Checking the math on personalized send time

Author: Christina Dedrick

Claps: 74

Date: Sep 26, 2019

Last month, we wrote an article arguing that an A/B test was the best way to find the optimal send time for a group of customers. We said it would be impractical to A/B test at the individual level, and in this post weâ€™ll explain why. Weâ€™ll start with some basic intuition and then take you through the full mathematical explanation of why it doesnâ€™t work.

Letâ€™s imagine a brand is trying to find the best time to send you messages. They want to A/B test send times to pick a winner among the different times of day. How do they set up this test? They canâ€™t send you the same email multiple times and see how you interact â€” youâ€™re going to report them for spamming if theyâ€™ve sent you the exact same subject line and contents multiple times (if your inbox doesnâ€™t already do that for you!). So theyâ€™ll need to organically send you different emails at different times of the day on different days to try and see how you respond.

Weâ€™ve already run into the first problem with A/B testing for a single person â€” weâ€™re not testing the same thing in each email. Each email has a different subject line. We have to be able to correct for different urgency or overall interest so we can compare ignoring a subject line like â€œOur fall favoritesâ€¦ to opening one that says â€œBack to school sale starts now, coupon inside.â€¦ We have to correct for product interest â€” even similar subject lines like â€œShirts you canâ€™t missâ€¦ and â€œPants styles you canâ€™t missâ€¦ might cause problems in comparisons. If you signed up for the newsletter with the intention of refreshing the pants in your wardrobe, youâ€™re likely to be much more interested in subject lines about pants and tend to click on those. To properly compare emails, we a way need to compare how an open of the subject line â€œThe Fall Lookbookâ€¦ compares to an open of â€œThe countdown is onâ€¦â€¦ and how those compare to an ignore of â€œ$17.99 Checked Shirtâ€¦ and an ignore of â€œwe heard you like surprisesâ€¦â€¦ We need to be able to normalize for the effectiveness of different subject lines so that our send time test doesnâ€™t become a test of what subject lines you respond to as a recipient.

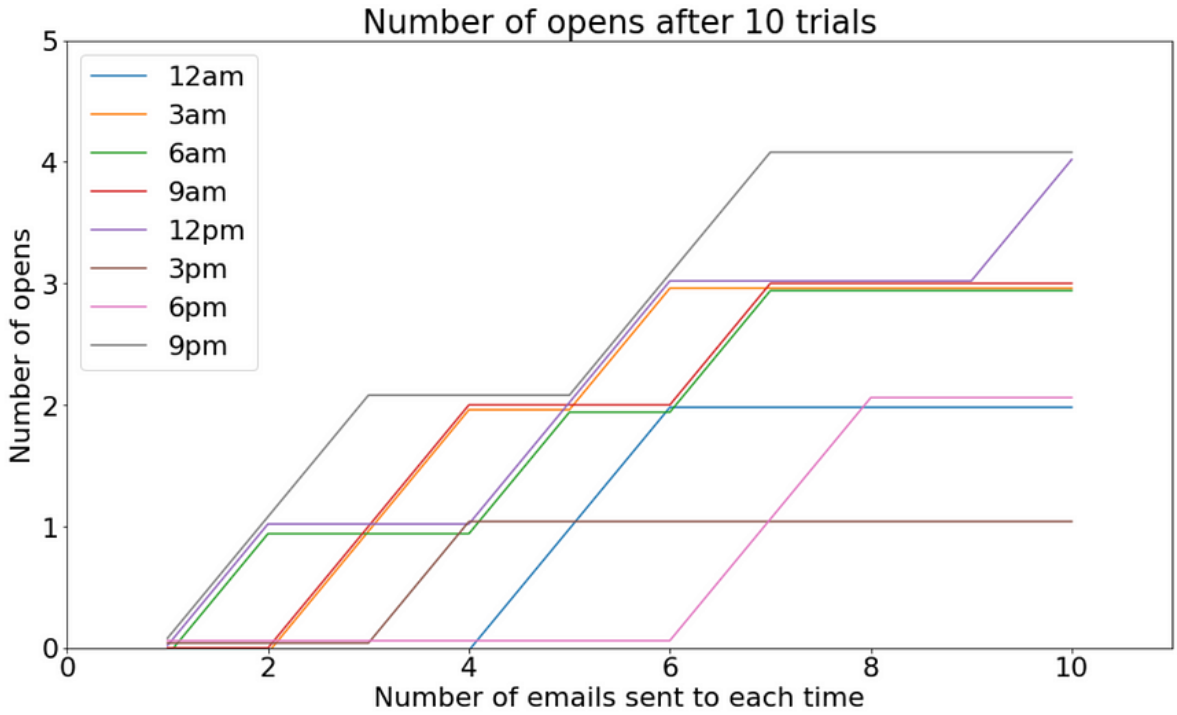
Back to the experiment. Our brand will pretend that subject line wonâ€™t be a problem. They have another problem to tackle, they need to test the entire day. Even if they have a lot of historical data that shows that you open emails theyâ€™ve sent at 10am at 10:30am, youâ€™ve never received an email at a different time of the day, so theyâ€™ll have to test those. They canâ€™t rule out that you wonâ€™t do better with an early morning, afternoon, evening, or even overnight send time. So theyâ€™ll have to test the entire 24 hour day â€” even if that breaks down into 3 hour blocks, thatâ€™s still 8 test periods they need to cover. But how many emails do they need to send you in each hour?

They start with one. Over the course of 8 campaign sends, they send one at 12am, 3am, 6am, 9am, 12pm, 3pm, 6pm, and 9pm. You ignore them all. What do they do now?

They send more emails to get more data. They send another 8 campaigns, this time you open the one sent at 6am at 9:30 am, the one sent at 3pm at 3:45 pm, and the one sent at 9pm at 10am the next day. You ignore the rest. Now theyâ€™ve seen that there are 3 hours of the day that youâ€™ve opened an email. But, they also know that one data point is not enough to make a decision, so they send another batch of 8 emails.

This time you open the ones sent at 3am at 9:45am, 12pm at 5pm, and 9pm at 6:30am the next day. The score is now 12am: 0, 3am: 1, 6am: 1, 12pm: 1, 3pm: 1, 6pm: 0, 9pm: 2. Itâ€™s still not enough to make a decision â€” there is no consistency between the three rounds of sending! The only thing weâ€™ve learned so far is that your open time isnâ€™t necessarily related to the time you receive an email.

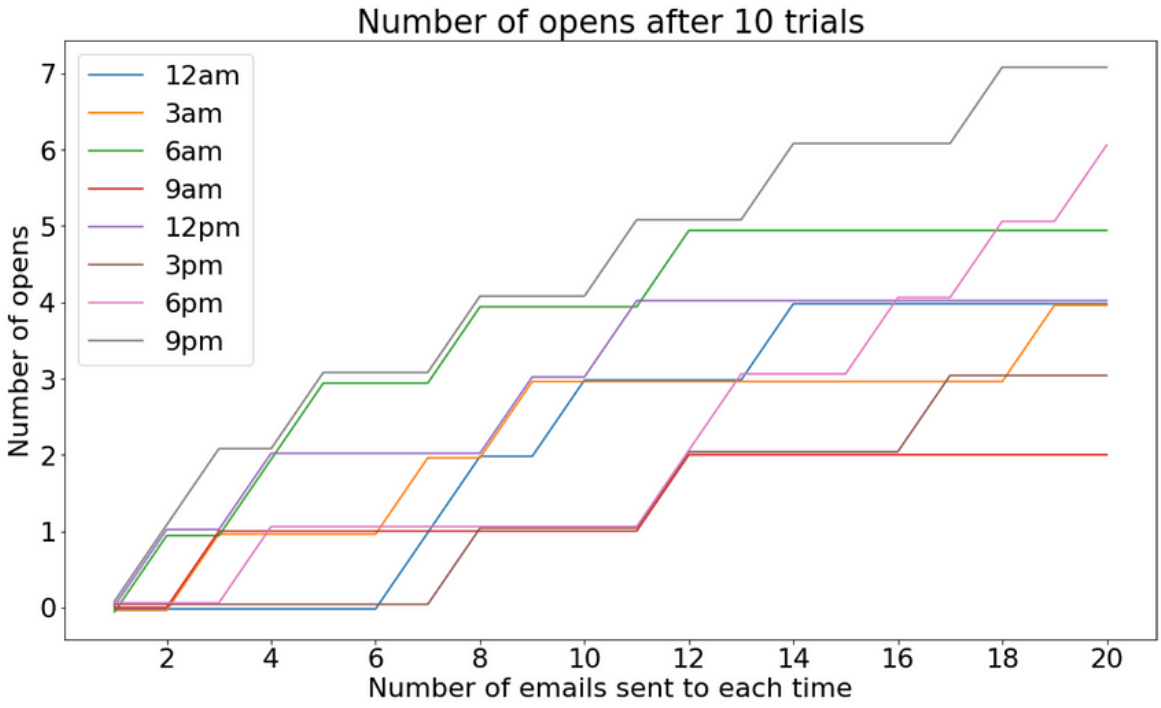
Letâ€™s plot what might happen if they keep repeating this process. Below we show the number of opens that might occur over time. On the x-axis, we have the number of emails we have cumulatively sent to each hour. On the y-axis, we show the number of total opens. Each colored line represents a different send time. On each send where a customer opens an email, that timeâ€™s total number of opens jumps up by 1.



At the end of this trial, 12pm and 9pm are tied at 4 opens each. 3am, 6am, and 9am are close behind with 3 opens each. 12am and 6pm have 2 each. So what is the optimal send time?

For the simulation that created this data, it was 9pm, where we simulated that you opened emails at a 40% rate at that hour. For all other hours, we assumed you opened 20% of the time. 10 trials and a total of 80 emails was not enough to pick an optimal send time for you.

If we run another simulation out to 20 emails (160 sends total, or approximately half a year of sending if the brand is a daily sender), we have a little more certainty about what the best send time was. We show this simulation below, but we see that itâ€™s still pretty close, just +1 more open, between the optimal send time and the next best hour.



Weâ€™ll explain the full math about why itâ€™s taking so long to discover your optimal send time by testing different times of the day below. Before that, letâ€™s address two more questions about our intuitive explanation.

Why canâ€™t we cut some hours off early? Stopping early would let us eliminate obviously terrible times from our experiment, so weâ€™d need to make fewer sends overall. However, it comes at a risk. What if we cut the optimal variation early? Looking at the graph of cumulative opens above, we see that if we had cut the worst performing variations off after 10 sends, we would have cut out the 9am, 3pm and 6pm variations. 6pm went on to perform second best out of all the variations at 20 emails. Because the numbers we are comparing in terms of total open counts are so small, it doesnâ€™t make sense to eliminate variations early because they are not doing that much worse than the ones performing best.

Why canâ€™t we ignore unrealistic times like 3am? As it turns out, 3am isnâ€™t an unrealistic time to send an email to get it read. Based on our data, 3am outperformed 9am in terms of open rates in 39% of Smart Send Time 24 hour exploratory sends. 9am and 10am are among the lowest open rate hours to send an email, contrary to a lot of the advice non-data driven marketing materials provide. Many opens happen at in the 9am-10am range, but significant numbers of people are still opening emails from the day and night before.

Picking an optimal send time for a single recipient through continuous A/B testing requires two things marketers donâ€™t have: perfect subject line comparison and time. Marketers would have to send hundreds of emails to attempt to learn about a recipientâ€™s behavior before being able to capitalize on the boost in open rate.

Letâ€™s get into the math about why itâ€™s so hard to pick out an optimal send time.

Weâ€™ll start assuming you have the 20% -> 40% boost in open rates as we did above. However, this +100% open rate is completely unrealistic based on the data weâ€™ve collected about optimal send time. The lift weâ€™ve seen has been +10%, and weâ€™ll show that makes

it orders of magnitude harder to identify a personalized optimal send time on any reasonable timescale.

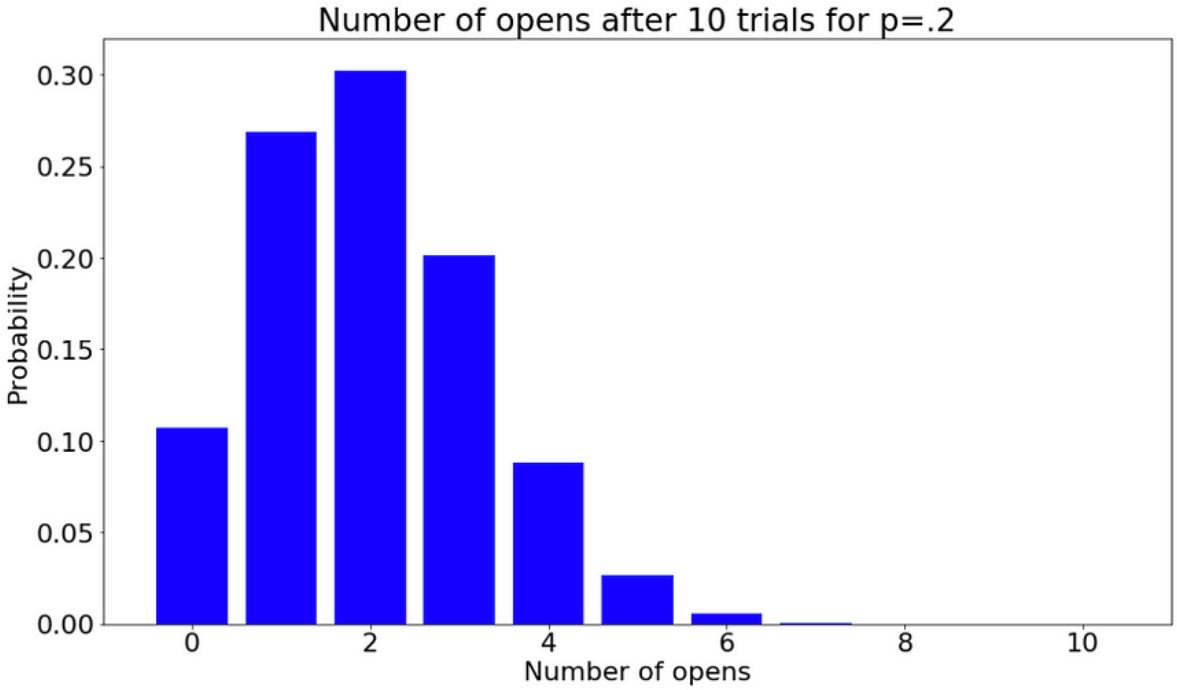
To model whatâ€™s happening mathematically, weâ€™ll treat each email we send is a separate, independent event. Weâ€™ll keep making the assumption that we have perfect subject line comparisons and test the entire day as 8 three hour blocks. Weâ€™ll also assume the 20% underlying open rate at the 7 non-optimal send times and 40% at the optimal send time. Weâ€™ll start by sending our recipient 10 emails to each block.

A 20% email open rate doesnâ€™t mean that the recipient opens exactly 1 in 5 emails they receive. Each email is an independent Bernoulli random event, so there is a lot of variation in how many emails they open. On average, they will open 2 of the 10 emails, they may actually open 0, 1, 3, 4, or even 5 or more of them.

The outcomes of multiple Bernoulli trials are modeled as a binomial random variable. So, the number of successful email opens, k, after $n = 10$ emails can be modeled as a binomial distribution $B(k,n,p) =$

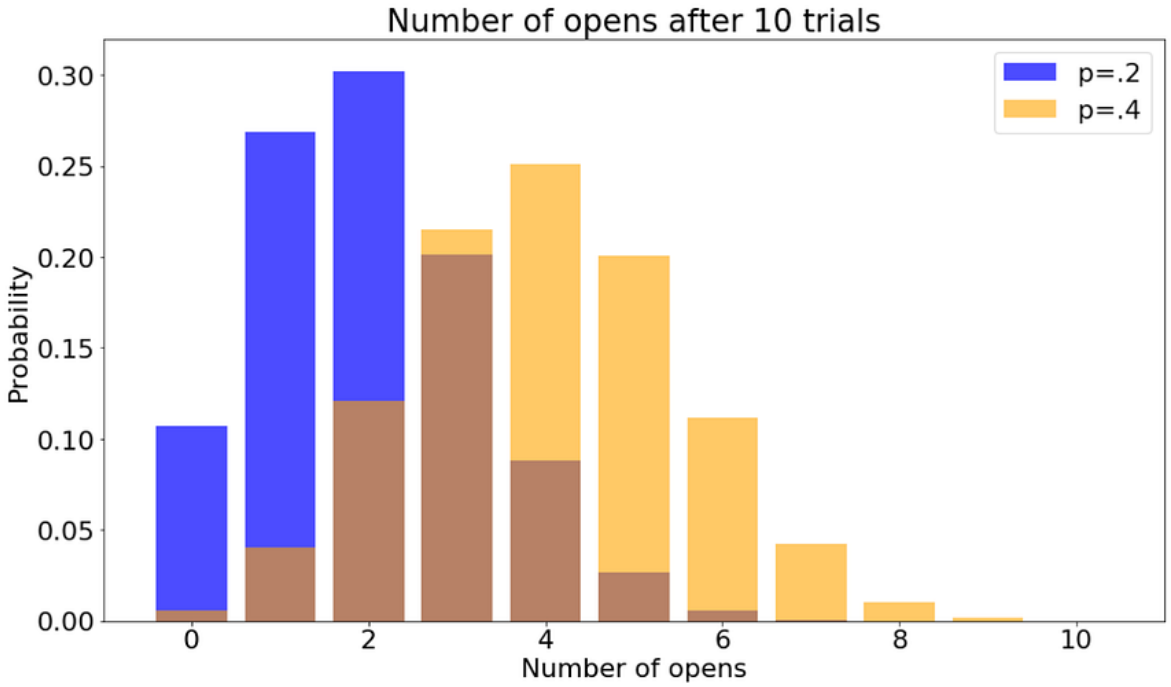
$$\binom{n}{k} p^k (1-p)^{n-k}$$

where $n = 10$ and $p = .2$. This distribution has an expected value of $n*p = 2$ opens and a variance of $n*p*(1-p) = 1.6$ opens. We show the probability mass function, the PMF, of the number of observed opens after 10 events below:



In the PMF, the height of the bars shows how likely each event is to occur â€” so itâ€™s most likely that our recipient opens 2 emails, but 1 and 3 are also very likely. 0 or 4 are fairly likely to be observed. Seeing 5 or more opens is possible but is a rare event for our non-optimal sending periods.

Weâ€™ll contrast the behavior in the non-optimal sending periods to the period of optimal sending, where there is a 40% chance of opening an email. For the 10 trial binomial distribution, the expected value is 4 opens and the variance is 2.4. At 10 emails, the variance is large enough such that the distributions are not very separated. We show an overlay of the two below:

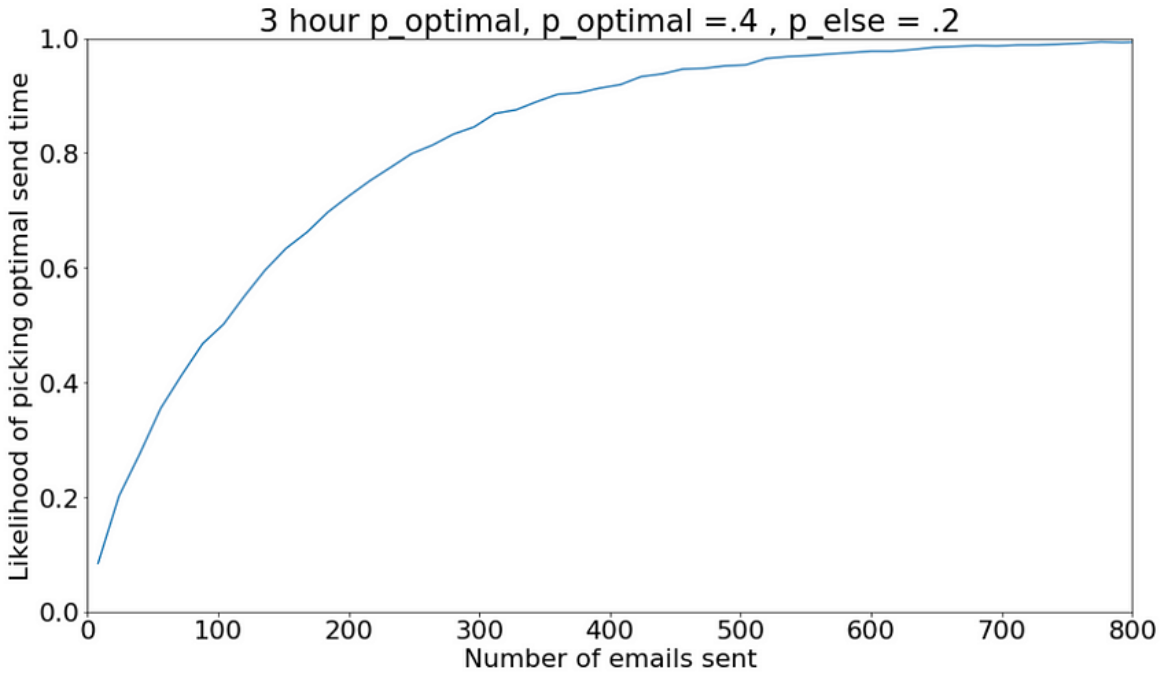


Since the two distributions overlap so much, it is possible to see fewer opens during the period of optimal sending. We might see 4 opens during one of our periods where p = .2 and only 3 opens during the optimal sending period where p = .4. In this case, weâ€™d incorrectly identify the optimal send time since weâ€™d see a higher open rate at a non-optimal time.

How often does the optimal send time have the highest number of opens? To do this, weâ€™ll start with a direct comparison of one of the non-optimal p = .2 hours to the optimal p = .4 hour. We can find the joint probability of observing exactly $k(p=.2)$ and $k(p=.4)$ opens as the product of the two binomial probabilities. Then, we can find all $k(p=.2)$ and $k(p=.4)$ pairs where $k(p=.2) < k(p=.4)$. For all those pairs, we sum the joint probabilities, and find that the probability of identifying the optimal send time in one comparison is 0.895.

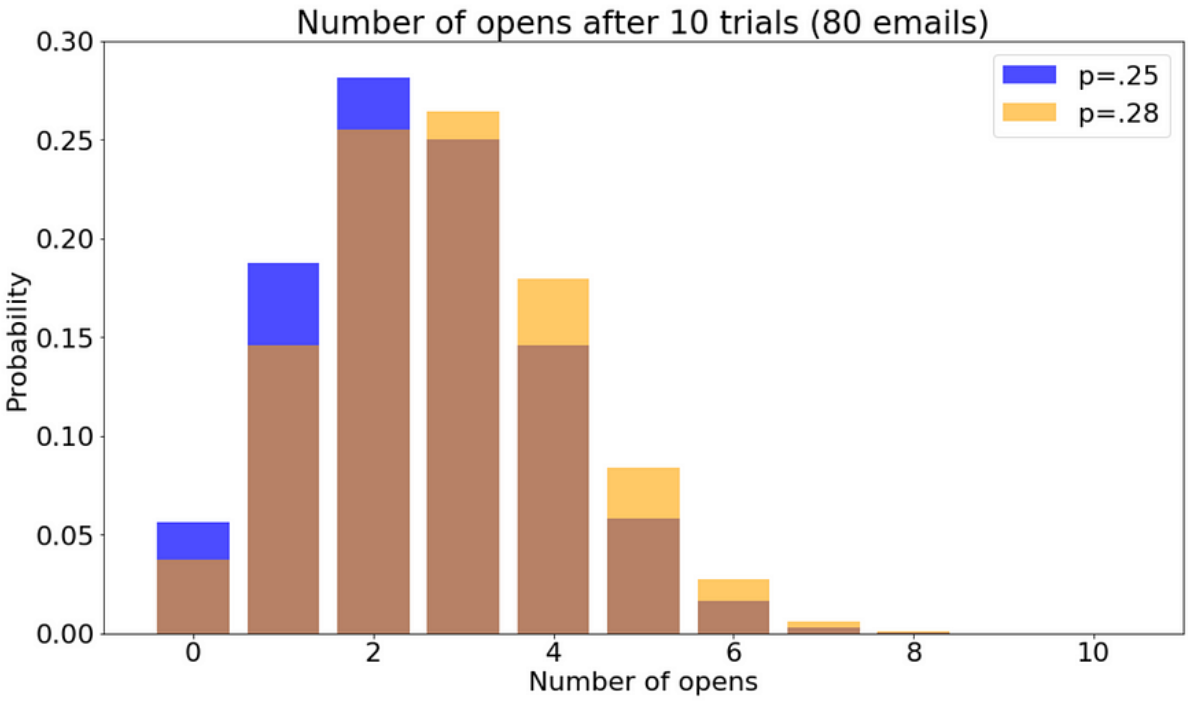
But thatâ€™s only one out of the seven comparisons we need to do since weâ€™re testing 8 different send times. The probability we make the correct comparison 7 times is (0.895)â·= 0.459. So, with 80 emails, weâ€™re at just under a coin flip at predicting our recipientâ€™s best send time.

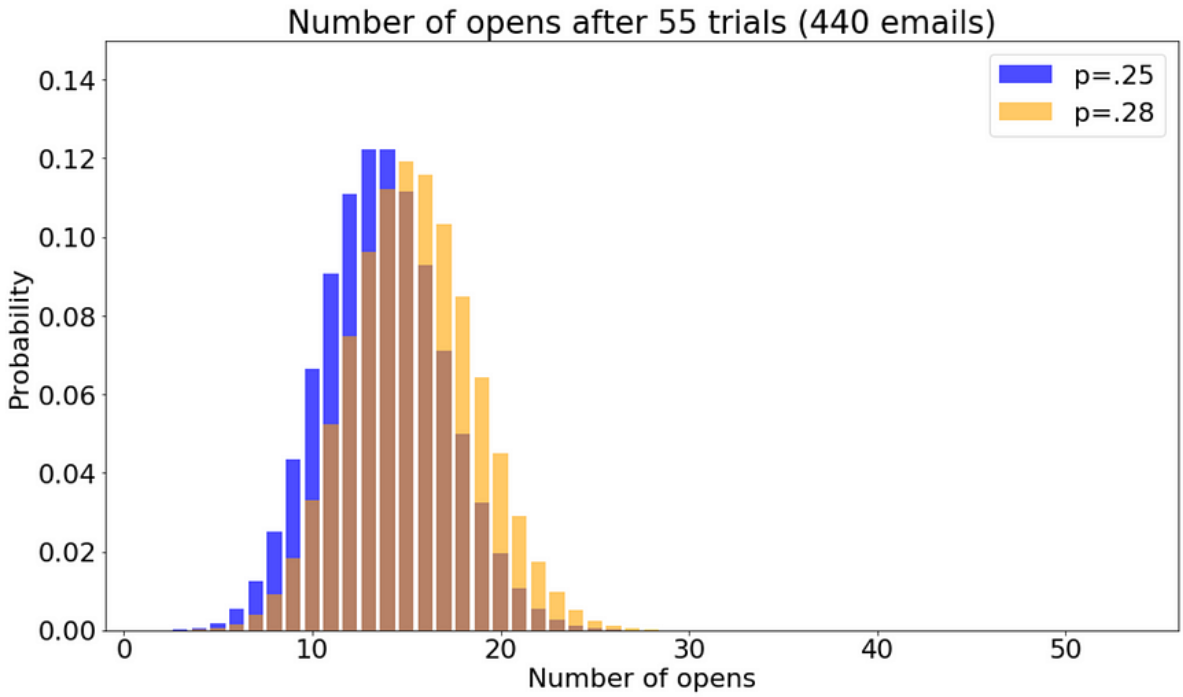
To visualize how the probability of picking the best send time changes as we increase the number of emails we send to a recipient, we plot the probability for picking the optimal send time vs the number of emails below. We generated this data by running a Monte Carlo simulation of 10,000 customers opening different numbers of emails at the probabilities of p= .2 and p = .4. For each number of emails, we counted the number of customers where the highest open rate time was the optimal send time with the p = .4 open rate.



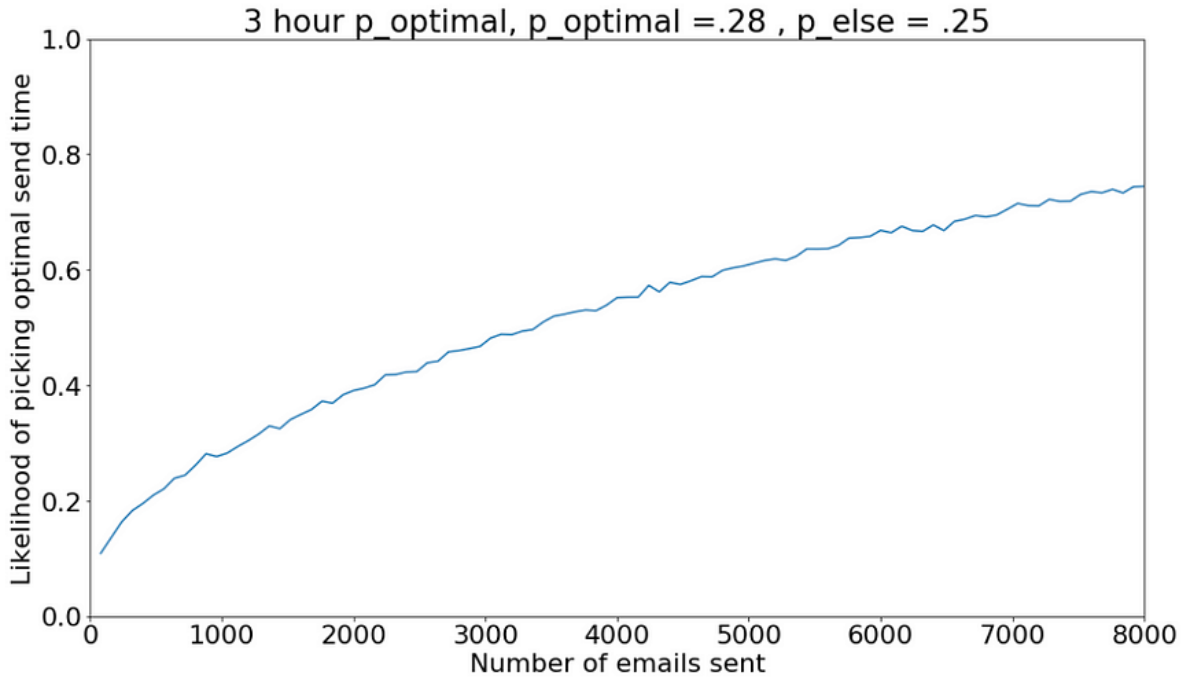
It takes 440 emails to guess a personâ€™s optimal send time correctly 95% of the time. So, even at a daily sending rate, thatâ€™s still over a year of experimenting before we can gather enough data to predict the personalized best send time with any sort of accuracy. But this is the best case scenario. The 20% to 40% lift in open rate, a +100% lift, was overly optimistic. Based on our experiments, we expected recipients to have a +10% increase in open rate at optimal send time.

Letâ€™s do the math again, and this time weâ€™ll assume we have a more realistic 25% open rate at the non-optimal hours and a 28% open rate at the optimal send time. We can compare the binomial PDFs for the two open rates and we see that weâ€™re completely hopeless at distinguishing the two distributions at 80 or even 440 emails:

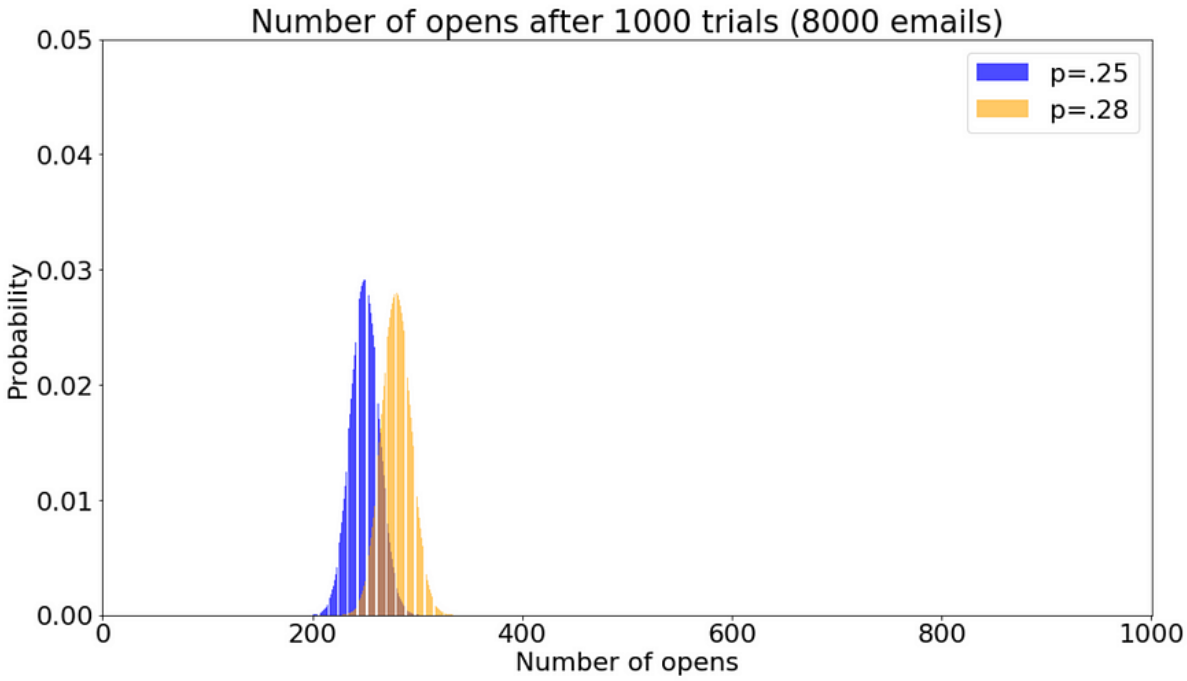




We show the probability of correctly identifying optimal send time below, again generated by Monte Carlo simulation:



It takes a lot of emails, about 3500, or 10 years of daily sending, to even reach the coin flip of picking optimal send time. By 8000 emails, weâ€™ve gotten some decent differentiation. We have about a 75% chance of correctly identifying the optimal send time for recipients.



A/B testing individualized send times will not pay off on any reasonable timescale. Individualized A/B testing or optimization strategies are more likely to overfit to a send time that occurs due to statistical noise than find a recipientâ€™s best send time, as we saw in our experiments.

Models that claim to predict personalized send time canâ€™t collect and train on enough data to make accurate predictions. Instead, they overfit to hours where higher open rates were observed by chance. While developing Smart Send Time, we tested how personalized send time models performed. It wasnâ€™t great â€” in fact, some groups of recipients had lower open rates at their personalized send times than control times. Overall, we only saw a +2.5% in open rate lift for personalized tests which was primarily due to shifting send times later in the day compared to the +10% we observed with the Smart Send Time list wide A/B test. Since both the theoretical math and experimental results agreed that it would not provide value to personalize send time, we focused on building a framework to optimize list wide send time instead, something that we could consistently A/B test.