ASA
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

NCTM
NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS

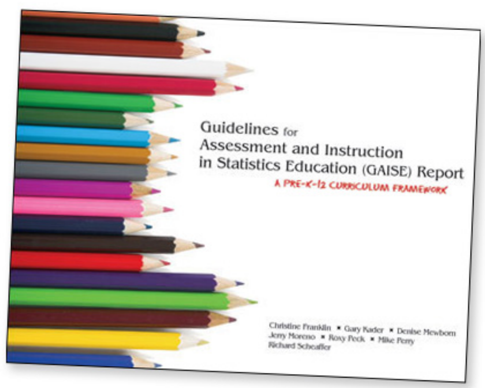
The Statistics Teacher Network

Number 84 | The **S**tatistics **T**eacher **N**etwork | **ASA/NCTM Joint Committee** on **Curriculum** in **Statistics** and **Probability** | Fall 2014

The Editor's Corner

Welcome to Issue 84 of the Statistics Teacher Network. Two articles in this issue offer valuable topics for inspiring your more capable elementary and secondary students. There is also a follow-up article about the LOCUS project. In addition, there is some exciting news about future issues.

The first article, reprinted with permission from the editor of *Understanding Our Gifted*, was written by me and three other colleagues: William P. Bintz of Kent State University, Sara Delano Moore of ETA/Cuisenaire, and Cheryll M. Adams of Ball State University. You'll find ideas for differentiating your instruction to respond to the learning needs of your gifted students. If you're interested in more ideas, check out *www.ourgifted.com,* which is dedicated to providing information to help you understand and work with gifted children.



The second article, from Jennifer Brown, assistant professor at Columbus State University, presents a lesson focused on logistic regression. While this may be a topic you've never considered teaching your high-school students, you might want to give it a try since learning how to predict the probability of winning a football game may inspire your students to go above and beyond. Jennifer provides step-by-step directions for using an Excel template spreadsheet for the calculations.

Catherine Case, doctoral fellow, and Tim Jacobbe, associate professor, at the University of Florida, write about the current NSF-funded Levels of Conceptual Understanding in Statistics (LOCUS) project. The project is focused on developing statistical assessments in the spirit of the GAISE framework. Catherine and Tim tell us about results from the pilot administration of the assessments.

As always, I think you will find all three articles worth the read!

Now, for the exciting news! I want to welcome Angela Walmsley as the Statistics Teacher Network's new editor. Angela becomes editor following this issue and has the credentials and experience necessary to ensure the Statistics Teacher Network stays relevant and engaging into the future. Welcome, Angela!

I encourage you to send any articles or ideas you have for consideration to me at *rpierce@bsu.edu* or Angela at *a.walmsley@neu.edu*.

Best Regards,
Rebecca Pierce, Editor, Ball State University

Associate Editors

Jessica Cohen – Western Washington University
David Thiel – Clark County School District, Retired
Angela Walmsley – Northeastern University, Seattle Graduate Campus
Derek Webb – Bemidji State University

In this issue …

Articles

Using Statistics to Lie, Distort, and Abuse Data

William P. Bintz – Kent State University
Sara Delano Moore – ETA/Cuisenaire
Cheryll M. Adams – Ball State University
Rebecca L. Pierce – Ball State University

(Editor's Note: This article is a reprint of "Using Statistics to Lie, Distort, and Abuse Data" from Understanding Our Gifted, *22(1):14–18, 2009)*

Lying is not telling the truth, and it is taboo in our society. Parents scold children and teachers punish students for lying. Bosses fire employees for lying on the job. Judges even incarcerate witnesses for committing perjury [lying] in court. And yet, lying has been, and continues to be, ubiquitous. For whatever reasons, and there are many, people lie, and they have been doing so for centuries. In large part, this is because there are so many ways for people to be deceptive. People use language, numbers, pictures, technology, body motions, facial expressions, and hand signals to deceive others. They also use statistics—sometimes clumsily, but most often expertly, cleverly, and even surreptitiously—to distort data.

There are three kinds of lies: lies, damned lies, and statistics.

— Disraeli

Statistics is a branch of mathematics that involves organization, presentation, and interpretation of data, both quantitative and qualitative. Unlike people, data don't lie. On the surface, quantitative data are basically inanimate objects, nothing more than lifeless and meaningless symbols that can appear on a page, calculator, computer, or just in the mind. That's why statistics can't state, suggest, or imply anything about numerical data. Only people can do this. Only people can metaphorically breathe life into numerical data and make them meaningful and comprehensible. Simply stated, statistics can't talk; rather, people use statistics to talk about what a set of numerical data means to them. From this perspective, statistics can't tell truths or lies. Only people can use statistics to tell honest truths or blatant lies, something in between, or even both.

The purpose of this article is to show how people do use statistics as a tool to lie, distort, and abuse data. We begin by describing the relationship between statistics and data and discuss what teachers need to do to help students better understand this relationship. Next, we illustrate how teachers can use literature to help students organize, graph, and analyze data so students understand working with data in a variety of ways. We end with several questions students and teachers can and should ask to make sure people who lie, distort, and abuse data aren't successful, or at least are questioned and challenged.

Statistics & Data

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."

— H. G. Wells

There is an important and interesting relationship between statistics and data. In general, people think of numbers as data, but technically in statistics, numbers are not data unless you place them in context. For instance, "69" is not considered data unless you include "69 inches," where it might represent the height of a person or "69 pounds," where it might represent the weight of your dog.

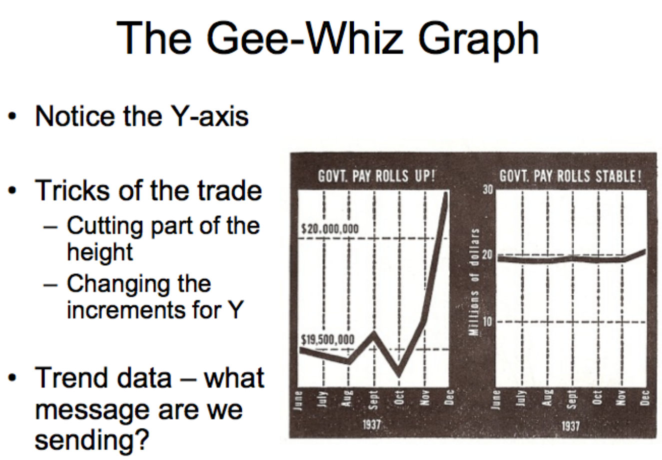
Statistics is considered the interface of mathematics with science and each academic discipline, or content area, has a unique language. Thus, data is *central to everything*. The two types of data, quantitative and qualitative, can cause confusion. Do our students know what to do? What's a teacher to do?

Organizing & Presenting Data

"He uses statistics as a drunken man uses lamp-posts.… For support rather than illumination."

— Andrew Lang (1844–1912)

Questions: How do we help students understand that presenting information can distort, rather than illuminate, data? How do we help students organize and present information? What role does literature play in both? For example:



In the graphs above (Huff, 1954), the same data are shown. The appearance of the data is quite different, however, because of the scale presented on the Y axis. In the graph to the right, the Y axis is shown down to a value of $0 and thus the changes appear quite small. In the graph on the left, the Y axis shows approximately $1 million (compared to $30 million on the right) so the same changes appear quite large. Depending on the message being sent, the same data can be used to suit either headline.

To understand the data well enough to see various ways of presenting it, students need experience collecting, organizing, and exploring data. For students learning to explore data and graph, the use of manipulatives can be an important component of understanding. Line graphs can be created using tools such as geoboards or the XY Coordinate Pegboard. Bar graphs can be created with linking cubes. Circle graphs can be created with colored links. In each case, children are literally constructing the data representation rather than sketching it. This makes the decisions about how to represent data more personal and real, especially for younger students. Students can then progress from these hands-on graphs to figures drawn on graph paper or created with a calculator or computer.

Literature is also important, especially informational text. Literature is a powerful tool to integrate literacy and mathematics, especially statistics (see Appendix A for a list of books that can be helpful here). It provides a human perspective on mathematics by helping students see mathematics operating in an authentic context by showing people putting mathematics to good use (Whitin and Wilde, 1995). In this way, literature helps students connect the abstract language of mathematics to the real world (Whitin and Whitin, 1997). It captures children's imagination, stimulates their mathematical thinking and reasoning (Burns, 1992, p. 166), and provides visualizations of mathematical concepts through illustrations. Finally, math scores increase when math strategies are combined with high-quality literature (Jennings, 1992).

Bush Babies by Kim Dale is an award-winning trade book that was short-listed for the 2003 picture book of the year by The Children's Book Council of Australia. It is a collection of rhyming poems, informational prose, and colorful illustrations that pays tribute to animals who suffered in the Australian drought and bush fires of 2002–2003. Each poem and illustration identifies and provides information about animals that live in the bush. These animals include dingo pups, brushtail possums, sugar gliders, emu chicks, wedgetailed eaglets, puggles, koalas, and wombats. Additional information about each animal is provided at the end of the book. Here's information about the Eastern Grey Kangaroo (Dale, 2003):

> *The mature male kangaroo stands approximately 2 m tall and weighs as much as 70 kg. Kangaroos' habitats range from semi-arid scrub to woodlands and forests. They are mainly grazing animals, preferring grass, which they eat from dusk to dawn. Eastern grey kangaroos are distributed throughout eastern Australia, including Tasmania. Kangaroos have extremely strong hind legs that enable them to hop at very fast speeds, using their tail for balance.*
>
> *The female gives birth to a tiny joey, only 2 cm long. It finds its way into the pouch, attaching itself to one of two teats. It feeds on milk for eleven months, then leaves the pouch but still suckles while grazing for another seven months. A new joey will be in the pouch by then. The mother produces different milk for each joey.*

This kind of information is interesting and informative, especially to young readers who prefer nonfiction to fictional text.

Graphing Data

A book like this provides data that can be graphed and explored. The chart below shows data presented in the book about the length of the tail and weight of a joey. Students can graph these data using manipulatives or on paper and explore the relationship (if any) between these two variables. In other words, they can explore a question like this: As a joey grows heavier, does the length of the tail increase at a predictable rate?

Graphing Data

Eastern Grey Kangaroo Growth Chart

Tail Length (Inches)	Weight (Ounces)
5	19
10	43
12	66
15	83
21	199

Analyzing Data

"Round numbers are always false."

— Samuel Johnson

Questions: What kind of data (nominal, ordinal, interval/ratio)? What is the best measure of central tendency (mean, median, or mode)? Are there outliers? What impact do they have on data?

Students need experience analyzing data. Surveys provide excellent examples of various types of data: "How long is your travel time to school" generates different information than "what is your favorite color." By exploring both, students come to understand when to use a mean versus a median or mode to describe the middle of data. They learn about the impact of outliers (one "weird" number) and how to respond without starting over.

- This is a weird number, teacher. Can I start my experiment over?
- What case am I trying to make?

How to Talk Back to a Statistic

"It ain't so much the things we don't know that get us in trouble. It's the things we know that ain't so."

— Artemus Ward

Although we can hope, it isn't likely that the act of using statistics to lie, distort, and abuse data will end anytime soon. People will continue to believe data talk and statistics helps us hear what the data are saying, as well as distort, if not ignore, the fact that people make data talk and statistics is one way people say what it means to them. Here, we share several questions students can ask to make sure people who lie, distort, and abuse data aren't successful, or are at least questioned and challenged. These questions include: Who says so? How do they know? What's missing? Did somebody change the subject? Does it make sense?

Who says so?
This question addresses the issue of bias. Who paid for the research to be conducted? Does the person writing/talking have a preferred solution to support? What confidence do we have that the author is presenting the data fairly?

How do they know?
This question addresses sampling. What kind of study was conducted? If the study was conducted via the Internet and was asking questions of nursing home residents, it is not likely to be a representative sample. Likewise, a study reporting to tell about all teenagers that was conducted at one exclusive shopping mall on one day will not result in a representative sample. Did the person conducting the study make a reasonable effort to get a good assortment of participants—teenagers who like shopping as well as those who like sports as well as those who like reading as well as those who like the arts and so forth?

What's missing?
What's not reported about the data? Knowing a mean without the range can distort perspective. When talking about the mean salary, knowing it is $30,000 with a range of $20,000 to $50,000 gives a very different picture than a mean of $30,000 with a range of $10,000 to $100,000.

Did somebody change the subject?
Do the data presented actually answer the question asked? The number of reported flu cases increased dramatically in spring 2009 because of concern over H1N1 (swine) flu. Only a limited number of cases were confirmed to be the new type, but a great many more cases were reported because people were more likely to go to the doctor if they had the flu. It's not that more people were sick, just that more people were talking about it.

Does it make sense?
This fundamental filter is often the best one of all. Take a look at the data presented and ask yourself, "Does this make sense to me?" If not, go back and look at some of the issues raised here to see where the problem might be.

Conclusion

What power and potential does this topic have for students who are gifted? Often, teachers have a difficult time selecting topics that challenge gifted students mathematically, particularly those who have been identified as gifted in mathematics. Differentiating the instruction is one way to respond to the learning needs of gifted students. Providing challenging activities that target the advanced learning needs of gifted students allows them to work at a level that is appropriate for them (Adams & Pierce, 2006).

A particular characteristic of gifted students is the need and ability to pursue a topic in depth. While most textbooks provide some extension activities for students who grasp the concepts quickly, these extensions are generally not challenging for gifted students, since the textbook extensions do not provide activities that foster in-depth learning. Because these students can make connections among a variety of subjects easily, combining literature, science, and mathematics provides a perfect vehicle for studying data and statistics.

Students who are mathematically gifted should work with advanced mathematical concepts at a variety of levels. These students can use the higher-level critical thinking skills of analysis, synthesis, and evaluation to apply key concepts and issues in their advanced study of statistics. Gifted students who are curious, persistent, and have flexible thought processes enjoy pursuing the kinds of questions mentioned earlier: Who says so? How do they know? What's missing? Did somebody change the subject? Does it make sense? They enjoy the opportunity to work as practicing professionals (e.g., scientists or statisticians) to determine if people have lied with or distorted statistics. Commercials and infomercials from television or magazine advertisements can provide a wealth of statistical data that could be perused for distortion. The opportunity to debate the pros and cons of a particular claim would be relished by many gifted students.

Whether a multidisciplinary unit on statistics is used as acceleration or enrichment does not matter. What does matter is how the unit is constructed: The material presented is advanced, requires students to think both critically and creatively, makes use of real-world issues and problems, and allows students to work as practicing professionals.

Appendix A

Best, J. 2001. *Damned lies and statistics: Untangling numbers from the media, politicians, and activists*. Berkley, CA: University of California Press.

Best, J. 2004. *More damned lies and statistics: How numbers confuse public issues*. Berkley, CA: University of California Press.

Best, J. 2008. *Stat-Spotting: A field guide to identifying dubious data*. Berkley, CA: University of California Press.

Gonick, L. 1993. *Cartoon guide to statistics*. New York: Collins.

Holmes, C. B. 1980. *The honest truth about lying with statistics*.

Huff, D. 1954. *How to lie with statistics*. New York: Norton.

Jones, G. E. 2006. *How to lie with charts*. BookSurge Publishing.

Seligman, D. 1997. *Lies, damned lies, and politically motivated statistics*.

Monmonier, M. 1996. *How to lie with maps*. Chicago, IL: University of Chicago Press.

Paulos, J. A. 1997. *A mathematician reads the newspaper*. Anchor.

Smithson, M. 2000. *Statistics with confidence*.

Takahasi, S. 2008. *The manga guide to statistics*. No Starch Press.

References

Adams, C. M., and R. L. Pierce. 2006. *Differentiating instruction: A practical guide to tiered lessons in the elementary grades*. Waco, TX: Prufrock Press.

Burns, M. 1992. *About teaching mathematics: A k-8 resource*. Sausalito, CA: Math Solutions Publications.

Dale, K. 2003. *Bush babies*. Melbourne, AU: Lothian Press.

Huff, D. 1954. *How to lie with statistics*. New York: Norton.

Jennings, C. M. 1992. Increasing interest and achievement in mathematics through children's literature. *Early Childhood Research Quarterly* 263–276.

Whitin, D. J., and S. Wilde. 1995. *It's the story that counts*. Portsmouth, NH: Heinemann.

Whitin, D. J., and P. Whitin. 1997. *Inquiry at the window: Pursuing the wonders of learners*. Portsmouth, NH: Heinemann.

K–12 Teachers:

Free Trial Membership

Sign up today for your **FREE** 3-month trial membership and receive:

- **Subscriptions to** *Amstat News*, the ASA's monthly membership magazine, now fully online and interactive, and *Significance*, an ASA and RSS partnership magazine aimed at international outreach to enhance both organizations, the statistics profession, and statisticians.
- **Members-only access** to the ASA's top journals and resources—including online access to *CHANCE* magazine, the *Journal of Statistics Education*, and *The American Statistician*—in addition to discounts on all ASA meetings and products.
- **Access to teaching resources**, including webinars, the *Statistics Teacher Network*, *GAISE: A Pre-K–12 Curriculum Framework*, and the *Statistical Significance* series.
- **Information about** upcoming events and products for K–12 teachers, including statistics education workshops, webinars, student competitions, publications, and online peer-reviewed lesson plans. Visit the Education section of the ASA website to learn more.

JOIN THE ASA

A special offer tailored for K–12 educators!

The American Statistical Association wants to help you enhance your students' statistical education.

Visit *www.amstat.org/membership/k12teachers* for details.

join

The Probability of Winning a High-School Football Game

Jennifer L. Brown – Columbus State University

We, as teachers, struggle to find instructional activities that actively engage the interests of our students. In the southeast, on any given Friday night during the months of August through November, the local high school's stadium is filled with football fans of all ages. Fortunately for teachers, after a weekend filled with football games, the media presents a wide range of statistics to explain the game's ultimate result: win or lose.

Thinkstock

During a high-school, college, or professional football game, the sports broadcasters will discuss probabilities, frequencies, and overall season descriptives regarding offensive or defensive plays (e.g., third down and four, field goal from 50 yards, or goal line stand by the defense). Even though the general game format is the same for high-school, college, and professional football, rules and regulations differ depending on the level, conference, and division. Often, the high stakes involved with these high-school games are overlooked compared to college and professional games.

There have been a number of studies (Wagner, 1985; Willoughby, 2002) about predicting the outcome of professional and college football games. Are the same variables used to determine the probability of winning a game with college and professional teams applicable to high-school football teams? This lesson provides a way to answer this question using logistic regression, a nonparametric statistical modeling method.

Background

What is logistic regression? Logistic regression is comparable to multiple linear regression, a parametric statistical method; however, the dependent variable is dichotomous. In addition, logistic regression uses a maximum likelihood estimation procedure instead of the ordinary least squares for linear regression (Tansey, White, Long, & Smith, 1996). Maximum likelihood estimation results in a linear equation with beta weights to maximize the predictability of the group membership. With this method, you are determining the probability of an event occurring (e.g., winning a football game) given a specific condition (e.g., number of passing yards or number of first downs). These plotted probabilities based on the event take an S shape. This method transforms the data into a linear shape by changing probabilities into odds then computing the natural logarithm of the odds, referred to as log-odds. Since the dependent variable is dichotomous, the assumptions of logistic regression include the independent and dependent variables have a nonlinear relationship, but the homogeneity of variance is not applicable and the errors are not normally distributed.

Lesson Details

This lesson is intended for an Advanced Placement Statistics course, an upper-level statistics course at the secondary level, or an introductory statistics course at the post-secondary level. It covers the following NCTM Standard under the Data Analysis and Probability Content Standard: develop and evaluate inferences and predictions that are based on data. According to the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report* (2007), the Pre-K–12 curriculum should incorporate more advanced statistical concepts to nurture the mature statistician and prepare him or her for post-secondary coursework. Within the Common Core Standards, students are expected to summarize, represent, and interpret data on two categorical and quantitative variables. In addition, students are expected to make inferences and justify conclusions from sample surveys, experiments, and observational studies.

For this lesson, students should have a basic understanding of the following concepts:

- Ordinary least squares regression
- Levels of measurement
- Nonparametric statistics (e.g., chi square)
- Summary statistics
- Probability
- Odds
- Natural logarithm

The instructional objective is for students to apply the logistic regression statistic to real-world data. The materials needed are Internet access, Microsoft Excel, and the logistic regression template spreadsheet (*www.forecastingsolutions.com/downloads%5CLogisticRegression.xls*).

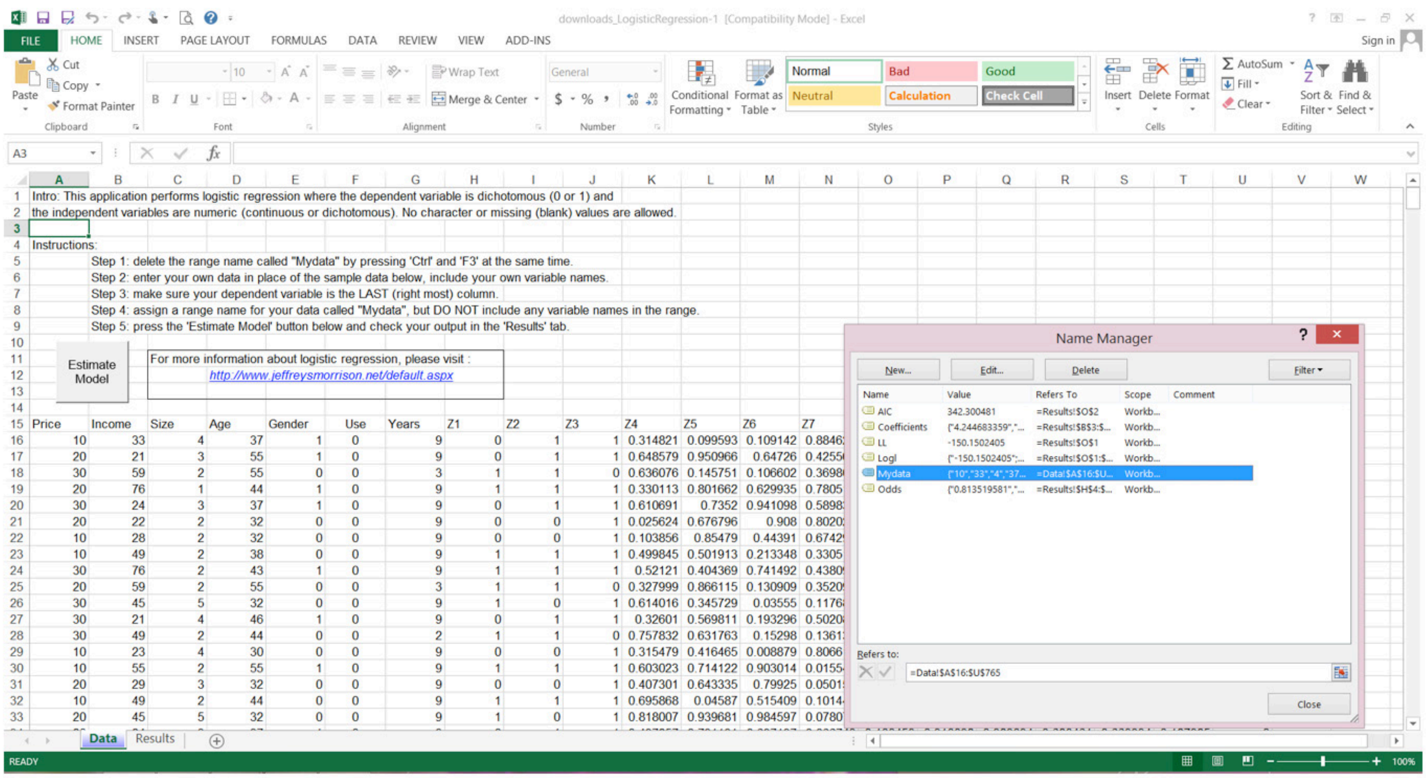
Instructional Procedures

The following sequential procedures outline the lesson using high-school football data easily available on the Internet. For purposes of this article, I selected four Division 5-A high-school football teams from a southeastern state.

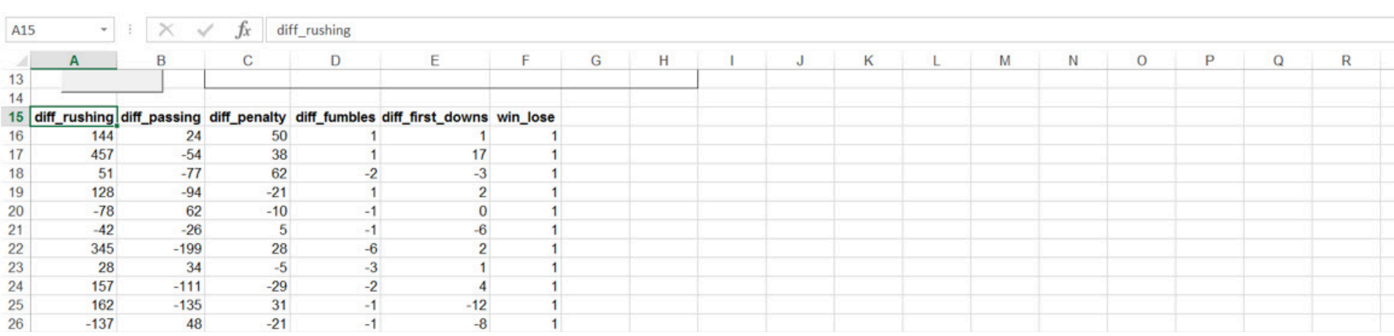
1. Collect data game box statistics from the team's website or other media source (e.g., newspaper). A total of 46 individual game box statistics were collected from the 2006 regular and postseason seasons with 31 overall wins (67.4%) and 15 overall losses (32.6%). If one team played another team within the same set of selected participants, that game was entered for the winner of the football game and not entered for the loser of the game.

2. Open the downloads_LogisticRegression.xls spreadsheet from *www.forecastingsolutions.com/downloads%5C LogisticRegression.xls.*

3. Follow Step 1 directions within the template: Delete the range name called "Mydata" by pressing 'Ctrl' and 'F3' at the same time.



4. Enter data into Excel file template. The numeric variables include (a) difference between the teams and the rushing yards; (b) difference between the teams and the passing yards; (c) difference between the teams and the penalty yards; (d) difference between the teams and the fumbles; (e) difference between the teams and the first downs; and (f) win/lose [(coded with 0 (lose) and 1 (win)], which is the dependent variable. The last five variables are the independent or predicting variables. (Note: the dependent variable).



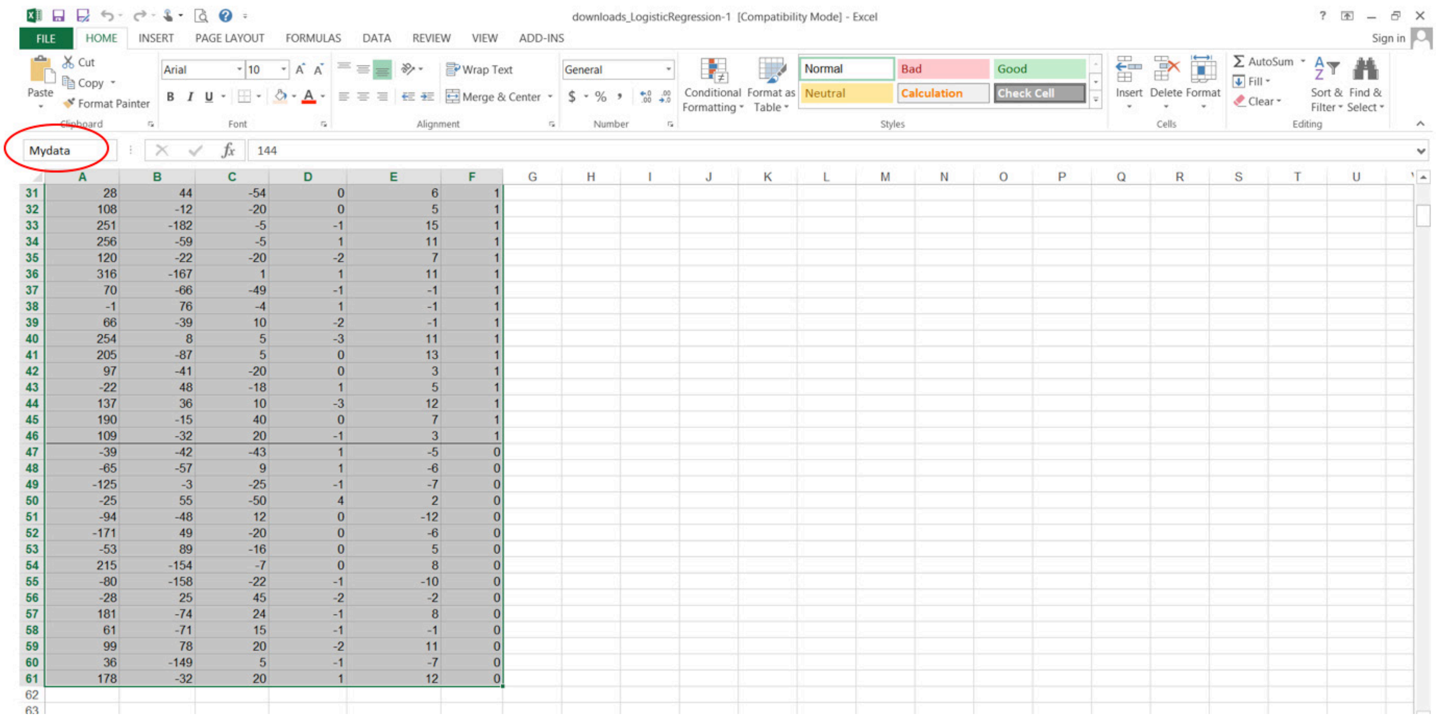
	diff_rushing	diff_passing	diff_penalty	diff_fumbles	diff_first_downs	win_lose
15	diff_rushing	diff_passing	diff_penalty	diff_fumbles	diff_first_downs	win_lose
16	144	24	50	1	1	1
17	457	-54	38	1	17	1
18	51	-77	62	-2	-3	1
19	128	-94	-21	1	2	1
20	-78	62	-10	-1	0	1
21	-42	-26	5	-1	-6	1
22	345	-199	28	-6	2	1
23	28	34	-5	-3	1	1
24	157	-111	-29	-2	4	1
25	162	-135	31	-1	-12	1
26	-137	48	-21	-1	-8	1

5. Follow Step 1 directions within the template: Assign a range name for your data called “Mydata,” but *do not* include any variable names in the range.

Highlight all the cells with entered data. (Note: Do not highlight the variable names.)

In the Name box (to the left of the Function box), type “Mydata.”

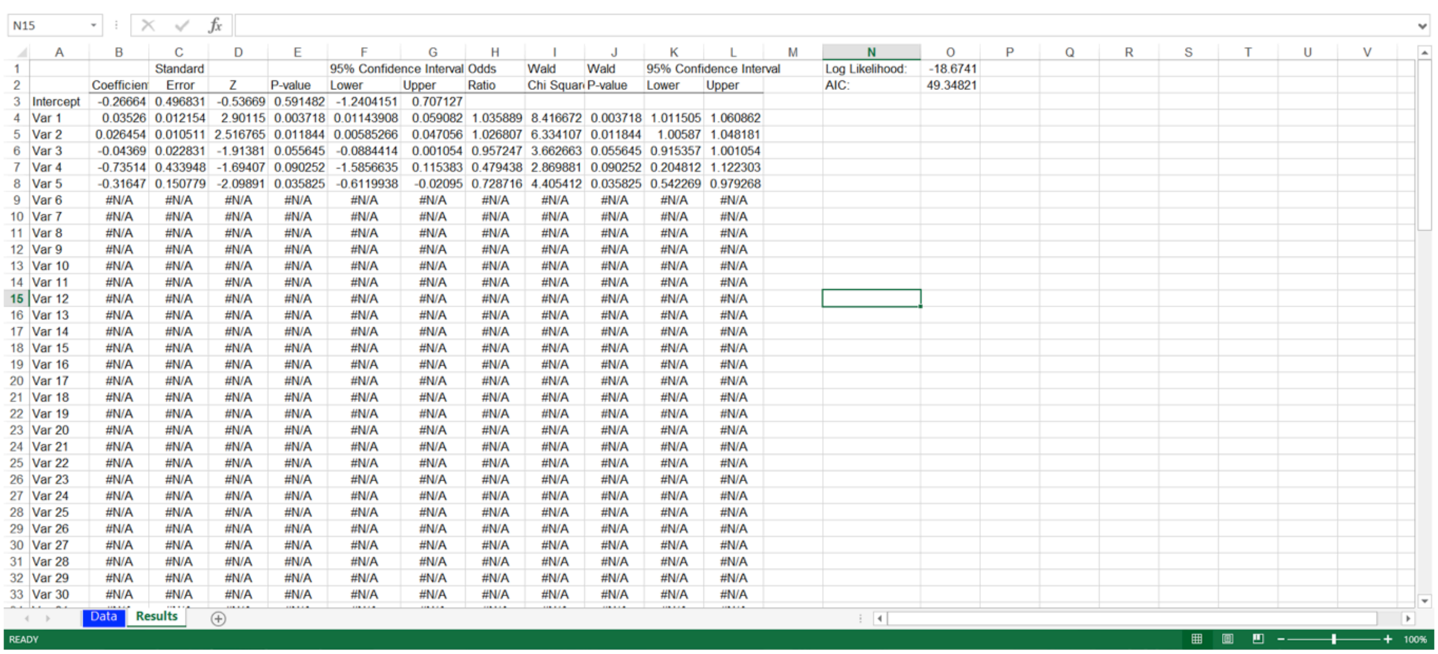
Press ENTER.



	A	B	C	D	E	F
31	28	44	-54	0	6	1
32	108	-12	-20	0	5	1
33	251	-182	-5	-1	15	1
34	256	-59	-5	1	11	1
35	120	-22	-20	-2	7	1
36	316	-167	1	1	11	1
37	70	-66	-49	-1	-1	1
38	-1	76	-4	1	-1	1
39	66	-39	10	-2	-1	1
40	254	8	5	-3	11	1
41	205	-87	5	0	13	1
42	97	-41	-20	0	3	1
43	-22	48	-18	1	5	1
44	137	36	10	-3	12	1
45	190	-15	40	0	7	1
46	109	-32	20	-1	3	1
47	-39	-42	-43	1	-5	0
48	-65	-57	9	1	-6	0
49	-125	-3	-25	-1	-7	0
50	-25	55	-50	4	2	0
51	-94	-48	12	0	-12	0
52	-171	49	-20	0	-6	0
53	-53	89	-16	0	5	0
54	215	-154	-7	0	8	0
55	-80	-158	-22	-1	-10	0
56	-28	25	45	-2	-2	0
57	181	-74	24	-1	8	0
58	61	-71	15	-1	-1	0
59	99	78	20	-2	11	0
60	36	-149	5	-1	-7	0
61	178	-32	20	1	12	0

6. Press “Estimate Model”

7. Select the Results tab at the bottom left.



	Coefficien	Standard Error	Z	P-value	95% Confidence Interval Lower	Upper	Odds Ratio	Wald Chi Squar	Wald P-value	95% Confidence Interval Lower	Upper
Intercept	-0.26664	0.496831	-0.53669	0.591482	-1.2404151	0.707127					
Var 1	0.03526	0.012154	2.90115	0.003718	0.01143908	0.059082	1.035889	8.416672	0.003718	1.011505	1.060862
Var 2	0.026454	0.010511	2.516765	0.011844	0.00585266	0.047056	1.026807	6.334107	0.011844	1.00587	1.048181
Var 3	-0.04369	0.022831	-1.91381	0.055645	-0.0884414	0.001054	0.957247	3.662663	0.055645	0.915357	1.001054
Var 4	-0.73514	0.433948	-1.69407	0.090252	-1.5856635	0.115383	0.479438	2.869881	0.090252	0.204812	1.122303
Var 5	-0.31647	0.150779	-2.09891	0.035825	-0.6119938	-0.02095	0.728716	4.405412	0.035825	0.542269	0.979268
Var 6	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 7	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 8	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 9	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 10	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 11	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 12	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 13	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 14	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 15	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 16	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 17	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 18	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 19	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 20	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 21	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 22	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 23	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 24	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 25	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 26	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 27	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 28	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 29	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 30	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A

Log Likelihood:	-18.6741
AIC:	49.34821

8. To predict log-odds, or logit (L), you would use the following equation based on the beta "coefficient" weights and constant value, which is similar to the linear regression modeling procedure:

L = -0.035(diff_rushing) + -0.026(diff_passing) + 0.044(diff_penalty) + 0.735(diff_fumbles) + 0.316(diff_first_downs) + 0.267

		Standard			95% Confidence Interval		Odds	Wald	Wald	95% Confidence Interval	
	Coefficien	Error	Z	P-value	Lower	Upper	Ratio	Chi Squar	P-value	Lower	Upper
Intercept	-0.26664	0.496831	-0.53669	0.591482	-1.2404151	0.707127					
Var 1	0.03526	0.012154	2.90115	0.003718	0.01143908	0.059082	1.035889	8.416672	0.003718	1.011505	1.060862
Var 2	0.026454	0.010511	2.516765	0.011844	0.00585266	0.047056	1.026807	6.334107	0.011844	1.00587	1.048181
Var 3	-0.04369	0.022831	-1.91381	0.055645	-0.0884414	0.001054	0.957247	3.662663	0.055645	0.915357	1.001054
Var 4	-0.73514	0.433948	-1.69407	0.090252	-1.5856635	0.115383	0.479438	2.869881	0.090252	0.204812	1.122303
Var 5	-0.31647	0.150779	-2.09891	0.035825	-0.6119938	-0.02095	0.728716	4.405412	0.035825	0.542269	0.979268
Var 6	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 7	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 8	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 9	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 10	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 11	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 12	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 13	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 14	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 15	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 16	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 17	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 18	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 19	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 20	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 21	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 22	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 23	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 24	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 25	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 26	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 27	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 28	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 29	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Var 30	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A

Log Likelihood:	-18.6741
AIC:	49.34821

Data | Results

READY

9. Select the Data tab at the bottom left.

10. Enter the equation into the next available column. For this example, here is the Excel formula, then CTRL + D to copy the formula down the column: =-0.035*(A16)+-0.026*(B16)+0.044*(C16)+0.735*(D16)+0.316*(E16)+0.267

COUNTIF | =-0.035*(A16)+-0.026*(B16)+0.044*(C16)+0.735*(D16)+0.316*(E16)+0.267

Intro: This application performs logistic regression where the dependent variable is dichotomous (0 or 1) and the independent variables are numeric (continuous or dichotomous). No character or missing (blank) values are allowed.

Instructions:

Step 1: delete the range name called "Mydata" by pressing 'Ctrl' and 'F3' at the same time.
Step 2: enter your own data in place of the sample data below, include your own variable names.
Step 3: make sure your dependent variable is the LAST (right most) column.
Step 4: assign a range name for your data called "Mydata", but DO NOT include any variable names in the range.
Step 5: press the 'Estimate Model' button below and check your output in the 'Results' tab.

Estimate Model

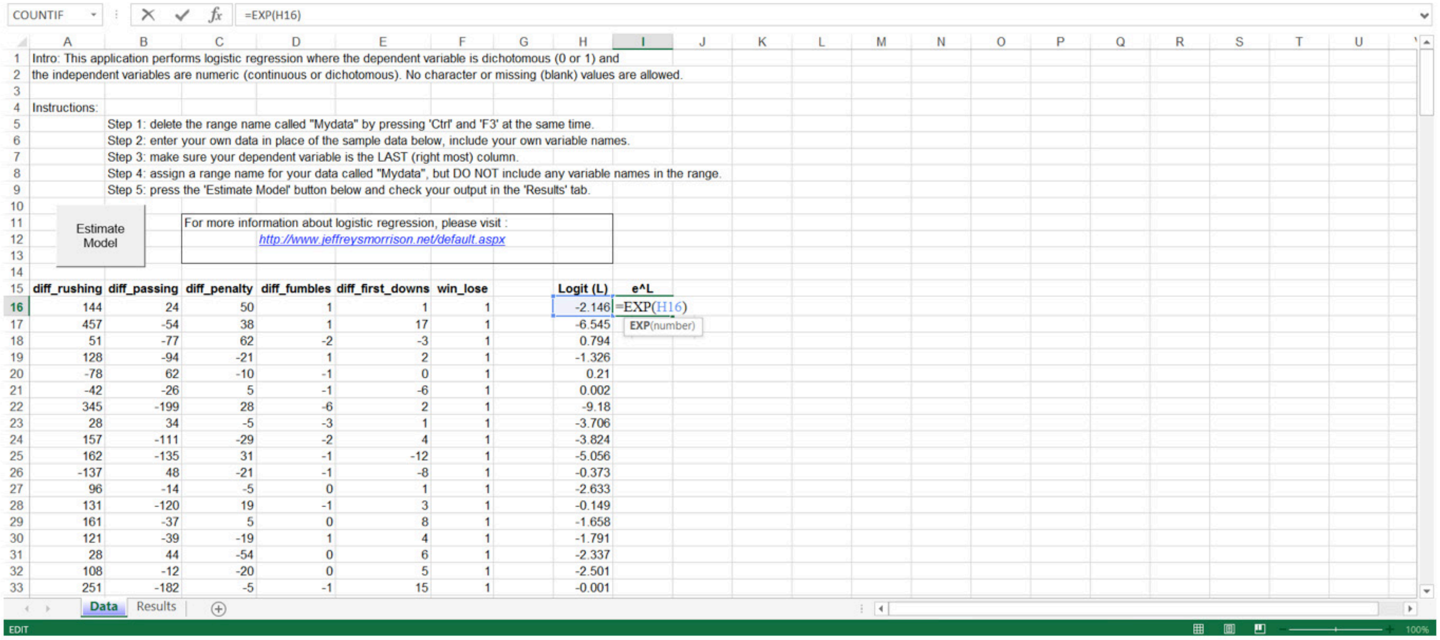
For more information about logistic regression, please visit : http://www.jeffreysmorrison.net/default.aspx

diff_rushing	diff_passing	diff_penalty	diff_fumbles	diff_first_downs	win_lose		Logit (L)
144	24	50	1	1	1		=-0.035*(A
457	-54	38	1	17	1		
51	-77	62	-2	-3	1		
128	-94	-21	1	2	1		
-78	62	-10	-1	0	1		
-42	-26	5	-1	-6	1		
345	-199	28	-6	2	1		
28	34	-5	-3	1	1		
157	-111	-29	-2	4	1		
162	-135	31	-1	-12	1		
-137	48	-21	-1	-8	1		
96	-14	-5	0	1	1		
131	-120	19	-1	3	1		
161	-37	5	0	8	1		
121	-39	-19	1	4	1		
28	44	-54	0	6	1		
108	-12	-20	0	5	1		
251	-182	-5	-1	15	1		

Data | Results

EDIT

11. In the next column, raise e to the L power. For this example, here is the Excel formula, then CTRL + D to copy the formula down the column: =EXP(H16)



COUNTIF | =EXP(H16)

Intro: This application performs logistic regression where the dependent variable is dichotomous (0 or 1) and
the independent variables are numeric (continuous or dichotomous). No character or missing (blank) values are allowed.

Instructions:
Step 1: delete the range name called "Mydata" by pressing 'Ctrl' and 'F3' at the same time.
Step 2: enter your own data in place of the sample data below, include your own variable names.
Step 3: make sure your dependent variable is the LAST (right most) column.
Step 4: assign a range name for your data called "Mydata", but DO NOT include any variable names in the range.
Step 5: press the 'Estimate Model' button below and check your output in the 'Results' tab.

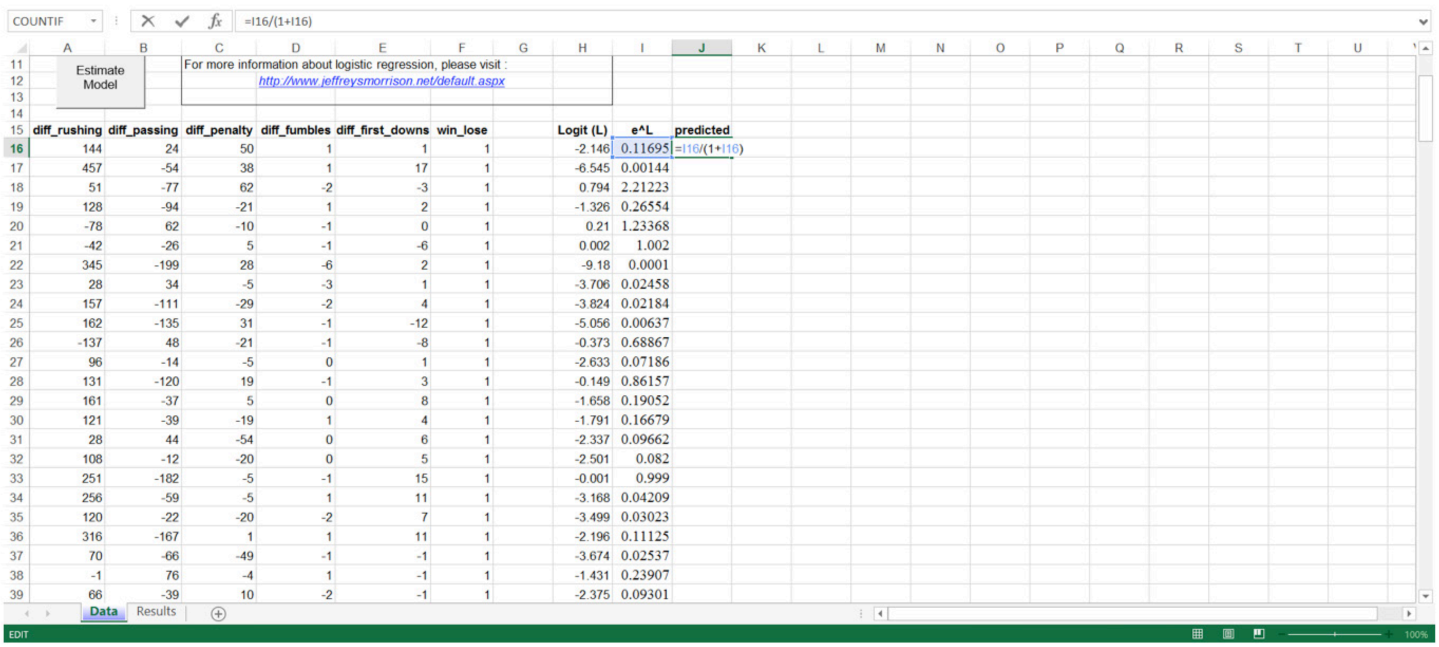
Estimate Model

For more information about logistic regression, please visit :
http://www.jeffreysmorrison.net/default.aspx

	diff_rushing	diff_passing	diff_penalty	diff_fumbles	diff_first_downs	win_lose		Logit (L)	e^L
16	144	24	50	1	1	1		-2.146	=EXP(H16)
17	457	-54	38	1	17	1		-6.545	EXP(number)
18	51	-77	62	-2	-3	1		0.794	
19	128	-94	-21	1	2	1		-1.326	
20	-78	62	-10	-1	0	1		0.21	
21	-42	-26	5	-1	-6	1		0.002	
22	345	-199	28	-6	2	1		-9.18	
23	28	34	-5	-3	1	1		-3.706	
24	157	-111	-29	-2	4	1		-3.824	
25	162	-135	31	-1	-12	1		-5.056	
26	-137	48	-21	-1	-8	1		-0.373	
27	96	-14	-5	0	1	1		-2.633	
28	131	-120	19	-1	3	1		-0.149	
29	161	-37	5	0	8	1		-1.658	
30	121	-39	-19	1	4	1		-1.791	
31	28	44	-54	0	6	1		-2.337	
32	108	-12	-20	0	5	1		-2.501	
33	251	-182	-5	-1	15	1		-0.001	

Data | Results

12. To determine the probability of the event occurring based on the Logit equation, enter L/(1+L) in the next column. For this example, here is the Excel formula, then CTRL + D to copy the formula down the column: =(I16/(1+I16)).



COUNTIF | =I16/(1+I16)

Estimate Model

For more information about logistic regression, please visit :
http://www.jeffreysmorrison.net/default.aspx

	diff_rushing	diff_passing	diff_penalty	diff_fumbles	diff_first_downs	win_lose		Logit (L)	e^L	predicted
16	144	24	50	1	1	1		-2.146	0.11695	=I16/(1+I16)
17	457	-54	38	1	17	1		-6.545	0.00144	
18	51	-77	62	-2	-3	1		0.794	2.21223	
19	128	-94	-21	1	2	1		-1.326	0.26554	
20	-78	62	-10	-1	0	1		0.21	1.23368	
21	-42	-26	5	-1	-6	1		0.002	1.002	
22	345	-199	28	-6	2	1		-9.18	0.0001	
23	28	34	-5	-3	1	1		-3.706	0.02458	
24	157	-111	-29	-2	4	1		-3.824	0.02184	
25	162	-135	31	-1	-12	1		-5.056	0.00637	
26	-137	48	-21	-1	-8	1		-0.373	0.68867	
27	96	-14	-5	0	1	1		-2.633	0.07186	
28	131	-120	19	-1	3	1		-0.149	0.86157	
29	161	-37	5	0	8	1		-1.658	0.19052	
30	121	-39	-19	1	4	1		-1.791	0.16679	
31	28	44	-54	0	6	1		-2.337	0.09662	
32	108	-12	-20	0	5	1		-2.501	0.082	
33	251	-182	-5	-1	15	1		-0.001	0.999	
34	256	-59	-5	1	11	1		-3.168	0.04209	
35	120	-22	-20	-2	7	1		-3.499	0.03023	
36	316	-167	1	1	11	1		-2.196	0.11125	
37	70	-66	-49	-1	-1	1		-3.674	0.02537	
38	-1	76	-4	1	-1	1		-1.431	0.23907	
39	66	-39	10	-2	-1	1		-2.375	0.09301	

Data | Results

13. To predict the dichotomous outcome, 50%, or 0.50, is the cut-off value. If the predicted probability is greater than 0.50, then a loss is predicted (dummy coded as 0). If the predicted probability is less than 0.50, then a win is predicted (dummy coded as 1). I highlighted the values that indicated a predicted outcome other than the observed outcome.

J57 =I57/(1+I57)

	diff_rushing	diff_passing	diff_penalty	diff_fumbles	diff_first_downs	win_lose		Logit (L)	e^L	predicted	predicted_outcome
16	144	24	50	1	1	1		-2.146	0.11695	0.104706	1
17	457	-54	38	1	17	1		-6.545	0.00144	0.001435	1
18	51	-77	62	-2	-3	1		0.794	2.21223	0.68869	0
19	128	-94	-21	1	2	1		-1.326	0.26554	0.209822	1
20	-78	62	-10	-1	0	1		0.21	1.23368	0.552308	0
21	-42	-26	5	-1	-6	1		0.002	1.002	0.5005	0
22	345	-199	28	-6	2	1		-9.18	0.0001	0.000103	1
23	28	34	-5	-3	1	1		-3.706	0.02458	0.023986	1
24	157	-111	-29	-2	4	1		-3.824	0.02184	0.021373	1
25	162	-135	31	-1	-12	1		-5.056	0.00637	0.006331	1
26	-137	48	-21	-1	-8	1		-0.373	0.68867	0.407816	1
27	96	-14	-5	0	1	1		-2.633	0.07186	0.067045	1
28	131	-120	19	-1	3	1		-0.149	0.86157	0.462819	1
29	161	-37	5	0	8	1		-1.658	0.19052	0.160031	1
30	121	-39	-19	1	4	1		-1.791	0.16679	0.14295	1
31	28	44	-54	0	6	1		-2.337	0.09662	0.088105	1
32	108	-12	-20	0	5	1		-2.501	0.082	0.075788	1
33	251	-182	-5	-1	15	1		-0.001	0.999	0.49975	0
34	256	-59	-5	1	11	1		-3.168	0.04209	0.040388	1
35	120	-22	-20	-2	7	1		-3.499	0.03023	0.029341	1
36	316	-167	1	1	11	1		-2.196	0.11125	0.10011	1
37	70	-66	-49	-1	-1	1		-3.674	0.02537	0.024747	1
38	-1	76	-4	1	-1	1		-1.431	0.23907	0.192943	1
39	66	-39	10	-2	-1	1		-2.375	0.09301	0.085099	1
40	254	8	5	-3	11	1		-7.34	0.00065	0.000649	1
41	205	-87	5	0	13	1		-0.318	0.7276	0.421163	1

Data Results

READY AVERAGE: 0.616844104 COUNT: 6 SUM: 3.701064622 100%

14. To make a classification table, set up a 2-by-2 table.

	diff_rushing	diff_passing	diff_penalty	diff_fumbles	diff_first_downs	win_lose		Logit (L)	e^L	predicted	predicted_outcome
16	144	24	50	1	1	1		-2.146	0.11695	0.104706	1
17	457	-54	38	1	17	1		-6.545	0.00144	0.001435	1
18	51	-77	62	-2	-3	1		0.794	2.21223	0.68869	0
19	128	-94	-21	1	2	1		-1.326	0.26554	0.209822	1
20	-78	62	-10	-1	0	1		0.21	1.23368	0.552308	0
21	-42	-26	5	-1	-6	1		0.002	1.002	0.5005	0
22	345	-199	28	-6	2	1		-9.18	0.0001	0.000103	1
23	28	34	-5	-3	1	1		-3.706	0.02458	0.023986	1
24	157	-111	-29	-2	4	1		-3.824	0.02184	0.021373	1
25	162	-135	31	-1	-12	1		-5.056	0.00637	0.006331	1
26	-137	48	-21	-1	-8	1		-0.373	0.68867	0.407816	1
27	96	-14	-5	0	1	1		-2.633	0.07186	0.067045	1
28	131	-120	19	-1	3	1		-0.149	0.86157	0.462819	1
29	161	-37	5	0	8	1		-1.658	0.19052	0.160031	1
30	121	-39	-19	1	4	1		-1.791	0.16679	0.14295	1
31	28	44	-54	0	6	1		-2.337	0.09662	0.088105	1
32	108	-12	-20	0	5	1		-2.501	0.082	0.075788	1
33	251	-182	-5	-1	15	1		-0.001	0.999	0.49975	0
34	256	-59	-5	1	11	1		-3.168	0.04209	0.040388	1
35	120	-22	-20	-2	7	1		-3.499	0.03023	0.029341	1
36	316	-167	1	1	11	1		-2.196	0.11125	0.10011	1
37	70	-66	-49	-1	-1	1		-3.674	0.02537	0.024747	1
38	-1	76	-4	1	-1	1		-1.431	0.23907	0.192943	1
39	66	-39	10	-2	-1	1		-2.375	0.09301	0.085099	1
40	254	8	5	-3	11	1		-7.34	0.00065	0.000649	1
41	205	-87	5	0	13	1		-0.318	0.7276	0.421163	1

	predicted	
observed	win (1)	lose (0)
win (1)		
lose (0)		

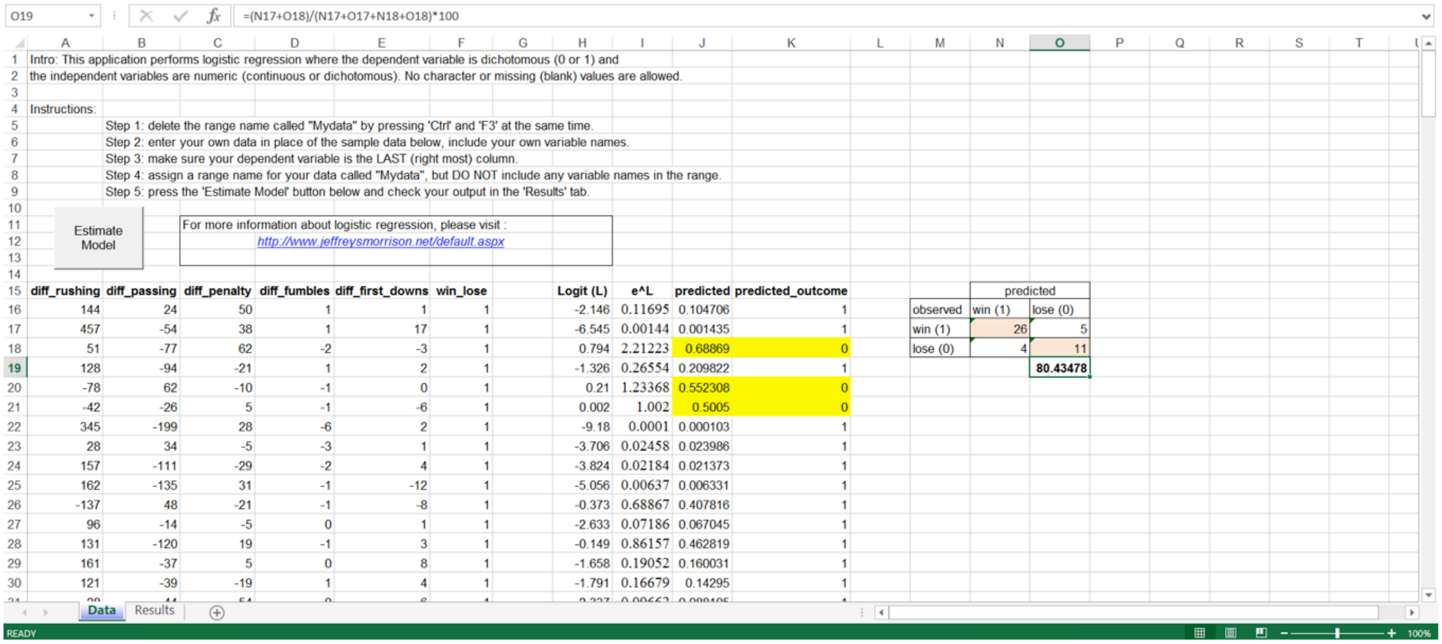
Data Results

READY 100%

15. Count the number of predicted outcomes for each cell. See the following table with Excel formulas for this example: =COUNTIF(range of cells, criteria value)

	Predicted	
Observed	win (1)	lose (0)
win (1)	=COUNTIF(K16:K46,1)	=COUNTIF(K16:K46,0)
lose (0)	=COUNTIF(K47:K61,1)	=COUNTIF(K47:K61,0)

16. To calculate the percentage of correctly predicted outcomes, add the win/win cell and lose/lose cell, divide by the sum of all cells, then multiply by 100. Here is the formula for this example: =(N17+O18)/(N17+O17+N18+O18)*100.



O19 =(N17+O18)/(N17+O17+N18+O18)*100

Intro: This application performs logistic regression where the dependent variable is dichotomous (0 or 1) and the independent variables are numeric (continuous or dichotomous). No character or missing (blank) values are allowed.

Instructions:
Step 1: delete the range name called "Mydata" by pressing 'Ctrl' and 'F3' at the same time.
Step 2: enter your own data in place of the sample data below, include your own variable names.
Step 3: make sure your dependent variable is the LAST (right most) column.
Step 4: assign a range name for your data called "Mydata", but DO NOT include any variable names in the range.
Step 5: press the 'Estimate Model' button below and check your output in the 'Results' tab.

Estimate Model

For more information about logistic regression, please visit : http://www.jeffreysmorrison.net/default.aspx

diff_rushing	diff_passing	diff_penalty	diff_fumbles	diff_first_downs	win_lose	Logit (L)	e^L	predicted	predicted_outcome
144	24	50	1	1	1	-2.146	0.11695	0.104706	1
457	-54	38	1	17	1	-6.545	0.00144	0.001435	1
51	-77	62	-2	-3	1	0.794	2.21223	0.68869	0
128	-94	-21	1	2	1	-1.326	0.26554	0.209822	1
-78	62	-10	-1	0	1	0.21	1.23368	0.552308	0
-42	-26	5	-1	-6	1	0.002	1.002	0.5005	0
345	-199	28	-6	2	1	-9.18	0.0001	0.000103	1
28	34	-5	-3	1	1	-3.706	0.02458	0.023986	1
157	-111	-29	-2	4	1	-3.824	0.02184	0.021373	1
162	-135	31	-1	-12	1	-5.056	0.00637	0.006331	1
-137	48	-21	-1	-8	1	-0.373	0.68867	0.407816	1
96	-14	-5	0	1	1	-2.633	0.07186	0.067045	1
131	-120	19	-1	3	1	-0.149	0.86157	0.462819	1
161	-37	5	0	8	1	-1.658	0.19052	0.160031	1
121	-39	-19	1	4	1	-1.791	0.16679	0.14295	1

	predicted	
observed	win (1)	lose (0)
win (1)	26	5
lose (0)	4	11
		80.43478

Data Results

READY

Assessment

Students will conduct another study using pre-existing data where logistic regression could be used to analyze the data and answer the research question. The students would need to collect the data, enter it into Excel, analyze it using logistic regression, and interpret the results.

Our Question

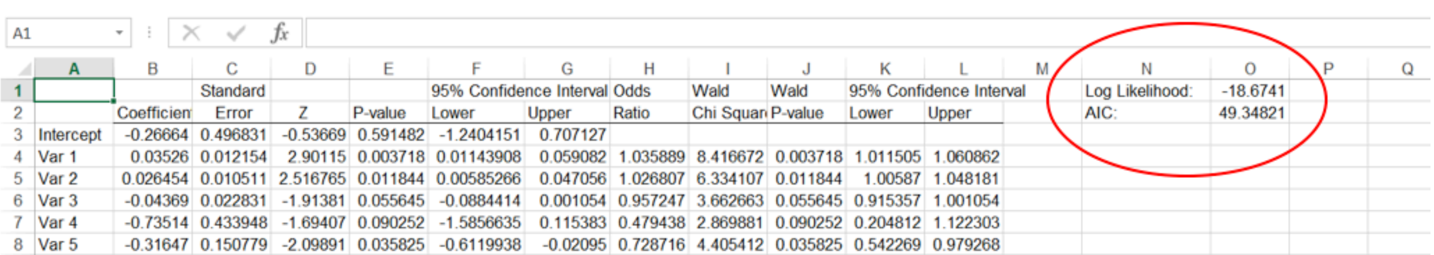
Are the same variables used to determine the probability of winning a football game with college and professional teams applicable to high-school football teams?

For college football games, Wagner (1985) found passing and rushing yards increased the margin of victory. Moreover, rushing yards increased the margin of victory more than passing yards, but turnovers decreased the margin of victory. Willoughby (2002) examined Canadian Football League (CFL) games, and his results suggested the better teams in the CFL were more likely to win the game based on rushing yardage, passing yardage, and the number of interceptions, which was parallel to the findings of Wagner (1985).

Results Based on the Excel Output

After conducting the logistic regression analysis, the resulting Excel output can be interpreted as follows to answer our question.

■ The Log Likelihood value is -18.67, and Akaike Information Criterion (AIC), a measure of whether the model is a good fit for the data, is 49.35. Based on the chi square distribution table, with 5 degrees of freedom at the .05 criteria, any value greater than 11.070 would be considered significant, meaning the change that resulted from adding all five predicting variables was not due to chance.



	Coefficien	Standard Error	Z	P-value	95% Confidence Interval Lower	Upper	Odds Ratio	Wald Chi Squar	Wald P-value	95% Confidence Interval Lower	Upper
Intercept	-0.26664	0.496831	-0.53669	0.591482	-1.2404151	0.707127					
Var 1	0.03526	0.012154	2.90115	0.003718	0.01143908	0.059082	1.035889	8.416672	0.003718	1.011505	1.060862
Var 2	0.026454	0.010511	2.516765	0.011844	0.00585266	0.047056	1.026807	6.334107	0.011844	1.00587	1.048181
Var 3	-0.04369	0.022831	-1.91381	0.055645	-0.0884414	0.001054	0.957247	3.662663	0.055645	0.915357	1.001054
Var 4	-0.73514	0.433948	-1.69407	0.090252	-1.5856635	0.115383	0.479438	2.869881	0.090252	0.204812	1.122303
Var 5	-0.31647	0.150779	-2.09891	0.035825	-0.6119938	-0.02095	0.728716	4.405412	0.035825	0.542269	0.979268

Log Likelihood:	-18.6741
AIC:	49.34821

■ Overall predicting percentage for the logistic regression model was 80.4%, which was a 30% improvement above the 50/50 chance of the null model. From the output, you can see five games were predicted as a loss, but they were won. Four games were lost, but were predicted as a win.

The most significant predicting variables ($p < .05$) were the difference in total rushing yards, the difference in total passing yards, and the difference in the number of first downs. The Wald statistic determines if the predicting variable made a significant contribution. (Note: Var 1 = diff_rushing; Var 2 = diff_passing; Var 3 = diff_penalty; Var 4 = diff_fumbles; Var 5 = diff_first_downs.)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1			Standard			95% Confidence Interval		Odds	Wald	Wald	95% Confidence Interval			Log Likelihood:	-18.6741
2		Coefficien	Error	Z	P-value	Lower	Upper	Ratio	Chi Squar	P-value	Lower	Upper		AIC:	49.34821
3	Intercept	-0.26664	0.496831	-0.53669	0.591482	-1.2404151	0.707127								
4	Var 1	0.03526	0.012154	2.90115	0.003718	0.01143908	0.059082	1.035889	8.416672	0.003718	1.011505	1.060862			
5	Var 2	0.026454	0.010511	2.516765	0.011844	0.00585266	0.047056	1.026807	6.334107	0.011844	1.00587	1.048181			
6	Var 3	-0.04369	0.022831	-1.91381	0.055645	-0.0884414	0.001054	0.957247	3.662663	0.055645	0.915357	1.001054			
7	Var 4	-0.73514	0.433948	-1.69407	0.090252	-1.5856635	0.115383	0.479438	2.869881	0.090252	0.204812	1.122303			
8	Var 5	-0.31647	0.150779	-2.09891	0.035825	-0.6119938	-0.02095	0.728716	4.405412	0.035825	0.542269	0.979268			
9	Var 6	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A			
10	Var 7	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A			

■ Based on the results of this study, the likelihood of winning a high-school football game depends on increased rushing yardage and decreased passing yardage. The results of this study suggest the predictor variables used in determining the probability of winning college and professional games (Wagner, 1985; Willoughby, 2002) are applicable to high-school football games.

References

Franklin, C., G. Kader, D. Mewborn, J. Moreno, R. Peck, M. Perry, and R. Scheaffer. 2007. *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K–12 curriculum framework*. Alexandria, VA: American Statistical Association.

Tansey, R., M. White, R. G. Long, and M. Smith. 1996. A comparison of loglinear modeling and logistic regression in management research. *Journal of Management, 22*(2):339–358.

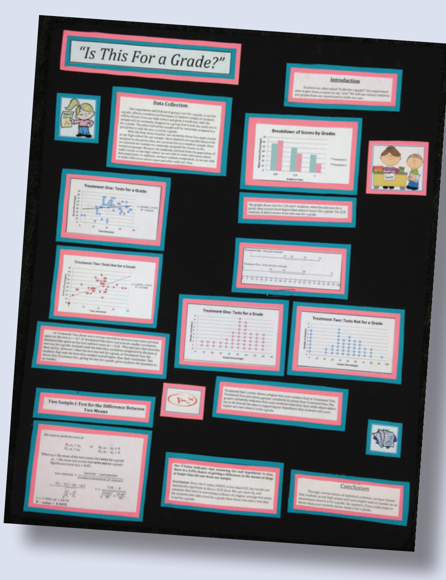
Wagner, G. O. 1985. College and professional football scores: A multiple regression analysis. *American Economist, 31*(1):33–37.

Willoughby, K. A. 2002. Winning games in Canadian football: A logistic regression analysis. *The College Mathematics Journal, 33*(3):215–220.

ANNOUNCEMENTS

ASA 2014 Poster and Project Winners Announced

The ASA is pleased to announce the winners of the 2014 poster and project competitions at *http://magazine.amstat.org/blog/2014/08/01/poster-winners*. The competitions offer opportunities for students to formulate questions and collect, analyze and draw conclusions from data. Winners were recognized with plaques, cash prizes, certificates and calculators, and their names were published in Amstat News. To view the winning posters and projects or for more information, visit *www.amstat.org/education/posterprojects*.



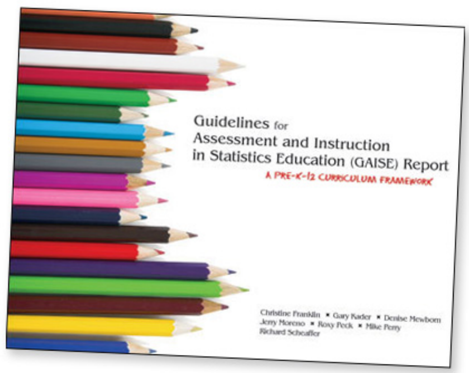
2015 Poster and Project Competitions

Introduce your K–12 students to statistics through the annual poster and project competitions directed by the ASA/NCTM Joint Committee on the Curriculum in Statistics and Probability. The competitions offer opportunities for students to formulate questions and collect, analyze and draw conclusions from data. Winners will be recognized with plaques, cash prizes, certificates and calculators, and their names will be published in Amstat News. Posters (grades K-12) are due every year on April 1. Projects (grades 7-12) are due on June 1. For more information, visit *www.amstat.org/education/posterprojects*.

Lessons from the LOCUS Assessments (Part 2): Center, Spread, and Informal Inference

Catherine Case – University of Florida
Tim Jacobbe – University of Florida

Since its publication, the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K–12 Curriculum Framework* (Franklin et al., 2007) has been influential in statistics education. The developmental levels—A, B, and C—through which students are hypothesized to progress provide convenient touchstones for curriculum and lesson design.



Despite the impressive contributions to statistics education in terms of instructional recommendations and the aforementioned developmental progression, the GAISE framework says less about what types of assessments are recommended or should be considered as a model. The NSF-funded Levels of Conceptual Understanding in Statistics (LOCUS) project focused on developing statistical assessments in the spirit of the GAISE framework. These assessments emphasize conceptual (rather than procedural) understanding and can be used to classify students as having understanding at level A, B, or C.

The assessments consist of four forms: a pre- and post-test targeting the A and B levels and a pre- and post-test targeting the B and C levels. The A/B assessment was designed for students in grades 6–9, and the B/C assessment was designed for students in grades 10–12. Two versions of these are available: one with 23 multiple choice items and five free response items and another with 30 multiple choice items only.

The items from which these forms were constructed were piloted in spring of 2013 with a total of 2,075 students for the A/B assessment and 1,249 students for the B/C assessment. (Although not every item was piloted with every student, each item was piloted with several hundred students.) While the pilot administration was large and included students of many backgrounds and ability levels, it was not selected to be a representative sample of students in the United States. We do report some overall performance indicators, but we caution these should not be over-interpreted. Rather, the indicators are included to paint a more complete picture of the students and item.

Student work can be a valuable resource for teachers. The size and scope of the LOCUS pilot assessments yielded considerable variation in student responses. While there were some 'textbook' correct answers, students also were able to demonstrate correct statistical reasoning in imaginative ways. Incorrect answers often illustrated specific misunderstandings and, if identified as such, can suggest areas for more attention.

The LOCUS item being examined here is shown in Figure 1. This item addresses the following Common Core State Standards (CCSS):

- *6.SP.5 Summarize and describe distributions*
- *7.SP.3 Draw informal comparative inferences about two populations*
- *S-ID.2 Use statistics to compare center and spread*
- *S-ID.3 Interpret differences in shape, center, and spread in context*

The "Analyze Data" component of the GAISE framework at Levels A and B also is addressed by this item.

Student Responses

This free response item was piloted with a total of 523 students. Of those, 505 were students in grades 6–8, and the remaining 18 were students in grades 9–12 who took both the A/B and B/C forms of the assessment. Free response items were scored out of 4 points, with a 4 indicating a "complete" response (allowing for mistakes such as minor computational errors that are not indicative of a misunderstanding). For each item, a small team of graders established a rubric and conducted initial item scoring verbally as a group. Once every grader felt comfortable applying the rubric, scoring continued individually. Any discrepancies or questions were brought to the group's attention. The distribution of students' scores for the item is given in Table 1: 0.2% earned a 4, 0.4% earned a 3, 2.3% earned a 2, 7.1% earned a 1, and the remaining 90.0% earned either a 0 or did not provide a response. Although the scores on this item were very low, the concepts it assesses are explicitly mentioned in the CCSS for middle-school students and in levels A and B in the GAISE Framework. The responses given provide useful illustrations of students' varying conceptions.

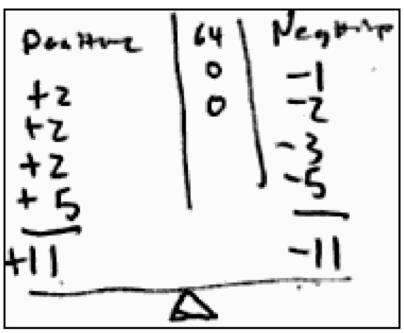
Table 1—The Distribution of Student Scores for the Item

Score	Percent of Students
4	0.2%
3	0.4%
2	2.3%
1	7.1%
0 or no response	90.0%

Questions Related to the Free-Response Item in Figure 1 (Next Page)

Explain why the mean height of the plants in the no-compost bed is 59 inches without performing any calculations.

A student noted that the sum of the positive deviations from the mean for plants grown in the no-compost bed was +11 inches, which balances the sum of the negative deviations from the mean, which was -11 inches. Show that the positive and negative deviations for the compost bed plants are also balanced about the mean of 64 inches.



These two parts ask students to provide a conceptual explanation of the mean without relying on the standard computational algorithm. Student responses to these questions are informative, because Common Core (S-ID.3) requires students to interpret measures of center in context and Mokros and Russell (1995) have suggested students who only understand average as algorithm often are unable to make contextual connections. Thus, Mokros and Russell recommend instruction focused on the development of richer conceptions of average. The GAISE Framework specifically mentions illustrating the mean as "balance point" to promote conceptual understanding of the mean as a measure of central location (Franklin, et al., 2007, pp. 41–43).

Most students were unable to provide a complete explanation for the mean height of the plants in the no-compost bed without performing the familiar calculations. Some responses included the algorithm for computing the mean in spite of the instructions. Other responses indicated lower-level conceptions of the mean. For example, many students' responses demonstrated modal thinking, a common early conception of average: "There are two points of data on the 59, while there is only one on all other points." Other students described the mean as the midpoint or midrange value: "It is right in the center, with an equal range on either side." Because of the perfect symmetry of the "no-compost" distribution, students could make true statements about why the mean is 59 that demonstrate their current conceptions of mean. To score full marks, a student had to attend to this symmetry in some way: "The mean of the no-compost bed is 59 inches because the ones that are shorter than 59 and the ones that are taller than 59 are in the same position or symmetrical. So we know 59 inches is the mean because it is right in the middle of the data."

When asked to show that the mean of 64 inches in the compost bed was the balance point, many students were unable to follow the directions and gave responses similar to those given for part (a). For example, "54 to 69 with a mean of 64 inches" was a response given to part (b) that again seems to refer to the midrange. Note that some explanations that were acceptable in part (a) are not appropriate interpretations for the mean of an asymmetrical distribution. However, some students did successfully find the balance point by summing the positive and negative deviations, and of those, several provided interesting representations of that process.

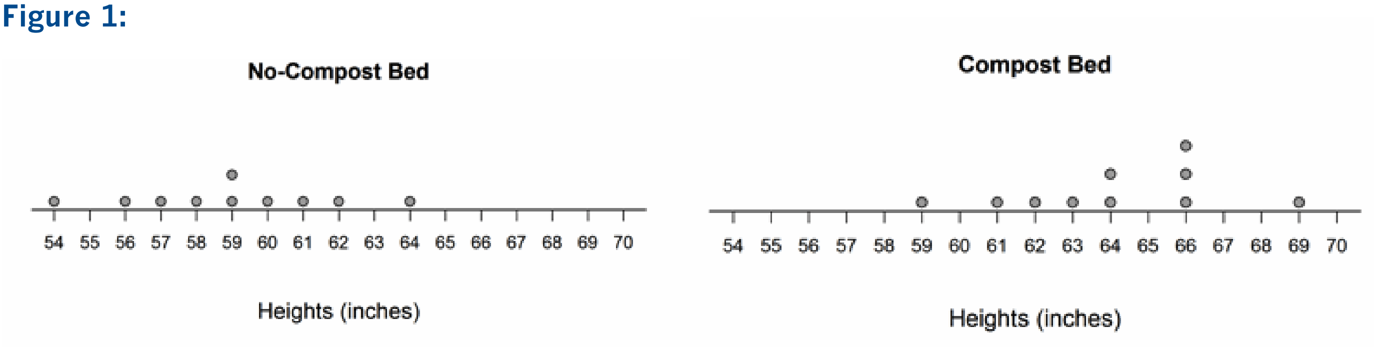
Question Related to the Free-Response Item

The MAD (mean absolute deviation) for each group is 2.2 inches. What does this value tell us about the plant heights in the no-compost bed?

This question asks the students to interpret a measure of spread in the context of the problem. Variability is a central concern of statistics, so it is important for students to be able to recognize and interpret measures of variability, even at early levels. Because the average distance of each value from the mean is a relatively intuitive measure of spread, the GAISE Framework recommends using the MAD "as a precursor to the standard deviation" (p. 44). The CCSS mentions using the MAD as a measure of spread beginning in grade 6 (6.SP.5c).

Many students provided responses that demonstrated a lack of understanding of the MAD as a measure of spread. For example, one student said that both groups having a MAD of 2.2 means the plants in the no-compost bed "grew at a similar rate." However, a few students did provide complete interpretations of the MAD: "The variation is the same as the compost bed, and on average, every point is 2.2 away from the mean." Others gave less precise interpretations that still demonstrated understanding of the MAD as a measure of spread: "Their data is spread out just as much as the compost bed data."

Figure 1:



Middle-school science students wanted to compare the effect of compost on the plant height of their favorite flower, the dahlia. They had two similar planting areas (called beds), except one was treated with compost and the other was not. In designing their experiment, they tried to control as many variables as they could so any observed height difference in the plants would be due to the no-compost/compost treatments and not something else. Of 20 similar dahlia plants of the same variety and height, they randomly assigned 10 to plant in the no-compost bed and 10 in the compost bed. The heights (inches) of the plants after eight weeks are displayed in the dotplots above.

(a) Explain why the mean height of the plants in the no-compost bed is 59 inches without performing any calculations.

(b) A student noted that the sum of the positive deviations from the mean for plants grown in the no-compost bed was +11 inches, which balances the sum of the negative deviations from the mean, which was -11 inches.

Show that the positive and negative deviations for the compost bed plants also are balanced about the mean of 64 inches.

(c) The MAD (mean absolute deviation) for each group is 2.2 inches. What does this value tell us about the plant heights in the no-compost bed?

(d) Note that the mean height of plants grown in the compost bed exceeded the mean height of plants grown in the no-compost bed by 5 inches. In statistics, we often describe a difference in means in terms of how many measures of spread, such as the MAD, separate them. If two means differ by more than two MADs, then some say the two means are "very different" from each other.

i. How many MADs separate the mean height of the plants grown in the no-compost bed from the mean height of the plants grown in the compost?

ii. Is the mean height of the plants grown in the compost bed "very different" from those grown in the no-compost bed? Why or why not?

Question Related to the Free-Response Item

Note that the mean height of plants grown in the compost bed exceeded the mean height of plants grown in the no-compost bed by 5 inches. In statistics, we often describe a difference in means in terms of how many measures of spread, such as the MAD, separate them. If two means differ by more than two MADs, then some say the two means are "very different" from each other.

How many MADs separate the mean height of the plants grown in the no-compost bed from the mean height of the plants grown in the compost bed?

Is the mean height of the plants grown in the compost bed "very different" from those grown in the no-compost bed? Why or why not?

This question asks students to draw informal inferences about an observed difference in plant heights taking into account the amount of variability in each bed. This type of informal inference is clearly spelled out in CCSS-M 7.SP.3, which requires students to assess the overlap of two distributions, "measuring the difference between the centers by expressing it as a multiple of a measure of variability." This reasoning is a precursor to more formal inference methods that use test statistics to summarize sample differences in light of variability.

Although the difference of means and the MAD for each group was given, many students did not successfully calculate the number of MADs separating the mean heights for each group. Some students confused the MAD with an ordinary unit of length, specifically inches, and said the two means were 5 MADs apart. Other students divided the number of MADs by 2.2, but didn't complete the arithmetic to reach a fraction or decimal value; for example, one student showed division work and concluded that "the means are separated by a little more than 2 MADs." Because the LOCUS assessments are intended to measure conceptual understanding, not computational skills, these responses with minor computational errors could still receive full credit.

Part (ii) required students to compare the number of MADs to the benchmark of 2 and indicate that the means are "very different." Some students were able to complete this part, even if they had an incorrect answer in part (i): "Yes, because means that are even 2 MADs from each other are considered very different and the No-Compost and Compost Bed's were 5 MADs apart."

Note that this question includes scaffolding to help students informally assess the difference in two distributions by expressing the difference in the mean heights as a multiple of a measure of variability. This type of reasoning, expected under CCSS, may not come naturally. Other LOCUS free-response questions simply asked students to consider the amount of variability to decide whether a difference was meaningful, and very few students took this approach. Since this type of reasoning is expected, but not necessarily intuitive, it must be scaffolded during instructional interactions and included in formative assessment.

Discussion

This item broadly targets the "Analyze Data" component of the GAISE Framework at levels A and B and specifically targets several CCSS standards (6.SP.5, 7.SP.3, S-ID.2, S-ID.3). For teachers who want to focus instruction on the content covered in this item, there are several resources available. The video series Learning Math: Data Analysis, Statistics, and Probability (WGBH, 2001) is intended to introduce K–8 teachers to statistical concepts through classroom case studies. In particular, video 5, "Variation about the Mean," explores various conceptions of the mean and introduces the MAD as a measure of variability. Also, several lesson plans on the STEW website (*www.amstat.org/education/stew*) address the key aspect of comparing two groups using graphical displays. For example, "Bubble Trouble!" and "Don't Spill the Beans!" both use boxplots to compare distributions of data collected during in-class experiments. These lessons could easily be extended to cover the informal inferential reasoning addressed in this item by prompting students to describe the difference between centers as a multiple of a measure of variability. The lessons also could be modified to incorporate other graphical representations (e.g., dotplots, histograms) and numerical summaries (e.g., mean, MAD) of quantitative data.

STatistics Education Web
STEW
online journal of K-12 statistics lesson plans

Because the pilot forms of the LOCUS assessments were administered before full implementation of Common Core, it is likely students were not taught the concepts of center, spread, and informal inference at the level of rigor required by the new standards. The scores on this piloted item suggest these concepts are widely misunderstood, so it is critical that instruction focused on richer conceptions of mean, variability, and comparative inference become a part of the enacted curriculum. The student work presented here illustrates students' various conceptions and serves as a resource to inform instruction.

References

Franklin, C. A., G. Kader, D. Mewborn, J. Moreno, R. Peck, M. Perry, and R. Scheaffer. 2007. *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K–12 curriculum framework*. Alexandria, VA: American Statistical Association.

Mokros, J., and S. J. Russell. 1995. Children's concepts of average and representativeness. *Journal for Research in Mathematics Education* 26:20–39.

National Governors Association Center for Best Practices and Council of Chief State School Officers. 2010. *Common core state standards for mathematics*. National Governors Association Center for Best Practices and Council of Chief State School Officers: Washington, DC.

WGBH Educational Foundation (Producer). 2001. *Learning math: Data analysis, statistics, and probability*. Available from *www.learner.org/resources/series158.html*.

Featured STEW Lesson Plan
For this and other free, peer-reviewed lessons, please visit www.amstat.org/education/stew.

Additional resources accompanying this lesson also are posted.

How Wet Is the Earth?

Laura Ring Kapitula – Grand Valley State University
Paul Stephenson – Grand Valley State University

In this activity, random sampling is used to estimate the proportion of the Earth's surface covered with water. Students use an Internet site to select random points on the Earth's surface and to see them on a map. After selecting their sample of points, the students record whether each point is on water. Each student then uses their data to calculate the sample proportion of points that are on water and compute a confidence interval for the proportion of the Earth's surface covered by water.

After each student or student group finishes their calculations, the class data can be used to illustrate the sampling distribution of the sample proportion and the long-term behavior of confidence intervals. This lesson also can be adapted for use with middle-school students if the confidence intervals are discussed briefly or in a simplified manner.

GAISE Components

This activity follows all four components of statistical problemsolving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. The four components are: formulate a question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level C activity.

Common Core State Standards for Mathematical Practice

2. Reason abstractly and quantitatively

4. Model with mathematics

5. Use appropriate tools strategically

8. Look for and express regularity in repeated reasoning

Common Core State Standards Grade-Level Content (High School)

S-IC. 1. Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

S-IC. 4. Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling.

Common Core State Standards Grade-Level Content (Grades 6 and 7)

6. SP. 1. Recognize a statistical question as one that anticipates variability in the data related to the question and accounts for it in the answers.

6. SP. 4. Display numerical data in plots on a number line, including dot plots, histograms, and box plots.

6. SP. 5. Summarize numerical data sets in relation to their context.

7. SP. 1. Understand that statistics can be used to gain information about a population by examining a sample of the population; generalizations about a population from a sample are valid only if the sample is representative of that population. Understand that random sampling tends to produce representative samples and support valid inferences.

7. SP. 2. Use data from a random sample to draw inferences about a population with an unknown characteristic of interest. Generate multiple samples (or simulated samples) of the same size to gauge the variation in estimates or predictions.

7. SP. 5. Understand that the probability of a chance event is a number between 0 and 1 that expresses the likelihood of the event occurring. Larger numbers indicate greater likelihood. A probability near 0 indicates an unlikely event, a probability around 1/2 indicates an event that is neither unlikely nor likely, and a probability near 1 indicates a likely event.

7. SP. 6. Approximate the probability of a chance event by collecting data on the chance process that produces it and observing its long-run relative frequency, and predict the approximate relative frequency given the probability.

NCTM Principles and Standards for School Mathematics

Data Analysis and Probability Standards for Grades 9–2

Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them:

- Know the characteristics of well-designed studies, including the role of randomization in surveys and experiments

- Understand the meaning of measurement data and categorical data, of univariate and bivariate data, and of the term variable

• Understand histograms, parallel box plots, and scatterplots and use them to display data

• Compute basic statistics and understand the distinction between a statistic and a parameter

Develop and evaluate inferences and predictions that are based on data:

• Use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions

• Understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference

Understand and apply basic concepts of probability:

• Use simulations to construct empirical probability distributions

Prerequisites

Students should have some knowledge of sampling and estimating unknown parameters.

Learning Targets

Upon completion of the activity, students will be able to:

• Randomly sample from a population

• Estimate the proportion in a population with a certain characteristic

• Understand variability in estimated proportions

• Compute a confidence interval for a population proportion

Time Required

This activity can be completed in one 50-minute class period.

Materials Required

Each student or student group needs to have a computer connected to the Internet with a word-processing program such as Microsoft Word. Having a printer in the classroom makes it a bit easier for students, but it is not necessary. Each student will need a copy of the worksheet given at the end of the lesson. Alternatively, each student or student group needs to have an iPad or other tablet connected to the Internet.

Instructional Lesson Plan

The GAISE Statistical Problemsolving Procedure

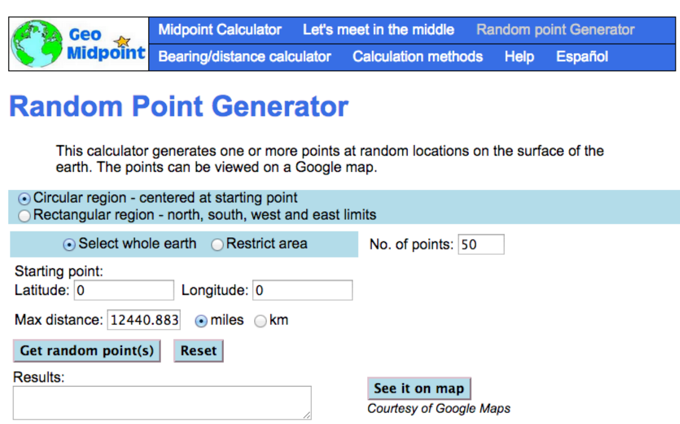
I. Formulate Question(s)

What proportion of the Earth's surface is covered by water? The true proportion is what we are estimating, and we will use sample data to estimate this proportion. How will the sample proportions vary? What proportion of 90% confidence intervals calculated would we expect to contain the true proportion? Note that the true proportion is about 71% (see *http://ga.water.usgs.gov/edu/earthhowmuch.html*).

II. Design and Implement a Plan to Collect the Data

Students should be instructed to go to *www.geomidpoint.com/random* and select circular region and whole earth (see Figure 1). It may be a good idea to start by having students randomly generate a point on Earth and click to see it on a map. The map will come up centered on a blue pin. Students should be instructed to ignore the blue pin and look for the red pin. Then they should determine whether the red pin is on water (they may need to zoom in). [Note: This single point is generated to illustrate the process employed to collect data, and this point will not be used for any future calculations.] Next, have students randomly generate 50 points on Earth and click to see them on a map (see Figure 2).

Figure 1. Generating 50 random points



Remind students to be careful to not count points more than once. Given the rectangular nature of the map they may have to zoom in to make sure each pin is on the map only once. Students can then use the Snipping Tool to copy the map with the 50 points.

For example:

Figure 2. Example of snipped image of map for printing

After the map is generated and snipped, the students should open Word (or some other word processor) and paste the map into a Word file. They then can print the Word file and use the webpage and zooming to determine how many of the 50 points are on water, and circle the pins that are on the water. Once they do this, they should record the number of pins on water and on land on their worksheet.

III. Analyze the Data

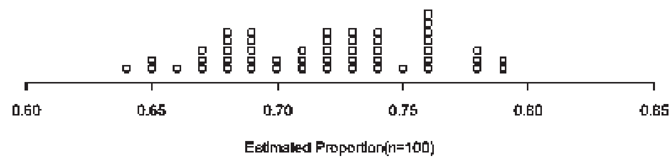
Students are then asked to pair with a neighbor and combine results to create a sample size of 100. Using the combined data and the following formulas, students calculate a sample proportion of points on water and a 90% confidence interval for the proportion of the Earth covered with water:

$n =$ total number in sample, $\hat{p} = \dfrac{\text{number in sample on water}}{n}$,

and the corresponding 90% CI equals: $\hat{p} \pm 1.645\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$, given the large sample method is appropriate in this case. Students could check the appropriateness of the large sample method by confirming that $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$. It would be extremely unlikely to get a sample that would violate these conditions if a sample size of 100 is used, given the true proportion is around 71%.

Once students have an estimated proportion, they can share their proportion with the class. A dot plot or stem-and-leaf plot with the sample proportions can be drawn on the board. Typically, we allow students to come up to the board and write in their value on a plot rounded to the nearest hundredth. The plot of the sample proportions is an empirical sampling distribution for the sample proportion. See Figure 3.

Figure 3. Example dot plot of student's estimated proportions for 50 samples of size 100.

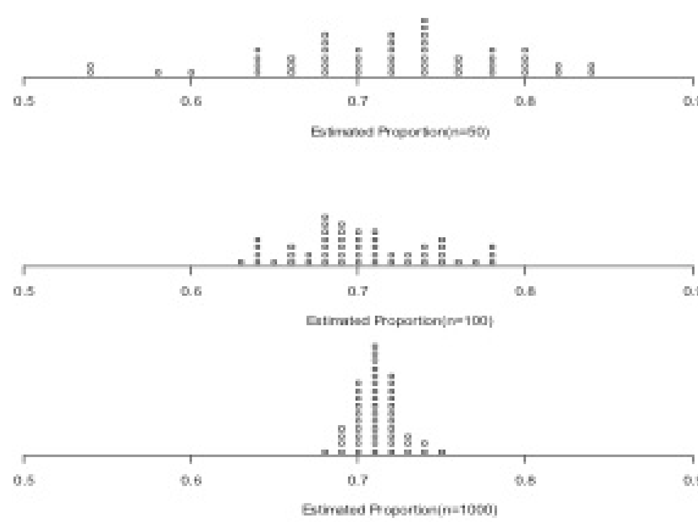


After students generate their confidence intervals, ask the class groups whether their confidence interval contains the true of value 0.71. Tally the results and see what proportion of the 90% confidence intervals computed by the class contain the true value.

IV. Interpret the Results

Ask students how much spread they see in the empirical sampling distribution. Ask them why not everyone got the true value of 71% and ask them about the shape and center of the distribution of sample proportions. In the dot plot in Figure 3, the values are centered at about 71% and look fairly symmetric with no outliers. Ask them how they would expect the sampling distribution to change if instead of everyone in the class having a sample of 100, they had a sample size of 50, or a sample size of 1,000. Simulated data can be used to illustrate this concept or the instructor can show the plot given in Figure 4. R-code to simulate data and make dot-plots and simulated data are attached to this lesson.

Figure 4. Simulated sampling distributions for the sample proportion.



If time is available, the instructor can calculate a confidence interval using data from the whole class and talk about how the size of the margin of error is smaller when the sample size is larger.

Assessment

1. Suppose we did the activity again, but had each student look at 500 points, instead of 50, resulting in a sample size of 1,000 when they combined their data with their partner. How would this change affect the distribution of the sample proportions? How would this change affect the 90% confidence interval?

Answer: Increasing the sample size would result in the sample proportions being less variable. They would be in general closer to 71% but still centered at 71%. They would be more bell shaped as well, but students may not get that from this lesson since, given the true population proportion is moderate, the sampling distribution should be fairly bell shaped with an $n = 100$. Increasing the sample size would cause the confidence intervals to be less wide, but they still would have a success rate of 90%, since the success rate is fixed by the method.

2. Suppose a student obtained 75 points out of 100 on the water. Use these data to calculate a sample proportion and both a 90% and a 95% confidence interval for the proportion of water on the surface of the Earth. Which interval is wider? Will this always be the case?

Answer: $\hat{p} = 0.75$, the number of failures is 25, and the number of successes is 75 so it is reasonable to use the large sample method to calculate the confidence interval for the proportion.

The 90% confidence interval is $\hat{p} \pm 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = (0.679, 0.821)$.

And the 95% confidence interval is $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = (0.665, 0.835)$.

Students should notice that the 95% confidence interval is wider

than the 90% confidence interval, and looking at the formulas,

they should notice that this will always be true.

Possible Extensions

- What proportion of the surface area of the southern or northern hemisphere is covered with water?
- Is the proportion of water in the southern and northern hemispheres different?
- The STEW activity "What Percent of the Continental U.S. Is Within One Mile of a Road?" by Stoudt, Cao, Udwin, and Horton is an excellent extension or alternative to this activity for students who are using R in class. We developed this activity independently before reading Stoudt, et. al., but there are many similarities with the Stoudt, et. al. activity.
- This activity could tie in with learning to calculate the surface area of a sphere and using that value to estimate the surface area of land mass on the Earth's surface.

Possible Simplifications

- Do not estimate confidence intervals and only use the activity to illustrate the sampling distribution for a proportion.
- Elementary- or middle-school students could get into groups, look at 100 points, and use the information to calculate and understand the idea of a percentage.
- Younger children could sample points and learn to make a pie chart or bar chart using their sample or a larger sample for the whole class.

Acknowledgements

We wish to thank Randall Pruim who inspired this lesson and implemented the "rgeo()" and "googleMap()" functions in the mosaic package for the software package R. Although we did not use R in designing our lesson, his work and a personal discussion inspired the activity. If you wish to implement this activity in R, modifications can be made to the STEW activity "What Percent of the Continental U.S. Is Within One Mile of a Road?" by Stoudt, Cao, Udwin, and Horton.

References

Franklin, C. A., G. Kader, D. Mewborn, J. Moreno, R. Peck, M. Perry, and R. Scheaffer. 2007. *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K–12 curriculum framework*. Alexandria, VA: American Statistical Association.

National Governors Association Center for Best Practices and Council of Chief State School Officers. 2010. *Common core state standards for mathematics*. National Governors Association Center for Best Practices and Council of Chief State School Officers: Washington, DC.

National Council of Teachers of Mathematics. 2000. *Principles and standards for school mathematics*. Reston, VA: NCTM.

Stoudt, Cao, Udwin, and Horton. 2014. What percent of the continental U.S. is within one mile of a road? *www.amstat.org/education/stew*.

Water drop image taken from *http://naturesan.com/images/earth_water_drop.png*.

How Wet Is the Earth?

Activity Sheet

We will take random samples of points on the Earth's surface and see if we can use them to learn about estimating the proportion of the Earth's surface that is water.

What numeric value (or parameter) are we estimating?

Go to *www.geomidpoint.com/random* and select circular region and whole Earth.

Now randomly generate a point on Earth and click to see it on a map. The map will come up centered on a blue pin. Ignore that blue pin and look for the red pin. Determine whether the pin is on water (you may zoom in if you need to).

Now randomly generate 50 points on Earth and click to see the points on a map.

Use the Snipping Tool to copy the map with the 50 points.

Open Word and paste the map into a Word file.

Print the Word file with the map.

Go back to the web page and determine how many of the 50 points are on water; zoom in if you need to.

Circle pins on the map that are in the water.

Record the number of pins on water and on land.

On Water	On Land

Pair up with one of your neighbors.

Combine the results with your neighbor.

On Water	On Land

Calculate the sample proportion of points on water for these combined samples.

Compute a 90% confidence interval for the proportion of points on water. This is an estimate for the proportion of the Earth covered by water.

As you may or may not know, about 71% of the Earth's surface is covered by water. See *http://ga.water.usgs.gov/edu/earthhowmuch.html* for more information.

ANNOUNCEMENTS

ASA PR Campaign Launched

The ASA public relations campaign—designed to increase interest in the field of statistics and recruit the next generation of statisticians—has launched. The "This Is Statistics" campaign (*www.thisisstatistics.org*) will educate students, parents, teachers, and counselors about the abundant and well-paying career opportunities in statistics. The informational website features videos of "cool" statisticians doing interesting work, career and salary information, sectors that employ statisticians, and an interactive map of select organizations across the world at which statisticians work. We encourage teachers to share the website link with students and colleagues, follow the latest news on Twitter (*https://twitter.com/ThisisStats*), find additional information on Facebook (*www.facebook.com/ThisIsStats*), and view campaign videos on YouTube (*www.youtube.com/user/ThisIsStats*). The more we spread the word, the more we can correct the frustrating misconceptions about statistics.

Census Bureau's Statistics in Schools Program

The U.S. Census Bureau's Statistics in School program provides math, statistics, and history resources to K–12 teachers. In his blog, Census Bureau Director John H. Thompson writes about the efforts to work with teachers to evaluate the materials: *http://bit.ly/statisticsinschools*.

Useful Websites for Statistics Teachers

The ASA currently hosts a listing of websites useful for teachers of statistics. The list was updated recently, though it is still a work in progress. Visit the site at *www.amstat.org/education/usefulsitesforteachers.cfm*. If you have recommendations or additions, contact Rebecca Nichols at *rebecca@amstat.org*.

Episode 10 of STATS+STORIES Available

Episode 10 ("Workplace Safety") of S+S is available at *www.statsandstories.net*. S+S guest Chris Whittaker Sofge, chief of the Risk Evaluation Branch in the Education and Information Division at the National Institute for Occupational Safety and Health, joined the Stats+Stories regulars to talk about evaluating risks associated with workplace hazards. You can subscribe to the Stats+Stories podcast on iTunes.

New Online Community for ASA K–12 Teacher Members

A new online community for ASA K–12 Teacher Members will allow participation in online discussions and sharing resources with other members. More information is available at *http://community.amstat.org/participate/helpfaqs*. Not yet an ASA K–12 Teacher Member? You can start your free three-month trial online at *www.amstat.org/membership/K12teachers*.

ASA Writing Statistics Education of Teachers Report

In light of the Common Core State Standards, the Conference Board of the Mathematical Sciences (CBMS) released *The Mathematical Education of Teachers II* (*MET2*), which focuses on the mathematics and statistics preparation of K–12 teachers. The ASA review of *MET2* was well received by CBMS, which encouraged the ASA to expand recommendations to a white paper. The ASA board recently funded this project to create a companion report on the statistics education of teachers, led by Christine Franklin and Tim Jacobbe, both members of the ASA/NCTM Joint Committee. For more information, see *http://magazine.amstat.org/blog/2014/03/01/education-of-teachers*.

Census at School Reaches More Than 25,000 Students

The ASA's U.S. Census at School program (*www. amstat.org/censusatschool*) is a free, international classroom project that engages students in grades 4–12 in statistical problemsolving. The students complete an online survey, analyze their class census results, and compare their class with random samples of students in the United States and other participating countries. The project began in the United Kingdom in 2000 and now includes Australia, Canada, New Zealand, South Africa, Ireland, South Korea, and Japan. The ASA is seeking champions to further expand the U.S. Census at School program nationally. For more information about how you can get involved, visit *http://magazine.amstat.org/blog/2012/02/01/censusatschool-2* or email Rebecca Nichols at *rebecca@amstat.org*.

ANNOUNCEMENTS

World of Statistics Website and Resources

The free international statistics education resources created during the 2013 International Year of Statistics are available and ongoing through The World of Statistics website. Teachers everywhere can access a wealth of statistics instruction tools and resources from around the world at *www.worldofstatistics.org*.

Free Statistics Education Webinars

The ASA offers free webinars on K–12 statistics education topics at *www.amstat.org/education/webinars*. This series was developed as part of the follow-up activities for the Meeting Within a Meeting Statistics Workshop. The Consortium for the Advancement of Undergraduate Statistics Education also offers free webinars on undergraduate statistics education topics at *www.causeweb.org*.

PROJECT-SET

PROJECT-SET is an NSF-funded project to develop curricular materials that enhance the ability of high-school teachers to foster students' statistical learning regarding sampling variability and regression. All materials are geared toward helping high-school teachers implement the Common Core State Standards for statistics and are closely aligned with the learning goals outlined in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE): A Pre-K–12 Curriculum Framework*. For more information, visit *http://project-set.com*.

LOCUS

LOCUS (*http://locus.statisticseducation.org*) is an NSF-funded project focused on developing assessments of statistical understanding across levels of development as identified in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE)*. The intent of these assessments is to provide teachers, educational leaders, assessment specialists, and researchers with a valid and reliable assessment of conceptual understanding in statistics consistent with the Common Core State Standards.

Enable Students to Explore Real Data on TuvaLabs

Tuva Labs enables teachers to teach statistics and quantitative reasoning in the context of real-world issues and topics. They empower students to think critically about data, ask meaningful questions, and communicate their observations and findings. Educators can conduct inquiry-driven activities, lessons, and projects related to topics such as sports, environment, entertainment, and latest news that are aligned to the Common Core Content and Practice Standards. With Tuva Labs, educators can equip their students with the knowledge, values, and skills to become active members of their own communities and global citizens of the world. Tuva Labs is 100% free for teachers and students. As they curate and make available hundreds more data sets, they would like to open them up to the ASA community to tag them appropriately so it is easy for educators around the world to find data sets based on topic, concept, standard, or grade level. You can check out Tuva Labs at *www.tuvalabs.com*.

STEW Lesson Plans

For free, peer-reviewed lessons, visit *www.amstat.org/education/stew*.

STatistics Education Web

STEW

Lesson Plans Available on Statistics Education Web for K–12 Teachers

Statistics Education Web (STEW) is an online resource for peer-reviewed lesson plans for K–12 teachers. The lesson plans identify both the statistical concepts being developed and the age range appropriate for their use. The statistical concepts follow the recommendations of the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework*, *Common Core State Standards for Mathematics*, and *NCTM Principles and Standards for School Mathematics*. The website resource is organized around the four elements in the GAISE framework: formulate a statistical question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the data in the context of the original question. Teachers can navigate the site by grade level and statistical topic. Lessons follow Common Core standards, GAISE recommendations, and NCTM Principles and Standards for School Mathematics.

Lesson Plans Wanted for Statistics Education Web

The editor of STEW is accepting submissions of lesson plans for an online bank of peer-reviewed lesson plans for K–12 teachers of mathematics and science. Lessons showcase the use of statistical methods and ideas in science and mathematics based on the framework and levels in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE)* and *Common Core State Standards*. Consider submitting several of your favorite lesson plans according to the STEW template to *steweditor@amstat.org*.

For more information, visit *www.amstat.org/education/stew*.

Teaching Statistics Through Data Investigations

New Online Course Coming Spring 2015

www.mooc-ed.org/tsdi

Learn with colleagues near and far in this eight-week, online professional development course, designed for teachers of statistics in grades 6–12 and post-secondary contexts. This course can help you learn to teach statistics using investigations, with real data and real cool tools! The course is FREE and can lead to continuing education credits.

Offered through the Friday Institute for Educational Innovation at NC State University Lead instructor:
Dr. Hollylynne Lee.

MOOC-Ed
MASSIVE OPEN ONLINE COURSES FOR EDUCATORS

CENSUS at SCHOOL

FREE international classroom project to engage students in statistical problemsolving

Teach statistical concepts, statistical problemsolving, measurement, graphing, and data analysis using your students' own data and data from their peers in the United States and other countries.

Complete a brief online survey (classroom census)

13 questions common to international students, plus additional U.S. questions

15–20-minute computer session

Analyze your class results

Use teacher password to gain immediate access to class data.

Formulate questions of interest that can be answered with Census at School data, collect/select appropriate data, analyze the data—including appropriate graphs and numerical summaries for the corresponding variables of interest—interpret the results, and make appropriate conclusions in context relating to the original questions.

Compare your class census with samples from the United States and other countries

Download a random sample of Census at School data from United States students.

Download a random sample of Census at School data from international students (Australia, Canada, New Zealand, South Africa, and the United Kingdom).

International lesson plans are available, along with instructional webinars and other free resources.

www.amstat.org/censusatschool

For more information about how you can get involved, email Rebecca Nichols at *rebecca@amstat.org*.

Bridging the Gap

Between Common Core State Standards and Teaching Statistics

Twenty data analysis and probability investigations for K–8 classrooms based on the four-step statistical process as defined by the Guidelines for Assessment and Instruction in Statistics Education (GAISE)

www.amstat.org/education/btg

MAKING SENSE OF STATISTICAL STUDIES

Making Sense of Statistical Studies consists of student and teacher modules containing 15 hands-on investigations that provide students with valuable experience in designing and analyzing statistical studies. It is written for an upper middle-school or high-school audience having some background in exploratory data analysis and basic probability.

www.amstat.org/education/msss