLAC Grant Review Committee for an Equipment Grant that supported the present research and, more generally, the development of DATASIM for the Macintosh. DATASIM was developed by the author and is available at nominal cost from Desktop Press, 90 Bardwell St., Lewiston, ME 04240 (telephone: 207-786-4113 or 6180). Requests for reprints should be addressed to Drake R. Bradley, Department of Psychology, Bates College, Lewiston, ME 04240.

statistical tests (normality, homogeneity of variance), the problem of unequal n with the 2×2 ANOVA, and an explicit comparison of the results obtained with questionnaires as opposed to field methods for collecting data.

The Doob and Gross (1968) study is just one of a large number that the instructor might wish to discuss in lecture. For this purpose, it is extremely useful to be able to generate, display, and analyze simulated data from such studies. A PC or Macintosh computer connected to an overhead projection panel provides the hardware to accomplish this, and a general purpose data simulator, such as DATASIM (Bradley, 1988), provides the software. By storing the relevant commands in a file, the instructor can quickly initialize a simulation of a particular study, generate a simulated dataset, and display and analyze the results. Parameters of the simulation can be easily changed, and new datasets generated and displayed in order to assess what effects, if any, the changes have produced (see Bradley, 1989a, 1989b; Bradley, Senko, & Stewart, 1990).

Of course, to be able to generate simulated data for a study, one must know the values of the relevant parameters (n , μ , σ , ρ , etc.). In the absence of other information, the sample statistics actually obtained in the study may be used to estimate the corresponding population parameters: for example, $s \rightarrow \sigma$, $r \rightarrow \rho$, and so on. Unfortunately, the relevant statistics are not always cited in the published report. Sample standard deviations, for example, are often not available, and investigators who use repeated measures designs almost never cite the actual intercorrelations among conditions. In some cases, it is possible to derive the values of parameters from information that is provided: for example, σ from MS_{within} or ρ from MS_{residual} (see Bradley, 1988, pp. 53–56). In other cases, one must simply employ a reasonable “guess-timate.” Yet another problem arises in the attempt to ascertain distribution shapes: Unless the published report provides histograms of the data, inferences about the presence of nonnormality must be made indirectly, on the basis of the means, standard deviations, and scale limits of the measured variables.¹ Consequently, to simulate a research study requires a certain amount of “detective” work. In the present paper, I shall illustrate how to use such methods to estimate the simulation parameters for the Doob and Gross (1968) study. This study proves to be a good choice, because it presents a number of interesting challenges for anyone who wishes to implement the simulation successfully.

Estimating Simulation Parameters

Let us consider the results of the Doob and Gross (1968) experiment in greater detail. We need to determine first what information is provided in the published report, and we need to compare that to the information needed to initialize the simulation. Table 1 contains the cell means, sample sizes, and F ratios reported by Doob and Gross for the 2×2 ANOVA that they conducted on the latency data. It was not possible from the published report to ascertain whether Doob and Gross had conducted a least squares

Table 1
Means,* Sample Sizes, and F Ratios† Reported by Doob and Gross (1968)

Frustrator (A)	Sex of Subject (B)		$\bar{A}_i$
	Male (b_1)	Female (b_2)	
Low status (a_1)	6.8 ($n=23$)	7.6 ($n=15$)	7.2
High status (a_2)	8.5 ($n=25$)	10.9 ($n=11$)	9.7
$\bar{B}_j$	7.65	9.25	8.45 ($\bar{G}$)

*The row, column, and grand means are *unweighted* means.

† $F_a = 10.71$, $F_b = 4.49$, and $F_{ab} = 1.17$.

or an unweighted means analysis of variance. For the present illustration, we will assume that an unweighted means ANOVA was conducted (see Winer, 1971, p. 445).² Consequently, *unweighted* row and column means are shown in the margins of Table 1. The harmonic mean of the four sample sizes, employed in the unweighted means analysis, is $n = 16.5938$.

The sample means in Table 1 may be used to estimate the μ_{ij} for the $p \times q$ cells of the design. With respect to σ_{ij} , Doob and Gross (1968) did not, unfortunately, report the standard deviations for the four cells, nor did they report the error term for the F ratios from which an estimate could be obtained ($\sqrt{MS_{\text{error}}} \rightarrow \sigma$). Consequently, the cell standard deviations have to be estimated indirectly. This can be done using the relation $F_{\text{effect}} = MS_{\text{effect}}/MS_{\text{error}}$. Isolating MS_{error} on the left side of the equality, we have $MS_{\text{error}} = MS_{\text{effect}}/F_{\text{effect}}$. The value of MS_{effect} for Factor A, Factor B, and the AB interaction can be calculated from the means in Table 1. The F ratios are directly available from the published report (see Table 1, notes). The following shows the application of this method to the results of the $p \times q$ design employed by Doob and Gross:

$$MS_a = nq \sum (\bar{A}_i - \bar{G})^2 / (p-1) = 103.711$$

$$MS_b = np \sum (\bar{B}_j - \bar{G})^2 / (q-1) = 42.4801$$

$$MS_{ab} = n \sum \sum (\bar{AB}_{ij} - \bar{A}_i - \bar{B}_j + \bar{G})^2 / (p-1)(q-1) = 10.620$$

$$MS_{\text{error}} = MS_a / F_a = 106.25 / 10.71 = 9.68358$$

$$MS_{\text{error}} = MS_b / F_b = 43.52 / 4.49 = 9.46105$$

$$MS_{\text{error}} = MS_{ab} / F_{ab} = 10.88 / 1.17 = 9.07695$$

The three values obtained for MS_{error} differ somewhat because of rounding error arising from the authors' cell means, which were reported to only one decimal place of accuracy (Table 1). The mean of these three values 9.40719, will be used as our best guess as to the pooled within-cell variance, MS_{error} . The pooled within-cell standard deviation for the latency data would therefore be $\sigma = 3.106711$. (And indication of the degree of rounding error is given by recomputing the F ratios by dividing each MS_{effect} by 9.40719: $F_a = 11.02$, $F_b = 4.52$, and $F_{ab} = 1.13$.)

We now have most of the information that we need to initialize a simulation of the Doob and Gross (1968) study. For each cell of the design, we know the mean, standard deviation, and sample size. (If the study had included any repeated measures factors, we would need to know the correlation across levels as well.) Unless otherwise specified, DATASIM will generate data following a normal distribution. In the present case, we can see that this will be a problem. In cell AB₂₂ of the design, the mean was 10.9 and the upper scale limit for the latency data was 12 sec. Since our estimate of the within-cell standard deviation is $\sigma \approx 3.107$, we see that "off-scale" values will frequently be generated for this cell. In fact, since the upper limit is $z = (12 - 10.9)/3.107 = .35$ standard deviations above the mean, X values exceeding 12 sec will occur about 36% of the time. If we wish to avoid such off-scale values and still retain a normal distribution, the standard deviation for this cell would have to be reduced to $S_{22} = (12 - 10.9)/3 = .367$!

All of this implies that the data collected by Doob and Gross (1968) were nonnormally distributed and that the variances across cells were heterogeneous. There is an obvious reason for this. Since the cycle of the spotlight had a duration of 12 sec, any subject whose latency would have exceeded this value was assigned a score of $X = 12$. Even if the underlying distribution were normal, this would have the effect of truncating the upper tail. Given the nature of the measure, however, it is unlikely that the underlying distribution would be normal. Most drivers will not honk initially, but as the elapsed time increases, the probability of honking should rise steadily. Near the end of the cycle, as the driver realizes the blocking car might not vacate the intersection before the light changes back to red, the probability of honking should rise dramatically. This would cause the latencies to "bunch up" at the upper end of the scale, near 12 sec. All of this suggests that the underlying distribution is negatively skewed, with a few people honking right away, but most waiting until near the end of the cycle before honking. If so, the distribution of latencies will have its mode at or near 12 sec, the upper scale limit.

Now, if high-status cars inhibit honking, and if females are less likely to honk quickly than males, we would predict the mean latency in cell AB₂₂ (high status/female subjects) to be quite high. This is just what Doob and Gross (1968) found. Cell AB₁₁ (low status/male subjects) had a mean latency of 6.8 sec. By comparison, the mean latency of 10.9 sec in cell AB₂₂ indicates that the data are "packed" right up against the upper scale limit. It seems reasonable to suppose, then, that the variance of the data in AB₂₂ was probably less than that in AB₁₁. For these reasons, it is likely that Doob and Gross had both nonnormally distributed data *and* heterogeneous variances. The question is, how can we model this on the basis of the information that we have?

One approach is to assume that the distribution shape in cell AB₂₂ is a negatively skewed exponential, with the mode being equal to the maximum score of 12 sec. The exponential distribution allows us to place the mean as

close as possible to the maximum score value: about 1 standard deviation away, as opposed to the 3 standard deviations of a normal distribution. Nevertheless, if we use the pooled within-cell standard deviation of 3.107 sec, X values exceeding 12 sec will frequently be generated. The maximum value of the standard deviation for cell AB₂₂ that would allow the exponential distribution to consistently generate on-range data is $S_{22} = (12 - 10.9)/1 = 1.1$. This corresponds to a variance of 1.21, as opposed to the value of MS_{error} estimated above to be 9.407! Consequently, even if we assume that the data follow highly nonnormal distributions, there would still have to be substantial heterogeneity of variance across cells. Given the small variance assumed for cell AB₂₂, the variances of one or more other cells must be correspondingly larger in order for the pooled variance to be 9.407.

Therefore, instead of specifying a common standard deviation of 3.107, we will specify different standard deviations for each cell of the design. This must be done in such a way that the cell variances pool to the value of MS_{error} estimated above. In the absence of information to the contrary, it was assumed that the cell variances were inversely proportional to the cell means. This seems reasonable, because as the cell means increase, the data will pack more and more tightly against the upper scale limit of 12 sec, and the variances will decrease. After some trial and error, the regression equation $Y = 33.19525 - 2.92647 X$ was found to yield cell variances as a function of the cell mean, such that the pooled variance was $MS_{\text{error}} = 9.40719$. Substituting the cell means from Table 1 in this expression produces cell variances of 13.30, 10.95, 8.32, and 1.30 for AB₁₁-AB₂₂. The pooled average (weighted by df) of the unrounded values of these cell variances is 9.40719. The corresponding within-cell standard deviations are 3.65, 3.31, 2.88, and 1.14 (rounded). Note that this solution yields a standard deviation for cell AB₂₂ that is somewhat larger than the value of 1.1 (see above) needed to generate on-range values for the exponential distribution. The larger value was necessary in order to avoid having the standard deviations of cells AB₁₁ and AB₁₂ be too large. As we shall see, the larger standard deviations needed for these two cells will sometimes produce X values that are *smaller* than the lower scale limit of .2 sec. Consequently, in terms of generating on-range data, the standard deviations selected represent the best compromise that could be achieved across all four cells.

The last thing that we need to do is to select appropriate nonnormal distributions for generating the data in each of the four cells. This is done using the generalized lambda distribution (GLD) developed by Ramberg and Schmeiser (1972, 1974) and Ramberg, Dudewicz, Tadikamalla, and Mykytka (1979). The GLD is a four-parameter distribution that can model distributions with various degrees of skewness (α_3) and kurtosis (α_4). The percentile function of the GLD is given by (Ramberg et al., 1979, p. 202):

$$R(p) = \lambda_1 + [p^{\lambda_3} - (1-p)^{\lambda_4}] / \lambda_2 \quad (0 \leq p \leq 1).$$

Substituting a value of p into this expression produces a value of x corresponding to the p th percentile of the GLD.

Figure 1. Probability density functions for the GLDs used to generate simulated data for each cell of the Doob and Gross (1968) experiment. The distributions become progressively more negatively skewed and leptokurtic as we move from AB₁₁ to AB₂₂ (see Table 2).

The parameters λ_1 and λ_2 are location and scale parameters, respectively, whereas λ_3 and λ_4 jointly determine the skewness (α_3) and kurtosis (α_4) of the distribution. The values of λ_1 – λ_4 therefore control the first four moments of the GLD (Ramberg et al., 1979, pp. 204–205). Note that the percentile function, $R(p)$, is simply the inverse of the cumulative distribution function, $F(x)$. The GLD is ideal for Monte Carlo research, because the percentile function lends itself to use with random-number generators. As noted by Ramberg et al. (1979, p. 202): “If X is a continuous random variable with percentile function R , and U is a uniform random variable on the interval zero to one, then the transformation $X = R(U)$ yields a

random variable with the percentile function R .” Since many algorithms exist for generating uniform random numbers (see Lewis & Orav, 1989, chap. 5), it is a simple matter to generate values of X following any particular GLD.

The specific GLDs employed to simulate the Doob and Gross (1968) study were selected in the following way. First, we assumed that no driver would honk before .2 sec, even if extremely impatient. Second, the maximum latency, as noted above, was 12 sec. The distance of the lower and upper scale limit from each cell mean was then computed in standard deviation units (using the cell standard deviations derived above). Finally, a GLD was selected (from Ramberg et al., 1979, pp. 210–214) that had 1% and 99% percentiles in standard score units that matched, as closely as possible, those of the scale limits. A GLD chosen in this way will generate non-normally distributed data that will usually fall within the bounds of the scale. In Figure 1 are shown the probability density functions of the GLDs used to model cells AB₁₁–AB₂₂. The distributions are all negatively skewed, with modes equal to the upper scale limit of 12 sec. (The one shown for cell AB₂₂ models an exponential distribution with $\alpha_3 = -2$ and $\alpha_4 = 9$.) As is evident from the plots, it was not possible to find GLDs that completely avoided the generation of off-scale values. In cells AB₁₁ and AB₁₂, in particular, it is apparent that score values less than .2 sec will arise occasionally. Nevertheless, the

Table 2
Parameters Defining the Generalized Lambda
Distributions (GLDs) for Cells AB₁₁–AB₂₂

Parameters*	Cell			
	AB ₁₁	AB ₁₂	AB ₂₁	AB ₂₂
μ	6.8	7.6	8.5	10.9
σ	3.65	3.31	2.88	1.14
α_3	-.60	-.75	-.90	-2.00
α_4	2.40	2.80	3.20	9.00
λ_1	1.4111	1.334	1.277	1.000
λ_2	.2347	.2104	.1880	-.00058
λ_3	.4951	.3903	.3160	-.00058
λ_4	.0000	.0000	.0000	.0000

* $\mu = E(X)$, $\sigma = \sqrt{E(X-\mu)^2}$, $\alpha_3 = E(X-\mu)^3/\sigma^3$, $\alpha_4 = E(X-\mu)^4/\sigma^4$.

GLDs do a good job of limiting the maximum value of X to 12 sec in each cell, and of modeling the desired non-normal distributions. As the cell mean increases from AB_{11} to AB_{22} , the data pack more and more tightly against the upper scale limit of 12 sec, and the variability in the data decreases. The values of λ_1 – λ_4 defining each of the GLDs in Figure 1 are listed in Table 2, along with the corresponding values of μ , σ , α_3 , and α_4 .³

We are now in a position to actually initialize the simulation in DATASIM.⁴ This is done by specifying the research design and the values of the relevant parameters. The necessary commands are shown in Figure 2: DESIGN specifies a 2×2 between-groups factorial design, MU and SIGMA specify the cell means and standard deviations, NOBS specifies the cell sample sizes, and LAMBDA specifies the parameters λ_1 – λ_4 for each cell. The DECIMAL command sets the number of decimal places to be retained in the simulated data, and LOWER and UPPER set lower and upper limits for generating the data (if off-range values occur, they are set equal to the limit). The remaining commands are used to label Factors A and B, and the dependent variable. Note that in the Macintosh version of DATASIM shown in Figure 2, commands may be either entered after a command prompt, or selected by pulling down the appropriate menu from the menu bar at the top of the screen. In the latter case, DATASIM builds and then enters the selected command on the command line, just

as if it had been entered by the user manually. This feature allows a complete record of the command history to be maintained, regardless of the way in which the commands were built initially (i.e., through typing or through menu selection).

Having initialized the simulation, the user can plot the distribution shapes of the GLDs by entering the PDF command (see Figure 1, above). As is shown in Figure 3, the population trends in the cell means are plotted by entering PLOT MU.⁵ These trends reflect the longer latencies reported by Doob and Gross (1968) for the high-status car and the female subjects. The trends also reveal a weak interaction between the two factors, even though the test on F_{ab} reported by Doob and Gross was not significant. If we should prefer *not* to incorporate this weak interaction in the simulation, then we could compute adjusted cell means (removing the interaction) and use these for μ_{ij} rather than the actual values reported by Doob and Gross.⁶ In the present case, we have chosen to initialize a literal replication of their study, including the weak but nonsignificant interaction between status and sex of subject.

Generating Simulated Data

Once the simulation is initialized, there are a variety of ways to generate simulated data. The SEED and DATA commands are used to generate a *single* simulated dataset: the SEED command simply generates the data,

Figure 2. Commands for initializing a simulation of the Doob and Gross (1968) study. The screen capture shows DATASIM MacVersion 1.0 running on a Macintosh IIfx under MultiFinder.

Figure 3. DATASIM plot of the cell means (PLOT MU) initialized for the Doob and Gross (1968) simulation. The figure shows the mean latency to honk as a function of status of frustrator and sex of subject.

whereas the DATA command both generates and displays the data. Yet another command, SIMULATE, is used to generate *multiple* simulated datasets. If other DATASIM commands are entered on the command line along with SIMULATE, they will be executed on each of the simulated datasets as they are generated. Consider the following example:

SIMULATE 6 458692301, AOV, PLOT MEANS.

This command list tells DATASIM to generate six simulated datasets based on a master seed of 458692301, and to display a two-way ANOVA and a plot of the means for each dataset.⁷ The output for the first simulated dataset is shown in Figure 4. Note that the DATASIM command and plotting windows have been sized and positioned so that the results of the ANOVA and the corresponding plot of the cell means are visible at the same time. For this particular dataset, the same general results were obtained as reported by Doob and Gross (1968, p. 216). Both status and sex of subject are significant [$F_A(1,70) = 15.66$, $p = .0002$, and $F_B(1,70) = 4.15$, $p = .0454$], whereas the interaction, which is roughly comparable to that evident in the population trends of Figure 3, is not significant. The mean squares in the ANOVA summary table can be compared with those derived above for the Doob and Gross study: 141.024 versus 103.711 for MS_A , 37.365 versus 42.4801 for MS_B , 18.472 versus 10.620 for MS_{AB} , and 9.003 versus 9.407 for MS_{error} .

As the execution of the SIMULATE command continues, the ANOVAs and plots for the remaining five datasets are displayed. The plots of the means obtained for all six datasets are shown in Figure 5. Given the sample sizes and variances specified for the four cells (Figure 2), the variation in the plots simply reflects the sampling error in the cell means. This variation provides a good indication of the reliability of the original study. For Dataset 2, for example, the results differ from those of Doob and Gross (1968) in that sex of subject is *not* significant, whereas the interaction *is*. Altogether, status is significant ($p \leq .05$) in Datasets 1, 2, 4, 5, and 6; sex of subject is significant in Datasets 1, 3, 4, and 5; and the interaction is significant only in Dataset 2. (As will be shown in subsequent sections, these results are in accord with the theoretical power levels of the tests.) Only three of the simulated datasets (1, 4, and 5) produce the same pattern of significance on F_A , F_B , and F_{AB} as reported by Doob and Gross. Given this degree of sampling error, therefore, researchers performing exact replications of Doob and Gross would be expected to duplicate their results, in terms of significance tests on A, B, and the AB interaction, about half of the time.

The six datasets plotted in Figure 5 represent a small-scale *sampling experiment*. DATASIM is designed to conduct large-scale sampling experiments involving thousands of simulated datasets. The statistical results for each dataset (F ratios, t tests, etc.) can be saved on disk and subsequently

Figure 4. The results of a 2x2 ANOVA on the first of six simulated datasets generated for the Doob and Gross (1968) study. The sample cell means are plotted in the window at the upper right.

processed in various ways. For example, the number of Type I or Type II errors can be tabulated, the sampling distributions of test statistics can be plotted and compared to theoretical distributions using the Kolmogorov-Smirnov test, and so on (Bradley et al., 1990). These capabilities are illustrated in subsequent sections of this paper.

Testing the Simulation

After one has initialized a simulation in DATASIM and generated a few simulated datasets, it is good practice to assess the quality of the simulation by conducting additional tests. To see if the program generates data that have the desired central tendency, variability, and distribution shape, it is helpful to temporarily increase the cell sample sizes. This is done with the NOBS command. We can then have DATASIM generate a large dataset, compute basic descriptive statistics on the data, and display plots of the histograms. If the simulator is functioning properly, we would expect the sample statistics to be close to the population values specified in the initialization, and the histograms to be similar to the corresponding PDFs.

To assess the quality of the Doob and Gross (1968) simulation, the cell sample sizes were scaled up by a factor of 100 by entering NOBS 2300 1500 2500 1100. A single simulated dataset was then generated by entering SEED 101682141 (the value of the seed was selected at random). In Figure 6, the histograms of the simulated data are pre-

sented for each of the four cells. Comparison of these with the theoretical distributions plotted in Figure 1 shows that the data follow the appropriate distributions. The only problem is the "upswing" in frequencies near the bottom end of the scale for cells AB₁₁ and AB₁₂. This problem was anticipated earlier, when we noted that the GLDs and variances selected for these cells could not completely avoid the generation of off-range values. As is apparent from Figure 1, latencies of less than .2 sec will sometimes be generated, and resetting these values to the lower limit of .2 sec causes the upswings seen in the histograms. This problem can be avoided by deleting the commands that set limits on the data (LOWER, UPPER). When this is done and the data are regenerated, the resulting histograms look the same as the PDFs in Figure 1.

Setting limits on the data also has the effect of biasing the means and standard deviations to some extent. In cell AB₁₁, for example, setting off-range values equal to the lower limit (.2 sec) will cause the mean to shift upward and the standard deviation to decrease somewhat. This is evident from Table 3, in which the mean and standard deviation of each cell are presented for (1) the theoretical distribution specified in the initialization, (2) the simulation shown in Figure 6, which incorporated lower and upper limits on the data, and (3) an identical simulation in which no limits were imposed.⁸ Note that, in general, the simulation without limits produces cell means and standard

Figure 5. Plots of the cell means for the six simulated datasets of the Doob and Gross (1968) study.

deviations that are closer to the theoretical values. (There is no difference in the results for cell AB₂₂ across the two simulations, because this cell did not generate off-range values.) The pooled variances computed from the standard deviations are 8.86 and 9.34 for the simulations with and without limits, respectively. These compare favorably to the theoretical value derived above of $MS_{\text{error}} = 9.407$. As expected, setting limits constrains the variability of the simulated data to some degree. Nevertheless, the bias introduced by setting limits is relatively small, and the

initialization in Figure 2 does a good job of producing data with the desired theoretical properties.

Conducting Sampling Experiments

To further assess the quality of the simulation, we can conduct large-scale sampling experiments by using the SIMULATE command, and we can compare the empirical sampling distributions of F_a , F_b , and F_{ab} to the corresponding theoretical distributions. For the initialization of the Doob and Gross (1968) study shown in Figure 2,

Figure 6. Histograms of the simulated data for each cell of the Doob and Gross (1968) study (seed = 101682141). The sample size for each cell was scaled up by a factor of 100, in order to better assess the quality of the simulation.

the violations of the normality and homogeneity assumptions will cause the actual sampling distributions to deviate to some extent from the expected noncentral F distributions. On the other hand, if simulations of the Doob and Gross study are conducted in which normal distributions and homogeneous variances are used, then the sampling distributions of F should conform to the expected theoretical distributions. In either case, the degree of fit between the empirical and theoretical distributions can be assessed by plotting their cumulative distribution functions (CDFs). As an additional check, we can calculate the theoretical power $(1 - \beta)$ for F_a , F_b , and F_{ab} and com-

Table 3
Means and Standard Deviations of the Simulated Data*
for the Doob and Gross (1968) Study

Frustrator (A)	Sex of Subject (B)			
	Male (b_1)		Female (b_2)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Low status (a_1)				
Theoretical	6.80	3.65	7.60	3.31
Simulation†	6.87	3.51	7.60	3.24
Simulation‡	6.80	3.65	7.56	3.34
High status (a_2)				
Theoretical	8.50	2.88	10.90	1.14
Simulation†	8.44	2.83	10.87	1.16
Simulation‡	8.44	2.85	10.87	1.16

* Cell sample sizes were $n_{ij} = 2300, 1500, 2500$, and 1100 for AB₁₁-AB₂₂; seed = 101682141. †Conducted with lower and upper scale limits of .2 and 12 sec, respectively. ‡Conducted without scale limits.

pare this to the actual number of significant F ratios obtained in the sampling experiment. Or, if simulations are conducted in which the population cell means are equal (so H_0 is true), then the actual Type I error rate of the F test can be compared to α . Violations of assumptions should produce some degree of departure of actual from nominal α . Of greatest concern is the possibility that the violations will *increase* the probability of a Type I error; in this case, the researcher will be more likely to reject H_0 (when it is true) than he or she would otherwise think.

Since the latency data for the horn-honking study showed very substantial departures from the normality and homogeneity of variance assumptions, sampling experiments can help us to assess what effects, if any, these violations had on the validity of the statistical tests reported by Doob and Gross (1968, p. 216). For example, did these violations produce Type I errors with respect to their conclusions about the significant effects of status (F_a) and sex of subject (F_b)? Alternatively, did the violations produce a Type II error with respect to the nonsignificant finding reported for the interaction (F_{ab})? As a general purpose statistical simulation program, DATASIM is ideally suited to answer such questions. Simulations are initialized in which H_0 is true or false for the various effects, and in which assumptions are or are not violated in various ways. The actual Type I and Type II error rates are then assessed by conducting sampling experiments. If the effects of the violations are minimal, the actual and nominal error rates should be about the same. If so, the

investigator need not worry about the violations' producing incorrect inferences with respect to H_0 . On the other hand, if the actual error rates differ substantially from the nominal level, it is possible to use the empirical sampling distributions to derive critical values that "adjust" for the effects of the violations.

To consider a concrete example, suppose DATASIM has been initialized to perform a simulation such as the one shown in Figure 2. A sampling experiment is conducted by executing a DATASIM command macro that generates, analyzes, and records the results of a large number of simulated datasets. The following is an example:

```
EXACT ON, CLOSE SCREEN, OPEN T1
SIMULATE 10000 739920247, AOV 00
CLOSE T1, STRIP T1, CGET T1 C2-C4
NONE 3, NOBS 10000, QREAD T1
STAT 1, COUNT C1-C3 >= 3.97778
```

To see how this macro performs the sampling experiment, we will briefly review the function of each command. First, the EXACT command tells DATASIM to output the exact (unrounded) values of any test statistics computed on the data. This is important, because the test values will subsequently be compared to a critical value (see COUNT command) and a decision made to reject or not reject H_0 . The next command, CLOSE, closes the screen for output. This is done in order to speed up the simulation. The third command, OPEN, opens a file on disk (called T1) for storing the output of the sampling experiment. Following this, the second line in the macro conducts the actual sampling experiment: SIMULATE generates 10,000 datasets starting from a master seed of 739920247, and AOV conducts a two-way ANOVA on each.⁹ The 00 option after AOV is an output switch that causes the F ratios to be displayed in *brief* mode rather than an ANOVA summary table, and that suppresses the computation of exact p levels for the tests (this speeds up the simulation quite substantially). Consequently, for each simulated dataset, only the current value of the seed and the values of F_a , F_b , and F_{ab} are output to T1. After the sampling experiment finishes executing, the third line in the macro closes the file T1, strips it of any text, and retains just the three columns (CGET C2-C4) containing the values of F_a , F_b , and F_{ab} . The fourth command line redefines the design to be NONE with 3 columns and 10,000 rows, and reads the F ratios from T1 into DATASIM. The last line computes the sample statistics (mean, standard deviation, etc.) for each column of F values, and then counts the number of F values in each column that are equal to or greater than 3.97778. This latter value is the exact $\alpha = .05$ critical value of F with 1 and 70 df , as computed by DATASIM.¹⁰ (Since F_a , F_b , and F_{ab} have the same numerator df in 2×2 designs, the critical value is the same for each.) Note that when H_0 is true, these counts represent the Type I error rate (α) of the test; when H_0 is false, they represent the power ($1 - \beta$) of the test.

Sampling Experiments with and without Violations: H_0 False

The macro shown above was used to conduct a sampling experiment that was based on the original initialization of the Doob and Gross (1968) study (Figure 2). Since this initialization incorporates nonzero effects for A, B, and the AB interaction, H_0 is false for each. The left column of Figure 7 shows the sampling distributions of F_a , F_b , and F_{ab} that resulted (each histogram is based on 10,000 F ratios). Next, the initialization was modified by removing the lower and upper scale limits, by specifying normal distributions to generate the data, and by employing homogeneous variances across cells ($\sigma^2 = 9.40719$, as derived above).¹¹ The command macro was then used to conduct a second sampling experiment. Since the *same* master seed was specified, the uniform random numbers generated and supplied to the GLDs for this sampling experiment were *identical* to those of the first. However, because of the changes in the initialization, these values were now transformed to X values following normal distributions having constant variance across cells, and which were not truncated by setting off-range values to a lower or upper limit. Consequently, the resulting F ratios should follow the appropriate noncentral F distributions. The right column of Figure 7 shows the histograms of the sampling distributions of F_a , F_b , and F_{ab} that were obtained. One immediately apparent difference is that the distributions on the right show greater variability than those on the left. So, one effect of the violations appears to be that extreme values of F are somewhat less likely to occur in either tail. Depending on the location of the critical value ($F_{.05,1,70} = 3.97778$) relative to the sampling distribution, this could either increase or decrease the power of the test. For F_a and F_b , the critical value falls in the left tail, and therefore the power of these tests will be slightly higher for the initialization having violations than the one not having violations. For F_{ab} , on the other hand, the critical value falls in the right tail of the sampling distribution. Consequently, the reduced variability of the sampling distribution will result in fewer F values exceeding the critical value, and power will therefore be somewhat less.

A better way to assess the results of the two sampling experiments is to compare the CDFs of the sampling distributions in Figure 7 to the CDFs of the appropriate noncentral F distributions. This is done in Figure 8, in which the CDFs of the empirical and theoretical distributions are plotted for F_a , F_b , and F_{ab} . The left and right columns show the CDFs for the sampling experiments with and without violations, respectively. For the latter, the empirical and theoretical CDFs are so close that it is difficult to distinguish them individually. This is not the case for the left column; in all three plots, the theoretical CDF is slightly above the empirical CDF for small values of F , about the same for intermediate values, and slightly below the empirical CDF for large values of F . This pattern reflects the reduced variability of the F ratios noted

Figure 7. Empirical sampling distributions of F_a , F_b , and F_{ab} . The SIMULATE command employed a master seed of 739920247 to generate 10,000 datasets for each of two simulations: one incorporating the nonnormality and heterogeneity of the Doob and Gross (1968) study (see Figures 1 and 2), and one employing normal distributions, homogeneous variances, and no scale limits. The results for the two simulations are shown in the left and right columns, respectively.

in Figure 7. Two additional points should be made with respect to the CDFs in Figure 8. First, when assumptions are met, DATASIM generates sampling distributions of noncentral F that correspond quite closely to theoretical expectation. Second, even given the very substantial violations of assumptions that characterized the Doob and Gross (1968) study, the resulting F distributions do not differ very much from the theoretical distributions. As shown below, the power of the tests for F_a , F_b , and F_{ab} are altered only slightly.

The theoretical CDFs in Figure 8 were computed using the following approximation for the noncentral F distribution (Owen, 1978, p. 165):

$$\Pr \{F^N(f_1, f_2, \lambda) \leq c\} \cong \Pr \{F(f^*, f_2) \leq c[f_1/(f_1 + \lambda)]\}$$

$$f^* = (f_1 + \lambda)^2 / (f_1 + 2\lambda).$$

To obtain the probability that F will be less than or equal to some value c in a noncentral F distribution with f_1 and f_2 *df* and noncentrality parameter λ , the value of c is mul-

Figure 8. Cumulative distribution functions (CDFs) for the empirical and theoretical sampling distributions of F_a , F_b , and F_{ab} (see caption accompanying Figure 7).

multiplied by $f_1/(f_1 + \lambda)$ and the result evaluated against a central F distribution with $f^* = (f_1 + \lambda)^2 / (f_1 + 2\lambda)$ df for the numerator and f_2 df for the denominator. For the initialization of the Doob and Gross (1968) study, the values of λ are as follows:¹²

$$\lambda_a = nq \sum \alpha_i^2 / \sigma^2 = 103.711 / 9.40719 = 11.0247$$

$$\lambda_b = np \sum \beta_j^2 / \sigma^2 = 42.4801 / 9.40719 = 4.51571$$

$$\lambda_{ab} = n \sum \alpha \beta_{ij}^2 / \sigma^2 = 10.620 / 9.40719 = 1.12893.$$

The value of n (the harmonic mean) is 16.5938, and α_i , β_j , and $\alpha\beta_{ij}$ are the main and interaction effects computed

from the means in Table 1. The nonzero value of λ_{ab} reflects the weak interaction alluded to above. The Type II error probability (β) of F_a , F_b , and F_{ab} can be approximated by setting $c = F_{.05,1,70} = 3.97778$, and then computing the probability using the approximation shown above. Subtracting this value from 1.0 gives the power ($1 - \beta$) of the test. These calculations result in power values of .9243, .5441, and .1725 for F_a , F_b , and F_{ab} , respectively. For the two sampling experiments reported above, the COUNT command revealed that the empirical power of F_a , F_b , and F_{ab} was .9393, .5779, and .1617 when assumptions are violated, and .9064, .5555, and

.1797 when assumptions are met. As expected, the violations slightly increase the power of F_a and F_b , and decrease the power of F_{ab} .

It is instructive to compare the means and standard deviations of the sampling distributions in Figure 7 to those of the corresponding theoretical distributions. The mean and variance of noncentral and central F are given by (Owen, 1978, pp. 164–165):

$$\mu_{NF} = \frac{f_2(f_1 + \lambda)}{(f_2 - 2)f_1}$$

$$\sigma_{NF}^2 = \frac{2(f_2)^2[(f_1 + \lambda)^2 + (f_1 + 2\lambda)(f_2 - 2)]}{(f_2 - 2)^2(f_1)^2(f_2 - 4)}$$

$$\mu_F = \frac{f_2}{(f_2 - 2)}$$

$$\sigma_F^2 = \frac{2(f_2)^2(f_1 + f_2 - 2)}{f_1(f_2 - 2)^2(f_2 - 4)}.$$

These formulas may be used to obtain the theoretical mean and standard deviation of the noncentral F distributions for the Doob and Gross (1968) study. (As illustrated below, central F distributions arise if the population cell means, μ_{ij} , are the same.) The first row of values in Table 4 shows the theoretical mean, standard deviation, and power for the noncentral distributions of F_a , F_b , and F_{ab} , given $f_1 = 1$, $f_2 = 70$, and λ_a , λ_b , and λ_{ab} as computed above. The next two rows, labeled S1 and S2, present the means and standard deviations of each of the sampling distributions in Figure 7. (The empirically determined power levels are also shown.) The S1 simulation represents the one with no violations of assumptions—the one

with normal distributions, homogeneous variances, and no scale limits. Given the agreement between the empirical and theoretical CDFs in the right column of Figure 8, we would expect the results of the S1 simulation to closely match the theoretical values, and they do.¹³

The S2 simulation represents the one that incorporated the violations characterizing the Doob and Gross (1968) study. The nonnormal distributions and heterogeneous variances employed in the simulation are clearly evident from the PDFs plotted in Figure 1. Recall that this simulation also set scale limits of .2 and 12 sec for the minimum and maximum possible scores. Table 4 shows that the means and the standard deviations of the S2 sampling distributions are smaller than the corresponding theoretical values or S1 empirical values. Comparing the distributions in the left and right columns of Figure 7, we see that when violations are present (left column), less extreme values of F are generated in the tails of the sampling distributions than would be expected, particularly in the upper tails. Hence, these distributions are both less variable and less skewed to the right than the distributions that arise when assumptions are met (right column). These facts are consistent with the smaller means and standard deviations observed for the S2 versus S1 sampling distributions (Table 4).

As explained above, the reduced variability of the S2 distributions increases the power of F_a and F_b , and it decreases the power of F_{ab} . The impact of the violations, therefore, is to produce a shift in the Type II error rate (β) and power ($1 - \beta$) of the tests. Since it is not clear to what extent this shift is due to the nonnormality and heterogeneity in the data, or to what extent it is due to the imposition of scale limits, a third simulation was conducted. The S3 sampling experiment was identical to the

Table 4
Effects of Violations on Noncentral and Central F
(Based on Doob & Gross, 1968)

Simulation*	F Distribution								
	$F_a(1,70)$			$F_b(1,70)$			$F_{ab}(1,70)$		
Noncentral† F	M	SD	$1 - \beta$	M	SD	$1 - \beta$	M	SD	$1 - \beta$
Theoretical	12.38	7.41	.9243	5.68	4.78	.5441	2.19	2.69	.1725
S1: no violations	12.36	7.31	.9064	5.66	4.77	.5555	2.17	2.69	.1797
S2: both violations‡	12.18	6.29	.9393	5.53	4.13	.5779	2.03	2.35	.1617
S3: both violations	11.99	6.12	.9394	5.40	4.05	.5685	1.93	2.29	.1516
S4: heterogeneous	12.25	6.62	.9308	5.51	4.26	.5683	1.97	2.35	.1521
S5: nonnormal	12.68	7.77	.8924	5.80	4.85	.5645	2.21	2.65	.1867
Central§ F	M	SD	α	M	SD	α	M	SD	α
Theoretical	1.03	1.49	.0500	1.03	1.49	.0500	1.03	1.49	.0500
S1: no violations	1.03	1.45	.0473	1.01	1.46	.0475	1.04	1.47	.0517
S2: both violations‡	.829	1.18	.0312	.820	1.20	.0288	.836	1.20	.0300
S3: both violations	.813	1.17	.0298	.814	1.20	.0275	.824	1.18	.0286
S4: heterogeneous	.807	1.15	.0276	.808	1.16	.0267	.822	1.18	.0277
S5: nonnormal	1.03	1.44	.0489	1.01	1.43	.0477	1.03	1.45	.0510

* Each simulation is based on $k = 10,000$ datasets generated from a master seed of 739920247. †Based on the μ_i , μ_j , and μ_{ij} in Table 1, so that $\lambda_a = 11.025$, $\lambda_b = 4.516$, and $\lambda_{ab} = 1.129$. ‡Conducted with lower and upper scale limits set to .2 and 12 sec, respectively. §Based on $\mu_i = \mu_j = \mu_{ij} = \mu = 8.45$, so that $\lambda_a = \lambda_b = \lambda_{ab} = 0$.

S2 simulation, except that no lower or upper limits were specified for the data, and hence no truncation of the GLDs could occur. The results are shown in Table 4. Since the means, standard deviations, and power levels obtained are quite similar to those of the S2 simulation, this implies that it is the nonnormality and heterogeneity per se that produces the observed shifts.¹⁴

Two additional simulations were conducted, to pinpoint the source of the shifts in power more precisely. As in the S3 simulation, no scale limits were imposed; in contrast to the S3 simulation, however, only one of the violations was incorporated at a time. Simulation S4 was conducted with heterogeneous variances and normal distributions, whereas Simulation S5 was conducted with nonnormal distributions and homogeneous variances ($\sigma^2 = 9.40719$). As can be seen in Table 4, the S4 simulation (heterogeneous variances) produced power levels for the F tests that are similar to those of the S3 simulation. This implies that the heterogeneity of variance is primarily responsible for the observed differences between the power levels of the S1 and S3 sampling experiments. Consistent with this, the S5 simulation shows that nonnormality by itself produces only slightly different power levels than those obtained when assumptions are completely satisfied (S1). Moreover, nonnormality and heterogeneity seem to have opposite effects: S5 power for F_a is *lower* than S1 power, whereas S4 power is *higher* (the reverse trend is apparent for F_{ab}). Note that the effects of the two violations are in no sense *additive*: while nonnormality lowers the power of F_a slightly (.8924 vs. .9064) and heterogeneity increases it (.9308 vs. .9064), both violations together produce the highest power of all (.9394 vs. .9064). If the individual effects of nonnormality and heterogeneity were additive, the power of F_a when both are present would be $.9064 + (.8924 - .9064) + (.9308 - .9064) = .9064 - .0140 + .0244 = .9168$. Given the actual value of .9394 obtained in the S3 simulation, we see that the joint effects on power are greater than what is predicted by the individual effects. Since the uniform random numbers generated for Simulations S1–S5 (and subsequently transformed by the GLD) were identical, this lack of additivity cannot be ascribed to sampling variation in the results. Instead, it suggests that the effects of violations in combination cannot be accurately predicted from their effects in isolation.

Sampling Experiments with and without Violations: H_0 True

Let us now consider the case in which the population cell means (μ_{ij}) are equal, so that H_0 is true for each of the effects in the design (A, B, and the AB interaction). Provided that assumptions are met, F ratios computed on the simulated datasets will now follow central F distributions, and the Type I error rates will be $\alpha = .05$. For simulations having violations and/or scale limits, the sampling distributions of F will deviate to some extent from the central F distribution, and the actual Type I error rate will not be exactly α . Sampling experiments conducted

under the latter circumstances provide a way to assess the effects of violating assumptions on the Type I error rate of the F test (Glass, Peckham, & Sanders, 1972). In the context of the present example, we are interested in seeing whether the specific violations that characterized the Doob and Gross (1968) study produce a substantial inflation in the Type I error rate. If so, the possibility would exist that the significant effects reported by Doob and Gross for status (F_a) and sex of subject (F_b) are simply Type I errors. To answer this question, sampling experiments were initialized as in Figure 2, except that the cell means were set to $\mu_{ij} = 8.45$ (the grand mean from Table 1). As above, simulations were conducted with and without violations, and with and without scale limits. To facilitate comparison with the previous results, the same master seed (739920247) was employed for the sampling experiments.¹⁵

The results are presented in the lower half of Table 4. Once again, there is a close correspondence between the theoretical values and the empirical values when the normality and homogeneity assumptions are met (S1). The means and standard deviations of the sampling distributions are quite similar to the theoretical values, and the empirical Type I error rates are close to α . As demonstrated previously (Bradley et al., 1990, p. 241), this verifies that DATASIM generates sampling distributions of central F that correspond to the expected theoretical distributions. For the S2 and S3 simulations, both of which involved nonnormal distributions and heterogeneous variances, the means and standard deviations are once again smaller than the theoretical or S1 empirical values, and the actual Type I error rates are less than α (about .03 as compared to .05). This effect is slightly more pronounced when scale limits are not imposed (S3). The reduction in the Type I error rate follows from the reduced variability of the sampling distributions, which in turn results in fewer F ratios exceeding the upper tail critical value of $F_{.05,1,70} = 3.97778$. The S4 and S5 simulations show, as above, that primarily the heterogeneity of variance is responsible for the deflation in the Type I error rate.¹⁶ Also, note that the effects of nonnormality (S5) and heterogeneity (S4), as assessed through comparison with the S1 simulation (no violations), are not additive with respect to the S3 simulation (both violations). This lack of additivity is most evident for F_{ab} ; if the individual effects of nonnormality and heterogeneity were additive, the Type I error rate of F_{ab} when both are present would be $.0517 + (.0510 - .0517) + (.0277 - .0517) = .0517 - .0007 - .0240 = .0270$. This contrasts with the S3 error rate of .0286. Hence, the joint effects of the violations produce less deflation in the Type I error rate than would be predicted from the individual effects.

Adjusting the Critical Value for Violations

The results of the S3 sampling experiment in the lower half of Table 4 can be used to obtain critical values for F_a , F_b , and F_{ab} that *adjust* for the violations present in the data. To illustrate, consider Figure 9, in which the

Figure 9. CDFs for the empirical and theoretical sampling distributions of F_a in simulation S3: nonnormal distributions, heterogeneous cell variances, and H_0 true.

empirical CDF of F_a and the corresponding theoretical CDF for central $F_{1,70}$ are plotted. From Table 4, we see that $F_{.05,1,70} = 3.97778$ cuts off only .0298 in the upper tail of the sampling distribution of F_a . Figure 9 shows that this discrepancy is due to the fact that the empirical CDF rises more rapidly than the theoretical CDF. To obtain a critical value of F_a that adjusts (empirically) for the effects of the violations on the CDF, we can draw a horizontal line from the ordinate at $1 - \alpha = .95$ over to the empirical CDF, drop a vertical line to the abscissa, and read the corresponding value of F . A more precise method would be to use the KS (Kolmogorov-Smirnov) command in DATASIM to rank-order the F ratios in the sampling distribution, to compute the cumulative proportions, and to locate the value of F_a falling at the 95th percentile rank. The latter technique produces an adjusted critical value of $F'_{.05,1,70} = 3.26117$, in contrast to the unadjusted value of $F_{.05,1,70} = 3.97778$. If the adjusted critical value is used to test F_a , the long run risk of a Type I error should be about .05 instead of .0298. Furthermore, if H_0 is false, using the adjusted critical value will increase the power of F_a because a greater proportion of the F ratios will exceed this value. This is verified by using the COUNT command to determine the power of F_a for the S3 simulation in the top half of Table 4: the power is $1 - \beta = .9611$, using the adjusted critical value, as compared with .9394 for the unadjusted value.

There are a number of advantages to using DATASIM to obtain adjusted critical values for test statistics when assumptions are violated. If the pattern of violations is complex, it is generally quite difficult to determine the exact effect of the violations by using analytical methods. Sampling experiments like those reported in Table 4 can provide reasonably accurate estimates of these effects, provided that the number of simulated datasets employed in the simulation is large. As noted elsewhere (Bradley et al., 1990, p. 244), DATASIM estimates of Type I error

rates have, in several cases, been more accurate than the "exact" values reported in previous analytical studies (Box, 1954; Hsu, 1938). So, the fact that DATASIM estimates are not *exact* is less of a concern than might seem the case on first consideration. Moreover, when an investigator is faced with substantial and complex patterns of violations of assumptions, most of the standard methods for dealing with such problems (e.g., scale transformations) simply do not work. Although nonparametric methods may be employed in lieu of the one-way ANOVA, there is no well-accepted nonparametric analogue to the two-way ANOVA. As demonstrated above, however, the researcher can initialize a DATASIM simulation that incorporates the specific pattern of violations actually observed in the data, conduct sampling experiments to assess the effects of these violations on the Type I (or Type II) error rates of the tests, and then obtain critical values that adjust for the effects of the violations. If the researcher's F ratios exceed the adjusted critical values, then he or she can reject H_0 and know that the risk of a Type I error is being maintained at or near α .

Let us apply this reasoning to the Doob and Gross (1968) study. Using the adjusted critical value of $F'_{.05,1,70} = 3.26117$ derived above, Doob and Gross would have made the same decision with respect to F_a as they actually reported—that is, reject H_0 . The adjusted critical values of F_b and F_{ab} also result in the same decisions—that is, reject H_0 for B, but do not reject H_0 for AB. This shows that the presence of the violations did not cause Doob and Gross to draw different conclusions than they would have drawn if the effects of these violations had been taken into account. More specifically, the simulations in the bottom half of Table 4 show that the substantial violations encountered in the data (i.e., exponential distributions and a maximum variance ratio of 10:1) did not increase the probability of a Type I error. In fact, if H_0 is really true, using an unadjusted critical value for the tests will *reduce* the probability of Type I errors. If H_0 is false, on the other hand, the simulations in the top half of the table show that the effects of the violations depend on the particular effect being tested: power increases slightly for F_a and F_b , and it decreases slightly for F_{ab} . Taken together, these various results suggest that Doob and Gross did *not* make errors of inference with respect to F_a , F_b , or F_{ab} that could be uniquely attributed to the violations themselves. Put differently, they ran pretty much the same risk of making a Type I or Type II error as they would have if the normality and homogeneity assumptions had been satisfied.

Although the preceding conclusions are reassuring, they are specific to the particular patterns of violations investigated here, and it should not be assumed that the F test will be equally robust with respect to other patterns of violations. Moreover, Table 4 makes it clear that generalizations about the effects of heterogeneity of variance, nonnormality, and the like, as reported by previous Monte Carlo and analytical studies in which these effects were investigated individually (Boneau, 1960; Box, 1954; Collier, Baker, Mandeville, & Hayes, 1967; Hsu, 1938;

Hsu & Feldt, 1969; Norton, 1952; Scheffé, 1959), cannot be used to predict what will happen when these violations are combined. For researchers like Doob and Gross (1968), who are confronted with a specific pattern of serious violations, the easiest way to determine the possible impact of these violations is to initialize an appropriate simulation in DATASIM, conduct sampling experiments under H_0 and H_1 , and then compare the actual and nominal error rates of the tests. As is shown in Table 4, the results of such sampling experiments can reassure researchers that their statistical inferences are probably valid, despite the violations present in the data.

Other Simulation Methods

In the DATASIM simulations illustrated in this paper, the generalized lambda distribution (Ramberg et al., 1979) was employed to model the data in each of the four cells of the Doob and Gross study (1968). As explained above, it was not possible to find GLDs that completely avoided the generation of off-range values, even though highly skewed distributions with heterogeneous variances were specified for the different cells. Since the GLD is a "well-behaved" continuous probability distribution, it cannot easily incorporate the existence of range truncation effects in the data. Although DATASIM can approximate such effects by setting scale limits, this has the undesirable effect that the expected values of the mean and standard deviation will now be biased relative to those originally specified for the GLD.

It would be desirable to be able to generate data that follow arbitrary distributions, that never have off-range values, and that have the correct mean and standard deviation in the long run. This can be accomplished using the CATVAR command (Bradley, 1988, p. 66). CATVAR allows the user to specify a *discrete* probability distribution for generating the data. For example, suppose that Doob and Gross had measured horn-honking latencies to the nearest second. In that case, the possible values of X would range from 0 to 12 in increments of 1. The probability of generating each of these values can be specified by the CATVAR command, and this can be done independently for each of the four cells of the design. The initialization would look similar to that shown in Figure 2, except that the CATVAR command would now replace the LAMBDA command. In addition, there would be no need to specify MU and SIGMA, because these are completely determined by the probability distribution. The latter fact means that the discrete probabilities, P , specified by CATVAR have to be selected carefully in order to assure that the expected values of the mean and variance are equal to the values desired for the particular cell of the design:

$$\begin{aligned}\mu &= E(X) = \sum xP(x) \\ \sigma^2 &= E(X-\mu)^2 = E(X^2) - [E(X)]^2 \\ &= \sum x^2P(x) - [\sum xP(x)]^2.\end{aligned}$$

Generating simulated data with discrete probabilities eliminates the problems noted earlier and allows the user to

specify distributions with very unusual shapes. The main disadvantage is that this approach is applicable only when a relatively small number of discrete values of X are to be generated. Also, some trial and error is necessary to find values of P that produce the desired mean and standard deviation.

A final point to be made concerns the 2×3 chi-square analysis reported by Doob and Gross (1968). In this analysis, the low- and high-status conditions were compared by testing the homogeneity of proportions of drivers honking zero, once, or twice during the 12-sec cycle of the light. They found that 47% of the drivers blocked by the low-status car honked twice, whereas only 19% of the drivers blocked by the high-status car did so. Conversely, 16% of the drivers never honked in the low-status condition, compared to 50% in the high-status condition. The chi-square test reported by Doob and Gross (p. 216) for the overall 2×3 table was significant [$\chi^2(2) = 11.14$, $p < .01$]. For instructors or researchers interested in simulating this part of the study, DATASIM can generate frequency data in tables following whatever trends are desired (Bradley, 1988, pp. 59–62). The design is declared as a TABLE design (e.g., DESIGN TABLE 2×3), and the individual cell probabilities are specified by using the CPROBS command (similar to CATVAR). If desired, either the row or the column marginal frequencies can be *fixed* in the sampling scheme (RFIX, CFIX). This allows simulated datasets to be generated that have a predesignated number of observations in each row or column of the table. Finally, we note that sampling experiments based on TABLE designs can be used to assess the effects of violating the minimum expected frequency requirement on the Type I error rate of the chi-square test (Bradley, Bradley, McGrath, & Cutcomb, 1979; Bradley & Cutcomb, 1977).

Summary

The classic "horn-honking" study by Doob and Gross (1968) provides many interesting challenges to successfully implementing a DATASIM simulation. The study consisted of a 2×2 factorial design with unequal n s, the latency data followed nonnormal distributions with heterogeneous variances across cells, and scale limits were imposed that potentially truncated the distributions. Moreover, the original report did not supply all of the information that was needed to specify the appropriate population parameters. Nevertheless, it was possible to derive reasonable values for these parameters indirectly, and certain assumptions were made about the distribution shapes and variances across cells. With this information, the parameters of the Doob and Gross study were initialized in DATASIM, and a number of simulations were conducted. The results show that: (1) DATASIM is able to generate data for each cell of the Doob and Gross study that have the desired central tendency, variability, and distribution shape; (2) DATASIM generates sampling distributions of noncentral and central F that conform closely to theory, and (3) the violations of assumptions present in the Doob and Gross study probably did not lead them to make errors

of inference. The simulation techniques illustrated in the paper should be of interest to instructors looking for novel laboratory applications of computers, and to Monte Carlo researchers conducting large-scale sampling experiments on the effects of violating assumptions.

REFERENCES

- BONEAU, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin*, 57, 49-64.
- BOX, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, 25, 290-302.
- BRADLEY, D. R. (1988). *DATASIM*. Lewiston, ME: Desktop Press.
- BRADLEY, D. R. (1989a). Computer simulation with DATASIM. *Behavior Research Methods, Instruments, & Computers*, 21, 99-112.
- BRADLEY, D. R. (1989b). A general purpose simulation program for statistics and research methods. In G. Garson & S. Nagel (Eds.), *Advances in social science and computers* (Vol. 1, pp. 145-186). Greenwich, CT: JAI Press.
- BRADLEY, D. R., BRADLEY, T. D., McGRATH, S. G., & CUTCOMB, S. D. (1979). Type I error rate of the chi-square test of independence in $R \times C$ tables that have small expected frequencies. *Psychological Bulletin*, 86, 1290-1297.
- BRADLEY, D. R., & CUTCOMB, S. (1977). Monte Carlo simulations and the chi-square test of independence. *Behavior Research Methods & Instrumentation*, 9, 193-201.
- BRADLEY, D. R., SENKO, M. W., & STEWART, F. A. (1990). Statistical simulation on microcomputers. *Behavior Research Methods, Instruments, & Computers*, 22, 236-246.
- COLLIER, R. O., BAKER, F. B., MANDEVILLE, G. K., & HAYES, T. F. (1967). Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. *Psychometrika*, 32, 339-353.
- DOOB, A. N., & GROSS, A. E. (1968). Status of frustrator as an inhibitor of horn-honking responses. *Journal of Social Psychology*, 76, 213-218.
- GLASS, G. V., PECKHAM, P. D., & SANDERS, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42, 237-288.
- HSU, P. L. (1938). Contribution to the theory of "Student's" *t*-test as applied to the problem of two samples. *Statistical Research Memoirs*, 2, 1-24.
- HSU, T. C., & FELDT, L. S. (1969). The effect of limitations on the number of criterion score values on the significance level of the *F*-test. *American Educational Research Journal*, 6, 515-527.
- LEWIS, P. A. W., & ORAV, E. J. (1989). *Simulation methodology for statisticians, operations analysts, and engineers*. Pacific Grove, CA: Wadsworth.
- NORTON, D. W. (1952). *An empirical investigation of some effects of non-normality and heterogeneity of the F-distribution*. Unpublished doctoral dissertation, Iowa State University, Ames, IA.
- OWEN, D. B. (1978). Special continuous distributions. In W. Kruskal & J. Tanur (Eds.), *International encyclopedia of statistics* (Vol. 1, pp. 161-168). New York: Free Press.
- RAMBERG, J. S., DUDEWICZ, E. J., TADIKAMALLA, P. R., & MYKYTKA, E. F. (1979). A probability distribution and its uses in fitting data. *Technometrics*, 21, 201-214.
- RAMBERG, J. S., & SCHMEISER, B. W. (1972). An approximate method for generating symmetric random variables. *Communications of the ACM*, 15, 987-990.
- RAMBERG, J. S., & SCHMEISER, B. W. (1974). An approximate method for generating asymmetric random variables. *Communications of the ACM*, 17, 78-82.
- SCHÉFFÉ, H. (1959). *The analysis of variance*. New York: Wiley.
- WINER, B. J. (1971). *Statistical principles of experimental design* (2nd ed.). New York: McGraw-Hill.

NOTES

1. For example, if the upper limit is less than 3 standard deviations from the mean but the lower scale limit is not, then the data are probably skewed to the left (negative skew). This kind of distribution might arise, for instance, when students rate a highly popular instructor on a bipolar rating scale (-3 to +3).

2. It is possible to argue that either a least squares or an unweighted means analysis should be conducted on the data, depending on the goal of the researchers. If the goal is to obtain estimates of *overall* horn-honking latencies in the population of drivers, which take into account the relative incidence of male and female drivers, then a least squares (or weighted means) solution should be employed. On the other hand, if the goal is to assess differences in the behavior of male and female drivers, regardless of their relative incidence in the population of drivers, an unweighted means analysis would be appropriate. From the latter point of view, the investigator would prefer to have an orthogonal design with equal numbers of males and females in the low- and high-status conditions. However, due to lack of control in soliciting and assigning subjects in field experiments, this may be difficult to achieve in practice. The unweighted means analysis provides an alternative solution: it attempts to determine what effects the independent variables would have had in an orthogonal design with equal numbers of subjects per cell. (It is assumed that the inequality of cell sample sizes is due to random factors, and not to the experimental treatments themselves.)

3. The values of λ_1 - λ_4 shown in Table 2 define GLDs with $\mu_z = 0$ and $\sigma_z = 1$ (see Ramberg et al., 1979, p. 205). DATASIM transforms the nonnormal deviates (*z*) generated by the GLD by computing $X = \mu + \sigma z$, where μ and σ are the mean and standard deviation specified for a particular cell. The PDFs plotted in Figure 1 show these transformed distributions.

4. In this paper, I describe the most recent versions of DATASIM: 1.1 for the IBM, and 1.0 for the Macintosh. These versions have been specifically enhanced for conducting statistical simulations and large-scale sampling experiments.

5. Note that in the Macintosh version of DATASIM, a separate plotting window is opened in front of the command window, and the plot is displayed. The plotting window can be resized and positioned like any Macintosh window, and the user can move back and forth between the command and plotting windows by simply clicking on them.

6. The adjusted cell means would be computed as follows:

$$\overline{AB}_{ij}' = \overline{A}_i + \overline{B}_j - \overline{G}.$$

7. The master seed is used to initialize DATASIM's uniform random-number generator (Bradley et al., 1990, pp. 237, 241). Subsequent calls to the random-number generator fetch uniform random numbers generated from this initial seed. Consequently, an entire set of simulated datasets can be regenerated at any time by reinitializing the simulation, and then reseeding the random-number generator with the value of the master seed originally employed. (The original value is selected at random by entering the MASTER command.) DATASIM also allows the user to conduct simulations in which each successive dataset is generated from a seed composed at random, rather than one ultimately derived from a master seed.

8. By identical simulation, we mean that the same seed was employed to initialize the random-number generator, and therefore precisely the same random numbers were generated for both simulations. The only difference was that in the first simulation, off-range values were set equal to the scale limit, whereas in the second they were not.

9. If the cell sample sizes specified in the initialization are unequal, DATASIM conducts an unweighted means analysis of variance on the data.

10. Exact critical values and probabilities for *z*, *t*, *F*, *r*, and chi-square are output by the CRIT and PROB commands available in DATASIM (Bradley, 1988, p. 47).

11. The parameters for a GLD approximating a standard normal distribution are: $\lambda_1 = 0$, $\lambda_2 = .1974$, $\lambda_3 = .1349$, and $\lambda_4 = .1349$.

12. Note that λ refers here to the noncentrality parameter defining a noncentral *F* distribution, *not* to the parameters of the generalized lambda distribution (GLD). In the remaining sections of the text, the intended usage should be clear from the context.

13. One comparison in Table 4 deserves comment. The theoretical power of F_a is .9243, whereas the empirical power obtained in the S1 simulation is .9064. Using the normal approximation to the binomial, these two values differ significantly ($z = -6.77$, $p < .0001$). This difference reflects in part the inherent sampling error in the simulation, but it also reflects any approximation error present in the technique (Owen, 1978, p. 165) used to compute theoretical power. Using power tables in Winer (1971, p. 886), nonlinear interpolation between the table entries bracketing $\phi = \sqrt{(\lambda_a/p)} = \sqrt{(11.0247/2)} = 2.3478$ provides estimated power of .90 for 1/30 df and .92 for 1/ ∞ df . Since this range does not include the value of .9243 obtained with the approximation method ($F_{1,70}$ with $\lambda_a = 11.0247$), the approximated value would appear to be too high. The empirical estimate of .9064, on the other hand, falls in between the two values obtained with Winer's table.

14. Recall that the initialization in Figure 2 employed cell variances that were proportional to the cell means. In this case, a square root transformation, such as $X' = \sqrt{(X + .5)}$, can stabilize the variances (Winer, 1971, p. 399). To see whether this would improve the situation, the S2 simulation in Table 4 was repeated, and the data in each simulated dataset were subjected to a square root transformation prior to conducting the two-way ANOVA. The resulting means for F_a , F_b , and F_{ab} were 10.53, 4.61, and 1.51; the standard deviations were 5.40, 3.61, and 1.96; and the empirical power levels were .913, .486, and .103, respectively. The means and standard deviations cannot be directly compared to the theoretical values in Table 4, because the noncentrality parameters for the transformed distributions will differ from those of the un-

transformed distributions. However, it is evident from the power results that the square root transformation has reduced the sensitivity of the tests on F_b and F_{ab} . In this particular instance, therefore, one is better off *not* transforming the data.

15. In conducting Monte Carlo research, it is generally good practice to use different (randomly selected) seeds to initialize the random-number generator for each sampling experiment. In this case, sampling error associated with the estimates of Type I or II error rates in one sampling experiment will be "independent" of those obtained in other sampling experiments. (This presumes that the sequence of random numbers generated for the two sampling experiments do not overlap, and that the uniform random-number generator is functioning properly.) On the other hand, using the same master seed for each sampling experiment allows the investigator to more easily separate the effects that are due to changes in parameter specification (e.g., nonnormality, heterogeneity) from those that are due to sampling variation. Researchers would be well advised to employ both techniques in the course of conducting exhaustive Monte Carlo investigations.

16. The Type I error rates are deflated because in the Doob and Gross study the larger cell variances occur in the cells that have the larger sample sizes. Conversely, the smallest cell variance ($\sigma^2 = 1.30$ for AB₂₂) occurs in the cell that has the smallest sample size ($n = 11$). The greater weight given to the larger variances when pooling the cell variances causes the value of MS_{error} to be inflated. As shown in Table 4, this decreases both the mean and the variance of the sampling distribution of F values (relative to theoretical expectation), and therefore the Type I error rate as well.