**Berkson's Paradox in Machine Learning**

Understanding Hidden Biases in Data Analysis

[[Olivier Caelen](https://medium.com/@ocaelen?source=post_page-----113818ac7657--------------------------------)](https://medium.com/@ocaelen?source=post_page-----113818ac7657--------------------------------)

[[Towards Data Science](https://towardsdatascience.com/?source=post_page-----113818ac7657--------------------------------)](https://towardsdatascience.com/?source=post_page-----113818ac7657--------------------------------)

[Olivier Caelen](https://medium.com/@ocaelen?source=post_page-----113818ac7657--------------------------------)

·

[Follow](https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fsubscribe%2Fuser%2Fd7268030c8a8&operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fberksons-paradox-in-machine-learning-113818ac7657&user=Olivier+Caelen&userId=d7268030c8a8&source=post_page-d7268030c8a8----113818ac7657---------------------post_header-----------)

Published in

[Towards Data Science](https://towardsdatascience.com/?source=post_page-----113818ac7657--------------------------------)

·

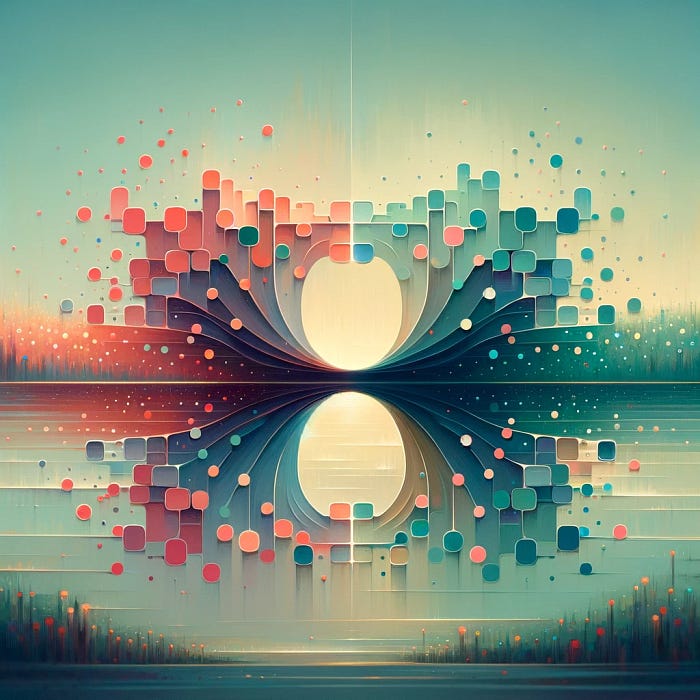
8 min read

·

5 days ago

253

1



Generated with DALL-E

Sometimes, statistics show surprising things that make us question what we see daily. Berkson's Paradox is one example of this. This Paradox is strongly related to the sampling bias problem and occurs when we mistakenly think that two things are related because we don't see the whole picture. As a machine learning expert, you should be familiar with this Paradox because it can significantly impact the accuracy of your predictive models by leading to incorrect assumptions about the relationship between variables.

**Let us start with some examples**

* Based on [Berkson](https://en.wikipedia.org/wiki/Joseph_Berkson)'s original example, let's imagine a retrospective study conducted in a hospital. In this hospital, researchers are studying the risk factors for cholecystitis (a disease of the gallbladder), and one of these risks could be diabetes. Because samples are drawn from a hospitalized population rather than the general population, there is a sampling bias, and this can lead to the mistaken belief that diabetes protects against cholecystitis.
* Another well-known example comes from [Jordan Ellenberg](https://en.wikipedia.org/wiki/Jordan_Ellenberg). In this example, Alex creates a dating pool. This group does not represent all men well; we have a sampling bias because she picks either very friendly men, attractive, or both. And, in Alex's dating pool, something interesting happens… Among the men she dates, it seems like the nicer they are, the less attractive they appear, and vice versa. This sampling bias can lead to a mistaken belief for Alex that there is a negative association between being friendly and attractive.

**Let's try to formalize a bit the problem**

Suppose we have two independent events, *X* and *Y*. As these events are independent:

These random events can be, for example, having the disease cholecystitis or having diabetes, as in the first example, or being nice or beautiful in the second example. Of course, it's important to realize that when I say the two events are independent, I'm talking about the entire population!

In the previous examples, the sampling bias was always of the same type: there were no cases where neither event occurred. In the hospital samples, no patient has neither cholecystitis nor diabetes. And in the Alex sample, no man is both unfriendly and ugly. We are, therefore, conditioned to the realization of at least one of the two events: event *X* has occurred, or event *Y* has occurred, or both. To represent this, we can define a new event, *Z,* which is the union of events *X* and *Y*.

And now, we can write the following to indicate that we are under the sampling bias hypothesis:

That is the probability that event *X* occurs, knowing that events *X* or *Y*(or both)have already been realized. Intuitively, we can feel that this probability is higher than *P(X)* … but it is also possible to show it formally.

To do that, we know that:

By assuming that it is possible for the two events not to occur at the same time (e.g., there are ugly and unfriendly people), the previous statement can become a strict inequality; because the set *(X*∪ *Y)* is not the sample space Ω:

Now, if we divide both sides of this strict inequality by *P(X ∪ Y)* and then multiply by *P(X)*, we get:

where

Therefore, we have indeed that the probability under sampling bias *P(X|Z)* is higher than *P(X),*in the entire population:

Okay, fine … But now let us return to our Berkson's Paradox. We have two independent random variables, *X* and *Y*, and we want to show that they become dependent under the sampling bias *Z* described above.

To do that, let's start with *P(X | Y ∩ Z)*, which is the probability of observing event *X*, given that we know event *Y* has already occurred and that we are under-sampling bias *Z.*Note that *P(X | Y ∩ Z)*can also be written as *P(X | Y, Z).*

As *(Y ∩ Z) = (Y ∩ (X ∪ Y)) = Y,*and as *X* and *Y* are independent variables, we have:

And … finally, knowing that *P(X) < P(X | Z)*, we get what we’re looking for:

This equation shows that under the sampling bias defined by Z, the two initially independent events, *X* and *Y,* become dependent (otherwise, we would have had equality rather than ">").

Go back to the example of Alex's dating pool, if

* Z is the event of being in Alex's dating pool
* X is the event of selecting a friendly guy
* Y is the event of selecting an attractive guy

then *(X | Z)* is the event that Alex meets a nice man and *(X | Y ∩ Z)*is the event that Alex meets a nice man given that he is beautiful. Because of the selection process used to build Alex's dating pool, and because of Berkson's Paradox, Alex will feel that when she meets good-looking boys, they won't be so nice, whereas these could be two independent events if there were taken from the whole population….

**Perhaps a numerical example will help to be more concrete**

To illustrate Berkson's Paradox, we use two dice:

* Event X: The first die shows a 6.
* Event Y: The second die shows either a one or a 2.

These two events are clearly independent, where *P(X)=1/6* and *P(Y)=1/3*.

Now, let's introduce our condition (*Z*), representing the biased sampling by excluding all outcomes where the first die is not six and the second is neither 1 nor 2.

Under our biased sampling condition, we need to calculate the probability that the event *X* occurs, given that at least one of the events (*X* or *Y*) has occurred, and this is denoted by *P(X|Z)*.

First, we need to determine the probability of Z = (X ∪ Y) … and sorry, but from now we'll have to do a bit of calculation… I'll do it for you…. :-)

Next, we calculate the probability of X given Z:

To see if there is a dependence between *X* and *Y* under the assumption that *Z* occurs, we have to compute *P(X | Y ∩ Z).*

As

We have

To demonstrate Berkson's Paradox, we compare *P(X|Z)*with *P*(*X*∣ Y ∩ Z) and we have:

* *P(X | Z) = 0.375*
* *P(X |Y ∩ Z) ≈ 0.1666…*

We retrieve indeed the property that under Berkson's Paradox, due to the sampling bias Z, we have *P(X | Z) > P(X ∣ Y ∩ Z)*.

I personally find it surprising! We had two dice … Two clearly independent random events… and we can obtain the impression that dice rolls become dependent through a sampling process.

**To finish convincing us, let's finish with a bit of simulation**

In the code below, I will simulate dice rolls with Python.

The following code simulates one million experiments of rolling two dice, where for each experiment, it checks if the first dice roll is a 6 (event X) and if the second dice roll is a 1 or 2 (event Y). It then stores the results of these checks (True or False) in the lists X and Y, respectively.

import random  
  
#Get some observations for random variables X and Y  
def sample\_X\_Y(nb\_exp):  
 X = []  
 Y = []  
 for i in range(nb\_exp):  
 dice1 = random.randint(1,6)  
 dice2 = random.randint(1,6)  
 X.append(dice1 == 6)  
 Y.append(dice2 in [1,2])  
 return X, Y  
  
nb\_exp=1\_000\_000  
X, Y = sample\_X\_Y(nb\_exp)

Then, we have to check if these two events are indeed independent. To do that, the following code calculates the probability of event X and the conditional probability of event X given event Y. It does this by dividing the number of successful outcomes by the total number of experiments for each probability.

# compute P(X=1) and P(X1=1|Y=1) to check if X and Y are independent  
p\_X = sum(X)/nb\_exp  
p\_X\_Y = sum([X[i] for i in range(nb\_exp) if Y[i]])/sum(Y)  
  
print("P(X=1) = ", round(p\_X,5))  
print("P(X=1|Y=1) = ", round(p\_X\_Y,5))

P(X=1) = 0.16693  
P(X=1|Y=1) = 0.16681

As we can see, both probabilities are close; therefore (as expected ;-) ) or two dice are independent.

Now, let's see what happens when introducing the sampling bias *Z.* The following code filters the results of the experiments, keeping only those where either X = 1, Y = 1, or both. It stores these filtered results in the lists XZ and YZ.

# keep only the observations where X=1, Y=1 or both (remove when X=0 and Y=0)  
XZ = []  
YZ = []  
for i in range(nb\_exp):  
 if X[i] or Y[i]:  
 XZ.append(X[i])  
 YZ.append(Y[i])   
nb\_obs\_Z = len(XZ)

And now, let's check if these new variables are still independent.

# compute P(X=1|Z=1) and P(X1=1|Y=1,Z=1) to check if X|Z and Y|Z are independent  
p\_X\_Z = sum(XZ)/nb\_obs\_Z  
p\_X\_Y\_Z = sum([XZ[i] for i in range(nb\_obs\_Z) if YZ[i]])/sum(YZ)  
  
print("P(X=1|Z=1) = ", round(p\_X\_Z,5))  
print("P(X=1|Y=1,Z=1) = ", round(p\_X\_Y\_Z,5))

P(X=1|Z=1) = 0.37545  
P(X=1|Y=1,Z=1) = 0.16681

We have an inequality (the same values as in the previous section), meaning that if Z is true, then having information on Y changes the probabilities for X; therefore, they are no longer independent.

**What are the implications of this Paradox for experts in machine learning?**

I don't think experts in machine learning pay enough attention to this type of bias. When we talk about Berkson's Paradox, we're diving into a critical topic for people working in machine learning. This idea is about understanding how we can be misled by the data we use. Berkson's Paradox warns us about the danger of using biased or one-sided data.

**Credit Scoring Systems**: In finance, models trained on data featuring applicants with either high income or high credit scores, but rarely both, could falsely infer a negative correlation between these factors. This risks unfair lending practices by favoring certain demographic groups.

**Social Media Algorithms**: In social media algorithms, Berkson's Paradox can emerge when training models on extreme user data, like viral content with high popularity but low engagement and niche content with deep engagement but low popularity. This biased sampling often leads to the false conclusion that popularity and engagement depth are negatively correlated. Consequently, algorithms may undervalue content that balances moderate popularity and engagement, skewing the content recommendation system.

**Job Applicant Screening Tools**: Screening models based on applicants with either high educational qualifications or extensive experience might incorrectly suggest an inverse relationship between these attributes, potentially overlooking well-balanced candidates.

In each scenario, overlooking Berkson's Paradox can result in biased models, impacting decision-making and fairness. Machine learning experts must counteract this by diversifying data sources and continuously validating models against real-world scenarios.

**Conclusion**

In conclusion, Berkson's Paradox is a critical reminder for machine learning professionals to scrutinize their data sources and avoid misleading correlations. By understanding and accounting for this Paradox, we can build more accurate, fair, and practical models that truly reflect the complexities of the real world. Remember, the key to robust machine learning lies in sophisticated algorithms and the thoughtful, comprehensive collection and analysis of data.

**Thanks for reading!**

Please consider following me if you wish to stay up to date with my latest publications and increase the visibility of this blog.