The proportion of crashes involving vehicle type X, compared to distance travelled by vehicle X.

Henk Stipdonk, May 2020

1. Introduction

Suppose we suspect that a specific type of vehicles is overrepresented in the set of reported road crashes, how can we find out if this is true? Of course, *overrepresented* suggests we need to take into account the presence in traffic of this specific type of vehicle. We could use the proportion of these vehicles in the vehicle fleet as a reference, or the proportion of distance travelled of these vehicles in the total distance travelled, and compare this with the proportion of crashes which involve a vehicle of this specific type.

If we use distance travelled as the reference for our comparison, we actually try to compare *risk* values, i.e. numbers of fatal crashes (or road deaths) divided by distance travelled. This is a common way to compare the safety of e.g. travel modes. In that case we calculate the risk of cycling by dividing deaths by distance cycled, and present the result as the risk of cycling (deaths/km). Such a risk can be compared to the risk of riding a powered two wheeler (P2W), or driving a car. This common way to express risk yields a probability to die in traffic when cycling, riding a P2W or driving etcetera, given a specific distance travelled. In this risk, both single vehicle crashes and two vehicle crashes may be combined. Mark that for two vehicle crashes, this risk does not involve the road deaths in the other vehicle. Also, it does not involve the distance travelled by the other vehicle that may be involved in a crash. It is just the risk that some person becomes a road death while this person is travelling, either walking, cycling, riding a P2W or driving. Bear in mind that this risk therefore increases when there is more traffic of other vehicles. Let us call this risk1

A not so common but conceptually equally valid way to use risk for comparing the safety of travel modes, is by calculating the risk that your travel is associated with *someone else* becoming a road death while you are riding or driving. We might compare the number of road deaths outside our car, truck or van, as compared to distance travelled by our car, truck or van. When calculating this risk, we need the number of road deaths outside our vehicle, and divide by the distance travelled of this vehicle. We do not take into account the road deaths inside the vehicle, or the distance travelled of the other vehicle. Mark that single vehicle crashes that do not involve anyone outside the vehicle play no role here. Again, bear in mind that this risk increases when there are more other vehicles or pedestrians around. Let us call this risk2

There seems to be a third way: why not look at all road deaths in crashes where vehicle x is involved (whether inside the vehicle or outside, in another vehicle or as pedestrians), and divide by distance travelled of the vehicle. It seems as this amounts to the overall risk of there being a death in a crash where a specific vehicle type is involved. But there is a catch, once we start to compare the result with the risk of there being a death in all other crashes. This is because when there are two vehicles involved in a crash, comparison of distance travelled of *one* of these vehicles cause a bias. Let us call this risk3.

This paper explains how this bias, this catch, may perhaps surprisingly yield misleading assessments of accident-proneness of specific types of vehicles (or drivers, for that matter), and how it can be overcome.

2. The research question, based on an example from road safety practice

In ETSC's Pin Flash (PF) 39, the risk (risk3) of there being a road death in a crash involving an HGV is compared to the risk of there being a road death in a crash where no goods vehicles is involved. This risk is expressed as risk3 as described in the introduction. When ETSC carried out the preliminary analyses, it turned out that while HGVs take 5% of all distance travelled, a disproportionate 15% of all road deaths are associated with a crash involving an HGV. This suggests a risk3 ratio of even more than a factor 3, as can be understood from the following reasoning. In this reasoning we assume, for reasons of simplicity, that there are just two kinds of vehicle, HGV and others and that all crashes are two vehicle crashes. In the case of the analysis of HGV in Europe, as described in PF39, this latter assumption must be close to the actual situation:

the PF explains that 14% of all crashes with an HGV involve deaths from inside the HGV. These can have been victims in a crash between an HGV and another kind of vehicle, or in an HGV-HGV crash, or in a single HGV crash. Hence, it is unlikely that many of these 14% were the victim of a single vehicle HGV crash.

The reasoning needed to calculate this risk3 ratio involves the relations between distance travelled of HGV (d_{HGV}) , of other vehicles (d_{other}) and of the total ($d_{total} = d_{HVG} + d_{other}$) where $d_{HGV}=0.05d_{total}$ and $d_{other}=0.95d_{total}$. And further we use the number of fatal crashes where an HGV is involved as n_{HGV}, the number of such crashes where no HGV is involved as n_{other} , and the total as $n_{total}=n_{HGV}+n_{other}$, noting that $n_{HGV}=0.15n_{total}$ and $n_{other}=0.85n_{total}$. With these we can calculate the three values of risk3 for all vehicles, and for the two groups.

- For all vehicles we have $risk3_{total}=n_{total}/d_{total}$.
- For HGV we have $risk3_{HGV}=n_{HGV}/d_{HGV}=0.15/0.05\ risk3_{total}=3\ risk3_{total}$
- For the remaining group we have $risk3_{other}=n_{other}/d_{other}=0.85/0.95\ risk3_{total}=0.89risk3_{total}$.
- Now, the ratio of the two risk3 values yields 3/0.89=3.37.

The question I want to answer in this paper is: to what extent could this result be a consequence of the fact that we mix road deaths in two vehicles with distance travelled in just one vehicle.

To do so I will start with an overly simplified example, in which all vehicles are equally accident prone, drive the same distance annually, are equally often involved in crashes with road deaths outside the vehicle and have the same number of road deaths inside the vehicle.

3. The most simple case.

Consider a country where all trips have equal risk, i.e. the probability of there being a road death in the vehicle while travelling a certain distance is equal for all vehicles, all trips and all travelers. The same holds for the risk to be involved in a crash with a road death outside the vehicle. So all values for risk1 are equal and all values of risk2 are equal. A perfectly homogeneous traffic system (and hence no bicycles, pedestrians, drink driving etcetera). Further: let's assume there are no single vehicle crashes. This assumption is just to make it easier to show where the bias we are going to demonstrate comes from.

Now let's assume that while all vehicles are technically equal, there is just one property in which they differ. Some irrelevant difference, not important for road safety. For example the number plate could start with either an A or a B. Just that. Let's assume that a small proportion p are of type A, and the remaining (1-p) are of type B. For example p could be equal to 10% (i.e. 0.1).

In the following we will express both the number of road deaths and distance travelled as its proportion of the total. Hence, when expressing the average risk in this manner, we find that 100% of crashes divided by 100% of distance travelled leads to a risk of 1, for all vehicles.

Clearly, for risk1 and risk2 nothing amazing happens: for vehicle A, both risk1 and risk2 are also equal to 1: 10% of all road deaths divided by 10% of distance travelled. The same holds for vehicle B: the risk equals 90%/90%=1. But what about risk3? Let's see what happens with our calculation of risk 3. To do that, we need to consider three types of crashes in this country: Crashes between A and A, between B and B, and between A and B.

To start with, let's think about crashes between two vehicles of type A. For such a crash, we need both vehicles to be of type A. As all risks are equal, for each one of the vehicles involved there is a 10% probability that this one vehicle is of type A. This means that there is a 1% probability that an average crash involves *two* vehicles of type A. If you have difficulty imagining this, think of a large bag with balls marked either A or B, of which 10% is marked A and 90% is marked B, and blindly select two balls. What is the probability both are A?). On the other hand, the probability of a crash where both vehicles are of type B is 81% (0.9 times 0.9). Hence we have 1% of crashes with two vehicles A, and 81% with two vehicles B. The remaining 18% of crashes combine a vehicle of type A and one of type B. SO we have 1% AA, 81% BB and 18% AB

This has a far reaching consequence for the calculation of risk3. On the one hand, 99% of all crashes involve at least one vehicle of type B (18%+81%), and 19% of all crashes involve at least one vehicle of type A (1%+18%). So where are the 10% and the 90% gone? If we now calculate risk3 (disregarding units of distance) for vehicle type A, we find risk3 = 19%/10%=1.9. If we calculate the risk for all other crashes and use distance travelled for all other vehicles, we have 81% of the deaths and 90% of the distance travelled, or a risk3 ratio of 81%/90%=0.9. So this way, we find that vehicles of type A are more than twice (1.9/0.9) as dangerous as vehicles of type B. It gets even more astonishing if we change the order of our comparison. If we start with vehicle B, and calculate risk3, we find it is 99%/90% = 1.1. The remaining risk for crashes with no vehicle B involved is 1%/10%=0.1. So now vehicle B is 11 times as dangerous as vehicle A.

4. Can we put this right?

Clearly, it is not fair to first select all crashes with vehicle A involved, calculate risk3, and then apply similar calculations to all other crashes to calculate the risk of vehicle type B. in the first situation we find that vehicle A is more than twice as dangerous as vehicle type B, while in the second situation vehicle type B is 11 times more dangerous than type A.

This effect shows when comparing proportions of crashes, or when comparing risks in a similar way. The cause is that we shouldn't divide the number of casualties in all crashes involving vehicle type A, calculate the risk, and then continue with the remaining group. This remaining group of crashes is disproportionally smaller than the original group of crashes.

Is there a way to do it right, persisting in using risk3? What if we calculate risk3 for both A and B while looking at all crashes that involve one vehicle of each kind twice? The answer is that even that doesn't work: Starting with vehicle A, we find a risk3 of 1.9 (19% of crashes and 10% of distance). If we start with B, we find a risk3 of 1.1 (99% of crashes and 90% of distance). The conclusion should be that using risk3 always leads to bias

In PIN flash 39 we analysed crashes with and without HGV, where we had 5% of distance travelled by HGV and 85% of the crashes not involving an HGV. Now, suppose HGV would be just as dangerous as any other vehicle, *and* suppose all crashes are two vehicle crashes. What would we have expected then?

With 5% of distance travelled by HGV, we would have expected that (0.05^2+2.0.05.0.95 = 9.75% of the crashes involves a HGV, and 90.25% would not. Hence, if HGV were as safe as any other vehicle, with 5% of distance travelled, we would have expected HGV would be involved in almost 10% of the crashes. The risk ratio would have yielded 9.75/5 divided by 90.25/95 = 2.05, slightly more than 2. The observation that HGVs were involved in 15% of the crashes would still indicate a higher risk3 for HGVs than for other vehicles, but not as much higher as the ratio of 3.37 given by the risk3 values.

5. Can we correct for this bias?

If we know all crashes are two vehicle crashes, and we know that vehicle A accounts for a proportion of p of distance travelled, we know we must expect $p^2 + 2p(1-p)$ as the proportion of two vehicle crashes involving at least one vehicle of type A. If p=5%, this means we can expect almost 10% of crashes involving vehicle A. Now, if in fact this proportion is higher (of lower) than this value, we may conclude that vehicle A is more (or less) dangerous than vehicle B. But take care. If we want to compare the result with the risk of vehicle B we should not calculate the risk from the remaining group, as this leads to results that are difficult to interpret.

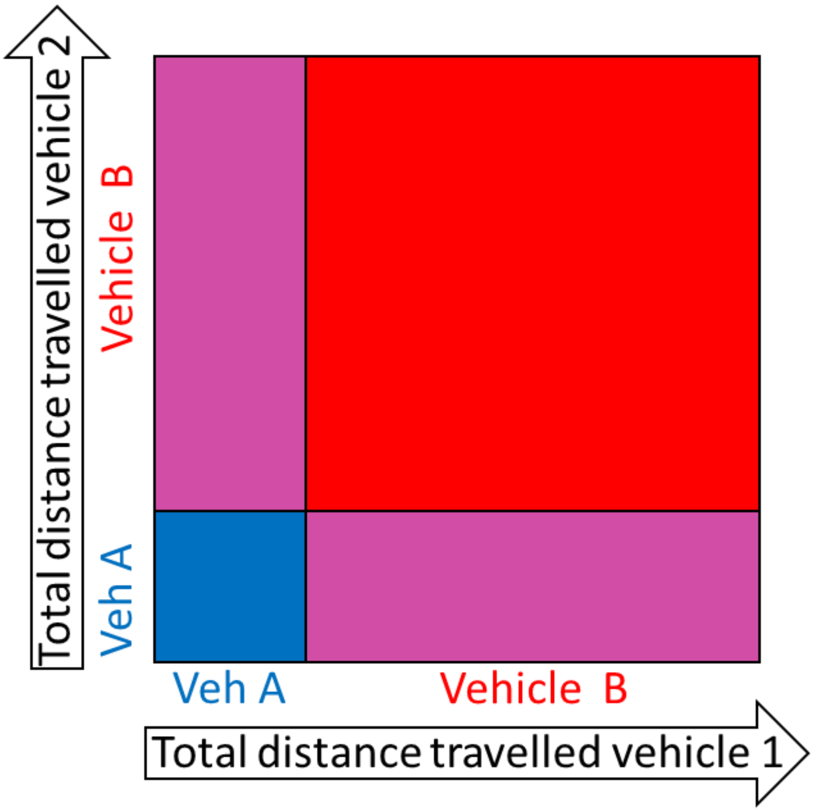
There is a more straightforward way to correct the number of victims in two vehicle crashes for distance travelled, but this involves the product of the distances travelled by the relevant vehicle types. The easiest way to make this clear is by visualization of the influence of distance travelled in two vehicle crashes.

6. Visualization of distance travelled impact in two vehicle crashes.

A way to understand the problem that arises from combining the victims in both vehicles, and attributing them to the distance travelled of one vehicle, is that this yields mixing up the way we select victims and the way we select distance travelled. Let us try to visualize the situation, by showing the crashes in a diagram with axes denoting distance travelled of both vehicles.

We assume that the probability of having a crash with a vehicle is proportional to distance travelled, and the same holds for every other vehicle. Hence, the probability for two vehicles to be in a crash is proportional to both values of distance travelled. Thus, we can visualize the set of two vehicle crashes with two vehicle types A and B, and the expected proportion of crashes that we can associate with combinations of vehicles, by using a two dimensional area, spanned by distance travelled of the two vehicles involved in the crash.

In the figure below, each part stands for a combination of vehicles that can be involved in a crash. The blue square suggests a crash between two vehicles A. The red square corresponds to crashes between two vehicles B. The purple rectangles correspond to crashes with vehicles A and B. In each part of the figure, its area corresponds with the proportion of the number of crashes that can be expected in each of the parts of the graph. If we can assume that all vehicles are equally accident-prone, the density of the number of crashes in each area is homogeneous. Mark that each purple area equals distance travelled of vehicle A times distance travelled by vehicle B.


Total distance travelled vehicle 2
Vehicle B
Veh A
Veh A
Vehicle B
Total distance travelled vehicle 1

Now suppose that we mark every victim in these crashes as a single dot in the appropriate region, where we assume the victim to have travelled in vehicle 1. Mark that in all crashes associated with the blue square, the victim must have been in vehicle A, whereas in the red square, the victims must have been in vehicle B. In the purple area, we define the vehicle the victim was travelling as vehicle 1. In the case there are several victims, we place a dot for every victim, and for every dot the vehicle that the victim was in is chosen as vehicle 1 for that victim.

Crashes between two vehicles A would show as dots in the blue square, crashes between two vehicles B would show as dots in the red square. Crashes between A and B with a victim in vehicle B would show in the lower right purple rectangle for each victim in Vehicle B. Crashes between A and B with the victim in vehicle A would show as dots in the upper left purple rectangle for each victim in vehicle A.

Suppose that vehicle A is more accident prone than vehicle B, this would somehow show in the number of dots in this diagram. Either in the blue, or the purple regions, or both, there would have to be a higher-than-proportional number of dots. Hence, more accident-prone would mean: a higher density of dots. More dots per unit of area means, more dots per product of distances travelled.

Now, if we look at crashes involving at least 1 vehicle A, we see that the area associated with these crashes is the blue square and the two purple rectangles. The area of these regions equals the square of the distance travelled by vehicle A, plus twice the distance travelled by A times distance travelled by B. The result would be a proper denominator to express the relative accident-proneness of vehicle A in traffic with both types of vehicles.

Under the assumptions we have made, a logical denominator for the number of crashes between two vehicles A would be the square of the distance travelled by vehicle A. Something similar holds for the number of crashes between two vehicles B.

7. Conclusion

In this analysis we studied the relations between distance travelled and number of road deaths it two vehicle crashes. We discarded crashes with more than two vehicles, and only briefly mention single vehicle crashes

When comparing the number of road deaths for travelers using a specific travel mode (vehicle, pedestrian) with a number of road deaths for travelers using another travel mode, it is common to take distance travelled into account. We calculate risk, i.e. the number of road deaths using a travel mode, divided by distance travelled by that travel mode. Comparing these risks is straightforward and doesn't cause any bias. These risks, as calculated this way, tell us something about the safety of travelling using a specific travel mode. In the above analysis this risk is denoted as risk1. Mark that we can also apply this definition to single vehicle crashes and calculate the total risk as the risk for single vehicle crashes and the risk for two vehicle crashes.

For some travel modes, the vehicle concerned is much more likely to be involved in a two vehicle crash where the road death is in the other vehicle (or a pedestrian). An example is trucks. Also, some travel modes are more likely than others to be associated with crashes where the road death is someone not using a vehicle – usually a pedestrian. If we want to compare the number of road deaths outside a specific travel mode with the number of road deaths outside another travel mode, we should also take into account distance travelled. In the same way as with risk1, we can do this by defining risk2, as the number of road deaths in crashes involving a certain travel mode, but where the road deaths are outside this travel mode, divided by distance travelled of that travel mode. When we do this for all travel modes, the outcomes show us which travel modes are more dangerous to people other than their users. Again, this type of risk can be applied straightforwardly and doesn't cause any bias.

When we want to combine the two types of road deaths (those inside and outside the vehicle), things get much more complicated. When we want to compare this total number of road deaths associated with a specific travel mode with a similar number for another travel mode, we can't just use distance travelled of the relevant travel mode to correct for the presence of these vehicles in traffic. Instead, we need to take into account the distances travelled of each travel mode involved in specific two vehicle crashes. For crashes involving travel mode A, and another travel mode being either also A or any other vehicle (and let all others be travel mode B), we need to calculate a complex denominator to correct for the presence of these travel modes in traffic. The assumptions made in Section 6 above imply that for crashes between two vehicles A, we need to use the square of distance travelled by travel mode A. For crashes between travel mode A and travel mode B, we need to use twice the product of distance travelled by travel mode A and travel mode B.

When calculating the number of road deaths per distance travelled squared, we actually calculate a risk per distance travelled. This is necessary because here we must realize that being in a crash with a specific type of other vehicle, not only depends on the distance travelled of oneself, but also the presence of the other vehicle. This presence is proportional to the other vehicle's distance travelled (and many other factors).

If you think this gets too complicated, it may be wise to take great care when trying to compare the number of road deaths inside and outside a specific travel mode, with that of another travel mode.

8. Acknowledgement

The author wishes to thank prof. dr Richard Allsop for his tireless efforts to help me remove all errors from this document. Of course, the remaining ones are solely my responsibility.