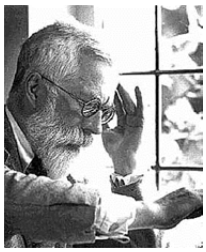


A statistical primer: The pluses and minuses of going beyond addition and subtraction

Jennifer LaFleur, CIR

Much to the chagrin of my engineer brother, my 15-year-old niece pronounced upon returning home one afternoon last week that because she planned to be a journalist, she didn't need to know math. He got me on the phone immediately. I told her that no matter what, journalists need to know basic math: addition, its tricky partner subtraction, multiplication and division. They also need to know some very important calculations:

Percent of total
This one is easy to get turned around. The rule is: the amount divided by the total.

Percent change
When you're dealing with two years worth of data and want to figure percent change from one year to the next, take the difference of the two values divided by the amount from the earlier year.

Per capita and rates
Looking at occurrences by city or county or state can be interesting. But remember to keep things on common ground. I can't compare raw numbers if the "bases" are different. That means if I'm looking at the number of murders or the number of home loans by city in the United States, Los Angeles and New York probably will come out on top, but only because they have more people. A more useful measure would be per capita murders. In this case, per capita murders would be the number of murders divided by population.

Sometimes these measures will be adjusted a little if the number of occurrences is low. Crime statistics, for example often are reported per 100,000 or per 10,000 people. These measures are used because the number of murders in a city is low compared to the number of people. If you have multiple murders per person – you might want to check the source of your data. When the per capita measure might be 0.0005, the figure per 100,000 people it becomes easier to read: 50 murders per 100,000 people.

Per-plexed? When you're trying to figure if persons per household is households divided by population or population divided by households, do one simple thing: turn your per into a division sign. This means that persons PER households becomes persons / households or population/households.

Mean, median and mode
Why do we often see things such as home values and income reported as median rather than just the average (mean)? That's because the median (the middle value – the point at which half the values are above and half are below) isn't swayed by extremes like the average (total of everything divided by the number) is. So one big fancy house on a block will throw off the average value for the neighborhood.

If you're trying to decide which to use, test to see whether there is a big difference between the mean and the median. If there isn't, use the mean. It's easier to explain to readers.
Why care about the mode? The mode (the most frequently occurring value) can be important to show clusters.

Make some adjustments
If you are dealing with money from year to year, you need to know how to adjust for inflation. To compute inflation, just look at the percent change in the Consumer Price Index (CPI) from one year to the next and adjust the previous years "up" to reflect them in current dollars. Because CPI varies regionally, use the CPI for your area.

Exploring your data

When you go beyond calculations and want to start studying your data, here are some things to look at:

Range

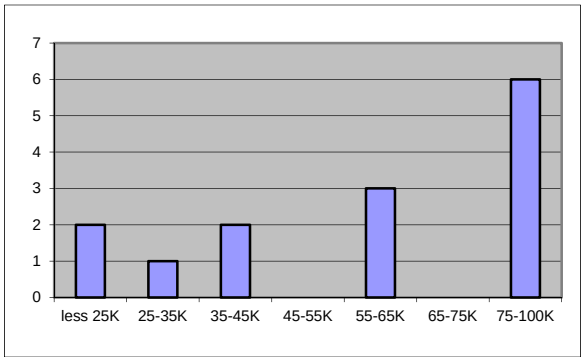
By examining the range of your data – the highs and the lows -- you'll find extremes. This can help you show part of the story about your data. We report on ranges a lot. A good example is with census data where we report the "richest" census tract or the "oldest city."

Distribution

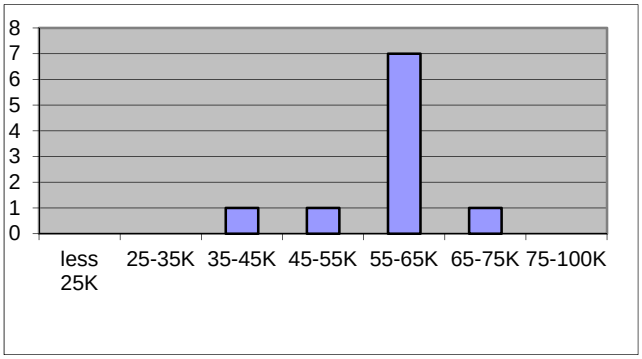
The range gives you part of the picture, but how the rest of the data is distributed can make a big difference in the story that it tells. Let's examine two income distributions:

Group 1	Group 2
$22,132.00	$44,264.00
$24,156.00	$48,312.00
$35,462.00	$55,252.40
$36,187.00	$55,511.40
$56,532.00	$56,532.00
$56,732.00	$56,732.00
$56,932.00	$56,739.20
$78,932.00	$56,932.00
$79,302.00	$57,899.20
$87,543.00	$61,280.10
$87,543.00	$61,280.10
$87,543.00	$61,280.10
$99,000.00	$69,300.00

These groups have the same median, but the ranges of each are pretty different. Group 1 incomes are concentrated on the extremes, so the distribution looks like this:



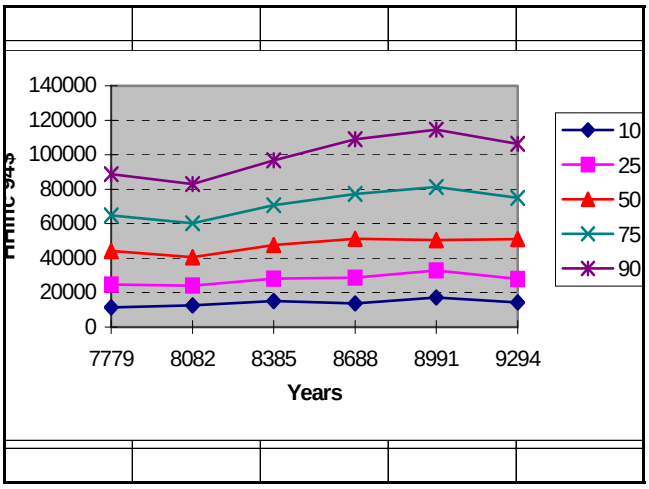
Group 2 is more concentrated toward the middle:



N-tiles

Another tool for looking at a distribution, such as incomes for a given geography are N-tiles. These are especially useful when you're trying to look at how well different income groups fare over time. (Keep in mind that your numbers would need to be adjusted for inflation.)

Here's an example of income groups over time in Santa Clara County, Calif.



Statistics

Many of you are here because you know the basics and you're ready to move beyond basic math. Much of what do with computer-assisted reporting is sifting and sorting through data and doing calculations, but sometimes you need a more powerful tool.

First, let's figure out what a statistic is:

1 : *a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data*
2 : *a collection of quantitative data*

That's not so scary – you've been doing that all along. You might think of statistics as using statistical software such as STATA or SPSS or SAS to run statistical tests or analyses. You might think of statistics as dealing with survey or poll data.

Why would you want to use something beyond sifting and sorting and calculating? Some day, you'll run into data with which you need other tools to tell the whole story. It's sort of like the different between showing a list of census tracts and mapping the data – it shows you the patterns that you can't see any other way.

Let's think about a particular type of data: school test scores.

You have many ways you can provide your readers information about the data. You could:

1. Just print the scores and let them figure things out for themselves.
2. Print the scores showing increases and decreases from previous years
3. Categorize the data -- the percent that scored above the mean, below the mean
4. Break them into N-tiles

You might even to go further and ask WHY. Why do certain schools do better or worse: Smarter kids? Better teachers? Environmental considerations? Mobility? Language? Substitute teachers? Computers?

If you're not an expert on school test scores, you might see out an expert. You may ask experts why – what do you they look at? They may tell you that factors, such as poverty, mobility, teacher experience, all play into how well kids perform on standardized tests. But should you just believe them?

With statistical software, you actually can test these notions for yourself. You can tell in your area how strong a factor teacher experience is on how well kids perform.

Here are some tools and calculations that may help you and your readers better understand your data:

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.

Punxy Phil's Weather Index

To test the accuracy of Pennsylvania's famed meteorological marmot, one might need to come up with a sort of weather index combining temperature, frozen precipitation and hours of sunshine for six weeks following Feb. 2. Your index could be made up of the number of days below and above the median temperature, the number of days with snow and the number of hours fewer or more sunshine than normal.

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:

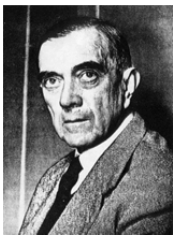
Step 1:
Probability that two persons chosen from a population at random will be members of the same racial group: $P_R = (A^2+B^2+C^2+D^2)$ where A, B, C & D are the proportions in the population of whites, blacks, native Americans, and Asians or Pacific islanders. (Using the single-race total as the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

Step 2:
Compute the probability that two persons chosen from a population at random will either be 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.

Let's 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 market, which was claiming 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 was that:

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.

Regression

This is a statistical tool that you can compute using either Excel or a statistical program such as SPSS or SAS. It does a couple of things. First, it looks at the impact of external forces known as independent variables on another variable (the dependent one). It also allows you to level the playing field and measure the true gain or loss of one entity.

This tool frequently is used in school test score analyses to find out what variable has the biggest impact on school test scores and to use that "model" to determine how well schools should be doing. This will show you who's performing up to their abilities (given the factors you've determined) and who's not.

And things get fancier from here. There's your basic linear regression, logistic regression, multivariate regression and so on. Before applying this tool to a story, it's good idea to study up by taking an IRE stats course, taking a local college stats course and befriending a local professor to help verify your results.

A little light reading...

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

The New Precision Journalism by Philip Meyer. Indiana University Press, Bloomington. 1991.

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.

Other things you can do to learn more:

Read research papers in your beat area – boring – but it can lead you to data and methodologies.

Attend a workshop for data analysts in your beat area. (Crime analysts, educators, etc…)

Read past tipsheets and stories from the IRE resource center.

Attend a stats workshop or take a class at a local college or university.