The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year university library subscription	$125	1-year university library subscription	$165
2-year university library subscription	$215	2-year university library subscription	$295
3-year university library subscription	$315	3-year university library subscription	$435
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2013)
13, Number 1, pp. 3–20

Stata as a numerical tool for scientific thought experiments: A tutorial with worked examples

Theresa Wimberley
Department of Economics and Business
National Centre for Register-Based Research
Aarhus University
Aarhus, Denmark
theresa@ncrr.au.dk

Erik Parner
Department of Public Health
Biostatistics
Aarhus University
Aarhus, Denmark

Henrik Støvring
Department of Public Health
Biostatistics
Aarhus University
Aarhus, Denmark

Abstract. Thought experiments based on simulation can be used to explain the impact of the chosen study design, statistical analysis strategy, or the sensitivity of results to fellow researchers. In this article, we demonstrate with two examples how to implement quantitative thought experiments in Stata. The first example uses a large-sample approach to study the impact on the estimated effect size of dichotomizing an exposure variable at different values. The second example uses simulations of datasets of realistic size to illustrate the necessity of using sampling fractions as inverse probability weights in statistical analysis for protection against bias in a complex sampling design. We also give a brief outline of the general steps needed for implementing quantitative thought experiments in Stata. We demonstrate how Stata provides programming facilities for conveniently implementing such thought experiments, with the advantage of saving researchers time, speculation, and debate as well as improving communication in interdisciplinary research groups.

Keywords: st0281, quantitative thought experiments, simulations

1 Introduction

A primary obligation for applied statisticians working in larger, interdisciplinary research groups is to provide guidance on study design, choice of statistical model, and explanation of results. The impact of the chosen study design, statistical analysis strategy, or sensitivity of results to certain assumptions must often be explained to the entire research group, even when some members have no formal statistical training. While the statistical literature may often provide solid results for preferring one approach over another, it will typically be in the form of abstract, mathematical, or probabilistic reasoning, which is difficult to communicate in plain language. Consequently, the the-

oretical results risk being perceived as unconvincing—magical arguments originating from the black hat of a statistician. This perception is only reinforced if the result is counterintuitive or controversial.

In such situations, we have found that implementing quantitative thought experiments may serve as a valuable pedagogical instrument to better explain what the theory means. These numerical experiments can often be further updated to account for ensuing "what if?" questions, which can be crucial in making sure that all arguments put forward in the group have been heard and fairly evaluated. The main prerequisite for this to be of practical value, however, is the ability to easily implement the thought experiment in a flexible and convenient software program that allows numerical analysis. In this article, we will demonstrate with two examples that Stata provides such programming facilities.

Arguably, this is just another example of how to use the computational muscles of Stata to overcome analytical shortcomings or solve problems that are mathematically intractable. A well-known example of this is to use Stata for estimation of statistical power via stochastic simulation (Feiveson 2002), which is now a commonly used and often-cited strategy. Therefore, the main objective of this article is not to give a general and detailed account on stochastic simulation in Stata. Rather, our main objective is to provide two illustrative case studies where the thought-experiment approach allowed us to present convincing arguments to our fellow researchers—in this case, epidemiologists—even in situations where the statistical theory is complex.

Both examples presented here originate from our work within the Lifestyle During Pregnancy Study (LDPS). The LDPS is a large epidemiological study on the effect of low-to-moderate alcohol consumption during early pregnancy on a child's neurodevelopment at age 5 (Kesmodel et al. 2010). The study was based on a complex stratified sample conducted within the Danish National Birth Cohort (Olsen et al. 2001), where the stratification was used to ensure adequate representation of women with higher exposures in terms of both average alcohol intake (weekly number of alcoholic drinks) and binge episodes (drinking at least five alcoholic drinks at a single occasion). Originally, the sampling was based on 20 different strata (Kesmodel et al. 2010) defined in a complex way; for pedagogical reasons, we here choose a more simple design consisting of 18 different strata, all defined by categories of average alcohol intake during pregnancy (0, 1–4, 5–8, 9+ drinks per week) and timing of binge drinking (no binge, occurrence in weeks 1–2, 3–4, 5–8, 9, or later), where the last two categories of timing of binge drinking were collapsed when the average alcohol intake was 5–8 and 9+ drinks per week. Note that average intake was defined such that it was possible for women to have an average intake of 0 (the typical intake) and yet have one or more binge episodes.

The first example in this article studies the impact on the estimated effect size due to dichotomizing an exposure variable at different cutoff values. The second example illustrates the necessity of using sampling fractions as inverse probability weights in the statistical analysis for protection against bias in this complex sampling design. After the two examples, we give a brief outline of the general steps needed for implementing quantitative thought experiments in Stata.

2 Example 1: Does dichotomizing an exposure variable at higher values always lead to larger effect sizes?

2.1 Scientific setting

In the LDPS, the cutoff for defining a binge drinking episode was set at five drinks at a single occasion. With this definition, the analysis of the data yielded a rather small binge effect on IQ, and so speculation naturally arose in the research group on the causes for this. In the literature, it is well known that dichotomization of a predictor variable may impair statistical efficiency and cutoffs should be chosen carefully (Senn and Julious 2009). In a previous article by Olsen (1994), a cutoff of eight drinks had been used and a larger effect estimate reported; thus one of the epidemiologists argued that using a higher cutoff value in the LDPS would likely have led to a larger estimated effect (in absolute value).

We wanted to investigate whether this was invariably true in realistic scenarios, that is, whether a higher cutoff value will always lead to a larger effect estimate (in absolute terms) when the effect on outcome is monotonically increasing with higher values of the continuous explanatory variable. To answer this, we set up a quantitative thought experiment.

2.2 Implementation

In general, the estimated effect size will depend on the distribution of the explanatory variable and the dose–response relationship; thus these two characteristics must be varied to create the relevant scenarios.

Assume that IQ is the outcome of interest and that binge drinking is the binary exposure (yes or no) defined from the actual number of drinks consumed at a single occasion based on a cutoff value. To mimic the actual setting, we consider using five and eight drinks as cutoff values. The distribution of the number of drinks can take many different forms, but here we will consider three simple forms. Either most women only have a few drinks, the women have a uniform distribution of drinks, or most women consume many drinks. Assume without loss of generality for all the settings that 14 drinks is the maximum number of drinks consumed at a single occasion. The uniformly distributed exposure variable can then be generated straightforwardly by taking integer values of uniform random numbers after multiplication with an appropriate factor, here 15.

When most women drink a small number of drinks, the exposure variable can conveniently be generated by raising a uniform random variate to a power larger than 1 before taking the integer value. Similarly, the situation with most women drinking high numbers of drinks can be obtained by choosing a power smaller than 1. Regardless of its distribution, the variable can subsequently be dichotomized into a binary binge variable. For example, the situation with most women having a small intake can be generated as follows:

```
. set obs 1000000
obs was 0, now 1000000
. generate ndrinks = int(runiform()^3*15)
. generate binge5 = ndrinks >=5
. generate binge8 = ndrinks >=8
```

Notice that we chose to generate a very large dataset (n = 1,000,000) so as to virtually eliminate random error in subsequent results. The resulting distribution of exposure is shown in figure 1.

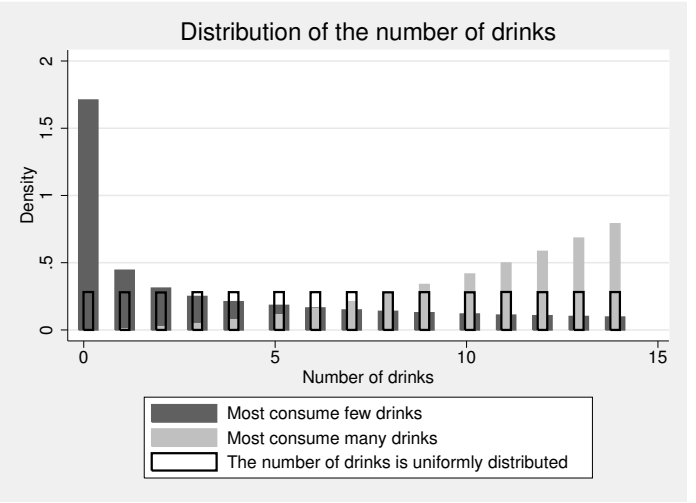

Figure 1. The distribution of the number of drinks generated from 1,000,000 observations as a decreasing, uniform, and increasing integer function ranging from 0 to 14. The three different distributions are generated by raising a uniform random variate to the power of 3, 1, and 0.33, respectively.

2.3 The shape of the dose–response curve

When specifying the dose–response curve between IQ (the response) and the number of drinks (the dose), we consider three different types of dose–response curves, namely, concavely declining, linearly declining, or convexly declining. By definition, the IQ in a general population is defined to follow a normal distribution with a mean of 100 and a standard deviation of 15; so for the unexposed, we assume a higher IQ—say, 105—and we then subtract the effect of alcohol from the mean for higher intakes. Figure 2 shows the shapes of the three types of relationships we considered.

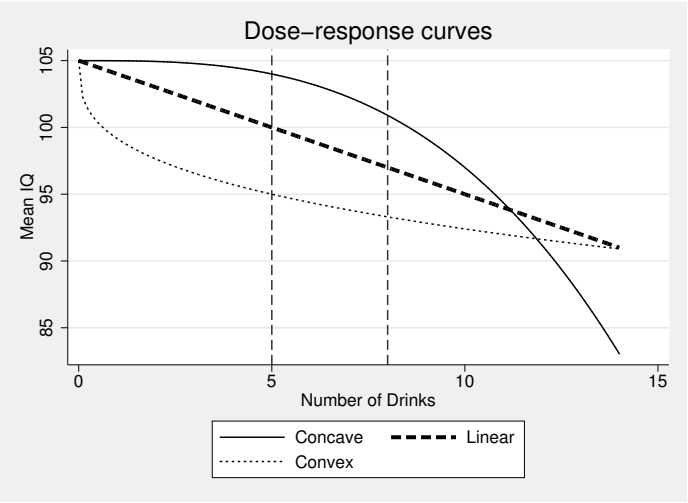

Figure 2. Three different shapes of the dose–response curve, where the mean IQ is plotted against the number of drinks. The vertical lines illustrate the values used for dichotomizing the exposure. Concave: $\text{IQ} = 105 - (x/5)^3$; linear: $\text{IQ} = 105 - x$; convex: $\text{IQ} = 105 - \sqrt[3]{200x}$.

In Stata syntax, the concavely declining relationship is specified as

```
. generate IQ = rnormal() * 15 + 105 - (ndrinks/5)^3
```

2.4 Comparison of the effect sizes in different settings

For the actual implementation, let us first consider the single setting, where most women have a low intake and the shape of the dose–response curve is concavely declining. Because IQ is not, according to the above definition, normally distributed given binge status (mean IQ varies within binge categories), we use robust variance estimation to obtain standard error estimates, which implies that the model must be formulated as a linear regression with IQ as the response variable and binge drinking as a binary covariate. The following output shows the estimated effect sizes for each of the two cutoff values defining binge drinking:

```
. regress IQ i.binge5, vce(robust)
Linear regression                               Number of obs = 1000000
                                                F(  1,999998) =45431.21
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0463
                                                Root MSE      =  15.432
------------------------------------------------------------------------------
             |               Robust
          IQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.binge5 |  -7.376778   .034609  -213.15   0.000    -7.444611   -7.308946
       _cons |   104.9271   .0180216  5822.29   0.000     104.8918    104.9624
------------------------------------------------------------------------------

. regress IQ i.binge8, vce(robust)
Linear regression                               Number of obs = 1000000
                                                F(  1,999998) =68834.50
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0699
                                                Root MSE      =   15.24
------------------------------------------------------------------------------
             |               Robust
          IQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.binge8 |  -10.6854   .0407275  -262.36   0.000    -10.76522   -10.60557
       _cons |   104.6848    .016679  6276.45   0.000     104.6521    104.7175
------------------------------------------------------------------------------
```

In this setting, the effect size is largest in absolute value when binge exposure is defined from the high cutoff of eight drinks.

To efficiently estimate all effect sizes corresponding to the different definitions of the outcome and exposure, we wrap the code into two `foreach` loops:

```
. set seed 198598
. foreach power of numlist 3 1 0.33 {
  2.   set obs 1000000
  3.   generate ndrinks = int(runiform()^`power´ * 15)
  4.   generate dr_concave = - (ndrinks / 5)^3
  5.   generate dr_linear = - ndrinks
  6.   generate dr_convex = - (ndrinks * 200)^(1 / 3)
  7.   generate binge8 = ndrinks >=8
  8.   generate binge5 = ndrinks >=5
  9.   foreach dr_fct of varlist dr_concave dr_linear dr_convex {
 10.     generate IQ = rnormal() * 15 + 105 + `dr_fct´
 11.     regress IQ i.binge5, vce(robust) noheader
 12.     regress IQ i.binge8, vce(robust) noheader
 13.     drop IQ
 14.   }
 15.   drop _all
 16. }
```

(*output omitted*)

2.5 Results

The code above results in nine different linear regression analyses for each of the two different definitions of the exposure; these analyses are listed in table 1.

Table 1. Estimated effect size and robust standard errors from the linear regression of IQ on the dichotomized binge exposure on a large sample (n = 1,000,000) in different settings. For the concave scenario with decreasing distribution of exposure, the results differ slightly from those presented earlier, which is simply the consequence of using different random seeds.

		Distribution of exposure		
Dose–response	Exposure cutoff	Decreasing	Equal	Increasing
Concave	5	−7.38 (0.03)	−8.57 (0.03)	−12.00 (0.08)
	8	−10.75 (0.04)	−10.91 (0.03)	−12.19 (0.04)
Linear	5	−8.06 (0.03)	−7.49 (0.03)	−7.91 (0.08)
	8	−9.23 (0.04)	−7.51 (0.03)	−6.18 (0.04)
Convex	5	−8.95 (0.03)	−6.03 (0.03)	−4.48 (0.08)
	8	−8.71 (0.04)	−5.10 (0.03)	−3.12 (0.04)

As expected, both the shape of the dose–response curve and the distribution of the exposure variable determine which cutoff value yields the larger effect. When the shape of the dose–response is concave, a cutoff value of eight drinks results in the largest effect size in absolute value in this example, but this is reversed when the shape is

convex. When the dose–response relationship is linear, the distribution of the exposure determines which is larger. The conclusion is thus that the magnitude of the estimated effect depends not only on the cutoff value but also on the distribution of the covariate and the shape of its association with the outcome. There is thus no guarantee that choosing a higher cutoff value would have led to a higher effect estimate. Note that the high number of observations (n = 1,000,000) has virtually eliminated the random variation (all standard errors are between 0.03 and 0.08), so any variation in effect sizes can be considered systematic effects.

3 Example 2: Can use of sampling weights be avoided in statistical analyses of complex sampling designs?

3.1 Scientific setting

In the LDPS, the sample was stratified in a complex manner to ensure adequate representation in every stratum. When one analyzes such complex sampling designs, the standard strategy according to the statistical literature is to calculate the sampling fractions for each stratum and use these as inverse probability weights in a weighted analysis with robust variance estimation (see [U] **20.22.3 Sampling weights**). This strategy was followed in the LDPS, but weighting the analyses seemed to decrease precision, which in the research group raised the question of whether an appropriate analysis could be conducted omitting weights.

While the use of sampling weights may in some specific situations be abandoned without inducing bias (Winship and Radbill 1994), this is not true in general. In the LDPS, the focus was on estimating a marginal effect of average alcohol intake while accounting for the sampling categories defined by average and binge drinking but without considering an interaction term between the two. Therefore, the main objective for this example was to illustrate how large the cost could be in terms of bias and reduced coverage probabilities of confidence intervals by omitting weights in such an analysis. We again used a quantitative thought experiment implemented in Stata to answer this.

3.2 Implementation

To mimic the setup for the actual LDPS and yet keep the model simple, we imagine a study where the sample is stratified by both average alcohol intake (four categories: 0, 1–4, 5–8, 9+ drinks per week) and binge drinking (yes or no). Our aim is now to estimate the effect of maternal average alcohol intake on a child's IQ by conducting a linear regression analysis. Note that just as in the real LDPS, the sampling design in our example is based on categorized average intake, but the actual average number of drinks consumed per week is recorded and can be used as a covariate in the statistical analysis.

To define the sample, we must specify the joint distribution of average alcohol intake and binge drinking. Suppose that average alcohol intake per week lies between 0 and 14 drinks with most women having a low intake; that is, the random variable describing average intake is generated as a declining integer function ranging from 0 to 14. Suppose that the probability of binge drinking during pregnancy increases with average alcohol intake such that among those with an average intake of 0, 20% will be categorized as binge drinkers, whereas for those with an average intake of 14, approximately 50% will be categorized as binge drinkers.

So that we can mimic the original setup where the LDPS is a subsample within the Danish National Birth Cohort, a dataset with 100,000 observations is first generated for reference, and from this, all the subsamples are then drawn.

```
. set seed 1508776
. set obs 100000
obs was 0, now 100000
. generate avalco = int(runiform()^3 * 15)
. generate binge = runiform() < (.2 + avalco/(14*2))
```

As in the LDPS, a higher fraction is sampled among those having a high alcohol intake, be it on average or as binging.

```
. recode avalco (0 = 1) (1/4 = 2) (5/8 = 3) (9/20 = 4), generate(alcocat)
(92637 differences between avalco and alcocat)
. generate sampfrac = (alcocat / 10 + binge / 2) / 20
. table alcocat binge, c(mean sampfrac) format(%5.3f)
```

RECODE of avalco	binge 0	binge 1
1	0.005	0.030
2	0.010	0.035
3	0.015	0.040
4	0.020	0.045

Using the sampling weights, we select stratified samples from the cohort of 100,000 observations. In table 2, an example of one of these stratified samples is shown—the sample is more balanced across the strata than the full cohort. In a more realistic sample, it would be likely that more subjects have an average intake of 0, but because the average intake is defined by a simple decreasing function, the first group (0 drinks per week) contains fewer individuals than the second group (1–4 drinks per week) in this hypothetical setup. This is, however, of no consequence for the subsequent results.

Table 2. The distribution of subjects across the strata in the full cohort and in the sample

avalco	binge	Full cohort, n (%)		Stratified sample, n (%)	
1	0	32,401	(32.4)	160	(8.8)
	1	8,206	(8.2)	222	(12.2)
2	0	20,770	(20.8)	205	(11.3)
	1	7,890	(7.9)	287	(15.8)
3	0	8,621	(8.6)	133	(7.3)
	1	6,426	(6.4)	268	(14.7)
4	0	6,171	(6.2)	113	(6.2)
	1	9,515	(9.5)	430	(23.7)
Total		100,000	(100)	1,741	(100)

For the outcome variable IQ, we assume that its mean decreases slightly with increasing alcohol intake and that it is normally distributed with a standard deviation of 15 and a mean of 105 for nondrinkers, just as in the previous example. In Stata, this becomes

```
. generate IQ = rnormal()*15 + 105 - (avalco/7)^3
> - 4 * binge - .4 * (avalco/7)^3 * binge

. save sourcepop, replace
file sourcepop.dta saved
```

Note that the data defining the full cohort is saved in `sourcepop.dta` for later use.

3.3 Bias in unweighted versus weighted analyses

A simple linear regression of IQ on average alcohol intake in the full cohort of 100,000 observations yields an effect estimate of -0.618. We take this to be the true value with which the estimated coefficients in the smaller stratified samples are to be compared.

First, we again construct a simple program that saves the estimated regression coefficient and standard error for the unweighted and weighted analysis, respectively. This can then be fed to `simulate`. Second, the results are evaluated by calculating and summarizing the relative bias and the coverage probability from the 2,500 simulated regression coefficients and standard errors.

```
. * A program selecting a subsample with sample fraction sampfrac
. * and running an unweighted regression of IQ on average alcohol.
. use sourcepop

. program define alcononpw, eclass
  1. preserve
  2. keep if runiform() < sampfrac
  3. regress IQ avalco
  4. restore
  5. end

. simulate _b _se, reps(2500) saving(nonpwres, replace): alcononpw

      command:  alcononpw

Simulations (2500)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
  (output omitted)
..................................................  2500

. * A program selecting a subsample with sample fraction sampfrac,
. * but running a weighted regression of IQ on average alcohol.
. use sourcepop

. program define alcopw, eclass
  1. preserve
  2. keep if runiform() < sampfrac
  3. regress IQ avalco [pw = 1/sampfrac]
  4. restore
  5. end

. simulate _b _se, reps(2500) saving(pwres, replace): alcopw

      command:  alcopw

Simulations (2500)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
  (output omitted)
..................................................  2500

. * The estimated regression coefficient used as the true value.
. preserve

. use sourcepop

. quietly regress IQ avalco

. matrix truecoefs = e(b)

. local trueval = truecoefs[1, 1]

. display `trueval´
-.61761482

. restore
```

```
. ********************************************************************
. * Results
. ********************************************************************
. * Summarizing the results by calculating the relative bias
. * and coverage probability.
. foreach dataset in nonpwres pwres {
  2.   display "Dataset used `dataset´"
  3.   use `dataset´, clear
  4.   generate relbias = (_b_avalco - `trueval´) / `trueval´
  5.   ci _b_avalco
  6.   centile relbias _se_avalco
  7.   generate coverage = (_b_avalco - 1.96 * _se_avalco) <= `trueval´
> & (_b_avalco + 1.96 * _se_avalco) >= `trueval´
  8.   ci coverage, bin
  9. }
Dataset used nonpwres
(simulate: alcononpw)
```

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]
_b_avalco	2500	-.6621544	.0014757	-.6650481 -.6592607

				— Binom. Interp. —	
Variable	Obs	Percentile	Centile	[95% Conf. Interval]	
relbias	2500	50	.0727582	.0672506	.0785631
_se_avalco	2500	50	.0761998	.0761099	.0762694

				— Binomial Exact —
Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]
coverage	2500	.9188	.0054628	.9073955 .9292116

```
Dataset used pwres
(simulate: alcopw)
```

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]
_b_avalco	2500	-.6209438	.0019018	-.624673 -.6172146

				— Binom. Interp. —	
Variable	Obs	Percentile	Centile	[95% Conf. Interval]	
relbias	2500	50	.0066028	.0001758	.0142346
_se_avalco	2500	50	.0967832	.0966078	.0970025

				— Binomial Exact —
Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]
coverage	2500	.9552	.0041373	.9463414 .9629713

From table 3, it is clear that the weighted analysis yields less bias and a coverage probability closer to the nominal value than the unweighted analysis.

Table 3. The mean estimate with its median standard error, median relative bias, and coverage probability based on 2,500 simulations. As a true value, we used the estimate −0.618 obtained in the linear regression in the entire cohort of 100,000.

	Mean (standard error)	Relative bias (%)	Coverage probability with 95% confidence interval
Unweighted	−0.662 (0.076)	7.3	0.919 [0.907, 0.929]
Weighted	−0.621 (0.097)	0.7	0.955 [0.946, 0.963]

We thus conclude that using sample fractions as inverse probability weights substantially reduces bias while maintaining the coverage probability close to the nominal value of 95% even when the model is misspecified. These features are not shared by the unweighted analysis. Although the use of sampling weights results in larger standard errors and less power, the protection against bias outweighs this, and when it is not possible in a study to gain both high power and unbiased estimates, a less precise but unbiased estimate is preferred. As a follow-up (not shown but available upon request), we found in another worked thought experiment that the power for detecting an interaction effect between average intake and binging was low, so a simple strategy for avoiding the use of weights does not exist.

4 Outline of the process for constructing a quantitative thought experiment

In this article, two different approaches have been used to construct quantitative thought experiments. One approach is to generate one very large dataset and use this to compare the estimated effect size in different situations, as done in example 1, with the objective of virtually eliminating random error. The other approach is to generate many datasets of a realistic size in simulations, analyze them separately, and summarize results across them with the objective of estimating bias, precision, or coverage probability, all in finite samples with random error, as illustrated in example 2. For both approaches, the recipe for constructing a quantitative thought experiment is rather similar, as outlined below.

4.1 Generating datasets

After a random-number seed is set to make the results reproducible, the chosen number of observations is generated; this number may either be very large—say, 1,000,000—or reflect realistic sample sizes actually available. Thus the following two lines are typical when initiating a thought experiment:

```
. set seed 2083675
. set obs 1000000
obs was 0, now 1000000
```

The thought experiment should be both realistic, illustrative, and easy to follow. Therefore, it is important to consider how to build up a realistic scenario but still keep it simple: do not include more variables than necessary, keep distributions of the variables as simple as possible, etc.

Often the only two variables needed to be defined are the exposure and the outcome. Before constructing the variables, one should define some characteristics for each of the variables. What should the range be? Is it categorical or continuous? What shape does its distribution have? Is it normally distributed or skewed, and what is the relation between the variables? Because outcome typically depends on the exposure variable, the exposure variable must be defined first. The simplest way of generating a random variable with a given distribution is by inversion of its distribution function, because the procedure can then be based on generating uniform random variates. For example, an exposure with a range from 0 to 10 and with a decreasing distribution function can be defined as $F(x) = \{(1/10)x\}^{1/2}$, which in Stata becomes

```
. generate exposure = runiform()^2*10
```

To check whether the distribution of the variable is as desired, we suggest plotting a histogram of the variable.

The outcome is then defined based on the expected association with exposure. For example, the association may be as a normally distributed outcome with mean 50 and a standard deviation of 10 for the unexposed, $X = 0$, and linearly decreasing with exposure:

```
. generate outcome = rnormal() * 10 + 50 - exposure
```

The association can, of course, be more complicated, and the outcome may depend on more than one variable; see example 2.

4.2 Estimation/simulation

When the scenario is established, the generated dataset can be used to investigate different expected properties of the data. For the first approach with a single large dataset, the only result of interest is the estimated effect size. This estimate can straightforwardly be found with a single estimation command, for example, a simple linear regression with

robust standard errors, to account for departures from normality. By using a very large dataset, we ensure that random variation is virtually eliminated (see example 1):

```
. quietly regress outcome exposure, vce(robust)
```

For the other approach, where several datasets are generated and analyzed, more steps are needed to obtain and present results. For each of the characteristics, the following procedure should be followed:

1. Choose the number of simulations.
2. Define a program.
3. Run the simulations on the defined program.
4. Present the results.

In the following, we give a short description of the procedure for simulating the bias and the coverage probability. For both, it is essential to define the true value, and if this cannot be determined analytically, it could instead be taken as the estimate found in a very large simulated dataset, where random error is negligible.

Simulating the bias

As in example 2, the bias is estimated as the relative difference between the estimated effect and the true value. The true value is here taken to be the estimate in the generated dataset containing 1,000,000 observations. The number of simulations is not easy to determine beforehand but can be adjusted according to the resulting standard error of the bias. If we wish to get a standard error of the relative bias of 0.005, we could simply keep increasing the number of simulations until we reach this precision. Next the program used to actually estimate the parameter is defined. In this example program, the estimation is done on a simple 1% random sample of the observations, and the results are saved by specifying the `eclass` option. Strictly speaking, the observations are not independent, but given the large size of the source dataset, this is a problem that may be ignored. The program could look like this:

```
. program define pr_est, eclass
  1. preserve
  2. keep if runiform() < 0.01
  3. regress outcome exposure, vce(robust)
  4. restore
  5. end
```

Using the `simulate` command in Stata, the program is applied several times—say, 2,500—and the results of each simulation are saved in a new dataset.

```
. simulate _b, reps(2500) saving(datares, replace): pr_est
      command:  pr_est
Simulations (2500)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
  (output omitted)
..................................................  2500
```

To evaluate the median relative bias, we calculate the relative bias from each of the 2,500 saved effect estimates. The mean of these and their corresponding confidence intervals are estimated and presented with the `centile` command.

```
. local trueval = -0.9944
. generate relbias = (_b_exposure -`trueval´)/`trueval´
. centile relbias
                                                      -- Binom. Interp. --
    Variable |      Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
     relbias |     2500          50      .0007486       -.000921    .0025195
```

Simulating the coverage probability

Because the coverage probability is an estimated proportion with a nominal value of 95%, it can be shown from the formula for the standard error that if we choose the number of simulations to be 2,500, this will result in a standard error of less than 0.5%.

To present the coverage probability, we must generate a variable that records whether the true value is within the computed 95% confidence interval. If we use the `ci` command with the binary option on this variable, the coverage probability can be estimated and compared with the nominal value, say, 95%:

```
. simulate _b _se, reps(2500) saving(datares, replace): pr_est
      command:  pr_est
Simulations (2500)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
  (output omitted)
..................................................  2500
. local trueval = -0.9944
. generate cp = (_b_exposure - 1.96 * _se_exposure) <= `trueval´
> & (_b_exposure + 1.96 * _se_exposure) >= `trueval´
. ci cp, binomial
                                                      -- Binomial Exact --
    Variable |       Obs        Mean    Std. Err.      [95% Conf. Interval]
-------------+-------------------------------------------------------------
          cp |      2500        .954    .0041897       .9450405    .9618754
```

5 Discussion

In this article, we presented two examples of how to use Stata for answering questions that could otherwise easily result in extensive speculation and debate. The main prerequisite for both examples is a rigorous specification of the scenario of interest. Invariably, we started by formulating a full statistical model for the outcome, its relationship with explanatory variables, and their distribution. Values for these are often naturally available from the actual application in which the discussions arise, so it is possible to accept arguments at face value, code them in Stata language, and simply observe what the results are.

There are, however, a few caveats to this seemingly limitless inventory of tools. The choice of contrasting scenarios often requires experience and intuition that allow one to select the features that need to be varied to obtain general results. If the scenarios do not span the relevant variation, it is easy to become misled by seemingly general results. A case in point is the choice of shapes for dose–response curves in example 1. Had the concave curve been omitted, one might have concluded that departures from a linear relationship always led to larger effect estimates with lower cutoff values.

The major advantage of using worked thought experiments is the ability to engage constructively in an ongoing development of thoughts within the study group. When, for example, the results of example 2 are presented—the complex stratified design leads to a lower precision than one would find if just running an ordinary analysis ignoring sampling weights—the following question may naturally arise: What precision would be anticipated in other studies by using the data from the LDPS, but with, say, different exposure variables from other register data? Such a question would lend itself directly to a new thought experiment where the observed sampling fractions are used to weight the anticipated analysis.

We thus hope the examples may serve as inspiration for applied statisticians who need to engage with subject matter specialists and provide intelligible guidance to them on the statistical planning and analyses on a given project. In our own practice, we have found that once we adopted this strategy, it quickly became compelling and widespread because it is very flexible, convenient, and applicable to a huge range of situations.

6 Acknowledgment

The authors want to thank the research group in the Lifestyle During Pregnancy Study for lively and engaging discussions on application of statistical methods, in particular Erik Lykke Mortensen and Ulrik Schiøler Kesmodel.

7 References

Feiveson, A. H. 2002. Power by simulation. *Stata Journal* 2: 107–124.

Kesmodel, U. S., M. Underbjerg, T. R. Kilburn, L. Bakketeig, E. L. Mortensen, N. I. Landrø, D. Schendel, J. Bertrand, J. Grove, S. Ebrahim, and P. Thorsen. 2010. Lifestyle during pregnancy: Neurodevelopmental effects at 5 years of age. The design and implementation of a prospective follow-up study. *Scandinavian Journal of Public Health* 38: 208–219.

Olsen, J. 1994. Effects of moderate alcohol consumption during pregnancy on child development at 18 and 42 months. *Alcoholism: Clinical and Experimental Research* 18: 1109–1113.

Olsen, J., M. Melbye, S. F. Olsen, T. I. Sørensen, P. Aaby, A.-M. N. Andersen, D. Taxbøl, K. D. Hansen, M. Juhl, T. B. Schow, H. T. Sørensen, J. Andresen, E. L. Mortensen, A. W. Olesen, and C. Søndergaard. 2001. The Danish National Birth Cohort—Its background, structure and aim. *Scandinavian Journal of Public Health* 29: 300–307.

Senn, S., and S. Julious. 2009. Measurement in clinical trials: A neglected issue for statisticians? *Statistics in Medicine* 28: 3189–3209.

Winship, C., and L. Radbill. 1994. Sampling weights and regression analysis. *Sociological Methods and Research* 23: 230–257.

About the authors

Theresa Wimberley has a master's degree in statistics from the University of Aarhus. Between 2009 and 2011, she worked as a research assistant, particularly on the Lifestyle During Pregnancy Study, in the Department of Biostatistics at the University of Aarhus. Since 2012, she has been employed at the National Centre of Register-based Research as a PhD student, doing research within the field of pharmacoepidemiology.

Erik T. Parner has a PhD in statistics from the University of Aarhus. He is a professor of biostatistics at the University of Aarhus. His research fields are time-to-event analysis, statistical methods in epidemiology and genetics, and the etiology and changing prevalence of autism.

Henrik Støvring has a PhD in biostatistics from the University of Southern Denmark. He is an associate professor of biostatistics at the University of Aarhus. His research fields are statistical methods in pharmacoepidemiology, risk communication to patients, and the use of health care services. He has served as a senior statistician on the Lifestyle During Pregnancy Study.