Newsroom Statistics in the Digital Age

A lot of journalists will tell you they don't do math. But there is a growing number of journalists that not only do math, but use it as an integral part of their work. In this digital age, part of being a resourceful journalist is knowing how to find data then use statistical methods to reveal the structure and mechanics of that data. Those methods vary in complexity depending on the goal, but a passing familiarity is necessary when dealing with data of any sort.

I spoke with 9 data-savvy journalists from across the nation to explore the best practices for dealing with data statistics. The emerging landscape of data journalism has given rise to a litany of new job titles, from "data specialist" to "quantitative editor." Despite the disparate titles, they all have the same aim: using the language of numbers to tell stories. These journalists use statistics as a reporting tool to interview the data, acquiring information to inform their narratives. At this point, their applications diverge. Some will create graphics for presentation, while some do further analysis to explore the significance of the data. They carefully catalog its details as evidence to buttress or disassemble an argument. Some of the most advanced practitioners use statistics to engineer predictions about elections, sporting events, and laws being passed.

The explanation of these techniques in the midst of a story can be challenging both for writer and reader. Most journalists won't try to weave specifics about methods into the narrative, but many feel it's critical to describe the effects of the analysis. Journalists

employ various explanatory techniques while striving for clarity and accuracy with their individual readerships.

Numbers are a language. Just like any foreign tongue, they can be cryptic and incomprehensible to the uninitiated. A journalist uses language to tell stories, and numbers are no exception. Statistical methods are a way to describe groups of numbers, as adjectives might describe a noun. Thus, statistics are an extension of a journalist's vocabulary for telling stories with data. And data are literally everywhere now.

"It's not what I imagined at j-school, but the more I learn (about statistics) it's part of how I report. I think it's just exciting that there's so much more data," Tom Meagher, data editor at The Marshall Project, said.

The bounty of data Meagher describes has allowed a number of new data driven journalism sites — such as Vox, The New York Times' The Upshot, and fivethirtyeight.com — to employ data to tell stories about a vast array of topics. These news operations embrace an empirical approach to the news, a practice rooted in many investigative and data teams at more traditional news organizations.

"The rest of the New York Times is very, very good at answering the basic question, 'What happened?', and that's a pretty important duty," Derek Willis, a reporter for The Upshot who focuses on campaign finance, said. "The Upshot can, when we're really good, be more about how or why. We ask if there is a data set or research that helps shed some light on this topic."

Data for a story can be acquired in many ways, from scraping websites to downloading from government repositories. No matter where it comes from, reporters

will familiarize themselves intimately with what the data describes. Jon McClure, a news application specialist for the Dallas Morning News, describes how he grows familiar with data for a story.

"We tend to personify it almost; you spend time getting to know your data, back to front," he said.. "Every field. Understand the purpose of data, understand the pipe through which it's collected, stored, and used. Have a grasp of the data that's holistic and well-rounded."

This intimate familiarity with what data describe can help the journalist make judgments about points in the data that might not fit. Data can be dirty. Knowledge and common sense are the first line of defense in ensuring a journalist is working with good, accurate information.

Reporters such as Meagher and McClure use statistics to literally interview data. They probe and poke at the datasets with statistical methods to elicit answers. They ask questions about how much there is, what occurs most often, what the average is, the range of the data, and so on. These processes are loosely termed "summary statistics" and comprise the starting point for almost all analyses.

With this set of background information, reporters are then able to dive deeper into the shape of the data and see what might be of interest.

Steven Rich, the database editor for investigations at The Washington Post, uses these techniques to peer into phenomena that he can't observe directly.

"Stats is a problem-solving tool first and foremost," he said. "It helps to get at things you can't get at in other ways."

Using statistics, Rich was able to conclude that several companies were colluding in purchasing tax liens at Washington, D.C., auctions. "There was no video, no records about who posted and didn't win. We only had winning bids," he said, describing the challenge of analysis. "We ended up using stats to look at the patterns of how they were doing it." The text of the story states, "…the newspaper found clusters of back-and-forth bidding among top competitors that did not appear randomly in the running of 1,000 simulations." The non-random bidding showed there was a very high likelihood of collusion among the top bidders, an illegal practice that falls under criminal conspiracy. The analysis was conducted in concert with a team of economists and anti-trust experts in Boston, and it gave the Post an empirical stance from which to make the bid-rigging allegations.

To think about how reporters use statistics in another way, imagine an object of interest in an opaque plastic bag. If the air were removed, the bag would shrink and conform to the contours of the object, and its shape would be revealed. The proportions and dimensions would be pronounced and could now be described. Statistical methods are akin to removing the air from the bag and allowing the shape of the data to appear. To extend the metaphor a little farther: A reporter appraising the shape of the object through the bag would be especially drawn to any strange protrusions. These protrusions, in statistical terms, are called outliers. And they are often of interest to journalists as they describe a large divergence from the rest of the data.

McClure, in working with Medicare data for a story on doctors in Texas, was examining rates of procedures, and finding the outliers was integral to his story.

"We use (summary statistics) for just about every long-term original work, checking for outliers and how tightly things are clustered around the mean." The outliers he found represented the Texas doctors that were prescribing certain procedures far more often than their contemporaries, then collecting Medicare money for it. Using statistics, McClure was able to quantify exactly how much more these doctors were recommending procedures. Under a section of the story labeled "Texas Outliers" he wrote: "In 2012, Medicare paid Walker for 1,302 medium-complexity office visits with 60 patients. That rate, 21 visits per patient, was nearly 80 percent higher than the next-leading neurologist in the nation and over 15 times the national average."

Variance goes hand in hand with outliers. It's an important statistical concept when using numbers to describe the world. In his 1972 book "Precision Journalism," Philip Meyer describes how to harness the power of statistics in journalism and, specifically, how to look for variance. "It is the things that vary that interest us," Meyer says. "Things that do not vary are inherently boring."

Meyer goes so far as to say that variance makes news. He cites weather as an example. Temperature varies across the United States, in some places much more than others. Kansas has weather that's very hot in the summer, very cold in the winter, and everything in between. Arizona, on the other hand, has only two seasons: hot and really hot. The local weather is not nearly as newsworthy in Arizona as it is in Kansas. It simply does not vary as much.

A journalist must look at measures of this variation. Kansas' average temperature will have to encompass the freezing cold winter and the blazing hot summer. The average is not a very informative statistic in Kansas, as the spread of temperatures is very wide, whereas in Arizona, there is no real winter to speak of, and the average would likely be much closer to the current weather. This spread of numbers is termed as "variance" and is an important way to describe the shape of data as a whole. Variance, as McClure mentioned, tells you "how tightly things are clustered around the mean."

When looking into data, it is imperative to identify exactly which data points vary, how profoundly, and in what fashion. Many journalists, including reporter and data specialist Jeff Ernsthausen of the Atlanta Journal-Constitution, will employ a graphic display as a way to see outliers, variance, and many other summary measures at a glance. "I think it's really useful to visualize in segments and quintiles," Ernsthausen said. "Visualizing and turning it over as much as possible is the most valuable thing going into any kind of data analysis. That way you can see anomalies and outliers.."

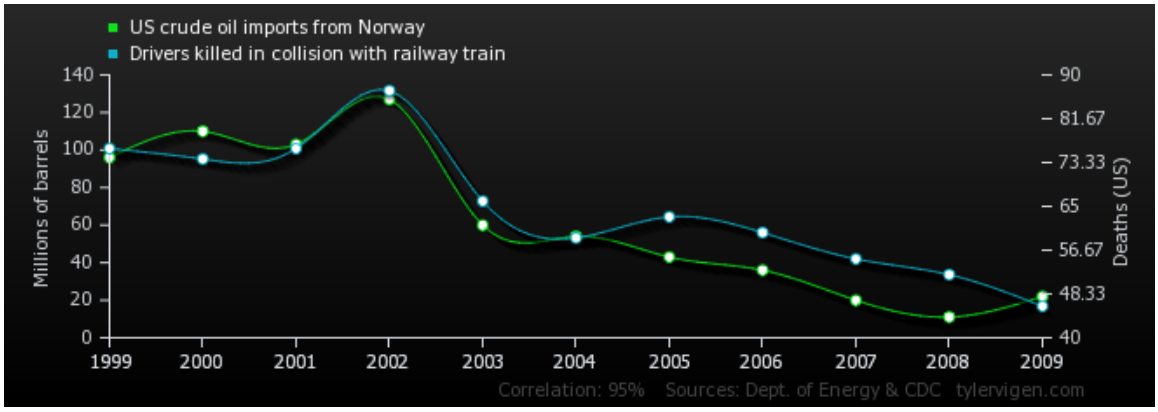
In a similar vein, McClure once made a number of charts showing different facets of the same data. "I had this 'book' I was carrying around with 300 different (medical) procedures, and each had a significant outlier. I was showing it around in order to get different points of view on them." His set of charts was valuable for his own interpretation as well as for easy communication of the findings to his colleagues.

Holly Hacker, an education reporter and data specialist at The Dallas Morning News, also used a chart to great effect while prompting a school district official to comment about highly variant test scores at a particular school. "I was trying to give them

enough information to understand and comment. When we showed it to them (in a chart) it was much harder to brush off," she said. The school district conducted its own investigation of that school based on the evidence from her chart.

Graphics are also a good way to show correlation, which is when the variance in one variable is in sync with another. This is a slippery slope, but correlation might be an indication that one variable actually causes the movement of the other, more or less. Going back to the weather example, temperature roughly correlates with the quantity of daylight. Thus the longer the sun is out, the warmer it generally will be, and the less time the sun is in the sky, the colder it generally will be. The sun certainly is not the only variable causing higher daily temperatures, but the two variables generally are correlated. Now, here's the slippery part: Just because two numbers vary in tandem does not necessarily mean the variance in one causes the other. In other words, correlation does not (always) equal causation. Please see the chart below for a splendid example of two variables that correlate, but one certainly does not cause the other.

US crude oil imports from Norway
correlates with
Drivers killed in collision with railway train



Credit: www.tylervigen.com

This is where a journalist's knowledge comes in. Because significant correlations sometimes happen when there is no relationship in the real world, it is incumbent upon the journalist to use judgment and knowledge of the subject to interpret the findings. The statistical method called regression helps measure the significance of this relationship between two variables. Ernsthausen tends to use regression as a diagnostic to make sure he's on the right track when dealing with complex datasets. "When you have a mix of variables, especially when they move in tandem, (regression) can be a great tool for teasing out which variable matters more," he said. Ernsthausen understands that the results of a regression analysis are more meaningful when stated in terms that the reader can quickly grasp.

"I do try to avoid if possible having to describe something on the basis of purely regression. I try to tell a story that makes more sense to people and be more powerful and dig though the issues," he said. "I use (regression) as more as a check of myself. I usually try to visualize or to turn it over enough times to describe the story it tells."

Rich also shores up his knowledge by relying on the expertise of others in the field he's investigating. "It's always better to call someone to see if I'm using the data correctly rather than publish a story that's wrong," he said. "The different statistical methods are like tools in a toolbox. You gotta know which ones to use. Part is experience, and part is calling experts and saying, 'I'm thinking about using this data to describe this.' I rely on them to advise me and guide me."

Framing statistical results through words a layperson can understand is often challenging. A balancing act often ensues when reporters are walking the line between accuracy and readability. Jon Perry, a colleague of Ernsthausen's in Atlanta, worked on a story published March 25, 2012, titled, "Cheating our children: Suspicious school test scores across the nation." Information from this story led to the eventual conviction of 10 public school educators, with each facing up to 20 years in jail. Perry translated the results of a linear regression model into simple probabilities that allowed a reader to understand exactly how improbable the events that occurred really were. "In nine districts," he wrote, "scores careened so unpredictably that the odds of such dramatic shifts occurring without an intervention such as tampering were worse than one in 10 billion."

Ernsthausen saw this as a very effective way to discuss the results with readers who might not be fluent in the language of statistics but are certainly capable of understanding the resulting improbabilities. Rich echoes this sentiment. "It's so hard to explain to readers," he said. "Statistical words will stop the readers cold. It's the sausage, and most people don't want to see how the sausage is made."

Some readers, however, want to know how it all works. For those people, publications employ a variety of strategies to work through the methodology. This often is referred to as a "nerd box" or "geek box," or at the Post, a "did box." Andrew Flower, the quantitative editor for fivethirtyeight.com, said the scale at which the methodology is explained relates to how much the story hinges upon the findings from the analysis.

"Is it a quick-and-dirty, super-simple logistic regression? Then the question becomes, do you even need to mention it or explain how it works? If it's the entire conceit of a piece and it hinges on the statistical findings, then yes, explain it's got a binary outcome, X,Y, Z control, and these results with these standard errors…" In terms of weaving the explanations of the methodology into the text, Flower will push and pull with the editors through analogies that are accurate but use general phrasing. "If you work hard, you can elevate it beyond a wiki or stats textbook definition and make it lively," he said.

In Neil Paine's story about the hottest goalie in the Stanley Cup playoffs for fivethirtyeight.com, he discussed the superstition that athletes featured prominently in high-profile media are "cursed" and rapidly fall from grace after the attention. Paine explains his point though statistics in language that is relatable and informative.

> "But in these kinds of cases, regression to the mean is the more likely culprit. To appear on the cover of the "Madden NFL" video game or Sports Illustrated, a player had to play at an incredibly high level and was usually helped along by luck (which includes staying healthy). When that luck dissipates, it seems there's a curse attached to the accolade.
>
> This is more true for the hottest goalie list, because I set up that metric to find players who were playing above a level that could be explained by their previous performance baselines and even the shooting skill of the opposing team.

Whatever's left over is, by definition, going to be fueled largely by luck, and therefore primed for regression."

These statistical concepts arrive in the midst of a narrative, but even if it's in a footnote or a nerd box, discussing the methods points to transparency of thought. Many publications further this aim for transparency by posting their data online along with their methodology, and any code needed to transform the data. Flower offered the most concise reasoning for when his publication decides to publish the data. "The data set has to rise to the occasion that merits publishing it. Are we going to publish data for every little unemployment chart? No. We say, 'Yeah the average reader is not going to be able to scrape or munge this data.' So it's more valuable."

Many publications seem to be moving toward sharing data for these reasons of transparency and validity. Amanda Cox, a graphics editor at The New York Times, is a firm proponent, especially when the methods venture into making predictions based on data. "Release in full all of your data and code if you want to play in this space, " she said. "It's a very responsible thing to do if you want to claim it's not opinion. It's the price of admission in some ways."

Cox, who has a master's degree in statistics, sees little practical use for presenting predictive statistics in the news. She describes most of the statistics in use at The Times are akin to accounting. "It ends up being relatively simple and straight-forward enough for people to understand. Being debatable is not so great." Cox noted that most of the complex statistics at The Times often end up under the hoods of their more sophisticated

news apps. These helper methods are prevalent in their very popular linguistics quiz, the Facebook sports fan maps, and the Senate prediction model, Leo.

Cox posed a philosophical question about how statistics might be presented: "Should a regression line belong in the news or the opinion section? After all, it's something that you made up based on data, and not different than a collection of facts or arguments." She sees the rarity of subject matter expertise in journalism as a reason to be cautious in the judgment of statistical analysis. "Another reason we're conservative in the use of (regression) is that there's no editing structure for how good or bad a model is, and that is a deep burden to put upon your readers, to force them to judge if your model is good."

In conclusion, my research was focused on finding the current best practices regarding statistics with journalism and how to best explain the use of statistical methods in journalistic work. I worked with a field of data journalists to explore these topics and found that working with statistics in journalism had established practices that mirrored basic social science analysis. The communication of the findings however was much more varied and depended greatly upon the publication's understanding of their audience. News sources like fivethirtyeight.com or The Upshot, which are dedicated to a more empirical view of journalism, has greater leeway to include basic statistical terms and methods in their stories, while the Dallas Morning News or the Atlanta Journal Constitution will often give an overview of their analysis in the story and defer to a 'geek box' to explain the details of the methodology.

The ability to employ statistical methods is a skillset that has applications that traverse the entire spectrum of journalism. The amount of data in our world is growing larger every day, and the benefit from knowing how to explore and evaluate data applies to every discipline in the newsroom. These techniques help to uncover stories that cannot be observed directly and find evidence of trends that occur over long periods of time.

“It's one thing to say, here's a problem but not know what to do about it,” Rich, of The Washington Post, said. “Stats can help to build a case with data.”