Double takes: some statistical examples with surprise twists

Andrew Gelman* and Deborah Nolan†

June 21, 2001

Abstract

We present five statistics examples with surprise twists, on topics ranging from descriptive statistics to sampling to multiple regression. Some of these examples are well known; our contribution here is to collect them in a convenient place and to discuss in some detail how we present them to a class in an interactive manner.

1 Introduction

After teaching statistics for several years, we all develop our own collection of favorite examples. We collect here some examples we have encountered that have surprise twists. In addition to being fun, the shock of confusion, followed by recognition, is a good way for students to learn and remember the material. In this article, we present the examples in the order that we do them in the class.

2 World record times for the mile run

At the beginning of our introductory course, we teach descriptive statistics, the simplest example of which are time plots. We illustrate the difficulties of extrapolation with a low-tech multimedia trick using a transparency projector and a blackboard.

We blow up and photocopy Figure 1 on a transparency and project it on the blackboard, but with the right half of the graph covered, so that the students see the world record times for the mile run from about 1900 to 1950. We then ask how well a straight line fits the data (reasonably well) and get a student to come up and draw a straight-line fit with chalk on the blackboard, extending to the year 2000 and beyond. We discuss interpolation and extrapolation: is the straight-line prediction for the year 2100 reasonable? What about the year 2000? The students agree that these extrapolations will be too optimistic since there is some minimum time that will never be attained, and they agree on an extrapolated curve that flattens out a bit between 1950 and 2000. We draw this extrapolation on the blackboard, and then prepare to reveal the covered-up half of the curve, so that the students can compare their extrapolation to what actually happened.

*Department of Statistics, Columbia University, New York
†Department of Statistics, University of California, Berkeley

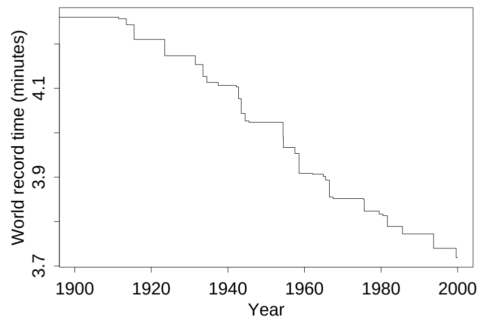

Figure 1: World record times in the mile run since 1900. Blow up this graph onto a transparency and display it on the blackboard with the portion after 1950 covered up. Then ask the students to extrapolate to the year 2000. It is surprising that the approximately linear trend continues unabated.

We then reveal the right half of the curve, and the students do a double-take. Amazingly, the linear extrapolation fits pretty well all the way to the year 2000—in fact, the time trend is actually slightly steeper for the second half of the century than for the first half. We were pretty shocked when we first saw this. Of course, we wouldn't expect this to work out to the year 2050 . . .

The mile run example is also discussed in Anderson and Loynes (1987, 128–130). If your students might be interested in other work on world record times of men and women in running and swimming, see Wainer et al. (2000).

3 Heights of men and women

A simple and standard example of the normal distribution is the height of adult men or women (in the United States, men's heights have mean $5'9.1''$ and standard deviation $2.9''$, women's have mean $5'3.7''$ and standard deviation $2.7''$, and both distributions are well-approximated by the normal. (The normal distribution is understandable here if we think of the height of a man or woman as a sum of many small factors, so that the Central Limit Theorem applies.) The numbers and more discussion of the normal model appear in Brainard and Burmaster (1992).

When students are asked to sketch the distribution of heights of adults, they often start with a bell-shaped curve. But when they are reminded that the population is divided into men and women, they almost always draw a strongly bimodal picture. We then show them the combined distribution of the heights of all adults which, surprisingly, has only one mode—the means of the two sexes are close enough that the modes overlap. See Figure 2.

The height distributions appear in Brainard and Burmaster (1992). Schilling et al. (2001) provide a detailed discussion of why there is a common misconception that the distribution of heights is bimodal.

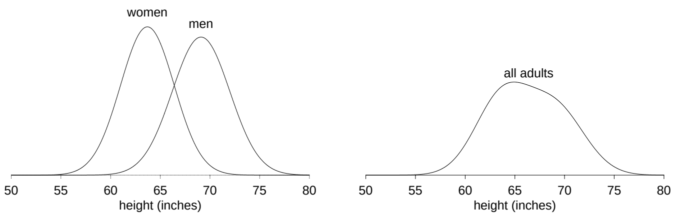

Figure 2: (a) Distributions of heights of adult men and women in the United States. (b) Distribution for all adults (52% women, 48% men). Surprisingly, the combined distribution has only one mode because the mean heights of men and women are so close together.

Family size (# of siblings, including self)	Count
1	3
2	3
3	5
4	5
5	3
6 or more	0

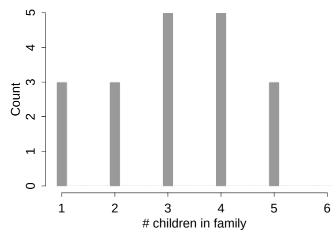

Figure 3: Data on number of children in families (displayed as frequency table and histogram) from students in a typical class. The average is 3.1, which is at first a surprise, given that the average family has about 2 children. See Section 4 for the explanation.

4 How large is your family?

At some point when we are covering sampling, we ask the students to each write down how many children are in his or her family ("How many brothers and sisters are in your family, including yourself?"). We write the results on the blackboard as a frequency table and a histogram, and then compute the mean, which is typically around 3 (see Figure 3 for an example).

We tell the students that the average number of children in families that were having children 20 years ago (about the age of the students in the class) was about 2.0. Why is the number for this class so high? Students give various suggestions such as, perhaps larger families are more likely to send children to college. After some discussion, a student notes that if the family had zero children, they certainly did not send any to college. The 2.0 figure is the average number of children when sampling *by family*; 3.0 is the average number of children when sampling *by child*. When sampling by child, a family with n children is n times more likely to be sampled than a family with 1 child. This illustrates the general point that it is not enough to say you sampled at random; you must also know the method of sampling. It can also be considered as an example of *sampling bias*. We sometimes also ask the students to provide the number of children in his or her oldest uncle's family, which gives an estimate of family size when sampling by family. Although some form of sampling bias still exists, the mean number of uncle's children is closer to 2. For example, once when we asked this question, we found the mean number of children for those students with uncles who have children to be around 2.3.

This example appears in Gelman and Glickman (2000) and also appears in Madsen (1981). Size-based sampling in waiting times is a standard example in probability theory; a classroom demonstration based on this idea appears in Scheaffer et al. (1996, pp. 157–160). Also see Anderson and Loynes (1987, pp. 123–124).

5 Weighing a "random" sample

It is well known among statisticians that when you take a "haphazard" sample without using any formal probability sampling, you are likely to oversample the more accessible units. We have found students respond well to the following demonstration based on estimating the weight of a collection of objects.

We pass around the room a small digital kitchen scale along with a bag which, we (truthfully) tell the class, we filled ahead of time with 100 wrapped candies of different sizes (for example, 20 full-sized candy bars and 80 assorted small candies). The class is divided into pairs, and each pair of students is instructed to estimated the total weight of the candies in the bag by first selecting a "representative or random sample" of 5 candies out of the bag, then weighing the sample and multiplying by 20 to estimate the total weight. Each pair of students is told to write down their measurement and estimate silently (so as not to influence the other students), then put their sample back in the bag, shake up the bag, and pass to the next pair. At the end, we tell the students, we will weigh the bag, and whoever has the closest estimate will get to keep all the candy.

This demonstration takes about two minutes to explain, and then it proceeds while the lecture takes place, thus giving all the students the opportunity to participate without taking away lecture time. As usual, we have the students work in pairs so they will focus more consciously on the task.

When all the pairs of students have weighed their candies, we then ask each pair of students to state their estimated total weight; we write all these weights on the blackboard and then display them as a histogram, as shown in Figure 4. This histogram illustrates the *sampling distribution* of the estimated weights. We then get another student to weigh the entire bag (a digital kitchen scale with enough accuracy to weigh 5 candies and enough range to weigh all 100 can be bought for about $50) and state the total weight. It is invariably lower than most or even all of the sample-based estimates, and this shocks the students.

Why did this happen? The students realize that the larger candies are more accessible (and also

Estimated weights		
2040	2600	2760
3240	2600	4200
3500	2827.7	3500
4200	2840	3060
2640	3600	2640
2880	2520	

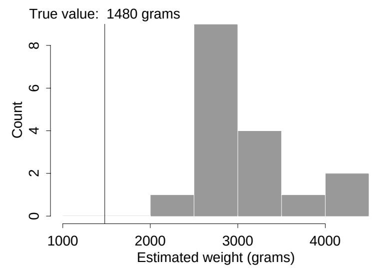

Figure 4: Results of 17 pairs of eighth-graders independently estimating the weight of a bag of 100 candies of varying size by selecting a sample of 5, weighing them (and returning them to the bag), and multiplying the weight by 20. (The estimate of 2827.7 was from a student trying something clever.) The histogram represents the *sampling distribution* of the estimates, and the spread in the histogram shows the *variance* of the estimate. The true weight was 1480 grams; thus the sampling distribution also includes a large *bias*. Students tend to overestimate the weight, even when they are motivated to guess accurately, because it is easier to grab the larger candies out of the bag.

are more likely to remain on the top of the bag after it has been shaken). Even though they tried to get a representative or random sample, they could not help oversampling the large candies.

This example leads to the topic of random sampling. We begin the discussion by asking the students how they would take a random sample of size 5 from the 100 candies. In addition, this is an excellent way to introduce the concepts of bias and variance of a sampling distribution: we indicate the bias of the sampling distribution by drawing a vertical line at the true weight on the histogram of estimates. We can then discuss how the bias and variance would change if (a) we switched to a random sampling approach, or (b) we increased the sample size from 5 to 10 or 20.

Variations of the subjective sampling demonstration appear in the literature; for example, Scheaffer et al. (1996, pp. 149–156) present a version based on rectangles drawn on paper.

6 Tall people have higher incomes

We use the following example to explain lurking variables in the context of linear regression. As with our other examples, there is a surprise twist at the end.

Before class begins, we set up Figure 5 on a transparency and project onto the blackboard. We trace the lines and label the axes of the graph, then turn off the projector, but leave it in place on the table.

We begin by asking students if they think that taller people have higher earnings (that is, income excluding unearned sources such as interest income). If so, by how much? We draw on the blackboard a pair of axes representing earnings and height, and a point at $(66.5, 20000)$: the average height of adults in the United States is about $5'6.5''$ and their average earnings (in 1990) were about \$20,000. We then draw a line through this central point with slope 1560 (we carefully draw this by drawing

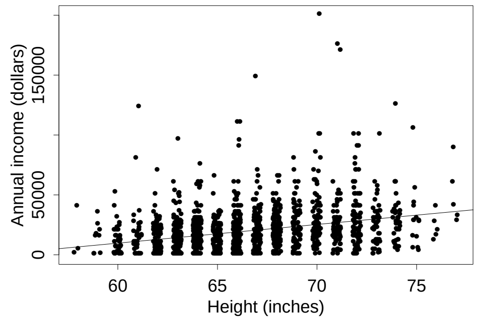

Figure 5: Annual earnings (in thousands of dollars) vs. height (in inches) for a random sample of adult Americans in 1990, along with the least-squares regression line. The heights have been jittered slightly so the points do not overlap.

the line through the point $(56.5, 20000 - 10 * 1560)$ and $(76.5, 2000 + 10 * 1560))$.

We explain that this line has equation $y - 20000 = 1560(x - 66.5)$, or $y = -84000 + 1560x$. What are the interpretations of -84000 and 1560? We tell the students that this is the regression line predicting earnings from height.

We now ask the students to work in pairs and sketch a scatterplot of data that are consistent with this regression line. They need one more piece of information: the standard deviation of the residuals, which is 19000. We show this on the graph by two dotted lines lines, parallel to the regression line, with one line 19000 above and the other line an equal distance below. Approximately 68% of the data should fall in this region, but plotting the points is tricky given that earnings cannot be negative.

We then turn on the projector and display the graph of the actual survey data (Figure 5) on the blackboard. (We downloaded the data from the Work, Family, and Well-Being Survey (Ross, 1990).) On the scale of the actual data, the regression slope ($1560 per inch of height) is small but undeniably positive (the regression output in Stata format appears below):

```
. regress earn height
  Source |       SS       df       MS                  Number of obs =    1379
---------+------------------------------               F(  1,  1377) =  137.21
   Model |  4.8773e+10     1  4.8773e+10               Prob > F      =  0.0000
Residual |  4.8948e+11  1377   355470204               R-squared     =  0.0906
---------+------------------------------               Adj R-squared =  0.0900
   Total |  5.3826e+11  1378   390606004               Root MSE      =   18854
------------------------------------------------------------------------------
    earn |      Coef.   Std. Err.       t     P>|t|       [95 Conf. Interval]
---------+--------------------------------------------------------------------
  height |   1563.138   133.4476     11.713   0.000       1301.355    1824.92
   _cons |  -84078.32   8901.098     -9.446   0.000      -101539.5   -66617.15
------------------------------------------------------------------------------
```

We then ask the students how do we interpret the constant term in the regression: that is, the value -84000 in the equation, $y = -84000 + 1560x$ (see the regression table above)? The answer is, -84000 is the y-value of the regression line where $x = 0$—that is, predicted value of income for an adult who is zero inches tall. In this example, such an extrapolation is meaningless. That is why we prefer to work with the form, $y = \bar{y} + b(x - \bar{x})$.

The next question is what does this mean—why do taller people have higher earnings? The students realize that men are, on average, taller than women and tend to make more money; thus, sex is a lurking variable.

The "lurking variable" story makes a lot of sense here, but it does not tell the whole story! We see this by running the multiple regression of earnings on height and sex:

```
. regress earn height sex
  Source |       SS       df       MS                  Number of obs =    1379
---------+------------------------------               F(  2,  1376) =  101.73
   Model |  6.9335e+10     2  3.4667e+10               Prob > F      =  0.0000
Residual |  4.6892e+11  1376   340785180               R-squared     =  0.1288
---------+------------------------------               Adj R-squared =  0.1275
   Total |  5.3826e+11  1378   390606004               Root MSE      =   18460
------------------------------------------------------------------------------
    earn |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  height |   550.5448   184.5701      2.983   0.003       188.4756    912.614
     sex |  -11254.57   1448.892     -7.768   0.000      -14096.85   -8412.295
   _cons |   1617.938    14059.5      0.115   0.908      -25962.44    29198.32
------------------------------------------------------------------------------
```

By including sex in the regression, we have reduced the coefficient of height by about a factor of 3, but it is still positive and statistically significant. Apparently, tall people do, on average, earn more, even after controlling for sex.

At the conclusion of the discussion, we hand out copies of the Stata log file, which shows what we had to do to clean the data file before running the regressions. They can also use this as a template when doing their computer homework assignments. We explain to the students that linear regression is not the best model for this sort of data (economists might use a logarithmic model, or a tobit regression), but it is in some ways more useful to illustrate the concept in an example for which it is not completely appropriate.

7 Discussion

We have shown how we make examples with twists into activities in which the students participate. Using data collected on themselves (Section 4 on the number of siblings) or collected by them (Section 5 on weighing a random sample) involves students directly. In examples where students do not generate data, we have them work in pairs to find estimates (Section 2 on the mile run), discuss reasons for the phenomenon observed (Section 6 on the regression of height and income), or sketch plots (Section 3 on heights of men and women). With this approach, the students process the problem before we tell the story, and when we reveal the surprise twist, we find the students are now intrigued with the problem. They want to find our what went wrong with their seemingly correct procedure, why their data do not match the standards, or what was the flaw in their logic. We are careful to leave time to help them figure it out. The impact of the surprise twist is lost if the revelation is carried over to the next class meeting.

Although we enjoy surprising students with these examples, we avoid giving them too many of these types of examples. If there are too many, they lose their impact, and students become distrustful of the material. But overall, we have found that double takes can be effective at promoting classroom discussion, and they can help students remember a statistical lesson.

References

Anderson, C. W., and Loynes, R. M. (1987). *The Teaching of Practical Statistics*. New York: Wiley.

Brainard, J., and Burmaster, D. E. (1992). Bivariate distributions for height and weight of men and women in the United States. *Risk Analysis* **12**, 267–275.

Gelman, A., and Glickman, M. E. (2000). Some class-participation demonstrations for introductory probability and statistics. *Journal of Educational and Behavioral Statistics* **25**, 84–100.

Gelman, A., and Little, T. C. (1998). Improving upon probability weighting for household size. *Public Opinion Quarterly* **62**, 398–404.

Madsen, R. W. (1981). Making students aware of bias. *Teaching Statistics* **3**, 2–5.

Ross, C. E. (1990). Work, family, and well-being in the United States. Survey data available from Inter-university Consortium for Political and Social Research, Ann Arbor, Mich.

Scheaffer, R. L., Gnanadesikan, M., Watkins, A., and Witmer, J. (1996). *Activity-Based Statistics: Instructor Resources*. New York: Springer-Verlag.

Schilling, M. F., Watkins, A. E., and Watkins, W. (2001). Is human height bimodal? Submitted to *The American Statistician*.

Wainer, H., Njue, C., and Palmer, S. (2000). Assessing time trends in sex differences in swimming and running (with discussion). *Chance* **13** (1), 10–21.