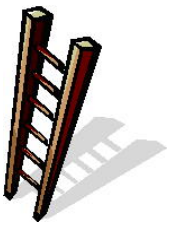
Statistics for stories

Jennifer LaFleur
ProPublica
Jennifer.lafleur@propublica.org

David Donald
Center for Public Integrity
ddonald@publicintegriyt.org

Before doing any analysis on your data, you must know something about your data, so you know which tool is the best to use.

What does your data look like?



You may have heard of the data ladder. It's a conceptual tool used to explain differences in data. We'll start at the bottom of the data ladder:

Categorical data

Categorical data comes in three flavors. The basic, most simple is DICHOTOMOUS. This means that the field has two choices: yes or no, 1 or 0, etc... Here are the sorts of analyses you can do?

Frequencies

Yes	123	34%
No	234	66%

Crosstabs

	Live in Texas	
Like Bush	Yes	No
Yes	382	200
No	125	307

The second type of categorical data is that with more than two possibilities such as age groups or race groups.

Frequencies

Race	Frequency
Asian	4766
Black	12807
Hispanic	9129
Native Amer	250
Multiple	1756
Unknown	3745
White	133915

Crosstabs

	Ticket Type 1	Type2	Type 3
Asian	1	3	0
Black	0	15	2
Hispanic	0	21	1
Native Amer	0	0	0
Unknown	7	22	3
White	7	71	31

The third type is ORDINAL. That means that the order of the categories means something, but that the actual values don't. Sound confusing? Consider some examples. School grade levels are categories. Four is definitely higher than one, but four is not necessarily four times one. Census age groups are ordinal as well (while age itself would be in our next category).

Continuous data

You've reached the top of the data ladder. You can perform math on this data. You'll often have so many numbers, such as all the individual incomes of people who applied for a mortgage, that running frequencies is impractical. Most continuous data is ratio data, meaning it's based on a scale that starts with zero and you can say a $50,000 salary is twice as large as a $25,000 income.

You have many analysis options with continuous data.

- Mean – This is a type of stat that statisticians call a "measure of central tendency." While you should never use that phrase in a story, it's good to think of the mean as an attempt to find the middle. The mean is what we usually think of as "the average" a very common stat you've probably already used in stories. You calculate the mean by adding all of your numbers (say incomes, ages or loan amounts) and dividing the sum by the number of instances. For example, we say the average age is X by taking five ages, summing them, and dividing by 5.

- Median – Another attempt to find the middle, the median is simply the middle number of all the numbers you have. Another way of expressing it is to say that half the numbers are above the median and half below. The median is especially useful when your data can be "skewed," meaning that a few really large numbers will make the average a less reliable view of the middle. Think of this as the "Bill Gates effect." If you live in a neighborhood with a bunch of houses that value about $100,000 each and Bill builds a $26 million mansion in your neighborhood, Bill's house will affect the average a lot but the median very little if at all.

- Range – Simply the difference between the highest and lowest values. If the highest mortgage loan in your community was $200,000 and the lowest $35,000, the range is $165,000. The range tells you whether your numbers are spread widely or clustered.

- Rank: We're used to sorting. Ranking is putting in order based on a score or other measure.

- Regression- Some characteristics help predict others. For example, people growing up in a lower-income family are more likely to score lower on standardized tests than those from higher-income families. Regression helps us see that connection and even say about how much characteristics affect another.

  To start, you need to determine which characteristics are what we call independent variables. These are the predictors. Next, the characteristics they help predict are the dependent variables.

  When you run a regression, you get a result called an R-square. That will help you see how much the independent variable predicts the dependent variable.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.911(a)	.830	.829	10.8330

a Predictors: (Constant), PCTPOOR

In this case, an R-square of .83 says about 83 percent of the poor test scores is related to the level of poverty.

- Correlation – But don't get into the habit of saying that the poverty CAUSED the lower test score. Most social researchers, from sociologists to political scientists, get hung up about causation – and they should. While we can see that poverty affects test scores, (The opposite doesn't make sense: We don't become poor based on our test scores.) we don't know that it actually causes low test scores.

  It's clear, though, that poverty and low test scores and other continuous data can be related. Correlation helps us see that relationship and the nature of that relationship. Does one characteristic increase as the other increases? Or are they inversely related, with one decreasing as the other increases? Even knowing the amount of correlation between two characteristics can make us more comfortable with our regression analysis. A Correlation Matrix shows how many variables correlate to each other.

  Dateline NBC wanted to investigate racial profiling. It looked at traffic violations recorded by many major city police departments and the race of those getting the citation. Showing a correlation between race and the violations helped make the point and make the reporters more confident about further analysis used for the story.

Mixing it up

Mixed categorical and continuous: Sometimes you get categorical data and continuous data, just as Dateline did when it analyzed race and traffic violations. Race and people's incomes are likely to show correlation.

- ANOVA, or Analysis of Variance, helps us look more closely when your independent variable is categorical and the dependent continuous.

ANOVA uses a number of tests to measure the main effects the independent has on the dependent. It can also look at the interaction effects of two or more independent variables, say race and sex, on someone's income. How do the groups compare is a typical question for ANOVA.

Categorical and continuous AND your outcome is dichotomous

Clearly, you can do frequencies, crosstabs and the calculations that you can do with continuous variables. But when you want to look at how some variables affect a dichotomous outcome, logistic regression is the tool.

That allows you to control for certain variables and compare groups. The Boston Globe used logistic regression on traffic stop data to be able to say that whites were less likely to get tickets than minorities and that men were more likely to get tickets.

> On city boulevards and rural lanes, whites are far more likely than minorities to receive written warnings instead of tickets when stopped for identical traffic offenses, according to a Boston Globe study of newly released state records. And women, especially young women, get breaks that aren't afforded to men. The price tag for this unequal treatment amounts to an estimated $25 million a year in traffic fines and higher insurance premiums.

Those statements come from a fairly complicated analysis. Below is an example of some of the output. Look at the last column. This tells us that minorities have more than 8 times the odds of getting stopped as whites. You'll need more information on logistic regression to do one.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	minority	2.127	.694	9.383	1	.002	8.386
	female	-.786	.453	3.010	1	.083	.456
	mphover	.292	.096	9.348	1	.002	1.339
	Constant	-4.640	1.473	9.918	1	.002	.010

a. Variable(s) entered on step 1: minority, female, mphover.

What it also let's you do is control for other factors. For example, in analyzing home loans a logistic regression would let you see if minorities were more or less likely to get loans and control for differences in income and other factors.

Safety First: Up and down the ladder

DOWN
Say you want categorical data and your data are continuous. For example, you want age groups, but you have ages. Recode you data into categories. Look at the distribution to decide where to make the breaks.

UP
If you want continuous data, but you only have categorical, the ONLY way to go up the ladder is to aggregate to another level. For example, if you have people and races and their census tract. You can create continuous data by aggregating by tract.

Some notes about dealing with polls and surveys

- Get the questionnaire and the methodology. If they won't give it to you, that's a red flag
- Beware of nonscientific methods: Web surveys, man on the street or other self-selection
- Know the sample size, which will give you the sampling error
- Again, know the source
- Account for margin of error and non-response or "don't know" when drawing conclusions
- If possible, run statistical tests on the data. What may look significant to you, may not be.
- When reporting, avoid false precision. Saying 52.18 percent of people think "blah, blah, blah" is portraying an impossible accuracy to readers.

Don't run with scissors

- Make sure you know how many records you should have and that you have them all. If you or someone in your newsroom tells you they have 65,536 records, be concerned. That is the exact limit of Excel.
- Double-check totals or counts. Check for studies or summary reports
- Consistency-checked all fields. Are all city names spelled the same way? How about other important fields?
- Other basic checks: make sure all states are included, all cities/counties are included, the range of fields is possible (for example, check for DOBs that would make people too old or too young.)
- Check for missing data or blank fields
- Check your methodology (if necessary) against other similar research

Consider your goal

Show Counts
Use frequencies and cross tabs. Check to see if there are other tools, such as an index you should use. Find out how that data is typically reported.

Change/growth
Make sure that some underlying factor does not account for the increase. Let's take an example of an increase in measles among people 65 and older in New Jersey:

- Use a rate, raw numbers could simply mean that the 65+ population has increased

- Check how the data is reported. Could reporting changes account for the increase?
- Has anyone else studies this?

Rank
Are you using the right variable to rank? If you are using sample data, be careful, the ranks may change within the margin of error. Make sure that what you're ranking on doesn't mimic some other factor (otherwise you may need an index). For example, an increase in rich Asians, could just mean an overall income increase.

Correlate
Use your statistical software to run a correlation matrix. Keep in mind that things may correlate, but not necessarily explain each other. For example, there is a correlation between the number of taxis in a city and the number of pigeons – why? Bigger populations in both, but no meaning.

Show how much of a factor is explained by other factors: Regression
If you're working with a continuous outcome variable, you'll use linear regression. If it's dichotomous, you'll want to use logistic regression.

Stats on the beat

Education
- *The Dallas Morning News* used linear regression to find schools that scored much higher than expected on tests. This lead to stories about potential cheating

- Several newspapers have used demographics and test score data to what factors impact whether scores go up or down and then use those factors to see who is doing as they should, taking those things into consideration.

Crime and justice
- *The Boston Globe* and Dateline NBC used logistic regression to examine traffic stop patterns.

- *The Dallas Morning News* used logistic regression to show that blacks were more likely to be struck from juries by prosecutors.

- Several newspapers have used linear regression along with local crime statistics to show what factors drive crime rates.

Sports
- The Pittsburgh Tribune-Review analyzed NFL data to look at player injuries. They also looked at the effect of players in certain positions being injured.

Business
- Chris Schmitt, then of the San Jose Mercury News used the Herfindahl-Hirshman Index, which measures competition in a market, along with regression to show that the NASDAQ stock market was not as competitive as it was supposed to be and that because of that, investors lost money.

Politics
- In analyzing polls and voting results, news organizations have used statistics to look at who votes particular ways.

Indexes

An index is used to simplify the measurement of movements in a numerical series or to compare a set of numbers and is usually on an understandable scale (0-1 or 0-100). Indices also can be used to combine several variables.

Indexes are all around us: The Dow Jones Industrial Average or the FBI's crime index.

A good example of an index that combines several values is the consumer price index (CPI). The CPI represents all goods and services purchased for consumption by urban households. We have classified all expenditure items into over 200 categories, arranged into 7 major groups.

You could come up with your own sort of CPI, depending on what you needed to look at. Let's say you're supposed to look at costs for skiing at various resorts. Rather than just comparing lift tickets, you might realize that skiing is more than just the price of a lift ticket. Your Cost-of-skiing index might be made up of the cost of a lift ticket, ski rental, transportation and an after skiing drink. That would give you one number to compare across all ski areas.

Measuring Diversity
If Washington reporters can be accused of pack political journalism, CAR reporters can be accused of pack nerdiness. One particular index hit its hey-day after the 2000 Census – the Diversity Index, a cool tool developed years earlier by Philip Meyer and Shawn McIntosh for *USA Today*. The index measures the probability that two people pulled at random from a given area would be of a different race. The higher the diversity, the more likely the two would of different races.

"The USA Today diversity index solves this problem with a probability-based index. The index has a range from 0 to 1, and its value represents the probability that two people chosen at random from the study population will differ along at least one ethnic dimension." (International Journal of Public Opinion Research, Spring 1992).

Here's the formula to calculate the Index of Diversity:

Step 1:
Probability that two persons chosen from a population at random will be members of the same racial group: $P_R = (A^2+B^2+C^2+D^2)$ where A, B, C & D are the proportions in the population of whites, blacks, native Americans, and Asians or Pacific islanders. (Using the single-race total as the base.)

Step 2:
Compute the probability that two persons chosen from a population at random will be either both Hispanic or both not Hispanic. This is a separate value because Census questionnaires ask Hispanic origin in a different question from race. Therefore it is possible to be both white and Hispanic or black and Hispanic.
$P_H=(H^2+N^2)$

Step 3:
Calculate the probability that two randomly chosen persons are the same in both race and Hispanic/non-Hispanic status: $P_R * P_H$

Subtract the result from 1: $1-(P_R * P_H)$

The Gini coefficient

First of all, it's not every day that you get to use a big word like "coefficient." But other than impressing your nerdy friends, the Gini is a useful tool. This tool was developed by Mr. Gini (Italian economist Corrado Gini) in 1912 to estimate the inequality of incomes and wealth.

Special tools are needed to measure income inequality because standard income measurements, such as median, don't give you the whole picture.

Consider two income distributions:

County 1	County 2
100	18000
888	19001
1000	19300
1200	20000
48800	30000
50000	31600
51800	32000
70000	34000

The median for both of these is $25,000; however, the range is significantly different. County 1 has incomes at extreme ends of the spectrum, while county 2 tends to be in the middle.

Because I didn't have time to come up with a good tool to measure inequality and Mr. Gini had already done some all the work, I used a spreadsheet to use the Gini formula. (which J.J. Thompson had developed at the University of North Carolina and provided to students in an advanced NICAR boot camp a few years ago - thanks J.J.!)

The data used for the Gini is categorical income data such as that provided in Census figures. For example: 500 people in the $10,000 to $14,999 category and so on... By multiplying the midpoints of each category by the number of people in the category, you can derive the weighted income and therefore the other parameters.

The result of the Gini is one number you can use to compare your county to other counties or to your county over time. Here's the formula:

Gini=1-SUM(Xi-Xj)(Yi+Yj)

Where:

X is the cumulative proportion of recipients
Y is the cumulative proportion of income
i is a particular income category
j is i-1 (the previous income category)

This is repeated for all the income categories and then totaled so you end up with one number for the county or whatever area you're calculating. You can calculate that same number for other counties to compare geographically or to other years to compare over time.

Competition

When he worked for the San Jose Mercury News, Chris Schmitt wanted to look at market competitiveness in the NASDAQ, which claimed to be more competitive because of the computerized system it used. Schmitt did some research and started talking to experts.

He used a tool used by economists, the Herfindahl-Hirschman Index. This index is based on market shares of the players in a particular marketplace. Viewing the dealers who trade a particular NASDAQ stock as a marketplace, the Mercury News applied this measure to the market shares held by each dealer in that stock. The HHI index is calculated as the sum of the squares of the market shares held by each market participant.

Index values are interpreted as follows:

‹1,000 Competitive
1,000-1,800 Moderately Concentrated
›1,800 Highly concentrated

The H index is obtained by squaring the market-share of each of the players, and then adding up those squares. For example:

(%Share of company1)+(%share of company2)+...

The higher the index, the more concentration and (within limits) the less open market competition. A monopoly, for example, would have an H index of 100^2, or 10,000. By definition, that's the maximum score. By contrast, an industry with 100 competitors where each has 1 percent of the market would have a score of $1^2 + 1^2 + 1^2 + ...1^2$ or a total of 100.

What Schmitt found:
NASDAQ is supposed to keep investors' costs low through spirited competition among the firms that execute your orders to buy or sell stocks. But

NASDAQ frequently doesn't operate that way, a San Jose Mercury News examination shows. Often, trading of a company's stock is concentrated in the hands of only a few dealers. And when that happens, the transaction costs for investors are higher.

Your own index

Sometimes, you want to know something but no one has a ready-made index to measure what you're looking at. You can create your own index.

When looking at people's attitudes toward pro-development versus pro-environment, no index measures this. And it's hard to ask people up front, which are you? The socially acceptable thing is to say that you're pro-environment, even if you're not.

One way is in a survey to ask a bunch of questions on a scale, such as whether more material consumption makes happier people. You ask on a scale of 1-5 if you agree or disagree with the statement.

Say you ask seven questions along this line. Then you can see if they correlate. For your index, you want some correlation but not too much. No correlation means your questions are not measuring the same thing. Too much correlation and there's no need for the different questions. So a Pearson Correlation or "r" between .2 and .5 is good between the questions.

Next, take the questions that correlate and run something called Alpha, which shows how much effect each question has on the others. You want the score to be above .7 at least, .8 is better. If you find that to be met, then you can average the scores on the seven questions (or however many correlate) and come up with one score that measures your difficult concept.

Using that, the Savannah Morning News showed that those who are most pro-development were lower income people, those who had been left out of the American Dream. Material success apparently allows you to become more pro-environment.

For more information...

Numbers in the Newsroom: Using Math and Statistics in News by Sarah Cohen for Investigative Reporters and Editors, Inc.

Precision Journalism, 4th ed., by Philip Meyer. Rowman & Littlefield Publishers: Lanham, MD, 2002.

News and Numbers by Victor Cohn. Iowa State University Press, Ames. 1989.

How to Lie with Statistics by Darrell Huff. W. W. Norton & Company, New York. 1954 (renewed 1984)

Innumeracy: Mathematical Illiteracy and Its Consequences by John Allen Paulos. Vintage Books, New York. 1990.

A Mathematician Reads the Newspaper by John Allen Paulos. Anchor Books, New York. 1995. (Also, check out the tape from Paulos keynote address at NICAR 2002 in Philadelphia)

The Journalist and the Gini Coefficient: A Statistical Approach, by J.J. Thompson. Master Thesis, University of North Carolina-Chapel Hill School of Journalism.