Statistical Pitfalls in Data Careers

Yeng M. Miller-Chang

March 19, 2019

About Me

- Research Analysis Specialist, Office of Research and Planning (2nd floor, by Java Junction), C2105
  - Serve employees and students by supporting evidence-based decision making
- Business cards
- Quick announcement...

Four Statistical Pitfalls in the Professional World

Things that you should **never** do with statistics:

Four Statistical Pitfalls in the Professional World

Things that you should **never** do with statistics:

1. Rely on only one statistic or procedure to make decisions

Four Statistical Pitfalls in the Professional World

Things that you should **never** do with statistics:

1. Rely on only one statistic or procedure to make decisions
2. Assume that all data are normally distributed (bell-shape curve)

Four Statistical Pitfalls in the Professional World

Things that you should **never** do with statistics:

1. Rely on only one statistic or procedure to make decisions
2. Assume that all data are normally distributed (bell-shape curve)
3. Assume that trends that apply to a larger group apply to a smaller group

Four Statistical Pitfalls in the Professional World

Things that you should **never** do with statistics:

1. Rely on only one statistic or procedure to make decisions
2. Assume that all data are normally distributed (bell-shape curve)
3. Assume that trends that apply to a larger group apply to a smaller group
4. Assume that machine learning (data science, artificial intelligence, predictive analytics, etc.) is a magic bullet

Reliance on only one statistic/procedure to make decisions

The simplest way this pitfall appears is usually in the form of averages.

Reliance on only one statistic/procedure to make decisions

The simplest way this pitfall appears is usually in the form of averages. Consider two data sets:

Data set 1: 50 zeroes, 50 tens
Data set 2: 20 fives

These both have an average of 5, but these data sets look significantly different (in values and in sample size)!

Reliance on only one statistic/procedure to make decisions

Let's put this in a real-world context.

- Suppose you're a manager of a retail store. You're trying to implement a change to cashier training.

Reliance on only one statistic/procedure to make decisions

Let's put this in a real-world context.

- Suppose you're a manager of a retail store. You're trying to implement a change to cashier training.
- You survey customers before and after the training, and ask them to rate their experience with their cashiers on a scale from 0 to 10: 0 being worst, 10 being best.

Reliance on only one statistic/procedure to make decisions

Let's put this in a real-world context.

- Suppose you're a manager of a retail store. You're trying to implement a change to cashier training.
- You survey customers before and after the training, and ask them to rate their experience with their cashiers on a scale from 0 to 10: 0 being worst, 10 being best.
- Many customers decline to take your survey, and the best you can get is the following:

	Number responded to survey	Average Rating
Pre-Training	50	5.5
Post-Training	25	5.75

Reliance on only one statistic/procedure to make decisions

Let's put this in a real-world context.

- Suppose you're a manager of a retail store. You're trying to implement a change to cashier training.
- You survey customers before and after the training, and ask them to rate their experience with their cashiers on a scale from 0 to 10: 0 being worst, 10 being best.
- Many customers decline to take your survey, and the best you can get is the following:

	Number responded to survey	Average Rating
Pre-Training	50	5.5
Post-Training	25	5.75

Did the cashier training improve customer interaction?

Reliance on only one statistic/procedure to make decisions

	Number responded to survey	Average Rating
Pre-Training	50	5.5
Post-Training	25	5.75

Reliance on only one statistic/procedure to make decisions

	Number responded to survey	Average Rating
Pre-Training	50	5.5
Post-Training	25	5.75

The averages are not comparable: the post-training sample is half the size of the pre-training sample!

- Why does this matter?

Reliance on only one statistic/procedure to make decisions

	Number responded to survey	Average Rating
Pre-Training	50	5.5
Post-Training	25	5.75

The averages are not comparable: the post-training sample is half the size of the pre-training sample!

- Why does this matter?
  - Integer scales (or "Likert scales") are subject to **measurement error**: it is difficult to explain the difference between why one would choose a 4 over a 5, for example.

Reliance on only one statistic/procedure to make decisions

	Number responded to survey	Average Rating
Pre-Training	50	5.5
Post-Training	25	5.75

The averages are not comparable: the post-training sample is half the size of the pre-training sample!

- Why does this matter?
  - Integer scales (or "Likert scales") are subject to **measurement error**: it is difficult to explain the difference between why one would choose a 4 over a 5, for example.
  - If **one person** changed their rating by one point in the pre-training survey, it would change the average by 0.02, as opposed to 0.04 in the post-test survey!

Reliance on only one statistic/procedure to make decisions

	Number responded to survey	Average Rating
Pre-Training	50	5.5
Post-Training	25	5.75

The averages are not comparable: the post-training sample is half the size of the pre-training sample!

- Why does this matter?
  - Integer scales (or "Likert scales") are subject to **measurement error**: it is difficult to explain the difference between why one would choose a 4 over a 5, for example.
  - If **one person** changed their rating by one point in the pre-training survey, it would change the average by 0.02, as opposed to 0.04 in the post-test survey!
  - A person in the post-training survey has **twice the influence on the average** than a person in the pre-training survey!

Reliance on only one statistic/procedure to make decisions

	Number responded to survey	Average Rating
Pre-Training	50	5.5
Post-Training	25	5.75

The averages are not comparable: the post-training sample is half the size of the pre-training sample!

- Why does this matter?
  - Integer scales (or "Likert scales") are subject to **measurement error**: it is difficult to explain the difference between why one would choose a 4 over a 5, for example.
  - If **one person** changed their rating by one point in the pre-training survey, it would change the average by 0.02, as opposed to 0.04 in the post-test survey!
  - A person in the post-training survey has **twice the influence on the average** than a person in the pre-training survey!
  - All it would take to get the difference of 0.25 in the averages would be seven people in the post-training survey scoring one point less than what they stated in the survey (completely plausible).

Reliance on only one statistic/procedure to make decisions

More exploration, outside of just comparing averages, needs to be done.

Reliance on only one statistic/procedure to make decisions

More exploration, outside of just comparing averages, needs to be done. If we return to our previous two data sets:

Data set 1: 50 zeroes, 50 tens
Data set 2: 20 fives

one could also look at the standard deviation of the two data sets.

Reliance on only one statistic/procedure to make decisions

Data set 1: 50 zeroes, 50 tens
Data set 2: 20 fives

The **standard deviation** is a metric used to roughly determine how far a data point is, on average, from the mean. It is generally used to measure how spread out data sets are.

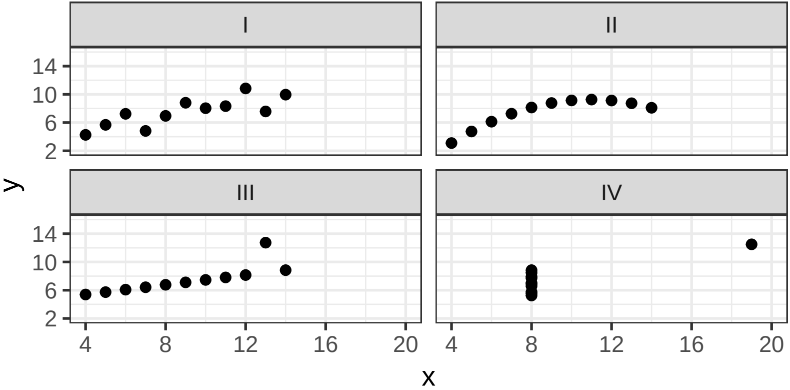
The standard deviation of data set 1 is larger than that of data set 2. Notice that the points in data set 1 are more spread out than those of data set 2.

If these two data sets represented pre- (data set 1) and post-training ratings of cashiers, one *could* conclude that post-training ratings are better because of the smaller standard deviation.[1]

[1] Although I would not recommend it in such a situation.

Reliance on only one statistic/procedure to make decisions

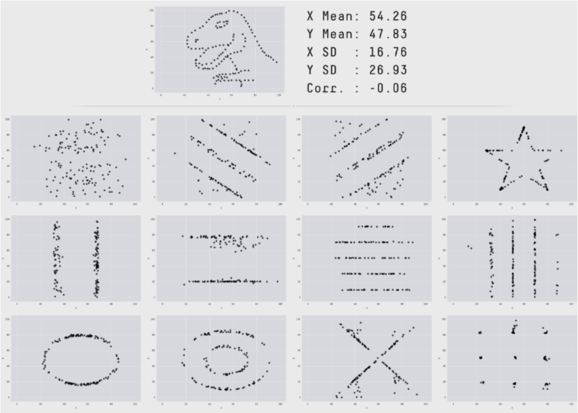
Anscombe's Quartet: four data sets with the same *x*-means, *x*-standard deviations, *y*-means, *y*-standard deviations[2]



[2] The correlation coefficients, R^2s, and OLS regression lines are also (essentially) the same! Sources: https://gist.github.com/ericbusboom, Wikipedia

Reliance on only one statistic/procedure to make decisions

Datasaurus Dozen:[3]



[3] Matejka, J. and Fitzmaurice, G. (2017), “Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing,” *Autodesk Research*, https://www.autodeskresearch.com/publications/samestats

Reliance on only one statistic/procedure to make decisions

The use of a few statistics is not sufficient to adequately explain the differences between data sets!

But... we don't have time to compute absolutely every possible statistic. For practical purposes, how to deal with this problem?

Reliance on only one statistic/procedure to make decisions

The use of a few statistics is not sufficient to adequately explain the differences between data sets!

But... we don't have time to compute absolutely every possible statistic. For practical purposes, how to deal with this problem?

1. Have the **domain knowledge** to understand where to focus your statistics. Do you understand the question of interest well enough so that you can focus your analysis?

Reliance on only one statistic/procedure to make decisions

The use of a few statistics is not sufficient to adequately explain the differences between data sets!

But... we don't have time to compute absolutely every possible statistic. For practical purposes, how to deal with this problem?

1. Have the **domain knowledge** to understand where to focus your statistics. Do you understand the question of interest well enough so that you can focus your analysis?
2. Understand the **limitations of your metrics**. Averages are sensitive to measurement error and sample size.

Reliance on only one statistic/procedure to make decisions

The use of a few statistics is not sufficient to adequately explain the differences between data sets!

But... we don't have time to compute absolutely every possible statistic. For practical purposes, how to deal with this problem?

1. Have the **domain knowledge** to understand where to focus your statistics. Do you understand the question of interest well enough so that you can focus your analysis?
2. Understand the **limitations of your metrics**. Averages are sensitive to measurement error and sample size.
3. Have the domain knowledge to understand what metrics are **best communicated**, given the situation.

Reliance on only one statistic/procedure to make decisions

The use of a few statistics is not sufficient to adequately explain the differences between data sets!

But... we don't have time to compute absolutely every possible statistic. For practical purposes, how to deal with this problem?

1. Have the **domain knowledge** to understand where to focus your statistics. Do you understand the question of interest well enough so that you can focus your analysis?
2. Understand the **limitations of your metrics**. Averages are sensitive to measurement error and sample size.
3. Have the domain knowledge to understand what metrics are **best communicated**, given the situation.
4. Spend a substantial time doing **exploratory data analysis (EDA)** to understand the structure of your data.

Reliance on only one statistic/procedure to make decisions

When U.S. air force discovered the flaw of averages:[4]

“In the late 1940s, the United States air force had a serious problem: its pilots could not keep control of their planes.”

“Back in 1926, when the army was designing its first-ever cockpit, engineers had measured the physical dimensions of hundreds of male pilots... and used this data to standardize the dimensions of the cockpit. For the next three decades, the size and shape of the seat, the distance to the pedals and stick, the height of the windshield, even the shape of the flight helmets were all built to conform to the average dimensions of a 1926 pilot.”

[4] Rose, T. (2016), “When U.S. air force discovered the flaw of averages,” *The Star*, https://www.thestar.com/news/insight/2016/01/16/when-us-air-force-discovered-the-flaw-of-averages.html

Reliance on only one statistic/procedure to make decisions

“So when the air force put him to work measuring pilots, Daniels harboured a private conviction about averages that rejected almost a century of military design philosophy. As he sat in the Aero Medical Laboratory measuring hands, legs, waists and foreheads, he kept asking himself the same question in his head: *How many pilots really were average?*”

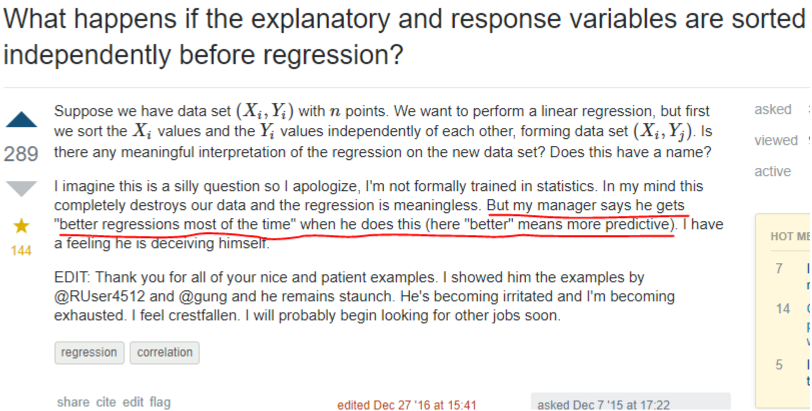
“The scientists also expected that a sizable number of pilots would be within the average range on all 10 dimensions. But even Daniels was stunned when he tabulated the actual number.

Zero.

Out of 4,063 pilots, not a single airman fit within the average range on all 10 dimensions.”

Reliance on only one statistic/procedure to make decisions

Changing the data to get a “better” prediction:[5]



What happens if the explanatory and response variables are sorted independently before regression?

289

144

Suppose we have data set (X_i, Y_i) with n points. We want to perform a linear regression, but first we sort the X_i values and the Y_i values independently of each other, forming data set (X_i, Y_j). Is there any meaningful interpretation of the regression on the new data set? Does this have a name?

I imagine this is a silly question so I apologize, I'm not formally trained in statistics. In my mind this completely destroys our data and the regression is meaningless. But my manager says he gets "better regressions most of the time" when he does this (here "better" means more predictive). I have a feeling he is deceiving himself.

EDIT: Thank you for all of your nice and patient examples. I showed him the examples by @RUser4512 and @gung and he remains staunch. He's becoming irritated and I'm becoming exhausted. I feel crestfallen. I will probably begin looking for other jobs soon.

regression correlation

share cite edit flag

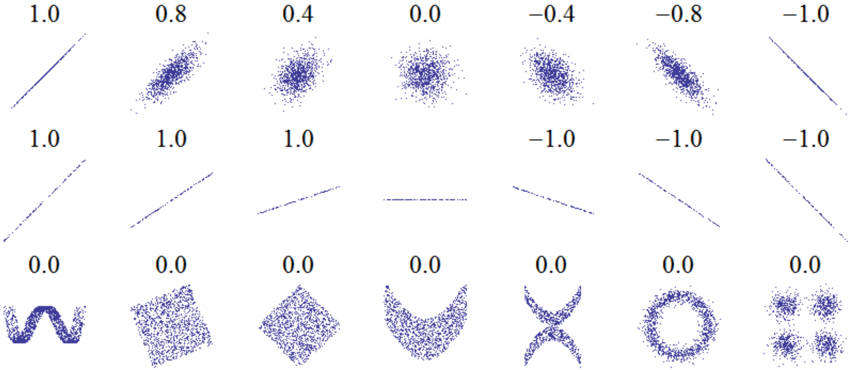
edited Dec 27 '16 at 15:41

asked Dec 7 '15 at 17:22

[5] https://stats.stackexchange.com/q/185507/46427

Reliance on only one statistic/procedure to make decisions

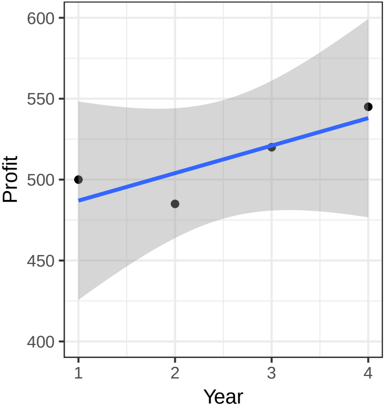
Correlation coefficient *R* (for **linear model** fit):[6]



[6] https://en.wikipedia.org/wiki/File:Correlation_examples.png

Reliance on only one statistic/procedure to make decisions

Year	Profit
1	500
2	485
3	520
4	545

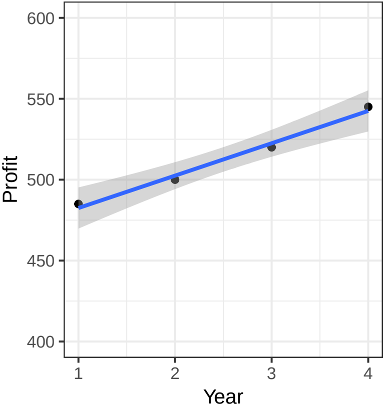


$R =$0.8447368

Reliance on only one statistic/procedure to make decisions

After "sorting:"

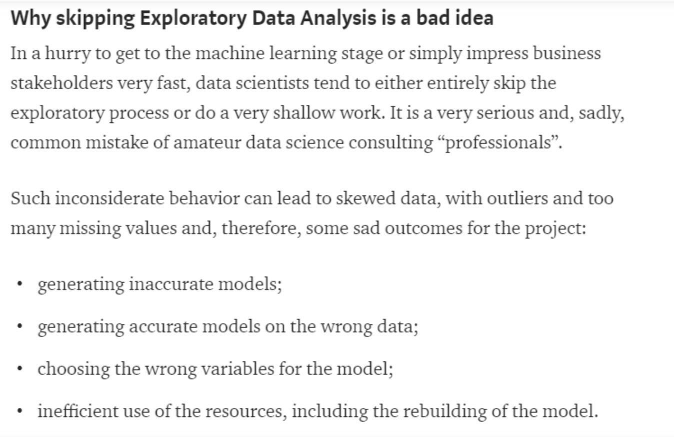
Year	Profit
1	485
2	500
3	520
4	545



R =0.993808

Reliance on only one statistic/procedure to make decisions

On EDA:[7]



Why skipping Exploratory Data Analysis is a bad idea

In a hurry to get to the machine learning stage or simply impress business stakeholders very fast, data scientists tend to either entirely skip the exploratory process or do a very shallow work. It is a very serious and, sadly, common mistake of amateur data science consulting "professionals".

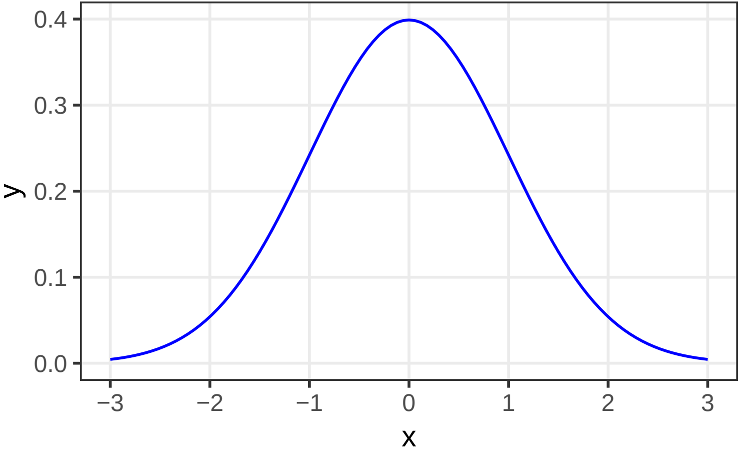
Such inconsiderate behavior can lead to skewed data, with outliers and too many missing values and, therefore, some sad outcomes for the project:

- generating inaccurate models;
- generating accurate models on the wrong data;
- choosing the wrong variables for the model;
- inefficient use of the resources, including the rebuilding of the model.

[7] https://medium.com/@InDataLabs/why-start-a-data-science-project-with-exploratory-data-analysis-f90c0efcbe49

Assuming that all data are normally distributed

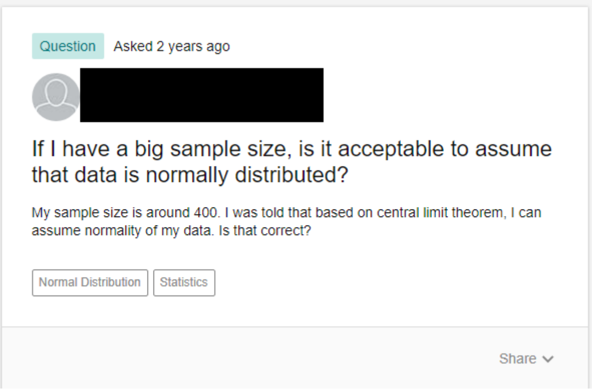
"All models are wrong; some models are useful." (George E. P. Box)[8]



[8] Box et al. (2005), *Statistics for Experimenters* (2^{nd} ed.), Hoboken, NJ: John Wiley & Sons

Assuming that all data are normally distributed

Common misconception: large sample ($n > 30$) $\implies$ normality:[9]



[redacted]

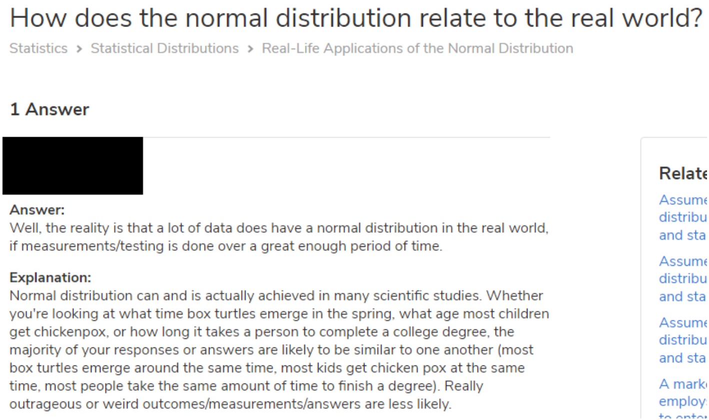
If I have a big sample size, is it acceptable to assume that data is normally distributed?

My sample size is around 400. I was told that based on central limit theorem, I can assume normality of my data. Is that correct?

[9] https://www.researchgate.net/post/If_I_have_a_big_sample_size_is_it_acceptable_to_assume_that_data_is_normally_distributed

Assuming that all data are normally distributed

Common misconception: large sample ($n > 30$) $\implies$ normality:[10] (the answer below is based on this misconception)



How does the normal distribution relate to the real world?

Statistics › Statistical Distributions › Real-Life Applications of the Normal Distribution

1 Answer

[redacted]

Answer:
Well, the reality is that a lot of data does have a normal distribution in the real world, if measurements/testing is done over a great enough period of time.

Explanation:
Normal distribution can and is actually achieved in many scientific studies. Whether you're looking at what time box turtles emerge in the spring, what age most children get chickenpox, or how long it takes a person to complete a college degree, the majority of your responses or answers are likely to be similar to one another (most box turtles emerge around the same time, most kids get chicken pox at the same time, most people take the same amount of time to finish a degree). Really outrageous or weird outcomes/measurements/answers are less likely.

[10] https://socratic.org/questions/5640dd3f581e2a6f21bb1f0a

Assuming that all data are normally distributed

Why do stats courses emphasize the normal distribution so much?

Assuming that all data are normally distributed

Why do stats courses emphasize the normal distribution so much?

- It is a **mathematically convenient** distribution: sums and differences of (jointly) normal distributions are also normally distributed, which is **not** the case for many distributions.

Assuming that all data are normally distributed

Why do stats courses emphasize the normal distribution so much?

- It is a **mathematically convenient** distribution: sums and differences of (jointly) normal distributions are also normally distributed, which is **not** the case for many distributions.
- It justifies the **rationale for many statistical procedures** (e.g., least-squares regression lines).

Assuming that all data are normally distributed

Why do stats courses emphasize the normal distribution so much?

- It is a **mathematically convenient** distribution: sums and differences of (jointly) normal distributions are also normally distributed, which is **not** the case for many distributions.
- It justifies the **rationale for many statistical procedures** (e.g., least-squares regression lines).
- Central Limit Theorem: **sums** and **averages** of large samples are approximately normal for large sample sizes. This does **NOT** mean that the underlying data are normal, but only sums and averages!

Assuming that all data are normally distributed

Examples of data that are **not** normally distributed:

- Website traffic (extremely skewed: a select few websites for an organization get more traffic than many other websites, especially if there's an established front page)

Assuming that all data are normally distributed

Examples of data that are **not** normally distributed:

- Website traffic (extremely skewed: a select few websites for an organization get more traffic than many other websites, especially if there's an established front page)
- Stock prices (there are “skew-normal” distributions to try to model the volatility of stock prices more appropriately)

Assuming that all data are normally distributed

Examples of data that are **not** normally distributed:

- Website traffic (extremely skewed: a select few websites for an organization get more traffic than many other websites, especially if there's an established front page)
- Stock prices (there are "skew-normal" distributions to try to model the volatility of stock prices more appropriately)
- Standardized test scores (these exhibit "floor" and "ceiling" effects; there are significant populations that perform very poorly or very well)

Assuming that all data are normally distributed

Misconceptions related to the assumption that all data are normally distributed:

Assuming that all data are normally distributed

Misconceptions related to the assumption that all data are normally distributed:

- Mean = median = mode

Assuming that all data are normally distributed

Misconceptions related to the assumption that all data are normally distributed:

- Mean = median = mode
- Any distribution can be completely described by the mean and standard deviation (true for normal, false everywhere else; recall the discussion on Anscombe's quartet and the Datasaurus Dozen)

Assuming that all data are normally distributed

Suppose E represents expenses, R represents revenue, and both E and R are normally distributed and independent. $\dfrac{E}{R}$ is called the **efficiency ratio** (how much is spent to make \$1 in revenue). What is the distribution of the efficiency ratio?

Assuming that all data are normally distributed

Suppose E represents expenses, R represents revenue, and both E and R are normally distributed and independent. $\frac{E}{R}$ is called the **efficiency ratio** (how much is spent to make $1 in revenue). What is the distribution of the efficiency ratio?

It turns out that, with the above assumptions, the efficiency ratio $\frac{E}{R}$ is **not** normally distributed, has **no mean**, and has an **infinite standard deviation**!

Many of the assumptions in this section outline the need for a full background in probability.

Assuming trends that apply to a larger group apply to a smaller group

Assuming trends that apply to a larger group apply to a smaller group

"**Simpson's paradox** arises whenever an apparent trend in data, caused by a confounding variable, can be eliminated or reversed by splitting the data into natural groups."[11]

[11] Reinhart A. (2005), *Statistics Done Wrong: the Woefully Complete Guide*, San Francisco, CA: No Starch Press, Inc.

Assuming trends that apply to a larger group apply to a smaller group

"**Simpson's paradox** arises whenever an apparent trend in data, caused by a confounding variable, can be eliminated or reversed by splitting the data into natural groups."[11]

Sex bias in UC-Berkeley graduate admissions (Fall 1973):[12]

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8,442	**44%**	4,321	**35%**

[11] Reinhart A. (2005), *Statistics Done Wrong: the Woefully Complete Guide*, San Francisco, CA: No Starch Press, Inc.

[12] Freedman et al. (2007), *Statistics* (4th ed.), W. W. Norton.; see also Wikipedia on "Simpson's Paradox"

Assuming trends that apply to a larger group apply to a smaller group

"[W]hen we take account of the differences **among departments** in the proportions of men and women applying to them and avoid this problem by computing a statistic on each department separately, and aggregating those statistics, the evidence for campus-wide bias in favor of men is extremely weak; on the contrary, there is evidence of bias in favor of women... The proportion of women applicants tends to be high in departments that are hard to get into and low in those that are easy to get into."[13] (emphasis added)

Again, a mix of domain knowledge and EDA can help with this problem.

[13] Bickel P.J., Hammel E.A., and O'Connell, J.W. (1975). "Sex Bias in Graduate Admissions: Data From Berkeley," in *Science*. 187 (4175), pp. 398–404.

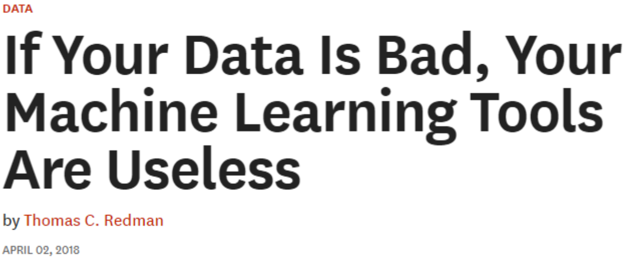
Assuming that machine learning is a magic bullet

Assuming that machine learning is a magic bullet

What do you need to run a machine learning algorithm?

Assuming that machine learning is a magic bullet

What do you need to run a machine learning algorithm? Data.[14]



DATA

If Your Data Is Bad, Your Machine Learning Tools Are Useless

by Thomas C. Redman

APRIL 02, 2018

[14] Redman, T. C. (2018), "If Your Data Is Bad, Your Machine Learning Tools Are Useless," *Harvard Business Review*, https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless

Assuming that machine learning is a magic bullet

“Yet today, most data fails to meet basic 'data are right' standards. Reasons range from **data creators not understanding what is expected**, to **poorly calibrated measurement gear**, to **overly complex processes, to human error**. To compensate, **data scientists cleanse the data** before training the predictive model. It is time-consuming, tedious work (**taking up to 80% of data scientists' time**), and it's the problem data scientists complain about most. Even with such efforts, cleaning neither detects nor corrects all the errors, and as yet, there is no way to understand the impact on the predictive model. What's more, data does not always meets [*sic*] 'the right data' standards, as reports of bias in facial recognition and criminal justice attest.” [15]

[15] Redman, T. C. (2018), “If Your Data Is Bad, Your Machine Learning Tools Are Useless,” *Harvard Business Review*, https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless

Assuming that machine learning is a magic bullet

"The second issue is that once these junior people get to the market, they come in with an unrealistic set of expectations about what data science work will look like. Everyone thinks they're going to be doing machine learning, deep learning, and Bayesian simulations.

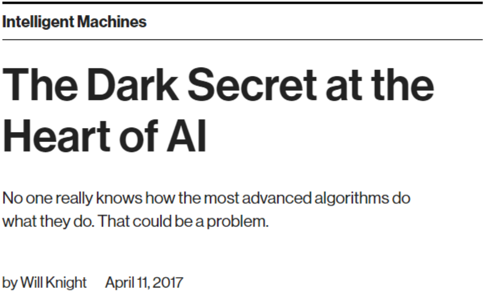
This is not their fault; this is what data science curriculums and the tech media emphasize. Not much has changed since I first glanced, starry-eyed, at Hacker News logistic regression posts many, many moons ago.

The reality is that 'data science' has never been as much about machine learning as it has about cleaning, shaping data, and moving it from place to place." [16] (emphasis added)

[16] Boykis, V. (2019), "Data science is different now," https://veekaybee.github.io/2019/02/13/data-science-is-different/

Assuming that machine learning is a magic bullet

Another problem with machine learning algorithms is that many of them do not provide easily interpretable results (interpretability vs. accuracy tradeoff). Beyond this, some of the most complicated algorithms are not understood by the vast majority of their users.[17]



Intelligent Machines

The Dark Secret at the Heart of AI

No one really knows how the most advanced algorithms do what they do. That could be a problem.

by Will Knight April 11, 2017

[17] Knight, W. (2017), "The Dark Secret at the Heart of AI," *MIT Technology Review*, https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

Solutions

Despite these pitfalls, they all have common solutions:

Solutions

Despite these pitfalls, they all have common solutions:

1. Have the **domain knowledge** to understand where to focus your statistics. Do you understand the question of interest well enough so that you can focus your analysis?

Solutions

Despite these pitfalls, they all have common solutions:

1. Have the **domain knowledge** to understand where to focus your statistics. Do you understand the question of interest well enough so that you can focus your analysis?
2. Understand the **limitations of your metrics**. Averages are sensitive to measurement error and sample size. **Machine learning and AI are not limitless**.

Solutions

Despite these pitfalls, they all have common solutions:

1. Have the **domain knowledge** to understand where to focus your statistics. Do you understand the question of interest well enough so that you can focus your analysis?
2. Understand the **limitations of your metrics**. Averages are sensitive to measurement error and sample size. **Machine learning and AI are not limitless**.
3. Have the domain knowledge to understand what metrics are **best communicated**, given the situation.

Solutions

Despite these pitfalls, they all have common solutions:

1. Have the **domain knowledge** to understand where to focus your statistics. Do you understand the question of interest well enough so that you can focus your analysis?
2. Understand the **limitations of your metrics**. Averages are sensitive to measurement error and sample size. **Machine learning and AI are not limitless**.
3. Have the domain knowledge to understand what metrics are **best communicated**, given the situation.
4. Spend a substantial time doing **exploratory data analysis (EDA)** to understand the structure of your data **and cleaning your data**.

Thank you for attending this presentation!

Any questions?