MISUSE OF STATISTICS

Author: Rahul Dodhia
Posted: May 25, 2007
Last Modified: October 15, 2007

This article is continuously updated. For the latest version, please go to
www.RavenAnalytics.com/articles.php

INTRODUCTION

Did you know that 54% of all statistics are made up on the spot?

Okay, you may not have fallen for that one, but there are plenty of real-life examples that bait the mind. For example, data from a 1988 census suggest that there is a high correlation between the number of churches and the number of violent crimes in US counties. The implied message from this correlation is that religion and crime are linked, and some would even use this to support the preposterous sounding hypothesis that religion causes crimes, or there is something in the nature of people that makes the two go together. That would be quite shocking, but alert statisticians would immediately point out that it is a spurious correlation. Counties with a large number of churches are likely to have large populations. And the larger the population, the larger the number of crimes.[1]

Statistical literacy is not a skill that is widely accepted as necessary in education. Therefore a lot of misuse of statistics is not intentional, just uninformed. But that does not mitigate its danger when misused. If we were to seek a positive slant on this, it means that there is some low-hanging fruit, some concepts that can be easily learnt so that one can have a deeper understanding of the myriad of reported statistics. The following sections point out some ways in which statistical literacy can be increased.

A PICTURE CAN MISLEAD YOU A THOUSAND WAYS

Graphics are a great way of communicating data, but a chart for a chart's sake is not always a good idea. Here are some common ways that charts confuse rather than elucidate.

WHEN CLARITY MAKES WAY FOR PRETTY

The following chart compares the return on investment for two mutual funds in successive years. To someone glancing at the chart, it would appear that Fund B outperformed Fund A slightly in three years.

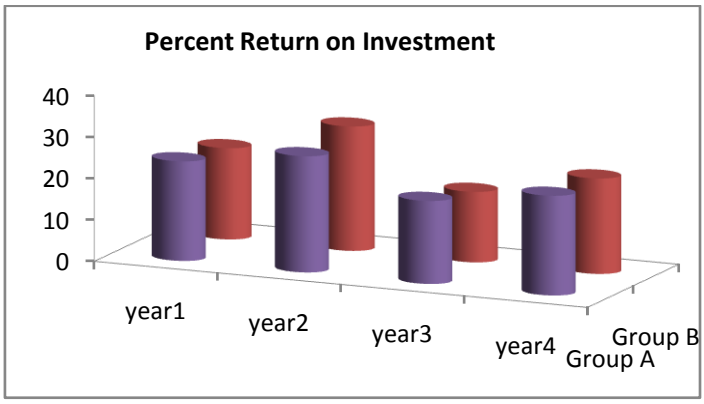

FIGURE 1

Here is the same data in more conventional but less pretty format. Now it is clear that Fund A outperformed Fund B in 3 out of 4 years, not the other way around.

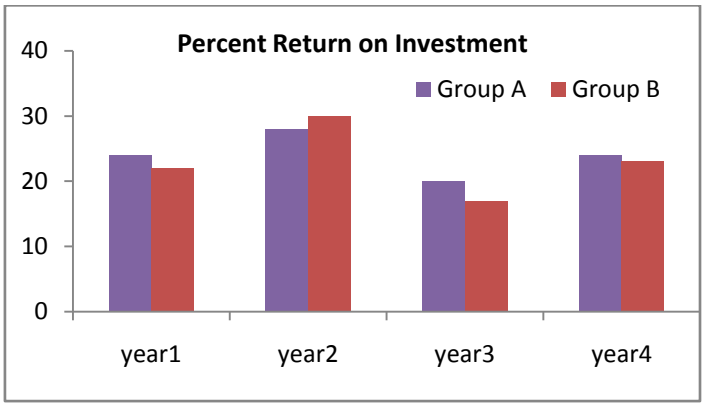

FIGURE 2

Three dimensional depictions of data are hard to perceive clearly, and the general rule of thumb is to only use as many dimensions as the data one is trying to show. In the example here, there are only two ordered dimensions: year and percent return. The third factor, Fund type, is unordered. Therefore only two dimensions should be used.

MISLEADING SCALE

The data from a poll question asking respondents which party would win in the presidential elections showed the following percentage responses.

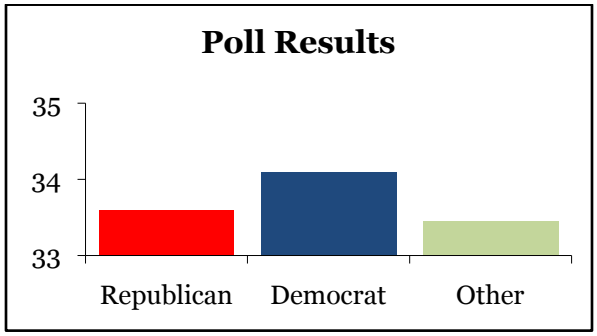

FIGURE 3

The democrats appear to have an advantage over the other two groups. But take a look at the scale on the vertical axis. It has a very narrow range, and one could argue that this scale exaggerates the differences among the three groups. Now take a look at the following graph which extends the scale to 0.

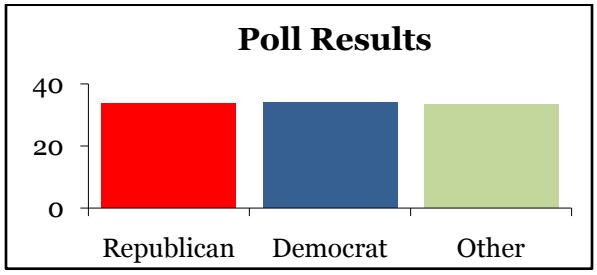

FIGURE 4

It would appear that there's a dead heat among all three groups. Which is the correct graph to show? I would argue the second graph is better, since it minimizes, probably correctly, the differences between the three groups. In general, it is better to show the end points of the scale, or at least one of them.

Another example of how the vertical scale can influence conclusions is shown in this graph by CNN: http://mediamatters.org/items/200503220005 .

LACK OF STATISTICAL COMPARISON

In the previous example, you would not have to be a statistician to suspect that the results are far from conclusive that Democrats would win. All we're doing is comparing means, and we have no idea how reliable the poll results are. What we need, and what a statistician would insist upon, is a measure of uncertainty. By adding standard error bars, the following graph provides more evidence that no group is really ahead of any of the others.

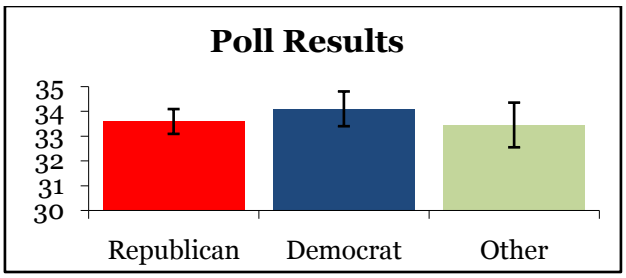

FIGURE 5

In this graph, the average response for each group is accentuated by the standard error bar. It provides evidence that the differences among the groups are likely due to chance. Another poll somewhere else or at a different time might show the Republicans leading the Democrats.

LEARNINGS

1. Avoid 3-D charts. If you have 2-dimensions to display, use a 2-d chart. Use 3-d charts if you have 3-dimensions to display, and then only if the third dimension can't be incorporated into a 2-d graph.
2. Have a scale that's relevant to the values being shown. Do not leave off the ends of the measured scale if they are not infinite. At least have one end of the scale to anchor the measurement.
3. The purpose of charts is usually to compare trends or general magnitudes rather than provide precise data points. Use a table to show precise data points.
4. Means by themselves do not always lend themselves to meaningful comparison. A measure of variability is also needed, such as error bars on a graph. This point is explored further in the next section.

HANDLE AVERAGES WITH CAUTION

NEED FOR STANDARD ERROR

The practice of statistics grew out of a need to make sense of inadequate data. The most basic comparison of all, comparing two magnitudes, is usually not enough. The polling example in the last section showed how standard errors lend weight to a comparison between means. Let's look at another example that illustrates the inadequacy of means.

A company wants to measure the impact of its customer service on its customers spending habits. One of the issues they are interested in is whether customers are more likely to increase their spending after receiving phone support rather than email support.

When the results are plotted as histograms, it looks as if phone outperforms email. The horizontal axis shows the average dollar increase in spending per customer after

receiving customer support, and the vertical axis shows indicates the number of customers. Since the blue graph is on the right, it looks clear that phone support almost always resulted in better outcomes than email support. The average of the blue graph is around $5 and the average of the red graph is about $3. The spreads of the two graphs hardly overlap, meaning that the highest email support customers increased their purchases only as much as the lowest phone support customers.

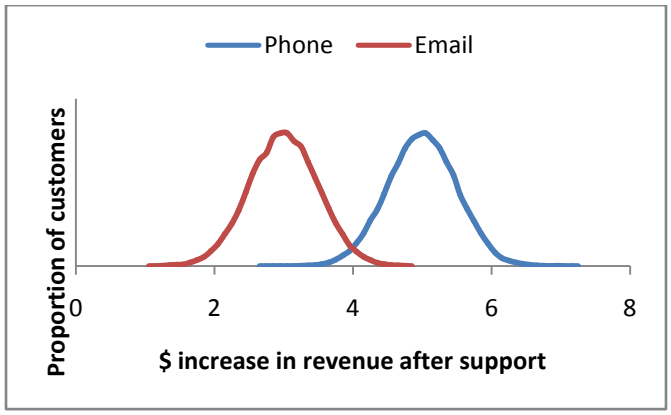

FIGURE 6

Now what if the averages remain the same, but the spreads are much wider?

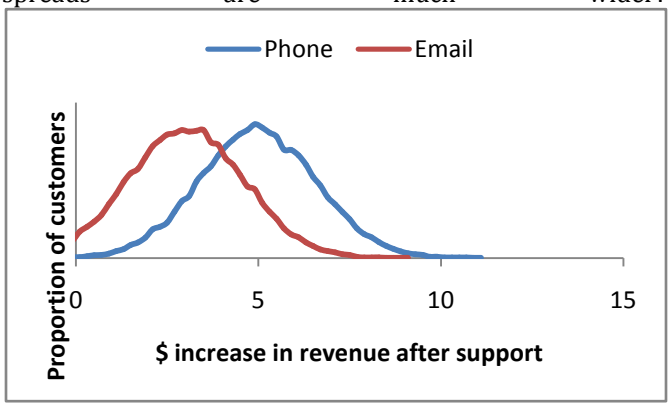

FIGURE 7

In both situations, by comparing only the averages, you might come to the same conclusion, that phone support is better. But comparing all the data and looking at the spread in the second graph, you see that the effects of phone and email support overlap a lot. In fact, they overlap so much that a statistician would say that phone support does not result in significantly greater customer spending, or that email support does not result in significantly less customer spending. A statistician can even compute roughly how likely the results from phone support are better than the results from email support– in this example it's only a 50% chance – even odds.

Here is another example when the standard error is necessary when comparing means. [2] This graphic appeared in an Ohio newspaper in 2000:

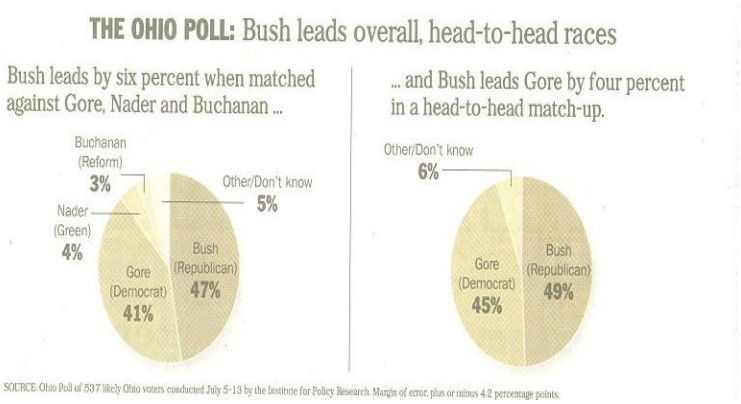

FIGURE 8

The results suggest that Bush is ahead of Gore whether or not other candidates are considered. However, the fine print under the charts gives some interesting details: the margin of error is ±4.2%. Therefore, in the second graph, we can say that Gore's support is really between 40.8% and 49.2%, while Bush's is between 44.8% and 53.2%. Bush's lead is not statistically significant, and this result suggests that it wouldn't be surprising to see another poll that showed Gore was ahead.

MEDIANS INSTEAD

Let's say you do a survey of the house values in your neighborhood, You compute the mean property value and find it is $492,000. With your own house valued at $463,000, you might feel a little depressed. But then your wife comes over and looks at your data. She points that the two houses on the hill are skewing your results, they are both worth over a million dollars and so they are pulling all the numbers north. When you remove those houses, and a couple of other peskily high-valued houses from the average calculation, the average drops to $465,000.

The problem with means is that they can be easily skewed by extreme high or low values. To get at a better representation of average value, use the median. The median house value would likely have remained around $460,000 with or without the high-valued properties. Whereas a mean can mislead, a median gives readers stronger assurance that the value is a good representation of the average. A median is preferable to calculating a mean with outliers removed because censoring data is a questionable step to take.

INTERVAL SCALES

A random group of people were asked to rank their favorite beverage, with 1 being the best. The results, when averaged across several respondents, came out as follows:

Beverage	Average Rank
Coca Cola	1.6

Pepsi	3.3
Dr. Pepper	4.4
7up	4.9
Ginger Ale	5.2

Coke proponents hailed this survey as proof that coke was twice as popular as pepsi, since pepsi's rank was twice as low as coke's. By the same reasoning, coke would be more than 3 times as popular as ginger ale.

Why is this not a correct conclusion to make? The flaw lies in the scale being used; the rankings from 1 to 5 were artificially induced. A respondent who gave coke and pepsi ranks of 1 and 2 respectively, but who liked Coke only marginally better than Pepsi, was forced to say he liked Coke twice as much as Pepsi – not a true divination of his preferences. One might argue that it washes out because there will be some other person who likes Coke more than 4 times as much as Pepsi, but will be forced to say the same thing, that she likes Coke only twice as much. But how representative of the population is this situation likely to be? There will probably not be enough of each type to average out, and so the results will not show the correct relative preferences.

Going from subjective preferences to objective data is never error free, but we try to remove as much bias as we can in order to do meaningful statistical comparisons. The difficulties in the example above can be mitigated by giving respondents a different way to answer. Maybe rate each beverage individually on a scale from 1 to 10?

LEARNINGS

1. Comparing only means may lead to conclusions that are not reliable. Adding standard errors to the comparison provides some idea of how reliable the difference in means actually is.
2. The proper choice of a summary statistic is not always the mean. Medians are often a more informative alternative.
3. Be aware of the scale used to measure data – is it artificial or real? If artificial, do the intervals from one value to another always mean the same thing?

SPURIOUS CORRELATIONS

Some remarkable claims are made by studies which get reported widely in the news

Cappuccino makers in home linked to healthier babies.[3]

Hopefully you had a twinge of suspicion when you read that. Even caffeine-addicted parents should follow that suspicion and think about why cappuccino makers in a home would be correlated with healthier babies. It is perfectly natural to find some a causal reason for why correlations exist, and the tendency is to find a positive explanation. Maybe something about coffee helps with the immune system? But a more sensible explanation is only a thought away: homes with cappuccino makers are likely to be wealthy, and undoubtedly a lot of that wealth is used to ensure their children's wellbeing. So yes, cappuccino makers and healthy babies are linked, but probably not in any meaningful, direct way.

Almost any day, you can open a major newspaper or magazine and find a study or a poll that has some remarkable conclusions. Some of these studies made headlines around the world. For example:

> Hormone Replacement Therapy helps prevent heart disease.[4]

This last statement was shown to be seriously flawed, and that HRT in fact increased the chances of heart disease. In some early studies, it appeared that HRT reduced the risk of heart disease. The studies were later realized to be flawed, and new studies showed that HRT was of no help, and maybe even increased the risk of heart disease.

Let's make a distinction between spurious correlations and mistaking correlation for causation. Spurious correlations occur when two visible variables are not really linked to each other, but are both linked to a third, hidden variable. In the case of the babies and the cappuccino makers, the hidden variable was the wealth of the household.

MISTAKING CORRELATION WITH CAUSALITY:

Showing that two factors are correlated sometimes inaccurately gets describes as A causes B. For example,

> Smoking reduces college achievement[5]

Opponents of smoking would take this as some vindication of their stand. But even when a reported correlation agrees with your belief, cultivate some skepticism. Could we just as well have said that low achievement drives students to smoke? Or, what third variable may explain this relationship? (Hint: partiers are less likely to study, more likely to be in social smoking situations.)

More often than not, mistaking correlation with causality is done innocently, when a writer makes a leap from the statistical facts to his or her own preconceptions. A causation link should only be put forward if there is reasonable evidence for it, and no compelling reasons against it. For example, the causal link between MMR vaccinations and autism can be considered just a correlation. A 1998 study spurred concern and controversy by linking MMR vaccines to autism, but was

retracted by most of the study's authors in 2004.[6] Children are identified with autism around the same age as when they are receiving vaccinations, and so the link may go both ways. Reasons against a causal link from vaccinations to autism include several scientific studies that did not find a link. Evidence for the causal link remains at the correlational level. Therefore, the CDC still maintains no causal connection between the two.

LEARNINGS

1. Do not take reported relationships at their face value, especially if you cannot see a direct, causal link between them. There may be a common-sense reason why they are linked through a third, unreported variable.
2. Do not confuse correlation (two variables seeming to have a relationship) with causation (one variable causes the other to change).
3. One way of figuring out that a causal link does not exist is to ask yourself if the "effect" may not in fact cause the "cause".
4. Beware of reports that give a causal explanation based on just one study.
5. Statistical analyses can never give a causal answer.
6. If you have access to the study, even a quick examination of the survey questions or study methodology may reveal that the correlation may be flawed.

STATISTICAL SIGNIFICANCE, BUT PRACTICAL INSIGNIFICANCE

Students of statistics courses often come away with the notion that attaining statistical significance is the paramount goal of statistical testing. That is true a lot of the time; the goal is indeed to show statistically significant changes in metrics or statistically significant differences between groups. But like so much in statistics, results should be taken with a healthy helping of common sense.

A company marketing a new brand wanted to increase the number of new customers, and wanted to know whether to it would be more cost effective to obtain them through affiliates or directly through their own advertising campaigns. Analyses of their customer acquisition channels showed that each customer cost $30 on average to acquire through affiliates, and $23 through advertising. The difference was $7 ± $2.20, a 95% confidence interval of ($4.80, $9.20). Since each customer was assumed to have the same lifetime value no matter how he or she was acquired, revenue considerations required the difference to be at least $10 for direct advertising to be worthwhile. So although this difference in acquisition costs between the two approaches was statistically significant, it wasn't large enough to justify the higher of affiliate incentive. It was not of *practical significance*.

A few years ago, a widely reported study of breast cancer and low fat diets stated that the difference in cancer returning between the control group and the low-fat diet group was 25%, a statistically significant difference.That is, people with low-fat diets had a 25% less chance of cancer returning. A closer look at the results showed they were not as salutary as they appeared. The actual return rates were 12.4% and 9.8% respectively, a difference of 2.6%. (The writer of the original report got 25% by dividing 9.8 by 2.6). So is this statistically significant difference of practical significance? A lot of people would say no, but maybe someone facing the risk of cancer returning would grasp this is definitely practically interesting. Other criticisms of this result are given in the next section.

William Kruskal, an eminent statistician long at the University of Chicago, an editor of the International Encyclopedia of the Social Sciences, and a former president of the American Statistical Association had this to say when asked "Why did significance testing get so badly mixed up, even in the hands of professional statisticians?" "Well," said Kruskal, who long ago had published in the Encyclopedia a devastating survey on "significance" in theory and practice[7], "I guess it's a cheap way to get marketable results." [8]

LEARNINGS

1. A statistically significant result is not an end in itself. Take a look at the actual numerical difference and decide whether or not it is of practical interest.
2. In traditional hypothesis testing, a result can be made statistically significant simply by increasing the number of subjects or sampled data points. Beware of results with small effects and large sample sizes.

LOOK UNDER THE HOOD

ERRONEOUS VERBAL CONCLUSIONS

Scientists and politicians are both known for decrying the misquotes in the media to which they are subjected. In the case of scientists, it occurs when the carefully structured world of jargon and assumptions is transmitted to the everyday world, with a lot lost in the translation. To be fair to journalists, sometimes the scientists help the hype, the publicity will result in more research funding.

The following example shows how results can get mistranslated. As an example of a scientific study that received much world press, and consequently much criticism for the erroneous reporting of its results, take a look at the study in 2005 that claimed that a low-fat diet reduces the risk of breast cancer returning[9]. This is an encouraging finding, and was grabbed on by health advocates everywhere. The newspaper reports made one

think that taking fat out of one's diet would drastically reduce the risk of getting breast cancer. Even a slightly skeptical reader would suspect that there are some qualifiers to this statement. In fact, if they were to go through the actual data, they might want to reconsider the impact of these results.

In the previous section, we saw that the size of the effect was reported as a 25% reduction in risk in women with low-fat diets. This reduction was not actually as dramatic as it sounded, since the actual difference in breast cancer reductions between the control and low-fat diet groups was 2.6%.

Then, consider the sample - the study was restricted to a particular subset of women – women who had had surgery and certain drugs. This brings up questions about the generalizability of the results to all women. Further details emerged that even within the sample, the difference was observed in only a small subset of women. The women in whom a significant reduction in tumors was observed were women whose tumors were not helped to regrow by estrogen, and they constituted only about 20% of the study sample.

Finally, we wonder whether this may not be a case of spurious correlation. The low-fat diet group lost on average 4 pounds, and lower weight has been linked to lower breast cancer risk. Exercise too has a similar link. So could exercise and lower weight be more direct links to breast cancer reduction? In consideration with other studies[10], the chances are that there is little evidence for a causal link between a low-fat diet and lower risk of cancer returning.

Why did the media make a big fuss about these results? We may put it down to drawing erroneous verbal conclusions from a misunderstanding of statistics and the research design.

Another example in which many of those affected have argued statistics lead to erroneous conclusions is the annual US News & World Report's ranking of colleges. These rankings are so well-known that they presumably drive a lot of applications to colleges ranked favorably. The impression one takes away from the rankings is that the top-ranked schools are the best places an undergraduate can prosper and gain a solid foothold into the post-college world. The typical reader assumes they are fine in taking the rankings at face value as an objective ranking of the best colleges in the US.

But what about the fine print? The rankings are a result of combining many factors together, and they are only as objective and comparable as the factors that go into making them. If one were to change the criteria that go into making these rankings, the rankings themselves may well change. This actually happened in 1999 when a new statistician at US News & World Report implemented a new methodology. However, the methodology was changed back and the old rankings reasserted themselves.[11] Many colleges feel that these factors are not valid and objective for various reasons – one of the criticisms is that the weightings used in the formula are changed every year without any empirical support, ensuring that the same schools appear at the top. In 2007 about 80 colleges resolved not to participate in the surveys that feed the rankings (none of the 80 comprise the report's very top colleges).[12]

So although it is tempting to take this surveys at face value, it is important not to be lulled into making erroneous verbal conclusions. The underlying methodology may not provide the objectivity that is assumed.

WHO'S PAYING FOR IT?

Research articles in medical journals often carry notes about possible conflicts of interest that the authors may have. This is important to maintain the impartiality of scientific research, and not to have certain results promoted because of financial interests. You would not want to buy drugs that are on the market because the researchers who validated these drugs have a financial stake in the companies manufacturing them. Just as you would be suspicious of a report stating that outsourcing helps grow America's economy, and finding that report was contracted by multinational corporations who need to cut costs.

Some studies make a big splash in the media without the interests of the study authors being fully revealed. A recent study made headlines by claiming that children who went to pre-school were more likely to be aggressive than children who stayed at home. This persuaded a lot of parents to keep their children at home.

The study failed to mention that children at this age would display the same behavior in any social situation. Even kids who stay at home, but meet other toddlers in other social situations would display the same behavior. "Aggressive" was defined as stealing toys, pushing others and starting fights.

The study was funded by a mother support group, who had an interest in getting the result that they did. A study done by another group found that toddlers who stayed at home in fact ended up being more aggressive later in life. But as of now, I don't know what affiliation this other group has, and their agenda, if any.

GUESSTIMATES

The news abounds with all sorts of numbers –

3 million illegal aliens cross over into the US annually.
$250,000,000 lost in productivity due to fantasy football.
The GDP of the United States was $13 trillion in 2006
Obesity kills 400,000 Americans annually

How do these numbers come about, Are there people assiduously counting and assessing, making thoughtful judgments about each one? It would be nice to think so, but reality is probably not as comforting. For example, the statistic of 3 million illegal aliens coming into the US appeared in a Time magazine story and was reproduced without question in other stories. Since we don't catch all these illegal aliens, the number must be an estimate – how would this statistic have been estimated? Professor Paulos in his ABC News column, "Who's Counting", explained it nicely.[13] This particular number started with border agents stating that they catch about 1 million people annually – then they guesstimated that for every person they catch, about 3 make it over safely. Thus we get 3 million. Since have no idea how accurate the 1:3 odds of making it across the border are, and the border agents did not give any rationale for it other than a hunch, the 3 million figure could be wildly off.

Then, since many people try multiple times to get across the border, 1 million apprehensions does not mean 1 million unique persons. We have no real way of knowing how many fewer it really is, but 1 million is a liberal upper bound. But whatever the machinations behind these estimates, the figure of 3 million illegal aliens is likely to stay and be repeated by news sources.

Another quick example is the $250 million lost because of fantasy football. The figure was reported by National Public Radio, and it makes one wonder how that was calculated. Various other news sources have given other estimates, with the NPR figure being among the lowest. Forbes magazine[14] reported $7.6 *billion*. The wide range of estimates should give one pause, and is evidence that different assumptions lead to very different estimates for what is supposedly the same quantity. The assumptions may include

A final example is the revelation that 400,000 Americans die of obesity every year[15]. The number was originally reported as 300,000 and it arose from a study published by Professor David Allison in 1999 in the Journal of the American Medical Association (JAMA). He had received funding from a large number of weight-loss companies like Jenny Craig, Weight-watchers, and pharmaceutical companies that manufactured weight-loss pills (see the previous section "Who's Paying For It"). In 2004, the Centers for Disease Control followed Professor Allison's methodology and upped the figure to 400,000. One of the consequences of this move was that a number of lawsuits were flung against restaurants for causing obesity. A few skeptics of this figure, including scientists at the CDC itself, were not loud enough to be hard over the storm.

After a lot of media furor, the statistical methodology in that particular report by the CDC was shown to be flawed and the data used was incomplete. The CDC rectified it by publishing another report that reduced the number to 112,000 deaths a year,[16] a dramatic difference.

LEARNINGS

1. A large round number is likely to be a an estimate based on a number of guesses. It takes just one guess to be off for the whole estimate to be wrong.
2. Look at who's providing the number – they may be ratcheting it up or down to suit their own agenda.
3. It's not always easy to get at the data that was used to provide the guesstimate, but if you can, see if the extrapolation makes sense to you.
4. Just because the number was printed by a well-known news source does not mean it's correct. Do not be overcome by the halo effect – just because a news source is widely read does not mean all its facts are correct.

Rahul Dodhia is the owner of Raven Analytics. Send comments to Rahul@RavenAnalytics.com

For the latest version of this article, please go to www.RavenAnalytics.com/articles.php

REFERENCES

1 http://www2.selu.edu/Academics/Education/EDF600/Mod11/sld030.htm

2 http://www.jcu.edu/math/faculty/DJH/poll.htm

3David Freedman, Robert Pisani, Roger Purves, and Ani Adhikari (1980). Statistics(2nd Ed.). p. 137.

4 http://www.nlm.nih.gov/medlineplus/ency/article/007111.htm

5 Brief report: relationships of cigarette smoking to academic achievement, cognitive abilities, and attitudes toward authority. Multivariate Behavioral Research. 1968, Vol. 3, No. 4, Pages 513-517. Dj Veldman and OH Brown.

6 S. Murch, A. Anthony, D. Casson, M. Malik, M. Berelowitz, A. Dhillon, M. Thomson, A. Valentine, S. Davies, J. Walker-Smith.. Retraction of an interpretation. The Lancet, Volume 363, Issue 9411, Pages 750-750

7 Kruskal, William S. 196a. "Tests of Statistical Significance." Pp. 238-250, in David Sills, ed., International Encyclopedia of the Social Sciences 14. New York: MacMillan.

8 CHANCE News 13.01, Jan 5, 2004 to Feb 5, 2004.
http://www.dartmouth.edu/~chance/chance_news/recent_news/chance_news_13.01.html

9 Gina Kolata and Lawrence K. Altman. Study of Breast Cancer Patients Finds Benefit in Low-Fat Diet. New York Times, May 17, 2005.

10 http://www.cnn.com/HEALTH/9801/11/fat.breast.cancer/

11 http://en.wikipedia.org/wiki/College_rankings#U.S._News_.26_World_Report_College_and_University_rankings

12 http://en.wikipedia.org/wiki/U.S._News_and_World_Report#Criticism_of_college_rankings

13 http://abcnews.go.com/Technology/WhosCounting/story?id=300038&page=1

14 http://www.forbes.com/leadership/2007/09/26/fantasy-football-office-lead-cx_tvr_0926productivity.html

15 http://www.obesitymyths.com/myth2.9.htm

16 http://www.news-medical.net/?id=9348