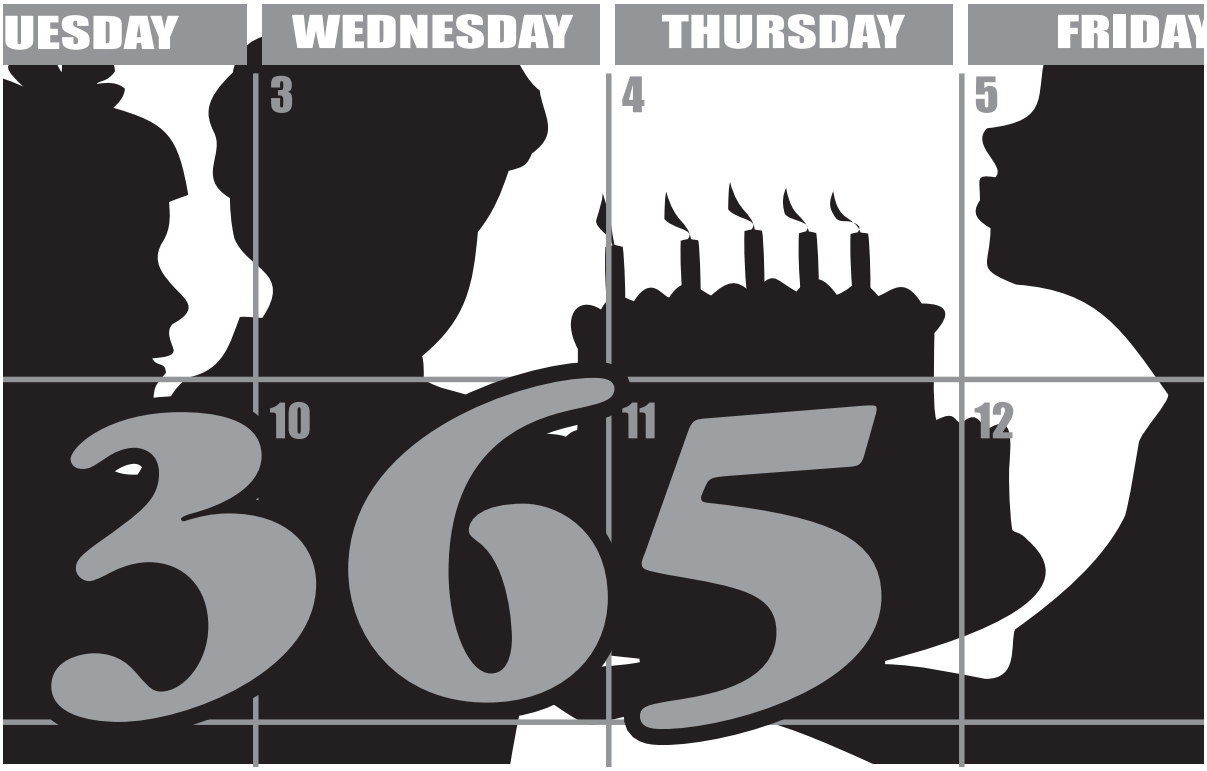
THE MAGAZINE FOR STUDENTS OF STATISTICS

SPRING 2005 • ISSUE # 43

STATS STATS STATS STATS STATS



Computing the Probabilities of MATCHING BIRTHDAYS

How to Make Make Millions on eBay

McDonald's French Fries—Large or Small?

PC Versus Mac

Statistics the EESEE Way!

Non Profit Org-
U.S. Postage
PAID
Permit No. 361
Alexandria,VA

Did you recently complete ***your statistics degree?***

Postgraduate Members pay only **$40**

For the first year after your graduation, you can join the ASA for only $40. That is over 50% off the regular ASA membership rate!

Postgraduate members receive discounts on all meetings and publications, access to job listings and career advice, as well as networking opportunities to increase your knowledge and start planning for your future in statistics.

JOIN NOW!

To request a membership guide and an application, call 1 (888) 231-3473 or join online at

www.amstat.org/join.

STATS

The Magazine for Students of Statistics

Spring 2005 • Number 43

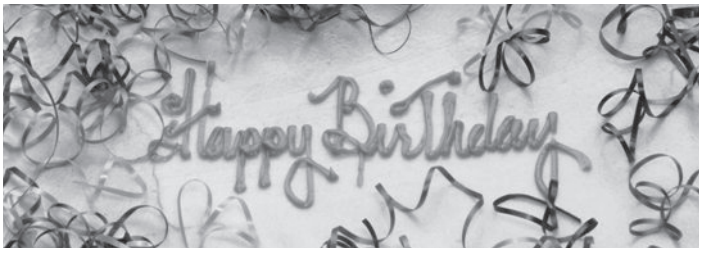
STATS: The Magazine for Students of Statistics (ISSN 1053-8607) is published three times a year, in the winter, spring, and fall, by the American Statistical Association, 1429 Duke St., Alexandria, VA 22314-3415 USA; (703) 684-1221; fax (703) 684-2036; Web site *www.amstat.org*.

STATS is published for beginning statisticians, including high school, undergraduate, and graduate students who have a special interest in statistics, and is distributed to student members of the ASA as part of the annual dues. Subscription rates for others: $13.00 a year for members; $20.00 a year for nonmembers.

Ideas for feature articles and material for departments should be sent to the Editors; addresses of the Editors and Editorial Board are listed above. Material can be sent as a Microsoft Word document or within an email. Accompanying artwork will be accepted in graphics format only (.jpg, etc.), minimum 300 dpi. No material in WordPerfect will be accepted.

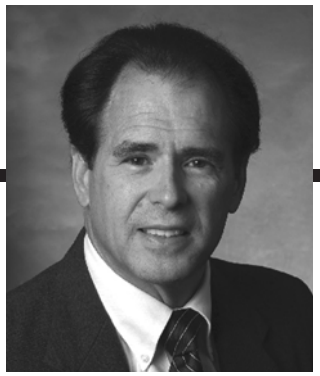
Requests for membership information, advertising rates and deadlines, subscriptions, and general correspondence should be addressed to *STATS* at the ASA office.

Features



Departments

EDITOR'S COLUMN



Paul J. Fields

In a group of people, have you ever wondered if anyone else in the group has your same birthday? Our lead article in this issue of *STATS* takes a look at how to answer that question and how to solve similar probability problems using simulation analysis.

Simulation is a powerful statistical technique. Often, real-world problems that are too difficult, too complex or too large to solve otherwise can be solved effectively and efficiently with simulation. Yet even for relatively simple problems, such as calculating the probability of people having matching birthdays, it is an extremely useful technique.

The flexibility and simplicity of the statistical software R make simulation an even more attractive tool. Bruce E. Trumbo, Eric A. Suess, and Clayton W. Schupp provide a step-by-step explanation of the basic concepts of simulation and how they can be implemented easily using R.

Visit their web site or the *STATS* web site to learn more about simulation, R, and the birthday matching problem. Try it, and have some fun. Then, send us an email message at *STATS* and tell us what interesting problems you have solved using simulation.

There are interesting and fun questions everywhere we can answer with statistics. How about this one: Can statistics help us estimate the value of items offered for sale on eBay? Kim Robinson, Eric Kamischke, and Josh Tabor illustrate using regression analysis to estimate the expected prices of violins for sale on the internet.

Nathan Wetzel and his students were curious about the french fries at McDonald's. They describe their project, which asked the question, "How do the masses of small and large orders of fries at McDonald's compare to the target values listed on the McDonald's web site?" Check it out!

Where do you stand on the PC versus Mac question? Alex Kim, Robert Lord, and Abdallah Salam—three students at Harvard—conducted a survey of the computer preferences of first-year Harvard students. They wondered, "What are the factors that influence students' computer preferences?" Be sure to read their article and see what they concluded from their study.

Chris Sroka, a PhD student in statistics, describes a great project at Ohio State—the *Electronic Encyclopedia of Statistical Examples and Exercises (EESEE)*. It is a library of interesting stories using statistics, and it contains a wealth of statistics examples based on real data. There are answers to questions about dogs, waiters, pizza, and much more. Visit the *EESEE* web site, *www.stat.ohio-state.edu/~eesee*, and see what's inside. Then, you can do statistics the EESEE way!

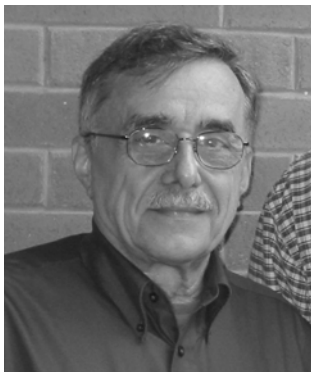
Under AP Statistics, Peter Flanagan-Hyde lists twelve types of variables we use in statistical analysis. That's a lot of variables! You might have asked, "How can I keep all of these variable straight?" He helps us by explaining the meaning and use of each of them.

This issue of *STATS* introduces a new department, *Ask STATS*. Jackie Miller is ready to take on your statistical questions and find the answers. So, send her an email with an interesting question.

In Chris Olsen's Statistical μ-sings, he poses the question, "What does it mean to be rational?" He reviews for us some of the history of probability as people have endeavored to make rational decisions in law, science, and politics. I am sure you will find his article as fascinating as I did.

Paul J. Fields

SIMULATION: Computing the Probabilities of Matching Birthdays



Bruce Trumbo

Eric Suess

Clayton Schupp

The Birthday Matching Problem

Sometimes the answers to questions about probabilities can be surprising. For example, one famous problem about matching birthdays goes like this: Suppose there are 25 people in a room. What is the probability two or more of them have the same birthday? Under fairly reasonable assumptions, the answer is greater than 50:50—about 57%.

This is an intriguing problem because some people find the correct answer to be surprisingly large. Maybe such a person is thinking, "The chance anyone in the room would have my birthday is very small," and leaps to the conclusion that matches are so rare one would hardly expect to get a match with only 25 people. This reasoning ignores that there are $(25 \times 24)/2 = 300$ pairs of people in the room that might yield a match. Alternatively, maybe he or she correctly realizes, "It would take 367 people in the room to be absolutely sure of getting a match," but then incorrectly concludes 25 is so much smaller than 367 that the probability of a match among only 25 people must be very low. Such ways of thinking about the problem are too fuzzy-minded to lead to the right answer.

As with most applied probability problems, we need to start by making some reasonable simplifying assumptions in order to get a useful solution. Let's assume the following:

• *The people in the room are randomly chosen.* Clearly, the answer would be very different if the people were attending a convention of twins or of people born in December.

• *Birthdays are uniformly distributed throughout the year.* For some species of animals, birthdays are mainly in the spring. But, for now at least, it seems reasonable to assume that humans are about as likely to be born on one day of the year as on another.

• *Ignore leap years and pretend there are only 365 possible birthdays.* If someone was born in a leap year on February 29, we simply pretend he or she doesn't exist. Admittedly, this is not very fair to those who were "leap year babies," but we hope it is not likely to change the answer to our problem by much.

The Solution Using Basic Probability

Based on these assumptions, elementary probability methods can be used to solve the birthday match problem. We can find the probability of no matches by considering the 25 people one at a time. Obviously, the first person chosen cannot produce a match. The probability that the second person is born on a different day of the year than the first is $364/365 = 1 - 1/365$. The probability that the third person avoids the birthdays of the first two is $363/365 = 1 - 2/365$, and so on to the 25th person. Thus the probability of avoiding all possible matches becomes the product of 25 probabilities:

$$P(\text{No Match}) = \prod_{i=0}^{24}\left(1 - \frac{i}{365}\right) = \frac{P_{25}^{365}}{365^{25}} = 0.4313$$

Bruce Trumbo (btrumbo@csueastbay.edu) is Professor of Statistics and Mathematics at California State University, East Bay (formerly CSU Hayward). He is a Fellow of ASA and holder of the ASA Founder's Award.

Eric Suess (esuess@csueastbay.edu), Associate Professor of Statistics at CSU East Bay, has used simulation methods in applications from geology to animal epidemiology.

Clayton Schupp, (cschupp@walk.ucdavis.edu) an MS student at CSU East Bay when this article was written, is currently a PhD student in statistics at the University of California, Davis.

since 365^{25} is the number of possible sequences of 25 birthdays and

$$P_{25}^{365} = 25!\binom{365}{25}$$

is the number of permutations of 365 objects taken 25 at a time, where repeated objects are not permitted. Therefore,

$$P(\text{At Least 1 Match}) = 1 - P(\text{No Match}) = 1 - 0.4313 = 0.5687$$

William Feller, who first published this birthday matching problem in the days when this kind of computation was not easy, shows a way to get an approximate result using tables of logarithms. Today, statistical software can do the complex calculations easily, and even some statistical calculators can do the numerical computation accurately and with little difficulty.

```
> prod(1 - (0:24)/365)
[1] 0.4313003

> factorial(25)*choose(365, 25)/365^25
[1] 0.4313003
```

Figure 1: Two ways to calculate the probability of no matching birthdays among 25 people selected at random

In Figure 1, we show two ways to use the statistical software R to calculate the probability of no matches.

Of course, different values of n would give different probabilities of a match. With a computer package like R that has built-in procedures for doing probability computations and making graphs, we easily can loop through various values of n to graph the relationship between n and P(At Least 1 Match). Figure 2 shows the small amount of R code required, and Figure 3 shows the resulting plot. (The title and the reference lines were added later.)

```
p <- numeric(50)
for (n in 1:50) {
          q <- 1 - (0:(n - 1))/365
          p[n] <- 1 - prod(q) }
plot(p)
```

Figure 2: R code to calculate the probability of matching Birthdays when the number of people in the room ranges from 1 to 50

By looking at the plot, we see the probability of at least one match increases from zero to near one as the number of people in the room increases from 1 to 50. We can see that n = 23 is the smallest value of n for which P(At Least 1 Match) exceeds 1/2. The computations show the probability for n = 23 to be 0.5073. A room with

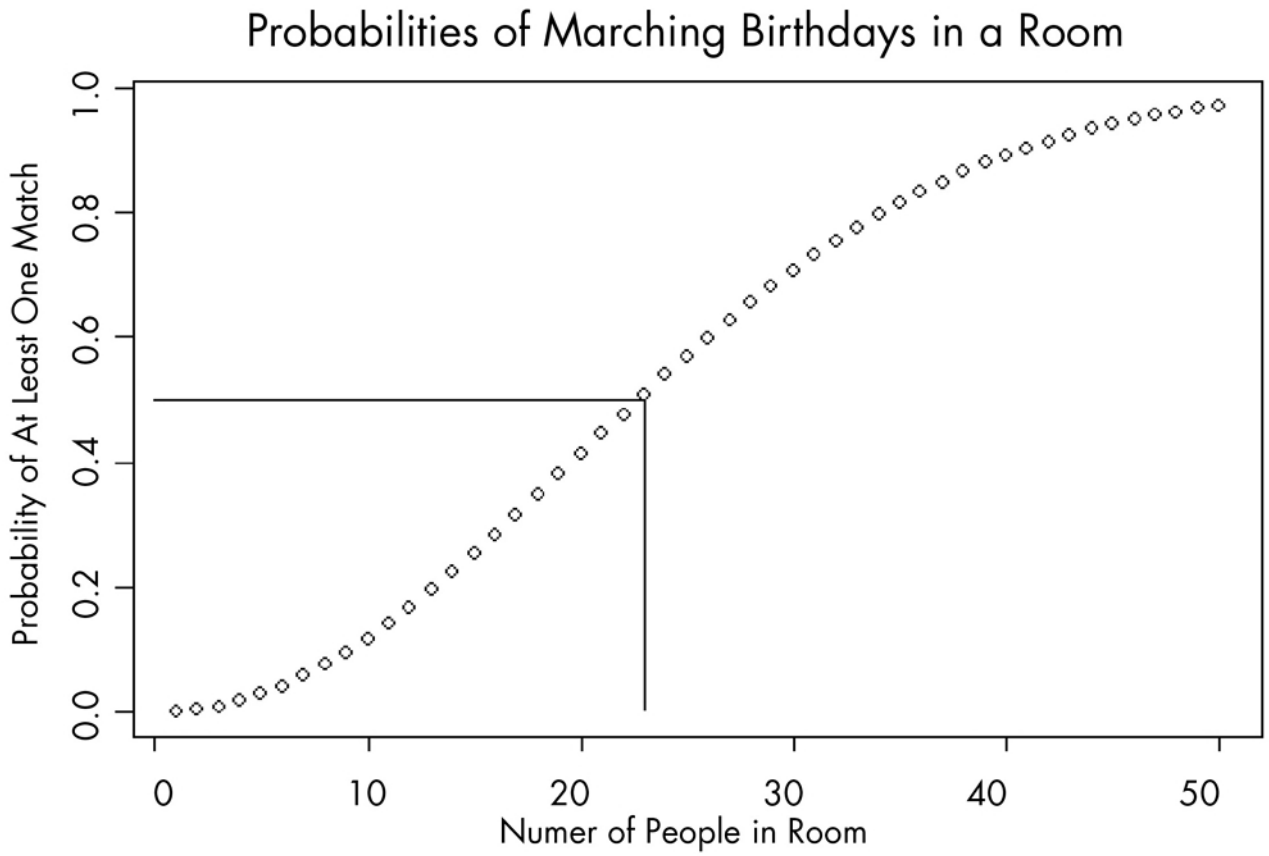
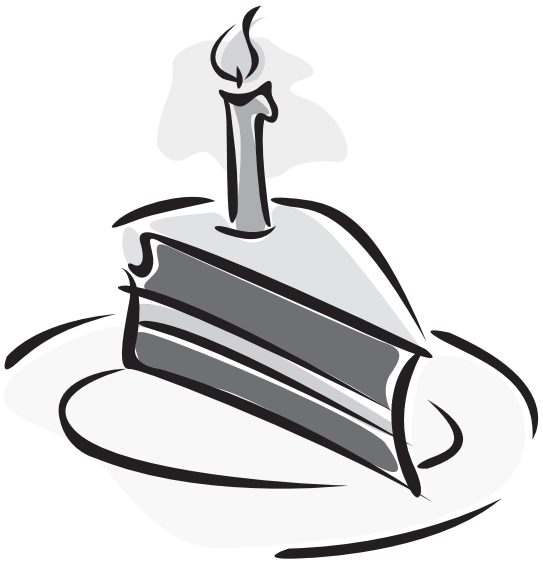


Figure 3: Plot from R of the probability of at least one pair of matching birthdays when the number of people in the room ranges from 1 to 50



n = 50 randomly chosen people is very likely to have at least one match. Indeed, for n = 50, the probability is 0.9704.

The Solution Using Simulation

A completely different approach to solving the birthday match problem is by simulation. Simulation is widely used in applied probability to solve problems that are too difficult to solve by combinatorics or other analytical methods. For example, we can use R to build a simulation model to approximate the probability that there are no matching birthdays among 25 people in a room.

This consists of first simulating the birthdays in many rooms, each with 25 people, and then checking to see what percentage of these rooms have matching birthdays. It is a little like taking a public opinion poll where the "subjects" are the rooms. We create the imaginary rooms by simulation, and then we "ask" each room, "Do you have any birthday matches?" If we ask a large number of rooms, the percentage of rooms with no match should be very near the true probability of no match in such a room.

This approach allows us to find the approximate distribution of the number of repeated birthdays (X). From this distribution, we can approximate P(X = 0), which we already know to be 0.4313. As a bonus, we also can approximate E(X), the expected number of matches among 25 birthdays. This expectation would be difficult to find without simulation methods.

Now let's build the simulation model step by step.

Step One: *Simulating birthdays for 25 people in one room*

Programmed into R is a function called "`sample`" that allows us to simulate a random sample from a finite population. To use this random sampling function, we need to specify three things.

First, we must specify the population from which to sample. For us, this is the 365 days of the year. In R, the notation `1:365` can be used to represent the list of these population elements.

Second, we have to specify how many elements of the population are to be drawn at random. Here, we want 25.

Third, we have to say whether sampling is to be done with or without replacement. Because we want to allow for the possibility of matching birthdays, our answer is "with replacement." In R, this is denoted as repl=T. We put the 25 sampled birthdays into an ordered list called b. Altogether, the R code is

```
b <- sample(1:365, 25, repl=T)
```

Each time R performs this instruction, we will get a different random list `b`. Below is the result of one run. For easy reference, the numbers in brackets give the position along the list of the first birthday in each line of output. For example, the 22nd person in this simulated room was born on the 20th day of the year, January 20.

```
 [1] 352 364 246 190 143 272 149
 [8] 206 154 272  61 199 357 141
[15] 264 157  42 340 287 166 335
[22]  20 123 214 149
```

You can see that there happen to be two matches in this list. The 6th and 10th birthdays both fall on the 272nd day of the year, and the 7th and 25th both fall on the 149th day of the year. Note that we also would have said there are two matches if, for example, the last birthday in the list had fallen on the 272nd day.

Step Two: *Finding the number of birthday matches among 25 people*

In a large-scale simulation, we need an automated way to find whether there are matching birthdays in such a room and, if so, how many repeats there are. In R, we can use the "`unique`" function to find the number of different birthdays, then subtract from 25 to find the number of birthday matches ("redundant" birthdays):

```
x <- 25 - length(unique(b))
```

For our run above, the list "`unique (b)`" is the same as `b`, but with the 10th and 25th birthdays removed. It is a list of the 23 unique birthdays since its "`length`" is 23. So the value of the random variable X for this simulated room is X = 25 – 23 = 2.

Step Three: *Using a loop to simulate X for many rooms*

If we repeat this process for a very large number of rooms, we obtain many realizations of the random variable X, and thus a good idea of the distribution of X. Counting the proportion of rooms with X = 0, we get the approximate probability of no match P(X = 0). Taking the average of these realizations of X, we get a good approximation to E(X).

When we simulated 10,000 such rooms, our result was P(No match) ≈ .4338, which is close to the exact value

0.4313 calculated using combinatorics. We also obtained $E(X) \approx 0.8081$. Additional runs of the program consistently gave values of $E(X)$ in the interval 0.81 ± 0.02.

The histogram in Figure 4 shows the approximate distribution of X – the Number of Birthday Matches. Our approximations would have been more precise if we had simulated more than 10,000 rooms, but the results seem good enough for practical purposes.

Testing Assumptions

With simulation, it is relatively easy to test the impact of the simplifying assumptions about 365 rather than 366 birthdays and that birthdays are equally likely. Actual 1997–1999 vital statistics for the United States show some variation in daily birth proportions. Monthly averages range from a low of about 94.9% of uniform in January 1999 to a high of about 107.4% in September 1999 (*www.cdc.gov/nchs/products/pubs/pubd/vsus/vsus.htm*). These fluctuations are illustrated in Figure 5. Daily birth proportions typically exceed 1/365 from May through September.

For nonuniform birthdays, computing the probability of no matches by analytical methods is beyond the scope of undergraduate mathematics; but using R, it is easy to modify our simulation so that 366 birthdays are chosen according to their true proportions in the United States population—rather than being chosen uniformly. We ran such a simulation and found that within the precision provided by 10,000 simulated rooms (about two decimal places), the results for the true proportions cannot be distinguished from the results for uniformly distributed birthdays. From these and related simulations on birthday matching, we conclude that, although birthdays in the United States are not actually uniformly distributed, it seems harmless in solving the birthday match problem to assume they are. However, important differences in the values of $P(X = 0)$ and $E(X)$ do occur if departure from uniform is a lot more extreme than in the United States population (See Nunnikhoven or Pitman and Camarri.

Using R Statistical Software

You can download R free of charge online at *www.r-project.org*. The program for doing the birthday matching problem with an explanation of the R code and an elementary tutorial on R are available online at *www.sci.csueastbay.edu/~btrumbo/bdmatch/index.html* and *www.amstat.org/publications/STATS/data.html*. Peter Dalgaard provides an introduction to statistics using R in *Introductory Statistics with R*. His book also is available as an electronic book, so check with your library.

Summary Comments on Simulation

Simulation is an important tool in modern applied

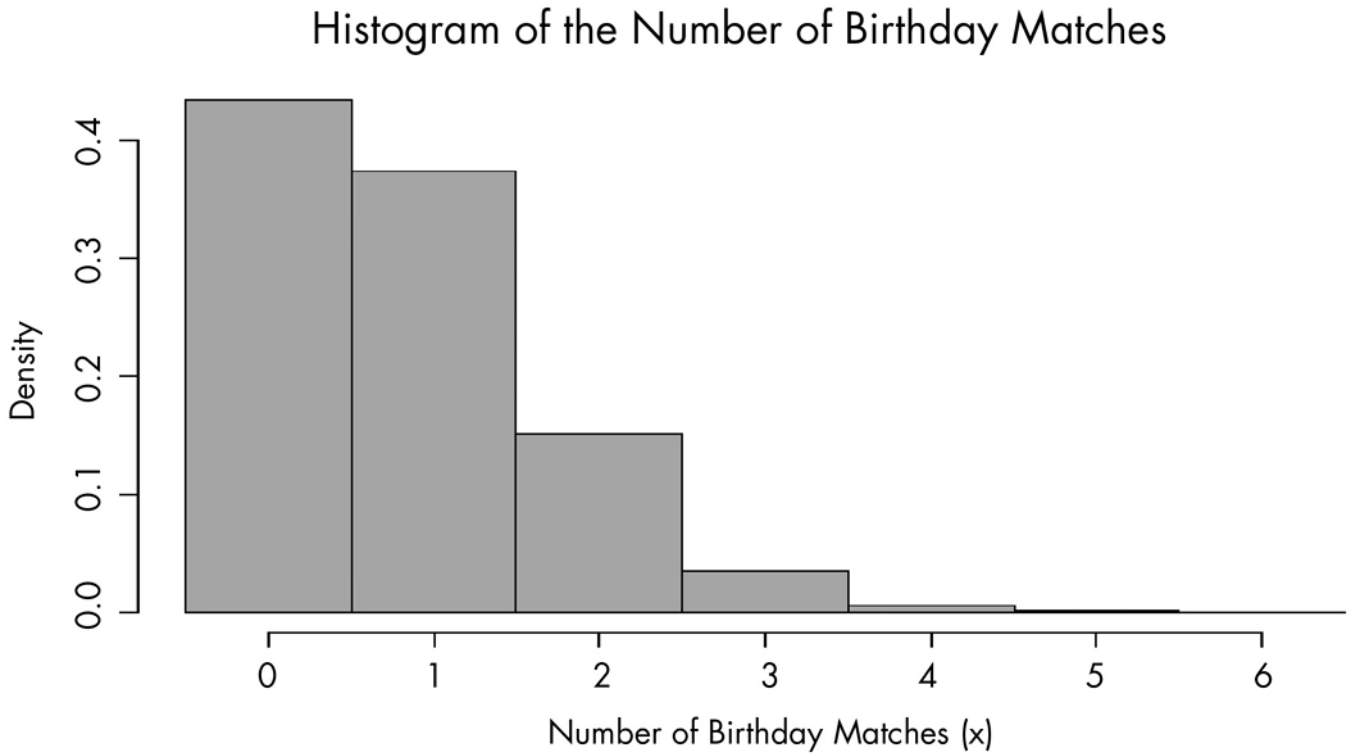


Figure 4: The simulated distribution of the number of birthday matches (X) in a room of 25 randomly chosen people

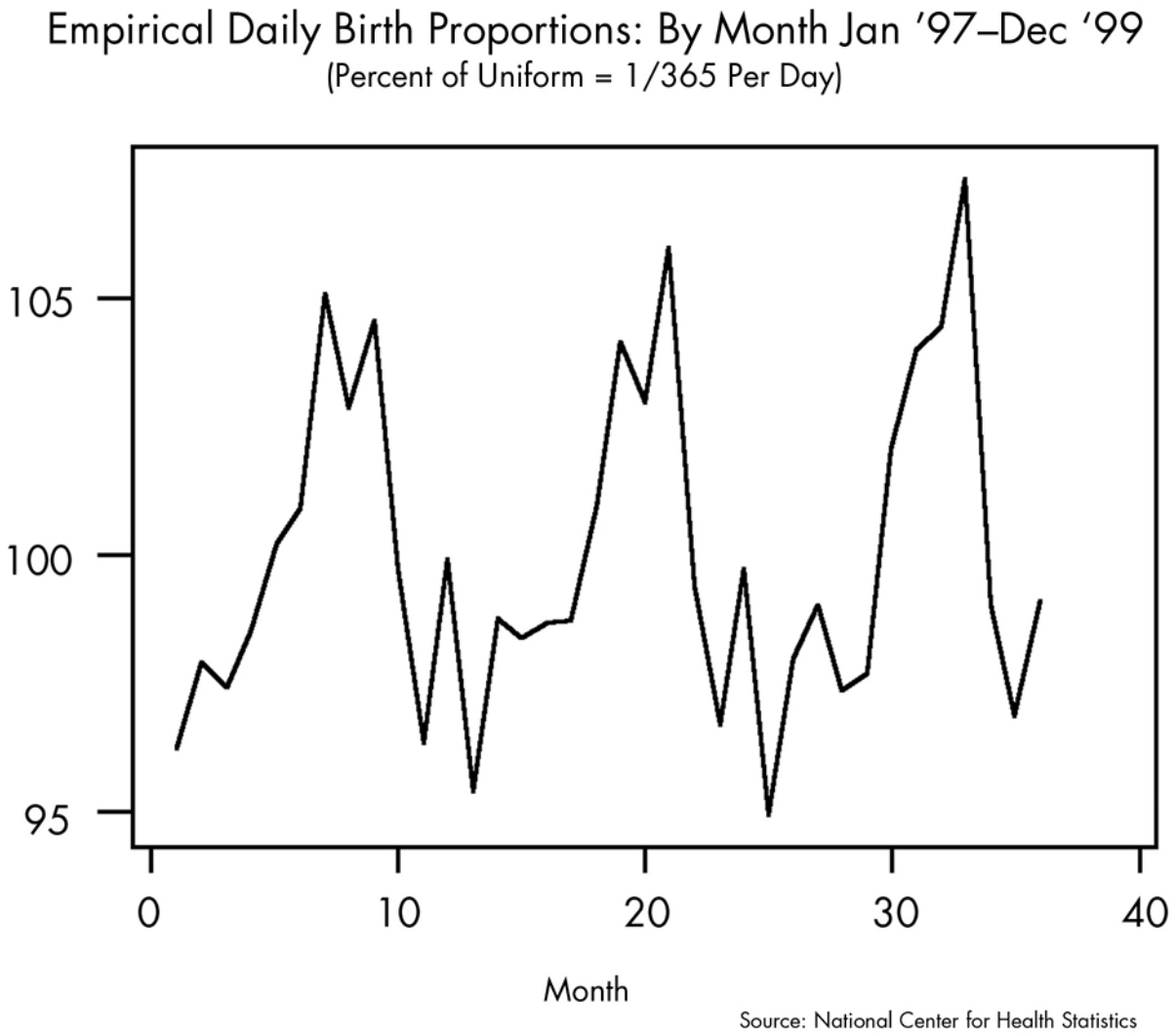


Figure 5: Cyclical pattern of birth frequencies in the United States for 36 consecutive months

probability modeling and in certain kinds of statistical inference. Many problems of great practical importance cannot be solved analytically.

From the birthday matching problem, we can see that—for practical purposes—simulation gives the same answer as does combinatorics for P(No Match) under the simplifying assumption that there are 365 equally likely birthdays. The same simulation provides a value for the expected number of matches, which would be difficult to find by elementary methods. Because this simulation gives what we know to be the correct answer for P(X = 0), credibility is given to the value it gives for E(X).

When we want to drop the uniformity assumption, we enter territory where analytic methods are much more difficult. But a minor modification of the simulation program provides us with values of P(X = 0) and E(X). This allows us to investigate the influence of our simplifying assumptions on results we intend to apply to real life.

In summary, we verified the correctness of a simulation method for an easy problem and then modified it to solve a closely related, but more difficult, problem. This process of building more complex simulation models based upon simpler trusted ones illustrates an important principle for using simulation reliably to solve a wide variety of important practical problems.

References

Peter Dalgaard. *Introductory Statistics with R*, Springer, 2002.

William Feller. *An Introduction to Probability Theory and Its Applications*, Vol. 1, 1950 (3rd ed.), Wiley, 1968.

Thomas. S. Nunnikhoven. "A birthday problem solution for nonuniform frequencies." *The American Statistician*, 46, 270–274, 1992.

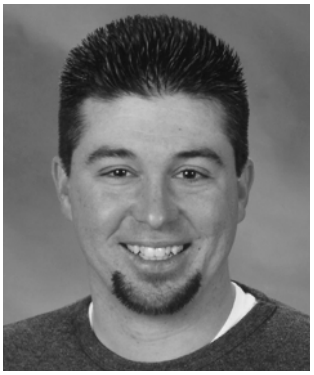
Jim Pitman and Michael Camarri. "Limit distributions and random trees derived from the birthday problem with unequal probabilities." *Electronic Journal of*

How to Make Millions on eBay,

or Using Multiple Regression to Estimate Prices

Kim Robinson

Eric Kamischke



Josh Tabor

Have you ever thought about buying or selling used items on the internet such as baseball cards, an electric train set, or a violin? What would be a reasonable price? Surely there are many variables that contribute to an item's value, depending on the product. For example, if you are selling baseball cards, you might create a model that includes variables for player, year, manufacturer, and condition. In our investigation, we decided to create a model to price used violins. While you may not be interested in violins, the same methods can be used to estimate the price of any product for sale online.

Used violin on the internet

Kim Robinson (KimRobinson@mail.clayton.edu) teaches at Clayton State University in Atlanta, Georgia. Her professional interests include online teaching and learning and the AP statistics course. She also enjoys gardening, jogging, and college football.

Eric Kamischke (Kamischkee1@yahoo.com) is a mathematics teacher at Interlochen Arts Academy in Michigan. He is also an author and consultant for Key Curriculum Press in California.

Josh Tabor (jtabor@hlpusd.K12.ca.us) is an AP statistics teacher at Wilson High School in Hacienda Heights, California.

Overall, when creating a model using multiple regression, there are two competing goals: reduction in variability and parsimony. That is, we would like to explain as much of the variability in price as possible and at the same time keep the model relatively simple.

To achieve both goals, it is important to find variables that are highly correlated with the response variable—in our case the price of a used violin. Of course, the hardest part of the process is figuring out which variables are the most useful. To get an idea of what variables affect the price of a violin, a brief history lesson is in order.

The first violin appeared about 1510 in Italy, with the current violin appearing at the end of the 16th century. Like the piano, early violins lacked appreciation and respect. It was only after prominent composers such as Bach, Mozart, Beethoven, Schubert, Brahms, Mendelssohn, Tchaikovsky, and Stravinsky utilized the violin in operas that the violin began to gain musical acclaim and admiration.

Master violinmakers Stradivari (1644–1737, Italian), Guarneri (1626–1698, Italian), and Stainer (1621–1683, Austrian) also contributed to increasing the popularity of the violin. These skilled masters created violins with unparalleled levels of musical quality. Later, the structure for both the instrument and bow changed in the 18th and 19th centuries, giving the violin a louder, deeper, and more brilliant tone. Recently, musicians who believe the original designs are better suited for early musical pieces have restored their instruments to the original designs.

(For additional historical background, refer to *http://encarta.msn.com* and *www.encyclopedia.com*.)

Based on the history of the violin, at least five variables seem worth investigating: the maker of the violin, the size of the violin, the age of the violin, the violin's classification (student or professional), and the country in which the violin was made. The condition of the instrument most likely also contributes to a violin's value, but this cannot be quantified easily so we did not consider this variable.

At present, there are hundreds—maybe thousands—of used violins for sale on the internet. Using a variety of search engines to look for "violin for sale" and "violins for sale," we selected sites at random. We limited our data selection to advertisements written in English that listed all the variables we suspected might be useful in the model. We found 68 violins. A frequency histogram of the distribution of prices revealed some very large observations. See Figure 1.

We discovered that all of the highest-priced violins were classified as "professional," as opposed to the cheaper "student" versions. Because extreme values can be very influential when fitting a model, we took the common logarithm of price to eliminate the skew. The distribution then appeared approximately normal, as shown in Figure 2. A log transformation is often useful when the data span several orders of magnitude (powers of 10).

Because it is obvious from the histogram that the prices of professional violins are distinctly different from the prices of student violins, we created a regression model using Log (Price) as the response variable and an indicator variable "Pro" (1 = professional, 0 = student) as the explanatory variable. The details of this model are displayed in Figure 3.

The model suggests the average price for student violins is:

$$10^{2.98409} = \$964$$

while the average price for professional violins is:

$$10^{(2.98409 + 1.38483)} = \$23{,}384.$$

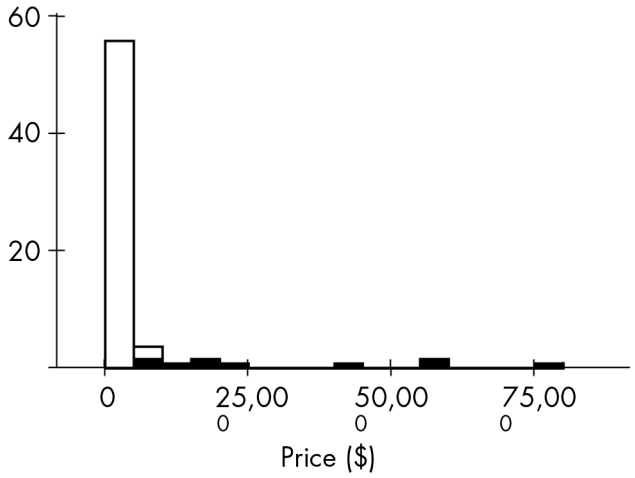


Figure 1. Distribution of prices of used violins offered for sale on the internet: "professional" violins are shown by shaded columns; "student" violins are shown by unshaded columns.

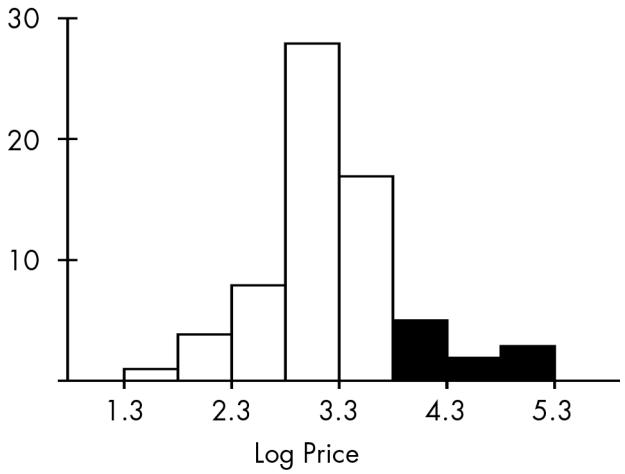


Figure 2. Distribution of the logarithms of prices of used violins offered for sale on the internet: "professional" violins are shown by shaded columns; "student" violins are shown by unshaded columns

DEPENDENT VARIABLE IS: LOG (PRICE)

R Squared = 51.8% Adjusted R Squared = 51.1%

Source	**Sum of Squares**	**df**	**Mean Square**	**F-Ratio**
Regression	16.3573	1	16.3573	71.0
Residual	15.2028	66	0.2303	

Variable	**Coefficient**	**Std. Error of Coeff**	**t-Ratio**	**Prob**
Constant	2.98409	0.0630	47.37	<0.0001
Pro	1.38483	0.1643	8.43	<0.0001

Figure 3. Regression model for Log (Price) with an indicator variable "Pro," where 1 = professional version and 0 = student version

PEARSON PRODUCT-MOMENT CORRELATION

	Log (Price)	Date	**SIZE**	Maker	Pro	Residuals
Log (Price)	1.000					
Date	–0.315	1.000				
Size	0.548	–0.373	1.000			
Maker	0.360	0.063	0.225	1.000		
Pro	0.720	–0.159	0.149	0.308	1.000	
Residuals	0.694	–0.283	**0.636**	0.200	0.000	1.000

Figure 4. Correlation matrix for residuals from the "Pro" model

PEARSON PRODUCT-MOMENT CORRELATION

	Log (Price)	**DATE**	Size	Maker	Pro	Residuals
Log (Price)	1.000					
Date	–0.315	1.000				
Size	0.548	–0.373	1.000			
Maker	0.360	0.063	0.225	1.000		
Pro	0.720	–0.159	0.149	0.308	1.000	
Residuals	0.532	**–0.081**	0.000	0.108	0.000	1.000

Figure 5. Correlation matrix for residuals from the "Pro and Size" model

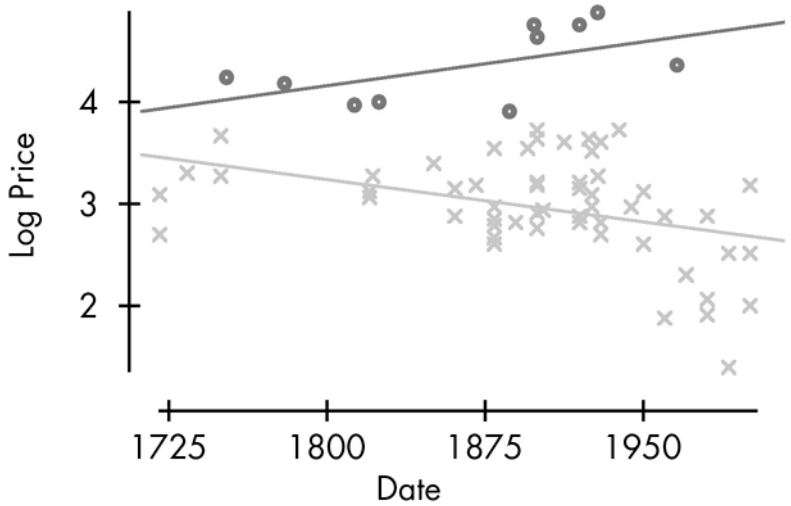


Figure 6. The dots are for professional violins and the "x" data points are for student violins in this scatter plot of "Date" versus Log (Price) for used violins.

Quite a difference! More importantly, we saw that the model explained about 52% of the variability in Log (Price), as $R^2 = 0.518$. Adjusted R^2 for this model was 0.511.

However, 48% of the variability still remained. To investigate which of the other variables could explain the remaining variability, we calculated the residuals of the "Pro" model—which measure the variability after the professional/student distinction is factored in—and created the correlation matrix shown in Figure 4.

"Size," which has the largest correlation with the residuals from the first model (r = 0.636), was chosen as the next main effect for the model. After adding "Size" as an explanatory variable, the updated correlation matrix shown in Figure 5 revealed a large change in correlation between "Date" and the residuals. "Date" is the year when the violin was made and represents the age of the violin.

This led us to investigate the relationship between Log (Price) and "Date." A scatter plot of the data is shown in Figure 6.

It appears that there are two relationships between Date and Log(Price). The professional violins are represented by the positively sloping line (the upper line), while the student violins are represented by the negatively sloping line (the lower line). Older student violins are worth more, while older professional violins are worth less. The difference in slopes indicates an interaction effect between age (Date) and the type of violin.

Because the interaction term ("Date*Pro") was significant (p = 0.034), but "Date" (p = 0.133) and "Pro" (p = 0.074) were not significant individually, both main effects were removed from the model leaving only the interaction term. Including the interaction term while removing both main effects reduced the length of our model by one term and slightly increased the R^2 value.

Next, to consider the effect of a violin's country of origin, we constructed the side-by-side box plots shown in Figure 7, which revealed Asian prices to be different from all other countries. As a group, the median price of Asian violins is lower.

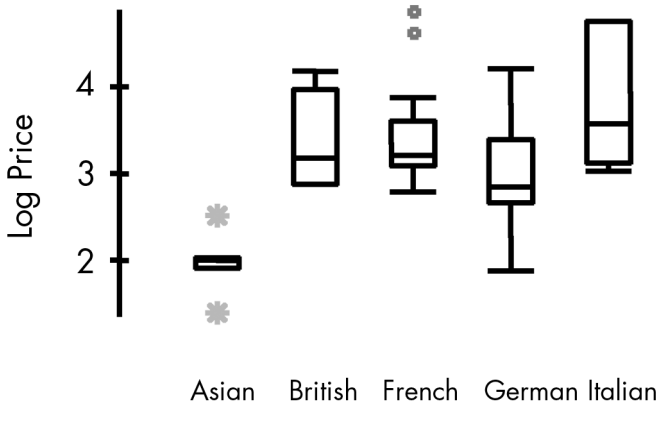


Figure 7. Box plots of Log (Price) of used violins by country. Note that the two French "outliers" were the only two French professional violins in the dataset.

Investigating this difference between Asian countries and all others, we inserted Asian as an indicator variable by coding it with a "1" for violins from Asia and with a "0" for violins from all other countries. Adding the Asian variable improved the model by increasing R^2 to 0.748 and adjusted R^2 to 0.736.

Considering the other countries, only the indicator variables for French (p = 0.006) and Italian (p = 0.022) violins had significant p-values. Including both of these variables in the model increased the adjusted R^2 to 0.769.

The last step was to consider including the only remaining variable, "Maker." However, when added to the model, "Maker" was not significant (p = 0.500), so we did not include it in the model. Therefore, our final model was:

Log (Price) = 2.22388 + 0.76781 Size + 0.000641 Date*Pro + 0.267288 French + 0.33665 Italian – 0.548708 Asian

The regression model is statistically significant ($F = 45.6$, $p < 0.0001$) and $R^2 = 0.786$.

Notice that if all of the country variables are set to zero, the model then predicts the log (Price) of either a German or British violin. This is because they are not distinguishable in the model. Our dataset included only six British violins. With a larger dataset containing more British violins, a different effect might emerge and a "British" term could be added to the model.

An example prediction using our model for a large (Size = 1), professional (Pro = 1) violin made in 1900 (Date = 1900) in Asia (French = 0, Italian = 0, Asian = 1) is:

$$10^{[2.22388 + 0.76781 * (1) + 0.000641 * (1900 * 1) + 0.267288 * (0) + 0.33665\ (0) - 0.548708 * (1)]} = \$5067$$

In conclusion, about 79% of the variation in the price of used violins for sale on the internet can be explained by the size of the violin, the interaction between the age of the violin and it's classification (student or professional), and its origin—specifically whether Italian, French, or Asian.

We need to point out that our model is based on a non-random sample in an observational study, and so the results should not be generalized beyond the context of our study. However, if you are in the market to buy or sell a used violin similar to the ones we found, this model could be a guide in estimating the price to expect.

So, if you are interested in buying or selling an electric train set, some baseball cards, or other collectible items on the internet, building a regression model could help you decide if a transaction is worthwhile. All you need are price data and relevant predictor variables to begin the process of building your own regression model. Then, with your model in hand, maybe you can find some bargains on eBay and maybe you can make a million dollars!

All data analysis was conducted using Data Desk, distributed by Data Descriptions, Inc. www.datadesk.com. eBay is a trademark of eBay, Inc. ■

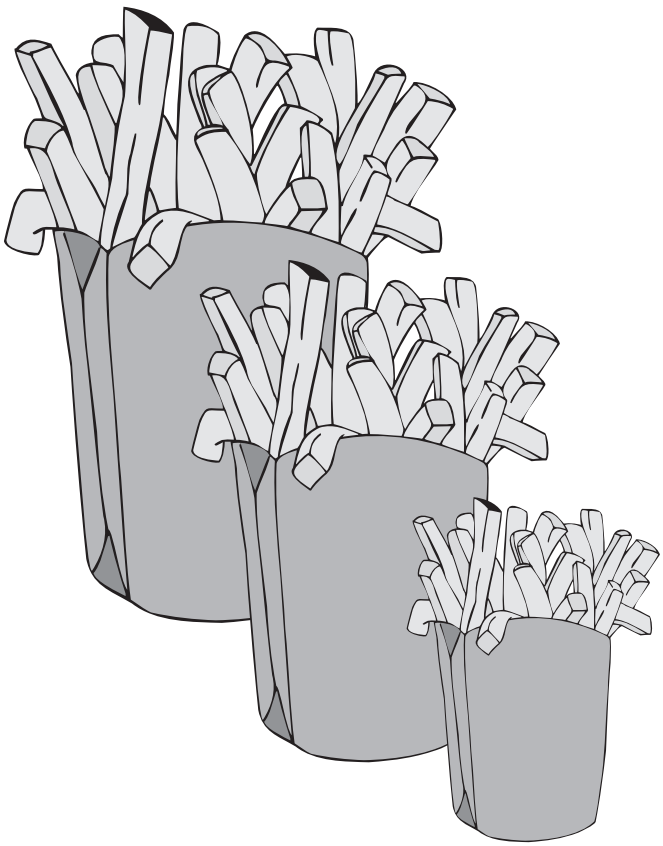
McDonald's French Fries

Would You Like Small or Large Fries?

Nathan Wetzel

Our Research Question

Early in the spring of 2004, the documentary *Super Size Me* came out. The documentary followed filmmaker Morgan Spurlock as he spoke with people about their experiences with fast food. Mr. Spurlock also documented his own health as he ate only at McDonald's for a month. Although not based on statistical evidence, the film increased the public's awareness of the link between health and diet.



In March 2004, CNN reported: "By the end of 2004, Supersize will no longer be available at the nation's 13,000-plus McDonald's outlets, except in certain promotions," McDonald's spokesman Walt Riker said.

He continued, "The move comes as the world's largest restaurant company and fast-food chains in general are under growing public pressure to give consumers healthier food options in a nation that has suddenly become aware of its bulging waistline and the health dangers that come with it."

At that time, McDonald's target serving sizes apparently were in flux. The brochure available at a local franchise listed 68 grams as the mass of a small order of french fries and 176 grams for a large order. However, McDonald's official web site indicated 74 grams for a small order and 171 grams for a large.

The currency of these issues provided my statistics classes with a good opportunity to investigate the serving sizes for large and small orders of french fries at McDonald's. In particular, we wanted to compare the average mass of a serving to McDonald's official target values. Our research question was, "Is there a difference between what McDonald's serves and its target value?" The null hypothesis was that there was no difference, and the alternative hypothesis was that there was a difference.

The Data

We collected data on two consecutive days within two half-hour time periods—1:00 p.m. to 1:30 p.m. and 5:15 p.m. to 5:45 p.m. In groups of two, one student ordered a small order of fries while the other ordered a large. We kept the fries in their bag and measured the mass of the fries and bag together. Then, we measured the mass of the bag without the fries. The difference between these two measurements gave the mass of the fries. The same scale was used for all measurements.

Some of the restaurant staff became curious about what we were doing. However, because the person filling the bags of fries often was not the same as the person

Nathan Wetzel (nwetzel@uwsp.edu) is an Associate Professor in the Department of Mathematics and Computing at the University of Wisconsin - Stevens Point. His primary interest is the teaching of statistics in the undergraduate curriculum. He enjoys crossword puzzles, gardening, and origami. His address is Dept. of Math and Computing, UW-Stevens Point, Stevens Point, WI 54481.

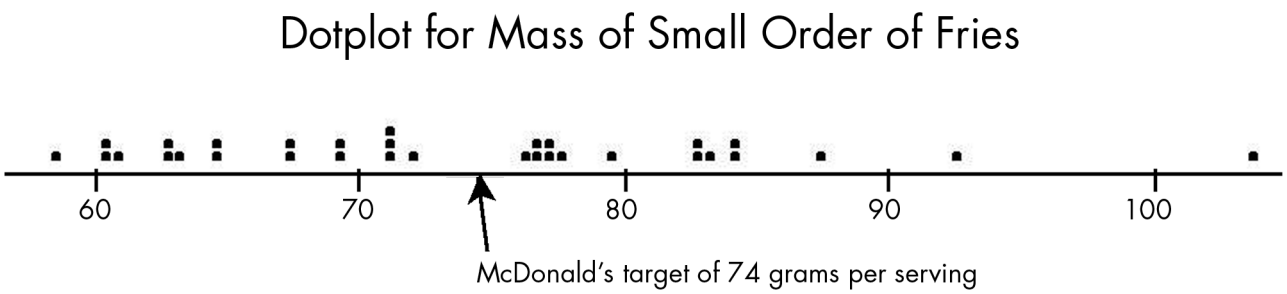


Figure 1. Mass measurements for small orders of french fries

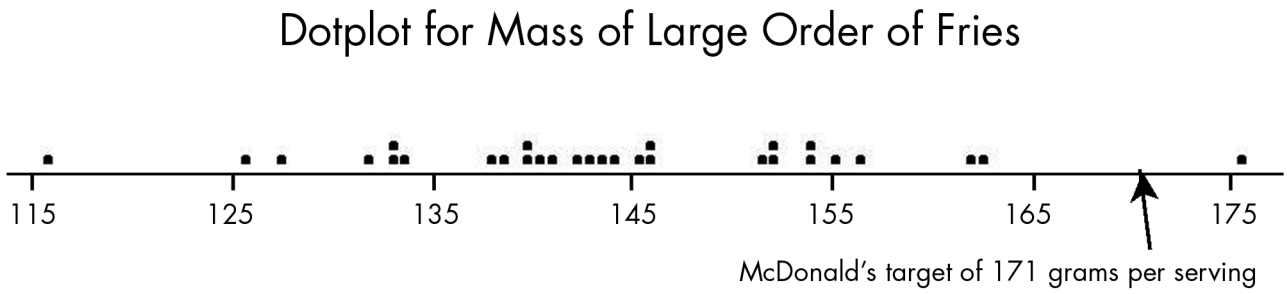


Figure 2. Mass measurements for large orders of french fries

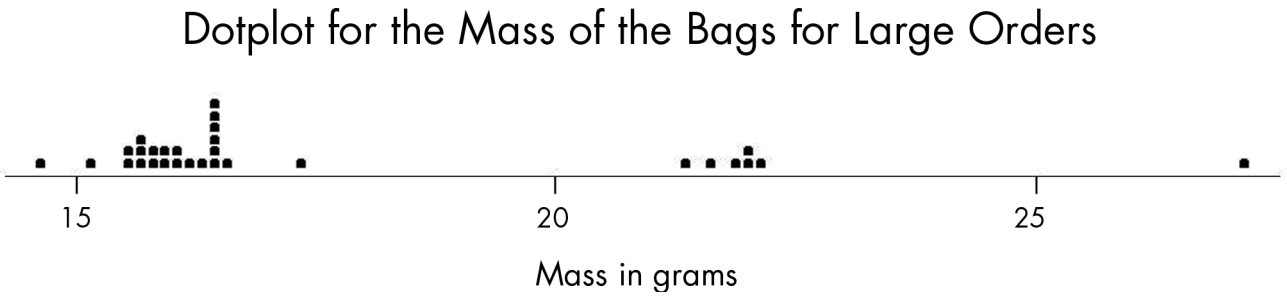


Figure 3. Mass measurements of the bags of large orders of french fries

taking the order, we assumed they did not change the serving sizes we received from what they typically serve.

The Results

In all, we measured 32 small orders and 31 large orders of fries. One person worked alone and only purchased a small order of fries. We noticed the mass of the bag for one of the large orders was inexplicably two and a half times larger than any other bag. This resulted in a data value with a mass much smaller than the others, so we omitted it from the analysis. Figure 1 displays graphically all of the mass measurements for the small orders, and Figure 2 shows the data for the large orders.

Table 1 summarizes the data and compares our sample information with McDonald's target values.

As you can see, the mean mass for the small orders of fries came very close to the target set by McDonald's. However, the mean mass of the large orders of fries was short of the target number. In fact, in Figure 2 you can see that only one of the 30 orders exceeded the target value set by McDonald's. In Table 1, you can see that the target value for a large order of fries does not lie within the

	McDonald's Target Serving Size	Data Collected		
Serving Size	Grams	Sample Size	Grams	95% Confidence Interval for the Mean
Small	74	32	Mean = 73.72	± 3.79
Large	171	30	Mean = 144.07	± 4.59

Table 1. Mean mass in grams of small and large servings of french fries

confidence interval for the mean. Therefore, we can reject the null hypothesis of no difference between the mean serving size and the target value.

Not surprisingly, a correlation was found between the number of fries and the mass of the fries, as shown in Table 2.

Some Observations

We also collected data regarding the frequency of students going to fast food restaurants. From these data, we estimated that about 25% of the students on our campus visit McDonald's each week. This fact helped to put the possible impact of our findings into perspective for the students involved in the project.

Although we did not notice as we were collecting the data, there could be two types of packaging used for large fries. A dotplot for the mass of the bag is shown in Figure 3, and two groupings can be seen in the data with one outlier can be seen in the data with one outlier to the right.

The difference between the two groups could be due to an extra napkin, a package of salt, or a different type of bag. Because the mass of a package of catsup is about 10 grams, the large outlier might have included a package of catsup that we did not notice.

There is some evidence of a difference in the mass when comparing small orders purchased in the afternoon compared with those purchased in the evening. When we omitted two possible outliers from the data we collected in the afternoon, there was a significant difference between the masses of the small fries in the afternoon compared to the evening ($p = .003$). We used a t-test for two samples with unequal variances for this comparison test.

Serving Size	Correlation Coefficient (Sample Size)
Small	r = 0.633 (n=32)
Large	r = 0.484 (n=30)

Table 2. Correlation between the number of french fries in a serving and the mass of a serving

Time of Day	Number of Samples	Mean (grams)	Standard Division (grams)
Afternoon	16	67.74	5.28
Evening	14	77.06	8.91

Table 3. Mean mass of small orders of french fries in the afternoon and in the evening

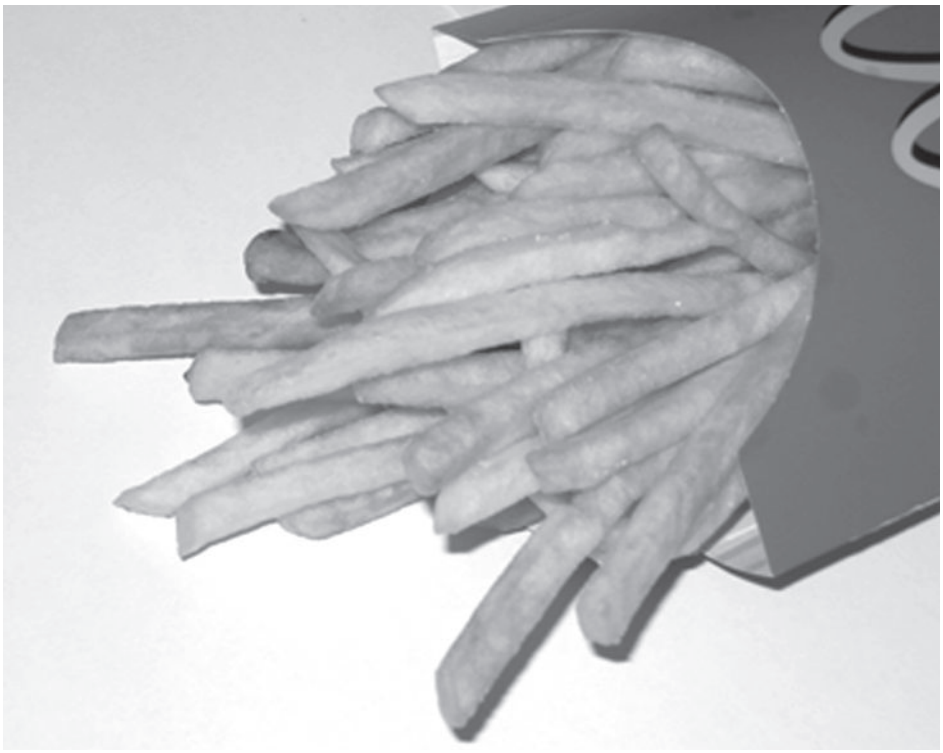


Table 3 shows that the mean mass of the afternoon samples is lower than in the evening by about 10 grams. It would be interesting to investigate this further.

Summary

In conclusion, we found that large orders of french fries from the restaurant in our study were on average 27 grams short of the target serving size. The university students were highly interested in the differences we found and the value shortfall they might represent.

However, in the context of our country facing an obesity problem, these results could be considered a good thing. Using the estimates from our analysis and an average of 3 calories per gram, we estimated that as a result of 27 fewer grams per large order, each serving contained about 81 fewer calories.

Finally, here are some questions we would like to investigate further:

- Would we get the same results if we gathered data from other McDonald's restaurants?
- What would be the results if we took measurements for a longer period of time?
- Did we gather data during a week that happened to be soon after McDonald's changed its target for large fries from 176 to 171 grams?
- Are there differences in the packaging of large fries?

References

McDonald's Corporation. *A Full Serving of Nutritional Facts*, 2003,

McDonald's Corporation. *McDonald's USA Nutrition Facts for Popular Menu Items*, *www.mcdonalds.com/app_controller.nutrition.index1.html*, 2004. ■

PC Versus Mac:
An Exploration of the Computer Preferences of Harvard Students

Alex Kirn

Robert Lord

Abdallah Salam

Computer Preferences

Steve Jobs, cofounder and CEO of Apple Computer, Inc., entered Reed College in 1972. Bill Gates, founder and chairman of the Microsoft Corporation, entered Harvard College in 1973. However, the alma mater of a computer executive is probably irrelevant when a student buys a computer. In an effort to actually shed light on the reasons underlying students' computer preferences, we conducted a survey at our college in the fall of 2004. We were interested in exploring the factors that might explain the purchase of a Macintosh (Mac) or a Personal Computer (PC) by Harvard students.

We selected the following variables to examine:

- Gender
- Desktop or laptop owner
- Importance of price
- Importance of performance
- Importance of design
- Importance of software compatibility
- Intended concentration (academic major)

We expected price and performance to strongly influence students' computer choices, as many students are on a strict budget but desire as much computing power as possible. We also suspected a student's intended concentration might necessitate the use of specific computing capabilities.

Alexander Kirn (alexander@kirns.de) studies economics at the University of St. Gallen in Switzerland. He currently is spending his junior year abroad at Harvard College. Alexander is passionate about entrepreneurship and enjoys playing tennis and snowboarding, which he also teaches.

Robert Lord (rlord@fas.harvard.edu) is a freshman at Harvard College and intends to study psychology and government. He is particularly interested in East Asian sociocultural studies, having grown up in Japan.

Abdallah Salam (salam@fas.harvard.edu) is a freshman at Harvard College and intends to major in economics. He received a mention très bien on the French Baccalaureate, specializing in mathematics. He enjoys debate, sleep, and spending time with family and friends.

Survey Methodology

In order to ensure that our survey instrument would be effective and provide data relevant to our investigation, we first conducted a small pilot survey. We sent a draft questionnaire to 14 of our friends by email. We received 12 cooperative responses. We learned many valuable lessons from the pilot survey, including how to improve the wording of our questions, and after several rounds of revisions, our final questionnaire was ready.

We sent identical messages to all Harvard freshmen through the community email system. The message contained a link to our questionnaire on the Harvard polling system. This online polling tool provided a simple and convenient way for students to respond anonymously and for us to collect and categorize the data.

Survey Data

The size of our target population was approximately 1,600 freshmen. Initially, we anticipated a response rate of about 5%; however, much to our surprise, we received a total of 404 responses—a 25% response rate and almost five times the amount of data we expected.

The survey results are summarized in Table 1. Respondents rated the "importance" of price, performance,

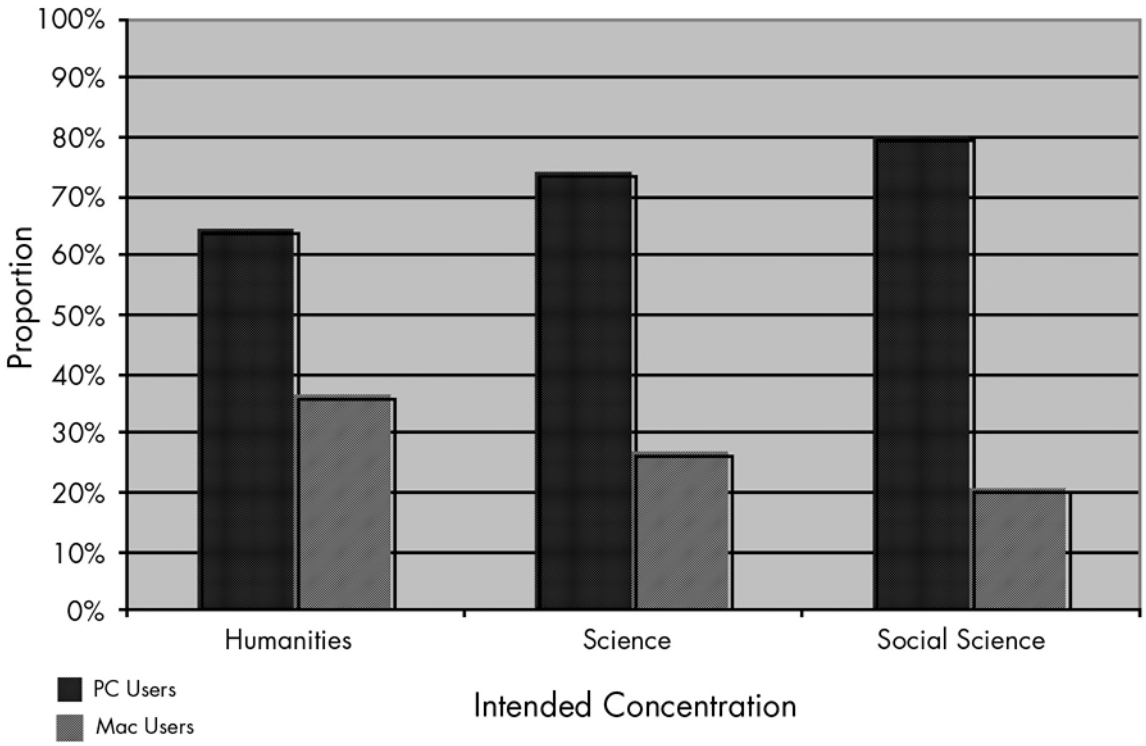


Figure 1. Proportion of students who own a PC or a Mac by intended concentration

design, and software compatibility on a scale of 1 to 5, with 5 being "high importance." The mean value of the ratings is shown in the table.

	Use a PC	Use a Mac
Responses	305	99
Male	137	44
Female	168	55
Desktop Owner	30	2
Laptop Owner	275	97
Price	3.82	3.45
Performance	4.51	4.61
Design	2.99	3.69
Software Compatibility	3.79	3.77

Table 1. Survey results

Figure 1 shows the proportion of students who own Macs and PCs grouped by their intended concentration. PCs are favored over Macs by about 3 to 1.

We chose chi-square tests to analyze the categorical variables, such as gender and intended concentration, and opted for logistic regression to assess the probability of students preferring a Mac or a PC.

Analysis: Chi-square

We were not surprised the chi-square test failed to reject the null hypothesis of no association between gender and Mac or PC preference (p=.934). So, we concluded there was no evidence of a difference in computer preference due to gender.

However, we were surprised the chi-square test revealed only a marginally significant association (p=.056, see Table 2) between intended concentration and computer preference. We expected it to be significant—at least at the $\alpha = .05$ level.

	Humanities	Science	Social Science
PC	34 (39.88)	114 (116.64)	153 (144.48)
Mac	19 (13.12)	41 (38.36)	39 (47.52)

$\chi^2 = 5.77$	$P(\chi^2_{df=2} = 5.77) = 0.056$

Table 2. Chi-square test of PC or Mac usage by intended concentration. The values in parentheses are the expected counts; the categories for intended concentration have been aggregated so all expected counts are ≥ 5.

In attempting to understand what might contribute to this relative lack of association, we noted the recent increases in platform cross-compatibility and the interdisciplinary nature of Harvard students' studies as possible factors.

Variable	Coefficient	z-Score	p-Value	95% Confidence Interval	
Importance of Design	0.6505	5.06	0.000	0.3984	0.9026
Prefer Desktop or Laptop	1.8262	2.38	0.017	0.3241	3.3283
Importance of Price	–0.2661	–2.35	0.019	–0.4878	–0.0445
Constant	–4.0621	–4.05	0.000	–6.0303	–2.0939

Table 3. Logistic regression results (variables are listed in order of decreasing statistical significance)

Analysis: Logistic Regression

Using logistic regression, we built a model to determine the probability of PC or Mac ownership among Harvard freshmen based on the variables we measured. Multiple regression would have been inappropriate, as the response variable needs to be normally distributed. In our case, the response variable was binary (PC = 0 and Mac = 1) and did not fulfill this requirement.

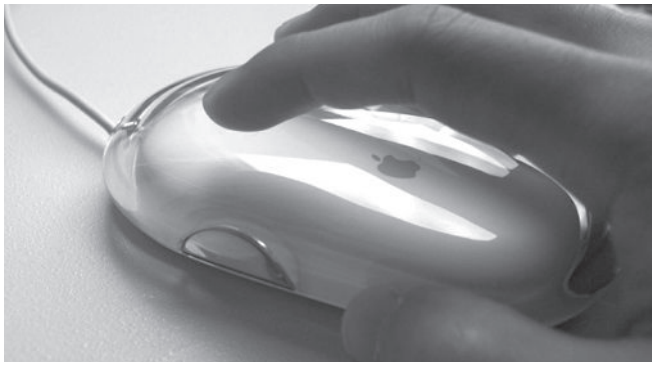
In building our model, we wanted to include all the variables in the model that helped predict the outcome, while at the same time keeping the model as parsimonious as possible. We performed a stepwise logistic regression on all the variables, essentially filtering out insignificant factors to create a "streamlined" predictive model.

Logistic regression allowed us to model the log of the odds of the response based on the explanatory variables. The Mac odds are the probability of a student having a Mac divided by the probability of not having a Mac. Unlike ordinary linear regression, which uses least-squares to estimate the regression coefficients, maximum likelihood is used to estimate the coefficients in logistic regression.

Table 3 shows the results of our logistic regression analysis. We found that the most important predictors of the probability of a student having a Mac were importance of design, preference for a desktop or laptop, and—of course—importance of price. Performance and software compatibility were not significant predictor variables, suggesting that perhaps students do not perceive a difference in performance or software compatibility between PCs and Macs.

The final logistic regression model was:

$$\ln\left(\frac{P(Mac)}{1-P(Mac)}\right) = -4.062 + 1.8262 * desklap + 0.6505 * design - 0.2661 * price$$



For the variable "desklap," a value of 1 indicates preference for a desktop, while a value of 0 indicates preference for a laptop.

A variable's coefficient shows the change of the log of the odds of the response as the variable changes. The negative sign on the price variable indicates the log of the odds of a student having a Mac goes down as the importance of price goes up.

We can rewrite the model to solve for the probability of a Harvard freshman having a Mac as:

$$P(Mac) = \left(\frac{1}{1 + e^{4.062 - 1.8262*desklap - 0.6505*design + 0.2661*price}} \right).$$

Figure 2 shows the probability of a student using a Mac as the importance he or she places on price and design change given a preference for a laptop.

Conclusions

A number of interesting effects were revealed regarding PC and Mac preferences among Harvard students. First, students who prefer laptops are more likely to prefer a Mac. Second, consistent with Apple's emphasis on the superior aesthetics of their machines, students concerned with design greatly favor Macs. Third, students for whom price is an important consideration prefer PCs.

There are many possibilities for additional research in this area. One obvious further exploration would be repeating this study for different class years, as our study only included freshmen. Perhaps, as students advance in their education, their computer preferences change due to academic specialization. It also would be interesting to see if the results would be the same at other colleges.

Additionally, computer preferences in the future might be influenced by further developments in the industry, such as the release of the Mac Mini—a far cheaper alternative for the price-conscious. Another scenario might be the introduction of an easier-to-use operating system based on Linux as an alternative to Windows, which could stimulate the demand for PCs.

With so many interesting possibilities, we hope other researchers will further our work in an effort to better understand the factors that influence the purchase of one of the most important items in a student's academic arsenal.

Finally, we would like to thank Professor Louise Ryan and the Harvard Class of 2008, without whom this study would have been impossible. ■

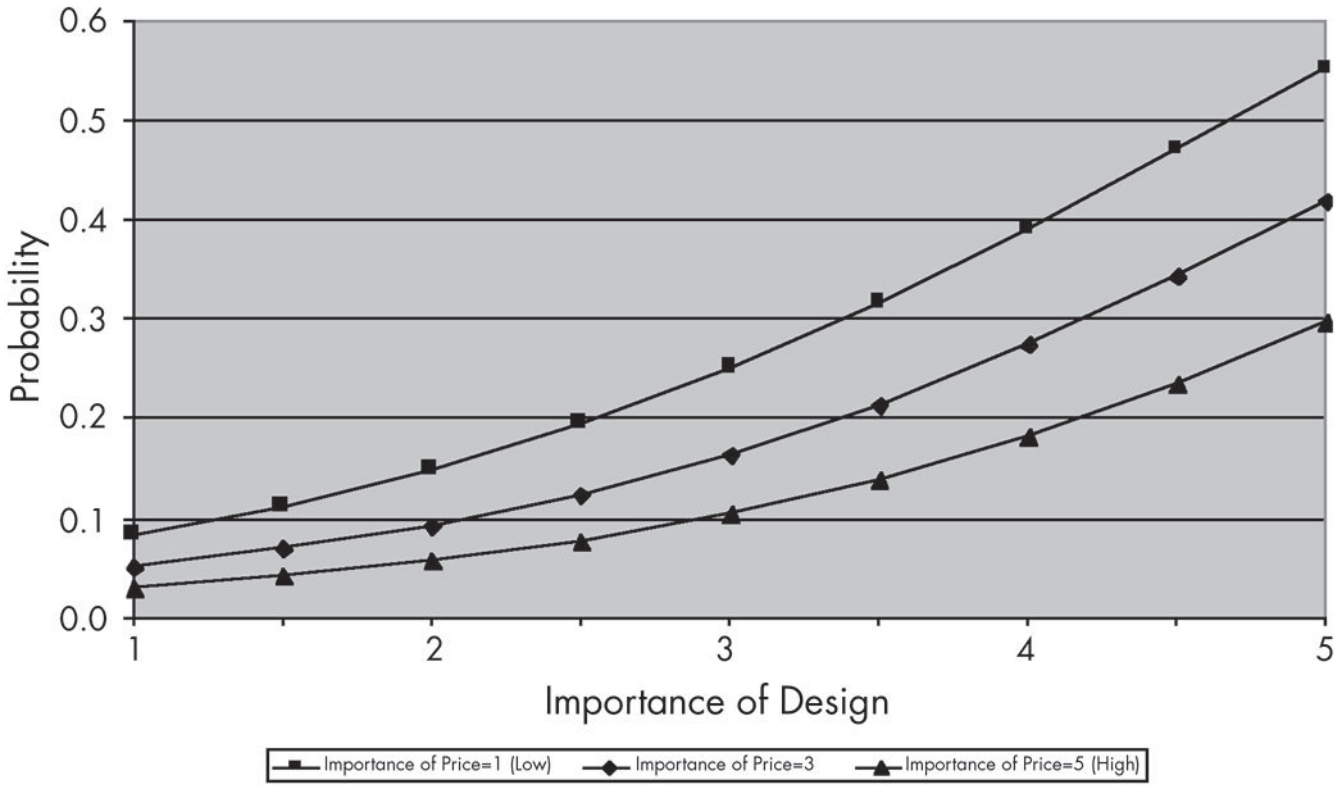


Figure 2. Predicted probability of owning a Mac among students who prefer a laptop

Statistics the EESEE Way!

Chris Sroka

Can dogs sniff out cancer? Do waiters who tell Eskimo jokes get better tips? Is Jessica Simpson a mathematical genius? These are some of the questions that have been asked for the newest stories being developed by EESEE.

EESEE (pronouced "easy") stands for "Electronic Encyclopedia of Statistical Examples and Exercises." It is a collection of stories that applies statistical concepts to real-world events using genuine data. EESEE illustrates the application of statistics to a variety of disciplines. Medicine, sports, business, humanities, and politics are just a few of the areas these stories touch upon. The collection provides instructors with quick access to real data and examples to use in their courses. Each story contains a series of thought-provoking questions that encourage students to carefully consider the relevant statistical concepts.

EESEE was developed at The Ohio State University by Professors William Notz, Dennis Pearl, and Elizabeth Stasny. My job as a research assistant is to help create new stories so EESEE stays interesting and fresh. Story ideas come from various sources: newspaper articles, the internet, and research performed by professors at Ohio State. Frequently, a television news report will call attention to an interesting study just published in a journal.

There is a plethora of studies and news reports citing some sort of statistic. Which ones make for a good story? A key element is that the story appeals to college students and motivates them to work through problems. Students will find a story more engaging if it pertains to some aspect of their lives. Some stories may be related to a topic a student is studying in another class. Other stories are based on everyday situations that are familiar to college students, such as tipping in a restaurant (or waiting tables for tips).

Chris Sroka (csroka@stat.ohio-state.edu) is a second-year PhD student in statistics at The Ohio State University. Prior to attending graduate school, Chris worked as an economist for the Congressional Research Service in Washington, DC. He holds a BA and MA in economics from Wayne State University. In addition to EESEE, his research interests include ranked set sampling and survey methodology.

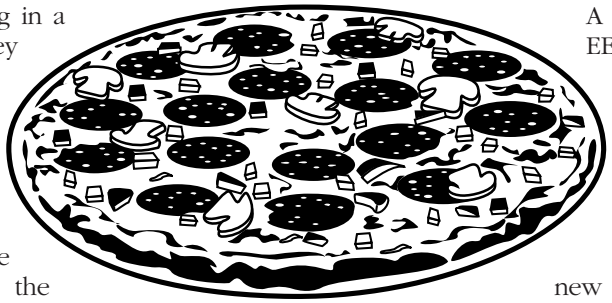
In addition to being interesting, a good story will have data with which students can work. It actually is quite difficult to find such stories. While many statistics are cited in the media, the data that generate them are often hard to come by. One reason is that news stories sometimes do not provide a detailed citation of where information was obtained. For story ideas generated by journal articles, the authors are not always willing to share their data; however, researchers have been very generous with letting us use their data for EESEE stories.

Data that are updated regularly, such as annual survey results, are particularly valuable. The quality of EESEE stories is enhanced when the stories contain the most recent information. We are in the process of developing a web-based tool that allows stories to appear online instantly. Once this is in place, new data can be uploaded as soon as they are available.

Another characteristic of a good story topic is one that generates a lot of interesting questions. If data are available, there are several computational questions that can be written for students. Most of our computational questions ask students to state their answers in terms a nonstatistician can understand. The questions I like writing the most ask students to think critically about the design of an experiment or survey. Some experimenters go to great lengths to avoid bias in their results. A few do not go to any length at all. Students can learn a great deal by dissecting a study, questioning the techniques used, and understanding the limitations of the results. Story ideas come from various sources: newspaper articles, the internet, and television news reports. Frequently, these sources call attention to an interesting study recently published in a journal.

A story recently developed by EESEE was based on a study that tested whether dogs can detect cancer. The experiment was motivated by an anecdote of a dog sniffing a particular part

of its owner's body. When the owner went to the doctor, cancer was discovered in the part of the body the dog was sniffing. The researchers trained seven dogs to identify a urine sample from a patient with bladder cancer. After the training, a randomized experiment was conducted where the dogs had to identify a single cancerous urine sample from among six control samples. As a group, the dogs were successful 41% of the time. If the dogs were acting in a completely random manner, they would have been successful only about 14% (1/7) of the time.



This story asks students to calculate the latter probability and to reach conclusions based on the answer. Stu-dents also are asked several questions about the design of the experiment. What makes the study interesting is the variety of steps the scientists took to ensure the dogs were not providing a conditioned response. The exercises for this story help students identify those steps.

Another recent EESEE story was motivated by three studies examining tipping behavior in restaurants. One study found that food servers receive better tips if they squat, rather than stand upright, when first speaking to a customer. See Figure 1.

A second study found that writing a favorable weather forecast on the back of the bill increased a server's tips, while a third found that servers received higher tips when they delivered a card printed with a joke on it with the bill. Here's the joke used in the experiment: "An Eskimo had been waiting for his girlfriend in front of a movie theater for a long time, and it was getting colder and colder. After a while, shivering with cold and rather infuriated, he opened his coat and drew out a thermometer. He then said loudly, 'If she is not here at 15°, I'm leaving!'"

For each of these studies, the researchers were kind enough to share their data with us. The questions for the tipping story ask students to draw inferences from the data using regression and chi-square tests.

A fellow research assistant for EESEE, Yifan Huang, recently completed a story inspired by a Pizza Hut commercial. Pizza Hut recently introduced a new product that lets a person order four pizzas in a single box.

An advertisement for this new product featured pop sensation Jessica Simpson. She was shown calculating the number of different pizzas that could be ordered, taking into account all the possible topping combinations. In the commercial, she states that more than 6 million combinations are available. Because many college students are huge fans of Jessica Simpson (or at least of pizza), we thought this story would captivate their interest. For this story, students are asked to calculate the actual number of combinations (which turns out to be in the billions).

EESEE contains numerous stories that invite students to explore the relevance of statistics in the world around them. We work hard to make these stories entertaining for students and instructors. With this approach, EESEE lives up to its name by making statistical learning easy. ■

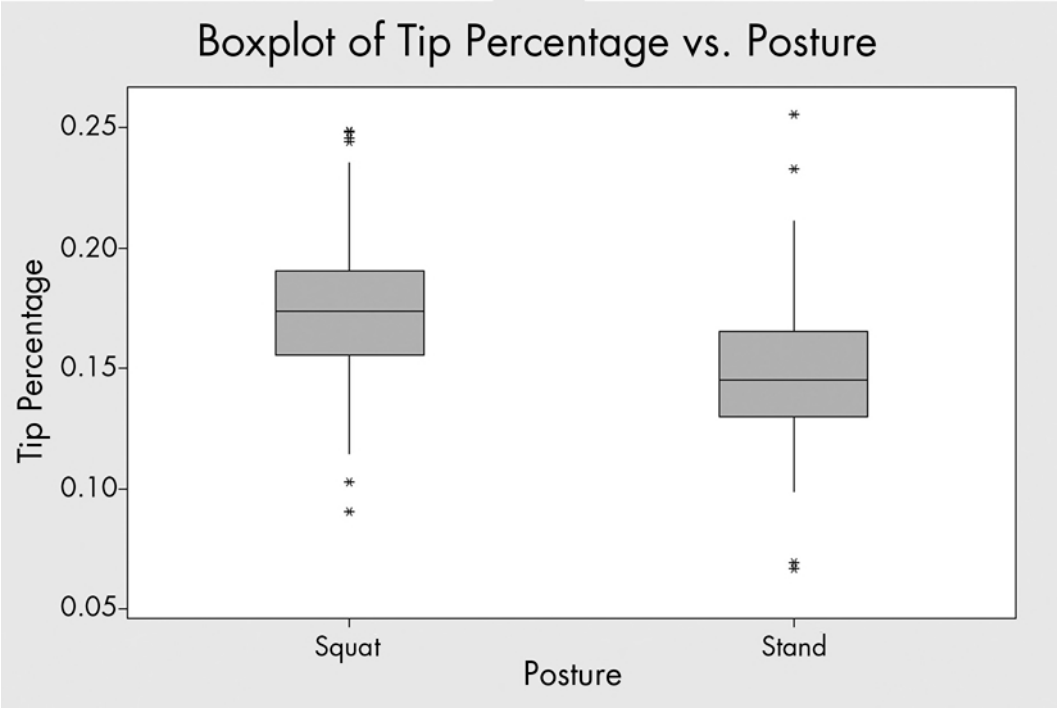
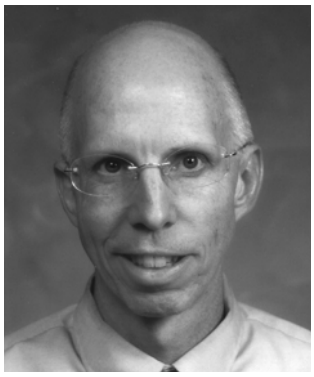


Figure 1. Tip percentage based on food server's posture

Confound It! I Can't Keep These Variables Straight!



Peter Flanagan-Hyde

One of the challenges facing any student when learning a new subject is mastering the vocabulary unique to the discipline. In statistics, there is a large set of vocabulary for beginning students to learn. This task is made more difficult by the fact that some of the words are familiar to them, but used in different contexts—often with quite subtle distinctions.

One of the most important concepts in statistics is the idea of a variable. Here is a definition that is consistent with that found in the first pages of many statistics books (see a sample of textbooks in the References):

A variable is any characteristic of an individual whose value can change from individual to individual.

This leads immediately to other distinctions that depend on the nature of the characteristic being measured: numerical versus categorical variables and discrete versus continuous numerical variables.

More interesting situations arise, however, when studying the relationships between the values of several variables recorded for each individual. The actual relationships are often quite complex, so it is no wonder that this is an area in which students often have trouble. The adjectives applied to the related variables are intended to illuminate the relationships, but for many students, they become a confusing array of similar-sounding terms. Over the course of a discussion, students might hear about:

- response variables
- dependent variables
- explanatory variables
- independent variables
- predictor variables
- associated variables
- confounded variables
- lurking variables
- extraneous variables
- interacting variables
- common response variables
- blocking variables

One issue in sorting this out is that there is not a universally agreed upon definition for some of these terms. That some of these often are called "factors" compounds the problem further. In this article, I'll give simple definitions for some of the terms I believe can help clarify the underlying concepts for students as they work to make sense of the vocabulary.

As a setting for this discussion, consider the issue of elevated cholesterol levels (hypercholesterolemia) in adults in the United States. The cholesterol level of an individual is a variable, because it's not the same for all individuals. Figure 1 is a graph that shows the distribution of values for the United States adult population (Schwartz 1999). The serum cholesterol levels, measured in milligrams per deciliter, vary from the lowest values of about 80 to the highest, near 400.

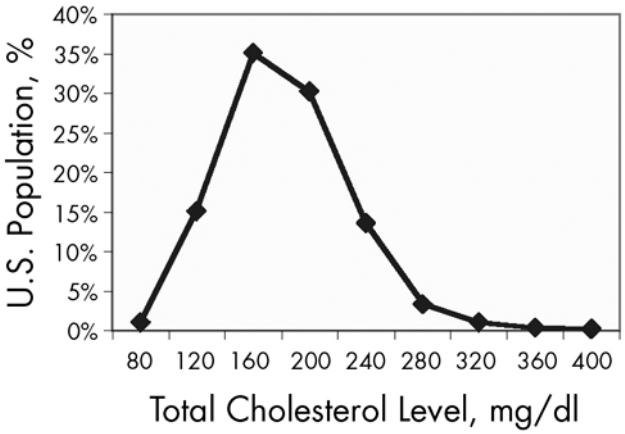


Figure 1: Serum cholesterol levels for adults in the United States

Hypercholesterolemia is defined as a cholesterol level of 240 mg/dL or higher, although some researchers advocate a lower threshold of 200 mg/dL. The risk of this disease seems to be in direct proportion to the cholesterol level. Researchers, therefore, want to be able to identify actions people can take to reduce their cholesterol levels. In this context, the cholesterol level of an individual is the response variable.

Peter Flanagan-Hyde (pflanaga@pcds.org) has been a math teacher for 27 years, the most recent 15 in Phoenix, Arizona. With a BA from Williams College and an MA from Teachers College, Columbia University, he has pursued a variety of professional interests, including geometry, calculus, physics, and the use of technology in education. Peter has taught AP statistics since its inception in the 1996–97 school year.

A response variable measures an outcome whose different values are thought to depend on the values of other variables. (Synonym: dependent variable.)

What aspects of an individual's behavior or makeup determine their cholesterol levels? As measured quantities, researchers are hoping to find one or more explanatory variables.

An explanatory variable is a variable whose different values cause different values in a response variable. (Synonyms: independent variable, predictor variable.)

This definition is narrower than is often found in textbooks, but I believe more focused. It's consistent with the dictionary definitions of explanatory as "serving to explain" and explain as "to make clear the cause or reason of." I would reserve the term explanatory variable for those variables whose causal relationship to a response variable has been established, or at least proposed for study.

In our cholesterol example, researchers ideally would like to be able to draw a figure such as Figure 2, where differences in cholesterol level can be attributed to one or more general categories of explanatory variables.

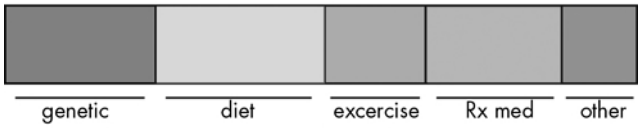


Figure 2: Categories of explanatory variables related to an individual's cholesterol level

For variables whose values are associated with one another, but not necessarily in a cause-and-effect way, I would use the more general term associated variables.

An associated variable is a variable whose different values tend to be matched with values of another variable.

The real issue in the cholesterol study is whether associated variables are explanatory variables. A number of variables might be determined to be associated with higher cholesterol levels through an observational study of the general population. Here is a plausible list of such variables, given current research about this topic:

1. Higher levels of dietary fats.
2. Lower levels of dietary fiber.
3. Lower levels of exercise.
4. Higher levels of body mass index (BMI)—a measure of appropriate weight for a given height.
5. Higher levels of hypertension (high blood pressure).
6. Higher levels of family history of cardiac disease.
7. Higher doses of statins—a class of pharmaceutical drugs.

Which, if any, of these are explanatory variables? The problem with using an observational study to determine if any of these associated variables are explanatory variables is that they are potentially confounded.

Confounded variables are two or more associated variables whose effects on a response variable cannot be separated from one another.

For example, the study might reveal that participants who have higher BMI levels have higher cholesterol levels. But, they also may have higher levels of dietary fats as well, or lower levels of exercise. So which factor is causing the difference in cholesterol levels? It is impossible to say. Note this important characteristic of confounded variables: Two variables can be confounded only if they are associated with each other as well as the response variable.

To establish that a variable is an explanatory variable, a randomized comparative experiment must be done. Participants are randomly assigned to groups, and the researcher fixes the potential explanatory variable at different levels for the different groups. This allows the researcher to isolate the effects of one variable because the random assignment means the effects of other variables, not of interest to the researcher at the moment, are distributed more or less evenly among the groups. These other variables, which are not the focus of the study, are lurking variables.

A lurking variable is a variable that is associated with the response variable, but is not examined as a potential explanatory variable in a study. (Synonym: extraneous variable.)

Here are several experiments that illustrate these issues. In each, the researcher begins with a group of volunteers with elevated cholesterol levels.

Study #1: Half of the group is randomly assigned to have a bowl of oatmeal and a banana for breakfast (7 grams of fiber), while the other half has bacon and eggs (0 grams of fiber). At the end of the study, the mean cholesterol level in the high-fiber group is lower by a statistically significant amount.

Analysis: The potential explanatory variable is dietary fiber. Randomization distributes the effects of many of the lurking variables. However, there are other dietary differences in the treatments, including fat (3 grams for the oatmeal and banana versus 16 grams for two strips of bacon and two eggs). This lurking variable is not accounted for in the study. No conclusion can be reached about the effect of dietary fiber on cholesterol level. This is a poorly designed study.

Study #2: Half of the group is randomly assigned to an exercise group, with 60 minutes of brisk walking done four times a week. The other half has no exercise program. At the end, the mean cholesterol level in the exercise group is lower than for the non-exercise group. The researcher notes that the exercise seems to have other benefits as well, because mean weight in the exercise group decreased lower by a statistically significant amount.

Analysis: The potential explanatory variable is exercise. The effects of many of the lurking variables (dietary differences, hypertension, family history, etc.) are distributed equally to the two groups through randomization. However, because the exercise group also lost weight, exercise is confounded with BMI. No

conclusion can be made about whether exercise reduces cholesterol level. This study was not necessarily poorly designed, but the fact that exercise and BMI are associated led to unanticipated confounding.

Study #3: Half of the group is randomly assigned to be given a dose of statins and the other half receives a placebo. The study is carried out in a double-blind manner so that both the participants and those who have contact with them don't know the true nature of the pills they are taking. At the end of the study, the mean cholesterol level in the statin group is lower by a statistically significant amount.

Analysis: The potential explanatory variable is the statins. The effects of lurking variables are distributed to the two groups through randomization. Since there are no other differences between the treatment groups, statins cause a reduction in cholesterol level.

Here are a couple of observations about these studies. In the first study, confounding occurred because the treatment, itself, included more than one variable, but only one was the variable of interest. The solution to this problem is to make the treatments as alike as possible in all respects except for a single explanatory variable.

In the second study, this could have been a problem as well if the nonexercise group was given less attention. The psychological effect of the attention provided to the exercise group cannot be overlooked as a source for the improvement.

The second study also illustrates what are sometimes called interacting variables.

Two variables are interacting variables if changes in one are associated with changes in the other, or the effect of one changes for different values of the other.

The third study shows a potential paradox. An observational study could show a positive association between statin use and high cholesterol, yet a randomized experiment as in Study #3 could show just the opposite. To understand this, we must remember that in the general population, the only people who are prescribed a statin are those who struggle to maintain a low cholesterol level. So even if the drug helps reduce a patient's cholesterol level, it still may be higher than in the rest of the population.

Several of the variables that can be associated with cholesterol level are not candidates for a randomized experiment. These include hypertension and a family history of cardiac disease. In the case of the family history, this is a characteristic of the individual that the researcher cannot control. To assess the effect of this variable, careful observational studies are the only option.

In the case of hypertension, this is not controlled easily, and the means for control include many of the same lifestyle changes that lower cholesterol—notably diet and exercise. Hypertension and cholesterol level are examples of common response variables.

Common response variables are two variables affected by the same lurking or explanatory variables and are associated with each other.

Let me end with one other type of variable that students may encounter: the blocking variable.

A blocking variable is used to divide an experimental group into groups (blocks) with similar characteristics prior to randomization. Participants are then randomly assigned to a treatment within each of the blocks.

Blocking variables have two purposes. The first is to more reliably counter the potential effects of lurking variables on the outcome of the study. Randomization can guarantee only treatment groups that are the same in the long run. In an individual study, one of the lurking variables may be unevenly divided among the groups, leading to a poor estimate of the effect of the explanatory variable. For example, in the experiment with statin above, a researcher might choose to block on BMI, insuring that both the statin group and the treatment group are identical in this regard. Then, observed differences can be more confidently attributed to the explanatory variable.

The second purpose is to determine if the effect of an explanatory variable is different for different groups. Using the statin example again, it is possible that the drug works differently, either better or less well for males and females. By blocking on gender, this difference can be assessed.

The types of variables we use in statistics correspond to concepts that are important to understand individually and in how they relate to one another. If you are having trouble with the vocabulary of variables, think about the concepts involved, reread the definitions in your textbook, and look at examples of statistical studies. These can help with the foggy sense that sometimes accompanies learning to match the words with the concepts they represent.

References

Moore, David S. *Statistics: Concepts and Controversies*. New York: W.H. Freeman Co., 2001.

Moore, David S. and George P. McCabe. *Introduction to the Practice of Statistics*. New York: W.H. Freeman Co., 2003.

Peck, Roxy, Chris Olsen, and Jay Devore. *Introduction to Statistics and Data Analysis*. Belmont, CA: Brooks/Cole, 2005.

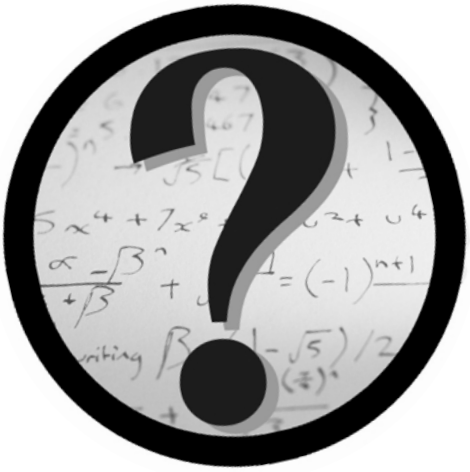
Schwartz, Lisa M. and Steven Woloshin, Changing Disease Definitions: Implications to Disease Prevalence. *Effective Clinical Practice 2* (March/April 1999): 76–83.

Utts, Jessica M. and Robert F. Heckard. *Mind on Statistics*. Belmont, CA: Brooks/Cole, 2004.

Watkins, Ann E., Richard L. Scheaffer, and George W. Cobb. *Statistics in Action*. Emeryville, CA: Key Curriculum Press, 2004.

Yates, Daniel S., David S. Moore, and Daren S. Starnes. *The Practice of Statistics*. New York: W. H. Freeman Co., 2003. ■

Correction: A web site address printed in "Microarray Data from a Statistician's Point of View" was printed incorrectly in the Winter 2005 issue. It should have read *www.bio.davidson.edu/courses/genomics/chip/chip.html*.



ASK STATS

Jackie Miller

Welcome to the first *STATS* question-and-answer column! Because this is the first time this column has appeared, I chose to ask an expert the following questions:

1. What do you love about statistics?
2. Why should I major in statistics?
3. What can I do with statistics once I get out of school?
4. What's cool about statistics?

For this issue, we have responses from Dr. Robert Gould, vice-chair of undergraduate studies at the Department of Statistics at UCLA.

What does Dr. Gould love about statistics? Here is his response:

When I was in graduate school, I loved the fact that I could tell my family and friends what I was working on. My math friends, when asked what they were studying would say, "Well, um, you see, uh, suppose H is a Hilbert space and you have a bounded function that, uh...." I, on the other hand, could say "I'm looking at improving methods for assessing mental performance of Alzheimer's patients and trying to find evidence of genetic links to some performance measures." And people could actually build a conversation on that. Now, I would say that I love the fact that it gives me insight into so many scientific fields.

Dr. Gould thinks you should major in statistics because...

- ***You get to play with computers***
- ***It provides you with a way of making sense of the world***
- ***You get to meet people from many walks of life with fascinating problems***
- ***It helps you solve problems that society finds valuable***
- ***It's employable.***

We have a little problem with Dr. Gould's response about what you can do with statistics once you get out of school. His response was, "You want out of school?"

Good point! Some of us have very exciting positions working with undergraduate and graduate students at colleges and universities across the country and around the world.

"Statistics is cool because it's expansive..."

Robert Gould

Finally, here's what Dr. Gould had to say about what is cool about statistics:

Statistics is cool because it's expansive—you get to "eavesdrop" on other people's professional lives, spend some time with their problems, and then move on. If you're really lucky and you really like their lives, you may not have to move on!

For example, once a week I work with a medical doctor and learn about alcoholism in the elderly and what we can do about it and how we can educate physicians to diagnosis it. Tomorrow, I'm going to meet with someone and learn all about the many

---

Dr. Jackie Miller (miller.203@osu.edu) is a Statistics Education Specialist and auxiliary faculty member in the Department of Statistics at The Ohio State University. She earned both a BA and BS in mathematics and statistics at Miami University, along with an MS in statistics and a PhD in statistics education from The Ohio State University. She is very involved in the statistics education community. When not at school, Dr. Miller enjoys a regular life (despite what her students might think!), including keeping up with her many dogs!

types of fish caught when fishing for tuna and why an international tuna commission is concerned about it. I've learned all about new methods of digitizing film so the major movie studios can "email" movies to theaters rather than send expensive reels. I've learned about the mating rituals of bees, the California prison system, and alleged racial discrimination within a large corporation. Statistics provides a way of approaching problems that's useful in so many endeavors. It allows you to constantly reinvent yourself.

Many thanks to Dr. Gould for responding to these questions and participating in our very first *Ask STATS*! Next time I see Dr. Gould, I will have to ask him to tell me more about some of those cool things he's learned about.... Wow!

Now we'd like to hear from you! What questions do you have about statistics? You can ask just about anything and I will find the answer for you, either by answering the questions myself or by getting expert opinions.

To have your questions answered, please email them to Dr. Jackie Miller at *miller.203@osu.edu*. In the subject line, please type "*Ask STATS*." In the body of the email along with your question, provide your school and level (high school, undergraduate, graduate) as well as the location of your school. Also, please let me know if I can use your name in the column. If we choose your question for publication, the American Statistical Association (ASA) will give you an ASA T-shirt of your choice!

We're looking forward to some exciting questions from our readers, so start emailing! ■

CALL FOR PAPERS

STATS: The Magazine for Students of Statistics is interested in publishing articles that illustrate the many uses of statistics to enhance our understanding of the world around us. We are looking for engaging topics that inform, enlighten, and motivate readers, such as:

- statistics in everything from sports to medicine to engineering.
- "statistics in the news," discussing current events that involve statistics and statistical analyses.
- statistics on the Internet, covering new web sites with statistical resources such as data sets, programs, and examples.
- interviews with practicing statisticians working on intriguing and fascinating problems.
- famous statisticians in history and the classic problems they studied.
- the "statistics almanac" that tell us what happened during this month in statistics history.
- how to use particular probability distributions in statistical analyses.
- examining surprising events and asking, "What are the chances?" and then providing the answers.
- "statistical data sleuth" problems.
- reviews of books about statistics that are not textbooks.
- student projects using statistics to answer interesting research questions in creative ways.

So think of some great ideas, and send a description of your concepts for feature articles that you would write to Dr. Paul J. Fields, Editor, *pjfields@byu.edu*.

What Does It Mean To Be Rational?

A Review of Classical Probability in the Enlightenment

Chris Olsen

It would seem one characteristic of humans in times of stress, depression, and/or out-and-out disapproval of the state of culture is to compare the current point in time with those halcyon times of yore. For example, children fondly wish for 10 minutes ago when they didn't have to go to bed; college students yearn wistfully for a few more minutes of the past when it wasn't yet time to get out of bed; older adults recall when society was civil and doctors made house calls. And some, with a perspective built on a still longer life, must look even further back for the fabled times of yore. Thus it was that I was attracted to *Classical Probability in the Enlightenment (CPE)* by Lorraine Daston, published in paperback by Princeton University Press.

The Enlightenment, as you will recall, was an intellectual movement that is identified with the 18th century. Individuals in the European population centers of the time—London and Paris—were of the opinion they were more enlightened than others and set off to enlighten those around them armed only with the conviction that reason, common sense, and tolerance could build a perfect society. This fixation with rationality survived until the French Revolution and subsequent development of Romanticism. As a mathematical person, this reader can see why one might look back to the Enlightenment with tears in one's eyes (not that there is anything wrong with Romanticism). "Classical" probability, as you will recall, is the probability of the Bernoullis, Pascal, and Fermat, finally set down by the Marquis de Laplace. (Modern statistics students sometimes get this wrong and confuse the Marquis de Laplace with the Marquis de Sade, presumed founder of "sadistics." But, I digress.)

Due to the excellent writings of Stephen M. Stigler and others, the history of the development of probability and statistics is probably not altogether unfamiliar. In many statistics classrooms, students are regaled with some of the history of probability, suggesting a genesis with the Fermat-Pascal correspondence. Teachers of statistics even engagingly shock students' sensibilities with the dirty little secret that probability was originally about gambling. This is not actually false, and yet it is not actually true either. The larger story of the birth of probability is less titillating, but nonetheless much more fascinating—or, at least much more fascinating when told by Ms. Daston.

Jakob Bernoulli

Pierre de Fermat

The story of probability that unfolds in *CPE* is not just a question of whether one should gamble, writ large—the story of probability develops as a 200-year odyssey of mathematicians attempting to discern just what it means for an individual to be rational in the face of incomplete knowledge and an uncertain future:

- When is it rational to buy a lottery ticket?
- When is it rational to accept a scientific hypothesis?
- When is it rational to sell off an expected future inheritance?
- When is it rational to invest in an annuity?
- When is it rational to accept A as the "cause" of B?

In statistics, we frequently "let the data speak for themselves," so let's let the author speak for herself....

Chris Olsen (colsen@cr.k12.ia.us) teaches mathematics and statistics at George Washington High School in Cedar Rapids, Iowa. He has been teaching statistics in high school for 25 years and has taught AP statistics since its inception.

"[*CPE*] is a study of mathematical theory, but a mathematical theory intoxicated by reason. It is consequently intended as a contribution to Enlightenment history as well as the history of science and mathematics. Philosophers, economists, and psychologists may also take an interest in the work of the classical probabilists, for many of the classical problems later migrated to other fields, along with the classical assumptions." [Introduction, p. xvii]

As can be seen, this is more than a book about the history of probability; it is a history of probability in a social and cultural context. One particularly interesting thread of context for me is Daston's description of the legal context within which probability grew. Every statistics teacher in the world has probably talked about Type I and Type II errors using the "innocent until proven guilty" analogy of the legal system. But there's more to the story of the relationship between probability concepts and legal concepts. What Daston brings out in *CPE* is that two of the most distinctive and enduring features of probability are the product of 17th century legal practices! These features—the interpretation of probabilities as degrees of certainty and the primacy of the concept of expectation—were standard legal constructions used in dealing with problems of contract law and annuities. In point of fact, the works of the early probabilists are riddled with legal references!

In closing, let me suggest that in *CPE* one gets the best of many worlds. For the readily offered price of an interest in the history of probability, one also receives a wonderfully written legal, scientific, social, and intellectual tour through the Age of Reason. With the broad perspective provided by *CPE*, your understanding of the history of probability will never again be merely mathematical.

References

Daston, Lorraine. *Classical Probability in the Enlightenment*, Princeton University Press, Princeton, New Jersey, 1995. ISBN 0-691-00644-X.

Stigler, Stephen M. *Statistics on the Table: The History of Statistical Concepts and Methods*, Harvard University Press, Cambridge, Massachusetts, 2002. ISBN 0-674-009797.

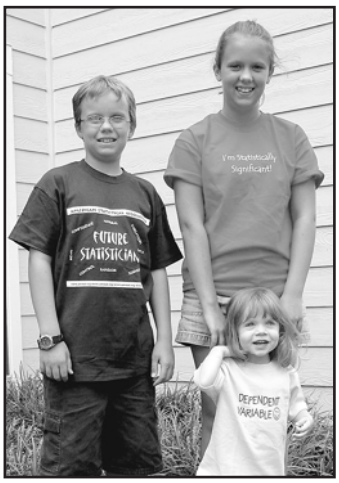
Stigler, Stephen M. *The History of Statistics: The Measurement of Uncertainty before 1900*, Harvard University Press, Cambridge, Massachusetts, 1990. ISBN 0-674-40341-X.

Photo of Pierre de Fermat provided by the University of Rochester, courtesy of AIP Emilio Segré Visual Archives. Photo of Jakob Bernoulli (public domain) provided by T.L. Fine, Cornell University. ■

ASA MARKETPLACE

Statistically Significant Gifts from the ASA

Circle the desired size & color for each selection (please indicate alternate color selection in case your first choice is out of stock). Allow 6–8 weeks for delivery of your items. Please call customer service at 1(888) 231-3473 for any questions.



Nephew Alex, and nieces Emily and Hannah of Carolyn Kesner, ASA Development and Grants Manager

Adult Shirts	Sizes Available	Price	Quantity	Total
Fleece Pullover with ASA Logo (SHIRT-FLECE) Navy	M L XL	$45.00 X	_____ =	_____
Denim Shirt with embroidered ASA Logo (SHIRT-DENIM) Denim	S M L XL XXL	$35.00 X	_____ =	_____
Polo Shirt with "ASA" (SHIRT-POLO2) White	S M L XL XXL	$35.00 X	_____ =	_____
"Got Data?" (SHIRT-DATA) Black White	M L XL XXL	$20.00 X	_____ =	_____
"Top 10 Reasons to be a Statistician" (SHIRT-TOP10) Navy White	M L XL XXL *Alternate Color(s):*_____	$20.00 X	_____ =	_____
"I'm Statistically Significant" (SHIRT-STAT-A) Navy Dark Red Steel Grey	M L XL XXL *Alternate Color(s):*_____	$20.00 X	_____ =	_____
"The Evolution of Statistics" (SHIRT-EVOLVE) White	M L XL XXL	$20.00 X	_____ =	_____
"What Part of Normal" (SHIRT-WHAT) Dark Red Steel Grey	M L XL XXL *Alternate Color(s):*_____	$20.00 X	_____ =	_____
"Approximately Normal" (SHIRT-APPROX) Denim Blue Yellow Steel Grey	M L XL XXL *Alternate Color(s):*_____	$20.00 X	_____ =	_____
"In God We Trust...All Others Bring Data" (SHIRT-INGOD) Denim Blue Black Hot Pink	M L XL XXL *Alternate Color(s):*_____	$20.00 X	_____ =	_____
"Absence of Evidence" (SHIRT-ABSEN) White	M L XL XXL	$20.00 X	_____ =	_____

Youth/Child T-Shirts		Sizes Available	Price	Quantity	Total
"Future Statistician" (SHIRT-FUTURE) Navy Red White	*Alternate Color(s):*_____	**Toddler**: 2T 3T 4T **Youth**: S M L	$10.00 X $10.00 X	_____ = _____ =	_____ _____
"I'm Statistically Significant" (SHIRT-STAT-C) Toddler: Lt Blue Lt Pink Lt Yellow Steel Grey	*Alternate Color(s):*_____	**Toddler**: 2T 3T 4T	$10.00 X	_____ =	_____
Youth: Red Forest Steel Grey	*Alternate Color(s):*_____	**Youth**: S M L	$10.00 X	_____ =	_____
"Dependent Variable" (SHIRT-DEPEND) Toddler: Lt Blue Lt Pink Lt Yellow Steel Grey	*Alternate Color(s):*_____	**Toddler**: 2T 3T 4T	$10.00 X	_____ =	_____
Youth: Red Forest Steel Grey	*Alternate Color(s):*_____	**Youth**: S M L	$10.00 X	_____ =	_____

Logo Items			
Denim Baseball Cap w/ Khaki Brim & ASA logo (HAT-DENIM)	$15.00 X	_____ =	_____
Stainless Steel Travel Mug w/ASA Logo (STEELMUG)	$10.00 X	_____ =	_____
Silvertone Chrome Keychain w/ASA Logo (SKEYCHAIN)	$8.00 X	_____ =	_____
Static Cling Car Window Decal w/ASA Logo (CARDECAL)	$1.00 X	_____ =	_____

SPECIAL CLEARANCE ITEMS								
Off-white Polo Shirt with embroidered ASA Logo (SHIRT-POLO)	S	M				$25.00 X	_____ =	_____
2003 JSM San Francisco (SHIRT-JSM03)		M	L	XL	XXL	$10.00 X	_____ =	_____
2004 JSM Toronto (SHIRT-JSM04)		M	L	XL	XXL	$10.00 X	_____ =	_____

Orders must be prepaid. Send order form and payment to:

ASA Souvenirs
American Statistical Association
1429 Duke Street, Alexandria, VA 22314-3415 USA
or fax to: (703) 684-2037. For more information, call 1 (888) 231-3473
Please make remittance payable in U.S. currency drawn on a U.S. bank.

Subtotal	$ _____
VA residents add 5%	$ _____
Postage & Handling (See postage chart below)	$ _____
TOTAL	$ _____

PAYMENT INFORMATION

❑ Check or money order enclosed for $_____, made payable to ASA.

❑ VISA ❑ MasterCard ❑ American Express for $ _____

Card Number:_____ CVS# (3-digit number on back of card) _____ Expiration Date: _____/_____

Name of Cardholder: _____

Cardholder's Signature:_____

ASA ID#: _____ Telephone # _____

Email: _____ Fax #_____

Ship to: Name_____

Address (No PO Boxes) _____

City:_____ State/ZIP or Postal Code/Country_____

POSTAGE & HANDLING CHART

For U.S. Orders Add:

Up to $10	**$3.00**
$10.01-$25	**$5.00**
$25.01-$50	**$8.00**
$50.01-$100	**$11.00**
Over $100	**$15.00**
CANADA	**$20.00**

INTERNATIONAL & EXPRESS SHIPPING Is available for the actual Shipping cost plus a $3 handling charge. Please call Customer Service at 1(888) 231-3473 to receive an estimate for your items.

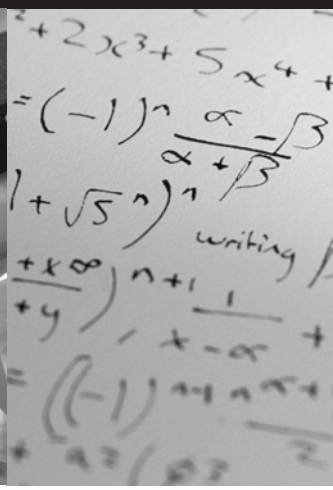
Seeking a **CAREER** in

Minneapolis, Minnesota
August 7–10

STATISTICS?







Are you nearing graduation and wondering about entry-level jobs?

Are you an experienced statistics professional interested in career information?

Register for the JSM Career Placement Service!

What can the CAREER PLACEMENT SERVICE do for YOU?

Each year, hundreds of companies, universities, recruiters, and government agencies search for applicants using the JSM Placement Service. At the 2004 JSM in Toronto, employers listed more than 230 positions for qualified statisticians. The JSM Placement Service provided the best opportunity for qualified applicants to meet employers, establish valuable contacts, and learn about organizations employing statisticians.

Career Placement Service BENEFITS

Applicant Reading Area—area for applicants to review binders containing complete job descriptions and contact information for all registered employers.

Visibility to Employers—applicants who register by July 21, 2005, will have their forms and résumés included in the Advance Applicant access database available to employers prior to the meeting. Employers may contact applicants whose forms are included in the advance database prior to meeting to schedule interviews.

Computerized Message Center—allows applicants and employers to communicate throughout the meeting.

www.amstat.org/meetings/jsm/2005/placement

for more information

Organizations Represented at Recent JSM Career Placement Services

Bureau of Labor Statistics • FDA • Centers for Drug Evaluation and Research • PPD
University of Maryland • Smith Hanley Associates • UCLA • Eli Lilly • Mayo Clinic