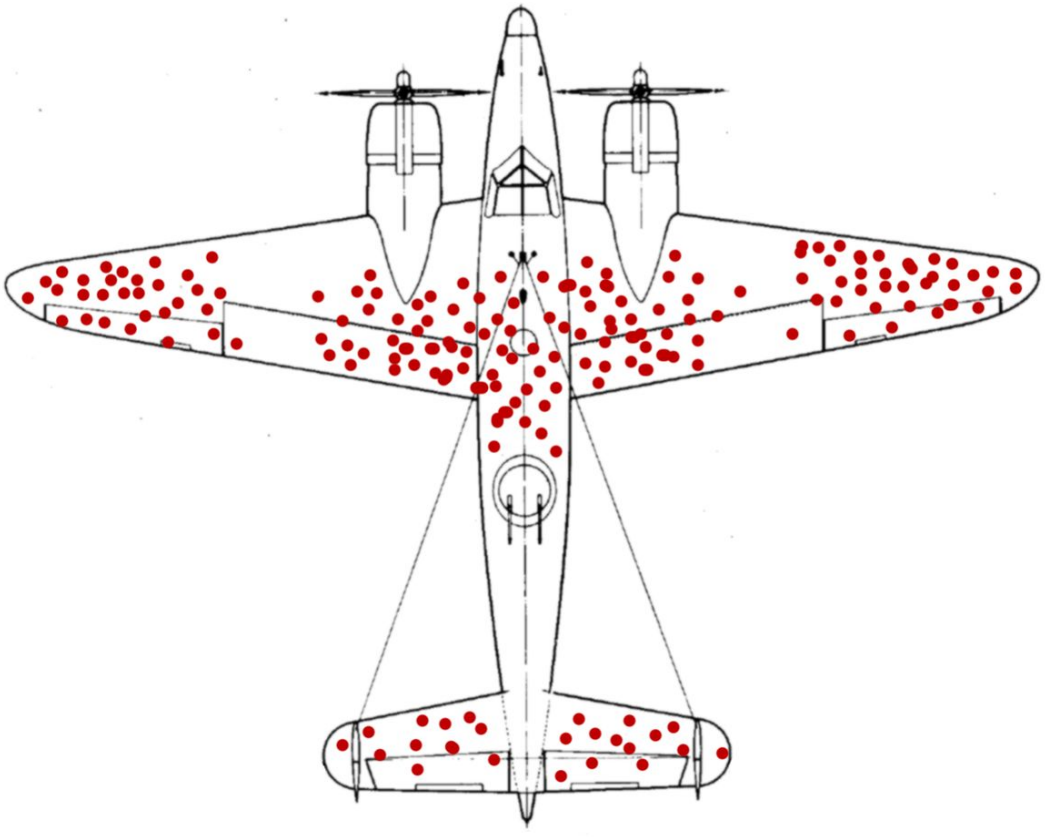
Survivorship bias & Simpson's paradox

Recently I worked on a project which is a company that wants to build a new movie studio and my job is to figure out what type of movie is the best. After analyzing multiple movie data sources I found out some type is the best choice but a new question showed up in my mind. Is this really the right answer?

I only work on movies which are public. Is it possible the type I found out is the best actually is the worst? Is it possible that the type of movie is actually the most common type that couldn't be completed? This is an interesting thought, so I did some research and I found a phenomenon called "Survivorship bias".

Survivorship bias

One of the most famous examples goes something like this (some artistic license has been applied): In WWII the US Military wanted to increase the likelihood their bombers would safely return from missions, knowing they couldn't just add more armor everywhere because it would be too heavy. They studied the problem internally and also asked the Statistical Research Group for their opinion. One chart they thought would be helpful was an aggregate shot chart: they mapped out where each returning aircraft that had been damaged by enemy fire was shot (see below).



The US Military took one look at this and decided they needed more armor in the middle of the plane and the tips of the wings, with maybe a bit on the tail if they could afford any extra armor. The Statistical Research Group was just about to sign off on this analysis when Abraham Wald spoke up: “You guys are idiots. You should add armor *literally* anywhere except the middle of the plane, the tips of the wings, or the tail.” The US Military laughed and asked why they’d add armor to places that had never even been shot instead of the places that get shot most often. Abraham pointed out those places hadn’t been shot *on planes that were able to return successfully*.

All of the red circles on the aggregate shot chart represent places the plane could be shot and still return to base safely. If you wanted to add armor to anything, it should be to a part of the plane that would prevent it from returning if it were damaged (engines, cockpit, etc.). The outcome of the analysis was 100% flipped once Abraham took into account “Survivorship bias”, which is the bias that can arise in data when you look at only a subset that has survived some process.

In my recent project, we only analyzed data from films that made it all the way through production and actually were released. Some movie projects stall out during filming or production and don’t make it to release after investing significant money. Many others go over budget even though they do end up with a successful release. Without any insight into these projects, it is very difficult to make recommendations on what types of movie projects to fund. One of our main findings was that low budget films have very good return on investment, but that is based on the final budget, not the amount the studio initially set aside for a film. Lacking data about projects that fail mid-production, and projects whose budget changes significantly during production makes it very difficult to say what the expected ROI of a new film project would be based on your initial budget, even if we know very well what the correlation between ROI and final budget is for released films.

When I read the page about “Survivorship bias” it looks so interesting and useful for data science so I clicked the other key word they mentioned: “Simpson's paradox”.

Simpson's paradox

In 1973, 44% of male applicants to UC Berkeley were admitted, while only 35% of female applicants were admitted. If I were alive then, I would have definitely demanded an explanation from the university. I imagine it would have gone something like this:

Me: I computed the odds of such a large difference in admissions rates occurring by random chance, and it’s nearly impossible! You have a sexist admissions process which admitted 277 more men and 277 fewer women than it should have. You need to change it!

UC Berkeley: I can assure you, we take any allegations of sexism very seriously. However, each department handles admissions separately. Could you help us identify the department responsible?

Me: Sure. I’ll look into it and get back to you!

<hours later>

Me: So, uhhh… Most of your departments are totally unbiased. Four departments appear to be have a statistical admissions bias against women, which resulted in 26 fewer women and 26 more men than expected.

UC Berkeley: I thought you said 277 women? Either way, give us the department names and we'll write them a VERY sternly worded email, so that we're sexism free going forward!

Me: That might not solve all the problems… There are actually 6 departments biased against men, who admitted 64 fewer men and 64 more women than they should have.

UC Berkeley: So our university is biased against women, but our departments are biased against men? How can that be?

Me: Well, let's use an example. The department that most men applied to, Department A, had 825 male applicants and 108 female applicants. It admitted 62% of men (511) and 82% of women (89). Another department, Department D, had 417 Male applicants and 375 Female applicants. It admitted 33% of men (138) and 35% of women (131). Both departments are admit more men than women right?

UC Berkeley: Yep.

Me: But if we add it together, 52% of male applicants across both departments were admitted (649 out of 1242). Meanwhile, only 46% of female applicants were admitted (220 out of 483).

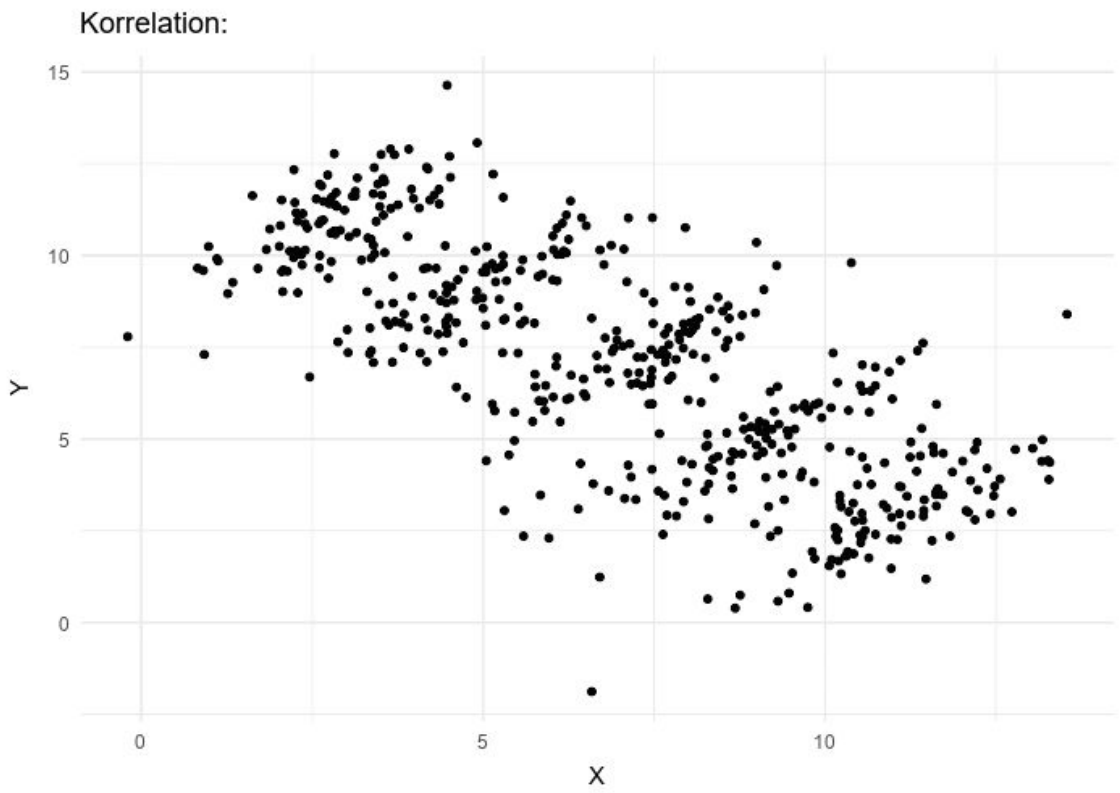
UC Berkeley: That doesn't make any sense.

Me: Basically, many more women are applying to hard departments with low admissions rates. Even if they do slightly better than men at getting into those departments, their overall admissions rate is low. Men are applying to departments with high admissions rates. Even if they do slightly worse than women getting into those departments, their overall admissions rate is very high!

UC Berkeley: That's still confusing.

Me: Check out this gif from 2017 (credit Pace~svwiki, CC BY-SA 4.0), it shows how trends can reverse when you mix across subgroups!

These stories were fun to read and think about, but they would be horrible if it happened during my analysis. Each of these biases can lead to completely flipped results, are easy to make even for intelligent people, and difficult to notice. Reading about, and learning from past mistakes in statistical analysis or Data Science is not only a good idea to improve the quality of my analysis, but it's also an interesting and fun exercise.



UC Berkeley: That is one super helpful gif. Thanks! Wait did you say 2017?