Numeracy and Statistics

Most journalists got into this business because they don't like math. Unfortunately, you'll encounter numbers all the time. Just look at any day's newspaper or TV broadcast and tally up how many stories involve numbers.

One guy did it for the St. Petersburg Times and found one-third of the 130 stories in that day's paper involved numbers.

"Too many of the stories display numeric overkill, a deadening procession of figures that overwhelm the reader and rob the writer of the opportunity to use numbers in ways that explain and illuminate" – source: "Avoiding numeric novocain" by Chip Scanlan

These stories were from all sections of the paper:

Some examples:
Professional sports team purchase
Legislation to limit consumer lawsuits
Questions about local govt revenues from new baseball stadium
Presidential campaign donations
Drug money laundering
Prostate and breast cancer screenings
Antitrust settlements
Company mergers
Sports statistics
Feature on cystic fibrosis sufferer's financial woes

We don't need advanced calculus or algebra or any of the more advanced math classes to survive as journalists. Just some good middle school math basics and the power of logical thinking. Check out this article by data journalist, Matt Waite, on how he faced his fear of math.

There are four areas we're going to look at today:
--Using the right numbers so that you make fair comparisons
--Helping readers understand numbers better
--Writing with numbers
--Avoiding "mutant statistics"

RATES:
Let's start with rates:
You may need to use rates when it's something that you've analyzed yourself or when somebody has dropped a pile of analysis in your lap in the form of a press release or report. Either way, you might need to decide whether a rate is the appropriate way to

express the information to your readers or viewers. Most often you will need to consult with an expert or other reliable source to find out the rate formulas that experts (or a particular industry) use.

There are 2 main reasons for using rates:
--To compare disparate groups
--To make very big or very small numbers comprehensible.

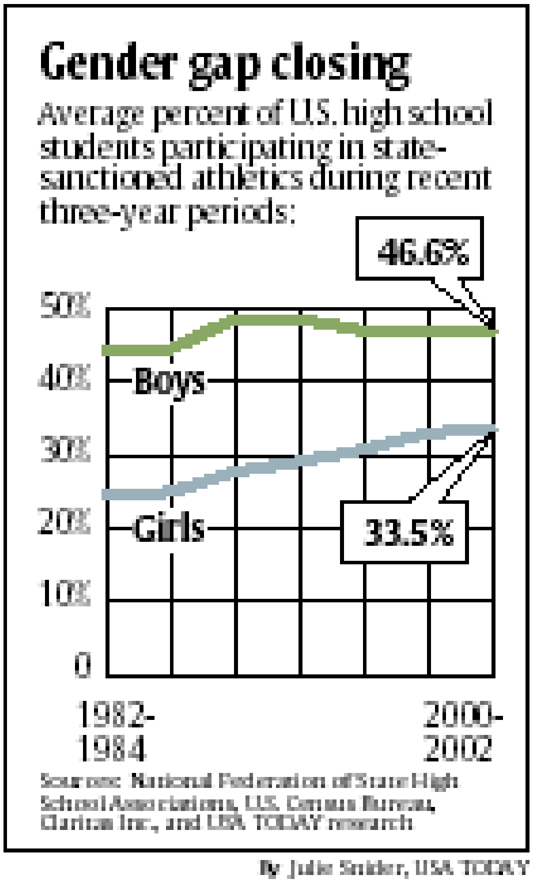
Rates can be a percentage or a ratio. Examples:
25 per 10,000
1 in a million
1 in 7
25%

Rates might be referred to as percentages, ratios, per capita rates
Examples where rates are necessary:
--Crimes in cities with different populations
--Deaths from various diseases (some diseases are more common in older people)
--Comparing deaths at different hospitals (some have more trauma patients)
--Number of births in different places (some cities have more younger people, others more older people)

Sometimes raw numbers are okay.

Murder is a good example. When talking about the total number of people killed in a given year, it would be okay to give the raw number because most people would agree that even 1 person killed is too many. But when comparing one city to another, a rate would be more appropriate because one city might have more people (more opportunities for deaths) than another.

Most of the times it depends on what your intention is with the numbers. Sports participation is an example here: we can say that 8.7 million boys and 8.3 million girls play sports in the US. In terms of accommodating all those kids – teams, coaches, equipment, etc., the total numbers are important. But if we want to see whether more kids are playing sports than not playing sports – and if that has changed over time – we have to use rates because the number of people in that age group changes from year to year.


Gender gap closing

Average percent of U.S. high school students participating in state-sanctioned athletics during recent three-year periods:

Source: National Federation of State High School Associations, U.S. Census Bureau, Claritas Inc., and USA TODAY research

By Julie Snider, USA TODAY

The National Federation of High School Sports Associations has put out a report every year about sports participation and, until recently, the press release every year praised the huge increases in the number of kids playing sports. Yes, it's true that there are more athletes, but there are also more kids because high schools have been feeling the "baby boomlet" When you adjust the participation numbers by population you find that the participation rate has leveled off in recent years, especially for boys. That's a very different picture than the one portrayed in the federation's press releases.

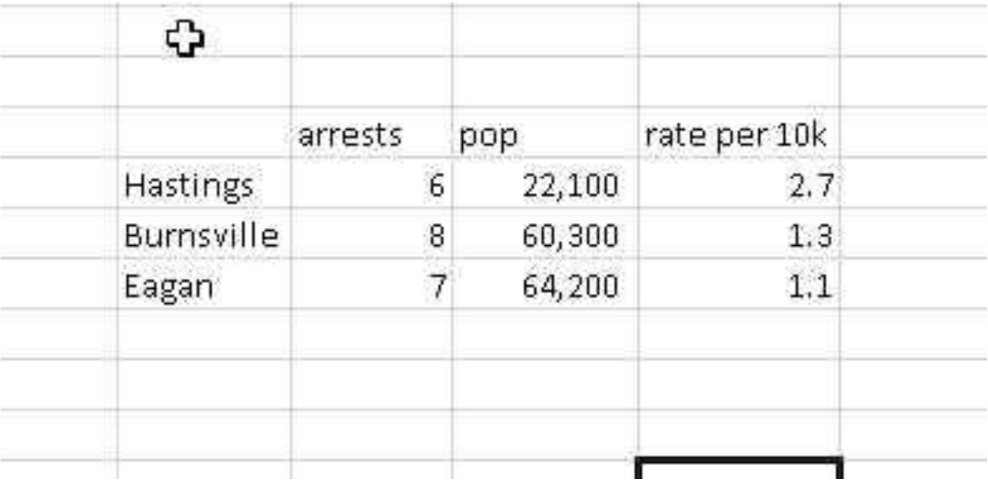
Another example:

Here's the exact wording from a story in the St. Paul Pioneer press:

"In Hastings, which has a population of about 22,100, there were six heroin arrests in 2011, according to a report from the drug task force. That's compared with eight arrests in Burnsville, a city with 60,300 residents. Eagan, a city of 64,200, had seven arrests."

That graph is almost incomprehensible. What are they trying to say?

What if we came up with a rate for each city, like this:



	arrests	pop	rate per 10k
Hastings	6	22,100	2.7
Burnsville	8	60,300	1.3
Eagan	7	64,200	1.1

Now we can say something like:

There were nearly 3 arrests per 10,000 people in Hastings, compared to about 1 arrest per 10,000 people in both Burnsville and Eagan, which are each nearly three times the size of Hastings.

Rates pull extreme numbers into reach:

Two examples from USA TODAY:

"[California] carried a relatively small amount of debt … Its $24.8 billion in debt in 2001 was equal to $733 per person compared with a national average of $820…."

Instead of just saying the state had $24.8 billion in debt – a number that is incomprehensible to almost everybody – the story calculated the per-person debt to come up with $733 per person. They went one step further and added context: how the state's per person rate compares to the national rate.

"America West and US Airways ….each had about 32 passengers of every 10,000 voluntarily give up their seats in the first quarter."

When you do the actual math on this: number of passengers giving up seats divided by total number of passengers served you come up with 0.0032 or 0.32%. Another number nobody can understand? So instead, the paper multiplied the rate by a factor – in this case they used 10,000. You can use any factor you want – 1 out of 1000, 1 out of 1 million, etc.

Not using rates when you're supposed to can be very misleading. The most widely-known example is about the "most stolen vehicles" list that gets put out each year by insurance industry groups.

Here's an example from a major news story:
"The Toyota Camry and Honda Accord, the best-selling cars in the USA recently, top the list of most stolen vehicles in 2000, as they have for several years."

If you just get the police data on stolen vehicles and tally up how many were stolen, the ones at the top will almost certainly be the most popular cars in general.

Recently – after much prodding from mathematicians and journalists refusing to publish these numbers – at least one group has started "adjusting" the thefts based on the number of registered vehicles of each type (unfortunately, the only "easy" way to get this data is by paying a small fortune to a company called Polk that collects it from all state DMVs)

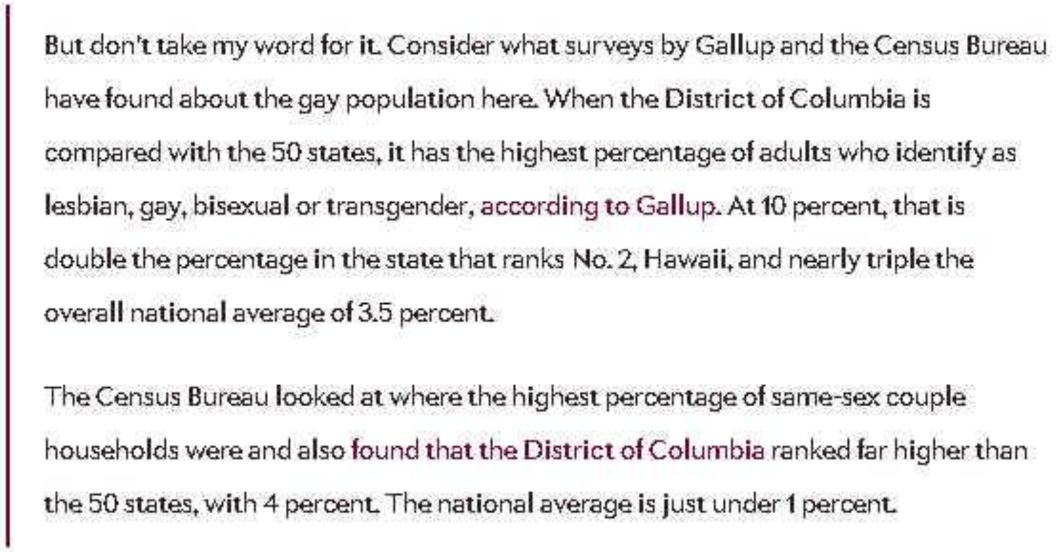
A more recent news story, using those rates, shows this:
"The 1995 Saturn SL was the nation's most-stolen vehicle last year based on thefts versus the number of models registered …. A new report shows …. One out of every 200 registered 1995 Saturn SLs was stolen in 2003, according to Chicago-based CCC Information Services, an insurance industry tracker…."

The examples I've given here are some of the more common ones that you'll encounter, but you'll certainly run into other cases.

Just remember to make sure you're using the right base or denominator. In the stolen car example, the base was the total number of cars of a particular make/model that were registered.

Some others that are tricky: when looking at accident or deaths rates on highways you have to use "vehicle miles traveled" which accounts for the number of drivers and distance driven. Also death rates, when comparing hospitals or comparing different diseases, get kind of tricky -- it's imperative to get guidance form an industry expert in those types of situations.

Another thing to keep in mind is to ensure you're comparing apples to apples. For example, this New York Times article from 2013, questioned the District of Columbia being dubbed the “gayest place in America”. It raised the question of whether DC is really comparable to states? Isn’t it more like a city?



> But don’t take my word for it. Consider what surveys by Gallup and the Census Bureau have found about the gay population here. When the District of Columbia is compared with the 50 states, it has the highest percentage of adults who identify as lesbian, gay, bisexual or transgender, according to Gallup. At 10 percent, that is double the percentage in the state that ranks No. 2, Hawaii, and nearly triple the overall national average of 3.5 percent.
>
> The Census Bureau looked at where the highest percentage of same-sex couple households were and also found that the District of Columbia ranked far higher than the 50 states, with 4 percent. The national average is just under 1 percent.

ADJUSTING FOR INFLATION:

Just like rates, you also need to think about adjusting for inflation when using currency figures. General rule of thumb is that you should adjust if you are comparing dollar values that are three or more years apart.

This might include salaries, city budgets, property values, prices of goods or services

How this works: You adjust the old number(s) into today’s dollars using the Consumer Price Index – this is a number that is generated on a monthly basis by the Bureau of Labor Statistics and can easily be found on the Web.

FORMULA:

(CPI Now/CPI Then) x Old value

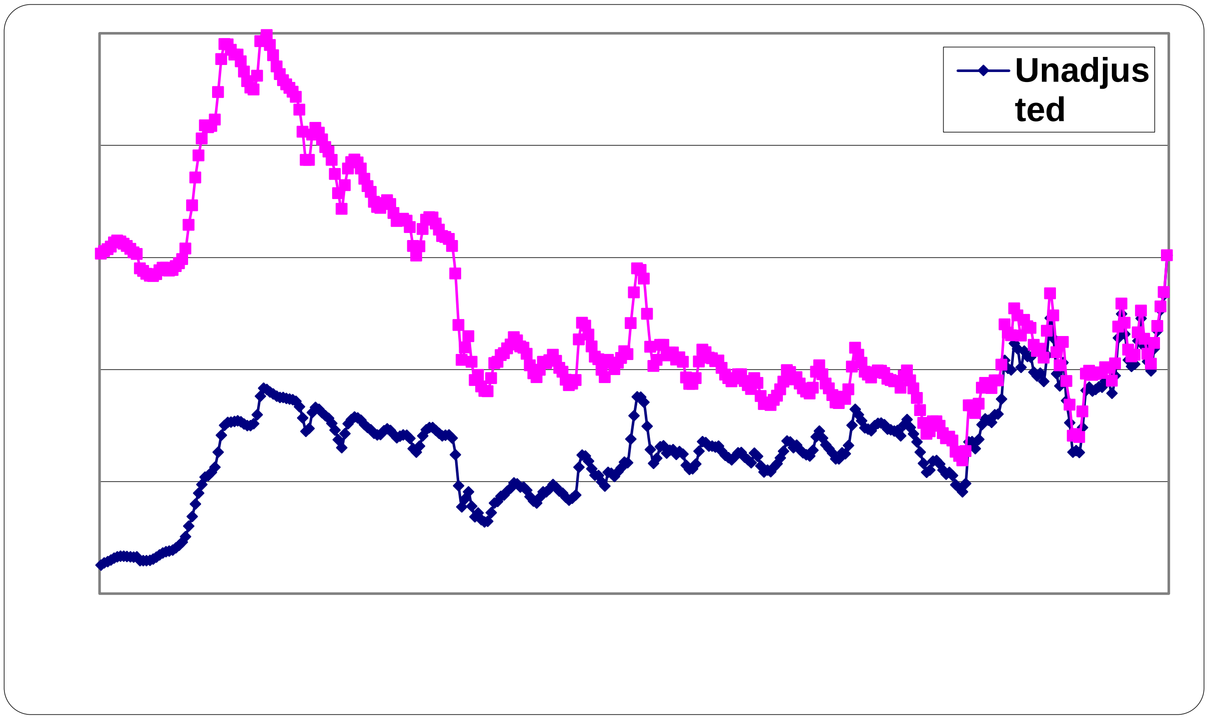
Example: Teachers made about $9,000 in 1970. By 1996 they made $38,000. Is that a big raise or a little one?

CPI Now (156.9)/ CPI Then (38.8) = 4.0438

It took about $4 in 1996 to buy what people bought for $1 back in 1970

Answer: 4.0438 x $9,000 = $36,394. That means their raise was only $1,600 after inflation

Gas prices are a good example of how a picture can be distorted if you don't use rates when you're supposed to.



This chart shows both unadjusted and adjusted gas prices from 1977 to 2004. In recent years, one of the big stories has been the "high" gas prices. Yes, they seem high compared to where we've been but when adjusting for inflation, today's prices look like nothing compared to the spike right around 1980. But it also shows that, yes, we had increases the last few years

AVERAGES AND MEDIANS:

This is another example of numbers being misused. The general public probably doesn't know what a "median" is, but they can grasp the idea of an average. Unfortunately, they are distinctly different when it comes to math.

The average (or mean) is the sum of values divided by the number of items.

The median is the middle value.

The big problem with an average is that it can be heavily influenced by “outliers” – one or two numbers that are extremely far above or below the rest of the numbers.

The general rule of thumb is to calculate both and if they are similar, you’re okay using the average (which is easier to explain to readers or viewers). If they are very different, you need to use the median.

Here’s a simple example. Imagine these are salaries of workers and you want to figure out what the “typical” salary is.

You have $40, $50, $45, $200, $250 and $50.

On the face of it, you would be inclined to say the “typical” worker makes something in the neighborhood of $50 – right?

If you calculate the average you come up with $105.
The median comes out at $50.

YARDSTICKS:

Once you’ve made sure you’re using the right numbers, you’ve got another hurdle: making sure your readers will understand them.

Very big numbers or very small numbers – as we saw with rates – can be very tricky. The other problem is making sure you don’t bombard your readers with too many numbers.

Let’s first talk about some other ways – besides rates – to help readers understand really big numbers.

That’s called using “yardsticks.” This means that you find a way to compare your number to something that people will readily understand.

For example: When the cleanup was going on at the World Trade Center in New York after 9/11, they talked about hauling away 3 billion pounds of debris.

You can’t even fathom what 3 billion pounds would be. I know that 1 pound is about how much ground beef I put in when I make tacos, but what would a billion look like?

So maybe we could convert it to how many tons? But would that be good enough? Do you know what a ton of debris looks like?

How about how many dump trucks?

Or how many train cars?

Is it enough to fill the Great Pyramids? Or the Grand Canyon?

Is it equivalent to a month's worth of NYC's garbage?

Make sure the analogy makes sense for your topic and your readers.

An article in the New Yorker magazine found a creative way to explain the quantity of available drinking water on earth to the quantity that is trapped in ways that we can't obtain. First they told us that only 3 percent of the earth's water is available for humans to drink.

Then they put it into a visual format:
"If a large bucket were to represent all the seawater on the planet, and a coffee cup the amount of freshwater frozen in glaciers, only a teaspoon would remain for us to drink."

Another approach is to use "compared to what?" …compare the number to some other total. In this example, the writer compared the total number of people killed in travel-related accidents to the total populations of certain communities.

"Travel in America claimed the lives of more than 44,000 people last year – roughly the population of Wilkes-Barre, Pa., Palatine, Ill., or Covina, Calif."

Something visual almost always helps: this example was used to explain the impact of the crash of 7,000 aircraft in the Netherlands during World War II – they showed that the whole state of New Jersey would have been affected.

"To put it another way, the crash of 7,000 aircraft would mean that every square mile of the state of New Jersey would have shaken to the impact of a downed plane."

In addition to thinking visually in words – to include in your story – also work with your graphics department to figure out if there's a way to display your quantitative information in graphics or maps that would help the reader visualize what you're talking about.

It also helps to always think about putting numbers into context.

--Use percentages. For example, instead of using points, say what percentage of the stock market increased.

--Give comparisons. For example, compare what's happening now to a point in the past, or compare one city to another, one player to another, etc.

--Adjust for inflation when necessary. A glaring example is box office receipts. You frequently hear about a movie bringing in the "most money ever" – but did they adjust for inflation? Look on Wikipedia for "highest-grossing films" and you'll find two lists – one showing "Avatar" made the most money ever with $2.7 billion worldwide. But then scroll down a bit and you'll see that, adjusted for inflation, the oldie-but-goodie, "Gone with the Wind" still holds the title with $3.3 billion in today's dollars. If you're talking about the shear reach of the movie, clearly more people must have watched "Gone with the Wind" in the theater, while "Avatar" just charged more money per ticket.

NUMBERS ARE NOT CONCRETE:

"The most important numerical fallacy is that people tend to think of numbers as known, constant and having no variability" said Donald Berry, a biostatistician at the University of Texas MD Anderson Cancer Center in Houston.

But think about where numbers come from. The majority that we, as journalists use, come from polls and surveys. Even census data, which we trust implicitly, is flawed. Every 10 years, the Census hires hundreds of thousands of people to knock on doors and hunt down homeless people in hopes of counting every head in the nation. They fail every time.

Nearly any dataset we use probably has flaws. Let's say you get the city's water billing records to see who is using the most water or what part of the city is using the most water – what if they accidentally have 2 records for the same person? Or what if they are missing a home or two?

This is why it's imperative for us to "characterize" numbers, rather than imply that they are super precise.

It's okay to round numbers (always up), avoid decimal places and to use words (like "one-third" or "half") to avoid using numeric digits. These kinds of practices help avoid giving the reader the impression that the numbers you are citing are scientifically accurate.

PRECISION:

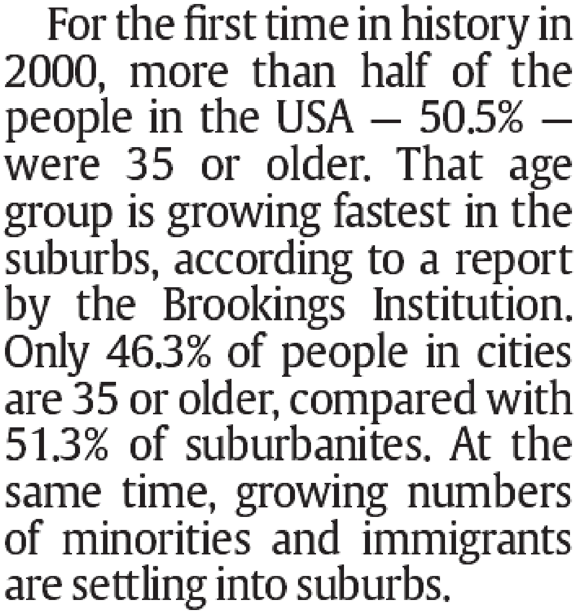
The most common error for journalists is to use too many decimal points – or too much precision. There are a couple problems with this:

Each decimal point is another number that the reader has to "trip" over as they are reading or listening to your story. Numbers slow your story down considerably – this is a scientific proven fact.

Precision usually isn’t necessary or might even be misleading.

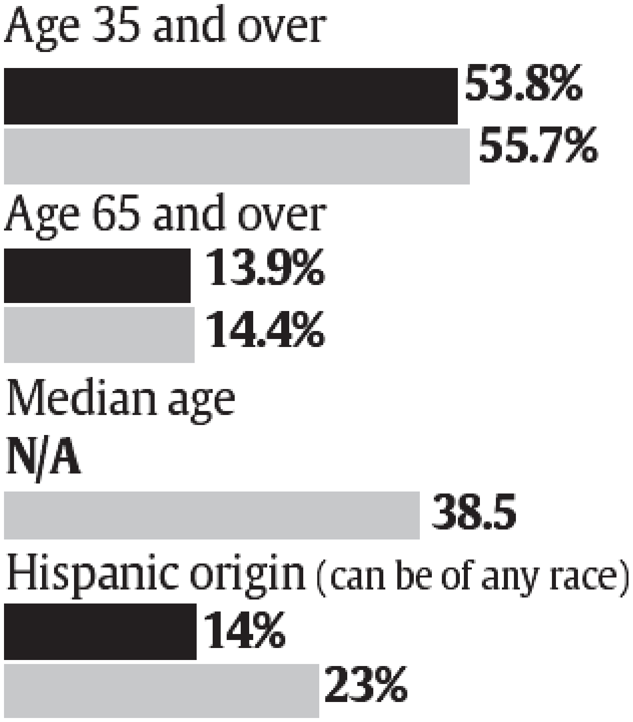
Let’s start with a couple examples:
The first is from USA TODAY and talks about findings from a Brookings Institution report on the aging population. Note how they says 50.5% were age 35 or older. Does that .5 percent really matter? Isn’t that the same as saying “half”?

Then it goes on to say 46.3% of people in cities are 35 or older and 51.3% of suburbanites – they are trying to compare cities versus suburbs. In this case, the decimal points don’t matter. It’s still 46% versus 51%. The reader isn’t going to care.



For the first time in history in 2000, more than half of the people in the USA — 50.5% — were 35 or older. That age group is growing fastest in the suburbs, according to a report by the Brookings Institution. Only 46.3% of people in cities are 35 or older, compared with 51.3% of suburbanites. At the same time, growing numbers of minorities and immigrants are settling into suburbs.

The next example is similar, showing bar graphs of age 35 and over, age 65 and over , median age and Hispanic origin. There are a couple problems here. The first is that they are inconsistent in when they use decimal points. The second goes back to the same problem in the previous example – does it really matter. Finally, is this really worth a graphic? The differences in the age 65 and over population are non-existent – they are both 14%.



The bigger issue underlying both of these examples is the source of the information to start with. There are two aspects to this point:

If the source is a survey, it's automatically flawed and therefore imprecise.

If the numerator and denominator used to calculate these percentages were whole numbers, it is also imprecise to start with.

In both those cases, using decimals implies a level of precision that is simply not there.

A lot of press releases and reports use decimal points to death. They want to prove that their work was rigorous and scientific. But do our readers really care? (okay, yes some of them do). But the vast majority do not.

Your job is to "characterize" the situation using numbers. In his book "Precision Journalism" Phil Meyer said "Decimal points are for meaning, not emphasis". So you need to use them ONLY when they impart true meaning.

TOO MANY NUMBERS:

The second most common numeracy error among journalists is to use too many numbers.

I've seen a couple different rules of thumb:

Keep the number of digits in a paragraph to no more than 8.

Use no more than 2 numbers in a paragraph.

I'm a big fan of really limiting the number of numbers in a story. I think of them more like quotes --- use only the best ones. Paraphrase when necessary.

When incorporating the results of my data analysis into a story, I also try to come up with a "star number." What's the one piece from my analysis that is most crucial to convey to my readers? What number really drives home our key point?

Here' an example from a Desert Sun story's "nut graph" where they did NOT use a star number:

"In a three-month investigation of water levels throughout the Coachella Valley, The Desert Sun found that the average depth of 70 existing wells across the valley in 1970 was 104.4 feet. As of this year, the average depth of 291 wells in the valley had dropped to 159.3 feet."

So what are they trying to say here?

Clearly they felt it necessary to explain that there are significantly more wells now than in the past – but does that need to be in the nut graph? Remember this is toward the top of the story, you're still trying to lure the readers into the story and get them hooked. You need to keep it simple and make them want to know more.

They did that in the next graph, but I think they could've eliminated a lot and just used this:

"The average loss of 55 feet of water depth reflects a significant depletion of the most precious resource in the California desert."

Use graphics to help provide the details of trends that you're writing about and then just use words in the story to say how that trend has changed over time.

"Sprinkle in numbers" like you do quotes. Remember basic reporting is that you should only use quotes where the person is saying something so profound or unique that you can't say it better in paraphrase — or it's something that legally you've got to put in quotes. Because of this rule, quotes should be few and far between in your story – how many sources do you know who sound "quotable" all the time?

So treat numbers the same way. Few and far between.

Example 1:

The "eyes glazed over" version: *"The Office of Redundancy's budget rose 48 percent in 2001, from $700.3 million to $1.03 million"*

A better way: *"Over the past year, the Office of Redundancy's budget grew by nearly half, to $1 billion."*

Source: "Numbers in the Newsroom" by Sarah Cohen

Example 2:
The "eyes glazed over" version: *"Statewide, lending institutions rejected 18.8 percent of black applicants, 15.7 percent of Hispanic applicants and 11.4 percent of white applicants for conventional mortgage loans in 2004, the latest year for which complete data are available….."* (source: South Florida Sun-Sentinel)

Notice there are 12 digits in this sentence, also they are just rattling off numbers. What are they really trying to convey to readers? The main point is that Black, and to a lesser extent, Hispanic, applicants are far more likely to be rejected than whites. So why not come up with a sentence that basically says that?

Good example: Murder by the numbers
"The oldest killer was 88; he murdered his wife. The youngest was 9; she stabbed her friend. The women were more than ***twice*** *as likely as men to murder a current spouse or lover. But once the romance was over, only the men killed their exes. The deadliest day was July 10, 2004, when eight people died in separate homicides. Five people eliminated a boss; 10 others murdered co-workers. Males who killed* ***favored*** *firearms, while women and girls chose knives* ***as often as*** *guns. More homicides occurred in Brooklyn than in any other borough. More happened on Saturday. And* ***roughly a third*** *are unsolved." (source: New York Times)*

There are lots and lots of numbers crammed into these first few grafs of the story but it's not as difficult to read as the previous example. I wouldn't necessarily recommend putting this many numbers in so few grafs, but this is an example of a particular style or voice that they were trying to achieve and the author does a really good job of mixing it up so that she isn't always using actual numeric digits.

MUTANT STATISTICS:
With the proliferation of sources of information – and lots of blogs and other publications copying other people's work – the problem of "mutant statistics" is even bigger now than ever before.

Mutant statistics occur when someone repeats a bad statistic that was already published. Or maybe they alter a good statistic to a point that it becomes wrong.

Let’s look at some examples….

The first is from the Pioneer Press. The reporter didn’t discover the error until a reader called to complain the next day and the newspaper ran a correction.

- *“The state received $2.5 million in federal funds to pay overtime costs for state troopers, sheriff’s deputies and city police officers throughout the state as part of the Highway Enforcement of Aggressive Traffic program….that should translate to about 1,400 hours of additional law enforcement during the next year…”*

The reporter’s biggest mistake was that he took numbers straight out of a press release and didn’t even do the math. If he had taken $2.5 million divided by 1,400 hours, he would have seen that this would result in those state troopers being paid $1,785 per hour. Clearly something is wrong with those numbers.

The second example is highlighted in the first chapter of a great book called “Damned Lies and Statistics”

- *“Every year since 1950, the number of American children gunned down has doubled”*

If we try to do the math on this, here’s what would happen:

Start with 1 child killed in 1950 and you “double” the number each year….

1951: 2 children
1952: 4 children
1965: 32,768 (total of 9,960 homicides that year)
1970: 1 million
1980: 1 billion (4x the total US population)

Turns out the reporter had messed with the wording of the original statistic.
It should have read: *“The number of children <u>killed each year</u> by guns has doubled since 1950”*
In other words, it has doubled between 1950 and current – not doubling every year.

The third example (marriage and terrorism) is one of the most famous out there. Back in 1986, Newsweek magazine and many other news outlets reported on new demographic research about marriage rates. According to the research, a woman who remained single at 30 had only a 20 percent chance of ever marrying. By 35, the probability dropped to 5 percent. Newsweek, however, inserted a phrase that shocked the world: *"a 40-year-old woman was more likely to be killed by a terrorist than to ever marry."*

The actual research didn't say that. It was a line that the writer dropped in as a sort of quip. It also turns out that the original research was put in doubt after this publication…. Other research showed the odds of never marrying were not nearly as dire as had been predicted.

- "Some demographers immediately doubted the dire odds. Within months Census researchers did their own study and concluded that a 40-year-old single woman really had a 17 to 23 percent probability of eventually marrying, not 2.6 percent." *(Newsweek, June 2006)*

A few years ago– the 20th anniversary of the famous Newsweek article – Newsweek magazine published an article talking about how it all came about and the fallout it caused. They also went on to show how society has changed in the last 20 years and what's happened with marriage rates.

CATCHING A BIG ONE:

Not all reporters get duped by bad stats. Paul Tosto, formerly the higher education reporter at the Pioneer Press, caught the University of Minnesota reporting very flawed statistics. It resulted in a front-page story.

The findings were part of a report by the Minnesota Commission on Out-of-School Time, a group launched by University of Minnesota President Robert Bruininks and funded by the U, State Education Department and McKnight Foundation. The panel called for more after-school opportunities and a new $12 million public-private fund among its recommendations.

The report said Minnesota had the highest percentage of children age 12 and older alone at home every afternoon – compared to other states. But when questioned about it, they couldn't come up with supporting data.

From the news story: "*The claims were eye-popping: Minnesota has the nation's highest percentage of teens home alone each afternoon. It has more young children taking care of themselves after school than any state in the country. Half its kids aren't part of any structured after-school activity."*

They also said 50 percent of young people were not participating in after-school programs. Turns out the researcher only went to 9 sites and talked to an unscientific sample of 101 kids. --- Not enough to base findings on.

And finally – the report said Minnesota had the highest rate of 10 to 12 years olds in “self care”. That came from a report by the Urban Institute. Minnesota was included in that report, but it was one of only 13 states. So you could say that Minnesota was the highest “among those 13 states” – but you certainly can’t say it’s the highest in the nation.