**Is Bigger Better? A Look at a Selection Bias that Is All Around Us**

**Arnold Barnett**
**Anna Teytelman**

***Opening dialog:***

AN: Hey Arnie!

AR: Oh, hi Anna. What's in the big bag?

AN: Plenty of trouble for me!

AR: Oh?

AN: My school had two performances this weekend and now I have to count all the tickets to see how many people actually went to see the play.

AR: No you don't! Here's an easy way of getting the answer. Pick someone at random who attended one of the shows. Ask him how many were in his show. Whatever answer he gives you, just double it and that's the answer. You don't have to count one by one.

AN: Oh, I don't know Arnie. I guess your method would work sometimes, but first of all, how would that person even know how many people want to see his show?

AR: Anna, this is MIT. Don't you assume that that person would know exactly how many seats there are in the theater and exactly what percentage were filled? So let's assume he'll give the right answer and then you just double it and you're done.

AN: I don't know. Even if that person knew exactly how many people were in his theater, I still don't think you can just take his answer and double it to get the exact number of people who went to both shows. You know there's a saying: "if something seems too good to be true, it probably is?" Well your rule is way too easy. I think it's too good to be true.

AR: Nonsense! If my rule is not correct, I'll suck a lemon dry. There's no need to make complicated that which is easy. There's no reason you have to count the tickets one by one.

So Arnie thinks that what we could do to find out the total number who went to see both shows is to ask one person at random from either of the two shows and get his number of the people in his theater, double it, and we'll get the total number of people who went to see both of the shows. Do you think this method works? Are there some cases in which Arnie's method actually works every time? And are there other cases in which Arnie's method is bound to fail?

We'll be back in just a few minutes after you've thought about it and Arnie is going to have a lemon with him just in case!

***Dialog:***

AN: Arnie, I was thinking about this and I still think your method won't work.

AR: Ohhhh?

AN: Well, I was thinking about it in terms of the show last year. Last year we also had two performances and one of them had 100 people and the other had 200 people in attendance.

AR: I see so there were 300 in total.

AN: 300 in total.

And I was thinking of what would happen if we were to try your method on this case. There are two things that could happen. The first thing that could happen is that we ask a random person from the first show. The first show had 100 people in it. So we ask him, "How many people were in your show?" And he would tell us 100. We'd double that and we'd get a total number of 200.

AR: Yes, that's true.

AN: The second thing that could happen is that we'd ask somebody from the second show and we'd ask him how many people were in his show and he'd say that 200 people attended. Then we'd double that and we'd get a 400 people total for all of the people.

AR: OK, yeah.

AN: Well, then the only guesses that we would have is 200 and 400, but we'll never actually have the right answer of 300 people.

AR: You know, I'm afraid you're right. I think my method would only work if both shows had the same number of people.

Do you have any salt and pepper? I was prepared for this possibility.

AN: Don't feel bad Arnie.

AR:Wait! I've got it. Instead of taking one person and doubling his answer, take six people at random from the two shows, get their individual answers, double them, average the six numbers together and you'll get 300 which is the correct answer. That method is bound to work. It's much easier than counting all the tickets and you know I am the greatest genius of all time!

AN: Well, that would be easier than actually counting all of the tickets. I still don't think it will work. I can't quite put my finger on it but I don't think that it will work every single time.

AR: Anna, the reason you can't put your finger on the problem is that there is no problem. The method is bound to work and I am the greatest genius of all time!

What do you think? Is Arnie the greatest genius of all time as he claims? Or is there something still wrong with his method?

***Dialog:***

AN: Arnie, I thought about this some more, and I'm afraid that your method still doesn't work.

AR: Ohhhh?

AN: Well, I thought about it again, about the case last year where you had one show with 100 people and one show with 200 people, with a total of 300 people right?

AR: Yes, 300 in total.

AN: Can I use the board?

AR: Sure.

AN: OK. So suppose that we have show one with 200 people and show two with 100 people. Now let's pick out six people just as you would suggest. Now we would expect that since show one has twice as many people as show two, then we would get twice as many people that we pick out at random from show one than from show two. So if we pick out six people, we expect four of them to be from show one and two of them to be from show two.

AR: Yes.

AN: So let's think now about how the process would actually play out. The four people from the big group would all say that 200 people attended the show. By your method we would double that and we would have four answers of 400.

AR: Yes, but we would also get other people giving us the answer 200.

AN: Indeed. So now we have two of these people who will tell us that there were 100 people in the show and doubling that, we have two answers of 200. And we want to average this.

AR: Yes, so you'll be averaging 200s and 400s and you'll get 300 just as I said.

AN: But we wouldn't expect the average to be 300. So how would we average this? What we would do is we would sum up all the answers and divide them by six. So the average would be
$(4 \times 400 + 2 \times 200) \div 6$ ,right?

So $4 \times 400 = 1600.$
$2 \times 200 = 400.$
So our answer is $2000 \div 6$ and that is 333.33, which is higher than 300 and it's not the right answer. Like I said, too high.

AR: But if you had three 400s and three 200s, then you would get the correct answer of 300. What's the problem?

AN: The problem is that we wouldn't expect a 3-3 split among those people that we pick at random. What we expect to see is a 4-2 split because the number of people in the first show is twice as big as the number of people in the second show. I guess what I'm saying is that if we pick somebody at random, then the number of people we will pick will be higher from larger groups.

Think of it this way. Suppose I had a big urn of green and red balls and I had 200 green balls and 100 red balls. If I were to pick out 6 balls on random, I would expect to see more green balls than red balls. In fact, I would expect to see 4 green balls and 2 red balls. Does that make sense?

AR: It does, but I hate it. What you're saying is that if we sample at random, we're not going to get equal numbers from the two groups, we're going to get more from the bigger group. And for that reason our answer is going to be too high. I guess I danced too soon. Ewwww.

AN: Don't feel bad Arnie. We can't all be good at math.

How about we go and catch a bus to Harvard Square and get some coffee because I still have to count all of those tickets and I have to be alert.

AR: OK. But let me ask you a question, though. What does it matter? Suppose we say the estimate is 333 when it should be 300. Why do we care?

AN: I would assume that the theater owner cares because he wants to know how much profit he made. And the profit he makes is — the amount he takes in is the number of tickets times the price of a ticket. And he would want to know if he covered his costs. This is America, don't you think the businessmen would really care about his profit?

AR: You're certainly right about that. Anna, what you're saying now reminds me of a situation related to the buses. Why don't I tell you about it before we walk to the bus stop.

AR: Anna, let me tell you my example about buses. Suppose I told you that 12 buses an hour reach a certain stop where people get on. How long do you think those people wait on average for the next bus?

AN: Well, if there are 12 buses an hour, then I would expect to see one bus every five minutes right?

AR: 60 over 12 is certainly 5 minutes, as you say.

AN: So I would expect that some people would just miss the bus, so they would have to wait an entire 5 minutes until the next bus shows up. And then some other people will arrive just in time and they would have zero wait time. Then the rest of the people would come somewhere in the middle of that period and they'd have some wait time between 0-5 minutes. So on average I think that somebody waits 2.5 minutes, 5 divided by 2. And as they say on *Who Wants to be a Millionaire*, "that's my final answer."

AR: Not a bad answer at all but not necessarily correct.

AN: What do you mean? You're the one that gives the wrong answers.

AR: Well, let me give an example. Let's suppose this is the schedule of the buses between 2 p.m. and 3 p.m. There's a bus at 2:00. The next bus is 2 minutes later at 2:02. Then it's 8 minutes until the next bus at 2:10. Two more minutes until 2:12. 2:20, 2:22. Would you agree with me that there are a total of 12 buses that have arrived in that hour?

AN: Yes, I do.

AR: OK. Let me make one more assumption: that one person per minute arrives at the bus stop and that everyone gets on the next bus that comes. OK? So we have these assumptions.

Now let's talk about the 2:02 bus, which is only 2 minutes after the previous bus at 2:00. Well, it's been 2 minutes since the previous bus, so in those 2 minutes 2 people arrived and they waited somewhere between 0-2 minutes, so they wait on average of 1 minute apiece. OK?

AN: Yes.

AR: Now let's consider the next bus at 2:10. That's 8 minutes since the previous one so 8 people have arrived. And their waits will vary from 0-8 and average out to 4. Are you with me?

AN: Yes.

AR: OK. So we have 10 people, in other words, who arrive between 2:00 and 2:10, some of them in the short interval, some of them in the much longer interval.

AN: So far so good.

AR: Now I'd like to work out with you the average amount of time those 10 people waited. Let me show you how I'll do it. First of all, I'll say let's think about the total number of minutes that they waited. There were two people who arrived in the first 2 minutes, between 2:00 and 2:02 and on average they waited 1 minute apiece. Right?

AN: Right.

AR: Now there were 8 people who arrived in the long interval between 2:02 and 2:10 and they waited an average of 4 minutes apiece. So the total amount they waited if we work this out is 34 minutes. OK?

AN: OK, but where are you going with this?

AR: Let's work out the average per person. If there are 10 people who wait 34 minutes in total, their average wait is 3.4 minutes apiece. But you said the average wait was 2.5 minutes.

AN: But when I said 2.5 what I meant was the average waiting time for passengers throughout the whole hour. What you just calculated was the average waiting time for people who arrived between the hours of 2:00 and 2:10.

AR: Ah! Anna, that's true but it doesn't make a difference because it's 3.4 minutes for the people who show up between 2:00 and 2:10. But wouldn't the very same argument apply between 2:10 and 2:20 and between 2:20 and 2:30? So the answer would be 3.4 minutes for the entire hour.

AN: OK. I agree with your calculations. But if the buses come on average every 5 minutes, why is your answer for the average passenger waiting time longer than mine?

AR: You're perfectly right to say they do run on average every 5 minutes. But let's think how we get that average of 5. Half of the intervals are 2 minutes long and the other half are 8 minutes long. So that's how we get the 5. Half of them are 2, half of them are 8. The average is 5. But here's what's happening to the passengers. The 8 minute intervals absorb 8 out of every 10 passengers. Most of the passengers arrive in the fat, ugly, long interval and have long waits. And that drags up the overall average. That's the problem we have here. We have two groups, those who are lucky and arrive in the short interval, and those who are unlucky who arrive in the long one. But many, many more people are unlucky and arrive in the ugly interval than arrive in the short, beautiful one.

AN: OK. I think I see now what you mean, but this is not a happy story. I want to talk to you more about this so let's go catch the bus to Harvard Square now.

AR: Well, we can certainly do so and I'll tell you why your example about the theater tickets reminded me of this, and then we'll be back for the grand finale!

***Grand Finale:***

AN: So Arnie, all of this stuff still kind of seems like magic to me. But I think that I'm getting the important part of what you're saying.

AR: Everything I say is important!

AN: Of course it is, but I think what you're trying to say in this case is that if we take people at random and ask them about their wait times, the majority of the people will come from the long interval, and they will tell us that they waited for a really long time. So if we average everybody together, then that will average passenger wait time will be really long.

AR: I think that's right. I also think that the situation comes up in your theater example, because if we pick people at random from the two shows, most of the people we see will be from the crowded show. And that's why if we average their responses together we're going to think there were more people in the two shows combined than there actually were.

AN: I see. So then our take-away from this is that it's a bad thing to pick somebody at random and ask them about their experience if we really want the experience of the entire group.

AR: Well, I don't think that's necessarily true because in the case of waiting for the buses, most people do have long waits in the long intervals. And picking people at random will accurately reflect that.

In the case of the theater, though, when we want to know how many were in each of the shows, estimates that are biased toward the busier show might be inaccurate. So it's not that it's necessarily wrong.

The real moral of the story is that big groups have loud voices. And there are formulas we have to take that into account. In some cases it's good to focus on the bigger group, in other cases it's not so good. But at a minimum we should recognize that random sampling goes toward the bigger group.

AN: I see. OK, Arnie. Is that our bus? I think we should catch it. Come on!

(Bus sounds)

AN: This is just my luck Arnie.

AR: Don't feel bad Anna. Do you know that there are some times when it's better just to miss a bus than to arrive at a random time?

AN: What are you talking about? That's impossible.

AR: No, it's not impossible. But it is one of the challenge problems that the website for this video which the students can discuss with their teachers. In the meantime, thanks so much for coming to visit us at MIT today.
Thank you!

***For Teachers***

Teachers, we want to thank you enormously for participating in this BLOSSOMS video with your classes, and we know you have a major role to talk to the classes between the individual segments to get them ready for the next segment and to reinforce what was said at the previous one. We did prepare a guide of suggestions for what you might do between the sections, although you may have your own ideas that might be better than ours. But let's quickly review what we suggest for what might happen in the five intervals that you take care of. Anna?

Between the first and second segments we hope that you can help the class test Arnie's first bad idea. We recommend that you can split the class into two equal halves and then you can select a student at random and ask him or her how big his half of the class is. Then you can double it and with the doubling method you will always get the correct answer for the total number of people in the class. Then however, if you split the class into two halves of not equal size, for example one is significantly larger than the second one, then the students who are picked at random will always give you the wrong answer when you double it. That way the class will hopefully have the intuition that Arnie's method only works when the two shows have an equal number of people.

After the third segment we wonder if you could get them ready for the segment about the buses by just asking about buses every 5 minutes, getting them to say we think the average wait is 2.5 minutes. But then just giving an example that raises questions about it. They don't have to come up with a numerical answer, but they sense that maybe 2.5 is not always the right answer. And then we have our segment about the buses.

The fourth segment actually asks you pretty much to replicate what we did in our example in the fourth segment because this is the kind of thing that can be a little elusive and it's good for them to see it more than once.

Then we have our grand finale. If there is time at the end, you could say to them, "Can you imagine a bus schedule where you're luckier if you just missed the bus than if you arrive at a random time?" And we were thinking suppose the schedule is 1:00, 1:02, 1:04, 2:00, 2:02, 2:04. And if you work it out as we do at the website, you'll see that those who just missed the bus on average wait 20 minutes, whereas those who arrive at a random time almost always arrive in this enormous interval and the average wait is more than 20 minutes. So missing a bus is better than just arriving at random in that extreme case.

Ladies and gentlemen what we've offered for what might happen in the breaks that the teachers are responsible for are literally just suggestions. You have more experience with your students and might well be able to come up with ways of presenting the material that are more effective with them. If that's the case, of course we'd be very grateful to learn what you did so that we can improve in future videos. We hope very much that the video will prove stimulating to your classes and are extremely grateful that you rode the bus with us! Thank you.

END OF VIDEO