Cohort Analysis – Useful in Theory, Colorful Noise in Practice?

Ronny Kohavi
Oct 16, 2023 (updated Oct 17, 2023)

This document is at https://bit.ly/CohortAnalysisNoisyRonnyK

LinkedIn Post:
https://www.linkedin.com/posts/ronnyk_is-cohort-analysis-useful-i-have-seen-this-activity-7119685152150077440-DNvB

TL; DR: Cohort analysis supposedly "allows you to see patterns clearly against the lifecycle of a customer" (Croll and Yoskovitz 2013). In Reforge's Growth Series Series (Fishman, Balfour and Chen 2023), cohort matrices are touted as a great way to measure retention over time. The theory seems to make sense, and the colorful graphs, such as the one in Figure 1, are beautiful. However, without any measure of statistical significance, and with such fine-grained segmentation of the population into small cells, it is likely to be showing colorful noise. Furthermore, the analysis of lagging metrics like retention is unlikely to provide many actionable insights.

Retention Cohort Percentage

Sign Up Date	New Users	Weeks After Sign Up										
		0	1	2	3	4	5	6	7	8	9	10
1/8	604	88%	79%	72%	65%	58%	55%	51%	49%	48%	47%	46%
1/15	734	91%	80%	72%	63%	55%	52%	50%	48%	48%	47%	45%
1/22	867	95%	89%	83%	77%	71%	67%	63%	60%	58%	57%	
1/29	912	88%	79%	72%	65%	58%	55%	51%	49%	48%		
2/5	544	92%	84%	74%	68%	64%	63%	61%	60%			
2/12	1,170	81%	75%	68%	62%	56%	55%	54%				
2/19	1,246	88%	74%	68%	60%	56%	54%					
2/26	1,385	94%	83%	72%	66%	59%						
3/5	1,429	86%	77%	70%	64%							
3/12	1,577	90%	82%	74%								
3/19	1,611	91%	81%									
Average		89%	80%	72%	65%	59%	57%	55%	53%	51%	51%	46%

Figure 1: Temporal cohort analysis of retention by week (Fishman, Balfour and Chen 2023)

Introduction to Cohort Analysis

Cohort studies are longitudinal studies that sample cohorts (a group of people who share a defining characteristic), such as an event (e.g., opening a user account) (Wikipedia Cohort Study 2023). The premise is reasonable: follow the cohort over time to see the long-term behavior (e.g., churn).

Dunkley (2022) writes that "cohort analysis can give us more details on who users are and thus help us narrow why they may be engaging more or less with the product based on cohort attributes." She writes about the following benefits:

1. It allows you to track the customer life cycle of specific groups to get a clearer view of how and when your customers engage with your product or business, and see customer lifetime value.
2. It can help you understand the long-term and current health of your business and predict possible churn.
3. It can tip you off to whether or not changes in your site or product are affecting users because you can zoom in and see if their change in behavior coincides with a launch of a new product feature or change to the website.
4. It can help your company make better and more informed decisions and predictions, as you can develop targeted campaigns for your most valuable users.

Stancil (2015) claims that retention rate by weeks after signup has become "one of the most important measures of health for many companies." He then notes that there is a major drawback to such temporal cohort analysis: they don't tell you what to do next. "If retention is falling, how do we fix it?" He points out that the categorization by user sign-up "isn't terribly helpful because you can't get more users to sign up last month." He recommends to cohort users in other ways, such as by device (10 phone types, 10 desktop types) or ads or browser (Chrome vs. Internet Explorer/Edge vs. Firefox). He then shows cohort matrix broken by 12 languages on the y-axis and 12 weeks on the x-axis as reproduced in Figure 2, and another matrix for devices, showing over 20 devices on the y-axis and 12 weeks on the x-axis.

Retention rate by language

Retention rates by weeks after signup

Language	New Users	1	2	3	4	5	6	7	8	9	10	11	12
arabic	105	13%	9.5%	9.5%	10%	10%	6.7%	8.6%	8.6%	2.9%	1.9%	1.9%	
chinese	93	17%	9.7%	5.4%	11%	8.6%	7.5%	11%	8.6%	6.5%	3.2%	3.2%	
english	1,451	18%	15%	15%	13%	12%	10%	9.2%	5.7%	3.8%	2.3%	1.2%	0.21%
french	242	19%	16%	16%	14%	9.9%	7.4%	6.2%	2.5%	2.5%	1.7%	0.41%	0.41%
german	160	19%	13%	8.1%	10%	9.4%	6.9%	6.3%	7.5%	5.0%	1.3%	3.1%	1.3%
indian	90	19%	17%	19%	12%	18%	8.9%	4.4%	6.7%	5.6%	2.2%		
italian	63	14%	7.9%	16%	7.9%	13%	13%	11%	11%	6.3%			
japanese	182	16%	15%	19%	12%	14%	9.9%	7.7%	7.1%	2.2%	1.6%	3.3%	1.1%
korean	28	7.1%	14%	7.1%	11%	7.1%	7.1%						
portugese	79	24%	11%	10%	13%	18%	14%	11%	6.3%	5.1%	3.8%	3.8%	1.3%
russian	76	16%	7.9%	6.6%	11%	12%	7.9%	13%	5.3%	7.9%	5.3%	2.6%	1.3%
spanish	254	16%	14%	15%	12%	10%	9.4%	4.7%	3.5%	1.6%	2.0%	2.0%	1.2%

Figure 2: Retention rates by weeks after signup (Stancil 2015)

The Statistics Say it is Mostly Noise

In A/B Testing, or randomized controlled experiments, a topic that received scientific attention for almost 100 years now (Fisher 1925), there are clear formulas for the minimum number of users needed to be able to detect a minimum effect. The power formulas tell us that for standard industry conventions (e.g., 80% power, alpha=0.05 threshold for p-values) if your conversion rate is 5% (typical of an e-commerce site, for example), and you want to detect a 5% relative delta, you need over 100,000 users in each variant of an A/B test (Kohavi, Tang and Xu 2020, Kohavi 2022).

Retention is hard to change; really hard. In (Kohavi 2017, slide 27), it was shared that at Bing the Sessions/User metric, which is a proxy for usage and retention, improved only once in about 5,000 controlled experiments. It's hard to introduce features that are so good that they improve retention materially. Note that while Bing is relatively mature, it also has a lot of users, so small effects can be detected. Startups may only be able to detect very large changes to retention.

There are three major problems with cohort analyses in relation to statistical power:

1. Analyses look at new users to understand their lifecycle.
2. Analyses look at new users in very limited durations, commonly a week.
3. The x-axis follows these new users for multiple weeks, and many churn.

Let's say you have about 1,000 new users per week, which matches the numbers in Figure 1. Several weeks later, it is likely that you lost about half the new users, so we are looking at under 500 users per cell and the cells are auto correlated. Figure 2 takes the new users and splits them into 12 languages. After English, Spanish has the most users, at 254 and the retention rates then take this down by about 90%, so most cells will have fewer than 25 users. These numbers are tiny relative to properly powered A/B tests with 100,000 users per variant. Many of the matrices are just a tapestry of noise.

Analyzing a Cohort Matrix

Here is an example of a cohort analysis matrix showing churn rates with 1,000 new sign-ups per week:

Sign-up week\week after signup	1	2	3	4	5	6	7	8	9	10
5/1/2023	6.2%	6.5%	5.0%	5.6%	7.3%	5.2%	5.2%	5.6%	4.8%	5.5%
5/8/2023	5.7%	5.3%	5.1%	6.4%	5.8%	4.5%	5.7%	6.7%	5.0%	6.8%
5/15/2023	6.3%	7.4%	6.3%	5.7%	5.8%	5.3%	6.5%	6.7%	6.6%	
5/22/2023	5.0%	5.7%	5.7%	5.2%	5.4%	6.0%	4.5%	6.1%		
5/29/2023	5.7%	6.8%	5.4%	5.7%	5.4%	6.4%	7.4%			
6/5/2023	5.9%	5.9%	5.6%	6.7%	5.9%	5.2%				
6/12/2023	5.7%	6.1%	5.5%	5.3%	6.7%					
6/19/2023	5.5%	5.3%	5.1%	6.2%						
6/26/2023	6.5%	6.3%	7.2%							
7/3/2023	7.3%	6.5%								

The patterns that one can see, and recommended action items:

1. New signups on the week of 5/22/2023 had lower than normal churn.
   Investigate whether there was an ad campaign, or holiday, or what product features were live at the time. Note that it appears we have degraded slightly the week after and again two weeks after.
   Was something turned off?
2. Major concern on the week of 7/3/2023, as churn was extremely high for week 1.
   We've never had such high churn.
   Investigate whether there was an outage, or what product features launched. Is it related to 4th of July holiday?
   Note that the week after is also higher churn than average. Very concerning.
3. The diagonal has generally high churn rates (mostly orange and red, except 6/5/2023 week 6).
   Since they all represent the same week of the year, investigate what happened that week.
4. It is now two weeks (6/26/2023 and 7/3/2023) where we have above average churn.
   Look at what features launched and understand why. Genuinely concerning.

If I convert the above into a retention matrix, where each cell is not independent, but impacted by the prior week, you get more patterns:

Sign-up week\week after signup	1	2	3	4	5	6	7	8	9	10
5/1/2023	93.8%	87.7%	83.3%	78.7%	72.9%	69.1%	65.5%	61.9%	58.9%	55.6%
5/8/2023	94.3%	89.3%	84.7%	79.3%	74.7%	71.4%	67.3%	62.8%	59.6%	55.6%
5/15/2023	93.7%	86.8%	81.3%	76.7%	72.2%	68.4%	63.9%	59.7%	55.7%	
5/22/2023	95.0%	89.6%	84.5%	80.1%	75.8%	71.2%	68.0%	63.9%		
5/29/2023	94.3%	87.9%	83.1%	78.4%	74.2%	69.4%	64.3%			
6/5/2023	94.1%	88.5%	83.6%	78.0%	73.4%	69.6%				
6/12/2023	94.3%	88.5%	83.7%	79.2%	73.9%					
6/19/2023	94.5%	89.5%	84.9%	79.7%						
6/26/2023	93.5%	87.6%	81.3%							
7/3/2023	92.7%	86.7%								

5. 5/15/2023 is a terrible cohort.
   What have we done to these poor people? Perhaps send them a free coupon.
6. 5/22/2023 is a great cohort right after.
   Is it possible that some churned people rejoined the week after after erasing their cookies?
   It's very interesting to see such consistent patterns week after week.
7. 7/3/2023 is terrible, consistent with above analysis.
   Fourth of July? Major downtime?

Now, lookup the definition of apophenia. The cohort matrix has the exact same underlying churn rate for all cells using a binomial distribution with 6% and 1,000 users. This is the first random set that came out and there was no attempt by me to find a particular interesting seed for the patterns.

In controlled experiments, we recommend running A/A tests. This is the equivalent for cohort analyses.

Retention Line Graphs are Dubious Averages

The Growth Series (Fishman, Balfour and Chen 2023) suggests generating a retention curve by averaging the columns in Figure 1, resulting in a line graph shown in Figure 3. For each week x on the x-axis, we're averaging column x in the matrix. Let's assume that the first row in the table represents the first week of the year. That means that we're averaging the retention of week1 of the year for cohort 1, then week2 of the year for cohort 2, then week3 of the year for cohort3.

There are several problems with this approach:

1. Given the upper diagonal matrix, the last columns are averages of less data. The data for x=10 coming from Figure 1 is an average of just two numbers, the users who were retained since earliest two weeks of the year.
2. If there is a holiday, or an important product improvement that increases retention, it will be averaged in some entries but not in others.
   Imagine, for example, that product improvement doubled retention by the middle of the year. The retention

graph will look decreasing, as the later numbers on the x-axis are not averaging retention numbers with these improvements.

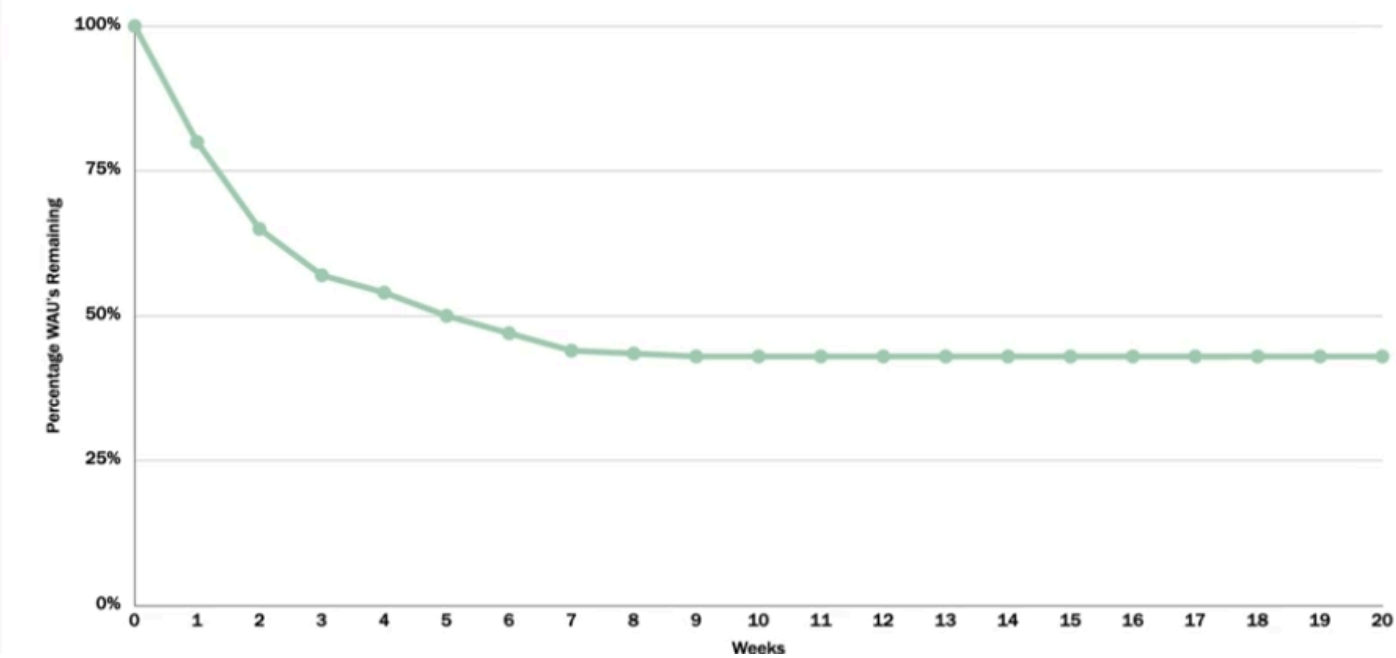

Figure 3: Retention line graph (Fishman, Balfour and Chen 2023)

The Matrices are not Very Actionable

As noted in Stancil (2015) the major drawback to such temporal cohort analysis: they don't tell you what to do next. "If retention is falling, how do we fix it?" Assume that you see a major drop in retention on row 5 for the first four weeks, what do you do?

Let's see why the analysis is hard and unreliable:

1. You released 15 features in the four weeks represented in row 5. Which of them is responsible? The data is observational and establishing causality is impossible.
2. Is it even changes to your product? It is possible that the competitor released a feature or did some marketing activity.

As noted in Wikipedia's Cohort study (2023) "[Randomized controlled experiments] are generally considered superior methodology in the hierarchy of evidence." A/B tests, or randomized controlled experiments, give you the ability to evaluate every feature change, but cohort analysis is highly limited in its actionability.

Summary

Retention is a lagging metric and cohort analysis is a lagging analysis. Dave Kellog (2022) summarized this point nicely as follows:

> While the CEO is on the bridge looking forward, many finance teams are on the stern, offering in-depth analyses of the ship's wake.

Unless you're running one of the largest sites in the world, where the new users in the cells will be large enough, the high variance of new user counts for most web sites and apps mean that the retention matrix will be very noisy. It's pretty, but not meaningful and not actionable.

References

Croll, Alistair, and Benjamin Yoskovitz. 2013. *Lean Analytics: Use Data to Build a Better Startup Faster.* O'Reilly Media.

Dave, Kellogg. 2022. *Thoughts on Lagging, Leading, and Predictive Indicators.* 10 27. https://kellblog.com/2022/10/28/slides-from-a-cfo-summit-on-leading-and-lagging-indicators/.

Dunkley, Chioma. 2022. *Cohort Analysis: An Introductory Guide for Better Retention.* March 28. https://mode.com/blog/cohort-analysis/.

Fisher, Ronald Aylmer. 1925. *Statistical Methods for Research Workers* .

Fishman, Adam, Brian Balfour, and Andrew Chen. 2023. *Growth Series.* October 3. https://program.reforge.com/cohorts/2023-fall-growth-series-revamp-eg.

Kohavi, Ron. 2017. *Online Controlled Experiments: Lessons from Running at Large Scale.* http://bit.ly/CH2017Kohavi.

Kohavi, Ron. 2022. *Practical Defaults for A/B Testing.* Nov 19. http://bit.ly/CH2022Kohavi.

Kohavi, Ron, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing.* Cambridge University Press. https://experimentguide.com.

Stancil, Benn. 2015. *Cohort Analysis That Helps You Look Ahead.* May 11. https://mode.com/blog/cohort-analysis-helps-look-ahead/.

Wikipedia Cohort Study. 2023. "Cohort Study." October 4. https://en.wikipedia.org/wiki/Cohort_study.