

Misinterpretation of Statistics – An Introduction

John Salter, Cerafor, United Kingdom

ABSTRACT

Some of the basic tenets of statistics are often misunderstood and in some cases, deliberately misused to lend weight to false assertions; the author seeks to rectify this with a review of some commonly used techniques and the misconceptions that are often applied to them. Examples include correlation versus causality; incidence with respect to prevalence; an interesting paradox surrounding the likelihood of sharing birthdays (and its clinical relevance) and the veridical paradox explained using Bayes Theorem - otherwise known as the Monty Hall problem.

INTRODUCTION

An amusing quote from the American humorist Evan Esar is “Statistics is the science of producing unreliable facts from reliable figures”. A less amusing - but nonetheless powerful - quote is “Results of the study showed that even students with formal instruction in statistics [at the undergraduate and graduate levels] continue to demonstrate misconceptions”, taken from Hirsch and O'Donnell's paper on identifying and assessing misconceptions in statistical reasoning.

This paper seeks to examine some common statistical theory and how the interpretation thereof can be misleading to the untrained eye, firstly by examining some commonly misconstrued statistical theory and then following up with some statistical ‘quirks’ or paradoxes to illustrate the idea that ‘adults have intuitions about probability and statistics that, in many cases, are at odds with accepted theory’.

COMMONLY MISCONSTRUED STATISTICAL METHODS

PERCENTAGE CHANGES

On April 21st 2012, the Daily Mail reported the pricing differences in Amazon Kindles™ between the UK and the US, taking this as an opportunity to show “yet more evidence of how international brands appear to see Britain as an easy target for inflated prices”. The article goes further detailing the discrepancy between prices based on currency conversions to highlight their headline that “it's 70% cheaper in the US”.

Except that it isn't. Let us examine the numbers quoted in the report:

	UK PRICE	US PRICE	Difference
Kindle	£109	£65.72	£43.28
Kindle Touch	£169	£98.05	£70.95

The report states the Kindle is 65.8% more expensive (“mark-up”) and the Kindle Touch “equates to a rise of 72.3%”. Now, these two statements are absolutely correct; but a 72.3% rise does not always equate to a 72.3% decrease (as implied by the title). The misinterpretation of the percentage change misleads the reader further – note that the true “percentage cheaper” for the Kindle and Kindle Touch is 39.7% and 42.0% respectively. Perhaps still evidence of an easy target for inflated brands, but not half (or to be precise 56%) as evidential as their initial presumption.

When calculating a percentage change, one must always be aware of the *source from which one calculates the change*. These may seem self-evident, but – in the clinical arena – we often consider percentage changes from baseline. Taking this into account, however, percentage change from baseline can be statistically inefficient; it does not always correct for imbalance between treatment groups at baseline. Furthermore, it may also create a non-normally distributed statistic from normally distributed data. A more accurate method of analysis would be use of analysis of covariance (ANCOVA) to calculate confidence intervals and a robust significance test; the results produced could then be converted to a percentage change using the relevant mean baseline and post-baseline scores.

(A post-script...the Daily Mail has since updated their on-line article to remove the errant headline)

PhUSE 2012

STATISTICAL SIGNIFICANCE AND CLINICAL SIGNIFICANCE

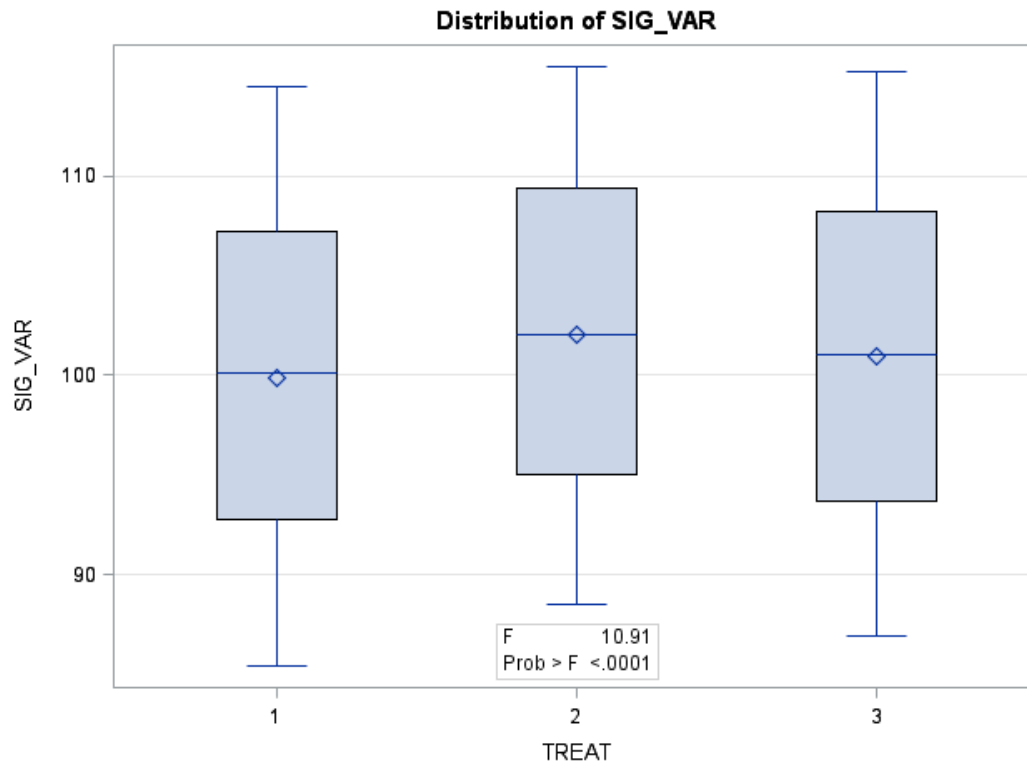
Statistical Significance is a statistical assessment of whether observations reflect a pattern, rather than just chance; the fundamental challenge being that any partial picture is subject to observational error. A result is deemed 'statistically significant' if it is unlikely to have occurred by chance, and hence provides enough evidence to reject the hypothesis of no effect.

Clinical Significance answers the question, how effective is the intervention or treatment, or how much change does the treatment cause? In terms of testing clinical treatments, practical significance optimally yields quantified information about the importance of a finding.

The misconception lies where people assume that statistical significance implies clinical significance. Consider a range of data that provides the following ANOVA table:

SOURCE	DF	SS	MS	F-value	Pr > F
Model	2	1467.3812	733.6906	10.91	<0.0001
Error	1797	120862.8415	67.2581		
Corrected Total	1799	122330.2227			

Although we have a highly significant p-value, one needs to consider the data. The following box-plot illustrates the data groupings:



One can see the data are similar; in fact, if one were to assume this data were Diastolic Blood Pressures, one could almost certainly state there is no clinical significance in the statistically significant result.

These differences that are small in magnitude (compared to the average result for a population) typically lack practical relevance and are unlikely to be clinically significant. Differences that are common in the population are also unlikely to be clinically significant, because they may simply reflect a level of normal variation. Additionally, clinicians should seek information in the assessment data and the patient history that may corroborate the relevance of the statistical difference, in order to establish the connection between performance on the specific test and the patient's general function.

CORRELATION AND CAUSALITY

In simplest terms, correlation is a statistical term that indicates whether two variables move together; for example, when it is snowing, it tends to be cold outside. The conventional dictum “correlations does not imply causation” means that correlation alone cannot be used to infer a causal relationship between two variables; it certainly should not be taken to mean that correlations cannot indicate the potential existence of causal relationships.

However, the causes underlying the correlation, if any, may be indirect and unknown, and high correlations also overlap with confounding factors where no causal process exists. Consequently, establishing a correlation between two variables is not a sufficient condition to establish a causal relationship. A correlation can be taken as evidence for a possible causal relationship, but cannot indicate what the causal relationship, if any, might be.

A correlation between age and height in children is fairly causally transparent (to a degree), but a correlation between various Quality of Life factors is less so.

Correlations and causalities can be grouped into four distinct categories (further categories may exist as combinations of these):

1. X can cause Y
2. Y can cause X
3. Z is causing both X and Y
4. Pure coincidence

Categories 1 and 2 are self-explanatory; there are several examples detailing compounding factors; in a well-known example of a confounding factor, numerous epidemiological trials indicated that women who were taking combined HRT also had a lower-than-average incidence of coronary heart disease, leading to the proposition that HRT was protective against CHD. However, several randomised clinical trials indicated that HRT caused a small but statistically significant *increase* in risk of coronary heart disease.

Re-analysis of the data from the epidemiological studies showed that women undertaking HRT were more likely to be from higher socio-economic backgrounds, with better than average diet and exercise regimens. The use of HRT and decreased incidence of coronary heart disease were coincident effects of a common cause (i.e. the benefits associated with a higher socioeconomic status), rather than cause and effect, as had been supposed.

Coincidental causality is often used to illustrate the spuriousness of correlation; a noted example follows the idea that as the incidence of pirate attacks has decreased, so global warming has increased; so is a lack of pirates causing global warming, or is global warming killing off the pirates?

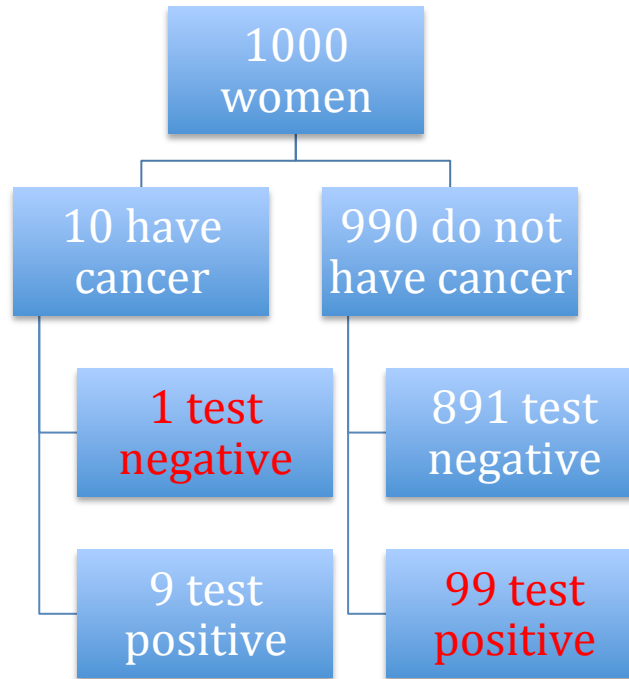
PREVALENCE AND INCIDENCE

Another commonly misinterpreted statistic is the understanding of the accuracy of screening techniques for serious diseases such as cancer. If we have a screening process that we know is 90% accurate, and a positive test is indicated, what are the chances that the test is accurate?

The answer is that it is impossible to answer that question with the information given, but unfortunately we are conditioned to think otherwise – in other words, a positive test means it's 90% likely that the person has the disease in question.

As with percentage changes above, it isn't as simple as that.

We need to know the incidence of the disease in a population to examine the true probability of having the disease. Let us talk in real numbers and with a real disease – breast cancer. The incidence of breast cancer across all ages is approximately 1% (rounding for simplicity). So, consider the following diagram:



Looking at the above approximation, most of the 'positive' tests from this initial screening are false-positive. In real terms, only about one in ten will truly have cancer, but they won't yet know which. That is how a test with 90% accuracy can result in only about 10% certainty, and why thinking about it in percentages can lead to confusion. In reality, these positive results would be followed up with further treatment, biopsies etc. and so this would reduce the number further.

Following on from this, another area of misrepresentation is the reporting of causal incidence of cancer; for example, reports of the effect of drinking on cancer often fail to say **how many** people are affected, presenting percentage increases – one recently stated figure was that a drink every day increases your chance of breast cancer by 12%. Referring to the previous section on percentage increases, we know that a percentage increase is dependent on the baseline variable; so in the instance of a 12% increase in a relatively low-incidence disease leads to a misleading interpretation of the facts. Using the same example - breast cancer - taking the 1% incidence shows that a 12% increase would take this from c.10 in 1000 to c.11 in 1000; an increase of 1 in 1000 does not appear as shocking as an increase of 12%. Either way this is reported, that one person in a thousand will always turn out to be important and relevant if that person turns out to be you or someone you know.

QUIRKS AND PARADOXICAL MISCONCEPTIONS

THE BIRTHDAY PARADOX

In probability theory, the birthday paradox concerns the probability that, in a set of randomly chosen people, some pair of them will have the same birthday. The probability reaches 100% when the number of people reaches 367 (since there are 366 possible birthdays, including February 29th during a leap year). However, 99% probability is reached with just 57 people, and 50% probability with 23 people. In other words, in a room full of 23 people, there is a greater than 50/50 chance that two people will share a birthday.

Let us consider the probabilities of the problem.

If $P(B)$ is the probability of at least two people sharing a birthday, then it is simpler to calculate the probability of no-one in the room sharing a birthday, $P(B')$; this way, we have two possibilities as these two events are mutually exclusive. For simplicity and purity, we disregard such variations such as leap years, twins and known seasonal variations. Also consider that the birthdays are independent events, therefore the probability of all events occurring is equal to the product of the probabilities of the events occurring.

- Consider the first person: assuming that the birthdays are likely to occur on each day of the (non-leap) year, the probability of this person **not** sharing his birthday with previous people will – of course – be 1 (365 / 365).
- Now consider the next person. Again, assuming equally likely probabilities of birthdays, the probability of this person not sharing a birthday with the first will be 364/365 (0.9972) – due to the fact that if this person was born on any of the other 364 days of the year, the first and second person would not share a birthday.
- Similarly for the third person...taking all of the previous assumptions and methodology into account, the probability of the third person not sharing a birthday with the previous two would be 363/365 (0.9945).
- Therefore, the probability of all three people not sharing the same birthday is:

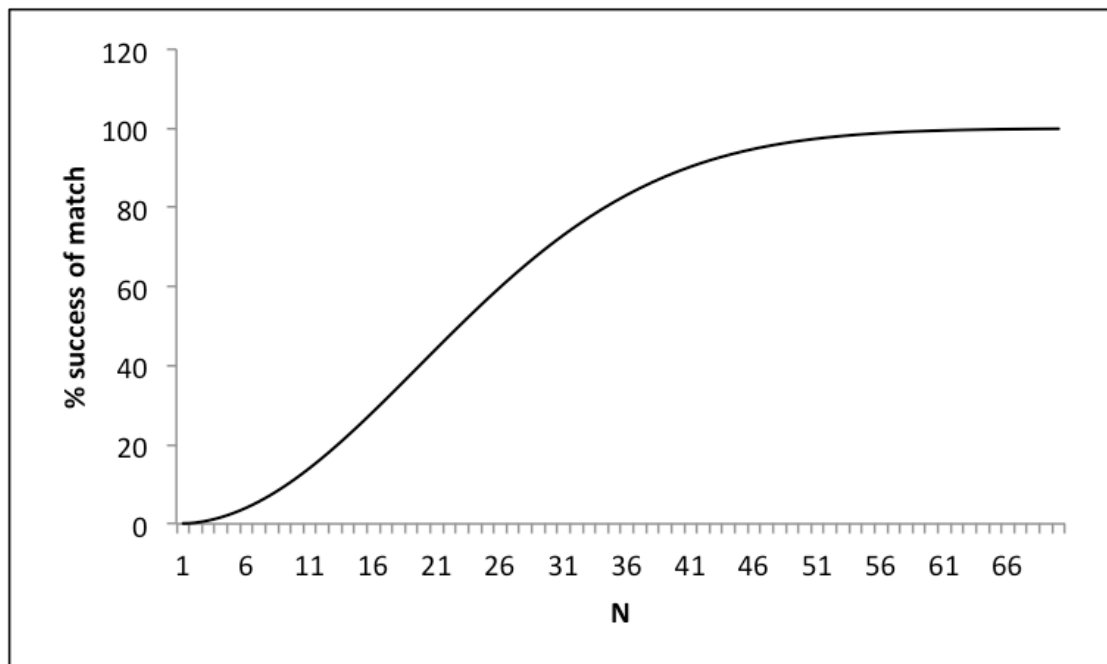
$$365/365 \times 364/365 \times 363/365 = 0.9917$$

Take this through to the 23rd person and we have the following probability:

$$365/365 \times 364/365 \times 363/365 \times 362/365 \times \dots \times 343/365$$

Evaluating this equation gives 0.4927; meaning that the chances of someone **not** not sharing a birthday is 0.5073 (or just over 50/50).

The data above can be plotted to show the following figure:



PhUSE 2012

Examining some of the numbers required for percentage change is interesting:

Number of people	Probability
10	0.1169
20	0.4114
30	0.7063
40	0.8912
50	0.9703
60	0.9941

The theory behind this process has been used in clinical trials for sample size estimation; the custom of single-arm, phase II oncology trials appears to date back to Gehan's 1961 article, which proposed a method to calculate the minimum sample size required to estimate the response rate with a given degree of precision, under the caveat that the study would be terminated early if there was 95% confidence that the response rate was less than some target response rate of interest (usually 15-20%).

In this classic design, 14 or 19 patients (corresponding to response rates of 20% or 15%) are initially treated, and if no responses are observed, the drug is considered inactive. On the other hand, if at least one response is observed, additional patients are enrolled and the response rate is estimated.

Illustrating this with the 20% response rate for this first stage, we compute the number of consecutive failures, i.e.:

N of failures	Probability (expanded)	
One failure:	0.8	0.8
Two failures:	0.8×0.8	0.64
Three failures:	$0.8 \times 0.8 \times 0.8$	0.512
...	..	...
Fourteen failures:	$(0.8)^{14}$	0.044

In other words, If the drug is 20% effective, there would be a 95% chance of at least one success.

PhUSE 2012

THE MONTY HALL PARADOX (OR “DEAL OR NO DEAL”)

The Monty Hall paradox is a probability-based method loosely based on the American television game show “Let’s Make A Deal” named after the show’s original host Monty Hall. It is a veridical paradox because the result appears odd but is demonstrably true. The modern day game show “Deal or No Deal” illustrates this problem further.

The classic problem is posed as follows:

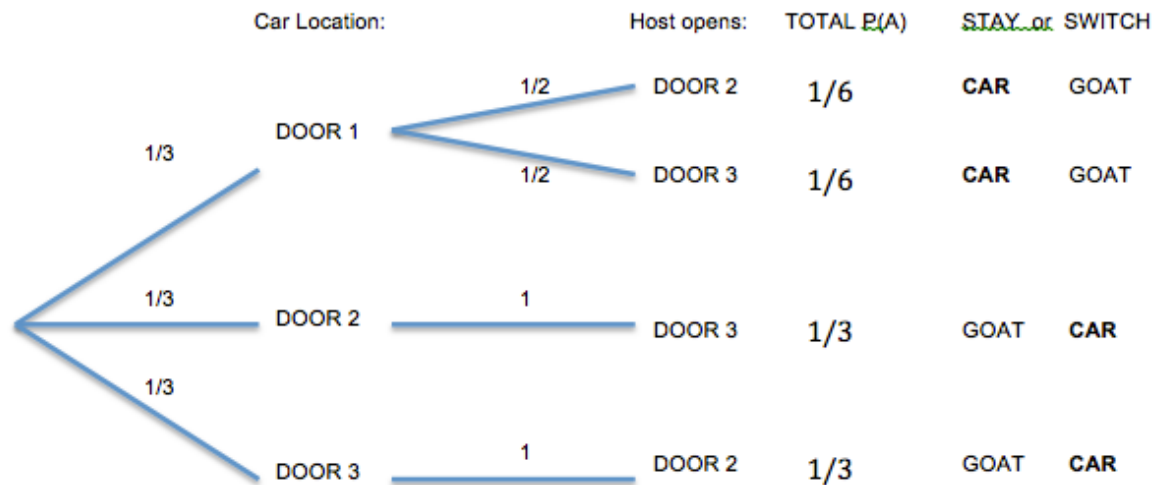
You are on a game show, and you are given the choice of three doors:

Behind one door is a car; behind the others, goats.

You pick a door, say No. 1 [but the door is not opened], and the host, who knows what is behind every door, opens another door, say No. 3, which has a goat. He then says to you, “Do you want to pick door No. 2?”

Is it to your advantage to switch your choice?

The (perhaps counterintuitive) response is that the contestant should always switch to the other door. If the car is initially equally likely to be behind each door, a player who picks Door 1 and doesn’t switch has a 1 in 3 chance of winning the car while a player who picks Door 1 and does switch has a 2 in 3 chance. The host has removed an incorrect option from the unchosen doors, so contestants who switch double their chances of winning the car.



If there are two doors left, then why isn’t each door “50/50”?

It may be easier to appreciate the solution by considering the same problem with 1,000,000 doors instead of just three. In this case there are 999,999 doors with goats behind them and one door with a prize. The player picks a door. His initial probability of winning is 1 out of 1,000,000. The game host goes down the line of doors, opening each one to show 999,998 goats in total, skipping over only the player’s door and one other door. The host then offers the player the chance to switch to the only other unopened door. On average, in 999,999 out of 1,000,000 times the other door will contain the prize, as 999,999 out of 1,000,000 times the player first picked a door with a goat—the chance that the player’s door is correct hasn’t changed. A rational player should switch.

The solution may also be described formally using Baye’s formula. Consider the discrete random variables, all taking values in the set of door numbers {1,2,3}:

C : the number of the door hiding the car,

P : the number of the door selected by the player, and

H : the number of the door opened by the host.

Then, if the player initially selects Door 1, and the host opens Door 3, the probability of winning by switching is:

$$P(C = 2|H = 3, P = 1) = \frac{P(H = 3|C = 2, P = 1)P(C = 2|P = 1)}{\sum_{i=1}^3 P(H = 3|C = i, P = 1)P(C = i|P = 1)}$$

$$= \frac{1 \times \frac{1}{3}}{(\frac{1}{2} \times \frac{1}{3}) + (1 \times \frac{1}{3}) + (0 \times \frac{1}{3})} = \frac{2}{3}$$

as due to the random placement of the car and the independence of the initial choice of the player and the placement of the car:

$$P(C = 1|P = 1) = P(C = 2|P = 1) = P(C = 3|P = 1) = \frac{1}{3}$$

and as part of the host's behaviour:

$$P(H = 3|C = 1, P = 1) = \frac{1}{2}$$

$$P(H = 3|C = 2, P = 1) = 1$$

$$P(H = 3|C = 3, P = 1) = 0$$

That switching has a probability of 2/3 of winning the car runs counter to many people's intuition.

CONCLUSION

Misconception, misunderstanding...call it what you will. The issue of 'misconstruing' statistical data and results continues to be a problem, be it through advertising or media manipulation of numbers to 'shock', or through data presented in a particular way that is not intended to mislead, but our perception of numbers is clouded by complexity - the bottom line is that something that appears to be particularly malignant more often than not is not quite so impressive...but of course, sometimes something that appears to be less impressive can be the opposite.

ACKNOWLEDGMENTS

Thanks to Martin Wilson and Stephen Bamford from Cerafor for review and guidance. Also to the authors below for the inspiration to write this paper and the source of some of the topics expanded upon.

RECOMMENDED READING AND REFERENCES

Freakonomics (2005; Levitt & Dubner)

ISBN 978-0-141-01901-7

Superfreakonomics (2009; Levitt & Dubner)

ISBN 978-0-141-03070-8

Statistical Misconceptions (2008, Huck)

ISBN 978-0-805-85904-1

PAPER: The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study (Andrew J Vickers)

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

John Salter
Cerafor
Kent Innovation Centre
Millennium Way
Broadstairs
Kent
CT10 2QQ
United Kingdom
W: www.cerafor.com
E: john.salter@cerafor.com

Brand and product names are trademarks of their respective companies.