Turning Fails into Wins

Grace Tang, Data Scientist

UBER

Image: pixabay

“Fails” in the title of this talk refers to failed experiments.

How many of us have painstakingly run experiments only for them show that your strategy had absolutely no effect? Or worse, harmful effects?

Image: https://pixabay.com/p-216986/?no_redirect

Image: pixabay

Many of us would label these experiments “failures”.

But in this talk I’m going to try to convince you that these “failures” can actually be extremely beneficial.
And conversely, some apparent “wins” can lead to detrimental effects.

First, let me give you a quick intro on what I do.
I'm part of a small team of behavioral data scientists at Uber, and one of the things we do is help various other teams in the organisation design and run experiments.
We all come from an academic research background, so we borrow best practices from academia and apply them within Uber. However, academia has a serious bias…

Image: http://www.publicdomainpictures.net/pictures/70000/velka/mortar-board-graduate-cap.jpg

“Win”

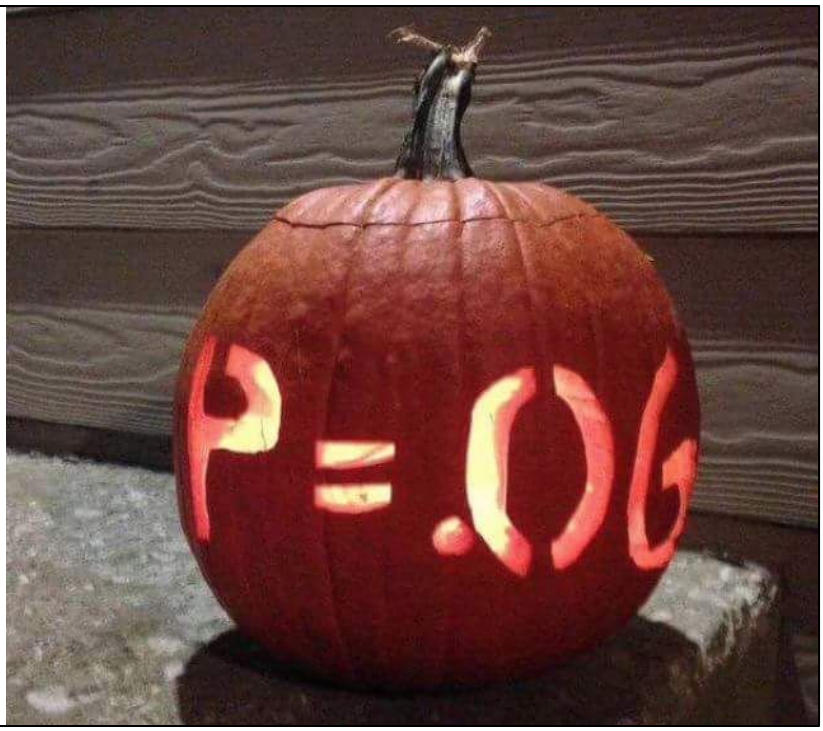
$p < 0.05$

*p-value: the probability that the data would be at least as extreme as those observed, if the null hypothesis were true

...and that bias is towards significant results.
In other words, an experiment that yields a p-value of less than 0.05 is considered a “win”.

“Fail”

$p > 0.05$



Conversely, if our test yields a p value of more than 0.05, this is usually labeled a “fail”.
So what happens to the strategies that don’t result in a significant effect?
Our result is swept under the rug, never to be spoken of again.

$p < 0.05$ ➜ **$p < 0.005$**

Redefine Statistical Significance (September 2017)

This number 0.05 is so ingrained in stats that sometimes we forget it's an arbitrary cut off.
In fact, this may not be the threshold for much longer - there was a proposal this year to change the cut off from 0.05 to 0.005.

https://www.nature.com/articles/s41562-017-0189-z

- Common experimentation **pitfalls**
- **Best practices** to ensure *all* experiments are useful

So, we celebrate when we get p less than 0.05 or 0.005, but sometimes these so-called wins can actually be failures in disguise, leading us to false conclusions and detrimental outcomes.
In the first part of my talk, I'll go over some common pitfalls that lead us to false conclusions.
And in the following part, I'll talk about best practices that we can use to avoid these pitfalls and ensure that all experiments are useful, regardless of their outcome.

Some “wins” are fails

Common experimentation pitfalls

Some “wins” are fails

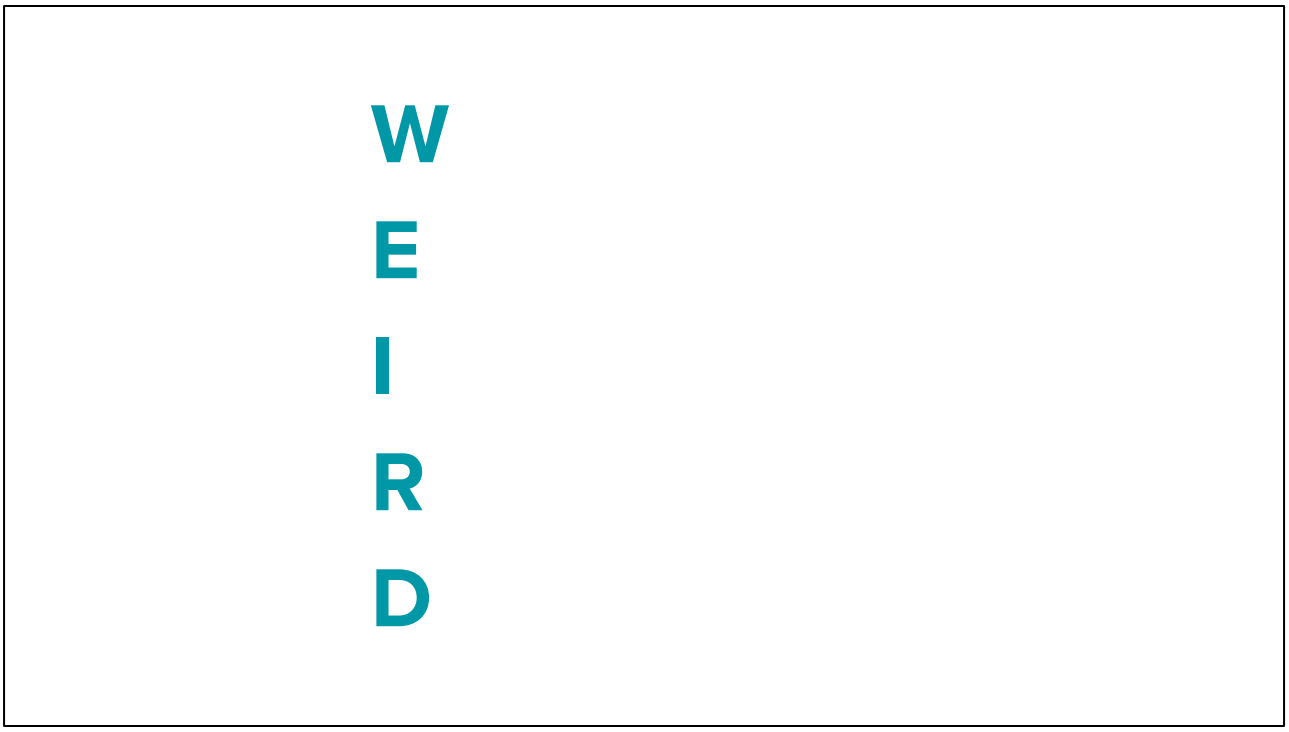
- Biased sampling
- Non-random assignment
- Sample size
- p-hacking

Biased sampling / Cherry picking

Image: wikimedia

Biased sampling, aka cherry picking.
Ideally, when we study a population, we want to ensure that we're selecting randomly from it so that our experiment sample is representative of the people we're trying to study. We want to ensure that there is no systematic bias that would make our sample different in some way from the population.

Images: https://upload.wikimedia.org/wikipedia/commons/b/bb/Cherry_Stella444.jpg



Sadly, academia again is the source of an example of a common bias.
Most academic studies run on human participants tend to have samples that consist mostly of WEIRD people

http://www.apa.org/monitor/2010/05/weird.aspx

Western

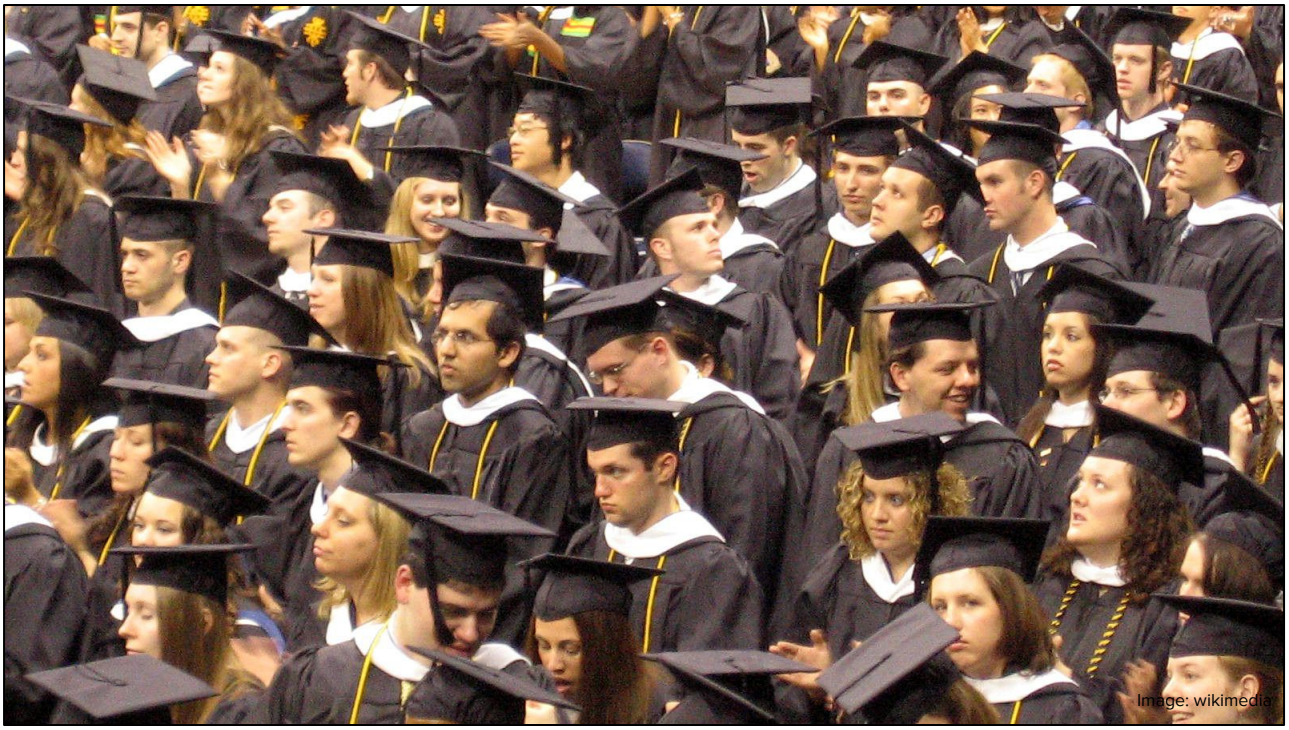
Educated

Industrialized

Rich

Democratic

http://www.apa.org/monitor/2010/05/weird.aspx



Why is this the case? Academic studies are run at universities, and the most convenient sample available to them are their students. About 80% of study participants fall into this WEIRD group, which is not representative of humankind when you consider they only make up 12 percent of the world's population

Images: https://upload.wikimedia.org/wikipedia/commons/3/3a/College_graduate_students.jpg

Non-random assignment

Image: wikimedia

Besides non-random or biased selection from the population, another pitfall is non-random assignment to treatment and control groups.
In other words, we want no systematic differences between treatment and control groups, so if we see a difference between the groups, we can be confident it's because of the treatment and not due to any other pre-existing differences between groups

Images:
https://upload.wikimedia.org/wikipedia/commons/7/7b/Orange-Whole-%26-Split.jpg
https://static.pexels.com/photos/102104/pexels-photo-102104.jpeg

Some the methods we use to assign people to groups do result in bias. Often, it's tempting to use the most convenient means possible to assign people to groups. Say I was running a study in a university testing the effectiveness of a new flu vaccine - the most convenient thing I could do would be to give the vaccine to everyone who showed up at the health center, and compare them to those who did not show up and therefore did not receive the vaccine. But this introduces a systematic difference between the two groups - the people who show up at the clinic might be more concerned for their health, or differ in other meaningful ways from the ones who didn't visit the clinic. So if this group showed a lower rate of catching the flu, we wouldn't know if it was due to the vaccine, or to pre-existing differences between the groups.

Images:
https://pixnio.com/science/medical-science/flu-vaccine-will-protect-against-the-three-influenza-viruses-that-research-indicates

Random assignment...?

- Email
- Phone number
- Name...

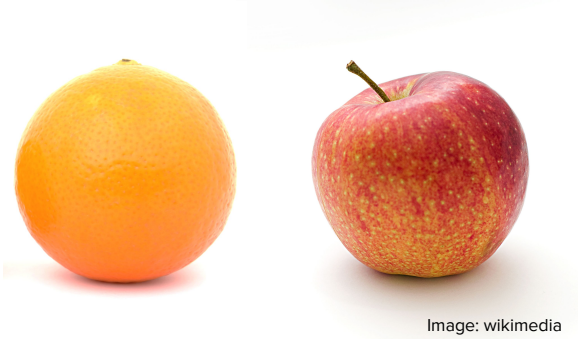


Image: wikimedia

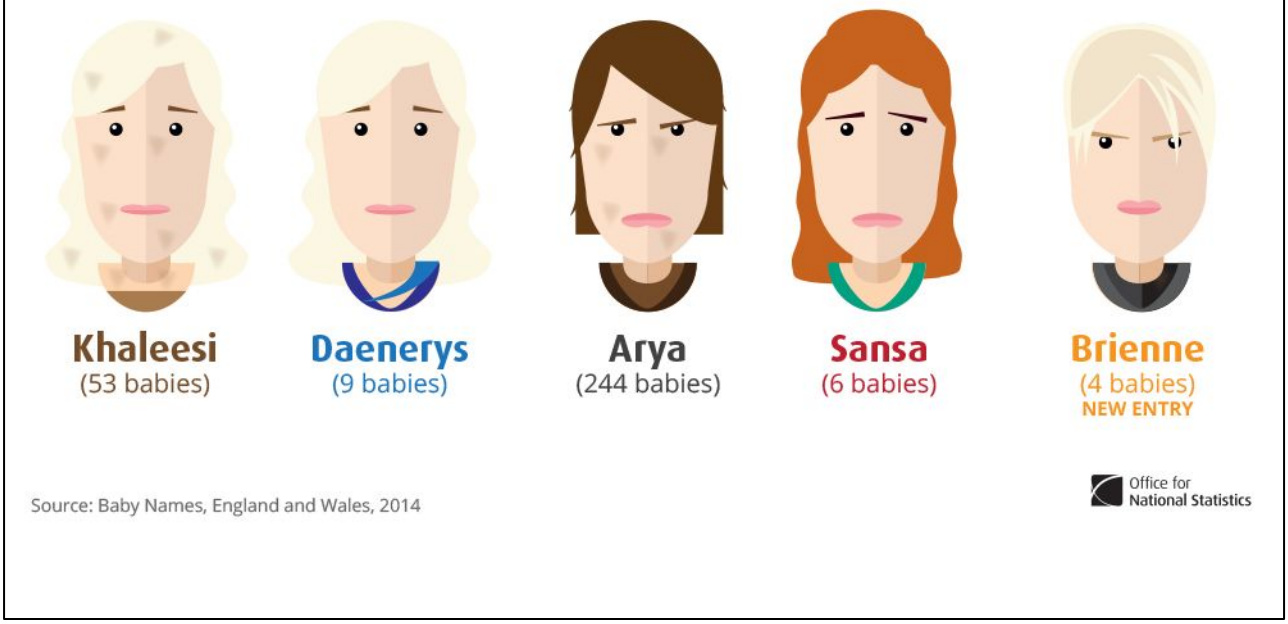
At other times, the features we choose to achieve random assignment might seem random, but aren't.
For example, we could sort people by name alphabetically, assigning everyone with names starting from a-m to the treatment group and everyone else to the control group. Seems innocent enough...

Images: https://upload.wikimedia.org/wikipedia/commons/7/7b/Orange-Whole-%26-Split.jpg
https://static.pexels.com/photos/102104/pexels-photo-102104.jpeg

'Z will be huge': Expert makes his predictions for hottest baby names

But names are not random!
Plenty of factors affect names, like culture, what decade it is...

Source:
http://www.dailymail.co.uk/femail/article-4060728/Z-huge-Expert-makes-predictions-hottest-baby-names-2017-Zander-Zephyr-Zyla-them.html



... and even which season of Game of Thrones is on.
So it's quite easy to accidentally introduce differences in our groups

Images: https://visual.ons.gov.uk/wp-content/uploads/2015/08/BN-post-image-1.png

Opt in bias

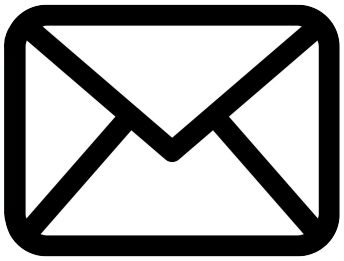


Image: pixabay

Even if we've achieved random selection and random assignment, there are still other ways for confounding variables to sneak into our experiments, and one such way is opt-in bias or the closely related non compliance bias.
In some experiments, people are going to have a choice about whether they opt in or comply with the treatment.
Say you're running an email experiment where one group gets an email, and the other doesn't.

Images:
https://pixabay.com/en/envelope-open-mail-postage-opened-35392/
http://simpleicon.com/wp-content/uploads/mail-5.svg

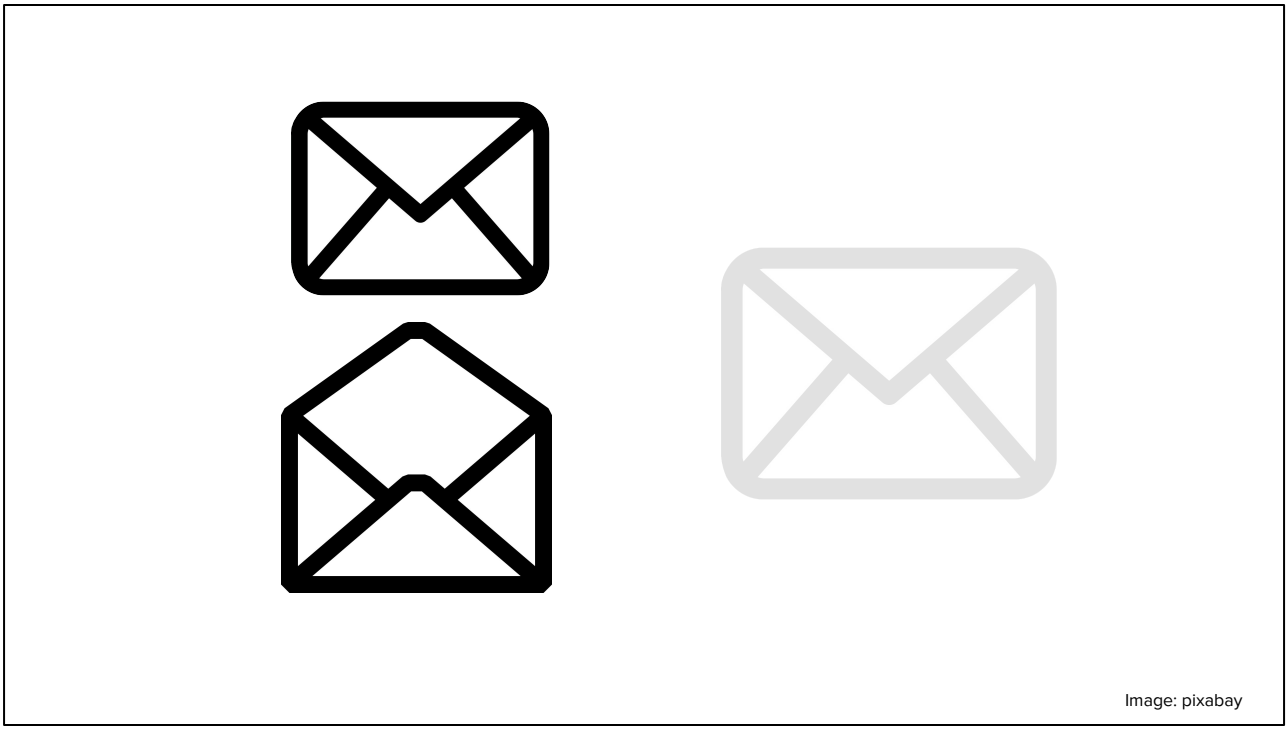


Image: pixabay

Some people in the email group will open your email, others won't. Who should we include in the analysis?

Images:
https://pixabay.com/en/envelope-open-mail-postage-opened-35392/
http://simpleicon.com/wp-content/uploads/mail-5.svg

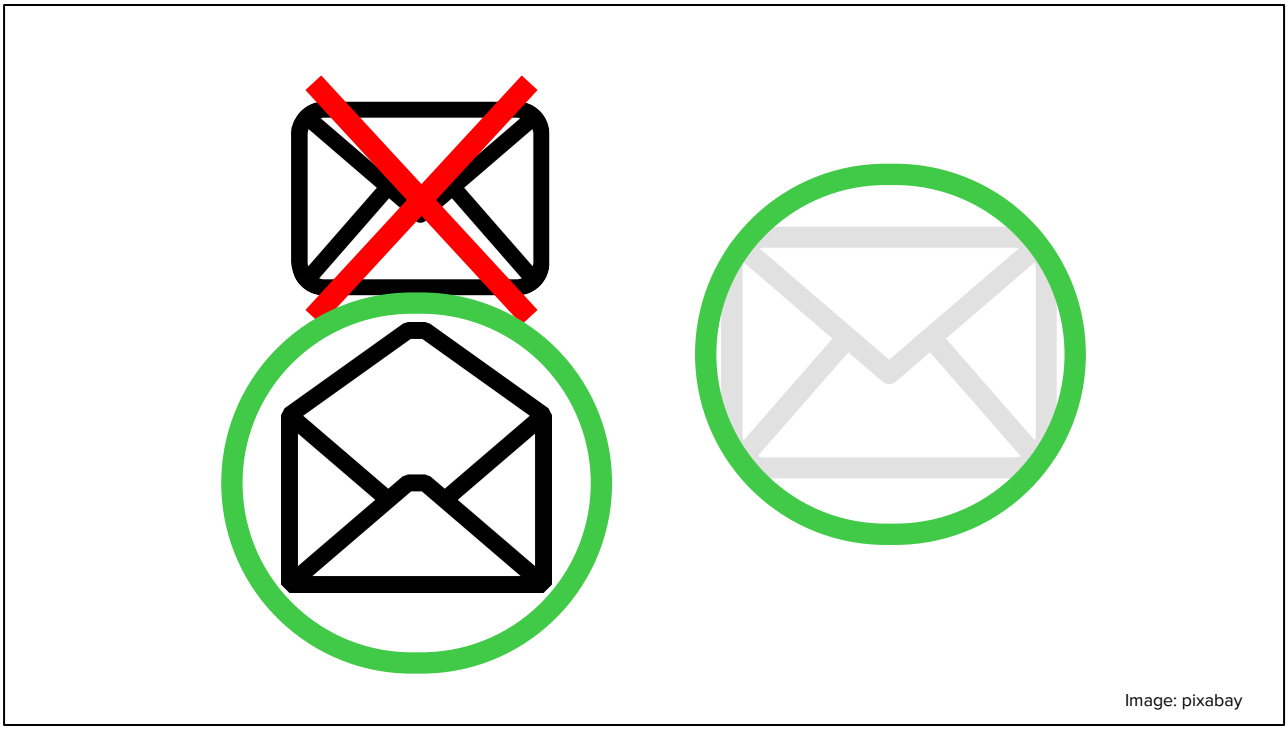


Image: pixabay

One way I've seen this analyzed is to compare only those who open the email in the treatment group against the entire control group, and this seems intuitive, after all, the ones who opened the email are the ones who received the actual treatment.

However, by excluding the people who didn't open the email, we're introducing a systematic difference. The people who opened the email may be more likely to check and open emails, or more engaged with your brand or topic already… so again, this wouldn't be a fair comparison.

Images:
https://pixabay.com/en/envelope-open-mail-postage-opened-35392/
http://simpleicon.com/wp-content/uploads/mail-5.svg

Sample size

Let’s move on to another topic - sample size. How many people should we include in an experiment?
In this case we face a goldilocks-type situation.

Images: http://1.bp.blogspot.com/-641szAIyK34/UzZFpEMzfJI/AAAAAAAABV4/p7qc9iUpYDs/s1600/bowls-with-oatmeal.png

Sample size **Not too small…**

Power

Reliability

Image: pexels / pixabay

Usually the problem is that our sample size is too small… we need a sufficiently large sample to have:

- enough statistical power, the power to detect an effect if it actually exists. If we don't observe a significant result, we want to be confident that it's because there really was no result, and not just that we didn't have enough power to detect it
- Reliability, i.e. if we do see a significant effect, we want to be confident we will see the effect again if we apply the treatment again. This is important if we're going to use our experiment results to decide whether or not to roll the treatment out to the whole population,

Images:
https://pixabay.com/p-2886223/?no_redirect
https://static.pexels.com/photos/116078/pexels-photo-116078.jpeg

Sample size Not too large...

V2?

One simple solution to avoid having too small a sample size is to make it extremely large. But this comes with its own set of problems:

- First it can be very expensive because usually treatments aren't free, and with large SS, costs can become very high.
- Furthermore, if the treatment is unexpectedly harmful, we have exposed more people to it.
- Thirdly, sometimes when we run experiments, it's important that subjects we include have not been included in a previous version of the experiment, and if we include everyone in the first version, we won't have any naive subjects left to include in the next iteration.
- Tiny effects are significant: this seems like a strange one, after all we want significant results right? However, sometimes the effects we detect might be so tiny that they're not very meaningful. e.g. If a doctor discovers that a treatment significantly decreases the chance of flu by 50%, that's great, but what about a significant 5% decrease? 1%? Or 0.1%? The effect is significant, but the impact or magnitude may not be large enough to be useful.

Images:
https://upload.wikimedia.org/wikipedia/commons/thumb/5/53/Skull_and_crossbones.svg/2000px-Skull_and_crossbones.svg.png
https://upload.wikimedia.org/wikipedia/commons/thumb/5/55/Magnifying_glass_icon.svg/1024px-Magnifying_glass_icon.svg.png
https://pixabay.com/p-2022440/?no_redirect

p-hacking / Data-fishing

Image: wikimedia / pixabay

The next one is my favorite
Let’s talk about p-hacking, or data fishing.

Images: https://pixabay.com/p-1617338/?no_redirect
https://upload.wikimedia.org/wikipedia/commons/1/1c/Fish_hook.png

If at first you don't succeed...

- Keep running the experiment until it's significant
- Check occasionally and stop if / when significant

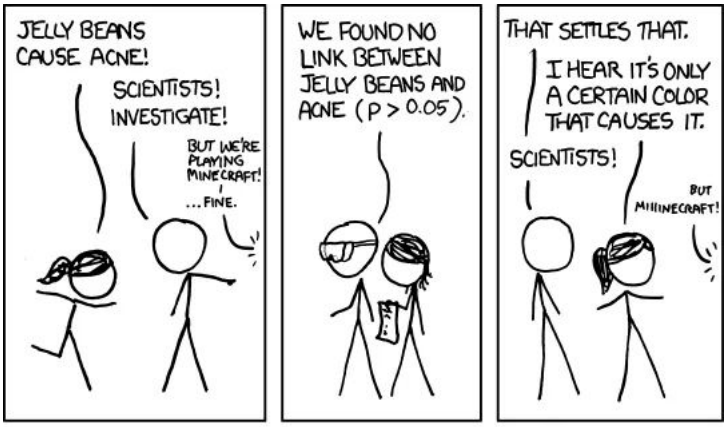
Segment by

- Gender
- City
- Lifecycle
- Week
- Age
- ...

- Test with different metrics
- Exclude outliers
- Add covariates
- ...

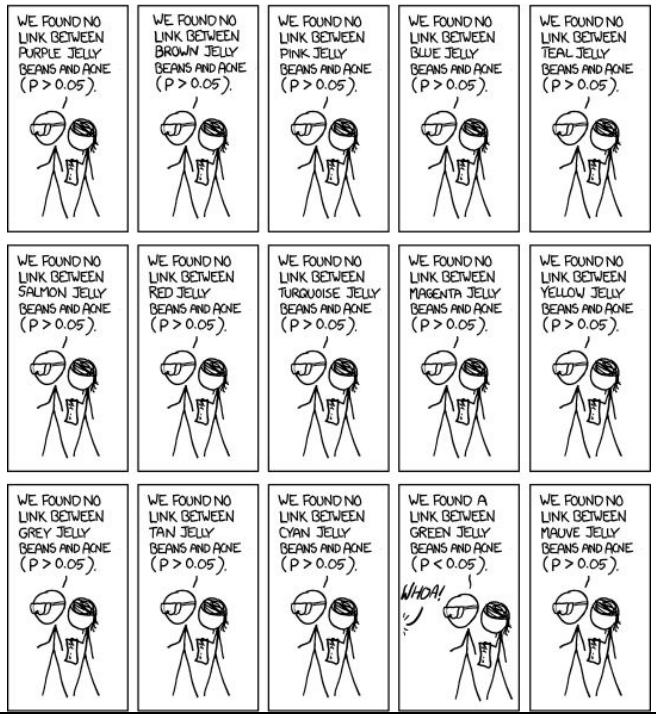
Let me tell you how to get significant results almost all the time, guaranteed. If at first you don't succeed, you can try a number of methods…

In case you didn't catch that, I was joking - don't do this.



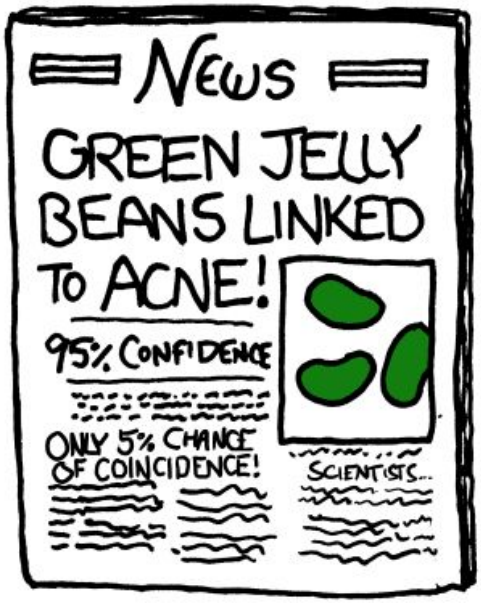
Source: XKCD

Why is this a problem to run many tests till we get a significant result?
There's a famous comic from XKCD that illustrates why this is a problem.
In this comic, someone has the hypothesis that jelly beans cause acne.
Their first experiment finds no link between jelly beans and acne, but someone further hypothesizes that only certain colors cause acne.



Source: XKCD

They run tests on 20 different colors… and find that one of those colors had a significant effect, $p < 0.05$



Source: XKCD

They happily conclude that green jelly beans cause acne.

Every time we run a test, there's a certain chance of getting a significant result just by chance.
False positive: the result shows up as significant even though there is no true effect.
As we run more tests, the chances of getting a false positive adds up.

# $p < 0.05$!!

Image: maxpixel

So, people sometimes do these things and celebrate when they get $p < 0.05$

Images: http://maxpixel.freegreatpicture.com/Brindisi-White-Background-Celebration-Champagne-2711895

# $p < 0.05$!!

Image: maxpixel / wikipedia

But in reality, we're wasting time, money, and effort by validating a strategy that may have no effect, or may even be detrimental.

Images: http://maxpixel.freegreatpicture.com/Brindisi-White-Background-Celebration-Champagne-2711895
https://upload.wikimedia.org/wikipedia/commons/thumb/9/90/Twemoji_1f4b8.svg/1000px-Twemoji_1f4b8.svg.png

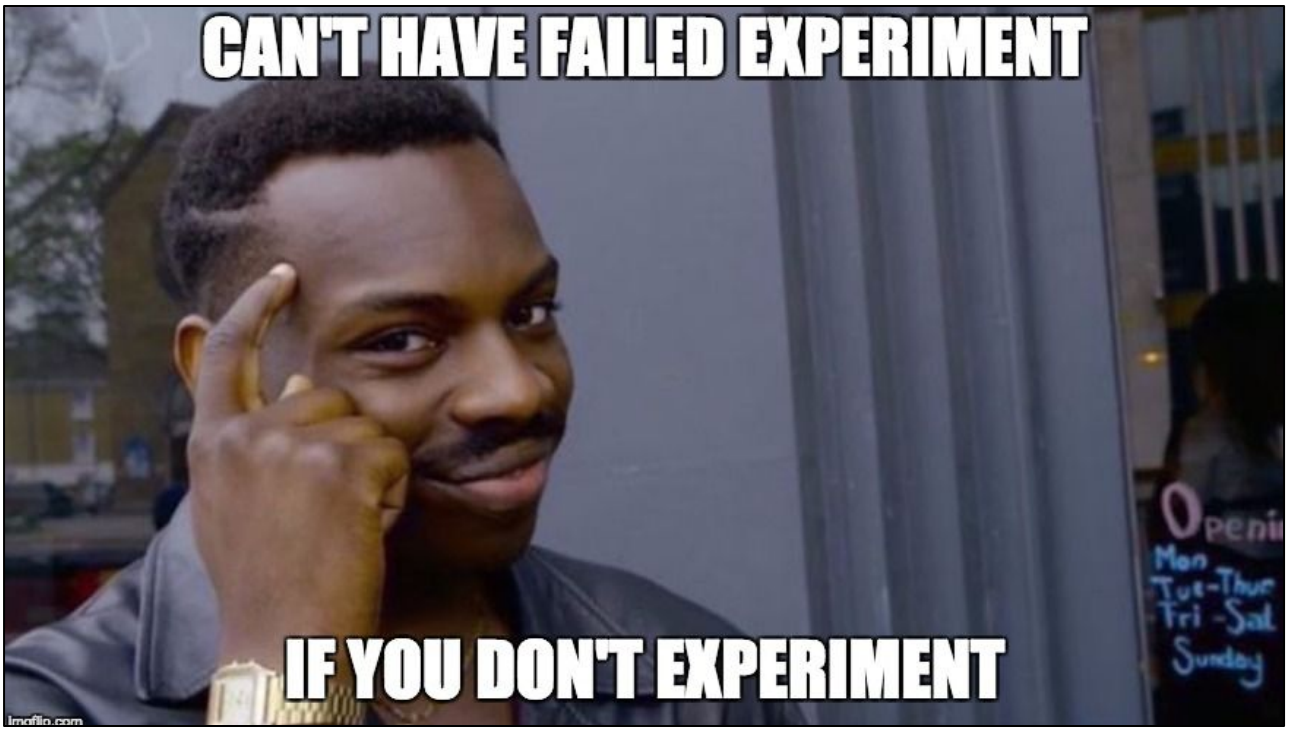
Not testing at all

- Don’t know better
- Faith in intuition
- Requires too much effort
- **Afraid of failure**

¯_(ツ)_/¯

Stats aside, I wanted to mention a whole other type of problem, which is that some people don’t test at all.
There are occasions where people don’t test because they’re afraid that the experiment won’t succeed.

Image: https://www.flickr.com/photos/blakeimeson/2743011812



After all… You can’t have a failed experiment if you don’t run the experiment in the first place

“Fails” can be wins

Best practices to ensure *all* experiments are useful

To address the last point, In this final section I’m going to go over why we should not be afraid of failed experiments. Failed experiments can actually be very useful in avoiding ineffective or even harmful tactics.

But for our fails or non-significant results to be useful, and actually for any results to be useful, we need to set up our experiment according to statistical best practices so we can trust our results and draw accurate conclusions, and more importantly, adopt the right company culture and process that encourages the reporting of all experiment results, even the null or harmful results.

Random sampling

Generalize conclusions only to included groups

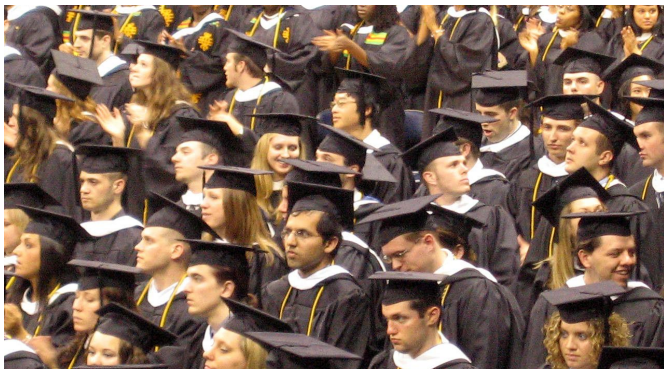


Image: wikimedia

Random sampling:
To recap, we want to sample randomly so that our sample is representative of the population.

Sometimes this isn't possible, or is very hard to do, and we need to balance best practices with business needs and efficiency, and that's fine. In these cases, we need to make sure that we only generalize our conclusions based only on the groups or traits we included, not the entire population.

E.g. In studies with WEIRD people, we need to be careful about not generalizing the findings to other cultures or socio economic groups

Image: https://upload.wikimedia.org/wikipedia/commons/b/bb/Cherry_Stella444.jpg

Random assignment

No systematic differences between treatment and control groups

- ~~Name~~
- ~~Email~~
- ~~Phone number~~

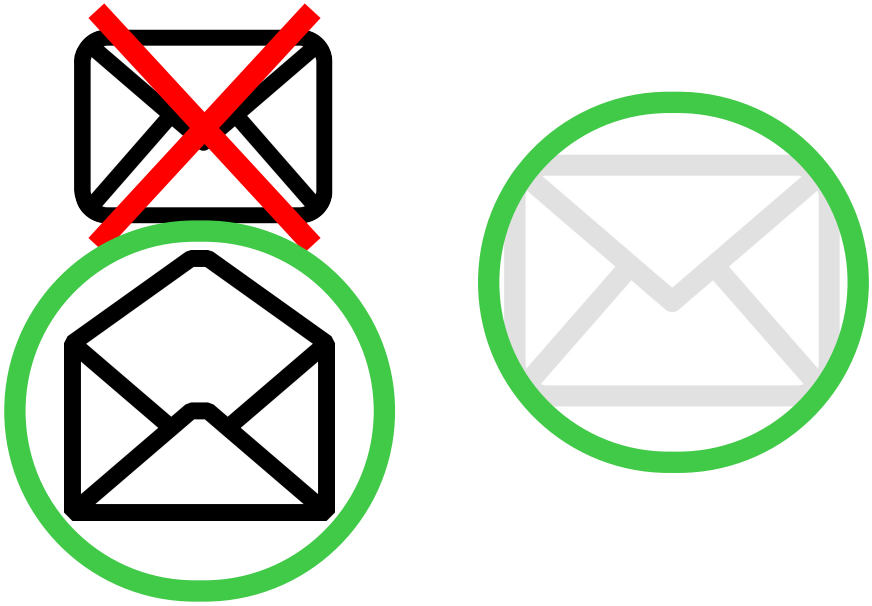
- UUID *universally unique identifiers*
- Random number generators

Secondly, we want random assignment to treatment and control groups such that there are no systematic differences between the two groups that could become a confounding variable.

To achieve this, We need to think about how we're randomly assigning people to each group so that we don't accidentally introduce bias.
E.g. instead of using properties like name, email, phone numbers, which may vary among individuals in a systematic way, we can use UUIDs or random number generators because they aren't tied to any properties of the participant, and as their name implies, they are random.

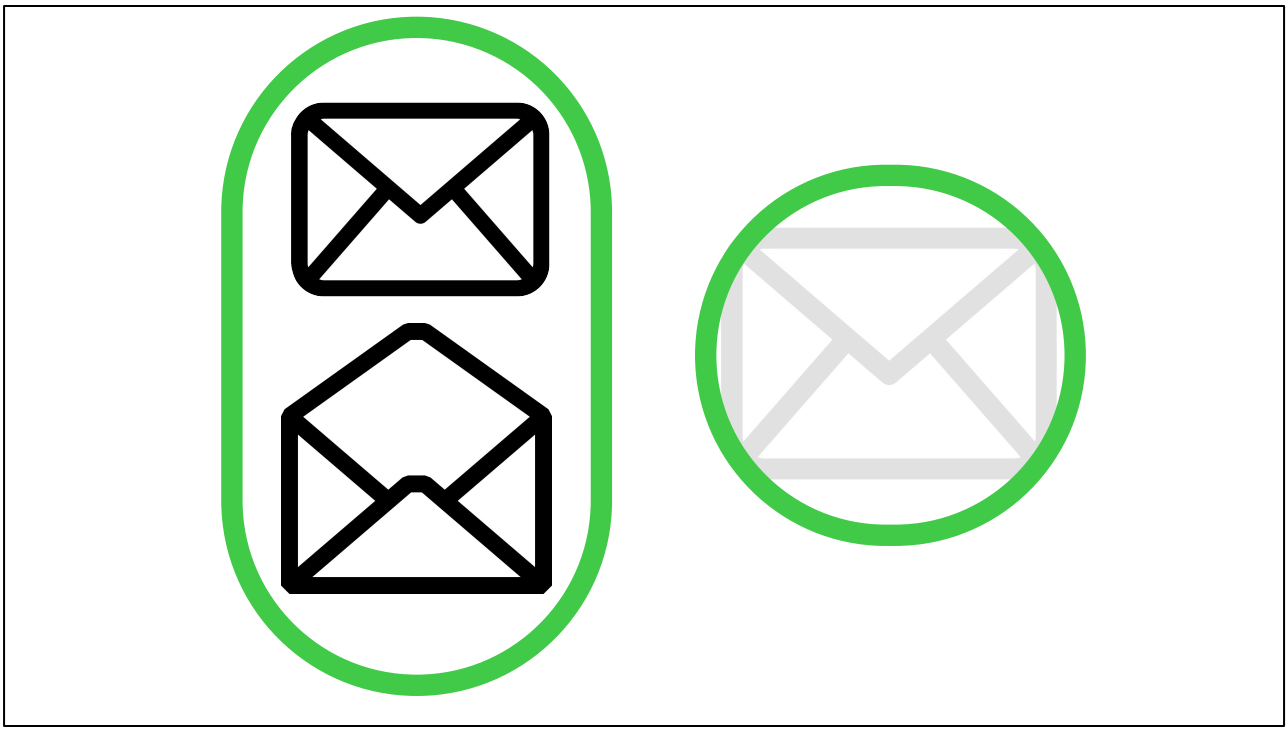
Image: https://pixabay.com/p-316473/?no_redirect

Opt in bias



To address the opt in bias problem we talked about just now:
Remember we were saying that one possibly biased way to run the analysis is to compare only the people who opened the email against the entire control group. However, this is biased because we're excluding people who are less likely to open your emails, for whatever reason.

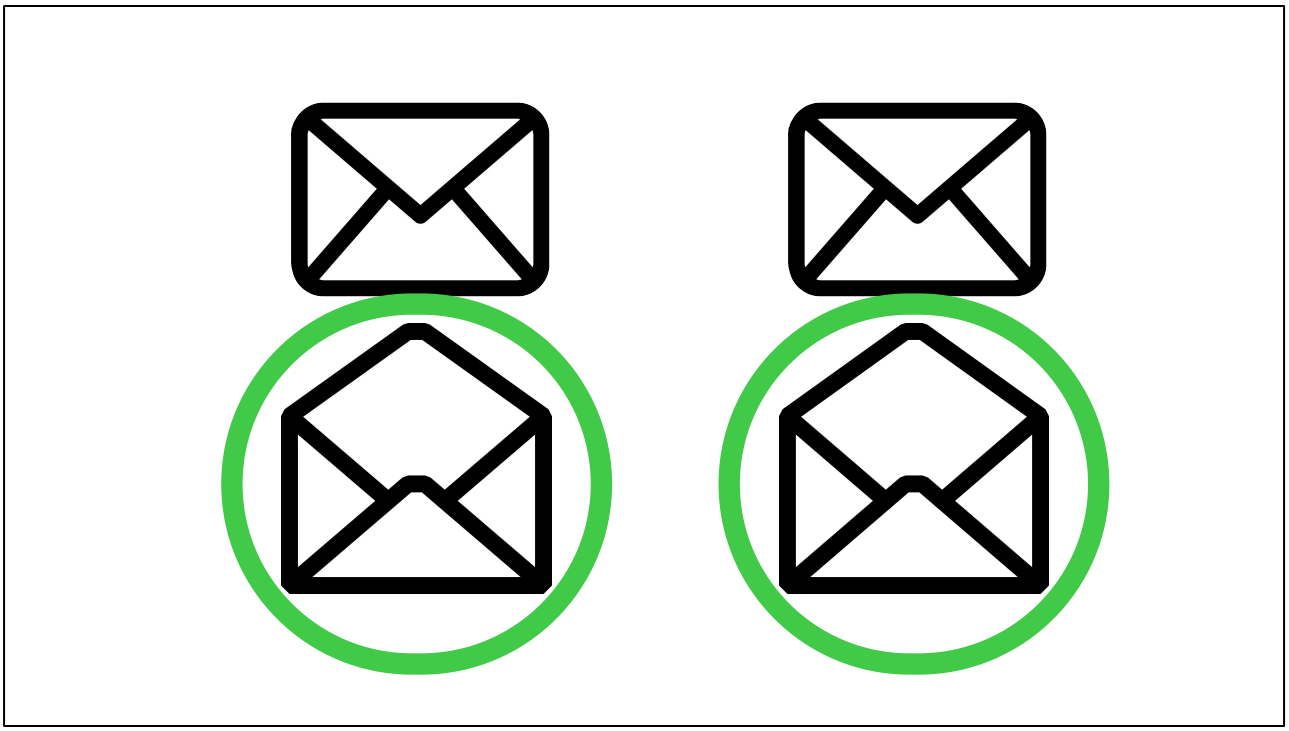
Images: https://pixabay.com/en/envelope-open-mail-postage-opened-35392/
http://simpleicon.com/wp-content/uploads/mail-5.svg



To achieve a more fair test, an alternative is to include everyone in the analysis, even the ones who did not open the email.

This is more conservative test, and it also takes into account the real world scenario where the effect of your email campaign may be weakened because some people will not open your email.

Images: https://pixabay.com/en/envelope-open-mail-postage-opened-35392/
http://simpleicon.com/wp-content/uploads/mail-5.svg



A second alternative is to send a control email to the control group too, and only compare email openers in the treatment group against email openers in the control group.
This controls for different levels of engagement and other factors that might cause some people to open your email, and not others.

An analogy here is giving the treatment group a drug and control a placebo, some will comply, others will not. We then only compare the compliers directly to each other.

Images: https://pixabay.com/en/envelope-open-mail-postage-opened-35392/
http://simpleicon.com/wp-content/uploads/mail-5.svg

Correcting for multiple comparisons

Bonferroni correction

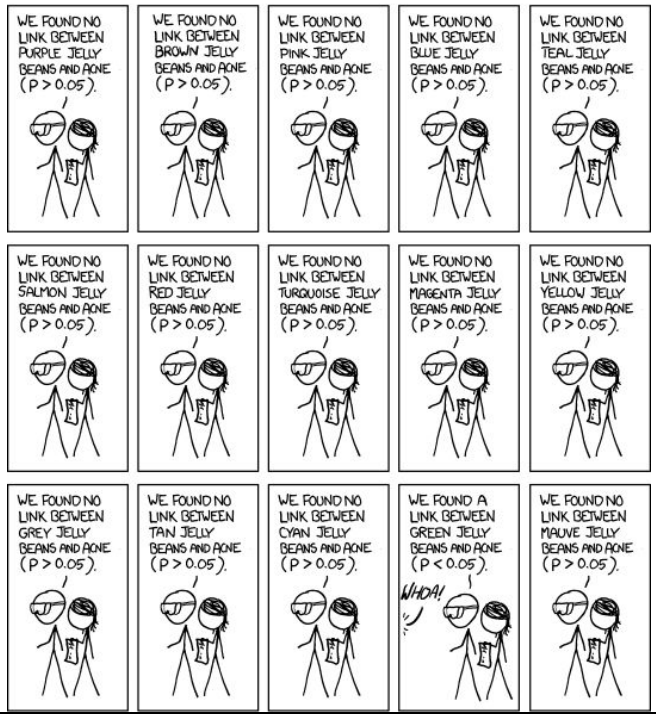
$$p < \frac{0.05}{[\text{number of hypotheses}]}$$

To keep the false positive rate under control, we need to do something called correcting for multiple comparisons.

One of the simplest methods is bonferroni correction, where you just divide 0.05 by the number of tests you're running, and use the resulting number as your new p-value threshold.

FDR (false discovery rate) - rate of discoveries that are false
family-wise error rate (FWER) is the probability of making one or more false discoveries

https://pxhere.com/en/photo/622134



Source: XKCD

So for the xkcd comic example where we ran 20 tests...

Correcting for multiple comparisons

Bonferroni correction

20 tests

$$p < \frac{0.05}{20} = 0.0025$$

We divide the original p value threshold of 0.05 by 20, which is 0.0025.

So if the p-value for the link between green jelly beans and acne was 0.01, earlier, we would have concluded that this result was significant, but now with bonferroni correction, we do not reject the null hypothesis.

Image: https://pxhere.com/en/photo/622134

Sample size

Calculate sample size requirements...

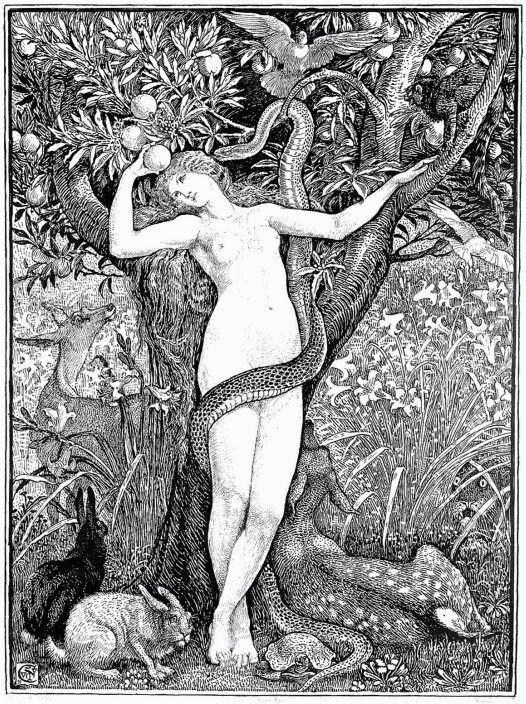
... then **don't peek** until sample size is reached

With regards to sample size, how do we know how many people to include in the experiment so that we have sufficient power and reliability, without throwing everyone in our population into the experiment?

To find the optimal number of people to include, there are formulae we can use to calculate the appropriate sample size, given the expected effect size and desired level of power.

The details are outside the scope of this talk, but once we know this number, we should ideally not peek at the results until the sample size is reached. In other words, we should only run the analysis when we've collected the minimum amount of data. This is so we're not tempted to end the experiment early if we are lucky enough to get a significant result by chance midway.

Image: https://pixabay.com/p-2089537/?no_redirect



Lead us not into temptation….

But we’re human, and in case we’re tempted to cheat, there are several measures we can adopt to control for that.

List hypotheses beforehand

"Is there a reason to believe the effect exists?"

Image: pixabay

The first is to list our hypotheses beforehand:

- This discourages p-hacking at the analysis stage
- The litmus test for including a hypothesis in the list is to ask ourselves, is there a reason to believe the effect exists? E.g. if we're segmenting by gender, there should be an underlying reason to believe there might be differences between men and women.

Image: https://pixabay.com/p-1464917/?no_redirect

Knowledge sharing

Share results regardless of outcome

Image: pixabay

Furthermore, sharing the results of each test, regardless of the outcome ensures that everyone in your organization can learn both from your successes and failures, and also prevents duplication of effort.

Unbiased analysts

- Neutral third party
- Peer review

Roping in unbiased analysts who have no vested interests in the outcomes of your analysis is another way to control cheating.
These people can act as a neutral 3rd party who can run your analysis in an unbiased way.
We can also set up a peer review system to have our peers check our analyses and keep us honest.

Image: https://www.flickr.com/photos/61056899@N06/5751301741

Culture / Attitude

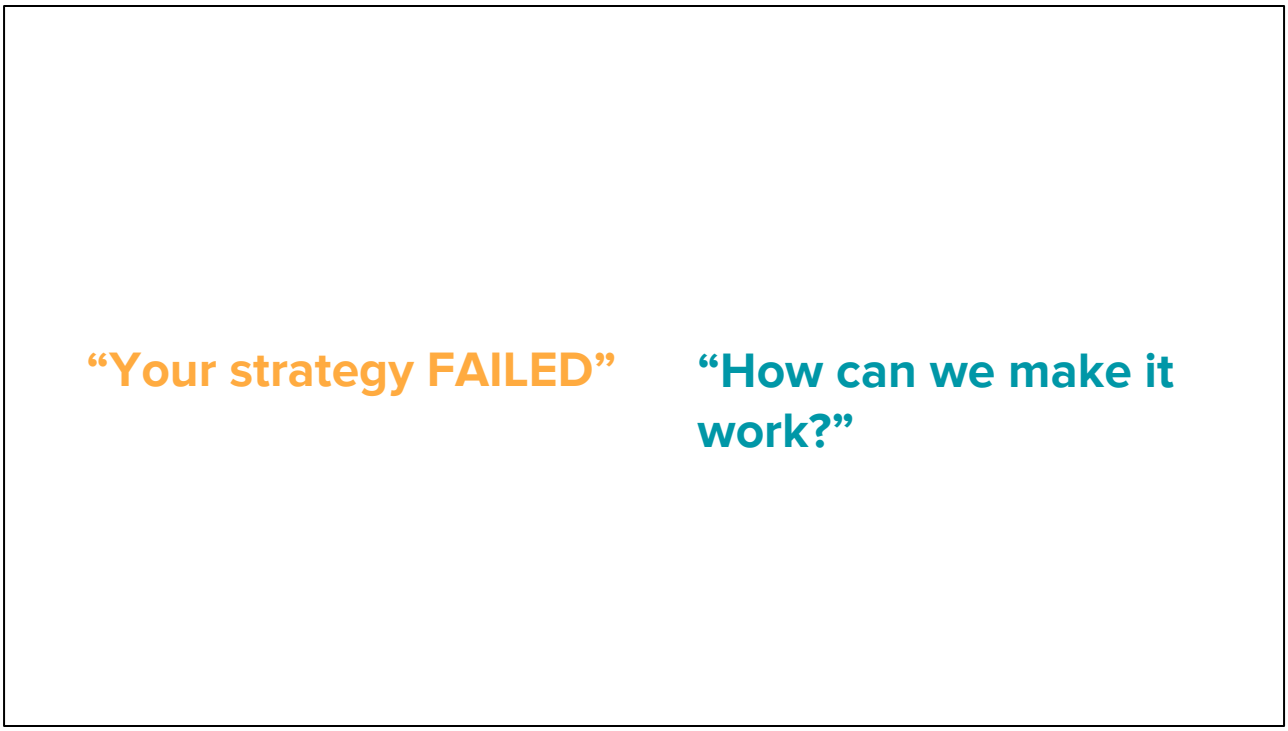
Seek the **Truth**

(NOT “Seek Significance”)

Lastly and most importantly, we need to create a culture that supports seeking the truth, NOT one that rewards significant results.

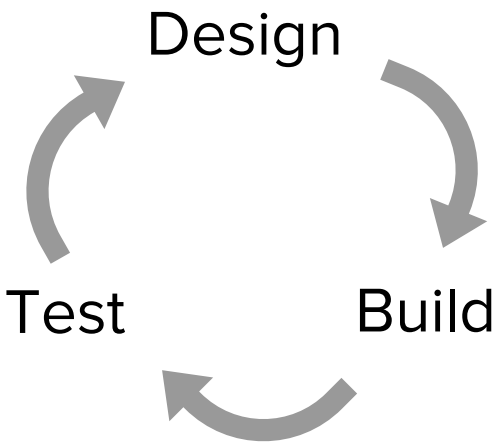
To do this, we need to encourage and reward the reporting of null or harmful effects. One way to do this is to reframe how we think and talk about failed experiments. E.g. Instead of saying our strategy failed, we can reframe this and recognize that we dodged a bullet.

What negative effects did we prevent by testing first?
How many hours of effort and dollars of spend did you save by not rolling out an ineffective strategy?



Secondly, instead of burying our failure, we can use the results of our experiment to find out how we can improve the next version of our strategy.

Iteration



By adopting this approach of testing our strategies after we design and build them, and then using the results of the test to improve the next version, regardless of whether they validated our expectations or not, we can iteratively improve our strategy over time instead of relying on gut feel.

失败是成功之母

Failure is the Mother of Success

If we adopt statistical best practices, and know that we can trust our results and conclusions, and on top of that also adopt a culture that celebrates failures, we can not only ensure our “so called” failures are regarded as wins in and of themselves, but can also lead us to future, bigger wins.

gracetang.me
www.linkedin.com/in/tsmgrace
grace.tang@uber.com

Grace Tang, Data Scientist
UBER