How To Avoid Data Analytics Pitfalls

Zhen Liu
zhen@dataforleaders.com

Do you only use metrics like ‘mean’ or ‘ratio’ to make data-driven business decisions? Do you know those ratio metrics can ‘lie’ and mislead you?

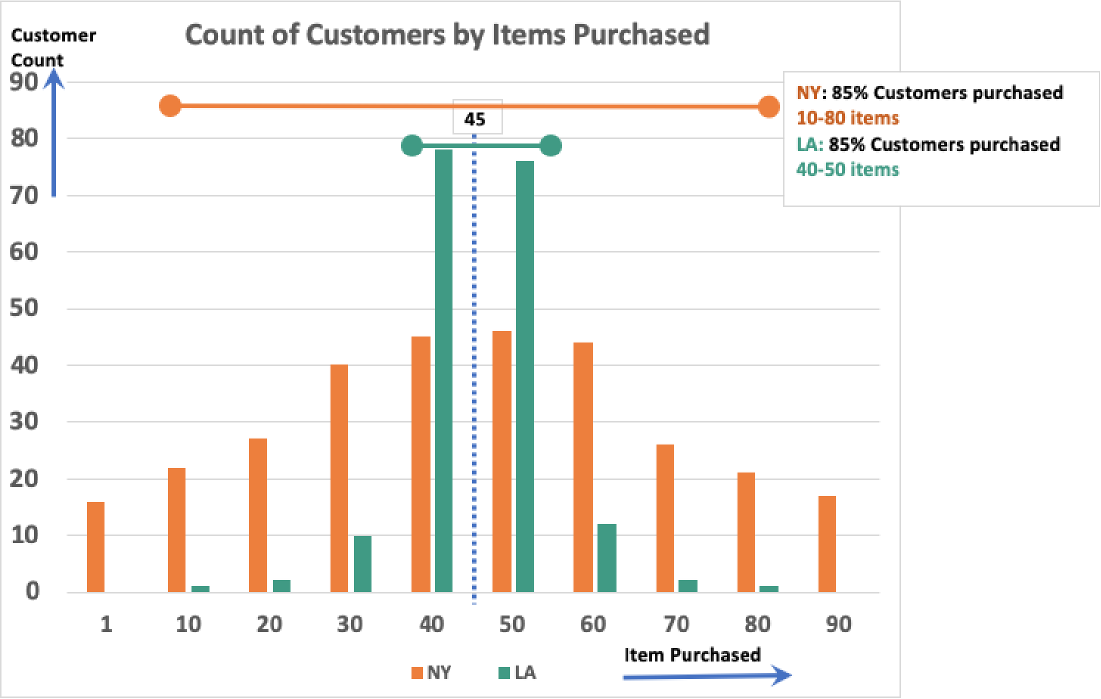
Now, I’ll use four data analytics cases to show you why those pitfalls are dangerous, and what you should do instead, so you don’t get ‘fooled’ by data.

Case 1: The Average Value Can’t Represent Your Customers

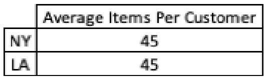
Why?

We use average a lot when analyzing product and business performance, but using the average alone actually creates blind spots. Because there are always variations due to different segments of the market or pure randomness. Unfortunately, the average value doesn’t tell you the variation of data.

Example: How many products do our customers buy on average?
A company is trying to understand the average amount of items purchased by a customer to develop a new marketing strategy. For New York and LA, they found that the average of the items purchased per customer is the same (45 items).



NY: 85% Customers purchased 10-80 items
LA: 85% Customers purchased 40-50 items



	Average Items Per Customer
NY	45
LA	45

Now, based on the plot below, should we apply the same marketing strategy for the customers in New York and LA?
NO.

In LA (green line), 85% customers purchased 40–50 items, which means the average amount(45) represents most customers' behavior. You might only need one big campaign to target the majority of the customers.

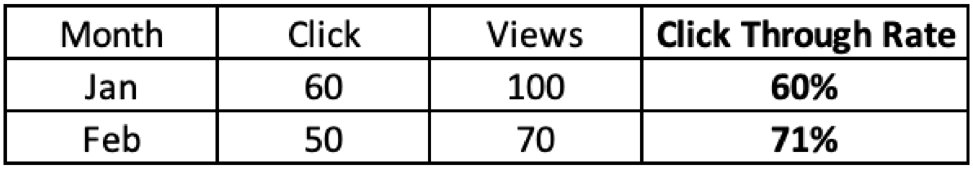
However, in New York, the average value can only represent 50% of the customers' behavior. The majority of customers, say 85%, lie between the buckets of purchasing between 10 items to 80 items, which we can observe from the large 'spread' of data as shown by the orange 'dumbbell' shaped line.

This means, the customers in New York have higher variance than those in LA, and you probably need multiple campaign strategies for New York when the customers' behaviors are more diversified.

What should we do?

When analyzing the average, also find out the range around the average by calculating variance.

A more accurate way to do this is to use Confidence Interval (CI) to estimate where the average lies in with a probability. (This link shows you how data scientists construct a Confidence Interval, and you can create it with Excel)



Month	Click	Views	**Click Through Rate**
Jan	60	100	**60%**
Feb	50	70	**71%**

An example for reporting is: the mean of item purchased per customer in New York is 45, and the 85% Confidence Interval is between 10 and 80.

Case 2. Ratio Metric Can Be Very Sensitive and Unreliable

Why?

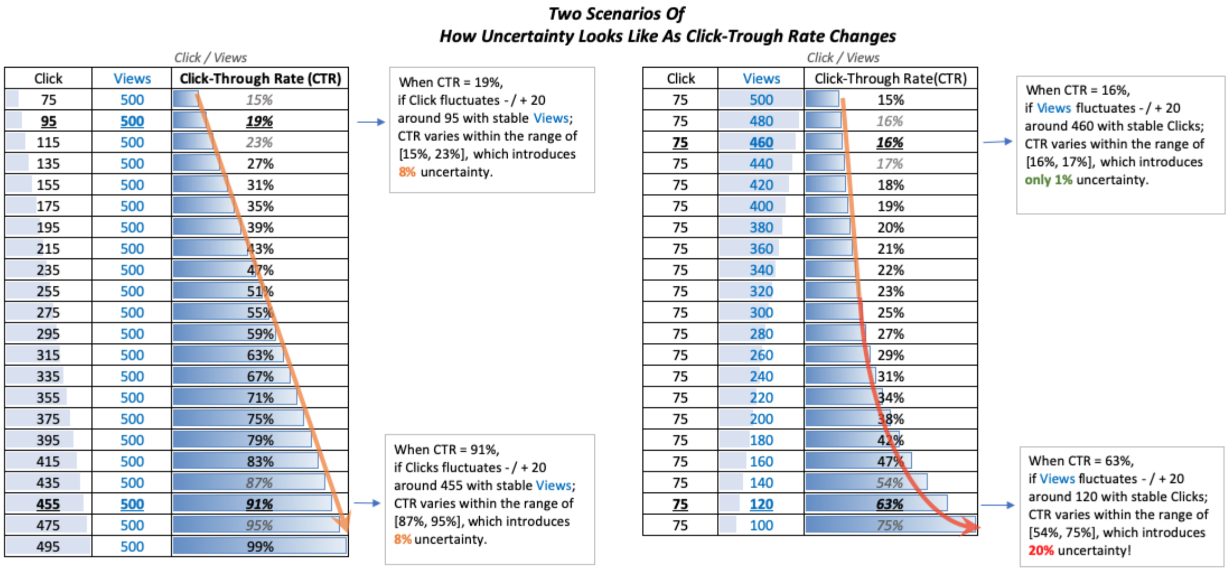
Ratio metric consists of at least two metrics; for example, Click-Through Rate (CTR) is Clicks divided by Views. With each metric's variation, ratio metric's variation is more complicated, and it doesn't follow any common distribution.

Example

Let's look at the table below first. You are measuring Click-Through Rate (CTR), from this table, it looks like Click-Through Rate increased from Jan to Feb. Sound great?

Well, actually both Clicks and Views decreased, the CTR increased because the Views decreased more. So, this increase is probably not what you want.

Now, let's look at the four more scenarios to see how CTR changes when we control one variable and change the other. Can we trust the ratio with the same level of certainty in each scenario?



Two Scenarios Of
How Uncertainty Looks Like As Click-Trough Rate Changes

Click / Views

Click	Views	Click-Through Rate (CTR)
75	500	*15%*
95	**500**	***19%***
115	500	*23%*
135	500	27%
155	500	31%
175	500	35%
195	500	39%
215	500	43%
235	500	47%
255	500	51%
275	500	55%
295	500	59%
315	500	63%
335	500	67%
355	500	71%
375	500	75%
395	500	79%
415	500	83%
435	500	*87%*
455	**500**	***91%***
475	500	*95%*
495	500	99%

When CTR = 19%,
if Click fluctuates - / + 20
around 95 with stable Views;
CTR varies within the range of
[15%, 23%], which introduces
8% uncertainty.

When CTR = 91%,
if Clicks fluctuates - / + 20
around 455 with stable Views;
CTR varies within the range of
[87%, 95%], which introduces
8% uncertainty.

Click / Views

Click	Views	Click-Through Rate(CTR)
75	500	15%
75	480	*16%*
75	**460**	***16%***
75	440	*17%*
75	420	18%
75	400	19%
75	380	20%
75	360	21%
75	340	22%
75	320	23%
75	300	25%
75	280	27%
75	260	29%
75	240	31%
75	220	34%
75	200	38%
75	180	42%
75	160	47%
75	140	*54%*
75	**120**	***63%***
75	100	*75%*

When CTR = 16%,
if Views fluctuates - / + 20
around 460 with stable Clicks;
CTR varies within the range of
[16%, 17%], which introduces
only 1% uncertainty.

When CTR = 63%,
if Views fluctuates - / + 20
around 120 with stable Clicks;
CTR varies within the range of
[54%, 75%], which introduces
20% uncertainty!

The Left Table shows that, if the denominator (View) is stable, the ratio metric (CTR) moves proportionally as the numerator (Click) moves, and the uncertainty of the data is easy to estimate, and the scale of uncertainty doesn't change much.

In the Right Table, when the denominator (View) is large enough as shown on the first few rows, ratio (CTR) is very stable with only 1–2% uncertainty. However, if you look at the bottom rows, the ratio can be very sensitive to changes, and it is very unstable when the denominator is small! When this is the case, it's better to monitor the Views and Clicks and expect a wide range of scenarios when you make decisions.

What should we do?

- Set a threshold for minimal acceptable value for the denominator. As the ratio can have high variance when denominator is small, we only trust the ratio when denominator is large enough. If you have to use the ratio metric to make decisions when denominator is small, make sure you report a range that covers the fluctuation.

- Use a moving average (take the average of a few nearby data points) to smooth the metrics when the variance is high.

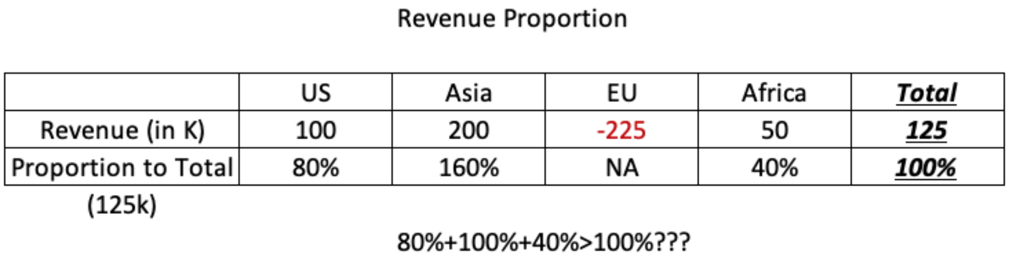
- Monitor the actual values (numerator, denominator) that we use for the ratio calculation. Understand the range of the ratio by simulating different scenarios of the numerator and denominator.

Case 3: The Blind Spots Behind Inflated Proportions

Say you have sales teams located in multiple regions, now a manager claims that the region's revenue is 160% of the overall revenue, would you trust that manager?

Example: What's the reasonable proportion of the revenue attributed to a region?

In the table below, we have the revenues of four regions: US, Asia, EU and Africa.



Revenue Proportion

	US	Asia	EU	Africa	***Total***
Revenue (in K)	100	200	-225	50	***125***
Proportion to Total (125k)	80%	160%	NA	40%	***100%***

80%+100%+40%>100%???

We want to see the regional revenues in proportion to the total. However, after calculation, we found that the sum of the proportions of each region exceeds 100%. How did it happen??

It's because the value in EU is negative, which drags the overall revenue down and inflates other location's proportion.

What should we do?

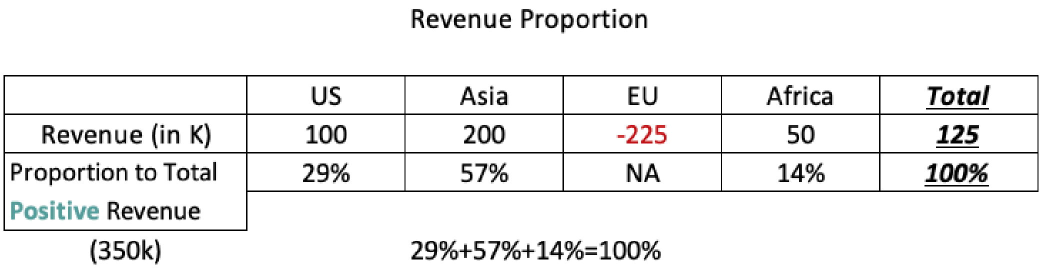
As a data analyst:

1. When calculating proportion, first check whether there are negative values.
2. If yes, calculate the proportion based on the positive ones, but also report how the negative sector influence the overall market.

As a decision maker:

If someone reports a proportion that exceeds 100%, ask for the detailed calculation and make sure you understand the interpretation. A 200% increase rate can make sense during a year-to-year comparison, but a 200% representation for a slice of a pie usually means you overlooked the negative factors.

In the case above, we can update the proportion by only using the total positive value as denominator like this:



Revenue Proportion

	US	Asia	EU	Africa	***Total***
Revenue (in K)	100	200	-225	50	***125***
Proportion to Total Positive Revenue (350k)	29%	57%	NA	14%	***100%***

29%+57%+14%=100%

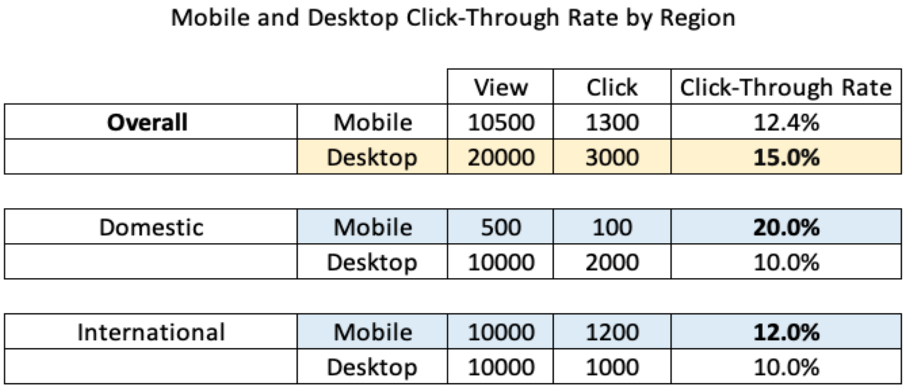
Case 4 : Your Subgroups Might Show Different Trend Than The Entire Population

Is it possible that the trend in the subgroups is completely different from the total population?

Example:

A company is launching a new marketing campaign. They want to measure the click-through rate (CTR) for Mobile and Desktop devices to see which one has a better response.

From the table below we can see that, the overall CTR is higher for desktop users (15%) than mobile users (12.4%). Should we conclude that this campaign is more effective for desktop users?

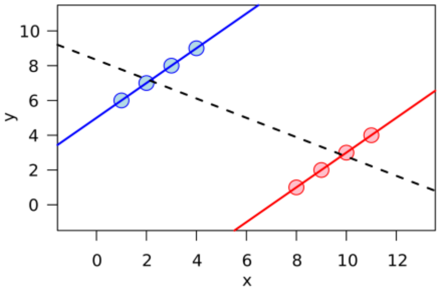


Mobile and Desktop Click-Through Rate by Region

		View	Click	Click-Through Rate
Overall	Mobile	10500	1300	12.4%
	Desktop	20000	3000	**15.0%**
Domestic	Mobile	500	100	**20.0%**
	Desktop	10000	2000	10.0%
International	Mobile	10000	1200	**12.0%**
	Desktop	10000	1000	10.0%

Let’s take a closer look. This campaign has launched in different regions. After we break it down to domestic and international, we found that both regions’ Mobile users had higher CTR than Desktop users, which is the opposite to the overall pattern! We would completely miss this effect if we didn’t break it down by regions.

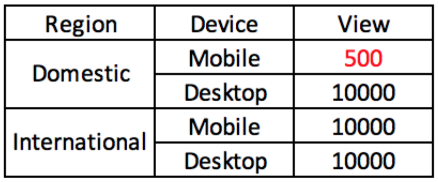
In Statistics, this effect is called Simpson’s paradox when subgroups’ trend is opposite to the overall trend.



This paradox is likely to occur when:

1. You are using a ratio metric.
2. There are different subgroups in your population.

3. There is an imbalance in the sizes of two subgroups. In the example above, the domestic mobile user views are significantly lower (only 500), although the domestic mobile CTR is higher, the count of total views for desktop is higher, which influenced the overall desktop CTR to be higher.



Region	Device	View
Domestic	Mobile	500
	Desktop	10000
International	Mobile	10000
	Desktop	10000

So, when the subgroups with smaller size have higher rate, the effect might get diluted when we calculate the overall rate of the entire population.

What should we do?

1. When you calculate a ratio metric, check the subgroups to see if the trends are the same as the overall population.

2. If you are designing an experiment and want to avoid this situation, make sure the sample is evenly distributed across different subgroups. You can have different variables like location, gender or age group, and you might not need to check the balance across all those subgroups exhaustively, but definitely check the distribution of those factors that might have an impact on the metric.

Takeaways For Your Data Analytics Strategy

Data analytics is not just calculation, it's also the measurement of uncertainty.

While summary statistics like mean or some ratio metrics help us 'Zoom Out' and see a big picture of data and our business, we also need to 'Zoom In' for the range, shape and subsets of data, to make sure we understand the uncertainty associated with the metrics.

- A data point is NOT enough! Create the range around it, and use variance to estimate uncertainty or different subgroups of the data.

- If your metric is a ratio like Click-Through Rate, analyze different scenarios to see how the metric changes as the denominator and numerator change. Be careful if your denominator is small, which means the ratio can be more sensitive to the change of data, and might not be reliable!

- Pay attention to how the proportion is calculated and whether there's any negative number involved in the entire pie.

- Be aware of "Simpson's Paradox", especially when the subgroups' sizes in your analysis are imbalanced, and you are using ratio metrics.

Need help with your data analytics strategy? Feel free to email zhen@dataforleaders.com

Want to get more free tutorials on data analytics for your business? Sign up on dataforleaders.com